RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2024, **14**, 33345

A formally exact theory to construct nonreactive forcefields using linear regression to optimize bonded parameters†

Thomas A. Manz *

This article derives theoretical foundations of force field functional theory (FFFT). FFFT studies topics related to the functional representation of nonreactive forcefields to achieve various desirable properties such as: (a) formal exactness of the forcefield's energy functional under certain conditions, (b) a formally exact ansatz separating the bonded potential energy from the nonbonded potential energy within a bonded cluster in a way that enables bonded parameters to be optimized using linear regression instead of requiring nonlinear regression, (c) the potential energy's continuous differentiability to various orders with respect to energetically accessible internal coordinate displacements within a subdomain defined by one electronic ground state, (d) forcefield design that guarantees the reference ground-state geometry is exactly reproduced as an equilibrium structure on the forcefield's potential energy landscape, (e) reasonably accurate and broadly applicable frugal model potentials, (f) computationally efficient embedded feature selection that identifies and removes unimportant forcefield terms, (g) well-designed methods to parameterize the forcefield from quantum-mechanically-computed and (optionally) experimental reference data, and (h) forcefields that approximately reproduce experimentally-measured properties. This article also introduces: (1) an angle-bending model potential that more accurately describes physical dynamics and is continuously differentiable to all orders with respect to internal coordinate displacements even when the bond angle is linear (*i.e.*, $\theta = \pi$ (180°)) and (2) a first-principles-derived stretch potential that accurately describes short-range Pauli repulsion and the long-range bond dissociation energy. This new angle-bending potential gave good agreement to CCSD quantum-chemistry calculations for CaH_2, CO_2, H_2O, HNO, Li_2O, NO_2, NS_2, SF_2, SiH_2, and SO_2 molecules. This new bond-stretch potential reproduced the first 12+ and 30+ vibrational energy levels of H_2 and O_2 molecules, respectively, within a few percent of experimental values. Studying the C–F bond stretch in C_6F_6 as an example, the new ansatz (item (b) above) reduced sensitivity of the optimized force constant's value to choice of nonbonded interaction parameters by an order of magnitude compared to the old ansatz. Normal mode analysis of optimized flexibility models for CO_2, H_2O, HNO, and SO_2 molecules yielded vibrational transition frequencies within a few percent of experimental values. These results demonstrate advantages of this new approach.

Received 11th March 2024
Accepted 23rd September 2024

DOI: 10.1039/d4ra01861c

rsc.li/rsc-advances

1. Introduction

In a nonreactive forcefield, the potential energy is often represented as the sum of bonded interactions and nonbonded interactions:[1–4]

$$U_{\text{total}}^{\text{FF}}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] = U_{\text{bonded}}^{(\text{scheme})}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] + U_{\text{nonbonded}}^{(\text{scheme})}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] \quad (1)$$

Chemical & Materials Engineering, New Mexico State University, Las Cruces, NM 88001, USA. E-mail: tmanz@nmsu.edu

† Electronic supplementary information (ESI) available: A PDF file containing: (a) two supplementary tables for the stretched H_2 molecule, (b) analytic first- and second-order derivatives of the damped nonbonded potential, (c) analytic first- through fourth-order derivatives of the Manz stretch potential, and (d) analytic first- and second-order derivatives of my new angle-bending potential. A zip archive containing: (i) optimized geometries of all molecules studied in this work, (ii) quantum-mechanically-computed and model potential angle-scan curves for ten triatomic molecules, (iii) a spreadsheet containing calculations comparing different bonded and nonbonded interaction models for the C–F stretch in C_6F_6. (iv) Outputs of the calculate_Manz_and_Morse_stretch_potential_exponents program, (v) spreadsheets and Matlab codes that optimized flexibility models for various molecules, and (vi) Matlab codes and outputs for computing the vibrational frequencies of H_2, O_2, CO_2, water, HNO, and SO_2 molecules from the parameterized flexibility models. See DOI: https://doi.org/10.1039/d4ra01861c

The independent variables, $\{\vec{R}_A, \mathbb{Z}_A\}$, define the material's chemical geometry. $\mathbb{Z}_A$ is the element number (aka 'atomic number') of atom A. The position of atom A's nucleus is

$$\vec{R}_A = (X_A, Y_A, Z_A) \tag{2}$$

The bonded interactions include flexibility terms such as bond stretches, angle bends, dihedral torsions, Urey–Bradley terms, cross terms, out-of-plane distances, *etc.* between first, second, third, and/or more distant bonded neighbors.[1,5–7] The nonbonded interactions account for interactions between atoms that are not directly bonded to each other. Nonbonded interactions include: (a) electrostatic interactions modeled by charges, dipoles and other multipoles, and/or polarizabilities, *etc.*, (b) short-range repulsion, (c) long-range dispersion interactions caused by fluctuating multipoles, *etc.*[8–14] The superscript '(scheme)' in eqn (1) reminds us that the partition of $U_{\text{total}}[\{\vec{R}_A, \mathbb{Z}_A\}]$ into bonded and nonbonded interactions depends on the particular scheme chosen to define such a partition.

At the material's equilibrium (aka optimized) ground-state geometry, the net force acting on atom A

$$\vec{F}_A = -\vec{\nabla}_A U_{\text{total}}^{\text{FF}} = \vec{F}_A^{\text{bonded,(scheme)}} + \vec{F}_A^{\text{nonbonded,(scheme)}} \tag{3}$$

is zero

$$\vec{F}_A\left[\left\{\vec{R}_C^{\text{eq}}\right\}\right] = \vec{F}_A^{\text{bonded,(scheme)}}\left[\left\{\vec{R}_C^{\text{eq}}\right\}\right] + \vec{F}_A^{\text{nonbonded,(scheme)}}\left[\left\{\vec{R}_C^{\text{eq}}\right\}\right] = 0 \tag{4}$$

where

$$\vec{F}_A^{\text{bonded,(scheme)}}\left[\left\{\vec{R}_C, \mathbb{Z}_C\right\}\right] = -\vec{\nabla}_A U_{\text{bonded}}^{\text{(scheme)}}\left[\left\{\vec{R}_C, \mathbb{Z}_C\right\}\right] \tag{5}$$

$$\vec{F}_A^{\text{nonbonded,(scheme)}}\left[\left\{\vec{R}_C, \mathbb{Z}_C\right\}\right] = -\vec{\nabla}_A U_{\text{nonbonded}}^{\text{(scheme)}}\left[\left\{\vec{R}_C, \mathbb{Z}_C\right\}\right] \tag{6}$$

The distance (d_{AB}) between atoms A and B is

$$d_{AB} = \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2 + (Z_A - Z_B)^2} \tag{7}$$

and has the equilibrium value d_{AB}^{eq} in the material's optimized ground-state geometry.

A popular strategy (aka the 'old' scheme) is to define the nonbonded potential as a sum of pairwise nonbonded potentials plus optional multibody[15] corrections:

$$U_{\text{nonbonded}}^{\text{old}}\left[\left\{\vec{R}_A, \mathbb{Z}_A\right\}\right] = \sum_{A} \sum_{B \notin \{\text{excluded}_A\}} U_{AB}^{\text{nonbonded}} + \left(U_{\text{multibody}}^{\text{nonbonded}}\right) \tag{8}$$

where $\{\text{excluded}_A\}$ is the set of atoms that are separated from atom A by 0 (*i.e.*, atom A itself), 1, 2, or (optionally) 3 bonds; that is, the set of atom A's zeroth (*i.e.*, atom A itself), first, second, and (optionally) third bonded neighbors.[16,17] As an example, consider a nonbonded potential between two atoms A and B having a simple form involving atomic charges and Lennard-Jones[18] parameters:

$$U_{AB}^{(q+\text{LJ})} = \overbrace{\frac{q_A q_B}{4\pi\varepsilon_0 d_{AB}}}^{\text{atomic charges}} + \underbrace{\varepsilon_{AB}^{\text{LJ}}\left(\left(\frac{d_{AB}^{\text{LJ}}}{d_{AB}}\right)^{12} - 2\left(\frac{d_{AB}^{\text{LJ}}}{d_{AB}}\right)^{6}\right)}_{\substack{\text{repulsion+dispersion} \\ (e.g.,\text{Lennard-Jones})}} \tag{9}$$

$\varepsilon_{AB}^{\text{LJ}}$ is the Lennard-Jones well-depth. d_{AB}^{LJ} is the distance at which the well-depth is reached. More sophisticated examples of $U_{AB}^{\text{nonbonded}}$ may involve terms containing atomic dipoles or quadrupoles, atomic polarizabilities, and/or non-Lennard-Jones van der Waals parameters, *etc.*[14,19–21]

Nonbonded energy schemes defined by eqn (8) and (9) or similar equations encounter a major disadvantage when optimizing the forcefield's bonded parameter values. When using such a scheme, the nonbonded interactions between atoms may exhibit a nonzero force on atom A even in the material's optimized ground-state geometry. This is evident, because atom A is acted upon by electrostatic forces (*e.g.*, Coulomb interactions between atomic charges) and van der Waals (*e.g.*, Lennard-Jones) forces exerted by atoms outside $\{\text{excluded}_A\}$ such as atoms in the same bonded cluster that are separated from atom A by ≥4 bonds but still inside the nonbonded interaction cutoff distance, $d_{\text{cutoff}}^{\text{nonbonded}}$. By eqn (4), this means the bonded interactions must exhibit a net force of equal magnitude in the opposing direction so as to make the total bonded plus nonbonded force acting on each atom zero in the optimized ground-state geometry. Unfortunately, this means the forcefield's flexibility terms such as the harmonic bond stretching potential

$$U_{AB}^{\text{harmonic\_stretch}} = \tfrac{1}{2}k_{AB}(d_{AB} - d_{AB}^{\text{ref}})^2 \tag{10}$$

have 'resting' values that are not necessarily equal to the equilibrium bond length

$$d_{AB}^{\text{ref}} = d_{AB}^{\text{resting}} \neq d_{AB}^{\text{eq}} \tag{11}$$

This has been pointed out in the prior literature by several authors who devised schemes to approximately estimate these resting values.[22–27] Because the resting values of bond lengths, angles, dihedrals, *etc.* enter the forcefield in a nonlinear fashion, this gives rise to a nonlinear optimization problem that may have several local minima.[26]

For the harmonic stretch, it is possible to rewrite eqn (10) as the linear model

$$U_{AB}^{\text{harmonic\_stretch}} = p_1(d_{AB})^2 + p_2 d_{AB} + p_3 \tag{12}$$

where the parameters $p_1 = \frac{1}{2}k_{AB}$, $p_2 = -k_{AB}d_{AB}^{\text{ref}}$, and $p_3 = \frac{1}{2}k_{AB}(d_{AB}^{\text{ref}})^2$. This allows the nonlinear optimization problem to be rewritten as a linear optimization problem. When using the old scheme, this kind of transformation from a nonlinear optimization problem into a linear optimization problem is not always feasible or practical, because it overly restricts the types of flexibility terms that could be included. For example, the old scheme could not be transformed into a linear optimization problem when the potential model includes the MM3 bond stretch[1,28] term or when using $\theta_{ABC}^{\text{resting}} \neq \theta_{ABC}^{\text{eq}}$ as the reference value in some angle-bending potentials (*e.g.*, when using $\theta_{\mathrm{ABC}}^{\mathrm{resting}}$ in place of $\theta_{\mathrm{ABC}}^{\mathrm{eq}}$ in eqn (165)). Consequently, the old scheme is typically a nonlinear optimization problem and only reduces to a linear optimization problem for a restricted set of special cases.

The distinction between linear optimization problems (also called linear regression) and nonlinear optimization problems (also called nonlinear regression) is as follows. In linear regression, all adjustable parameter values to be optimized enter the model linearly as coefficients multiplied by functions of the independent variables.[29] In nonreactive forcefields, the independent variables are the material's internal coordinates (*e.g.*, bond lengths, angles, dihedrals, *etc.*) that describe the material's geometry. These functions of the independent variables may contain fixed (*i.e.*, non-adjustable) parameters. For example, the equilibrium value of the bond angle between atoms A, B, and C ($\theta_{\mathrm{ABC}}^{\mathrm{eq}}$) in the material's optimized ground-state geometry as determined by a quantum chemistry calculation may be treated as a fixed (*i.e.*, non-adjustable) parameter in the model forcefield, because the value of this parameter can be directly computed without requiring regression.

All linear optimization problems are convex. Technically, this means the Lagrangian (*i.e.*, the loss function including Lagrange multiplier terms to enforce constraints (such as bounds) on the optimized parameters) has only a single minimum value not that the optimized parameter values are unique, because some of the material's internal coordinates (and hence model forcefield's bonded terms) may be redundant (*i.e.*, multicollinear, not linearly independent). The resulting degeneracy of optimized parameter values can be suppressed *via* techniques that minimize a norm of the optimized parameters vector.[29–31] Specifically, the least absolute shrinkage and selection operator (LASSO[32,33]) method minimizes the L_1 norm (*i.e.*, the sum of absolute values) of the optimized parameters, while the Moore–Penrose pseudo-inverse[31,34,35] and ridge regression[36] methods minimize the squared L_2 norm (*i.e.*, the sum of squares) of the optimized parameters.

In general, nonlinear optimization problems are more difficult to solve than linear optimization problems.[37,38] Nonlinear optimization problems have more complicated landscapes that may in some cases be nonconvex with multiple local minima.[37,38] As a consequence of this nonlinearity, it may be difficult to determine if the true global optimum of a model forcefield's bonded parameter values have been computed or if the optimizer only reached a local but not global optimum of the model forcefield's bonded parameter values.[38] Some nonlinear optimization problems are provably convex with a single minimum. However, whether a particular nonlinear optimization problem is provably convex must be derived on a case-by-case basis, which requires detailed theoretical analysis.[39] Sometimes it is not easily apparent whether a particular nonlinear optimization problem is convex or not. Furthermore, imposing bounds (or other constraints) on the regressed parameters (*e.g.*, constraining each bond stretch force constant to be non-negative) is generally more challenging for non-linear optimization problems than for linear optimization problems.[39,40] Moreover, multicollinearity arising from internal coordinate redundancy is more difficult to treat during nonlinear regression compared to linear regression. Algorithms for solving nonlinear optimization problems include deterministic methods (*e.g.*, conjugate gradient and steepest descent) to find a local minimum, deterministic global optimizers to find a global minimum, and stochastic methods (*e.g.*, genetic and particle swarm) to find a global minimum.[38–41]

In this article, I introduce a new theoretical framework that transforms the task of optimizing values for bonded parameters (aka flexibility parameters) in nonreactive forcefields from a nonlinear optimization problem into a linear optimization problem. As described in Section 2 below, this is accomplished by introducing a new ansatz for separating the forcefield's potential energy into bonded and nonbonded potential energy terms. My scheme formally decouples the bonded interactions from the nonbonded interactions in a way that zero-, first-, and second-order derivatives of an isolated bonded cluster's potential energy function at its optimized ground-state geometry depend only on the bonded interactions with no dependence on nonbonded interactions. This allows the bonded interaction terms to use equilibrium values directly from the material's quantum-mechanically-computed optimized ground-state geometry instead of separate 'resting values' that would require nonlinear regression. Moreover, this reduces sensitivity of the optimized force constants values appearing in the bonded interaction terms to the particular choice of nonbonded interaction model. Fortunately, this is done in a formally exact way under certain conditions that does not restrict the forcefield from exactly reconstructing the material's true potential energy function.

In practice, a finite cutoff distance for the nonbonded interactions is sometimes used to enhance computational efficiency.[42,43] When using a nonbonded interaction cutoff distance, my approach yields continuous zero-, first-, and second-order derivatives of the potential energy even at the cutoff distance. In contrast, most prior approaches yielded either discontinuous forces or discontinuous second-order derivatives at the cutoff distance.[42–44]

In most practical applications, approximations are introduced by using model forcefields containing a small finite number of terms to maximize computational efficiency at the expense of sacrificing exactness. As shown in Section 3 below, angle-bending model potentials described in prior literature have either derivative discontinuities or incorrect dynamics when the bond angle reaches a value of π radians (180°). In this article, I introduce a new angle-bending model potential that solves this problem and has continuous derivatives of all orders within the physically accessible region of bond angle values. This reduces model uncertainty and more accurately captures physical dynamics while still requiring relatively few force constants to be linearly optimized.

This article is part of a group of articles on the foundations of force field functional theory (FFFT). A companion article introduced the new SAVESTEPS protocol to optimize forcefield bonded parameters (aka 'flexibility parameters') and applied it to a materials dataset containing 116 metal–organic frameworks (MOFs).[45] That automated protocol used my new ansatz for separating bonded from nonbonded interactions and my new angle-bending model potential.[45]

Note: this article adopts the convention that function arguments are enclosed in square brackets, while parentheses denote multiplication. For example, $y[x+2]$ means 'y as a function of $(x+2)$' while $y(x+2)$ means 'y times $(x+2)$'.

2. Formally exact ansatz separating nonbonded from bonded interactions in a nonreactive forcefield

2.1 Foundations

Because an electron's mass is much smaller than an atomic nuclei's mass, electrons typically move much faster than atomic nuclei, and this gives rise to the Born–Oppenheimer approximation in which the electronic motions are approximately equilibrated (approximation # 1) for each geometric arrangement of the material's atomic nuclei.[46] Within the Born–Oppenheimer approximation, a chemical system's total energy $E_{\text{total}}^{\text{Born-Oppenheimer}}$ can be represented as the sum of nuclear kinetic energy and electronic energy, where the ground-state electronic energy $E_{\text{electronic}}^{0}$ is a functional of the chemical geometry:

$$E_{\text{total}}^{\text{Born-Oppenheimer}} = E_{\text{electronic}}^{0}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] + \sum_{\text{A}=1}^{N_{\text{atoms}}} \text{KE}_{\text{A}} \quad (13)$$

where KE_{A} is the nuclear kinetic energy of atom A. Eqn (13) applies whether or not relativistic corrections are included. Eqn (13) also assumes the electrons occupy the electronic ground state for the chemical geometry defined by $\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}$ (approximation # 2). However, eqn (13) allows the atoms to occupy excited and/or ground-state vibrational, rotational, translational states; that is, $\{\text{KE}_{\text{A}}\}$ can be either ground-state and/or excited-state kinetic energies of the atoms.

$E_{\text{electronic}}^{0}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ is the electronic ground-state energy output from an actual quantum chemistry calculation. The exact electronic ground-state energy, $E_{\text{electronic}}^{0,\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$, could conceivably be computed using a full configuration interaction calculation in the complete basis set limit; however, such a calculation may be too computationally expensive in practice.

Within the Born–Oppenheimer approximation, the ground-state electronic energy becomes the potential energy acting on the atomic nuclei.[46] Accordingly, $U_{\text{total}}^{\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ is the (hypothetical) potential energy functional that would formally reproduce the exact electronic energy exactly:

$$U_{\text{total}}^{\text{exact}}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] = E_{\text{electronic}}^{0,\text{exact}}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] \quad (14)$$

Our goal is to choose a forcefield model that has moderate computational costs and approximately reproduces the exact potential energy functional:

$$U_{\text{total}}^{\text{FF}}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] \approx U_{\text{total}}^{\text{exact}}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] \quad (15)$$

Almost all forcefield models are approximations. A formally exact forcefield corresponds to the (hypothetical) case in which the forcefield's potential energy model is $U_{\text{total}}^{\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$.

The total energy in an atomistic simulation parallels that of eqn (13):

$$E_{\text{total}}^{\text{atomistic\_simulation}} = U_{\text{total}}^{\text{FF}}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] + \sum_{\text{A}=1}^{N_{\text{atoms}}} \text{KE}_{\text{A}} \quad (16)$$

A non-reactive forcefield is limited to describing processes in which no existing chemical bonds are broken and no new chemical bonds are formed (approximation # 3), except that some non-reactive potentials (*e.g.*, Morse potential, QMDFF, *etc.*) can reproduce the energy of bond disassociation as the bond length is infinitely stretched.[47,48]

Although not an approximation of the forcefield itself, nonreactive forcefields are most commonly (but not always) used in simulations employing classical Newtonian mechanics. However, it is also possible to use these same forcefields in simulations involving relativistic mechanics (*e.g.*, special relativity) and/or quantum mechanics (*e.g.*, to describe the tunneling of hydrogen atoms during chemical reactions). For example, such forcefields can be used in Feynman path integral simulations (*i.e.*, path integral molecular dynamics, path integral Monte Carlo).[49,50]

Within a specific individual electronic ground state, continuity of the first derivatives of $E_{\text{electronic}}^{0}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ with respect to changes in the atomic positions $\{\vec{R}_{\text{A}}\}$ follows from the Hellmann–Feynman theorem, which states the atom-in-material forces can be computed from the forces the electron cloud exerts on the atomic nuclei.[51] Although I cannot yet provide a rigorous proof that $E_{\text{electronic}}^{0}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ is continuously differentiable to all orders (*i.e.*, 'infinitely differentiable') with respect to changes in the atomic positions $\{\vec{R}_{\text{A}}\}$ within a specific individual electronic ground state, this appears to be true if the system Hamiltonian is sufficiently well-behaved. Discontinuities in first (and/or high-order) derivatives of $E_{\text{electronic}}^{0}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ with respect to changes in the atomic positions $\{\vec{R}_{\text{A}}\}$ can arise where the ground state switches from one electronic ground state to another. Example ground state crossovers include singlet-to-triplet ground-state transitions, conducting to semi-conducting transitions, transitions from one magnetic ground state to another, charge-ordering transitions, transitions that change the crystal symmetry, and so forth.[52–55] Accordingly, $U_{\text{total}}^{\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ is conjectured to be continuously differentiable to all orders with respect to changes in the atomic positions $\{\vec{R}_{\text{A}}\}$ within each subdomain of $\{\vec{R}_{\text{A}}\}$ space that shares the same specific individual electronic ground state, but may exhibit derivative discontinuities at the boundaries where two or more such subdomains intersect. The value of $E_{\text{electronic}}^{0}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ and hence also of $U_{\text{total}}^{\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ varies continuously even at boundaries where the ground state switches from one electronic state to another, because the energy is equal for both electronic phases at the ground-state crossover. Therefore, a formally exact theory for $U_{\text{total}}^{\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ must be general enough to accommodate such behaviors.

To identify the individual bonded clusters in a simulation, we first construct the bond connectivity graph using atom typing radii.[10] Two atoms are classified as directly bonded to each other iff the distance between them is no greater than the sum of their atom typing radii.[10] Two atoms belong to the same bonded cluster iff a connected path of bonds exists between them (for example, if atom A is bonded to B, B is bonded to C, and C is bonded to D, then it follows atoms A and D belong to the same bonded cluster).

Without loss of generality, $U_{\text{total}}^{\text{FF}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ can be expressed as the sum of interactions occurring solely within each bonded cluster (*i.e.*, intracluster) and those involving any interactions between two or more bonded clusters (*i.e.*, intercluster):

$$U_{\text{total}}^{\text{FF}}\left[\left\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\right\}\right] = \sum_{\text{cluster\_}j=1}^{N_\text{clusters}} U_{\text{cluster\_}j}^{\text{intracluster}}\left[\left\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\right\}\right] + U_{\text{nonbonded}}^{\text{intercluster}}\left[\left\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\right\}\right] \quad (17)$$

For example, if a 3-body interaction involves two atoms from the same cluster and a third atom from a different cluster, then it is classified as an intercluster rather than an intracluster interaction.

By definition, 'bonded interactions' can occur only between atoms in the same bonded cluster. In contrast, 'nonbonded interactions' may occur between some atoms in the same bonded cluster and/or between atoms in different clusters. Accordingly, without loss of generality, the bonded and nonbonded interactions are expressed as:

$$U_{\text{bonded}}^{(\text{scheme})}\left[\left\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\right\}\right] = \sum_{\text{cluster\_}j=1}^{N_\text{clusters}} U_{\text{cluster\_}j}^{\text{bonded,(scheme)}}\left[\left\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\right\}\right] \quad (18)$$

$$U_{\text{cluster\_}j}^{\text{bonded,(scheme)}}\left[\left\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\right\}\right] = U_{\text{cluster\_}j}^{\text{intracluster}}\left[\left\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\right\}\right] - U_{\text{cluster\_}j}^{\text{nonbonded,(scheme)}}\left[\left\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\right\}\right] \quad (19)$$

$$U_{\text{nonbonded}}^{(\text{scheme})}\left[\left\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\right\}\right] = \sum_{\text{cluster\_}j=1}^{N_\text{clusters}} U_{\text{cluster\_}j}^{\text{nonbonded,(scheme)}}\left[\left\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\right\}\right] + U_{\text{nonbonded}}^{\text{intercluster}} \quad (20)$$

Eqn (19) allows for some flexibility in how we choose to define $U_{\text{cluster\_}j}^{\text{nonbonded,(scheme)}}$ so long as $U_{\text{cluster\_}j}^{\text{nonbonded,(scheme)}}$ and $U_{\text{cluster\_}j}^{\text{bonded,(scheme)}}$ sum to $U_{\text{cluster\_}j}^{\text{intracluster}}$. As derived below, choosing a specific ansatz to define $U_{\text{cluster\_}j}^{\text{nonbonded,(scheme)}}$ is not arbitrary, because some definitions (*e.g.*, the old scheme, eqn (8) and (9)) require nonlinear regression to optimize the forcefield's bonded parameters while the new scheme defined below is strongly preferred because it allows the forcefield's bonded parameters to be optimized using linear regression. Eqn (18)–(20) apply to both the old scheme and the new scheme; however, the definitions for the individual terms appearing in these equations depend on which scheme is chosen.

The exact intracluster force is given by

$$\vec{F}_\text{A}^{\text{intracluster,exact}}\left[\left\{\vec{R}_\text{C}, \mathbb{Z}_\text{C}\right\}\right] = -\vec{\nabla}_\text{A} E_{\text{cluster\_}j}^{0,\text{exact}}\left[\left\{\vec{R}_\text{C}, \mathbb{Z}_\text{C}\right\}\right] \quad (21)$$

Without loss of generality, the force acting on atom A in the system according to the forcefield model can be expanded as:

$$\vec{F}_\text{A} = \vec{F}_\text{A}^{\text{intracluster}} + \vec{F}_\text{A}^{\text{intercluster}} = -\nabla_\text{A} U_{\text{total}}^{\text{FF}}\left[\left\{\vec{R}_\text{C}, \mathbb{Z}_\text{C}\right\}\right] \quad (22)$$

$$\vec{F}_\text{A}^{\text{intracluster}}\left[\left\{\vec{R}_\text{C}, \mathbb{Z}_\text{C}\right\}\right] = -\vec{\nabla}_\text{A} U_{\text{cluster\_}j}^{\text{intracluster}}\left[\left\{\vec{R}_\text{C}, \mathbb{Z}_\text{C}\right\}\right] \quad (23)$$

$$\vec{F}_\text{A}^{\text{intercluster}}\left[\left\{\vec{R}_\text{C}, \mathbb{Z}_\text{C}\right\}\right] = -\vec{\nabla}_\text{A} U_{\text{nonbonded}}^{\text{intercluster}}\left[\left\{\vec{R}_\text{C}, \mathbb{Z}_\text{C}\right\}\right] \quad (24)$$

$$\vec{F}_\text{A}^{\text{bonded,(scheme)}} = \vec{F}_\text{A}^{\text{intracluster}} - \vec{\nabla}_\text{A} U_{\text{cluster\_}j}^{\text{nonbonded,(scheme)}} \quad (25)$$

where atom_A ∈ cluster_*j*.

2.2 A new ansatz for separating bonded interactions from nonbonded interactions

In the new scheme, the intercluster and intracluster nonbonded interactions are expanded using effective multibody pairwise potentials:

$$U_{\text{cluster\_}j}^{\text{nonbonded,new}}\left[\left\{\vec{R}_\text{C}, \mathbb{Z}_\text{C}\right\}\right] = \sum_{\text{A}\in\text{cluster\_}j} \sum_{\text{B}\in(\text{cluster\_}j-\{\text{excluded}_\text{A}\})} \Phi_{\text{AB}x}^{\text{intracluster}} \quad (26)$$

$$U_{\text{nonbonded}}^{\text{intercluster}}\left[\left\{\vec{R}_\text{C}, \mathbb{Z}_\text{C}\right\}\right] = \sum_\text{A} \sum_{\text{E}\notin\{\text{excluded}_\text{A}\}} \Phi_{\text{AE}x}^{\text{intercluster}} \quad (27)$$

In eqn (26), the summation over B includes all atoms in cluster_*j* except those in {excluded$_\text{A}$}. In eqn (27), the summation over E includes all atoms in the entire system (whether in cluster_*j* or any other cluster) except those in {excluded$_\text{A}$}.

Here, the subscript ABx refers to the nonbonded potential energy assigned to the atom pair AB. The lowercase x indicates this includes the 2-body AB interaction plus the portion of multibody interactions (*i.e.*, n-body interactions for $n \geq 3$) assigned to the AB atom pair. For example, the Axilrod-Teller 3-body dispersion energy[56] could be divided into equal thirds assigned to the pairs ABx, ACx, and BCx. If atoms A, B, and C belong to the same bonded cluster, then the 3-body ABC interaction is partitioned into contributions that go towards $\Phi_{\text{AB}x}^{\text{intracluster}}$, $\Phi_{\text{AC}x}^{\text{intracluster}}$, and $\Phi_{\text{BC}x}^{\text{intracluster}}$. If atoms A, B, and/or C belong to different bonded clusters, then the 3-body ABC interaction is partitioned into contributions that go towards $\Phi_{\text{AB}x}^{\text{intercluster}}$, $\Phi_{\text{AC}x}^{\text{intercluster}}$, and $\Phi_{\text{BC}x}^{\text{intercluster}}$.

We define $\Phi_{\text{AB}x}^{\text{intracluster}}[\{\vec{R}_\text{C}\}]$ as the two-body nonbonded interaction energy between atoms A and B belonging to the same bonded cluster plus the portion of intracluster multibody nonbonded interaction energy assigned to the AB pair. If atoms A and D belong to different bonded clusters, then $\Phi_{\text{AD}x}^{\text{intracluster}}[\{\vec{R}_\text{C}\}] = 0$.

In eqn (27), atom E may belong either to the same or a different bonded cluster as atom A. We define $\Phi_{\text{AE}x}^{\text{intercluster}}[\{\vec{R}_\text{C}\}]$ as the portion of the intercluster nonbonded energy, $U_{\text{nonbonded}}^{\text{intercluster}}[\{\vec{R}_\text{C}, \mathbb{Z}_\text{C}\}]$, that is assigned to the AE pair. $\Phi_{\text{AE}x}^{\text{intercluster}}[\{\vec{R}_\text{C}\}]$ includes intercluster interaction energies of all orders $n \geq 2$. For atoms A and B belonging to the same bonded cluster, the two-body (*i.e.*, $n = 2$) AB interaction counts exclusively towards $\Phi_{\text{AB}x}^{\text{intracluster}}$ and not towards $\Phi_{\text{AB}x}^{\text{intercluster}}$. If atoms A and D belong to different bonded clusters, then the two-body (*i.e.*, $n = 2$) AD interaction counts towards $\Phi_{\text{AD}x}^{\text{intercluster}}$.

For example, a system containing three water molecules, one carbon dioxide molecule, plus one MOF has five bonded clusters. In this case, cluster_1 is the first water molecule, cluster_2 is the second water molecule, cluster_3 is the third water molecule, cluster_4 is the carbon dioxide molecule, and cluster_5 is the MOF. In this case, $\{\vec{R}_C^{\text{eq\_1}}\}$, $\{\vec{R}_C^{\text{eq\_2}}\}$, and $\{\vec{R}_C^{\text{eq\_3}}\}$ are the optimized geometry of an isolated water molecule; $\{\vec{R}_C^{\text{eq\_4}}\}$ is the optimized geometry of an isolated carbon dioxide molecule; and $\{\vec{R}_C^{\text{eq\_5}}\}$ is the optimized geometry of the bare MOF (*i.e.*, the MOF containing no adsorbate molecules). For crystals, either experimentally-measured or theoretically-computed lattice constants (*i.e.*, unit cell lengths a, b, c and unit cell angles α, β, γ) can be used to define the unit cell's shape and volume when computing $\{\vec{R}_C^{\text{eq\_}j}\}$; normally, we use whichever data source is more accurate.

The new scheme is designed so that the zero-, first-, and second-order derivatives of the intracluster nonbonded potential energy are zero by construction within the equilibrium ('optimized') ground-state geometry of each isolated bonded cluster j:

$$U^{\text{nonbonded,new}}_{\text{cluster\_}j}\left[\left\{\vec{R}_C \Rightarrow \vec{R}_C^{\text{eq\_}j}\right\}\right] = 0 \tag{28}$$

$$-\vec{\nabla}_A U^{\text{nonbonded,new}}_{\text{cluster\_}j}\left[\left\{\vec{R}_C \Rightarrow \vec{R}_C^{\text{eq\_}j}\right\}\right] = 0 \tag{29}$$

$$\vec{\nabla}_{\text{atom\_1}}\vec{\nabla}_{\text{atom\_2}} U^{\text{nonbonded,new}}_{\text{cluster\_}j}\left[\left\{\vec{R}_C \Rightarrow \vec{R}_C^{\text{eq\_}j}\right\}\right] = 0 \tag{30}$$

where $\{\vec{R}_C^{\text{eq\_}j}\}$ are the optimized atom-in-material coordinates in the isolated bonded cluster j. Here, the notation $\{\vec{R}_C \Rightarrow \vec{R}_C^{\text{eq\_}j}\}$ means the subset of atoms from the full system which are contained within cluster_j are positioned where they would be located within the lowest energy configuration of the isolated bonded cluster j. $U^{\text{nonbonded,new}}_{\text{cluster\_}j}[\{\vec{R}_C, \mathbb{Z}_C\}]$ makes no contribution to (*i.e.*, does not affect) the energy (eqn (28)), atom-in-material forces (eqn (29)), and Hessian matrix (*i.e.*, matrix of second derivatives with respect to atomic displacements, eqn (30)) at the optimized geometry of isolated cluster_j.

Many classical molecular dynamics and Monte Carlo software packages have the option to use truncated and shifted nonbonded potentials that go to zero whenever the distance between two atoms A and B is greater than or equal to a nonbonded interaction cutoff distance, $d_{\text{cutoff}}^{\text{nonbonded}}$.[42,43] In general, such a 'bare' shifted nonbonded potential has discontinuous forces at the cutoff distance.[43] To avoid this discontinuity, both the forces and potential can be truncated and shifted at the cutoff distance; however, this still yields discontinuous second derivatives at the cutoff distance.[43]

The new scheme is designed so that the zero-, first-, and second-order derivatives of the potential energy are continuous even at the nonbonded interaction cutoff distance, and this also ensures that the atom-in-material forces and their first-order derivatives are continuous. Specifically, we will design the new scheme so that the following constraints hold for all systems

$$\lim_{d_{AB}\to d_{\text{cutoff}}^{\text{nonbonded}}} \Phi^{\text{intracluster}}_{ABx}\left[\left\{\vec{R}_C\right\}\right] = 0 \tag{31}$$

$$\lim_{d_{AB}\to d_{\text{cutoff}}^{\text{nonbonded}}} \vec{\nabla}_{\text{atom\_1}} \Phi^{\text{intracluster}}_{ABx}\left[\left\{\vec{R}_C\right\}\right] = 0 \tag{32}$$

$$\lim_{d_{AB}\to d_{\text{cutoff}}^{\text{nonbonded}}} \vec{\nabla}_{\text{atom\_1}}\vec{\nabla}_{\text{atom\_2}} \Phi^{\text{intracluster}}_{ABx}\left[\left\{\vec{R}_C\right\}\right] = 0 \tag{33}$$

$$\lim_{d_{AE}\to d_{\text{cutoff}}^{\text{nonbonded}}} \Phi^{\text{intercluster}}_{AEx}\left[\left\{\vec{R}_C\right\}\right] = 0 \tag{34}$$

$$\lim_{d_{AE}\to d_{\text{cutoff}}^{\text{nonbonded}}} \vec{\nabla}_{\text{atom\_1}} \Phi^{\text{intercluster}}_{AEx}\left[\left\{\vec{R}_C\right\}\right] = 0 \tag{35}$$

$$\lim_{d_{AE}\to d_{\text{cutoff}}^{\text{nonbonded}}} \vec{\nabla}_{\text{atom\_1}}\vec{\nabla}_{\text{atom\_2}} \Phi^{\text{intercluster}}_{AEx}\left[\left\{\vec{R}_C\right\}\right] = 0 \tag{36}$$

In eqn (32)–(36), atom_1 and atom_2 can be either atom A, B, E, or any other atom. Moreover, atom_1 and atom_2 can be either the same or different atoms.

To accomplish this, we first define a simple parameter-free smooth transition function $\tau_{AB}[s, t]$ that satisfies the following conditions:

(1) $\tau_{AB}[s, t]$ should satisfy the bound

$$-1 \le \tau_{AB}[s, t] \le 1 \tag{37}$$

(2) $\tau_{AB}[s, t]$ should smoothly turn on when s largely differs from t, while remaining mostly turned off when $s \approx t$. Specifically,

$$\lim_{s\to t} \tau_{AB}[s, t] = 0 \tag{38}$$

$$\lim_{\min[(s/t),(t/s)]\to 0} (\tau_{AB}[s, t])^2 = 1 \tag{39}$$

(3) $\tau_{AB}[s, t]$ should be independent of the choice of measurement units, because its value is a function of the dimensionless ratio t/s.

(4) $\tau_{AB}[s, t]$ should increase monotonically as the ratio t/s increases:

$$\frac{d\tau_{AB}[s, t]}{d[t/s]} \ge 0 \tag{40}$$

(5) By using various powers of $\tau_{AB}[s, t]$, the higher-order derivatives expand as follows

$$\lim_{s\to t} \frac{d[\tau_{AB}{}^2[s, t]]}{ds} = \lim_{s\to t} 2\tau_{AB}[s, t]\frac{d[\tau_{AB}[s, t]]}{ds} = 0 \tag{41}$$

$$\lim_{s\to t} \frac{d[\tau_{AB}{}^3[s, t]]}{ds} = \lim_{s\to t} 3\tau_{AB}{}^2[s, t]\frac{d[\tau_{AB}[s, t]]}{ds} = 0 \tag{42}$$

$$\lim_{s\to t} \frac{d^2[\tau_{AB}{}^3[s, t]]}{ds^2} = \lim_{s\to t}\left(3\tau_{AB}[s, t]\left(2\left(\frac{d[\tau_{AB}[s, t]]}{ds}\right)^2 + \tau_{AB}[s, t]\frac{d^2[\tau_{AB}[s, t]]}{ds^2}\right)\right) = 0 \tag{43}$$

where use has been made of eqn (38).

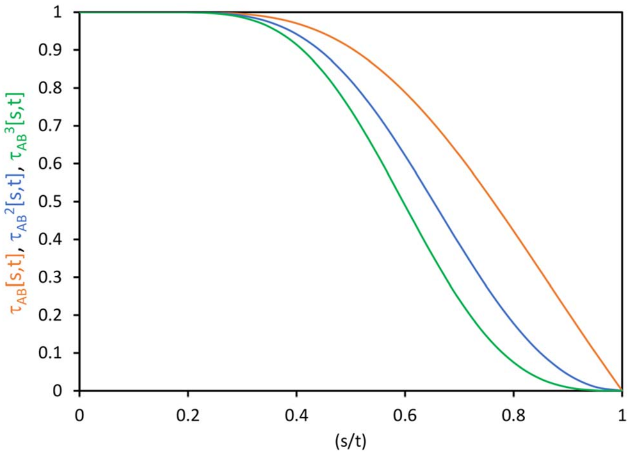

Fig. 1 Plot of the smooth transition function $\tau_{AB}[s, t]$ for the nonbonded potential. The square and cube of this function are also plotted.

(6) To achieve a balance between τ_{AB} increasing neither too quickly nor too slowly, $\tau_{AB}[s, t]$ should be mostly turned on (*i.e.*, $(\tau_{AB})^2 \approx 0.8$) when $t/s = 2$.

The following specific choice of smooth transition function

$$\tau_{AB}[s, t] = \tanh\left[\frac{t}{s} - \frac{s}{t}\right] \tag{44}$$

has a simple form satisfying all of the above conditions. When $t/s = 2$, $\tau_{AB}{}^2\ [s,\ t] = \tanh^2[2 - \frac{1}{2}] = 0.819\ldots$. This strikes a compromise between τ_{AB} increasing neither too quickly nor too slowly as t/s increases. Fig. 1 plots this smooth transition function.

Using this smooth transition function, we can now arrange the nonbonded interactions according to four cases to satisfy eqn (28)–(36) above. Case # 1: the two atoms A and B are inside the same bonded cluster j and a cutoff distance is used for their nonbonded interaction. In this case, we express the effective multibody pairwise potentials as follows:

$$\begin{aligned}\Phi_{ABx}^{\text{intercluster}} = {} & \Theta_H\left[d_{\text{cutoff}}^{\text{nonbonded}} - d_{AB}\right]\tau_{AB}{}^3\left[d_{AB}, d_{\text{cutoff}}^{\text{nonbonded}}\right] \\ & \tau_{AB}{}^2\left[d_{AB}, d_{AB}{}^{\text{eq\_}j}\right]\left(U_{ABx,\text{intercluster}}^{\text{nonbonded}}\left[\left\{\vec{R}_C\right\}\right]\right. \\ & \left.-U_{ABx,\text{intercluster}}^{\text{nonbonded}}\left[\left\{\vec{R}_C{}^{\text{eq\_}j}\right\}\right]\right)\end{aligned} \tag{45}$$

$$\begin{aligned}\Phi_{ABx}^{\text{intracluster}} = {} & \Theta_H\left[d_{\text{cutoff}}^{\text{nonbonded}} - d_{AB}\right]\tau_{AB}{}^3\left[d_{AB}, d_{\text{cutoff}}^{\text{nonbonded}}\right] \\ & \tau_{AB}{}^2\left[d_{AB}, d_{AB}{}^{\text{eq\_}j}\right]\left(U_{ABx,\text{intracluster}}^{\text{nonbonded}}\left[\left\{\vec{R}_C\right\}\right]\right. \\ & \left.-U_{ABx,\text{intracluster}}^{\text{nonbonded}}\left[\left\{\vec{R}_C{}^{\text{eq\_}j}\right\}\right]\right)\end{aligned} \tag{46}$$

Θ_H is the Heaviside step function, and $d_{AB}{}^{\text{eq\_}j}$ is the equilibrium distance between atoms A and B in the isolated bonded cluster j.

Case # 2: the two atoms A and B are inside the same bonded cluster j and a cutoff distance is not used for their nonbonded interaction. In this case, we express the effective multibody pairwise potentials as follows:

$$\begin{aligned}\Phi_{ABx}^{\text{intercluster}} = {} & \tau_{AB}{}^2\left[d_{AB}, d_{AB}{}^{\text{eq\_}j}\right]\left(U_{ABx,\text{intercluster}}^{\text{nonbonded}}\left[\left\{\vec{R}_C\right\}\right]\right. \\ & \left.- U_{ABx,\text{intercluster}}^{\text{nonbonded}}\left[\left\{\vec{R}_C{}^{\text{eq\_}j}\right\}\right]\right)\end{aligned} \tag{47}$$

$$\begin{aligned}\Phi_{ABx}^{\text{intracluster}} = {} & \tau_{AB}{}^2\left[d_{AB}, d_{AB}{}^{\text{eq\_}j}\right]\left(U_{ABx,\text{intracluster}}^{\text{nonbonded}}\left[\left\{\vec{R}_C\right\}\right]\right. \\ & \left.- U_{ABx,\text{intracluster}}^{\text{nonbonded}}\left[\left\{\vec{R}_C{}^{\text{eq\_}j}\right\}\right]\right)\end{aligned} \tag{48}$$

Case # 3: the two atoms A and D are not inside the same bonded cluster and a cutoff distance is used for their nonbonded interaction. In this case, we express the effective multibody pairwise potentials as follows:

$$\begin{aligned}\Phi_{ADx}^{\text{intercluster}} = {} & \Theta_H\left[d_{\text{cutoff}}^{\text{nonbonded}} - d_{AB}\right]\tau_{AB}{}^3\left[d_{AB}, d_{\text{cutoff}}^{\text{nonbonded}}\right] \\ & U_{ABx,\text{intercluster}}^{\text{nonbonded}}\left[\left\{\vec{R}_C\right\}\right]\end{aligned} \tag{49}$$

$$\Phi_{ADx}^{\text{intracluster}} = 0 \tag{50}$$

Case # 4: the two atoms A and D are not inside the same bonded cluster and a cutoff distance is not used for their nonbonded interaction. In this case, we express the effective multibody pairwise potentials as follows:

$$\Phi_{ADx}^{\text{intercluster}} = U_{ABx,\text{intercluster}}^{\text{nonbonded}}\left[\left\{\vec{R}_C\right\}\right] \tag{51}$$

$$\Phi_{ADx}^{\text{intracluster}} = 0 \tag{52}$$

Analytic first- and second-order derivatives of the nonbonded interactions for these four cases are shown in ESI Section S2.†

If the cutoff distance used for the nonbonded potential is infinite (*i.e.*, $d_{\text{cutoff}}^{\text{nonbonded}} \to \infty$), then

$$\begin{aligned}\lim_{d_{\text{cutoff}}^{\text{nonbonded}} \to \infty} & \left(\Theta_H\left[d_{\text{cutoff}}^{\text{nonbonded}} - d_{AB}\right]\tanh^3\left[\frac{d_{\text{cutoff}}^{\text{nonbonded}}}{d_{AB}} - \frac{d_{AB}}{d_{\text{cutoff}}^{\text{nonbonded}}}\right]\right) \\ & = 1\end{aligned} \tag{53}$$

Accordingly, Case # 2 can be regarded as the $d_{\text{cutoff}}^{\text{nonbonded}} \to \infty$ limit of Case # 1, and Case # 4 can be regarded as the $d_{\text{cutoff}}^{\text{nonbonded}} \to \infty$ limit of Case # 3.

According to these new definitions, $U_{\text{nonbonded}}^{\text{intercluster}}$ and $U_{\text{cluster\_}j}^{\text{nonbonded,new}}$ have continuous values and continuous first and second derivatives everywhere with respect to atom displacements. This should provide improved numeric precision when performing classical molecular dynamics and Monte Carlo simulations using a nonbonded interaction cutoff distance.

By convention, the nonbonded interaction energy goes to zero for two atoms infinitely far apart:

$$\lim_{d_{AB} \to \infty} U_{ABx,\text{intercluster}}^{\text{nonbonded}} = 0 \tag{54}$$

$$\lim_{d_{AB} \to \infty} U_{ABx,\text{intracluster}}^{\text{nonbonded}} = 0 \tag{55}$$

Specific models for $U^{\text{nonbonded}}_{\text{AB}x,\text{ intercluster}}$ and $U^{\text{nonbonded}}_{\text{AB}x,\text{ intracluster}}$ can include nonbonded interactions due to some or all of the following: atomic charges, atomic dipoles, atomic quadrupoles, atom-in-material polarizabilities, short-range repulsion, long-range dispersion interactions with short-range damping, *etc.*[8–12,14,19,56–63] A Lennard-Jones plus atomic charges model, $U^{\text{nonbonded}}_{\text{AB}x}[\{\vec{R}_{\text{C}}\}] \approx U_{\text{AB}}{}^{(q+\text{LJ})}$, or more sophisticated (and hopefully more accurate) models could be used for the nonbonded interactions. These are mentioned only as examples, because the possibilities are delineated only by the capacities of human ingenuity.

For an isolated bonded cluster j, the force on atom A is

$$\vec{F}_{\text{A}}{}^{\text{cluster\_}j} = -\vec{\nabla}_{\text{A}} U_{\text{total}}{}^{\text{cluster\_}j} = -\vec{\nabla}_{\text{A}} U^{\text{bonded,new}}_{\text{cluster\_}j} - \vec{\nabla}_{\text{A}} U^{\text{nonbonded,new}}_{\text{cluster\_}j} \quad (56)$$

At the optimized geometry of this isolated bonded cluster, the force on each atom in the cluster is zero:

$$\vec{F}_{\text{A}}{}^{\text{cluster\_}j}\left[\left\{\vec{R}_{\text{C}} \Rightarrow \vec{R}_{\text{C}}{}^{\text{eq\_}j}\right\}\right] = 0 \quad (57)$$

Substituting eqn (29) and (56) into (57) gives

$$\vec{F}_{\text{A}}^{\text{bonded,new}}\left[\left\{\vec{R}_{\text{C}} \Rightarrow \vec{R}_{\text{C}}{}^{\text{eq\_}j}\right\}\right] = -\vec{\nabla}_{\text{A}} U^{\text{bonded,new}}_{\text{cluster\_}j}\left[\left\{\vec{R}_{\text{C}} \Rightarrow \vec{R}_{\text{C}}^{\text{eq\_}j}\right\}\right] = 0 \quad (58)$$

Eqn (58) is the key result that enables the new scheme to directly use the equilibrium geometric parameters from isolated cluster_j's quantum-mechanically-computed optimized ground-state geometry as the 'resting values' in the bonded interaction terms. This enables the new scheme to use linear regression instead of requiring nonlinear regression to optimized cluster_j's bonded parameters.

Consider a system that contains only an isolated bonded cluster_j. In this case, the quantum-mechanically-computed optimized ground-state geometry $\{\vec{R}_{\text{C}}{}^{\text{eq\_}j}\}$ is always an equilibrium structure of the constructed forcefield; that is, the atom-in-material forces are zero as shown in eqn (57). With proper forcefield parameterization, the quantum-mechanically-computed optimized ground-state geometry $\{\vec{R}_{\text{C}}{}^{\text{eq\_}j}\}$ should preferably be at least a local energy minimum and more preferably a global energy minimum in the forcefield's potential energy surface for this isolated bonded cluster_j. In other words, an isolated bonded cluster's (*e.g.*, a molecule's or a MOF's) optimized ground-state geometry can still be predicted exactly by the forcefield even if the forcefield's potential energy function is an approximation! Near $\{\vec{R}_{\text{C}}{}^{\text{eq\_}j}\}$, eqn (28)–(30) show $U^{\text{nonbonded,new}}_{\text{cluster\_}j}$ only affects the third- and higher-order derivatives that control anharmonicity. This enables the forcefield's bonded force constant values (at least within the subdomain containing the cluster's optimized ground-state geometry) to be optimized to good approximation without requiring specific models for $U^{\text{nonbonded}}_{\text{AB}x,\text{intracluster}}[\{\vec{R}_{\text{C}}\}]$ or $U^{\text{nonbonded}}_{\text{AB}x,\text{intercluster}}[\{\vec{R}_{\text{C}}\}]$ to be chosen ahead of time. This facilitates directly comparing forcefields using different nonbonded interaction models without having to reoptimize the bonded parameters. (Since it is formally exact under the conditions described in Section 2.5 below, this new scheme can certainly also be used to precisely describe the anharmonicity; however, in that case a change in $U^{\text{nonbonded}}_{\text{AB}x,\text{intracluster}}[\{\vec{R}_{\text{C}}\}]$ requires also adjusting the bonded parameters to maintain an accurate description of the anharmonicity.)

The old scheme is more complicated than the new scheme, because the old scheme does not satisfy eqn (58). Consequently, the 'resting values' in the bonded interaction terms of the old scheme do not equal the equilibrium geometric parameters from isolated cluster_j's quantum-mechanically-computed optimized ground-state geometry. For example, a bond stretch under the old scheme could be constructed using the MM3 bond stretch potential, but this gives an optimization problem nonlinear in the parameter $d^{\text{resting}}_{\text{AB}}$. In stark contrast, the new scheme does not treat $d^{\text{resting}}_{\text{AB}}$ as an unknown to solve for during regression, because under the new scheme $d^{\text{resting}}_{\text{AB}} = d_{\text{AB}}{}^{\text{eq\_}j}$ has a known value before regression.

An analogy helps explain relationships between the old scheme and the new scheme. Suppose that we have a system composed of two fruit pies: one cherry pie and one apple pie. Each of these pies is analogous to the energy of a different bonded cluster in our system. Hypothetically, we could cut each pie into several pieces. By itself, this cutting operation does not introduce any approximations. For example, if we cut the cherry pie into two pieces, this does not introduce any approximations, because these two pieces still sum up to the entire pie. We notice that there are different ways we could choose to cut up the cherry pie. For example, we could cut the cherry pie into a left-side piece (called 'intracluster bonded interactions') and into a right-side piece (called 'intracluster nonbonded interactions'). The distinction between the old scheme and the new scheme is that they are different protocols for cutting the cherry pie into pieces. Although this choice affords some flexibility, it is not completely arbitrary, because some protocols (*e.g.*, the new scheme) for separating intracluster bonded interactions from intracluster nonbonded interactions yield a linear regression problem for the bonded parameters while some other separation protocols (*e.g.*, the old scheme) yield a nonlinear regression problem for the bonded parameters. While we get to choose how to cut up the cherry pie into pieces, the physical separation between the cherry pie and the apple pie, which is analogous to the 'intercluster nonbonded interactions', is defined by nature rather than being chosen by us. Suppose that we have a system comprised of several bonded clusters. In this case, the 'intercluster nonbonded potential energy' is the difference between the Born–Oppenheimer electronic energy of the total system and that of the isolated clusters:

$$E^{0}_{\text{intercluster}}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] = E^{0}_{\text{electronic}}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] - \sum_{\text{cluster\_}j=1}^{N_{\text{clusters}}} E^{0,\text{electronic}}_{\text{isolated\_cluster\_}j}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] \quad (59)$$

Each term in eqn (59) is physically defined in a non-subjective manner.

2.3 Expanding the bonded interactions

Because of eqn (30) and (58), we can construct $U_{\text{cluster\_}j}^{\text{bonded,new}}$ such that leading terms in its series expansion in the vicinity of $\{\vec{R}_{\text{A}} \Rightarrow \vec{R}_{\text{A}}^{\text{eq\_}j}\}$ are second-order in internal coordinate displacements:

$$U_{\text{cluster\_}j}^{\text{bonded,new}}\left[\left\{\vec{R}_{\text{A}}\right\}\right] - U_{\text{cluster\_}j}^{\text{bonded,new}}\left[\left\{\vec{R}_{\text{A}}^{\text{eq\_}j}\right\}\right] \approx \sum_h \sum_i \eta_{hi}\left(\alpha_h - \alpha_h^{\text{eq\_}j}\right)\left(\alpha_i - \alpha_i^{\text{eq\_}j}\right) + \text{h.o.t.} \quad (60)$$

where 'h.o.t.' are higher order terms. Here, α_i is an internal coordinate, and $\alpha_i^{\text{eq\_}j}$ is its equilibrium value in the quantum-mechanically-computed ground-state geometry of isolated cluster_*j*. η_{hi} is the corresponding expansion coefficient. Eqn (60) is not a Taylor series expansion, because the set of internal coordinates is partly redundant.

Alternatively, we can expand $U_{\text{cluster\_}j}^{\text{bonded,new}}$ as a linear combination of flexibility terms, such that each flexibility term has a Taylor series expansion whose leading term is second-order in internal coordinate displacements:

$$U_{\text{cluster\_}j}^{\text{bonded,new}}\left[\left\{\vec{R}_{\text{A}}\right\}\right] - U_{\text{cluster\_}j}^{\text{bonded,new}}\left[\left\{\vec{R}_{\text{A}}^{\text{eq\_}j}\right\}\right] \approx \sum_\gamma k_\gamma^{j,1} g_\gamma^{j,1}\left[\{\alpha_i\}, \left\{\left(\alpha_i - \alpha_i^{\text{eq\_}j}\right)\right\}\right] \quad (61)$$

$$g_\gamma^{j,1}\left[\{\alpha_i\}, \left\{\left(\alpha_i - \alpha_i^{\text{eq\_}j}\right)\right\}\right] = 1/2\sum_h \sum_i \left.\frac{\partial^2 g_\gamma^{j,1}}{\partial\alpha_h \partial\alpha_i}\right|_{\left\{\vec{R}_{\text{A}}^{\text{eq\_}j}\right\}} \left(\alpha_h - \alpha_h^{\text{eq\_}j}\right)\left(\alpha_i - \alpha_i^{\text{eq\_}j}\right) + \text{h.o.t.} \quad (62)$$

where $k_\gamma^{j,1}$ is the force constant and 'h.o.t.' are higher order terms. Eqn (62) is a Taylor series expansion, because the internal coordinates contributing to a single flexibility term are independent of each other (*i.e.*, non-redundant).

Owing to their continuous differentiability with respect to changes in the internal coordinates, the expansions shown in eqn (61) and (62) are only valid within the subdomain of $\{\vec{R}_{\text{A}}^{(j)}\}$ space that share the same electronic ground state type as $\{\vec{R}_{\text{A}}^{\text{eq\_}j}\}$. $\{\vec{R}_{\text{A}}^{(j)}\}$ is the set of atom-in-material coordinates for only those atoms contained in cluster_*j*. We can construct a formally exact expansion of $U_{\text{cluster\_}j}^{\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ that is globally valid over all accessible regions of $\{\vec{R}_{\text{A}}^{(j)}\}$ space by concatenating expansions for the various subdomains describing different electronic ground states:

$$U_{\text{cluster\_}j}^{\text{bonded,new}}\left[\left\{\vec{R}_{\text{A}}\right\}\right] - U_{\text{cluster\_}j}^{\text{bonded,new}}\left[\left\{\vec{R}_{\text{A}}^{\text{eq\_}j}\right\}\right] = \sum_{p=1}^{N_j^{\text{domains}}}\left(\Lambda_{j,p}\left[\left\{\vec{R}_{\text{A}}^{(j)}\right\}\right]\sum_\gamma k_\gamma^{j,p} g_\gamma^{j,p}\left[\{\alpha_i\}, \left\{\left(\alpha_i - \alpha_i^{\text{eq\_}j}\right)\right\}\right]\right) \quad (63)$$

$$\Lambda_{j,p}\left[\left\{\vec{R}_{\text{A}}^{(j)}\right\}\right] = \begin{cases} 1 & \text{if}\left\{\vec{R}_{\text{A}}^{(j)}\right\} \in \text{subdomain}(j,p) \\ 0 & \text{if}\left\{\vec{R}_{\text{A}}^{(j)}\right\} \notin \text{subdomain}(j,p) \end{cases} \quad (64)$$

Subdomain (*j*, *p*) means the *p*th electronic ground-state subdomain of cluster_*j*. Each subdomain (*j*, *p*) gets its own internal coordinate expansion and its own force constant values. The subdomains are chosen such that each single geometry, which is defined by $\{\vec{R}_{\text{A}}^{(j)}\}$, belongs to exactly one subdomain. Two different geometries of cluster_*j* may belong to the same or different subdomains, but a single geometry of cluster_*j* cannot simultaneously belong to two or more subdomains.

Since the optimized (aka 'equilibrium') ground-state geometry which resides in the $p = 1$ subdomain has relative potential and atom-in-material forces equal to zero, its zeroth-order (representing the potential contributions) and first-order (representing the force contributions) terms vanish in the Taylor series expansions of each flexibility term as shown in eqn (62). The $p \neq 1$ subdomains have no such constraint, because they do not contain the optimized (aka 'equilibrium') ground-state geometry. Accordingly, the $p \neq 1$ subdomains have the following Taylor series expansion:

$$g_\gamma^{j,(p\neq1)}\left[\{\alpha_i\}, \left\{\left(\alpha_i - \alpha_i^{\text{eq\_}j}\right)\right\}\right] = g_\gamma^{j,(p\neq1)}\left[\left\{\alpha_i = \alpha_i^{\text{eq\_}j}\right\}\right] + \sum_i \frac{\partial g_\gamma^{j,(p\neq1)}}{\partial\alpha_i}\left(\alpha_i - \alpha_i^{\text{eq\_}j}\right) + 1/2\sum_h \sum_i \frac{\partial^2 g_\gamma^{j,(p\neq1)}}{\partial\alpha_h \partial\alpha_i}\left(\alpha_h - \alpha_h^{\text{eq\_}j}\right)\left(\alpha_i - \alpha_i^{\text{eq\_}j}\right) + \text{h.o.t.} \quad (65)$$

By first choosing various types of flexibility terms (*e.g.*, bond stretches, angle bends, dihedral torsions) as $\{g_\gamma^{j,p}[\{\alpha_i\}, \{(\alpha_i - \alpha_i^{\text{eq\_}j})\}]\}$, we clearly have a linear optimization problem whose goal is to find the set of force constants values $\{k_\gamma^{j,p}\}$ such that $U_{\text{cluster\_}j}^{\text{intracluster}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ resembles $E_{\text{isolated\_cluster\_}j}^{0,\text{electronic}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ as closely as feasible subject to some optional constraints on the force constants. For example, we may want to constrain some of the force constants to have non-negative values.

2.4 Parameter optimization strategy

The exact bonded force is given by

$$\vec{F}_{\text{A,exact}}^{\text{bonded,old}} = -\vec{\nabla}_{\text{A}} E_{\text{cluster\_}j}^{0,\text{exact}} + \vec{\nabla}_{\text{A}} U_{\text{cluster\_}j}^{\text{nonbonded,old}} \quad (66)$$

$$\vec{F}_{\text{A,exact}}^{\text{bonded,new}} = -\vec{\nabla}_{\text{A}} E_{\text{cluster\_}j}^{0,\text{exact}} + \vec{\nabla}_{\text{A}} U_{\text{cluster\_}j}^{\text{nonbonded,new}} \quad (67)$$

A key distinction between the old and new scheme is that the new scheme obeys eqn (29). Accordingly, $\vec{F}_{\text{A}}^{\text{bonded,new}} = 0$ (eqn (58)) at the equilibrium geometry of the isolated cluster_*j*. This means $\vec{F}_{\text{A,exact}}^{\text{bonded,new}}$ can be expanded as

$$U^{\text{bonded,new}}_{\text{cluster\_}j\text{,exact}}\left[\left\{\vec{R}_{\text{G}}\right\}\right] - U^{\text{bonded,new}}_{\text{cluster\_}j\text{,exact}}\left[\left\{\vec{R}_{\text{G}}{}^{\text{eq\_}j}\right\}\right] = \sum_{h=1}^{\Xi_j}\sum_{i\geq h}^{\Xi_j}\left(\vartheta^{\text{new}}_{h,i}\left(\beta_h-\beta_h{}^{\text{eq\_}j}\right)\left(\beta_i-\beta_i{}^{\text{eq\_}j}\right)\right)+\text{h.o.t} \tag{68}$$

$$\vec{F}^{\text{bonded,new}}_{\text{A,exact}} = -\sum_{h=1}^{\Xi_j}\sum_{i\geq h}^{\Xi_j}\left(\vartheta^{\text{new}}_{h,i}\left(\left(\beta_h-\beta_h{}^{\text{eq\_}j}\right)\vec{\nabla}_{\text{A}}\beta_i+\left(\beta_i-\beta_i{}^{\text{eq\_}j}\right)\vec{\nabla}_{\text{A}}\beta_h\right)\right)+\text{h.o.t.} \tag{69}$$

where $\{\beta_i\}$ is a full set of non-redundant internal coordinates for cluster_*j*. H.o.t. is an abbreviation for higher order terms. $\{\vartheta_{h,i}\}$ are the associated constants. Ξ_j is the total number of non-redundant internal coordinates in cluster_*j*.

In contrast, $\vec{F}^{\text{bonded,old}}_{\text{A}}$ does not necessarily equal zero at the optimized geometry of isolated cluster_*j*. Consequently, it expands as

$$F^{\text{bonded,old}}_{\text{A,exact}}\left[\left\{\vec{R}_{\text{G}}\right\}\right] = \vec{F}^{\text{bonded,old}}_{\text{A,exact}}\left[\left\{\vec{R}_{\text{G}}{}^{\text{eq\_}j}\right\}\right] - \sum_{h=1}^{\Xi_j}\sum_{i\geq h}^{\Xi_j}\left(\vartheta^{\text{old}}_{h,i}\left(\left(\beta_h-\beta_h{}^{\text{eq\_}j}\right)\vec{\nabla}_{\text{A}}\beta_i+\left(\beta_i-\beta_i{}^{\text{eq\_}j}\right)\vec{\nabla}_{\text{A}}\beta_h\right)\right) + \text{h.o.t.} \tag{70}$$

$$\vec{F}^{\text{bonded,old}}_{\text{A,exact}}\left[\left\{\vec{R}_{\text{G}}\right\}\right] = -\sum_{h=1}^{\Xi_j}\sum_{i\geq h}^{\Xi_j}\left(\vartheta^{\text{old}}_{h,i}\left(\left(\beta_h-\beta_{h,i}{}^{\text{resting\_}j}\right)\vec{\nabla}_{\text{A}}\beta_i+\left(\beta_i-\beta_{i,h}{}^{\text{resting\_}j}\right)\vec{\nabla}_{\text{A}}\beta_h\right)\right)+\text{h.o.t.} \tag{71}$$

where $\beta_{i,h}{}^{\text{resting\_}j}$ is usually not equal to $\beta_i{}^{\text{eq\_}j}$.

For a series of small finite displacements (*e.g.*, 0.0001 Å) away from the equilibrium geometry of isolated cluster_*j*, the intracluster nonbonded force, $-\vec{\nabla}_{\text{A}}U^{\text{nonbonded,new}}_{\text{cluster\_}j}$, remains negligible because it is proportional to second-order and higher-order products of the finite displacements. In contrast, the bonded force is proportional to first-order and higher-order products of the finite displacements, as shown in eqn (69). Accordingly, the leading-order harmonic bonded force constants can be optimized by minimizing the following loss function

$$L^{\text{bonded,harmonic}}_{\text{cluster\_}j} = \sum_{\substack{\text{finite}\\ \text{displacement}\\ \text{geometries}}} \|\vec{\nabla}_{\text{A}}E^{0,\text{exact}}_{\text{cluster\_}j} - \sum_{h=1}^{\Xi_j}\sum_{i\geq h}^{\Xi_j}\left(\vartheta^{\text{new}}_{h,i}\left(\left(\beta_h-\beta_h{}^{\text{eq\_}j}\right)\vec{\nabla}_{\text{A}}\beta_i+\left(\beta_i-\beta_i{}^{\text{eq\_}j}\right)\vec{\nabla}_{\text{A}}\beta_h\right)\right)\|^2 \tag{72}$$

The vector inside $\|.\|$ has $3N_{\text{atoms}}$ force components for each geometry. This corresponds to an *x*, *y*, and *z* force component for each atom in the material's unit cell. Minimizing this *L* is a linear least squares optimization problem. Astonishingly, this means the new scheme provides a formally exact method to optimize the leading-order bonded force constants without having to explicitly pick a nonbonded interaction model (*i.e.*, without having to choose a specific model potential for $U^{\text{nonbonded}}_{\text{AG}x,\text{ intracluster}}$).

In practice, one often uses a set of redundant (rather than nonredundant) internal coordinates $\{\alpha_i\}$. In this case, the exact bonded force expands as

$$U^{\text{bonded,new}}_{\text{cluster\_}j\text{,exact}}\left[\left\{\vec{R}_{\text{G}}\right\}\right] - U^{\text{bonded,new}}_{\text{cluster\_}j\text{,exact}}\left[\left\{\vec{R}_{\text{G}}{}^{\text{eq\_}j}\right\}\right] = \sum_{h}^{N_{\text{RIC}}}\sum_{i\geq h}^{N_{\text{RIC}}}\left(\tilde{\vartheta}^{\text{new}}_{h,i}\left(\alpha_h-\alpha_h{}^{\text{eq\_}j}\right)\left(\alpha_i-\alpha_i{}^{\text{eq\_}j}\right)\right)+\text{h.o.t.} \tag{73}$$

$$\vec{F}^{\text{bonded,new}}_{\text{A,exact}} = -\sum_{i}^{N_{\text{RIC}}}\sum_{i\geq h}^{N_{\text{RIC}}}\left(\tilde{\vartheta}^{\text{new}}_{h,i}\left(\left(\alpha_h-\alpha_h{}^{\text{eq\_}j}\right)\vec{\nabla}_{\text{A}}\alpha_i+\left(\alpha_i-\alpha_i{}^{\text{eq\_}j}\right)\vec{\nabla}_{\text{A}}\alpha_h\right)\right)+\text{h.o.t.} \tag{74}$$

N_{RIC} is the number of redundant internal coordinates in cluster_*j*. This defines a linear least squares problem analogous to eqn (72) except that a regularization method (*e.g.*, LASSO[32,33]) should be used to handle the multicollinearity problem caused by redundancy in the internal coordinates. Again, this allows us to construct and optimize a model for the bonded force constants to leading order without having to explicitly pick a nonbonded interaction model.

As shown in eqn (61) and (62), it is possible to use a set of flexibility terms $\{g^{j,1}_{\gamma}\}$ that have a similar expansion to leading order as eqn (73). This defines a linear least squares problem with possible multicollinearity that should be addressed by using a regularization method (*e.g.*, LASSO[32,33]). Once again, this allows us to construct and optimize a model for the bonded force constants to leading order without having to explicitly pick a nonbonded interaction model. This amazing result is used to optimize bonded interaction models for 116 MOFs in the companion paper.[45]

Notably, the new scheme is formally exact to all orders, not merely to leading order. To compute the formally exact bonded force constants to all higher orders, the new scheme requires the loss function to also include the intracluster nonbonded interactions:

$$L^{\text{new,full}}_{\text{cluster\_}j} = W_{\text{E}}\sum_{\substack{\text{energy}\\ \text{training}\\ \text{geometries}}}\left(\Delta E^{0,\text{exact}}_{\text{cluster\_}j}-\left(\Delta U^{\text{nonbonded,new}}_{\text{cluster\_}j}+\Delta U^{\text{bonded,new}}_{\text{cluster\_}j}\right)\right)^2 + W_{\text{F}}\sum_{\substack{\text{force}\\ \text{training}\\ \text{geometries}}}\|\vec{\nabla}_{\text{A}}E^{0,\text{exact}}_{\text{cluster\_}j}-\vec{\nabla}_{\text{A}}U^{\text{nonbonded,new}}_{\text{cluster\_}j}-\vec{\nabla}_{\text{A}}U^{\text{bonded,new}}_{\text{cluster\_}j}\|^2 + \text{r.p.t.} + \text{constraints} \tag{75}$$

$$\Delta E^0_{\text{cluster\_}j} = E^0_{\text{cluster\_}j}[\{R_{\text{C}}\}] - E^0_{\text{cluster\_}j}[\{R_{\text{C}}{}^{\text{eq\_}j}\}] \tag{76}$$

$$\Delta U^{\text{nonbonded,(scheme)}}_{\text{cluster\_}j} = U^{\text{nonbonded,(scheme)}}_{\text{cluster\_}j}[\{R_{\text{C}}\}] - U^{\text{nonbonded,(scheme)}}_{\text{cluster\_}j}[\{R_{\text{C}}{}^{\text{eq\_}j}\}] \tag{77}$$

$$\Delta U^{\text{bonded,(scheme)}}_{\text{cluster\_}j} = U^{\text{bonded,(scheme)}}_{\text{cluster\_}j}[\{R_{\text{C}}\}] - U^{\text{bonded,(scheme)}}_{\text{cluster\_}j}[\{R_{\text{C}}{}^{\text{eq\_}j}\}] \tag{78}$$

where r.p.t. is the regularization penalty term that handles the multicollinearity problem. Here, W_E and W_F are observation weights applied to the energies and forces, respectively. A full series expansion (*e.g.*, eqn (63)) for $U^{\text{bonded,new}}_{\text{cluster\_}j}$ must be inserted into eqn (75). In practical applications, the following leading-order approximation is typically used

$$L^{\text{new,leading\_order}}_{\text{cluster\_}j} = W_E \sum_{\substack{\text{energy}\\\text{training}\\\text{geometries}}} \left(\Delta E^0_{\text{cluster\_}j}[\{R_C\}] - \Delta U^{\text{bonded,new}}_{\text{cluster\_}j}\right)^2 + W_F \sum_{\substack{\text{force}\\\text{training}\\\text{geometries}}} \left\| \vec{\nabla}_A E^0_{\text{cluster\_}j} - \vec{\nabla}_A U^{\text{bonded,new}}_{\text{cluster\_}j} \right\|^2 + \text{r.p.t.} + \text{constraints} \tag{79}$$

where one uses an approximate (*i.e.*, truncated) series expansion for $U^{\text{bonded,new}}_{\text{cluster\_}j}$.

The old scheme explicitly requires us to specify a particular nonbonded interaction model even if we only want to optimize the bonded force constants to leading order. Although this type of regularization has not been used with the old scheme in the prior literature,[17,22,23,27,64,65] one could construct the following type of loss function for the old scheme

$$L^{\text{old}}_{\text{cluster\_}j} = W_E \sum_{\substack{\text{energy}\\\text{training}\\\text{geometries}}} \left(\Delta E^{0,\text{exact}}_{\text{cluster\_}j} - \left(\Delta U^{\text{nonbonded,old}}_{\text{cluster\_}j} + \Delta U^{\text{bonded,old}}_{\text{cluster\_}j}\right)\right)^2 + W_F \sum_{\substack{\text{force}\\\text{training}\\\text{geometries}}} \left\| \vec{\nabla}_A E^0_{\text{cluster\_}j} - \vec{\nabla}_A U^{\text{nonbonded,old}}_{\text{cluster\_}j} - \vec{\nabla}_A U^{\text{bonded,old}}_{\text{cluster\_}j} \right\|^2 + \text{r.p.t.} + \text{constraints} \tag{80}$$

Compared to the new scheme, the old scheme requires additional terms and/or additional parameters to expand the bonded potential, because under the old scheme the bonded forces are not necessarily zero at isolated cluster_*j*'s optimized geometry. The following two requirements of the old scheme make it much more complicated than the new scheme:

(i) The old scheme requires explicitly choosing and including $U^{\text{nonbonded,old}}_{\text{cluster\_}j}$ even if we only want to optimize the bonded force constants to leading order. Under the old scheme, even the leading-order bonded force constant values depend on the particular choice of intracluster nonbonded potential model, $U^{\text{nonbonded,old}}_{\text{cluster\_}j}$.

(ii) The old scheme requires optimizing more bonded parameters than the new scheme. Specifically, the old scheme requires optimizing 'resting values' in the flexibility terms or including non-zero force intercept terms in the flexibility model.

In summary, this new scheme for separating bonded from nonbonded interactions in a nonreactive forcefield has so many compelling advantages that it should completely replace the old scheme. It is one of those cases where an important simplification (*i.e.*, turning a nonlinear optimization problem into a linear optimization problem for the bonded parameters) maintains formal exactness and simultaneously improves numeric precision and computational convenience (by providing continuous first- and second-order derivatives when using a nonbonded interaction cutoff distance), transferability (because the choice of $U^{\text{nonbonded}}_{\text{AB}x\text{,intracluster}}[\{\vec{R}_C\}]$ does not affect forces or Hessian at the isolated cluster's optimized geometry), and convergence robustness (because linear optimization problems do not have multiple local minima in the loss function's value).

Fig. 2 summarizes the key equations for my new forcefield parameterization process that consists of the following steps:

(1) First, we separate $E^0_{\text{electronic}}[\{\vec{R}_A, \mathbb{Z}_A\}]$ into intercluster and intracluster contributions. To do this, a series of quantum chemistry calculations are first performed to compute $E^0_{\text{cluster\_}j}[\{\vec{R}_A, \mathbb{Z}_A\}]$ for each individual bonded cluster_*j* by itself (*i.e.*, an isolated bonded cluster) over many geometries $\{\vec{R}_A^{(j)}\}$ allowing its internal coordinates (*e.g.*, bond lengths, bond angles, dihedrals, *etc.*) to vary.

(2) Second, a separate set of quantum chemistry calculations is then performed for the entire system that contains all of the bonded clusters together at various geometries. As shown in eqn (59), the intercluster energy is computed as the difference between the Born–Oppenheimer electronic energy of the total system and that of the isolated clusters.

(3) Following the method described in Section 2.2 above, intracluster nonbonded interactions are defined for each isolated bonded cluster. The remaining intracluster energy is assigned to the bonded interactions:

$$E^{\text{bonded}}_{\text{cluster\_}j}\left[\left\{\vec{R}_A, \mathbb{Z}_A\right\}\right] = E^{0,\text{electronic}}_{\text{isolated\_cluster\_}j}\left[\left\{\vec{R}_A, \mathbb{Z}_A\right\}\right] - U^{\text{nonbonded,new}}_{\text{cluster\_}j} \tag{81}$$

(4) Using linear regression with an appropriate Lagrangian ($L^{\text{bonded}}_{\text{cluster\_}j}$), the intracluster bonded energy model, $U^{\text{bonded,new}}_{\text{cluster\_}j}[\{\vec{R}_A, \mathbb{Z}_A\}]$, is fit to an internal coordinate series expansion for each electronic subdomain of cluster *j* to reproduce $E^{\text{bonded}}_{\text{cluster\_}j}[\{\vec{R}_A, \mathbb{Z}_A\}]$ as closely as feasible:

$$\text{minimize} \sum_{\substack{\text{cluster\_}j\\\text{geometries}}} L^{\text{bonded}}_{\text{cluster\_}j} \tag{82}$$

If the internal coordinate series expansion is complete, the minimum of the loss function will be zero, and this corresponds to an exact match between $U^{\text{bonded,new}}_{\text{cluster\_}j}[\{\vec{R}_A, \mathbb{Z}_A\}]$ and $E^{\text{bonded}}_{\text{cluster\_}j}[\{\vec{R}_A, \mathbb{Z}_A\}]$. In most practical applications, a truncated series expansion is used leading to approximation. In addition to energies, the training dataset and Lagrangian for bonded interactions may also include atom-in-material forces and/or constraints (such as bounds on some force constants) and/or regularization terms.

(5) Using linear or nonlinear regression with an appropriate Lagrangian ($L^{\text{intercluster}}_{\text{nonbonded}}$), the intercluster nonbonded energy

model, $U_{\text{nonbonded}}^{\text{intercluster}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$, is fit to an internal coordinate series expansion to reproduce $E_{\text{intercluster}}^{0}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ as closely as feasible by finding the minimum of $L_{\text{nonbonded}}^{\text{intercluster}}$. If the internal coordinate series expansion for $U_{\text{nonbonded}}^{\text{intercluster}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ is complete, the minimum of the loss function will be zero, and this corresponds to an exact match between $U_{\text{nonbonded}}^{\text{intercluster}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ and $E_{\text{intercluster}}^{0}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$. In most practical applications, a truncated series expansion is used leading to approximation. In addition to energies, the training dataset and loss function for nonbonded interactions may also include atom-in-material forces and/or constraints (such as bounds on some nonbonded parameters) and/or regularization terms. For example, this Lagrangian might take the form

$$L_{\text{nonbonded}}^{\text{intercluster}} = \left(\begin{array}{l} W_{\text{E}} \sum\limits_{\substack{\text{intercluster}\\\text{training}\\\text{geometry}\\\text{energies}}} \left(U_{\text{nonbonded}}^{\text{intercluster}}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] - E_{\text{intercluster}}^{0}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] \right)^2 \\ +W_{\text{F}} \sum\limits_{\substack{\text{intercluster}\\\text{training}\\\text{geometry}\\\text{forces}}} \left\| \vec{\nabla}_{\text{A}} U_{\text{nonbonded}}^{\text{intercluster}}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] - \vec{\nabla}_{\text{A}} E_{\text{intercluster}}^{0}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] \right\|^2 \\ +\text{r.p.t.} + \text{constraints} \end{array} \right) \tag{83}$$

(6) Conceptually, the forcefield's total potential energy can be reconstructed as the sum of three parts, which are the intracluster bonded interactions, the intracluster nonbonded interactions, and the intercluster nonbonded interactions:

$$\begin{aligned} U_{\text{total}}^{\text{FF}}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] &= \underbrace{\sum_{\text{cluster\_}j=1}^{N_{\text{clusters}}} U_{\text{cluster\_}j}^{\text{bonded,new}}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right]}_{\text{intracluster bonded interactions}} \\ &+ \underbrace{\sum_{\text{cluster\_}j=1}^{N_{\text{clusters}}} U_{\text{cluster\_}j}^{\text{nonbonded,new}}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right]}_{\text{intracluster nonbonded interactions}} + \underbrace{U_{\text{nonbonded}}^{\text{intercluster}}}_{\substack{\text{intercluster}\\\text{nonbonded}\\\text{interactions}}} \end{aligned} \tag{84}$$

2.5 Conditions under which this theoretical framework is formally exact

This section describes the conditions under which the theoretical framework described in the previous sections is formally exact, which means the forcefield's potential energy model asymptotically approaches the exact potential energy model:

$$U_{\text{total}}^{\text{FF}}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] \to U_{\text{total}}^{\text{exact}}\left[\left\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\right\}\right] \tag{85}$$

At first one may wonder whether formal exactness is a meaningless theoretical result, because in most practical situations some optional approximations will be chosen to make the forcefield easier to parameterize and use at the expense of losing some accuracy. However, closer analysis shows formal exactness is an extremely important property. If a theoretical framework is not formally exact, then the exact solution lies outside that theoretical framework; in this case, one reaches a wall beyond which the results cannot be further improved without leaving that particular theoretical framework. If a theoretical framework is formally exact, then the exact solution lies inside that theoretical framework; in this case, one can always improve the accuracy of solutions to get closer to and even reach the exact solution without having to leave that particular theoretical framework.

To begin, we must precisely define the problem statement whose exact solution defines the exact solution we seek. Here, the problem statement is defined as follows. For a specific material (aka 'chemical system') of precisely defined chemical composition and with precisely defined bond connectivity graph in the absence of externally applied fields, compute the exact Born–Oppenheimer ground-state electronic energy $E_{\text{electronic}}^{0,\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ for various sets of chemical geometries $\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}$ that match the defined bond connectivity graph, and use these results to construct the exact functional $U_{\text{total}}^{\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}] = E_{\text{electronic}}^{0,\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$.

The proof that such an exact functional $U_{\text{total}}^{\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ exists proceeds as follows. In the absence of externally applied fields (*e.g.*, in the absence of externally applied electric, magnetic, and gravitational fields), $\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}$ together with the chosen level of relativistic corrections defines the system's Hamiltonian. The system's Hamiltonian in turn defines its Born–Oppenheimer ground-state electronic energy, $E_{\text{electronic}}^{0,\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$. Since $U_{\text{total}}^{\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}] = E_{\text{electronic}}^{0,\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$, the existence of $E_{\text{electronic}}^{0,\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ means $U_{\text{total}}^{\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ (which is $E_{\text{electronic}}^{0,\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$) also exists. Hence, the exact $U_{\text{total}}^{\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ exists.

Let us examine the hypothetical counter-argument that no $U_{\text{total}}^{\text{exact}}[\{\vec{R}_{\text{A}}, \mathbb{Z}_{\text{A}}\}]$ exists or that it exists but does not equal

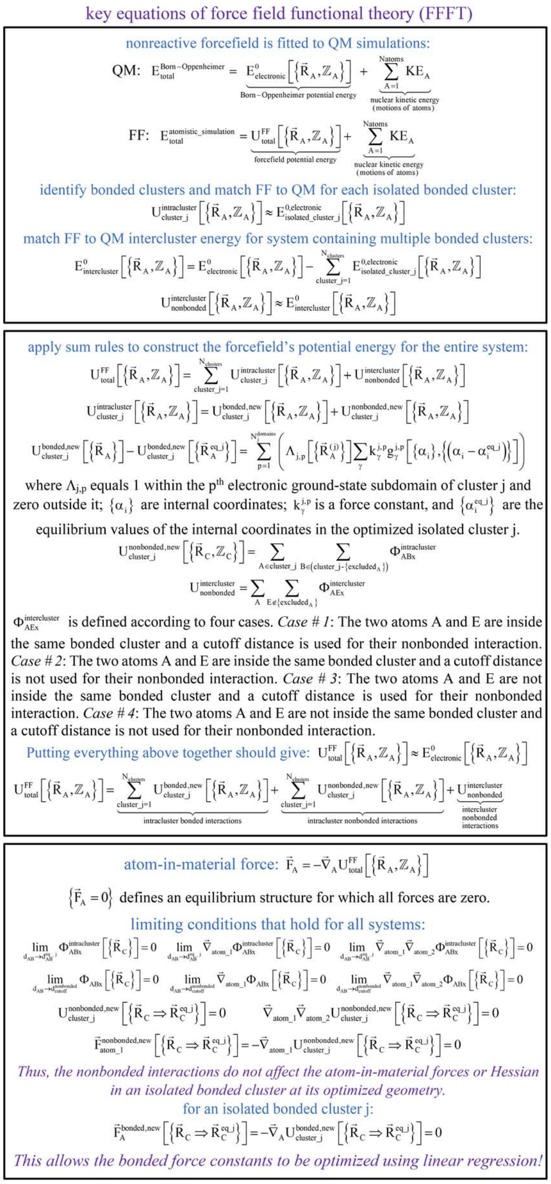

Fig. 2 A graphic summarizing relationships between key equations in force field functional theory.

$E^{0,\text{exact}}_{\text{electronic}}[\{\vec{R}_A, \mathbb{Z}_A\}]$, because the Born–Oppenheimer approximation is itself an approximation. This notion of non-exactness arises from choosing to define the problem statement as $U^{\text{exact}}_{\text{total}}$ being intended to match some experimental observable, and since the Born–Oppenheimer approximation is an approximation it does not exactly reproduce experimental observables. While it may be possible to redefine $U^{\text{exact}}_{\text{total}}$ in that way, I have chosen not to do so. The concept of using a forcefield's potential energy $U^{\text{FF}}_{\text{total}}[\{\vec{R}_A, \mathbb{Z}_A\}]$ intrinsically rests on the separation of electronic from nuclear motions. If these are strongly coupled so that the Born–Oppenheimer approximation is unreasonable, then in that case the forcefield's potential energy would need to include the electronic positions $\{\vec{r}_i\}$ as well, $U_{\text{total}}[\{\vec{r}_i, \vec{R}_A, \mathbb{Z}_A\}]$, and both the electrons and atomic nuclei would need to be included as explicit particles when using that forcefield in subsequent simulations. The fact that $U^{\text{exact}}_{\text{total}}[\{\vec{R}_A, \mathbb{Z}_A\}]$ omits the electronic positions $\{\vec{r}_i\}$ as explicit coordinates means it only applies within the Born–Oppenheimer approximation that allows the electronic relaxation to be performed for fixed nuclear positions. Hence, $U^{\text{exact}}_{\text{total}}[\{\vec{R}_A, \mathbb{Z}_A\}]$ must be defined as equal to $E^{0,\text{exact}}_{\text{electronic}}[\{\vec{R}_A, \mathbb{Z}_A\}]$. If we want something that goes beyond the Born–Oppenheimer approximation, that requires a different type of construct, $U_{\text{total}}[\{\vec{r}_i, \vec{R}_A, \mathbb{Z}_A\}]$, and a completely different type of forcefield that includes both the electrons and atomic nuclei as explicit particles in the forcefield. Here, I have chosen to define force field functional theory within the scope of the Born–Oppenheimer approximation. Within that defined scope, formal exactness is defined as constructing a forcefield model that exactly reproduces the Born–Oppenheimer potential energy surface as shown in eqn (14) and (85).

Table 1 lists the conditions that must be satisfied for this theoretical framework to reach the exact solution. Each of these conditions is now discussed.

To be exact, the bonded clusters and overall system being studied must be exactly the same ones as the forcefield was trained for. Within this theoretical framework, bonded interactions are parameterized for each isolated bonded cluster, and each bonded interaction is intracluster. On the other hand,

Table 1 List of conditions that must be met for this theoretical framework to be formally exact

(1) The bonded clusters and overall system being studied must be exactly the same ones as the forcefield was trained for
(2) No nuclear reactions, no nuclear decay processes, and no nuclear excitation processes take place
(3) Since the nuclear spin and local rotational orientation of an atomic nucleus is neglected in this type of nonreactive forcefield, this type of nonreactive forcefield is formally exact only when each atomic nucleus in the real physical system is spinless and spherically symmetric
(4) There are no externally applied fields, or the forcefield has been specifically parameterized for the precise configuration of externally applied fields that is present
(5) Within the Born–Oppenheimer approximation, an exact electronic structure theory is used to compute $E^{0,\text{exact}}_{\text{electronic}}[\{\vec{R}_A, \mathbb{Z}_A\}]$. This requires using an exact exchange–correlation theory together with appropriate relativistic corrections
(6) No new chemical bonds are formed and no chemical bonds are completely severed; however, the bond length of a bond may approach infinity
(7) For nonbonded interactions having theoretically infinite distance range, the $d^{\text{nonbonded}}_{\text{cutoff}} \to \infty$ limit must be used. For nonbonded interactions having theoretically limited distance range, the exact range $d^{\text{nonbonded}}_{\text{cutoff}} = d^{\text{exact\_range}}_{\text{cutoff}}$ must be used
(8) Because the forcefield is parameterized for the electronic ground state only, it does not describe processes involving excited electronic states or excited spin states
(9) The forcefield must be used only within the particular electronic ground-state subdomains for which it was parameterized. Since each electronic ground-state subdomain defines a region of atom-in-material positions, $\{\vec{R}_A\}$, this type of nonreactive forcefield must be used only within the general regions of $\{\vec{R}_A\}$ space for which it was trained
(10) $U^{\text{exact}}_{\text{total}}[\{\vec{R}_A, \mathbb{Z}_A\}]$ has been constructed to exactly match $E^{0,\text{exact}}_{\text{electronic}}[\{\vec{R}_A, \mathbb{Z}_A\}]$ over the relevant electronic ground-state subdomains

there are both intracluster nonbonded interactions and intercluster nonbonded interactions. For example, if the forcefield was trained on a system containing three water molecules, then formal exactness is lost and approximations manifest if we try to apply that same forcefield to a new system containing four water molecules. In this case, system # 1 containing 3 water molecules properly has exactly the same bonded interactions and intracluster nonbonded interactions as system # 2 containing 4 water molecules; however, these two systems have different intercluster nonbonded interactions. In other words, the bonded interactions and intracluster nonbonded interactions are strictly transferable between different systems comprised of the same kinds of bonded clusters, but the intercluster nonbonded interactions are not strictly transferable across such systems.

In practice, it is often more convenient to accept some level of approximation that results from applying a versatile intercluster nonbonded interaction model to different but similar systems. Consider as an example a series of systems in which a particular MOF is loaded with different combinations of molecules such as N_2, O_2, CO_2, methane, *etc.* In this case, a forcefield could be developed as follows. First, the bonded interactions and intracluster nonbonded interactions are calculated for each isolated bonded cluster. These are exactly transferable to the combined system containing several bonded clusters. Then, a versatile (but approximate) intercluster nonbonded interaction model is parameterized and applied to each system in the series. This is generally more convenient than the formally exact approach that requires separately parameterizing a new intercluster nonbonded interaction model for each specific combination of molecules in the MOF.

This type of nonreactive forcefield treats the atomic nuclei as immutable. Consequently, this type of nonreactive forcefield describes processes in which no nuclear reactions, no nuclear decay processes, and no nuclear excitation processes take place. Since the nuclear spin and local rotational orientation of an atomic nucleus is neglected in this type of nonreactive forcefield, this type of nonreactive forcefield is formally exact only when each atomic nucleus in the real physical system is spinless and spherically symmetric. In other words, this type of nonreactive forcefield does not describe nuclear magnetic resonance (NMR), the Mössbauer effect, and other phenomena related to nuclear spins or nuclear energy transitions. In line with the Born–Oppenheimer approximation that separates the electronic and nuclear motions, this immutability of atomic nuclei is considered to be part of the problem statement whose formally exact solution is sought.

A formally exact parameterization of the forcefield corresponds to one specific Hamiltonian. Consequently, there is a one-to-one correspondence between the system's Hamiltonian and the forcefield. Any modification that alters the system's Hamiltonian requires parameterizing a new forcefield to retain formal exactness. Since adding external fields (*e.g.*, external electric, magnetic, or gravitational fields) changes the system's Hamiltonian, to retain formal exactness a new forcefield would have to be parameterized for each combination of externally applied fields. Because reparametrizing the forcefield for each specific combination of externally applied fields would be extremely tedious, it is generally more convenient to accept some level of approximation that allows the same forcefield to be applied irrespective of the specific combination of externally applied fields. In general, polarizable forcefields can more accurately approximate responses to externally applied electric fields than nonpolarizable forcefields.[13,19,66–68]

Formal exactness requires that the Born–Oppenheimer electronic ground-state energy, $E^0_{\text{electronic}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$, be computed using an exact quantum chemistry method. This requires both that the exchange–correlation theory used is exact and that appropriate relativistic corrections[69–73] are included in the quantum chemistry calculations. Examples of formally exact quantum chemistry methods include full configuration interaction expansion and density functional theory (DFT) calculations in the complete basis set limit.[74–77] Since the exact DFT exchange–correlation functional is still unknown, in practice $E^0_{\text{electronic}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ is normally computed using a density functional approximation (DFA) or any other desired quantum chemistry method (*e.g.*, coupled-cluster, configuration interaction, *etc.*).[77–85] For best results, the quantum chemistry method chosen should include long-range dispersion interactions.[86–89] For convenience, a finite-sized basis set is normally used instead of the complete basis set limit.[90–92] If the finite-sized basis set is appropriately chosen, then this introduces an acceptable level of approximation. For heavy chemical elements, additional approximations such as freezing some of the core electrons or replacing some of the core electrons with a relativistic effective core potential (RECP) are sometimes used.[72,93,94] Even though the exact DFT exchange–correlation functional is still unknown, all of these approximations are formally optional, because we could conceivably (but not necessarily practically) choose to perform a full configuration interaction calculation in the complete basis set limit to obtain the exact quantum chemistry result; however, that would be extremely (and sometimes prohibitively) computationally expensive.

Each nonreactive forcefield operates only within the scope of a particular fixed bond connectivity graph; however, bonds (treated as harmonic or anharmonic springs) are allowed to stretch beyond the sum of their atom typing radii. Some nonreactive forcefields (*e.g.*, Morse potential,[48] QMDFF[47]) even allow bonds to stretch to infinite length. Accordingly, nonreactive forcefields do not describe complex chemical reactions. I classify this as a restriction on the scope of nonreactive forcefields, rather than treating it as an approximation. For a collection of atoms, a complete Born–Oppenheimer potential energy surface may contain separate regions (aka 'valleys') for reactants and products that are connected by reaction paths. Traversing these reaction paths involves forming and/or breaking chemical bonds. For example, a complete Born–Oppenheimer potential energy surface for four hydrogen, one carbon, and four oxygen atoms would contain separate regions and connecting reaction paths corresponding to: (a) one methane (CH_4) and two oxygen (O_2) molecules, (b) one carbon dioxide (CO_2) and two water (H_2O) molecules, (c) one formaldehyde (CH_2O) plus one water

(H_2O) plus one oxygen (O_2) molecule, and (d) many other regions and connecting reaction paths.

When using a nonbonded interaction cutoff distance, interactions having a theoretically infinite distance range (such as the Coulomb interaction between charged particles) will be undercounted between particles farther apart than $d_{\text{cutoff}}^{\text{nonbonded}}$. For nonbonded interactions that have theoretically infinite distance range, formal exactness requires that we use the $d_{\text{cutoff}}^{\text{nonbonded}} \to \infty$ limit for those interactions. We also allow the possibility that some (but not all) of the nonbonded interactions may have a theoretically limited distance range. For those particular nonbonded interactions, we should set $d_{\text{cutoff}}^{\text{nonbonded}} = d_{\text{cutoff}}^{\text{exact\_range}}$ (This possibility is included to accommodate multibody interaction models in fluids that have a finite-range of the multibody interactions. In such case, $d^{\text{nonbonded}}_{\text{cutoff}} = d_{\text{cutoff}}^{\text{exact\_range}}$ could be tuned so that the multibody interaction model reproduces the correct interaction energy.). This means the value of $d_{\text{cutoff}}^{\text{nonbonded}}$ can be different for different nonbonded interactions.

Because this type of nonreactive forcefield is parameterized for the electronic ground state only, it does not describe processes involving excited electronic states or excited spin states. I classify this as a restriction on the forcefield's scope, rather than treating it as an approximation. Manifestly, this type of nonreactive forcefield cannot describe optical transitions, fluorescence, phosphorescence, photoelectronic processes, electron excitation, electron transport, spin excitation, and spin transport phenomena.

To achieve formal exactness, this type of nonreactive forcefield must be used only within the particular electronic ground-state subdomains for which it was parameterized. It is possible to simultaneously parameterize this type of nonreactive forcefield for one, two, or more different electronic ground-state subdomains. Since each electronic ground-state subdomain defines a region of atom-in-material positions, $\{\vec{R}_\text{A}\}$, this type of nonreactive forcefield must be used only within the general regions of $\{\vec{R}_\text{A}\}$ space for which it was trained.

Finally, to achieve formal exactness, $U_{\text{total}}^{\text{exact}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ must be constructed to exactly match $E_{\text{electronic}}^{0,\text{exact}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ over the relevant electronic ground-state subdomains. To accomplish this, the process illustrated in Fig. 2 and described in the previous sections should be followed employing complete series expansions in terms of the internal coordinates for both the intracluster bonded interactions and the intercluster nonbonded interactions within each electronic ground-state subdomain. In this context, a 'complete series expansion' means a series expansion that employs all the independent degrees of freedom and also has enough functional representation (*i.e.*, a complete set of basis functions) to achieve an exact match between $U_{\text{total}}^{\text{exact}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ and $E_{\text{electronic}}^{0,\text{exact}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ within each electronic ground-state subdomain.

Both the electronic ground-state subdomains of the full system (which may contain multiple bonded clusters) and the electronic ground-state subdomains of each associated isolated bonded cluster must be included to construct an exact forcefield. Derivative discontinuities in $E_{\text{electronic}}^{0,\text{exact}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ and hence also in $U_{\text{total}}^{\text{exact}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ may occur at a boundary between two or more electronic ground-state subdomains of the full system. $U_{\text{total}}^{\text{exact}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}] = E_{\text{electronic}}^{0,\text{exact}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ is continuous but not necessarily continuously differentiable at such boundaries. Derivative discontinuities in $E_{\text{isolated\_cluster\_}j}^{0,\text{electronic}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ and hence also in exact_$U_{\text{cluster_}j}^{\text{intracluster}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ may occur at a boundary between two or more electronic ground-state subdomains of isolated cluster_*j*. exact_$U_{\text{cluster_}j}^{\text{intracluster}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}] = E_{\text{isolated_cluster_}j}^{0,\text{electronic}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ is continuous but not necessarily continuously differentiable at such boundaries. To be exact, $U_{\text{cluster\_}j}^{\text{bonded,new}}[\{\vec{R}_\text{A}\}]$ must be expanded locally within each electronic ground-state subdomain of the isolated cluster_*j* (see eqn (63)). Because $E_{\text{intercluster}}^{0,\text{exact}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ depends on both $E_{\text{electronic}}^{0,\text{exact}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ and $E_{\text{isolated\_cluster\_}j}^{0,\text{electronic}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ as shown in eqn (59), this means $E_{\text{intercluster}}^{0,\text{exact}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ may exhibit derivative discontinuities wherever either the full system or any of the associated isolated bonded clusters undergoes a change in electronic ground state. This means exact_$U_{\text{AB}x,\text{intercluster}}^{\text{nonbonded}}[\{\vec{R}_\text{C}\}]$ has the following piecewise expansion

$$\text{exact\_}U_{\text{AB}x,\text{intercluster}}^{\text{nonbonded}}\left[\left\{\vec{R}_\text{C}\right\}\right] = \sum_{w=1}^{N_{\text{pieces}}}\left(\Gamma_w\left[\left\{\vec{R}_\text{C}\right\}\right] U_{\text{AB}x,\text{intercluster}}^{\text{nonbonded,piece\_}w}\left[\left\{\vec{R}_\text{C}\right\}\right]\right) \quad (86)$$

where

$$\Gamma_w\left[\left\{\vec{R}_\text{C}\right\}\right] = \begin{cases} 1 & \text{if}\left\{\vec{R}_\text{C}\right\} \in \text{piece\_}w \\ 0 & \text{if}\left\{\vec{R}_\text{C}\right\} \notin \text{piece\_}w \end{cases} \quad (87)$$

All locations inside piece_*w* share the same electronic ground-state type of the full system. All locations inside piece_*w* also share the same electronic ground-state type of isolated cluster_*j*. Locations inside two different pieces have either different electronic ground-state types for the full system or for any associated isolated bonded cluster. $U_{\text{AB}x,\text{intercluster}}^{\text{nonbonded,piece\_}w}[\{\vec{R}_\text{C}\}]$ is continuous and continuously differentiable (up to some order) with respect to atom displacements (*i.e.*, changes in $\{\vec{R}_\text{C}\}$).

We normally only know values of $E_{\text{electronic}}^{0}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ that have been numerically computed for several chosen chemical geometries. Consequently, we have to use regression techniques to build a model for $U_{\text{total}}^{\text{FF}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$. In practice, approximate expressions are normally used to build the forcefield's $U_{\text{total}}^{\text{FF}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ functional, and this introduces a (hopefully small) difference between $U_{\text{total}}^{\text{FF}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ and $E_{\text{electronic}}^{0}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$:

$$U_{\text{total}}^{\text{FF}}\left[\left\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\right\}\right] \approx E_{\text{electronic}}^{0}\left[\left\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\right\}\right] \quad (88)$$

Although careful accounting of electronic ground-state subdomains is required to construct the exact forcefield, normally it is much easier to pursue an approximate treatment in which we focus on training the forcefield over particular region(s) of $\{\vec{R}_\text{A}\}$ space. $U_{\text{total}}^{\text{FF}}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ only provides reasonable approximation to $E_{\text{electronic}}^{0}[\{\vec{R}_\text{A}, \mathbb{Z}_\text{A}\}]$ within the general region(s) of $\{\vec{R}_\text{A}\}$

space for which it was trained (*i.e.*, fitted, parameterized). As an approximation, $U^{\text{nonbonded}}_{\text{AB}x,\text{intercluster}}[\{\vec{R}_{\text{C}}\}]$ is often oversimplified to have only a single piece (*i.e.*, $N_{\text{pieces}} = 1$) and $U^{\text{bonded,new}}_{\text{cluster}\_j}[\{\vec{R}_{\text{A}}\}]$ is often oversimplified to have only one subdomain (*i.e.*, $N^{\text{domains}}_{j} = 1$). These approximations make forcefield training easier.

2.6 Worked examples

2.6.1 The stretched H_2 molecule. Full configuration interaction (FCI) calculations were performed for the H_2 molecule. Because the H_2 molecule contains only two electrons, FCI calculation for this molecule corresponds to configuration interaction with single and double excitations (*i.e.*, CISD). CISD calculations were performed in Gaussian (ref. 95) software using the aug-cc-pVQZ[96–98] basis set. These calculations solved the time-independent multi-electronic Schrodinger equation

$$\hat{H}_{\text{el}}\Psi^{0}_{\text{electronic}} = E^{0}_{\text{electronic}}\Psi^{0}_{\text{electronic}} \tag{89}$$

using the following multi-electronic Hamiltonian operator (expressed in atomic units):

$$\hat{H}_{el} = \underbrace{-(1/2)\sum_{i=1}^{N_{\text{electrons}}}\nabla_i^{\,2}}_{\text{kinetic energy of electrons}} + \underbrace{\sum_{i=1}^{N_{\text{electrons}}}\sum_{A=1}^{N_{\text{atoms}}}\frac{-\mathbb{Z}_A}{|\vec{r}_i - \vec{R}_A|}}_{\text{electron-nuclei potential energy}}$$

$$+ \underbrace{\sum_{i=1}^{N_{\text{electrons}}}\sum_{j>i}^{N_{\text{electrons}}}\frac{1}{|\vec{r}_i - \vec{r}_j|}}_{\text{electron-electron potential energy}} + \underbrace{\sum_{A=1}^{N_{\text{atoms}}}\sum_{B>A}^{N_{\text{atoms}}}\frac{\mathbb{Z}_A\mathbb{Z}_B}{|\vec{R}_A - \vec{R}_B|}}_{\text{nucleus-nucleus potential energy}} \tag{90}$$

Eqn (89) and (90) are discussed in common quantum chemistry textbooks.[99–101] In eqn (89), $\Psi^{0}_{\text{electronic}}$ is the ground-state multi-electronic wavefunction. The multi-electronic Hamiltonian operator in eqn (90) is one of several possible choices that can be used in this theoretical framework. If desired, the multi-electronic Hamiltonian operator could include various relativistic corrections, spin–orbit coupling, and/or spin–spin magnetic coupling, *etc.*[99–101] For simplicity, those interactions were not included in the example studied in this section.

In the absence of externally applied fields, the only independent internal geometric coordinate for this molecule is its bond length. The FCI/aug-cc-pVQZ optimized bond length for the H_2 singlet spin state is 74.199 picometer (pm). The triplet energy was also computed at this same bond length. For both the singlet and triplet spin states, FCI/aug-cc-pVQZ calculations were also performed at a series of constrained bond lengths from 50 to 500 pm.

For each H atom, the aug-cc-pVQZ basis set contains six sets of s-type basis functions, four sets of p-type basis functions, three sets of d-type basis functions, and two sets of f-type basis functions.[97] One s-type basis function is a contraction of multiple Gaussian exponents.[97] The other five s-type basis functions and all of the p, d, and f basis functions contain one Gaussian exponent per basis function.[97] As shown in Table S1 of ESI,† the FCI/aug-cc-pVQZ singlet and triplet energies at 500 pm are less than 10^{-4} hartree away from the complete basis set limit value of exactly minus one hartree for two isolated hydrogen atoms. This shows the FCI/aug-cc-pVQZ results for the H_2 molecule are close to the complete basis set limit. Accordingly, the quantum-mechanically-computed results listed in Table S1 of ESI† are a nearly exact solution to eqn (89) and (90).

At each finite bond length, the exact (*i.e.*, FCI near the complete basis set limit) spin triplet electronic energy of H_2 is higher than the spin singlet electronic energy of the same bond length. At infinite bond length, the exact (*i.e.*, FCI near the complete basis set limit) spin triplet and spin singlet electronic energies of H_2 are equal. Accordingly, there is no singlet-to-triplet electronic ground-state crossover for this molecule at the FCI level of theory near the complete basis set limit. Thus, only one electronic ground-state subdomain (*i.e.*, the singlet electronic ground state) is needed to construct the nonreactive forcefield for this molecule.

For the H_2 molecule, there are no intracluster nonbonded interactions, because each H atom is directly bonded to the other H atom. Because the forcefield is nonreactive, this bond remains active even as the bond length is stretched to arbitrarily large distances.

For this molecule, the bond stretch energy can be represented exactly by the following expansion:

$$U^{\text{bonded,new}}_{\text{H}_2}[d_{\text{AB}}] - U^{\text{bonded,new}}_{\text{H}_2}[d^{\text{eq}}_{\text{AB}}] = \sum_{m=1}^{\infty} k_m g_m[d_{\text{AB}}, d^{\text{eq}}_{\text{AB}}] \tag{91}$$

$$g_m[d_{\text{AB}}, d^{\text{eq}}_{\text{AB}}] = \frac{(d_{\text{AB}} - d^{\text{eq}}_{\text{AB}})^{m+1}}{d_{\text{AB}}{}^{(m+1)} + d_{\text{AB}}{}^{\text{eq}(m+1)}} \tag{92}$$

Note that $g_m = 0$ and $\partial g_m/\partial d_{\text{AB}} = 0$ at $d_{\text{AB}} = d^{\text{eq}}_{\text{AB}}$; $g_m = -1$ (if m is odd) or +1 (if m is even) at $d_{\text{AB}} = 0$, and $g_m = 1$ at $d_{\text{AB}} = \infty$. Eqn (92) is not the only possible choice of flexibility terms to expand the bond stretch energy, but it is a reasonable and workable choice. The $d_{\text{AB}} = d^{\text{eq}}_{\text{AB}}$ datapoint yields

$$U^{\text{bonded,new}}_{\text{H}_2}[d^{\text{eq}}_{\text{AB}}] = E^{\text{el}}_{\text{singlet}}[d^{\text{eq}}_{\text{AB}}] \tag{93}$$

Truncating the summation in eqn (91) at $m = 18$ yields 18 force constants that can be computed by using linear regression to fit the bonded interaction model to a training dataset containing the 18 $d_{\text{AB}} \neq d^{\text{eq}}_{\text{AB}}$ quantum-mechanically-computed E_{singlet} datapoints from Table S1 of ESI.† To handle the multi-collinearity issue, I performed this linear regression using the LASSO method. This minimized the following loss function:

$$L = \sum_{i=1}^{18}\left(U^{\text{bonded,new}}_{\text{H}_2}[d_i] - E^{\text{el}}_{\text{singlet}}[d_i]\right)^2 + \lambda\sum_{m=1}^{18}|k_m| \tag{94}$$

where λ is the LASSO regularization parameter.

The Matlab lasso function was used with the following settings: intercept = false, standardize = false, RelTol = 10^{-8}, MaxIter = 10^9. As lambda decreased, the number of nonzero parameters increased and the root-mean-squared-error (RMSE) of the training dataset decreased. Table S2 of ESI† lists the resulting force constant values for several values of λ. The optimized force constants had the following numbers of nonzero values: 7 for $\lambda = 10^{-8}$, 9 for $\lambda = 10^{-9}$, 12 for $\lambda = 10^{-10}$, 17 for $\lambda = 10^{-11}$, and 18 for $\lambda \le 10^{-12}$. The sum of absolute values of the force constants increased as λ decreased.

Due to the multicollinearity issue, the solution for $\lambda = 0$ is ill-defined. Instead, one generally refers to a $\lambda \rightarrow 0$ result that means the smallest value of λ for which converged results were computed. As the LASSO regularization parameter λ becomes closer to zero in value, the number of iterations allowed for convergence (*i.e.*, MaxIter) needs to be increased and the RelTol needs to be decreased. Here, the tightest convergence achieved was for RelTol = 10^{-9}, MaxIter = 10^{10}, and $\lambda = 10^{-20}$ as shown in the last column of Table S1 of ESI.†

As shown in Table S1 of ESI,† this fitted forcefield nearly reproduced the quantum-mechanically-computed training data. The RMSE values (in hartree) for the training dataset decreased monotonically from 2.0×10^{-4} for $\lambda = 10^{-8}$ to 3.4×10^{-5} for $\lambda = 10^{-20}$. Fig. 3 plots the QM-computed singlet and triplet energies for the H_2 molecule as a function of bond length. For the singlet state, the model forcefield curves are shown for comparison and are in extremely close agreement to the FCI/aug-cc-pVQZ spin singlet data. However, the $\lambda = 10^{-20}$ model behaves erratically in the extrapolated region for bond lengths <40 pm. This clearly demonstrates that using λ values too close to zero causes the over-fitting problem that decreases the model's accuracy for describing regions outside the training data.

In summary, this example illustrates a practical implementation of parameterizing a formally exact nonreactive forcefield for an isolated bonded cluster. Formal exactness means that by improving the computational accuracy and precision, the parameterized model can be made infinitesimally close to the exact solution without having to leave the theoretical framework. Some ways to increase the computational accuracy and/or precision include:

(1) Increasing the basis set size is one aspect of ‘improving the computational accuracy and precision’. Although the aug-cc-pVQZ basis set already gets fairly close to the complete basis set limit for this molecule, increases in the basis set size would enable the quantum chemistry results to get even closer to the complete basis set limit. As the basis set size tends towards infinite, the complete basis set limit can be reached.

(2) The real number models used in the quantum chemistry calculations and the linear regression calculations have a finite storage size (*e.g.*, 64 bit real numbers) that determines the number of stored digits. Increasing the number of stored digits would enable these calculations to get even closer to the exact solution.

(3) Iterative quantum chemistry calculations (such as the FCI calculations performed here) employ convergence tolerances that allow the energy to be computed to some finite number of significant digits. Tightening these convergence tolerances would enable the quantum chemistry results to get even closer to the exact solution.

(4) The LASSO method for performing regularized linear regression uses a convergence tolerance, maximum number of

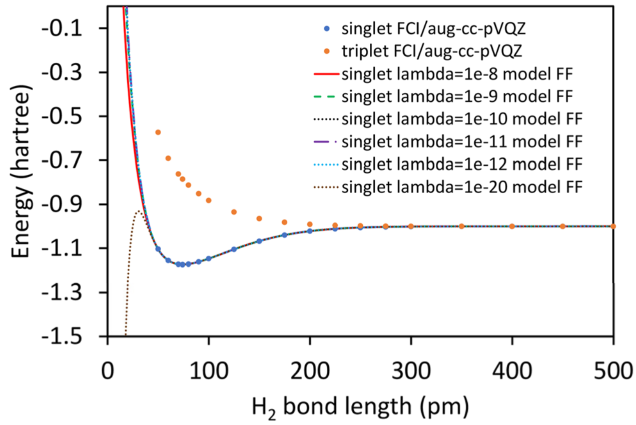

Fig. 3 Born–Oppenheimer potential energy surface for the H_2 molecule. Quantum chemistry results are compared to fitted forcefield models.

allowed iterations, and a regularization parameter (λ). $\lambda > 0$ values increase the model's transferability, robustness, and conciseness at the expense of introducing some approximation. To solve the linear regression problem exactly, an infinitesimal $\lambda \rightarrow 0$ solution is required together with extremely tight convergence tolerances, and this might require an extremely large number of iterations to converge.

(5) Here, the bonded interaction series expansion was truncated and a finite number of quantum-mechanically-computed datapoints were used to train the forcefield model. To achieve exactness for all geometries within the relevant connected region of the potential energy landscape, the training dataset would need to be expanded to include all such geometries (an infinite number) and an untruncated bonded interaction series expansion would need to be used. Additionally, the linear regression problem would need to be solved exactly so that the forcefield model exactly reproduced the training dataset.

My new ansatz for separating bonded interactions from nonbonded interactions works for developing forcefields using either machine-learning or non-machine-learning approaches. The example studied in this section used a series expansion containing many flexibility terms. Series expansion approaches can be useful to parameterize machine-learned forcefields that have been used to study many materials.[102–107] A key advantage of machine-learning strategies is that they can be applied across a wide range of different systems without requiring as much manual human labor to develop an effective working model. Such machine-learning methods allow high accuracy to be reached at the expense of typically requiring a relatively large number of fitted parameters.[102–107]

However, it is often desirable to construct and use frugal forcefields that contain relatively small numbers of fitted parameters associated with carefully chosen physically-motivated forcefield terms. This is desirable, because atomistic simulations (*e.g.*, classical molecular dynamics and Monte Carlo simulations) using frugal forcefields can run quicker than atomistic simulations employing parameter-heavy forcefields containing a relatively large number of forcefield terms. The next section revisits this example using a bond stretch model potential that achieves high accuracy using a small number of fitted parameters.

2.6.2 A new first-principles-derived bond stretch potential. The harmonic stretch model potential is simple to apply, but it only describes the shape of the bond stretch curve for small magnitude displacements (both bond compression and elongation) near the optimized bond length. As a bond is stretched to ever larger values, the energy of the harmonic bond stretch model potential increases proportional to the square of the displacement length, eventually becoming infinitely large in energy as the bond length is stretched to infinity.

For practical applications, it is often desirable to use a bond stretch model potential having a realistic shape. Here, I introduce a new first-principles-derived stretch model potential:

$$U_{\mathrm{AB}}^{\mathrm{Manz\_stretch}} = \frac{3k_{\mathrm{AB}}}{5\mathring{\gamma}_{\mathrm{AB}}^{2}}\left(1-\left(\frac{5}{2}\right)\exp\left[-\mathring{\gamma}_{\mathrm{AB}}\left(d_{\mathrm{AB}}-d_{\mathrm{AB}}^{\mathrm{ref}}\right)\right] + \left(\frac{3}{2}\right)\exp\left[-\frac{5}{3}\mathring{\gamma}_{\mathrm{AB}}\left(d_{\mathrm{AB}}-d_{\mathrm{AB}}^{\mathrm{ref}}\right)\right]\right) \quad (95)$$

The well-known Morse[48] potential

$$U_{\mathrm{AB}}^{\mathrm{Morse\_stretch}} = \frac{k_{\mathrm{AB}}}{2\gamma_{\mathrm{AB}}^{2}}\left(1-2\exp\left[-\gamma_{\mathrm{AB}}\left(d_{\mathrm{AB}}-d_{\mathrm{AB}}^{\mathrm{ref}}\right)\right]+\exp\left[-2\gamma_{\mathrm{AB}}\left(d_{\mathrm{AB}}-d_{\mathrm{AB}}^{\mathrm{ref}}\right)\right]\right) \quad (96)$$

has a related form, but with different coefficients and exponents. Although the Morse potential was originally proposed based on empirical arguments, later authors provided some physically-based rationalizations for its form.[108,109] As the bond length is stretched to infinity, these stretch potentials approach the predicted bond dissociation energy of

$$E_{\mathrm{dissociation}}^{\mathrm{Morse\_stretch}} = \frac{k_{\mathrm{AB}}}{2\gamma_{\mathrm{AB}}^{2}} \quad (97)$$

$$E_{\mathrm{dissociation}}^{\mathrm{Manz\_stretch}} = \frac{3k_{\mathrm{AB}}}{5\mathring{\gamma}_{\mathrm{AB}}^{2}} \quad (98)$$

The Morse and Manz stretch potentials have the same number of parameters, but they have different numbers of empirically-fitted parameters. In the Morse stretch potential, the exponent γ_{AB} is an empirically-fitted regression parameter that normally requires nonlinear optimization. In the Manz stretch potential, $\mathring{\gamma}_{\mathrm{AB}}$ is a quantum-mechanically-computed physical property not an empirically-fitted regression parameter. My new stretch potential provides a good tradeoff between accuracy and computational cost without requiring nonlinear regression when used with my new ansatz for separating bonded from nonbonded interactions. Analytic first-order through four-order derivatives of $U_{\mathrm{AB}}^{\mathrm{Manz\_stretch}}[d_{\mathrm{AB}}]$ are listed in ESI Section S3.†

Using my ansatz for separating bonded from nonbonded interactions, $d_{\mathrm{AB}}^{\mathrm{ref}}$ in eqn (95) or (96) always equals the (experimentally-measured or quantum-mechanically-computed) equilibrium bond length, $d_{\mathrm{AB}}^{\mathrm{eq}}$, in the isolated cluster's optimized geometry. Since my approach enables both $\mathring{\gamma}_{\mathrm{AB}}$ and $d_{\mathrm{AB}}^{\mathrm{ref}} = d_{\mathrm{AB}}^{\mathrm{eq}}$ to be computed directly, this facilitates using my stretch potential with forcefield parameterization protocols employing linear regression methods. In contrast, the old ansatz (for separating bonded from nonbonded interactions) requires the value of $d_{\mathrm{AB}}^{\mathrm{ref}}$ in eqn (95) or (96) to be an adjustable parameter $d_{\mathrm{AB}}^{\mathrm{resting}}$ that must be optimized using nonlinear regression techniques. However, iff the molecule is so small (*e.g.*, diatomic and triatomic molecules) that all intracluster nonbonded interactions are excluded, then in this limiting case $d_{\mathrm{AB}}^{\mathrm{resting}} = d_{\mathrm{AB}}^{\mathrm{eq}}$ even under the old ansatz.

My stretch potential is derived *via* the following observations and steps:

(1) Consider a chemical system comprised of atomic nuclei and electrons with no externally applied fields (no external potentials). Within the Born–Oppenheimer approximation, this system's electronic energy is the sum of the electronic kinetic energy and the nuclear plus electronic potential energies:

$$E_{\mathrm{electronic}}^{0} = T_{\mathrm{el}} + V_{\mathrm{e-e}} + V_{\mathrm{nucl-e}} + V_{\mathrm{nucl-nucl}} \quad (99)$$

When the AB bond length is at its optimized value,

$$\frac{\partial E_{\mathrm{electronic}}^{0}}{\partial d_{\mathrm{AB}}} = 0 \quad (100)$$

(2) The electron density of an isolated atom decays approximately exponentially in the atom's outer valence region (*i.e.* for large distances r_{A} from the atom's nucleus), and the value of this decay exponent b_{A} relates to the isolated atom's first ionization energy (I.E.) *via* I.E. $\approx -b_{\mathrm{A}}^{2}/8$.[110–112]

(3) The two major paradigms for assigning atoms in materials are: (a) the overlapping atoms-in-materials paradigm and (b) the non-overlapping atom-in-materials paradigm. Quantum Chemical Topology (QCT), which contains Bader's quantum theory of atoms in molecules (QTAIM) as a subpart, is currently the main theoretical and computational framework amongst those methods within the non-overlapping atoms-in-materials paradigm.[113–120] The Standard Atoms in Materials Framework (SAMF), which contains the Density-Derived Electrostatic and Chemical (DDEC) methods as a subpart, is currently the most accurate and versatile theoretical and computational framework within the overlapping atoms-in-materials paradigm.[121–123] To date, the best performing electron-density partitioning methods within the overlapping atom-in-materials paradigm assign atom-in-material electron density distributions $\{\rho_{\mathrm{A}}[\vec{r}_{\mathrm{A}}]\}$ such that their spherical averages $\{\rho_{\mathrm{A}}^{\mathrm{avg}}[r_{\mathrm{A}}]\}$ decay approximately exponentially with increasing distance (r_{A}) from the atom's nucleus.[9,122,124–127] Among these approaches, the DDEC6 method is the current state of the art.[9,60,61,128]

(4) It is useful to construct a set of atomic orbitals in materials (AOIMs) that describe the effective electronic spin–orbitals of each atom in the material according to the following criteria. Criterion 1: each AOIM is the product of an orbital function and a spin ket, and the orbital function is an exact atomic orbital angular momentum eigenfunction. This means that each AOIM

is a pure spherical harmonic function $Y_{\ell,m}[\theta_A, \phi_A]$ times a radial function $\zeta_i^A[r_A]$ times a spin ket $|s_i^A\rangle$.

$$\mathrm{AOIM}_i^A = \zeta_i^A[\vec{r}]\left|s_i^A\right\rangle = Y_{\ell,m}[\theta_A, \phi_A]\xi_i^A[r_A]\left|s_i^A\right\rangle \quad (101)$$

Criterion 2: AOIMs on the same atom are orthonormal to each other:

$$\left\langle \zeta_i^A[\vec{r}] \middle| \zeta_j^A[\vec{r}] \right\rangle \left\langle s_i^A \middle| s_j^A \right\rangle = \delta_{ij} \quad (102)$$

(AOIMs on two different atoms can have non-zero overlap) Criterion 3: these AOIMs have electron populations that satisfy the Pauli[129] exclusion principle:

$$0 \le n_i^{A,\uparrow}, n_i^{A,\downarrow} \le 1 \quad (103)$$

where $n_i^{A,\uparrow}$ and $n_i^{A,\downarrow}$ are the number of spin-up and spin-down electrons, respectively, on atom A occupying $\zeta_i^A[\vec{r}]$. Criterion 4: the occupation-weighted electron densities of AOIMs on atom A sum to the assigned atom-in-material electron density of atom A:

$$\rho_A[\vec{r}_A] = \sum_i n_i^A \zeta_i^{A*}[\vec{r}]\zeta_i^A[\vec{r}] \quad (104)$$

where

$$n_i^A = n_i^{A,\uparrow} + n_i^{A,\downarrow} \quad (105)$$

and * denotes complex conjugation. This requires that we include enough AOIMs so that all of the electron density will get projected onto the combined set of AOIMs. Criterion 5: the AOIMs are constructed according to some scheme that gives them good transferability for small changes in the material's geometry. This means that the shapes and occupations of individual AOIMs do not change drastically when a bond is slightly stretched or compressed. Of course, the center of each AOIM will move along with the position $\vec{R}_A$ of the atom to which it belongs. (Criterion # 5 can be satisfied by optimizing each AOIM to be an approximate energy eigenfunction. Together, these five criteria make each AOIM resemble atomic 1s, 2s, 2p, 3s, 3p, 3d, *etc.* orbitals. I recently developed such a method and programmed it into the Chargemol code. The details will be published in future work.)

(5) In the limit $\vec{r} \to \vec{r}'$, the products $\mathrm{AIOM}_i^{A*}[\vec{r}]\mathrm{AIOM}_i^A[\vec{r}']$ and $\mathrm{AIOM}_j^B[\vec{r}]\mathrm{AIOM}_j^{B*}[\vec{r}']$ approach the electron density in these orbitals. Because each AOIM approximately equals a polynomial function of radius times an exponential decay function, its spherically averaged electron density scales like:

$$\left(\xi_i^A[r_A]\right)^2 \propto (\mathrm{polynomial}[n, \ell, r_A])^2 \mathrm{e}^{-2\varpi_A^i} \quad (106)$$

Polynomial$[n, \ell, r_A]$ causes each AOIM to oscillate such that it has $(n - \ell)$ radial nodes, where n is the AOIM's principle quantum number and ℓ is its orbital angular momentum quantum number. For example, the 4p AOIM has $(4 - 1) = 3$ radial nodes. For valence AOIMs, these nodes occur in the core and semi-core regions. Well beyond the last radial node for large r_A (*i.e.*, in the valence region between atoms A and B), we can absorb the polynomial dependence into an effective exponent that approximately equals the effective decay exponent of the atom's valence density:

$$b_A = \left.\frac{\partial \ln[\rho_A^{\mathrm{avg}}[r_A]]}{\partial r_A}\right|_{\mathrm{valence}\ r_A} \quad (107)$$

$$(\xi_i^A[r_A])^2|_{\mathrm{valence}\ r_A} \propto \mathrm{e}^{-b_A r_A} \quad (108)$$

(6) The energy of the AB bond is affected by exchange, kinetic energy, coulombic, and dispersion interactions between AOIMs on atom A and AOIMs on atom B. Depending on the circumstances and the value of d_{AB}, the sum of energy contributions to bonding could be net attractive or net repulsive. For simplicity, we can classify these energy contributions into two major groups: (i) Group # 1 comprises interatomic kinetic and potential energy changes proportional to $\mathrm{AIOM}_i^{A*}[\vec{r}_A]\mathrm{AIOM}_j^B[\vec{r}_B]$. By the triangle distance inequality,

$$r_A + r_B \ge d_{AB} \quad (109)$$

Due to the exponential decay of these AOIMs' radial functions, the largest contributions to $\mathrm{AIOM}_i^{A*}[\vec{r}_A]\mathrm{AIOM}_j^B[\vec{r}_B]$ occur for points satisfying

$$r_A + r_B \approx d_{AB} \quad (110)$$

which correspond to points near the bond's axis. This allows us to approximate this type of term as

$$\mathrm{Sum}\left(\mathrm{AIOM}_i^{A*}[\vec{r}_A]\mathrm{AIOM}_j^B[\vec{r}_B]\mathrm{energy\ terms}\right) \propto \mathrm{e}^{-\mathring{\gamma}_{AB}\left(d_{AB} - d_{AB}^{\mathrm{eq}}\right)} \quad (111)$$

(ii) Group # 2 comprises the short-range repulsion (SRR) energy due to Pauli's[129] exclusion principle. When two atoms overlap, their electron orbitals must deform (change) to retain orthogonality between all of the molecular orbitals. This raises the energy of the electrons, thus leading to a repulsive force between the overlapping electron clouds. For $d_{AB} \ll d_{AB}^{\mathrm{eq}}$, the SRR term dominates the energy function. This SRR energy contains both kinetic energy and potential energy contributions and has an exponential dependence on d_{AB} with a decay exponent approximately equal to 0.83 (*i.e.*, (5/6)) times some weighted average between b_A and b_B:[8]

$$\mathrm{SRR} \propto \mathrm{e}^{-(5/6)\mathrm{weighted\_avg}[b_A, b_B]\left(d_{AB} - d_{AB}^{\mathrm{eq}}\right)} \quad (112)$$

(7) A precise value for the exponent $\mathring{\gamma}_{AB}$ can be derived by noting the scaling behavior of the overlap integral between two Slater functions

$$\mathrm{Overlap}\left[a_A, a_B, b_A, b_B, d_{AB} = \left|\vec{R}_A - \vec{R}_B\right|\right] = \oint \mathrm{e}^{a_A - b_A r_A}\ \mathrm{e}^{a_B - b_B r_B} d^3\vec{r} \quad (113)$$

When $b_A \approx b_B$, this integral has the value[8]

$$b_{\mathrm{eff}} \approx \sqrt{b_A b_B} \approx (b_A + b_B)/2 \quad (114)$$

$$\text{Overlap}[a_A, a_B, b_{\text{eff}}, d_{AB}] = e^{a_A+a_B}\left(\frac{\pi}{{b_{\text{eff}}}^3}\right)\left(\frac{(b_{\text{eff}}d_{AB})^2}{3} + b_{\text{eff}}d_{AB} + 1\right)e^{-b_{\text{eff}}d_{AB}} \quad (115)$$

On the other hand, when $0 < b_A \ll b_B$, then atom B looks almost like a point charge distribution so that the overlap becomes

$$\text{Overlap}[a_A, a_B, b_A \ll b_B, d_{AB}] \approx e^{a_A - b_A d_{AB}} N_B \quad (116)$$

where

$$N_B = \oint e^{a_B - b_B r_B} d^3\vec{r} = e^{a_B} 8\pi / {b_B}^3 \quad (117)$$

is the volume integral of the Slater function on atom B. For the overlap between the products of valence AOIMs, $\text{AOIM}_i^{A*}[\vec{r}_A]\text{AOIM}_j^B[\vec{r}_B]$, the effective decay exponents of these valence AOIMs is approximately $b_A/2$ and $b_B/2$, respectively. Choosing

$$\mathring{\gamma}_{AB} = \frac{{b_A}^{-1} + {b_B}^{-1}}{2({b_A}^{-2} + {b_B}^{-2})} = \frac{b_A b_B (b_A + b_B)}{2({b_A}^2 + {b_B}^2)} \quad (118)$$

yields the appropriate limits:

$$\min[(1/2)b_A, (1/2)b_B] \leq \mathring{\gamma}_{AB} \leq \max[(1/2)b_A, (1/2)b_B] \quad (119)$$

$$\underset{0 < b_A \ll b_B}{\text{when}} \ \mathring{\gamma}_{AB} \approx b_A/2 \quad (120)$$

$$\underset{b_A = (b_B + \varepsilon)}{\text{when}} \ \mathring{\gamma}_{AB} = (b_A + b_B)/4 \text{ for small } \varepsilon > 0 \quad (121)$$

(8) Expressed as a Taylor series:

$$\left(E^0_{\text{electronic}}[d^{\text{eq}}_{AB} + \Delta d_{AB}] - E^0_{\text{electronic}}[d^{\text{eq}}_{AB}]\right) = \Delta d_{AB} \frac{\partial E^0_{\text{electronic}}}{\partial d_{AB}} + \frac{1}{2}(\Delta d_{AB})^2 \frac{\partial^2 E^0_{\text{electronic}}}{\partial {d_{AB}}^2} + \text{h.o.t.} \quad (122)$$

Because of eqn (100), the first non-zero term in eqn (122) is proportional to $(\Delta d_{AB})^2$ instead of Δd_{AB}. Assembling the above results means the energy equation should take the form

$$\left(E^0_{\text{electronic}}[d_{AB}] - E^0_{\text{electronic}}[d^{\text{eq}}_{AB}]\right) = \text{coeff\_1}\left(1 - \text{coeff\_2}\ e^{-\mathring{\gamma}_{AB}(d_{AB} - d^{\text{eq}}_{AB})} + \text{coeff\_3}\ e^{-(5/3)\mathring{\gamma}_{AB}(d_{AB} - d^{\text{eq}}_{AB})}\right) \quad (123)$$

Since the right-hand side must equal zero when $d_{AB} = d^{\text{eq}}_{AB}$, this means

$$1 - \text{coeff\_2} + \text{coeff\_3} = 0 \quad (124)$$

Since the force (and first derivative of energy) must be zero when $d_{AB} = d^{\text{eq}}_{AB}$, this means

$$\text{coeff\_2} - (5/3)\text{coeff\_3} = 0 \quad (125)$$

Solving these two linear equations gives coeff_2 = 5/2 and coeff_3 = 3/2. Defining the force constant by

$$k_{AB} = \frac{d^2 E}{d{d_{AB}}^2} \quad (126)$$

gives

$$\text{coeff\_1} = \frac{3k_{AB}}{5{\mathring{\gamma}_{AB}}^2} \quad (127)$$

With these values for the coefficients, eqn (123) becomes my new stretch potential (eqn (95)).

My stretch potential is conceptually related to bond order changes. Bond orders computed using my bond order equation applied with DDEC6 partitioning (aka Manz/DDEC6 bond orders) decay approximately exponentially as the bond length is stretched beyond its equilibrium value.[128] My bond order equals the number of electrons that are dressed exchanged between two atoms in a material.[128] This bond order is also between approximately 1× and 2× the contact exchange (and DDEC6 overlap population) which also decay approximately exponentially as the bond length is stretched beyond its equilibrium value.[128] For bonds having similar type, the integrated crystal orbital Hamilton population (ICOHP) strongly correlated to the computed Manz/DDEC6 bond orders in various materials.[130] Moreover, Pauling proposed an empirical bond-distance-to-bond-order correlation in which the bond order decreases exponentially as the bond length increases.[131]

I now introduce a straightforward algorithm to compute b_A and b_B for the AB bond. First, we perform a quantum chemistry calculation on the material's optimized geometry. Then, we perform DDEC analysis on the quantum chemistry results. Starting with the $\{\rho^{\text{avg}}_A[r_A]\}$ printed by the Chargemol code, we first make sure these are monotonically decreasing functions by imposing

$$\tilde{\rho}^{\text{avg}}_A[r^{\text{nshells}}_A] = \rho^{\text{avg}}_A[r^{\text{nshells}}_A] \quad (128)$$

$$\tilde{\rho}^{\text{avg}}_A[{r_A}^j] = \max[\rho^{\text{avg}}_A[{r_A}^j], \tilde{\rho}^{\text{avg}}_A[{r_A}^{j+1}]] \quad \text{for } j = (\text{nshells} - 1), (\text{nshells} - 2), \ldots 1 \quad (129)$$

starting with the outer radial shell and proceeding inward to successively smaller radial shells. This procedure is done for all atoms in the material's unit cell. We next identify which atoms in the material are directly bonded to each other (*i.e.*, are nearest neighbors in the bond connectivity graph) using a chosen method. For example, we could consider two atoms to be directly bonded to each other iff the distance between them was no greater than the sum of their element-dependent atom-typing radii:[10]

$$d_{AB} \leq R^{AT}_A + R^{AT}_B \quad (130)$$

Alternatively, one could consider two atoms A and B to be directly bonded to each other if their bond order BO_{AB}, overlap population OP_{AB}, or contact exchange CE_{AB} is above a chosen threshold value. For a bond pair AB, find (a) the smallest value of r_A for which $\tilde{\rho}_B^{avg}[d_{AB} - r_A] > \tilde{\rho}_A^{avg}[r_A]$ and (b) the largest value of r_A for which $\tilde{\rho}_B^{avg}[d_{AB} - r_A] < \tilde{\rho}_A^{avg}[r_A]$; let D_A^{AB} equal the average of these two r_A values. Next, we perform linear regression to fit the model $a_A^{AB} - b_A^{AB} r_A$ to the datapoints $\ln[\tilde{\rho}_A^{avg}[r_A^j]]$ over the set of radial shells satisfying

$$((D_A^{AB} - 0.5 \text{ bohr})) \leq r_A^j \leq (D_A^{AB} + 2.5 \text{ bohr}) \quad (131)$$

The sampled range of $[r_A^j]$ values is asymmetric about D_A^{AB} to emphasize the outer valence region of atom A. This ensures the computed exponent b_A^{AB} is fairly characteristic of the $\tilde{\rho}_A^{avg}[r_A^j]$ decay behavior over the most relevant range of r_A values. This process is repeated for atom B in the AB pair to get b_B^{AB}. These are plugged into eqn (118) to compute $\mathring{\gamma}_{AB}$. Finally, this process is repeated for all bond pairs in the material.

This process assigns a different D_A^{AB} and hence different b_A^{AB} value for each different bond connected to atom A. For example, in the acetonitrile molecule ($H_3C–C{\equiv}N$), the central C atom is singly bonded to another carbon atom and triply bonded to a nitrogen atom. This has the effect of making the central carbon atom's D_A^{AB} slightly smaller for the triple bond than for the single bond, which means the resulting b_A^{AB} is fitted over slightly smaller r_A values for the triple bond compared to the single bond. As another example, we could consider a hydrogen atom that is covalently bonded to one oxygen atom and opportunistically 'hydrogen bonded' to another oxygen atom. Since the opportunistic 'hydrogen bond' has a greater length than the O–H covalent single bond, the procedure described above automatically fits b_A^{AB} for each bond over the relevant r_A values for that particular bond. Accordingly, this procedure should provide good results for a wide range of bond orders.

This procedure was used to analyze the stretched H_2 singlet molecule using the FCI/aug-cc-pVQZ quantum chemistry calculations introduced in the previous section. Also shown are CCSD calculations for the spin triplet O_2 molecule using the d-aug-cc-pVQZ[96–98] basis set (These were performed using Gaussian (ref. 95) software.) As shown in Fig. 4, both the Morse and Manz stretch potentials fit the quantum-mechanically-computed data well. These regression parameters were optimized in Excel using the Generalized Reduced Gradient (GRG) solver that works for both linear and nonlinear optimization problems. The Manz stretch potential has the advantage of requiring only a linear regression, while the Morse potential required a nonlinear regression to optimize its parameters.

The 'goodness of fit' (R-squared) was computed as follows:

$$R^2 = 1 - \text{SSE/SST} \quad (132)$$

where SSE = sum of squared errors and SST = sum of squares total. In this case,

$$\text{SST} = \sum_{\mu} \left(\text{E}_{\mu}^{\text{el}} - \text{E}_{\text{opt}}^{\text{el}} \right)^2 \quad (133)$$

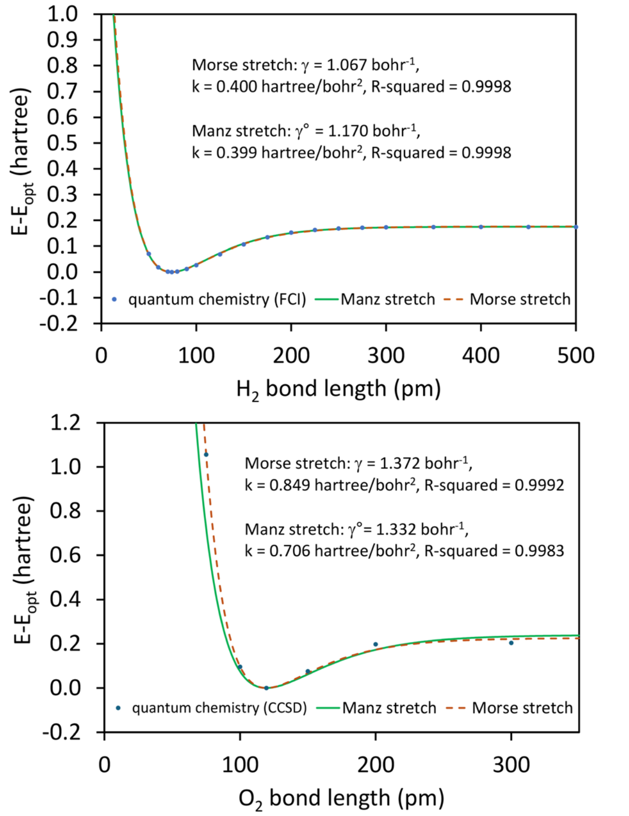

Fig. 4 Comparison of Morse and Manz stretch potentials fitted to the quantum-mechanically-computed H_2 singlet (top panel) and O_2 triplet (bottom panel) potential energy curves.

$$\text{SSE} = \sum_{\mu} \left(\left(E_{\mu}^{\text{el}} - E_{\text{opt}}^{\text{el}} \right) - \left(U_{\mu}^{\text{FF}} - U_{\text{opt}}^{\text{FF}} \right) \right)^2 \quad (134)$$

where the summation runs over the geometries in the dataset.

By fitting the empirical Morse stretch potential to the first-principles-derived Manz stretch potential, a hack was developed to accurately estimate the Morse potential exponent $\gamma_{AB}^{\text{Morse}}$. Equating the Morse potential's force constant (k) and dissociation energy ($E_{\text{dissociation}}^{\text{Morse\_stretch}}$, eqn (97)) to those of the Manz stretch potential (eqn (98)) yields:

$$\gamma_{AB}^{\text{Morse}} \approx \mathring{\gamma}_{AB} \sqrt{\frac{5}{6}} = \sqrt{\frac{5}{6}} \left(\frac{b_A b_B (b_A + b_B)}{2({b_A}^2 + {b_B}^2)} \right) \quad (135)$$

For the H_2 molecule, this equation predicts $\gamma_{AB}^{\text{Morse}} \approx 1.068$ bohr^{-1} compared to the optimized value of 1.067 bohr^{-1}. For the O_2 molecule, this equation predicts $\gamma_{AB}^{\text{Morse}} \approx 1.216$ bohr^{-1} compared to the optimized value of 1.372 bohr^{-1}. If using these predicted exponents in the Morse potential, the force constant k_{Morse} would be optimized (using linear regression) to yield 0.401 (H_2) and 0.952 (O_2) hartree per bohr^2 with R-squared values of 0.9998 (H_2) and 0.9881 (O_2).

The above procedure does not require a combining (aka 'mixing') rule that relates $\gamma_{AB}^{\text{Morse}}$ and $\gamma_{AB}^{\text{Manz}}$ to $\gamma_{AA}^{\text{Morse}}$, $\gamma_{BB}^{\text{Morse}}$, $\gamma_{AA}^{\text{Manz}}$, and $\gamma_{BB}^{\text{Manz}}$. In the above procedure,

$\gamma_{AB}^{\text{Morse}} = \overset{\circ}{\gamma}_{AB}\sqrt{5/6}$ and $\gamma_{AB}^{\text{Manz}} = \overset{\circ}{\gamma}_{AB}$ are extracted directly from quantum chemistry calculations. From eqn (118) and (135), the following combining rules emerge as approximations:

$$\gamma_{AB}^{\text{Morse}} \approx \frac{\gamma_{AA}^{\text{Morse}}\gamma_{BB}^{\text{Morse}}\left(\gamma_{AA}^{\text{Morse}} + \gamma_{BB}^{\text{Morse}}\right)}{\left(\gamma_{AA}^{\text{Morse}}\right)^2 + \left(\gamma_{BB}^{\text{Morse}}\right)^2} \tag{136}$$

$$\gamma_{AB}^{\text{Manz}} \approx \frac{\gamma_{AA}^{\text{Manz}}\gamma_{BB}^{\text{Manz}}\left(\gamma_{AA}^{\text{Manz}} + \gamma_{BB}^{\text{Manz}}\right)}{\left(\gamma_{AA}^{\text{Manz}}\right)^2 + \left(\gamma_{BB}^{\text{Manz}}\right)^2} \tag{137}$$

2.6.3 A bond stretch in the hexafluorobenzene molecule. We now consider the C–F bond stretch force constant in the C_6F_6 molecule. This molecule was chosen for two reasons:

(1) First, its symmetry means there is only one independent atom-in-material charge value. Specifically, if q_C is the net atomic charge assigned to each carbon atom, then $q_F = -q_C$ is the net atomic charge assigned to each fluorine atom.

(2) Second, this molecule includes first-, second-, third-, fourth-, and fifth-nearest neighbors. This ensures that some nonbonded interactions will still be present even when 1–2 (*i.e.*, first-neighbor), 1–3 (*i.e.*, second-neighbor), and/or 1–4 (*i.e.*, third-neighbor) nonbonded interactions are excluded in the forcefield model.

First, the geometry was fully optimized in Gaussian 16 (ref. 95) using the B3LYP[82,132,133] exchange–correlation functional and def2-TZVPD[134] basis set. The geometry was optimized to tight convergence criteria (*i.e.*, Gaussian keyword opt = tight). In the fully optimized geometry, the optimized bond lengths were 1.332 (C–F) and 1.389 Å (C–C), the optimized bond angles were 120.0° (both C–C–C and C–C–F), and all atoms were in the same plane.

Next, single-point energies were computed when the length of one C–F bond was changed by −0.14, −0.07, +0.07, and +0.14 Å compared to its length in the fully optimized geometry. This was done by moving one F atom in the C_6F_6 plane while holding the positions of all other atoms rigid at the same positions they had in the fully optimized geometry. All bond angle values were the same as in the fully optimized geometry. Since all atoms remained in the plane, no changes in dihedral values occurred. In summary, only the value of one internal coordinate (*i.e.*, the length of one and only one C–F bond) changed. This allows us to isolate the energy change of a single flexibility term; namely, the bond stretch for this one bond as shown in Fig. 5.

I wrote and used the calculate_Manz_and_Morse_stretch_potential_exponents program to compute the exponent $\gamma°$. This program uses the computational algorithm described in Section 2.6.2 above to analyze the DDEC6-computed $\{\rho_A^{\text{avg}}[r_A]\}$ printed from the Chargemol[60] program. The calculate_Manz_and_Morse_stretch_potential_exponents program can handle any number (*i.e.*, 0, 1, 2, or 3) of periodic boundary conditions and works for molecules, dense and porous solids, solid surfaces, polymers, nanotubes, nanosheets, ionic and covalent materials, opportunistically-hydrogen-bonded materials, magnetic and non-magnetic materials, *etc.* For the C–F bond in C_6F_6, the result was $\gamma° = 1.207$ bohr^{-1} using the fully optimized geometry.

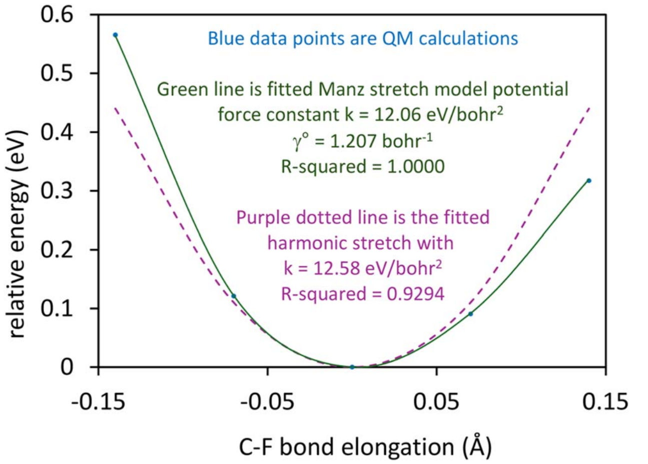

Fig. 5 Quantum-mechanically-computed Born–Oppenheimer potential energy curve for the C–F bond stretch in the C_6F_6 molecule. Fitted harmonic and Manz stretch model potentials are shown for comparison.

The stretch force constants were optimized by minimizing the following least-squares loss function:

$$L = \sum_{\substack{\mu\in\\ \text{training}\\ \text{geoms}}} \left(\left(E_\mu^{\text{el}} - E_{\text{opt}}^{\text{el}}\right) - \left(U_\mu^{\text{FF}} - U_{\text{opt}}^{\text{FF}}\right)\right)^2 \tag{138}$$

In eqn (138), μ is the geometry number in the training dataset (In this example, there are a total of four nonequilibrium geometries in the training dataset.). E_μ^{el} is the quantum-mechanically-computed energy of geometry μ. $E_{\text{opt}}^{\text{el}}$ is the quantum-mechanically-computed energy of the fully optimized ground-state geometry. U_μ^{FF} is the forcefield model's potential energy of geometry μ. $U_{\text{opt}}^{\text{FF}}$ is the forcefield model's potential energy of the fully optimized ground-state geometry. These optimizations were performed using Excel's GRG solver.

The forcefield's potential energy was expanded as the sum of bonded interactions plus non-bonded interactions:

$$U_\mu^{\text{FF}} = U_\mu^{\text{bonded}} + U_\mu^{\text{nonbonded}} \tag{139}$$

The atomic charges plus Lennard-Jones parameters model $U_{AB}^{(q+\text{LJ})}$ (eqn (9)) was used for these nonbonded interactions. Comparisons were made using different values for the atomic charges and Lennard-Jones parameters. No cutoff distance for the nonbonded interactions ($d_{\text{cutoff}}^{\text{nonbonded}}$) was used for these calculations. Models were built and compared using the harmonic stretch (eqn (10)) and Manz stretch (eqn (95) and (118)) potential for the bonded interaction.

Nonbonded interactions were always excluded between bonded first-neighbors (aka 1–2 interactions) and bonded second neighbors (aka 1–3 interactions). Comparisons were made between including or excluding nonbonded interactions for bonded third neighbors (aka 1–4 interactions). More remote nonbonded interactions (*e.g.*, 1–5, 1–6, *etc.*) were always included.

Comparisons were made using either the old ansatz or my new ansatz for separating bonded from non-bonded interactions. When using the old ansatz, the reference bond length, d_{AB}^{ref}, becomes a regression parameter, $d_{AB}^{resting}$, which leads to a nonlinear optimization problem. When using my new ansatz, the reference bond length, d_{AB}^{ref}, becomes the optimized bond length, d_{AB}^{eq}, which gives a linear optimization problem for both the harmonic and Manz stretch potentials. For the new scheme, $U_{ABx,intracluster}^{nonbonded} = U_{AB}^{(q+LJ)}$ was inserted into eqn (48) to compute $\Phi_{ABx}^{intracluster}$ which was inserted into eqn (26) to compute $U^{nonbonded,new}$ for each geometry. For the new scheme, the loss function of eqn (138) was minimized by varying the value of k_{AB}^{new}. For the old scheme, $U_{AB}^{nonbonded} = U_{AB}^{(q+LJ)}$ was inserted into eqn (8) to compute $U_{nonbonded}^{old}$ for each geometry. For the old scheme, the loss function of eqn (138) was minimized by varying the values of k_{AB}^{old} and $d_{AB}^{resting}$.

When using my new ansatz for separating bonded from nonbonded interactions, comparisons were made between the full potential model that included both intracluster bonded and intracluster nonbonded interactions (eqn (138) and (139)) and the leading-order potential model that included only intracluster bonded interactions. As explained in the previous sections, the leading-order potential model accurately describes the intracluster interactions up to and including second-order derivatives of the potential energy at the isolated cluster's optimized geometry.

The leading-order potential models for the harmonic and Manz stretch potentials are plotted in Fig. 5 and equal the full model potential results when all intracluster nonbonded interactions (*e.g.*, atomic charges and Lennard-Jones potential) are set to zero. *R*-Squared values were computed using eqn (132)–(134). As demonstrated by the results shown in Fig. 5, my stretch potential (*R*-squared = 1.0000) fit the QM data nearly perfectly while the harmonic stretch potential (*R*-squared = 0.9294) did not capture the bond's significant anharmonicity. For comparison, optimized values for the Morse stretch potential were: (a) γ (optimized) = 1.093 bohr^{-1}, k (optimized) = 12.01 eV bohr^{-2}, giving *R*-squared = 1.0000, and (b) γ (predicted using eqn (135)) = 1.102 bohr^{-1}, k (optimized) = 12.00 eV bohr^{-2}, giving *R*-squared = 1.0000.

The full potential model results are summarized in Table 2 (harmonic stretch) and Table 3 (Manz stretch). Results are compared for the new and old schemes using various parameters for the nonbonded interactions. *R*-Squared values were computed using eqn (132)–(134). Tests were performed for three charge values (q_C = 0, 0.10 (DDEC6), and 0.62 (QTAIM)) both with and without Lennard-Jones (LJ) interactions. The LJ interaction parameters were taken from the Universal Force Field (UFF): d_{C-C}^{LJ} = 3.851 Å, d_{F-F}^{LJ} = 3.364 Å, $d_{F-C}^{LJ} = \sqrt{d_{C-C}^{LJ} d_{F-F}^{LJ}}$, ε_{C-C}^{LJ} = 0.105 kcal mol^{-1}, ε_{F-F}^{LJ} = 0.050 kcal mol^{-1}, $\varepsilon_{F-C}^{LJ} = \sqrt{\varepsilon_{C-C}^{LJ} \varepsilon_{F-F}^{LJ}}$.[135] Please see ref. 136–138 for a further discussion of forcefield nonbonded parameters for this molecule.

As shown in Tables 2 and 3, the new scheme gave optimized force constant values that were practically the same irrespective of the non-bonded interaction model. Under the new scheme, only third-order and higher-order derivatives (*i.e.*, anharmonicities) of the potential energy are affected by fluctuations in the intracluster nonbonded parameters. For these reasons, the bonded force constant values are less sensitive to the particular choice of intracluster nonbonded potential model under the new scheme compared to the old scheme. This is an extremely important consideration, because nonbonded parameters such as atomic-in-material (AIM) charges, AIM multipole moments, AIM polarizabilities, AIM dispersion coefficients, Lennard-Jones parameters, *etc.* carry some uncertainties in their values.

As shown in Tables 2 and 3, the old scheme yielded force constants (k_{AB}^{old}) and resting values ($d_{AB}^{resting}$) that were moderately but not severely sensitive to the choice of nonbonded interaction model. My new stretch potential was able to describe the C–F bond's potential energy curve nearly perfectly (*i.e.*, *R*-squared = 1.0000) using both the new and old schemes for all of the nonbonded interaction models tested.

The harmonic stretch potential yielded higher *R*-squared values when using the old scheme compared to when using the new scheme. This was due to the old scheme's additional regression parameter (*i.e.*, $d_{AB}^{resting}$) compared to the new scheme which uses the quantum-mechanically-computed d_{AB}^{eq} value. However, the old scheme has the disadvantage of predicting the

Table 2 Sensitivity of the new and old schemes to the nonbonded parameter model. One C–F bond in the C_6F_6 molecule was modeled using the harmonic stretch potential. The equilibrium C–F bond length is 1.332 Å

Atom charges			New scheme		Old scheme		
q_C (method)	LJ parameters	1–4 nonbonded interactions included?	k_{AB}^{new} (eV bohr^{-2})	*R*-Squared	k_{AB}^{old} (eV bohr^{-2})	$d_{AB}^{resting}$ (Å)	*R*-Squared
0 (none)	0	Y, N	12.58	0.9294	12.58	1.349	0.9920
0.10 (DDEC6)	0	Y	12.58	0.9293	12.62	1.347	0.9919
0.10 (DDEC6)	0	N	12.58	0.9294	12.58	1.347	0.9919
0.62 (QTAIM)	0	Y	12.59	0.9285	12.80	1.340	0.9918
0.62 (QTAIM)	0	N	12.58	0.9295	12.56	1.338	0.9915
0 (none)	UFF	Y	12.58	0.9294	12.61	1.350	0.9921
0 (none)	UFF	N	12.58	0.9294	12.58	1.349	0.9920
0.10 (DDEC6)	UFF	Y	12.58	0.9292	12.65	1.348	0.9920
0.10 (DDEC6)	UFF	N	12.58	0.9294	12.58	1.347	0.9919
0.62 (QTAIM)	UFF	Y	12.59	0.9285	12.83	1.341	0.9918
0.62 (QTAIM)	UFF	N	12.58	0.9295	12.56	1.338	0.9915

Table 3 Sensitivity of the new and old schemes to the nonbonded parameter model. One C–F bond in the C_6F_6 molecule was modeled using the Manz stretch potential. The equilibrium C–F bond length is 1.332 Å

Atom charges			New scheme		Old scheme		
q_C (method)	LJ parameters	1–4 nonbonded interactions included?	k_{AB}^{new} (eV bohr^{-2})	R-Squared	k_{AB}^{old} (eV bohr^{-2})	$d_{AB}^{resting}$ (Å)	R-Squared
0 (none)	0	Y, N	12.06	1.0000	12.02	1.3323[a]	1.0000
0.10 (DDEC6)	0	Y	12.06	1.0000	12.17	1.331	1.0000
0.10 (DDEC6)	0	N	12.06	1.0000	12.15	1.331	1.0000
0.62 (QTAIM)	0	Y	12.07	1.0000	12.92	1.323	1.0000
0.62 (QTAIM)	0	N	12.06	1.0000	12.82	1.321	1.0000
0 (none)	UFF	Y	12.06	1.0000	11.99	1.333	1.0000
0 (none)	UFF	N	12.06	1.0000	12.03	1.332	1.0000
0.10 (DDEC6)	UFF	Y	12.06	1.0000	12.13	1.332	1.0000
0.10 (DDEC6)	UFF	N	12.06	1.0000	12.16	1.330	1.0000
0.62 (QTAIM)	UFF	Y	12.07	1.0000	12.88	1.324	1.0000
0.62 (QTAIM)	UFF	N	12.06	1.0000	12.82	1.321	1.0000

[a] An extra significant digit is shown here to explain why the optimized k value equals 12.02 instead of 12.06.

wrong value for d_{AB}^{eq}. By construction, the new scheme yields the correct value $d_{AB}^{eq} = 1.332$ Å. For the no charges and no LJ parameters model, the old scheme yields the value $d_{AB}^{eq} = 1.349$ Å, which is close but not exact.

In summary, the new scheme is preferable to the old scheme for the following three reasons. (1) The new scheme requires only linear regression to optimize the force constants, while the old scheme sometimes (*e.g.*, Manz and Morse stretch potentials) requires nonlinear regression to optimize the force constants and resting values. (2) The new scheme gives optimized force constant values that are almost insensitive to the choice of nonbonded interaction model. (3) The new scheme exactly reproduces the material's optimized reference geometry in which all atom-in-material forces are zero.

3. A better angle-bending model potential

3.1 Derivation and comparison to other popular angle-bending model potentials

An angle bending potential ($U_{\measuredangle ABC}[\theta_{ABC}]$) models the potential energy change from a change of bond angle θ_{ABC} where atom A is bonded to atom B, and atom C is bonded to atom B. The angle θ_{ABC} is defined as:

$$\theta = \cos^{-1}\left[\min\left[\max\left[-1, \frac{\vec{R}_{BA}\cdot\vec{R}_{BC}}{|\vec{R}_{BA}||\vec{R}_{BC}|}\right], 1\right]\right] \tag{140}$$

where ($\vec{R}_{BA}$) is the vector from atom B to atom A defined as:

$$\vec{R}_{BA} = \vec{R}_A - \vec{R}_B \tag{141}$$

We used min and max functions to cancel the effect of roundoff error on the results; this forces the argument of the arccosine function to be between −1 and 1. θ_{eq} is the equilibrium value of this bond angle in the optimized ground-state geometry. The physically allowed ranges are

$$0 < \theta_{eq} \leq \pi \tag{142}$$

$$0 < \theta_{ABC} \leq \pi \tag{143}$$

$\theta_{eq} = 0$ is not physically allowed, because this would correspond to either (i) two different atoms occupying the same nuclear position which is not allowed or (ii) all three atoms being collinear (*i.e.*, in a line) and in this case θ_{eq} would be interpreted as π instead of 0. Pauli repulsion[129] (aka 'short-range repulsion'[8]) prevents θ_{ABC} from getting close to zero.

As described in prior literature, the forces exhibited by the angle-bending potential on atoms A, B, and C are related to this potential's derivative $dU_{\measuredangle ABC}[\theta_{ABC}]/d\theta_{ABC}$.[1] At the linear angle value $\theta_{ABC} = \pi$, the angle-bending force should be zero by symmetry, because the angle decreases as either atom A or atom C moves in any direction perpendicular to the starting line ABC.[5] Accordingly, a physically viable angle-bending potential satisfies the constraint

$$\left.\frac{dU_{\measuredangle}[\theta]}{d\theta}\right|_{\theta=\pi} = 0 \tag{144}$$

For derivatives of all orders to be continuous at $\theta_{ABC} = \pi$, the angle-bending potential must be symmetric about $\theta_{ABC} = \pi$, which requires:

$$U_{\measuredangle}[\theta] = U_{\measuredangle}[2\pi - \theta] \tag{145}$$

This requirement arises, because a hypothetical bond angle of $\pi + \Delta$ where $\Delta \geq 0$ is actually computed (*via* eqn (140)) to be a bond angle of $\theta = \pi - \Delta$.

If an angle-bending potential does not satisfy eqn (144), the consequent spurious force discontinuity at $\theta = \pi$ could potentially degrade the accuracy of trajectories computed using numerical integrators (*e.g.*, Verlet integration[139,140]) for molecular dynamics calculations. As shown in Fig. 6, the following currently used angle-bending potentials violate eqn (144) when $\theta_{eq} \neq \pi$:

$$U_{harmonic\_bend}[\theta] = \tfrac{1}{2}k(\theta - \theta_{eq})^2 \tag{146}$$

$$U_{cosine\_bend}[\theta] = k(1 - \cos[\theta - \theta_{eq}]) \tag{147}$$

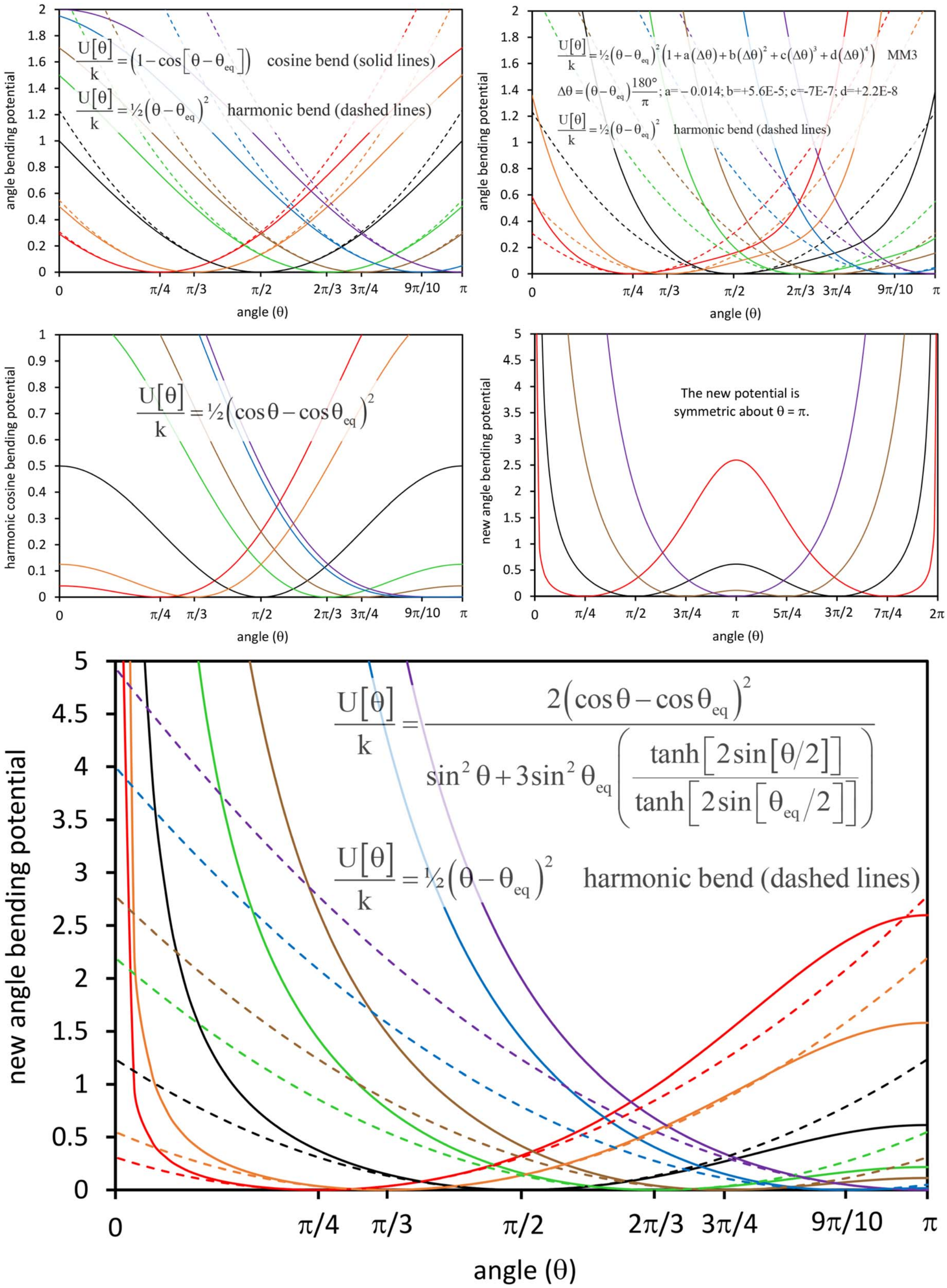

Fig. 6 The new angle-bending potential has continuous derivatives of all orders even for an angle of π radians while retaining a harmonic-like shape around the equilibrium angle (*i.e.*, $d^2U/d\theta^2|_{\theta=\theta_{eq}} > 0$) for all values of $\theta_{eq} > 0$. As shown in the top two and middle left panels, common angle-bending potentials described in prior literature do not achieve this. The middle right panel shows the new potential is symmetric about $\theta = \pi$. See the text for a complete description. The lines are colored as follows: $\theta_{eq} = \pi/4$ (red), $\pi/3$ (orange), $\pi/2$ (black), $2\pi/3$ (green), $3\pi/4$ (brown), $9\pi/10$ (blue), and π (purple).

$$U_{\text{MM3\_bend}} = (1/2)k\left(\theta - \theta_{\text{eq}}\right)^2$$

$$\left(1 + a(\Delta\theta) + b(\Delta\theta)^2 + c(\Delta\theta)^3 + d(\Delta\theta)^4\right)$$

$$\Delta\theta = \left(\theta - \theta_{\text{eq}}\right)\frac{180^\circ}{\pi};\ \ a = -0.014;\ \ b = +0.000056;$$

$$c = -0.0000007;\ \ d = +0.000000022 \tag{148}$$

The MM3 bend parameters in eqn (148) are from Allinger's MM3-2000 parameter set, which uses an updated value of the parameter d compared to the original 1989 value of Lii and Allinger.[1,141] The cosine bend in eqn (147) has not been widely used in forcefields to date; however, there have been some special cases of its use, especially for $\theta_{\text{eq}} \to \pi$.[23,142]

The harmonic cosine potential[142]

$$U_{\text{harmonic\_cosine}}[\theta] = \tfrac{1}{2}k(\cos[\theta] - \cos[\theta_{\text{eq}}])^2 \tag{149}$$

obeys eqn (144) but suffers the drawback that the potential's curvature for $\theta \to \theta_{\text{eq}}$ is zero when $\theta_{\text{eq}} = \pi$:

$$\left.\frac{d^2 U_{\text{harmonic\_cosine}}[\theta]}{d\theta^2}\right|_{\theta=\theta_{\text{eq}}=\pi} = 0 \tag{150}$$

This means $U_{\text{harmonic\_cosine}}$ exhibits too weak restoring force for small displacements when $\theta_{\text{eq}} = \pi$.[1] As shown in Fig. 6, the potential energy curve for $U_{\text{harmonic\_cosine}}[\theta]$ is too flat under these conditions.

The practical consequence of the above problems is that flexible forcefields have often used different forms of angle-bending potentials for nearly linear angles (*i.e.*, $\theta_{\text{eq}} \approx \pi$) compared to significantly bent angles (*i.e.*, $\theta_{\text{eq}} \ll \pi$). For example, van der Spoel *et al.* introduced a special angle-bending potential applicable to only linear bond angles (*i.e.*, $\theta_{\text{eq}} = \pi$).[143] The DREIDING forcefield used the harmonic_cosine potential for bent angles (*i.e.*, $\theta_{\text{eq}} < \pi$) and the cosine_bend potential for linear bond angles (*i.e.*, $\theta_{\text{eq}} = \pi$).[142] As another example, version 2 of the QuickFF protocol used the harmonic_bend potential for bent angles (*i.e.*, $\theta_{\text{eq}} < \pi$) and the cosine_bend potential for linear bond angles (*i.e.*, $\theta_{\text{eq}} = \pi$).[23] These workarounds raise the additional question of how small $(\pi - \theta_{\text{eq}})$ should be to trigger the linear bond angle potential. For such potentials, θ_{eq} must be rounded to π to remove the force discontinuity at $\theta = \pi$. For example, if $\theta_{\text{eq}} = 179^\circ(\pi/180^\circ)$ triggers the linear bond angle potential and gets rounded up to $\tilde{\theta}_{\text{eq}} = \pi$, then this can introduce a small non-zero change in the optimized equilibrium geometry.

A new angle-bending potential is required to resolve these problems. Its form was derived as follows. Since

$$\cos[\theta] = \cos[2\pi - \theta] \tag{151}$$

it follows that eqn (145) is satisfied by choosing

$$U[\theta] = \text{func}[\cos[\theta]] \tag{152}$$

If func[s] is an infinitely differentiable function with respect to the independent variable s, then $U[\theta]$ will be an infinitely differentiable function with respect to the independent variable θ.

The second derivative of the harmonic_cosine potential (eqn (149)) is

$$\left.\frac{d^2 U_{\text{harmonic\_cosine}}}{d\theta^2}\right|_{\theta=\theta_{\text{eq}}} = k\sin^2\left[\theta_{\text{eq}}\right] \tag{153}$$

Therefore, the angle-bending potential's curvature at $\theta = \theta_{\text{eq}}$ will equal its force constant k if we divide the harmonic_cosine potential by $\sin^2[\theta_{\text{eq}}]$:

$$U_{\text{modification\_1}}[\theta] = 1/2k\frac{\left(\cos[\theta] - \cos\left[\theta_{\text{eq}}\right]\right)^2}{\sin^2\left[\theta_{\text{eq}}\right]} \tag{154}$$

Unfortunately, this modified potential becomes infinite (aka 'blows up') when $\theta_{\text{eq}} = \pi$, because $\theta \neq \theta_{\text{eq}}$ makes the numerator greater than zero while the denominator is zero when $\theta_{\text{eq}} = \pi$. Close examination reveals this issue can be resolved by choosing

$$U_{\text{modification\_2}}[\theta] = k\frac{2\left(\cos[\theta] - \cos\left[\theta_{\text{eq}}\right]\right)^2}{\sin^2[\theta] + 3\sin^2\left[\theta_{\text{eq}}\right]} \tag{155}$$

In the denominator, the 1 to 3 ratio of coefficients for $\sin^2[\theta]$ relative to $\sin^2[\theta_{\text{eq}}]$ is required to make the potential's curvature equal to the force constant k at the equilibrium angle even when $\theta_{\text{eq}} = \pi$:

$$\left.\frac{d^2 U_{\text{modification\_2}}}{d\theta^2}\right|_{\theta=\theta_{\text{eq}}} = k \tag{156}$$

Eqn (156) holds for any possible value of the equilibrium angle $0 < \theta_{\text{eq}} \leq \pi$. Eqn (155) has the deficiency that the restoring force approaches zero as θ approaches zero. Unfortunately, this means the modification_2 potential does not have sufficient repulsive force to prevent the bond angle from reaching $\theta = 0$.

This problem is resolved by including a factor that makes the potential's denominator approach zero as θ approaches zero:

$$U_{\text{new}}[\theta] = k\frac{2\left(\cos\theta - \cos\theta_{\text{eq}}\right)^2}{\sin^2\theta + 3\sin^2\theta_{\text{eq}}\left(\dfrac{\tanh[2\sin[\theta/2]]}{\tanh\left[2\sin\left[\theta_{\text{eq}}/2\right]\right]}\right)} \tag{157}$$

The denominator of eqn (157) includes the factor

$$h[\theta] = \frac{\tanh[\nu\sin[\theta/2]]}{\tanh\left[\nu\sin\left[\theta_{\text{eq}}/2\right]\right]} \tag{158}$$

with the tanh multiplier ν having the specific value $\nu = 2$. $h[\theta]$ has the following important limits:

$$h[\theta = 0] = 0 \tag{159}$$

$$h[\theta = \theta_{\text{eq}}] = 1 \tag{160}$$

$$h[\theta > \theta_{\text{eq}}] > 1 \tag{161}$$

$$h[\theta < \theta_{eq}] < 1 \tag{162}$$

$$h[\theta = \pi] = \tanh[\nu]/\tanh[\nu \sin[\theta_{eq}/2]] \tag{163}$$

Fig. 6 plots the new potential shown in eqn (157). Because this new potential approaches infinite value as θ approaches zero, it prevents a physical system with finite energy from reaching $\theta = 0$. This behavior models the Pauli repulsion[129] that prevents bond angles in a real physical system from reaching bond angles of $\theta = 0$. Since

$$\sin[\theta/2] = \sqrt{\frac{1-\cos[\theta]}{2}} \tag{164}$$

$U_{new}[\theta]$ can be rewritten as a function of $\cos[\theta]$:

$$U_{new}[\theta] = k\frac{2(\cos\theta - \cos\theta_{eq})^2}{1-\cos^2\theta + 3(1-\cos^2\theta_{eq})\left(\dfrac{\tanh\left[\sqrt{2(1-\cos\theta)}\right]}{\tanh\left[\sqrt{2(1-\cos\theta_{eq})}\right]}\right)} \tag{165}$$

Accordingly, this new angle-bending potential has continuous well-defined derivatives of all orders for all $0 < \theta \leq \pi$. ESI Section S4† gives the analytic first- and second-order derivatives of $U_{new}[\theta]$.

As shown in Fig. 6, $U_{new}[\theta]$ has the same function value, first derivative, and second derivative (curvature) as both $U_{harmonic\_bend}[\theta]$ and $U_{cosine\_bend}[\theta]$ at the energy minimum $\theta = \theta_{eq}$. This leads to the following resolution. In every classical forcefield that uses $U_{harmonic\_bend}[\theta]$ and/or $U_{cosine\_bend}[\theta]$, the angle-bending potential can be upgraded to $U_{new}[\theta]$ without requiring a change in the angle-bending force constant values. If a forcefield has been optimized to use $U_{new}[\theta]$ but a molecular dynamics or Monte Carlo simulation program has not yet been updated to include this potential, in the meantime is it feasible to substitute either $U_{harmonic\_bend}[\theta]$ or $U_{cosine\_bend}[\theta]$ without requiring a change in the angle-bending force constant values, but in the long-term it is preferable to update the simulation code to use the more robust and general $U_{new}[\theta]$ potential.

A potential argument against using $U_{new}[\theta]$ is that its form is more complicated which will increase the computational costs during classical molecular dynamics and Monte Carlo simulations. However, a closer analysis shows the increased computational cost is likely to be insignificant in practical use, because the number of bonded interactions is typically much smaller than the number of non-bonded interactions during such simulations. Specifically, an atom-in-material A only shares bond angles with its first and second bonded neighbors, while it shares non-bonded interactions with a potentially much larger number of atoms within the non-bonded interaction cutoff distance (if used) or with all atoms in the entire simulation unit cell and their periodic images (if a non-bonded interaction cutoff distance is not used) except those within $\{\text{excluded}_A\}$. Accordingly, the increased robustness and generality of this new angle-bending potential outweighs its relatively insignificant increased computational cost.

3.2 Computational results for real molecules

Table 4 summarizes the optimized geometries for ten triatomic molecules. CCSD calculations were performed in Gaussian 16 (ref. 95) using the def2-TZVPD[134] basis set (In this article, CCSD not CCSD(T) calculations were used.). For molecules containing no elements heavier than neon, all electrons were correlated in the coupled-cluster calculation. For molecules containing one or more elements heavier than neon, the FreezeNobleGasCore keyword was used, which applies the coupled-cluster correlation to the valence shell electrons only on all atoms. Geometries were optimized to the following convergence criteria: (1) the maximum force is less than 0.00045 hartrees per bohr; (2) the root-mean squared (RMS) force is less than 0.0003 hartrees per bohr; (3) the maximum displacement is less than 0.0018 bohr; and (4) the RMS displacement is less than 0.0012 bohr.

Fig. 7 compares the new angle-bending model potential to quantum-mechanically-computed angle-bending energy curves for these ten molecules. CCSD/def2-TZVPD energy curves were computed by varying the bond angle with and without relaxing the bond lengths. The settings for these calculations were similar to those described in the previous paragraph, except that some of the geometric parameters were constrained. As shown in Fig. 7, relaxing the bond lengths (blue curves) lowered the energy by only a small amount compared to keeping the bond lengths fixed (orange curves) as the constrained angle varied. The angle-bending force constant used in the model potential (black curves) is displayed on each graph. The particular value for the angle-bending force constant was chosen by visual inspection to achieve approximate agreement between the quantum-mechanically-computed and model potential energy curves.

Even though this model potential requires only a single parameter (*i.e.*, the force constant value) to be adjusted, it was generally in reasonable agreement with the quantum-mechanically-computed energy curves. Notably, this model potential reasonably matched the slope, height, and curvature of the quantum-mechanically-computed energy curve as the bond angle approached the limiting value $\theta = \pi$. The reasons for this are understood. Specifically, the model potential has continuous well-defined derivatives of all orders over the entire range $0 < \theta \leq \pi$, and the values of these derivatives change at reasonable rates. Eqn (145) imposes reflection symmetry about

Table 4 Optimized geometries for ten triatomic molecules

Molecule	Angle (°)	Bond length (Å)
CaH_2 (HCaH)	180.0	2.065
CO_2 (OCO)	180.0	1.157
HNO	108.4	1.056 (HN), 1.201(NO)
H_2O (HOH)	104.7	0.962
Li_2O (LiOLi)	180.0	1.619
NO_2 (ONO)	135.0	1.185
NS_2 (SNS)	154.1	1.543
SF_2 (FSF)	97.7	1.586
SiH_2 (HSiH) spin singlet	92.4	1.515
SO_2 (OSO)	119.4	1.426

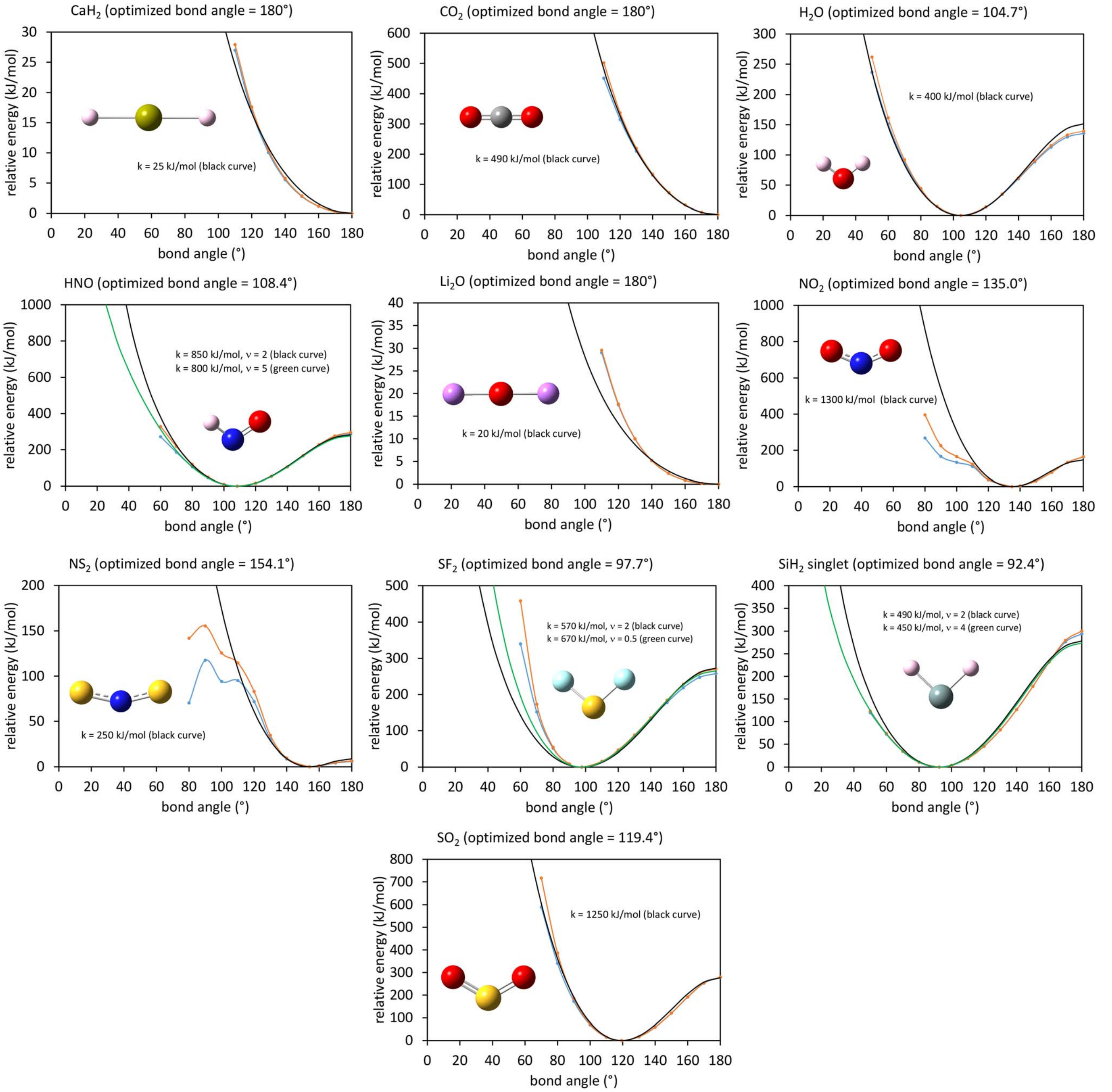

Fig. 7 Angle-bending energy curves for ten triatomic molecules. The orange curves show the quantum-mechanically-computed (CCSD/def2-TZVPD) values holding the bond lengths fixed as the angle varied, while the blue curves (CCSD/def2-TZVPD) relaxed the bond lengths. The black curves show the new angle-bending model potential with the displayed force constant value. In some of the panels, green curves show modified model potentials. (For purposes of reporting the force constant values, radians were treated as dimensionless units.)

$\theta = \pi$. The slope (*i.e.*, first derivative) varies from a value of zero at $\theta = \theta_{eq}$ to a value of approximately $k(\pi - \theta_{eq})/3$ at the midpoint $\theta = \frac{1}{2}(\pi + \theta_{eq})$ to a value of zero at $\theta = \pi$. These constraints on the slope approximately determine the third derivative's values over this range of bond angles. Together with the function's value of zero at $\theta = \theta_{eq}$ and the second-derivative's value of k at $\theta = \theta_{eq}$, these various conditions approximately determine the curve's shape over the range of angles between θ_{eq} and π.

For SF_2, the quantum-mechanically-computed energy curve rises more steeply than the model potential (eqn (157)) over the range $\theta < \theta_{eq}$. This can be partially but not fully resolved by modifying the model potential in eqn (157) such that the tanh multiplier used is smaller (*e.g.*, 0.5) instead of 2. The result is plotted as the green curve in the SF_2 panel of Fig. 7. For SF_2, further reducing the tanh multiplier value towards zero does not result in significant improvement compared to the $\nu = 0.5$ curve. For HNO, a slightly improved fit between the model potential and the quantum-mechanically-computed energy curve can be obtained by using a tanh multiplier value of 5 instead of 2 as shown by the green curve in the HNO panel of Fig. 7. For SiH_2, a slightly improved fit between the model potential and the quantum-mechanically-computed energy curve can be obtained by using a tanh multiplier value of 4 instead of 2 as shown by the green curve in the SiH_2 panel of Fig. 7. The tanh multiplier value of 2 (as shown in eqn (157)) was chosen as a compromise value that provides acceptably good results for most materials. Since $\sin[\theta_{eq}] = 0$ when $\theta_{eq} = 180°$, the value of the tanh multiplier ν has no impact on the model curves for CaH_2, CO_2, Li_2O, and other triatomic molecules having $\theta_{eq} = 180°$.

For NO_2, the quantum-mechanically-computed energy curves have a shoulder in the range of 90 to 110°. This appears to be due to some chemical hybridization changes within the molecule (aka 'chemical effects') that are not captured by the model potential. For NS_2, the quantum-mechanically-computed energy curves have a shoulder around 100° and a local maximum around 90°. This appears to be due to the formation of a S–S bond that lowers the energy as the bond angle is decreased to approximately 80°.

4. Flexibility parameters that approximately reproduce experimental vibrational frequencies

4.1 Homodiatomic molecules

Using the Manz stretch model potentials for H_2 and O_2 shown in Fig. 4, the following one-dimensional Schrodinger equation was solved for the vibrational eigenstates:

$$\left(U_{AB}^{Manz\_stretch}[d_{AB}] - \frac{\hbar^2}{2\mu_{AB}}\frac{d^2}{d{d_{AB}}^2}\right)\phi_\nu[d_{AB}] = \varepsilon_\nu\phi_\nu[d_{AB}] \quad (166)$$

The reduced mass is defined as

$$\mu_{AB} = \frac{m_A m_B}{m_A + m_B} \quad (167)$$

This Schrodinger equation corresponds to the situation in which the molecule is not rotating, its center-of-mass remains stationary, the Born–Oppenheimer approximation applies, relativistic effects are neglected, and the molecule is in the electronic ground state. The zero point energy (ZPE) corresponds to $\nu = 0$, while $\nu = 1$ is the first excited vibrational level.

I wrote a Matlab script to solve for the eigenvalues ε_ν and eigenfunctions $\phi_\nu[d_{AB}]$ for $\nu = 0, 1, 2, \ldots$. This script and its output results are provided in the ESI.† This script used a uniform grid for $(d_{AB}^{eq} - 1.5\text{ bohr}) \le d_{AB} \le (d_{AB}^{eq} + 5\text{ bohr})$ with a grid spacing of 0.001 bohr. The second-derivative in eqn (166) was computed using the central finite difference approximation. Computational tests with a finer grid spacing (0.0005 bohr) and the slightly larger range $(d_{AB}^{eq} - 2\text{ bohr}) \le d_{AB} \le (d_{AB}^{eq} + 6\text{ bohr})$ changed the $(\varepsilon_\nu - \varepsilon_{\nu-1})$ and ZPE results for H_2 and O_2 by less than 0.1 cm^{-1}.

As shown in Table 5, the computed ZPE and first dozen excited vibrational levels for the H_2 molecule differed by less than 10% from the experimental values. For different isotopes, the Born–Oppenheimer potential energy curve (and hence optimized force constant) remains the same, but the vibrational frequencies change owing to changes in reduced mass. Table 6 shows good agreement between the computed and experimental values for the first vibrational transition frequency of each hydrogen molecule isotope. As shown in Table 7, the model potential predicted the first 25 vibrational energy levels for the O_2 molecule within 3% of the experimental values. The relative deviations became larger closer to the bond dissociation energy (*e.g.*, 7% error for the $\nu_{29 \leftarrow 30}$ transition of O_2). Overall, these results show the Manz stretch model potential approximately reproduces experimental bond vibration frequencies.

4.2 Triatomic molecules

Why is it useful to compute both rigid and relaxed angle-bending scans as shown in Fig. 7? Comparing the rigid scan energy curve to the relaxed scan energy curve provides extremely valuable insights into the relative importance of some cross terms. Iff the relaxed scan curve is greatly below the rigid scan curve, then this indicates that changing bond lengths substantially lowers the energy at non-equilibrium angle values, and in this case bond-bend cross terms may be needed to construct an accurate forcefield. Iff the relaxed and rigid angle-scan curves are nearly identical, this suggests bond-bend cross terms are not required to construct an accurate forcefield model.

In this section, flexibility models were constructed for several triatomic molecules as examples, because these molecules do not require dihedral terms. Owing to the lengthy space required to thoroughly explain the dihedral terms, I decided that it would be easier for readers if the content related to dihedral model

Table 5 Comparison of vibrational frequencies (in wavenumber, cm^{-1}) calculated using the Manz stretch model potential to experimental values for the H_2 molecule. Each value is the energy of that vibrational level minus the energy of the prior vibrational level. ZPE = zero point energy

ν	Experiment[a]	Calculated	% error
ZPE	2179	2254	3%
1	4161	4312	4%
2	3926	4048	3%
3	3695	3782	2%
4	3468	3515	1%
5	3242	3247	0%
6	3014	2976	−1%
7	2782	2704	−3%
8	2543	2429	−4%
9	2293	2151	−6%
10	2026	1871	−8%
11	1737	1587	−9%
12	1415	1299	−8%

[a] Experimental data from ref. 144–146.

Table 6 Comparison of $\upsilon_{0 \leftarrow 1}$ transition frequency (in wavenumber, cm^{-1}) calculated using the Manz stretch model potential to experimental values for different isotopes of the hydrogen molecule: H_2, HD, D_2, HT, DT, and T_2

	Experiment	Calculated	% error
H_2	4161 (ref. 145 and 147)	4312	4
HD	3632 (ref. 145 and 147)	3765	4
D_2	2994 (ref. 145 and 147)	3104	4
HT	3435 (ref. 148)	3561	4
DT	2743 (ref. 148)	2845	4
T_2	2465 (ref. 148)	2556	4

potentials is presented in a subsequent companion article rather than incorporating it here.

As shown in Table 8, several flexibility models were parameterized and compared for the CO_2, water, HNO, and SO_2 molecules. For each molecule in this set, Fig. 7 shows that the relaxed angle-scan energy curve is approximately the same as the rigid angle-scan energy curve for the same molecule. Consequently, bond-bend cross terms were not required to build accurate flexibility models for these molecules. The constructed flexibility models contained bond-stretch and angle-bend terms. Flexibility models with and without Urey–Bradley or bond–bond cross terms were compared. My new angle-bending model potential (eqn (157)) was used for all of these flexibility models. Both bond and Urey–Bradley stretches were modeled using either the harmonic stretch (eqn (10)) or Manz stretch (eqn (95)) model potential. When present, the bond–bond cross term had the form:

$$U_{ABC}^{\text{bond–bond}} = k(d_{AB} - d_{AB}^{eq})(d_{BC} - d_{BC}^{eq}) \tag{168}$$

The force constants (*i.e.*, k values) were the only adjustable parameters in these flexibility models. Nonadjustable parameters included the equilibrium lengths and equilibrium bond angle, which were taken from the CCSD/def2-TZVPD optimized geometries. For the Manz stretch potential, the $\gamma°$ values computed using the method described in Section 2.6.2 were (in $bohr^{-1}$): (a) 1.203 (C–O) and 1.257 (O–O) for CO_2, (b) 1.276 (H–O) and 1.129 (H–H) for water, (c) 1.246 (H–N), 1.251 (N–O), and 1.212 (H–O) for HNO, and (d) 1.079 (S–O) and 1.151 (O–O) for SO_2. $\gamma°$ between the two outer atoms was only relevant when the flexibility model contained Urey–Bradley interaction.

The training and validation datasets contained quantum chemistry calculations at the CCSD/def2-TZVPD level of theory. For each molecule, the training dataset contained:

(1) The QM optimized geometry and energy.

(2) Both the relaxed angle-scan and rigid angle-scan geometries and energies for the subset of angles satisfying $(\theta_{ABC}^{eq} - 30°) \le \theta_{ABC} \le (\theta_{ABC}^{eq} + 30°)$. This subset of angles was chosen, because it focused the fit on angle values that are not extremely far away from θ_{ABC}^{eq}. The specific datapoints used were those plotted in Fig. 7 that satisfied the additional condition that they were within this angle range.

Table 7 Comparison of vibrational frequencies (in wavenumber, cm^{-1}) calculated using the Manz stretch model potential to experimental values for the O_2 molecule (^{16}O isotope). Each listed value is the energy of that vibrational level minus the energy of the prior vibrational level. ZPE = zero point energy

υ	Experiment[a]	Calculated	% error
ZPE	787	761	−3%
1	1556	1506	−3%
2	1533	1484	−3%
3	1510	1463	−3%
4	1486	1441	−3%
5	1463	1420	−3%
6	1440	1398	−3%
7	1419	1377	−3%
8	1395	1355	−3%
9	1372	1334	−3%
10	1350	1312	−3%
11	1329	1290	−3%
12	1304	1269	−3%
13	1280	1247	−3%
14	1258	1225	−3%
15	1236	1203	−3%
16	1212	1181	−3%
17	1188	1160	−2%
18	1166	1138	−2%
19	1141	1116	−2%
20	1117	1094	−2%
21	1092	1072	−2%
22	1067	1050	−2%
23	1040	1028	−1%
24	1013	1006	−1%
25	985	983	0%
26	956	961	1%
27	925	939	2%
28	891	917	3%
29	858	894	4%
30	818	872	7%

[a] Experimental data as compiled in ref. 149.

(3) QM-computed single-point energies for a set of geometries in which each bond length was changed by −0.14, −0.07, 0.00, +0.07, +0.14 Å relative to the fully-relaxed unconstrained geometry. For these structures, the bond angle was held rigid at θ_{ABC}^{eq}. For each structure, the single-point energy was computed without constrained geometry relaxation. For a symmetric triatomic (*e.g.*, CO_2, H_2O, SO_2), this yielded 14 distinct displaced geometries. For HNO, this yielded 24 distinct displaced geometries.

The GRG solver in Excel was used to solve this linear regression problem that minimizes the least-squares loss function shown in eqn (138) subject to the following force constant bounds. The angle-bending, bond stretch, and Urey–Bradley stretch (if present) force constants were constrained to be non-negative. No bounds were placed on the bond–bond cross (if present) force constant.

For each molecule, the validation dataset was constructed by using an uniform random number generator to generate random bond displacements in the interval −0.07 to +0.07 Å and random angle displacements in the interval −30 to +30°

Table 8 Flexibility parameters fitted for the CO_2, H_2O, HNO, and SO_2 molecules. UB = Urey–Bradley term. BBC = bond–bond cross term. Entries marked with "—" indicate that type of term was not considered for inclusion in the model. Entries marked with "0" mean that type of term was considered but converged to a value of zero. Please see the main text for a list of $\gamma°$ values. For each molecule, results for the recommended flexibility model are shown in boldface type

	Stretch type	$k_{stretch}$ (eV bohr^{-2})	k_{bend} (eV)	k_{UB} (eV bohr^{-2})	k_{BBC} (eV bohr^{-2})	*R*-Squared training	*R*-Squared validation
CO_2	Manz	30.58	5.17	—	—	0.9928	0.9940
CO_2	**Manz**	**27.26**	**5.03**	**2.31**	**—**	**0.9995**	**0.9998**
CO_2	Harmonic	31.58	5.17	—	—	0.9287	0.9911
CO_2	Harmonic	28.15	5.02	2.86	—	0.9341	0.9982
CO_2	Harmonic	31.01	5.17	—	2.86	0.9341	0.9929
H_2O	**Manz**	**14.95**	**4.26**	**—**	**—**	**0.9996**	**0.9974**
H_2O	Manz	14.87	4.11	0.10	—	0.9996	0.9978
H_2O	Harmonic	15.62	4.26	—	—	0.9456	0.9957
H_2O	Harmonic	15.62	4.26	0.00	—	0.9456	0.9957
H_2O	Harmonic	15.66	4.26	—	−0.16	0.9457	0.9957
HNO	**Manz**	**8.97 (HN), 22.34 (NO)**	**8.07**	**—**	**—**	**0.9902**	**0.9816**
HNO	Manz	8.55 (HN), 21.89 (NO)	6.23	0.61	—	0.9922	0.9751
HNO	Harmonic	9.03 (HN), 23.45 (NO)	7.99	—	—	0.9410	0.9787
HNO	Harmonic	7.47 (HN), 21.74 (NO)	4.07	2.50	—	0.9472	0.9907
HNO	Harmonic	9.03 (HN), 23.45 (NO)	8.00	—	1.71	0.9465	0.9814
SO_2	Manz	20.34	11.60	—	—	0.9970	0.9948
SO_2	**Manz**	**19.90**	**9.71**	**0.46**	**—**	**0.9986**	**0.9973**
SO_2	Harmonic	20.86	11.56	—	—	0.9648	0.9934
SO_2	Harmonic	19.88	9.42	1.11	—	0.9658	0.9943
SO_2	Harmonic	20.83	11.56	—	0.16	0.9648	0.9934

relative to the optimized geometry. In each validation geometry, three separate random numbers were used to independently displace each of the two bonds and the angle. For each molecule, nine validation geometries were prepared in this manner.

R-Squared values for the training and validation datasets were then computed using eqn (132)–(134). Table 8 lists the optimized force constant values, *R*-squared training, and *R*-squared validation.

For each flexibility model, normal vibrational mode analysis within the harmonic oscillator approximation was performed by diagonalizing the mass-weighted Hessian (MWH) matrix expressed in Cartesian coordinates:

$$\mathrm{MWH}_{(3(\mathrm{A}-1)+i),(3(\mathrm{B}-1)+j)} = \frac{1}{\sqrt{m_\mathrm{A} m_\mathrm{B}}} \frac{\partial^2 U}{\partial\left(\vec{R}_\mathrm{A}\right)_i \partial\left(\vec{R}_\mathrm{B}\right)_j} \quad (169)$$

where m_A is the mass of atom A. Here, $(\vec{R}_A)_i$ for $i \in \{1, 2, 3\}$ denotes the X, Y, or Z component of the nuclear position $\vec{R}_A$. The second derivatives can be computed either analytically or numerically; here, they were computed numerically using the central finite difference approximation. The eigenvalues $\{\lambda_i\}$ of the MWH matrix are related to the normal mode frequencies $\{\mathrm{freq}_i\}$ *via*:[101]

$$\mathrm{freq}_i = \sqrt{\lambda_i}\Big/(2\pi) \quad (170)$$

Each normal mode frequency was converted to wavenumber by dividing by the speed of light, *c*. Each eigenvector ($\vec{V}^{\mathrm{normal\_mode}}$) of the MWH matrix is the corresponding normal mode's mass-weighted differential displacement vector:

$$\varepsilon \vec{V}^{\mathrm{normal\_mode}} = \sum_\mathrm{A} \left(d\vec{R}_\mathrm{A}^{\mathrm{normal\_mode}} \sqrt{m_\mathrm{A}} \right) \quad (171)$$

for infinitesimal $|\varepsilon|$.

For linear molecules, five of the MWH eigenvalues are zero; these correspond to molecular rotation (2 modes) and center-of-mass translation (3 modes). For nonlinear molecules, six of the MWH eigenvalues are zero; these correspond to molecular rotation (3 modes) and center-of-mass translation (3 modes).

Table 9 lists the computed vibrational frequencies (in wavenumber, cm^{-1}) and their percent errors relative to experimental reference values. Examining Tables 8 and 9, flexibility models using the Manz stretch potential performed slightly better than those using the harmonic stretch potential. However, all of the parameterized flexibility models performed reasonably well. Including a Urey–Bradley term improved the results for CO_2 and SO_2 but had little effect for H_2O and HNO. Including a bond–bond cross term had little effect. In Tables 8 and 9, the 'recommended' flexibility model shown in boldface type achieves a good combination of high *R*-squared validation, high *R*-squared training, and accuracy for computed frequencies, while not introducing an excessive number of force constants.

For CO_2, the energy splitting between the asymmetric stretch and the symmetric stretch was not predominantly due to Urey–Bradley or bond–bond cross interactions as mistakenly suggested in the companion article.[45] That suggestion was based on the observation that without Urey–Bradley or bond–bond cross terms, the Hessian matrix is already diagonal (with equal eigenvalues for the two bond stretches) when expressed in terms of the internal coordinates (d_{AB}, d_{BC}, θ_{ABC}) as:[45]

Table 9 Computed vibrational frequencies (in wavenumber, cm^{-1}) using different forcefields for the CO_2, H_2O, HNO, and SO_2 molecules. For CO_2, the bend mode is 2-fold degenerate. The percent error relative to experimentally-measured value (ref. 150) is shown in parentheses. For each molecule, results for the recommended flexibility model are shown in boldface type

	Stretch type	UB?	BBC?	Bend	Stretch # 1[a]	Stretch # 2[a]
CO_2	Manz	No	No	694 (4%)	1363 (2%)	2609 (11%)
CO_2	**Manz**	**Yes**	**No**	**684 (3%)**	**1391 (4%)**	**2463 (5%)**
CO_2	Harmonic	No	No	694 (4%)	1385 (4%)	2651 (13%)
CO_2	Harmonic	Yes	No	684 (3%)	1434 (8%)	2503 (7%)
CO_2	Harmonic	No	Yes	694 (4%)	1434 (8%)	2503 (7%)
H_2O	**Manz**	**No**	**No**	**1634 (2%)**	**3885 (6%)**	**3942 (5%)**
H_2O	Manz	Yes	No	1629 (2%)	3889 (6%)	3932 (5%)
H_2O	Harmonic	No	No	1633 (2%)	3972 (9%)	4030 (7%)
H_2O	Harmonic	Yes	No	1633 (2%)	3972 (9%)	4030 (7%)
H_2O	Harmonic	No	Yes	1633 (2%)	3956 (8%)	4055 (8%)
HNO	**Manz**	**No**	**No**	**1451 (-3%)**	**3047 (14%)**	**1776 (13%)**
HNO	Manz	Yes	No	1407 (−6%)	3032 (13%)	1723 (10%)
HNO	Harmonic	No	No	1453 (−3%)	3058 (14%)	1807 (15%)
HNO	Harmonic	Yes	No	1434 (−4%)	3051 (14%)	1714 (10%)
HNO	Harmonic	No	Yes	1455 (−3%)	3051 (14%)	1798 (15%)
SO_2	Manz	No	No	550 (6%)	1259 (9%)	1468 (8%)
SO_2	**Manz**	**Yes**	**No**	**529 (2%)**	**1255 (9%)**	**1452 (7%)**
SO_2	Harmonic	No	No	549 (6%)	1275 (11%)	1487 (9%)
SO_2	Harmonic	Yes	No	553 (7%)	1272 (11%)	1452 (7%)
SO_2	Harmonic	No	Yes	549 (6%)	1279 (11%)	1480 (9%)

[a] For CO_2, H_2O and SO_2, stretch # 1 is the symmetric stretch, and stretch # 2 is the asymmetric stretch. For HNO, stretch # 1 is the H–N stretch, and stretch # 2 is the N–O stretch.

$$\text{Hessian} = \begin{pmatrix} \dfrac{\partial^2 U}{\partial {d_{AB}}^2} & 0 & 0 \\ 0 & \dfrac{\partial^2 U}{\partial {d_{BC}}^2} & 0 \\ 0 & 0 & \dfrac{\partial^2 U}{\partial {\theta_{ABC}}^2} \end{pmatrix} \tag{172}$$

However, the normal vibrational modes must be computed by diagonalizing the Hessian matrix defined by mass-weighted Cartesian coordinates, as shown in eqn (169).[101] For CO_2, this mass-weighted Hessian contains some non-zero off-diagonal elements even when the flexibility model contains no Urey–Bradley or bond–bond cross interactions, and this leads to a splitting between asymmetric and symmetric stretch frequencies even if no Urey–Bradley or bond–bond cross interactions are contained in the flexibility model.

5. Conclusions

In this article, I derived theoretical foundations of force field functional theory (FFFT). FFFT studies topics related to the functional representation of nonreactive forcefields to achieve various desirable properties such as:

(a) Formal exactness of the forcefield's energy functional under certain conditions.

(b) A formally exact ansatz separating the bonded potential energy from the nonbonded potential energy within a bonded cluster in a way that enables bonded parameters to be optimized using linear regression instead of requiring nonlinear regression.

(c) The potential energy's continuous differentiability to various orders with respect to energetically accessible internal coordinate displacements within a subdomain defined by one electronic ground state.

(d) Forcefield design that guarantees the reference ground-state geometry is exactly reproduced as an equilibrium structure on the forcefield's potential energy landscape.

(e) Reasonably accurate and broadly applicable frugal model potentials.

(f) Computationally efficient embedded feature selection that identifies and removes unimportant forcefield terms.

(g) Well-designed methods to parameterize the forcefield from quantum-mechanically-computed and (optionally) experimental reference data.

(h) Forcefields that approximately reproduce experimentally-measured properties.

Theoretical foundations of items (a), (b), and (d) were derived in Sections 2.1–2.5 above and demonstrated with examples in Sections 2.6, 4.1, and 4.2. Examples of (e) frugal model potentials include my new angle-bending and bond stretch model potentials that have (c) continuous differentiability to all orders. A companion article describes and applies several (f) embedded feature selection techniques including dihedral pruning, dihedral mode smart selection, and LASSO regression to identify and remove unimportant forcefield terms.[45] A companion article performs (g) quantum-mechanically-derived forcefield parameterization for more than a hundred MOFs that (h) exactly reproduces the experimental lattice constants.[45]

In general, a forcefield's potential energy is a functional of the material's chemical geometry and externally applied fields (if any) that exactly or approximately matches the quantum-mechanically-computed Born–Oppenheimer electronic energy as shown in eqn (14) or (88). This Born–Oppenheimer electronic energy surface is composed of subdomains such that chemical geometries within the same subdomain share the similar electronic ground state. Within each subdomain, $E^{0,\mathrm{QM}}_{\mathrm{electronic}}[\{\vec{R}_{\mathrm{A}},\mathbb{Z}_{\mathrm{A}}\}]$ has continuous first-order derivatives with respect to changes in the atomic coordinates $\{\vec{R}_{\mathrm{A}}\}$; however, its first (and/or higher-order) derivatives may be discontinuous at boundaries where two or more subdomains (*i.e.*, two or more different electronic ground states) intersect in energy. As shown in eqn (16), the system's total energy is obtained by adding the nuclear kinetic energy to the potential energy.

For convenience, $U^{\mathrm{FF}}_{\mathrm{total}}[\{\vec{R}_{\mathrm{A}},\mathbb{Z}_{\mathrm{A}}\}]$ is often partitioned into bonded and nonbonded interactions. In this article, I showed how to construct such a partition in a way that always guarantees the reference ground-state geometry of an isolated bonded cluster is exactly reproduced as a stationary point on the forcefield's potential energy landscape independently of the particular values to be assigned to the forcefield's force constants. At this optimized geometry, the new scheme's bonded interaction terms completely account for the isolated bonded cluster's geometry, potential energy, forces (first derivatives of potential energy), and Hessian (second derivatives of potential energy) with non-bonded interactions affecting only higher-order derivatives. This partitioning scheme is formally exact, because it does not introduce any new approximations into the forcefield model. In this partitioning scheme, the so-called 'resting values' contained in each flexibility term are precisely equal to the equilibrium values from the material's quantum-mechanically-computed ground-state geometry. Because these equilibrium values can be computed directly and do not need to be fitted during the forcefield parameterization, this transforms the task of optimizing the forcefield's bonded parameters from a nonlinear regression problem into a linear regression problem. Because linear regression problems are convex, this prevents separated regions in the optimization landscape from containing different local minima that can trap the optimizer. Moreover, multicollinearity issues can be more easily resolved (*e.g.*, by using the LASSO[32,33] method) in linear regression compared to nonlinear regression.

A key advantage of this new ansatz for separating intracluster nonbonded interactions from bonded interactions is that it reduces the sensitivity of optimized values for the bonded parameters on the particular choice of nonbonded interaction model. This allows the bonded interaction terms to be optimally parameterized to leading order without having to first choose specific values for the nonbonded interaction parameters. As an example, these important features were clearly demonstrated for the C–F bond stretch in the C_6F_6 molecule.

Section 2.5 above discusses the particular conditions that must be satisfied for $U^{\mathrm{FF}}_{\mathrm{total}}[\{\vec{R}_{\mathrm{A}},\mathbb{Z}_{\mathrm{A}}\}]$ to exactly equal $E^{0,\mathrm{QM}}_{\mathrm{electronic}}[\{\vec{R}_{\mathrm{A}},\mathbb{Z}_{\mathrm{A}}\}]$. Formal exactness requires that the forcefield was parameterized for the exact system being studied by the forcefield. The formally exact nonreactive forcefield requires a full series expansion of $U^{\mathrm{exact}}_{\mathrm{total}}[\{\vec{R}_{\mathrm{A}},\mathbb{Z}_{\mathrm{A}}\}]$ in terms of the material's internal coordinates; however, in most practical applications the explicit form of this exact expansion is unknown. In most practical applications, $U^{\mathrm{FF}}_{\mathrm{total}}[\{\vec{R}_{\mathrm{A}},\mathbb{Z}_{\mathrm{A}}\}]$ is represented by a weighted sum of model potentials to provide a pragmatic approximation of $E^{0,\mathrm{QM}}_{\mathrm{electronic}}[\{\vec{R}_{\mathrm{A}},\mathbb{Z}_{\mathrm{A}}\}]$. The coefficients in front of these model potentials are called force constants.

In practice, the formally exact series expansion is normally truncated by using model potentials having a finite number of interaction terms. This truncation introduces approximation. Careful choice of the model potentials can yield high computational efficiency, a relatively small number of required flexibility terms, continuous derivatives of all orders with respect to energetically accessible atom-in-material displacements, and excellent accuracy. This article introduced new angle-bending and bond-stretch model potentials that require only a small number of terms to achieve excellent accuracy, high computational efficiency, and continuous derivatives of all orders with respect to atom-in-material displacements.

The new angle-bending model potential was carefully derived to capture correct dynamics across a wide range of bond angles including the limiting value of $\theta = \pi$. In contrast, most previously used angle-bending model potentials have either a derivative discontinuity or incorrect dynamics when the bond angle reaches $\theta = \pi$. This new angle-bending model potential was compared to CCSD/def2-TZVPD quantum-mechanically-computed energy curves for ten triatomic molecules: CaH_2, CO_2, H_2O, HNO, Li_2O, NO_2, NS_2, SF_2, SiH_2, and SO_2. In all ten cases, the new angle-bending potential provided reasonably good results. However, some moderate discrepancies for $\theta < \theta_{\mathrm{eq}}$ were observed for NS_2 (due to chemical bonding changes), NO_2 (due to chemical hybridization changes), and SF_2 (due to steric repulsion between the two F atoms).

The new bond-stretch model potential was derived using first principles. This provides the key advantage that its exponent $\gamma^{\circ}_{\mathrm{AB}}$ is directly quantum-mechanically computed. This new bond-stretch model potential provides excellent accuracy for many bonds across a wide range of $\Delta d_{\mathrm{AB}} = d_{\mathrm{AB}} - d^{\mathrm{eq}}_{\mathrm{AB}}$ values, even as the bond length is stretched to infinity. Remarkably, this is accomplished with only one empirically-fitted parameter, which is the bond's force constant (k_{stretch}).

In this work, complete flexibility models (*i.e.*, bonded interaction models) were constructed for the H_2, O_2, CO_2, water, HNO, and SO_2 molecules. For each of these molecules, vibrational frequencies predicted by the parameterized flexibility model agreed closely with previously published experimentally-measured frequencies. For H_2 and O_2, these parameterized flexibility models agreed closely with the quantum-mechanically-computed bond-energy-*versus*-bond-distance curve. For CO_2, water, HNO, and SO_2, these parameterized flexibility models gave excellent *R*-squared values for approximately reproducing the quantum-mechanically-computed energies of independently chosen sets of validation geometries.

In a companion article, this new theory was used to optimize bonded parameters (aka flexibility parameters) for 116 MOFs.[45] As shown in that article, flexible forcefields constructed using FFFT and my new angle-bending and dihedral torsion model potentials gave excellent performance. Specifically, the model-predicted forces yielded goodness-of-fit (R-squared values) of 0.910 (avg across all MOFs) $\pm$ 0.018 (st. dev.) for atom-in-material forces across a quantum-mechanically-computed validation set of geometries generated using *ab initio* molecular dynamics in the NVE ensemble, where the parameterized forcefield model used dihedral pruning, individual equilibrium values, and no bond–bond cross terms.[45] This clearly demonstrates FFFT has enormous practical utility. That companion article introduces new best practices for: (a) typing bonds, angles, dihedrals, and other internal coordinates, (b) pruning dihedrals to reduce the redundancy of internal coordinates, (c) using the LASSO method in least-squares regression of the force constants to identify and eliminate unimportant forcefield terms, and (d) designing the forcefield to exactly reproduce experimental lattice constants defining the material's unit cell. That article introduces the well-designed SAVESTEPS protocol to parameterize the forcefield's bonded terms from quantum-mechanically-computed reference data.

Data availability

Optimized geometries of molecules, data analysis spreadsheets, Matlab codes and results, and outputs of the calculate_Manz_and_Morse_stretch_potential_exponents program are included as part of the ESI.† The calculate_Manz_and_Morse_stretch_potential_exponents program is available for download from **http://ddec.sourceforge.net**.

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgements

The author gratefully acknowledges financial support from NSF Career Award DMR-1555376. This work used the Expanse cluster at the San Diego Supercomputing Center (SDSC) through allocation CTS100027 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS[151]) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

1 D. Dubbeldam, K. S. Walton, T. J. H. Vlugt and S. Calero, Design, parameterization, and implementation of atomic force fields for adsorption in nanoporous materials, *Adv. Theory Simul.*, 2019, **2**, 1900135, DOI: **10.1002/adts.201900135**.
2 J. M. 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.*, 2004, **25**, 1157–1174, DOI: **10.1002/jcc.20035**.
3 P. Dauber-Osguthorpe and A. T. Hagler, Biomolecular force fields: where have we been, where are we now, where do we need to go and how do we get there?, *J. Comput.-Aided Mol. Des.*, 2019, **33**, 133–203, DOI: **10.1007/s10822-018-0111-4**.
4 J. A. Harrison, J. D. Schall, S. Maskey, P. T. Mikulski, M. T. Knippenberg and B. H. Morrow, Review of force fields and intermolecular potentials used in atomistic computational materials research, *Appl. Phys. Rev.*, 2018, **5**, 031104, DOI: **10.1063/1.5020808**.
5 J. R. Maple, M. J. Hwang, T. P. Stockfisch, U. Dinur, M. Waldman, C. S. Ewig and A. T. Hagler, Derivation of class II force fields .1. Methodology and quantum force field for the alkyl functional group and alkane molecules, *J. Comput. Chem.*, 1994, **15**, 162–182, DOI: **10.1002/jcc.540150207**.
6 H. C. Urey and C. A. Bradley, The vibrations of pentatonic tetrahedral molecules, *Phys. Rev.*, 1931, **38**, 1969–1978, DOI: **10.1103/PhysRev.38.1969**.
7 N. Foloppe and A. D. MacKerell, All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data, *J. Comput. Chem.*, 2000, **21**, 86–104.
8 M. J. Van Vleet, A. J. Misquitta, A. J. Stone and J. R. Schmidt, Beyond Born-Mayer: improved models for short-range repulsion in ab initio force fields, *J. Chem. Theory Comput.*, 2016, **12**, 3851–3870, DOI: **10.1021/acs.jctc.6b00209**.
9 T. A. Manz and N. Gabaldon Limas, Introducing DDEC6 atomic population analysis: part 1. Charge partitioning theory and methodology, *RSC Adv.*, 2016, **6**, 47771–47801, DOI: **10.1039/c6ra04656h**.
10 T. Chen and T. A. Manz, A collection of forcefield precursors for metal-organic frameworks, *RSC Adv.*, 2019, **9**, 36492–36507, DOI: **10.1039/c9ra07327b**.
11 T. A. Manz, T. Chen, D. J. Cole, N. G. Limas and B. Fiszbein, New scaling relations to compute atom-in-material polarizabilities and dispersion coefficients: part 1. Theory and accuracy, *RSC Adv.*, 2019, **9**, 19297–19324, DOI: **10.1039/c9ra03003d**.
12 S. Vandenbrande, M. Waroquier, V. Van Speybroeck and T. Verstraelen, The monomer electron density force field (MEDFF): a physically inspired model for noncovalent interactions, *J. Chem. Theory Comput.*, 2017, **13**, 161–179, DOI: **10.1021/acs.jctc.6b00969**.
13 M. M. Ghahremanpour, P. J. van Maaren, C. Caleman, G. R. Hutchison and D. van der Spoel, Polarizable Drude model with s-type Gaussian or Slater charge density for general molecular mechanics force fields, *J. Chem. Theory Comput.*, 2018, **14**, 5553–5566, DOI: **10.1021/acs.jctc.8b00430**.
14 J. W. Ponder, C. J. Wu, P. Y. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, M. Head-Gordon, G. N. I. Clark, M. E. Johnson and T. Head-Gordon, Current status of the AMOEBA polarizable force field, *J. Phys. Chem. B*, 2010, **114**, 2549–2564, DOI: **10.1021/jp910674d**.
15 J. G. McDaniel and J. R. Schmidt, First-principles many-body force fields from the gas phase to liquid:

a "universal" approach, *J. Phys. Chem. B*, 2014, **118**, 8042–8053, DOI: 10.1021/jp501128w.

16 B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York and M. Karplus, CHARMM: the biomolecular simulation program, *J. Comput. Chem.*, 2009, **30**, 1545–1614, DOI: 10.1002/jcc.21287.

17 V. Barone, I. Cacelli, N. De Mitri, D. Licari, S. Monti and G. Prampolini, JOYCE and ULYSSES: integrated and user-friendly tools for the parameterization of intramolecular force fields from quantum mechanical data, *Phys. Chem. Chem. Phys.*, 2013, **15**, 3736–3751, DOI: 10.1039/c3cp44179b.

18 J. E. Lennard-Jones, Cohesion, *Proc. Phys. Soc.*, 1931, **43**, 461–482, DOI: 10.1088/0959-5309/43/5/301.

19 Y. Shi, Z. Xia, J. J. Zhang, R. Best, C. J. Wu, J. W. Ponder and P. Y. Ren, Polarizable atomic multipole-based AMOEBA force field for proteins, *J. Chem. Theory Comput.*, 2013, **9**, 4046–4063, DOI: 10.1021/ct4003702.

20 J. A. Lemkul, J. Huang, B. Roux and A. D. MacKerell, An empirical polarizable force field based on the classical Drude oscillator model: development history and recent applications, *Chem. Rev.*, 2016, **116**, 4983–5013, DOI: 10.1021/acs.chemrev.5b00505.

21 J. P. Araujo and M. Y. Ballester, A comparative review of 50 analytical representation of potential energy interaction for diatomic systems: 100 years of history, *Int. J. Quant. Chem.*, 2021, **121**, e26808, DOI: 10.1002/qua.26808.

22 L. Vanduyfhuys, S. Vandenbrande, T. Verstraelen, R. Schmid, M. Waroquier and V. Van Speybroeck, QuickFF: a program for a quick and easy derivation of force fields for metal-organic frameworks from ab initio input, *J. Comput. Chem.*, 2015, **36**, 1015–1027, DOI: 10.1002/jcc.23877.

23 L. Vanduyfhuys, S. Vandenbrande, J. Wieme, M. Waroquier, T. Verstraelen and V. Van Speybroeck, Extension of the QuickFF force field protocol for an improved accuracy of structural, vibrational, mechanical and thermal properties of metal-organic frameworks, *J. Comput. Chem.*, 2018, **39**, 999–1011, DOI: 10.1002/jcc.25173.

24 R. X. Wang, M. Ozhgibesov and H. Hirao, Analytical Hessian fitting schemes for efficient determination of force-constant parameters in molecular mechanics, *J. Comput. Chem.*, 2018, **39**, 307–318, DOI: 10.1002/jcc.25100.

25 R. X. Wang, M. Ozhgibesov and H. Hirao, Partial Hessian fitting for determining force constant parameters in molecular mechanics, *J. Comput. Chem.*, 2016, **37**, 2349–2359, DOI: 10.1002/jcc.24457.

26 J. P. Durholt, G. Fraux, F. X. Coudert and R. Schmid, Ab initio derived force fields for zeolitic imidazolate frameworks: MOF-FF for ZIFs, *J. Chem. Theory Comput.*, 2019, **15**, 2420–2432, DOI: 10.1021/acs.jctc.8b01041.

27 A. Gabrieli, M. Sant, P. Demontis and G. B. Suffritti, Fast and efficient optimization of molecular dynamics force fields for microporous materials: bonded interactions via force matching, *Microporous Mesoporous Mater.*, 2014, **197**, 339–347, DOI: 10.1016/j.micromeso.2014.06.023.

28 N. L. Allinger, Y. H. Yuh and J. H. Lii, Molecular mechanics. The MM3 force field for hydrocarbons .1, *J. Am. Chem. Soc.*, 1989, **111**, 8551–8566, DOI: 10.1021/ja00205a001.

29 X. Su, X. Yan and C.-L. Tsai, Linear regression, *Wiley Interdiscip. Rev. Comput. Stat.*, 2012, **4**, 275–294, DOI: 10.1002/wics.1198.

30 H. Usefi, Clustering, multicollinearity, and singular vectors, *Comput. Stat. Data Anal.*, 2022, **173**, 107523, DOI: 10.1016/j.csda.2022.107523.

31 O. M. Baksalary and G. Trenkler, The Moore-Penrose inverse: a hundred years on a frontline of physics research, *Eur. Phys. J. H*, 2021, **46**, 9, DOI: 10.1140/epjh/s13129-021-00011-y.

32 R. Tibshirani, Regression shrinkage and selection via the LASSO, *J. R. Stat. Soc. Ser. B*, 1996, **58**, 267–288, DOI: 10.1111/j.2517-6161.1996.tb02080.x.

33 R. J. Tibshirani, The LASSO problem and uniqueness, *Electron. J. Stat.*, 2013, **7**, 1456–1490, DOI: 10.1214/13-EJS815.

34 R. Penrose, A generalized inverse for matrices, *Math. Proc. Cambridge Philos. Soc.*, 1955, **51**, 406–413, DOI: 10.1017/S0305004100030401.

35 A. Ben-Israel, The Moore of the Moore-Penrose inverse, *Electron. J. Linear Algebra*, 2002, **9**, 150–157.

36 A. A. E. Hoerl and R. W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, 1970, **12**, 55–67, DOI: 10.1080/00401706.1970.10488634.

37 D. S. Hochbaum, Complexity and algorithms for nonlinear optimization problems, *Ann. Oper. Res.*, 2007, **153**, 257–296, DOI: 10.1007/s10479-007-0172-6.

38 C. A. Floudas and C. E. Gounaris, A review of recent advances in global optimization, *J. Glob. Optim.*, 2009, **45**, 3–38, DOI: 10.1007/s10898-008-9332-8.

39 M. J. D. Powell, Convergence properties of algorithms for nonlinear optimization, *SIAM Rev.*, 1986, **28**, 487–500, DOI: 10.1137/1028154.

40 G. Venter and J. Sobieszczanski-Sobieski, Particle swarm optimization, *AIAA J.*, 2003, **41**, 1583–1589, DOI: 10.2514/2.2111.

41 S. Katoch, S. S. Chauhan and V. Kumar, A review on genetic algorithm: past, present, and future, *Multimed. Tools Appl.*, 2021, **80**, 8091–8126, DOI: 10.1007/s11042-020-10139-6.

42 A. Trokhymchuk and J. Alejandre, Computer simulations of liquid/vapor interface in Lennard-Jones fluids: some questions and answers, *J. Chem. Phys.*, 1999, **111**, 8510–8523, DOI: 10.1063/1.480192.

43 S. Toxvaerd and J. C. Dyre, Shifted forces in molecular dynamics, *J. Chem. Phys.*, 2011, **134**, 081102, DOI: 10.1063/1.3558787.

44 X. P. Wang, S. Ramírez-Hinestrosa, J. Dobnikar and D. Frenkel, The Lennard-Jones potential: when (not) to

use it, *Phys. Chem. Chem. Phys.*, 2020, **22**, 10624–10633, DOI: 10.1039/c9cp05445f.

45 R. Ghanavati, A. C. Escobosa and T. A. Manz, An automated protocol to construct flexibility parameters for classical forcefields: applications to metal-organic frameworks, *RSC Adv.*, 2024, **14**, 22714–22762, DOI: 10.1039/d4ra01859a.

46 M. Born and R. Oppenheimer, Quantum theory of molecules, *Ann. Phys.*, 1927, **84**, 457–484.

47 S. Grimme, A general quantum mechanically derived force field (QMDFF) for molecules and condensed phase simulations, *J. Chem. Theory Comput.*, 2014, **10**, 4497–4514, DOI: 10.1021/ct500573f.

48 P. M. Morse, Diatomic molecules according to the wave mechanics. II. Vibrational levels, *Phys. Rev.*, 1929, **34**, 57–64, DOI: 10.1103/PhysRev.34.57.

49 G. S. Fanourgakis, T. E. Markland and D. E. Manolopoulos, A fast path integral method for polarizable force fields, *J. Chem. Phys.*, 2009, **131**, 094102, DOI: 10.1063/1.3216520.

50 T. F. Miller and D. C. Clary, Torsional path integral Monte Carlo method for the quantum simulation of large molecules, *J. Chem. Phys.*, 2002, **116**, 8262–8269, DOI: 10.1063/1.1467342.

51 R. P. Feynman, Forces in molecules, *Phys. Rev.*, 1939, **56**, 340–343, DOI: 10.1103/PhysRev.56.340.

52 L. D. Sanjeewa, V. O. Garlea, M. A. McGuire, C. D. McMillen and J. W. Kolis, Magnetic ground state crossover in a series of glaserite systems with triangular magnetic lattices, *Inorg. Chem.*, 2019, **58**, 2813–2821, DOI: 10.1021/acs.inorgchem.8b03418.

53 F. Walz, The Verwey transition - a topical review, *J. Phys.: Condens.Matter*, 2002, **14**, R285–R340, DOI: 10.1088/0953-8984/14/12/203.

54 P. Haen and T. Fukuhara, Study of the crossover from ferromagnetic to antiferromagnetic ground state in $CeRu_2(Ge_{0.7}Si_{0.3})_2$ by resistivity measurements under pressure, *Phys. B*, 2002, **312**, 437–439.

55 T. Vogt, P. M. Woodward, P. Karen, B. A. Hunter, P. Henning and A. R. Moodenbaugh, Low to high spin-state transition induced by charge ordering in antiferromagnetic $YBaCo_2O_5$, *Phys. Rev. Lett.*, 2000, **84**, 2969–2972, DOI: 10.1103/PhysRevLett.84.2969.

56 B. M. Axilrod and E. Teller, Interaction of the van der Waals type between three atoms, *J. Chem. Phys.*, 1943, **11**, 299–300, DOI: 10.1063/1.1723844.

57 K. T. Tang and J. P. Toennies, An improved simple-model for the van der Waals potential based on universal damping functions for the dispersion coefficients, *J. Chem. Phys.*, 1984, **80**, 3726–3741, DOI: 10.1063/1.447150.

58 K. T. Tang and J. P. Toennies, The damping function of the van der Waals attraction in the potential between rare-gas atoms and metal surfaces, *Surf. Sci.*, 1992, **279**, L203–L206.

59 T. A. Manz and T. Chen, New scaling relations to compute atom-in-material polarizabilities and dispersion coefficients: part 2. Linear-scaling computational algorithms and parallelization, *RSC Adv.*, 2019, **9**, 33310–33336, DOI: 10.1039/c9ra01983a.

60 N. Gabaldon Limas and T. A. Manz, Introducing DDEC6 atomic population analysis: part 4. Efficient parallel computation of net atomic charges, atomic spin moments, bond orders, and more, *RSC Adv.*, 2018, **8**, 2678–2707, DOI: 10.1039/c7ra11829e.

61 N. Gabaldon Limas and T. A. Manz, Introducing DDEC6 atomic population analysis: part 2. Computed results for a wide range of periodic and nonperiodic materials, *RSC Adv.*, 2016, **6**, 45727–45747, DOI: 10.1039/c6ra05507a.

62 Q. Y. Yang, D. H. Liu, C. L. Zhong and J. R. Li, Development of computational methodologies for metal-organic frameworks and their application in gas separations, *Chem. Rev.*, 2013, **113**, 8261–8323, DOI: 10.1021/cr400005f.

63 J. A. Rackers and J. W. Ponder, Classical Pauli repulsion: an anisotropic, atomic multipole model, *J. Chem. Phys.*, 2019, **150**, 084104, DOI: 10.1063/1.5081060.

64 A. Gabrieli, M. Sant, P. Demontis and G. B. Suffritti, A combined energy-force fitting procedure to develop DFT-based force fields, *J. Phys. Chem. C*, 2016, **120**, 26309–26319, DOI: 10.1021/acs.jpcc.6b08163.

65 L. P. Wang, J. H. Chen and T. Van Voorhis, Systematic parametrization of polarizable force fields from quantum chemistry data, *J. Chem. Theory Comput.*, 2013, **9**, 452–460, DOI: 10.1021/ct300826t.

66 J. M. Wang, P. Cieplak, J. Li, T. J. Hou, R. Luo and Y. Duan, Development of polarizable models for molecular mechanical calculations I: parameterization of atomic polarizability, *J. Phys. Chem. B*, 2011, **115**, 3091–3099, DOI: 10.1021/jp112133g.

67 J. M. Wang, P. Cieplak, J. Li, J. Wang, Q. Cai, M. J. Hsieh, H. X. Lei, R. Luo and Y. Duan, Development of polarizable models for molecular mechanical calculations II: induced dipole models significantly improve accuracy of intermolecular interaction energies, *J. Phys. Chem. B*, 2011, **115**, 3100–3111, DOI: 10.1021/jp1121382.

68 O. Borodin, Polarizable force field development and molecular dynamics simulations of ionic liquids, *J. Phys. Chem. B*, 2009, **113**, 11463–11478, DOI: 10.1021/jp905220k.

69 T. Nakajima and K. Hirao, The Douglas-Kroll-Hess approach, *Chem. Rev.*, 2012, **112**, 385–402, DOI: 10.1021/cr200040s.

70 P. A. M. Dirac, The quantum theory of the electron, *Proc. R. Soc. London, Ser. A*, 1928, **117**, 610–624, DOI: 10.1098/rspa.1928.0023.

71 P. A. M. Dirac, The quantum theory of the electron - Part II, *Proc. R. Soc. London, Ser. A*, 1928, **118**, 351–361, DOI: 10.1098/rspa.1928.0056.

72 M. Dolg, Relativistic effective core potentials, in *Relativistic Electronic Structure Theory, Part 1: Fundamentals*, ed. P. Schwerdtfeger, Elsevier, Amsterdam, 2002, vol. 11, pp. 793–862.

73 Relativistic Effects and the Chemistry of Heavy Elements, in, *Comprehensive Computational Chemistry*, ed. K. Ruud, M. Yanez and R. J. Boyd, Elsevier, 2024, vol. 3, pp. 1–314.

74 P. Hohenberg and W. Kohn, Inhomogeneous electron gas, *Phys. Rev. B*, 1964, **136**, B864–B871, DOI: 10.1103/PhysRev.136.B864.

75 W. Kohn and L. J. Sham, Self-consistent equations including exchange and correlation effects, *Phys. Rev.*, 1965, **140**, 1133–1138, DOI: **10.1103/PhysRev.140.A1133**.
76 U. von Barth and L. Hedin, Local exchange-correlation potential for spin polarized case, *J. Phys. C: Solid State Phys.*, 1972, **5**, 1629–1642, DOI: **10.1088/0022-3719/5/13/012**.
77 J. M. Foster and S. F. Boys, Canonical configuration interaction procedure, *Rev. Mod. Phys.*, 1960, **32**, 300–302, DOI: **10.1103/RevModPhys.32.300**.
78 H. Nakatsuji and K. Hirao, Cluster expansion of wavefunction: symmetry-adapted-cluster expansion, its variational determination, and extension of open-shell orbital theory, *J. Chem. Phys.*, 1978, **68**, 2053–2065, DOI: **10.1063/1.436028**.
79 H. Nakatsuji, Cluster expansion of the wavefunction: calculation of electron correlations in ground and excited-states by SAC and SAC CI theories, *Chem. Phys. Lett.*, 1979, **67**, 334–342.
80 J. D. Watts, J. Gauss and R. J. Bartlett, Coupled-cluster methods with noniterative triple excitations for restricted open-shell hartree-fock and other general single determinant reference functions: energies and analytical gradients, *J. Chem. Phys.*, 1993, **98**, 8718–8733, DOI: **10.1063/1.464480**.
81 J. P. Perdew, K. Burke and M. Ernzerhof, Generalized gradient approximation made simple, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868, DOI: **10.1103/PhysRevLett.77.3865**.
82 A. D. Becke, Density-functional thermochemistry .3. The role of exact exchange, *J. Chem. Phys.*, 1993, **98**, 5648–5652, DOI: **10.1063/1.464913**.
83 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.*, 2008, **120**, 215–241, DOI: **10.1007/s00214-007-0310-x**.
84 L. G. Kong, F. A. Bischoff and E. F. Valeev, Explicitly correlated R12/F12 methods for electronic structure, *Chem. Rev.*, 2012, **112**, 75–107, DOI: **10.1021/cr200204r**.
85 A. J. Cohen, P. Mori-Sánchez and W. T. Yang, Challenges for Density Functional Theory, *Chem. Rev.*, 2012, **112**, 289–320, DOI: **10.1021/cr200107z**.
86 S. Grimme, A. Hansen, J. G. Brandenburg and C. Bannwarth, Dispersion-corrected mean-field electronic structure methods, *Chem. Rev.*, 2016, **116**, 5105–5154, DOI: **10.1021/acs.chemrev.5b00533**.
87 J. Hermann, R. DiStasio and A. Tkatchenko, First-principles models for van der Waals interactions in molecules and materials: concepts, theory, and applications, *Chem. Rev.*, 2017, **117**, 4714–4758, DOI: **10.1021/acs.chemrev.6b00446**.
88 A. D. Becke and E. R. Johnson, Exchange-hole dipole moment and the dispersion interaction revisited, *J. Chem. Phys.*, 2007, **127**, 154108, DOI: **10.1063/1.2795701**.
89 O. A. Vydrov and T. Van Voorhis, Nonlocal van der Waals density functional: the simpler the better, *J. Chem. Phys.*, 2010, **133**, 244103, DOI: **10.1063/1.3521275**.
90 L. Füsti-Molnar and P. Pulay, Accurate molecular integrals and energies using combined plane wave and Gaussian basis sets in molecular electronic structure theory, *J. Chem. Phys.*, 2002, **116**, 7795–7805, DOI: **10.1063/1.1467901**.
91 A. Ruiz-Serrano, N. D. M. Hine and C. K. Skylaris, Pulay forces from localized orbitals optimized in situ using a psinc basis set, *J. Chem. Phys.*, 2012, **136**, 234101, DOI: **10.1063/1.4728026**.
92 J. G. Hill and K. A. Peterson, Modern basis sets across the periodic table, in *Comprehensive Computational Chemistry*, ed. K. Raghavachari, M. Yanez, and R. J. Boyd, Elsevier, 2024, vol. 1, pp. 4–17, DOI: **10.1016/B978-0-12-821978-2.00127-6**.
93 P. E. Blochl, Projector augmented-wave method, *Phys. Rev. B*, 1994, **50**, 17953–17979, DOI: **10.1103/PhysRevB.50.17953**.
94 G. Kresse and D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, *Phys. Rev. B*, 1999, **59**, 1758–1775, DOI: **10.1103/PhysRevB.59.1758**.
95 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr, J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 16, Revision B.01*, Gaussian, Inc., Wallingford CT, 2016.
96 T. H. Dunning, Gaussian-basis sets for use in correlated molecular calculations. 1. The atoms boron through neon and hydrogen, *J. Chem. Phys.*, 1989, **90**, 1007–1023, DOI: **10.1063/1.456153**.
97 B. P. Pritchard, D. Altarawy, B. Didier, T. D. Gibson and T. L. Windus, New basis set exchange: an open, up-to-date resource for the molecular sciences community, *J. Chem. Inf. Model.*, 2019, **59**, 4814–4820, DOI: **10.1021/acs.jcim.9b00725**.
98 R. A. Kendall, T. H. Dunning and R. J. Harrison, Electron affinities of the 1st-row atoms revisited: systematic basis-sets and wave-functions, *J. Chem. Phys.*, 1992, **96**, 6796–6806, DOI: **10.1063/1.462569**.
99 A. Szabo and N. Ostlund, *Modern Quantum Chemistry*, Dover, Mineola, NY, 1996, pp. 1–466.
100 T. Helgaker, P. Jorgensen and J. Olsen, *Molecular Electronic-Structure Theory*, Wiley, New York, 2000, pp. 1–908.

101 P. Atkins and R. Friedman, *Molecular Quantum Mechanics*, Oxford University Press, Oxford, United Kingdom, 5th edn, 2011, pp. 1–537.
102 S. Wieser and E. Zojer, Machine learned force-fields for an ab-initio quality description of metal-organic frameworks, *npj Comput. Mater.*, 2024, **10**, 18, DOI: 10.1038/s41524-024-01205-w.
103 R. Jinnouchi, F. Karsai and G. Kresse, On-the-fly machine learning force field generation: application to melting points, *Phys. Rev. B*, 2019, **100**, 014105, DOI: 10.1103/PhysRevB.100.014105.
104 S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schutt and K. R. Muller, Machine learning of accurate energy-conserving molecular force fields, *Sci. Adv.*, 2017, **3**, e1603015, DOI: 10.1126/sciadv.1603015.
105 O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltaysky, K. T. Schütt, A. Tkatchenko and K. R. Müller, Machine learning force fields, *Chem. Rev.*, 2021, **121**, 10142–10186, DOI: 10.1021/acs.chemrev.0c01111.
106 K. Choudhary, B. DeCost and F. Tavazza, Machine learning with force-field-inspired descriptors for materials: fast screening and mapping energy landscape, *Phys. Rev. Mater.*, 2018, **2**, 083801, DOI: 10.1103/PhysRevMaterials.2.083801.
107 T. Han, J. Li, L. P. Liu, F. Y. Li and L. W. Wang, Accuracy evaluation of different machine learning force field features, *New J. Phys.*, 2023, **25**, 093007, DOI: 10.1088/1367-2630/acf2bb.
108 A. Mirzanejad and S. A. Varganov, Derivation of Morse potential, *Mol. Phys.*, 2024, **122**, e2360542, DOI: 10.1080/00268976.2024.2360542.
109 R. N. Costa Filho, G. Alencar, B. S. Skagerstam and J. S. Andrade, Morse potential derived from first principles, *Europhys. Lett.*, 2013, **101**, 10009, DOI: 10.1209/0295-5075/101/10009.
110 R. Ahlrichs, Long-range behavior of natural orbitals and electron density, *J. Chem. Phys.*, 1976, **64**, 2706–2707, DOI: 10.1063/1.432491.
111 M. Levy and R. G. Parr, Long-range behavior of natural orbitals and electron density – reply, *J. Chem. Phys.*, 1976, **64**, 2707–2708, DOI: 10.1063/1.432492.
112 M. M. Morrell, R. G. Parr and M. Levy, Calculation of ionization potentials from density matrices and natural functions, and long-range behavior of natural orbitals and electron density, *J. Chem. Phys.*, 1975, **62**, 549–554, DOI: 10.1063/1.430509.
113 P. Maxwell, N. di Pasquale, S. Cardamone and P. L. A. Popelier, The prediction of topologically partitioned intra-atomic and inter-atomic energies by the machine learning method kriging, *Theor. Chem. Acc.*, 2016, **135**, 195, DOI: 10.1007/s00214-016-1951-4.
114 P. L. A. Popelier, On Quantum Chemical Topology, in *Applications of Topological Methods in Molecular Chemistry*, ed. R. Chauvin, C. Lepetit, B. Silvi and E. Alikhani, Springer, New York, 2016, pp. 23–52.
115 P. L. A. Popelier, Quantum Chemical Topology, in *The Chemical Bond – 100 Years Old and Getting Stronger*, ed. D. M. P. Mingos, Springer, New York, 2016, pp. 71–117.
116 J. Andres, P. Gonzalez-Navarrete and V. S. Safont, Unraveling reaction mechanisms by means of Quantum Chemical Topology analysis, *Int. J. Quant. Chem.*, 2014, **114**, 1239–1252, DOI: 10.1002/qua.24665.
117 R. F. W. Bader, A quantum theory of molecular structure and its applications, *Chem. Rev.*, 1991, **91**, 893–928, DOI: 10.1021/cr00005a013.
118 R. L. Fulton, Sharing of electrons in molecules, *J. Phys. Chem.*, 1993, **97**, 7516–7529, DOI: 10.1021/j100131a021.
119 R. F. W. Bader, *Atoms in Molecules: A Quantum Theory*, Clarendon Press, New York, 1994.
120 S. Cardamone, T. J. Hughes and P. L. A. Popelier, Multipolar electrostatics, *Phys. Chem. Chem. Phys.*, 2014, **16**, 10367–10387, DOI: 10.1039/c3cp54829e.
121 T. A. Manz, Density Derived Electrostatic and Chemical Methods, in *Comprehensive Computational Chemistry*, ed. P. L. A. Popelier, M. Yanez, and R. J. Boyd, Elsevier, 2024, vol. 2, pp. 362–405, DOI: 10.1016/B978-0-12-821978-2.00072-6.
122 T. A. Manz, Seven confluence principles: a case study of standardized statistical analysis for 26 methods that assign net atomic charges in molecules, *RSC Adv.*, 2020, **10**, 44121–44148, DOI: 10.1039/d0ra06392d.
123 F. Heidar-Zadeh, P. W. Ayers, T. Verstraelen, I. Vinogradov, E. Vohringer-Martinez and P. Bultinck, Information-theoretic approaches to atoms-in-molecules: Hirshfeld family of partitioning schemes, *J. Phys. Chem. A*, 2018, **122**, 4219–4245, DOI: 10.1021/acs.jpca.7b08966.
124 T. A. Manz and D. S. Sholl, Improved atoms-in-molecule charge partitioning functional for simultaneously reproducing the electrostatic potential and chemical states in periodic and nonperiodic materials, *J. Chem. Theory Comput.*, 2012, **8**, 2844–2867, DOI: 10.1021/ct3002199.
125 T. Verstraelen, S. Vandenbrande, F. Heidar-Zadeh, L. Vanduyfhuys, V. Van Speybroeck, M. Waroquier and P. W. Ayers, Minimal basis iterative stockholder: atoms in molecules for force-field development, *J. Chem. Theory Comput.*, 2016, **12**, 3894–3912, DOI: 10.1021/acs.jctc.6b00456.
126 F. L. Hirshfeld, Bonded-atom fragments for describing molecular charge-densities, *Theor. Chim. Acta*, 1977, **44**, 129–138, DOI: 10.1007/BF00549096.
127 A. J. Misquitta, A. J. Stone and F. Fazeli, Distributed multipoles from a robust basis-space implementation of the iterated stockholder atoms procedure, *J. Chem. Theory Comput.*, 2014, **10**, 5405–5418, DOI: 10.1021/ct5008444.
128 T. A. Manz, Introducing DDEC6 atomic population analysis: part 3. Comprehensive method to compute bond orders, *RSC Adv.*, 2017, **7**, 45552–45581, DOI: 10.1039/c7ra07400j.
129 W. Pauli, On the connection of the arrangement of electron groups in atoms with the complex structure of spectra, *Z. Phys.*, 1925, **31**, 765–783, DOI: 10.1007/BF02980631.

130 R. Y. Rohling, I. C. Tranca, E. J. M. Hensen and E. A. Pidko, Correlations between density-based bond orders and orbital-based bond energies for chemical bonding analysis, *J. Phys. Chem. C*, 2019, **123**, 2843–2854, DOI: 10.1021/acs.jpcc.8b08934.
131 L. Pauling, Atomic radii and interatomic distances in metals, *J. Am. Chem. Soc.*, 1947, **69**, 542–553, DOI: 10.1021/ja01195a024.
132 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, Ab-initio calculation of vibrational absorption and circular-dichroism spectra using density-functional force-fields, *J. Phys. Chem.*, 1994, **98**, 11623–11627, DOI: 10.1021/j100096a001.
133 P. J. Stephens, F. J. Devlin, C. S. Ashvar, C. F. Chabalowski and M. J. Frisch, Theoretical calculation of the vibrational circular-dichroism spectra, *Faraday Discuss.*, 1994, **99**, 103–119, DOI: 10.1039/fd9949900103.
134 D. Rappoport and F. Furche, Property-optimized Gaussian basis sets for molecular response calculations, *J. Chem. Phys.*, 2010, **133**, 134105, DOI: 10.1063/1.3484283.
135 A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard and W. M. Skiff, UFF, a full periodic-table force-field for molecular mechanics and molecular-dynamics simulations, *J. Am. Chem. Soc.*, 1992, **114**, 10024–10035, DOI: 10.1021/ja00051a040.
136 A. F. Tillack and B. H. Robinson, Simple model for the benzene hexafluorobenzene interaction, *J. Phys. Chem. B*, 2017, **121**, 6184–6188, DOI: 10.1021/acs.jpcb.7b02259.
137 D. Dellis, I. Skarmoutsos and J. Samios, Molecular simulations of benzene and hexafluorobenzene using new optimized effective potential models: investigation of the liquid, vapor-liquid coexistence and supercritical fluid phases, *J. Mol. Liq.*, 2010, **153**, 25–30, DOI: 10.1016/j.molliq.2009.04.007.
138 W. L. Jorgensen, M. M. Ghahremanpour, A. Saar and J. Tirado-Rives, OPLS/2020 force field for unsaturated hydrocarbons, alcohols, and ethers, *J. Phys. Chem. B*, 2024, **128**, 250–262, DOI: 10.1021/acs.jpcb.3c06602.
139 L. Verlet, Computer experiments on classical fluids .1. Thermodynamical properties of Lennard-Jones molecules, *Phys. Rev.*, 1967, **159**, 98–103, DOI: 10.1103/PhysRev.159.98.
140 M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, Oxford, United Kingdom, 2nd edn, 2017, pp. 100–106.
141 J. H. Lii and N. L. Allinger, The MM3 force field for amides, polypeptides, and proteins, *J. Comput. Chem.*, 1991, **12**, 186–199, DOI: 10.1002/jcc.540120208.
142 S. L. Mayo, B. D. Olafson and W. A. Goddard, DREIDING - A generic force field for molecular simulations, *J. Phys. Chem.*, 1990, **94**, 8897–8909, DOI: 10.1021/j100389a010.
143 D. van der Spoel, H. Henschel, P. J. van Maaren, M. M. Ghahremanpour and L. T. Costa, A potential for molecular simulation of compounds with linear moieties, *J. Chem. Phys.*, 2020, **153**, 084503, DOI: 10.1063/5.0015184.
144 L. Wolniewicz, Vibrational-rotational study of electronic ground state of hydrogen molecule, *J. Chem. Phys.*, 1966, **45**, 515–523, DOI: 10.1063/1.1727599.
145 G. Herzberg and L. L. Howe, The Lyman bands of molecular hydrogen, *Can. J. Phys.*, 1959, **37**, 636–659, DOI: 10.1139/p59-070.
146 G. Herzberg and A. Monfils, The dissociation energies of the H_2, HD, and D_2 molecules, *J. Mol. Spectrosc.*, 1960, **5**, 482–498, DOI: 10.1016/0022-2852(61)90111-4.
147 G. D. Dickenson, M. L. Niu, E. J. Salumbides, J. Komasa, K. S. E. Eikema, K. Pachucki and W. Ubachs, Fundamental vibration of molecular hydrogen, *Phys. Rev. Lett.*, 2013, **110**, 193601, DOI: 10.1103/PhysRevLett.110.193601.
148 K. F. Lai, V. Hermann, T. M. Trivikram, M. Diouf, M. Schlösser, W. Ubachs and E. J. Salumbides, Precision measurement of the fundamental vibrational frequencies of tritium-bearing hydrogen molecules: T_2, DT, HT, *Phys. Chem. Chem. Phys.*, 2020, **22**, 8973–8987, DOI: 10.1039/d0cp00596g.
149 L. Bytautas, N. Matsunaga and K. Ruedenberg, Accurate ab initio potential energy curve of O_2. II. Core-valence correlations, relativistic contributions, and vibration-rotation spectrum, *J. Chem. Phys.*, 2010, **132**, 074307, DOI: 10.1063/1.3298376.
150 *CRC Handbook of Chemistry and Physics*, ed. W. M. Haynes, CRC Press, Boca Raton, FL, 2016–2017, p. 105, Section 9.
151 T. J. Boerner, S. Deems, T. R. Furlani, S. L. Knuth and J. Towns, ACCESS: Advancing Innovation: NSF's Advanced Cyberinfrastructure Coordination Ecosystem: Services and Support, in *Proceedings of the Practice and Experience in Advanced Research Computing (PEARC '23)*, Portland, Oregon, 2023, pp. 1–4.