JCTC Journal of Chemical Theory and Computation

ARTICLE

pubs.acs.org/JCTC

Automated Repulsive Parametrization for the DFTB Method

Zoltán Bodrog,* Bálint Aradi, and Thomas Frauenheim

Bremen Center for Computational Materials Science, University of Bremen, Am Fallturm 1, 28359 Bremen, Germany

ABSTRACT: The density-functional-based tight-binding method is an efficient scheme for quantum mechanical atomistic simulations. While the most relevant part of the chemical energies is calculated within a DFT-like scheme, a fitted correction function—the repulsive energy—is used to achieve results as close to *ab initio* counterparts as possible. We have developed an automatic parametrization scheme to ease the process of the repulsive energy fitting, offering a more systematic and much faster alternative to the traditional fitting process. The quality of the resulting repulsives can be tuned by selecting and weighting the fit systems and the important physical properties (energy, force, Hessian) of them. Besides driving DFT calculators in the fitting process automatically, the flexibility of our scheme also allows the usage of external data (e.g., molecular dynamics trajectories or experimental data) as a reference. Results with several elements show that our procedure is able to produce parameter sets comparable to handmade ones, yet requiring far less human effort and time.

1. INTRODUCTION

The density-functional-based tight-binding method (DFTB)[1] is an efficient quantum mechanical simulation method, which is an approximation to the density functional theory (DFT). While being typically orders of magnitude faster than its *ab initio* counterpart, it delivers results for many chemical problems with reasonable accuracy. The rigorously derived original DFTB method provides an excellent theoretical framework which can be systematically extended when higher accuracy is needed or some new chemical features should be described which could not be covered by previous schemes. The success of the framework can be judged by the huge amount of different systematic extensions which had been created over time (e.g., charge self-consistency for describing charge transfer,[2] inclusion of collinear and noncollinear spin,[3] time-dependent[4] and GW[5] formalism to calculate excited state properties, Green-function technique to describe electron transport,[6] etc.).

Common in all different DFTB extensions is the fact, that only the "most important" part of the total energy is calculated within an approximate quantum mechanical approach, while the rest, comprising the core−core repulsion and the double-counting terms, is taken into account as a fitted quasi-classical interaction energy (the so-called repulsive energy) between participating atoms, depending on the configuration of the atomic nuclei in the system. When carefully done, the fitted repulsive interaction can even compensate for parts of the error introduced with the approximations in the quantum-mechanically calculated parts of the DFTB energy. This division and the approximations in the quantum-mechanically calculated part allows calculations on chemical systems typically several orders of magnitude faster and using considerably less memory than *ab initio* calculations, while still maintaining a reasonable accuracy.

Due to its presence in all extensions and its effect on the accuracy of the total energies, the fitting of the repulsive interactions is a cardinal problem for the original DFTB scheme and all its extensions. However, the parametrization process for a broad range of chemical species is a rather tedious work, often taking months of valuable research time. Additionally, due to its pairwise nature, the work necessary to extend an existing set with a new element increases with the set size, as the interaction of all elements with the new one has to be created.

To lower the barrier of extending DFTB to new chemical systems, several attempts have been made. First, Knaup et al. demonstrated their evolutionary algorithm work at fitting specific repulsives for the proton transfer in imidazole.[7] Quite similar to our work and in time parallel to our early steps, an automated parametrization engine had been created by Gaus et al.[8] that is able to fit repulsive energies in molecules. Its applicability has been demonstrated by fitting repulsives for carbohydrogen interactions, giving an accuracy comparable to the mio parameter set.[9] Unfortunately, the fitting framework seems to have several limitations. First, it only considered molecular systems, making parametrizations for solids and surfaces rather difficult. Furthermore, it does not seem to have included any means of mass fit data production, i.e., does not seem to be capable of generating large energetic data sets with reasonably little human intervention. This is not an issue when the fit is done against experimental data, but it maintains a part of the "parametrization barrier" by making the usage of DFT calculator references and scanning off-equilibrium reference data much more difficult than necessary. Last, it is built around a fixed spline representation of the repulsives, which gives a limited flexibility.

In this paper, we describe a comprehensive automated fitting process for creating repulsive pair potentials. It was developed by trying to simplify the fitting procedure as far as possible while still keeping its applicability to almost any kind of chemical situation where one expects reasonable description by DFTB. We designed our algorithm to deal with a large variety of repulsive function shapes and to fit to energetic properties of not only molecules but also crystalline systems. Our algorithm contains interactive parts for not only the core fitting process but also defining and building fit and test systems and making large series of batch fits with various changing metaparameters (parameters

Received: May 13, 2011
Published: July 05, 2011

affecting the parametrization, e.g., preliminarily chosen cutoffs or polynomial degrees). It was designed to model the workflow and needs of an applied scientist creating new parameters, up to the point that it can be integrated into graphical chemistry and materials science program packages as a GUI-driven module. The applicability of the algorithm has been demonstrated by creating repulsive interactions for carbohydrogens, titanium-organic chemistry, and crystalline zinc-oxide compounds. As it will be shown later, the obtained sets are of comparable quality to the current best handmade ones for those systems, while significantly reducing the human effort involved in their creation.

The paper is structured as follows: first, we give a short overview of the DFTB method and the parametrization problem. Then, we introduce our parametrizer automaton in detail. This is followed by a comparison of the automatically created sets mentioned above against their existing handmade counterparts.

2. METHODOLOGICAL BACKGROUND

2.1. The DFTB Method. The original, noniterative, perturbative DFTB[1] is a tight-binding DFT method used to calculate electronic structures of chemical systems. It solves the Kohn–Sham equations

$$\hat{H}_{KS}|\mu\rangle = [\hat{T} + \hat{V}_{KS}]|\mu\rangle = \varepsilon_\mu|\mu\rangle \tag{1}$$

which in an LCAO basis of Slater-type orbitals lead to

$$\sum_\chi H_{\phi\chi}c_{\mu,\chi} = \varepsilon_\mu \sum_\omega S_{\phi\omega}c_{\mu,\omega} \tag{2}$$

where

$$|\mu\rangle = \sum_\phi c_{\mu,\phi}|\phi\rangle, S_{\phi\omega} = \langle\phi|\omega\rangle,$$
$$H_{\phi\chi} = \langle\phi|\hat{T} + \hat{V}_{KS}|\chi\rangle \tag{3}$$

with the effective Kohn–Sham potential

$$V_{KS}(\mathbf{r}) = \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,d\mathbf{r}' + \frac{\delta E_{xc}[\rho(\mathbf{r})]}{\delta\rho(\mathbf{r})} + V_{ext}(\mathbf{r}) \tag{4}$$

and V_{ext} being the potential of the atomic cores. The effective potential is decomposed into atom-centered contributions, and the integrals in the Hamiltonian matrix are calculated by taking only two-center terms into account. Since the two-center contributions can be tabulated *in advance*, the DFTB algorithm is extremely fast, but lacking any direct energy contributions involving more than two atoms (e.g., $\langle\phi|\hat{V}_{KS}|\chi\rangle$-like terms having ϕ, V_{KS}, and χ from three different atoms).

Using compressed atomic orbitals, one can already get good results for many systems with the non-self-consistent application of the above scheme. In this perturbative approximation, the Kohn–Sham or electronic energy of the system, up to the pairwise approximation of the Hamiltonian, is

$$E_{el} = \sum_\mu^{occ}\sum_{\phi,\chi}\bar{c}_{\mu,\phi}H^{(0)}_{\phi\chi}c_{\mu,\chi} = T + \int\left(\int\frac{\rho^{(0)}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,d\mathbf{r}' + \left.\frac{\delta E_{xc}}{\delta\rho(\mathbf{r})}\right|_{\rho^{(0)}} + V_{ext}(\mathbf{r})\right)\rho(\mathbf{r})\,d\mathbf{r} \tag{5}$$

with the first sum running over all occupied states (T is the total electronic kinetic energy). The total energy is

$$E = E_{el} - \frac{1}{2}\int\int\frac{\rho^{(0)}(\mathbf{r})\rho^{(0)}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,d\mathbf{r}\,d\mathbf{r}' - \int\left.\frac{\delta E_{xc}}{\delta\rho(\mathbf{r})}\right|_{\rho^{(0)}}\rho^{(0)}(\mathbf{r})\,d\mathbf{r} + E^{(0)}_{xc} + E^{(0)}_{core} \tag{6}$$

Here, E_{core} is the interaction energy of the atomic cores and the superscript (0) denotes quantities calculated from the nonperturbed superposition of starting atomic charge densities. The perturbative nature of the scheme ensures that the parts of the Hamiltonian and therefore the double-counting terms do not depend on the perturbatively calculated ρ charge density, but only on $\rho^{(0)}$.

As the ρ dependence and all sophisticated electronic properties are included in the E_{el} electronic part, the rest of the DFTB total energy can be treated as an effective potential between atomic nuclei, the so-called repulsive energy. Due to its corrective nature, we handle it in a simplified approach and break it down to a sum of pairwise potentials between the atoms:[1]

$$E_{rep} = E - E_{el} \approx \frac{1}{2}\sum_{i,j}^{nuclei} U_{type(ij)}(r_{ij}) \tag{7}$$

where $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ is the distance between atoms i and j.

A systematic enhancement to the original DFTB scheme is the self-consistent-charges (SCC) extension,[9] which makes the Hamiltonian depend on electronic density via a construction representing the charge fluctuations between atoms with point-like charges. This correction with respect to the total energy of the non-SCC DFTB is contained in the SCC total energy expression as

$$\Delta E = \frac{1}{2}\sum_{i,j}\gamma_{i,j}\Delta q_i\Delta q_j \tag{8}$$

where the γ's are the effective interaction profiles of spherically symmetric diffuse charges, q_i is the Mulliken charge of atom i, and Δq_i is its change with respect to the neutral atomic population. These γ's give back the Coulombic $1/r$ profile in large distances as well as the atomic chemical hardness at $r \to 0$. The above energy correction is realized in the Kohn–Sham Hamiltonian by

$$\Delta H_{\phi\chi} = \frac{1}{2}S_{\phi\chi}\sum_j(\gamma_{[\phi],j} + \gamma_{[\chi],j})\Delta q_j \tag{9}$$

where $[\phi]$ and $[\chi]$ represent the atomic centers of orbitals ϕ and χ, respectively.

Having a Hamiltonian depending on molecular charge distribution makes self-consistent iterative calculations possible. SCC-DFTB needs a reparametriztaion with respect to the non-SCC one, however, as E_{el} and thus the difference between DFT total energy and E_{el} changes with the addition of self-consistency.

2.2. Parametrization with Pair Potentials. According to eq 7, the repulsive energy is broken down to pairwise potentials:

$$E_{rep}(\{\mathbf{r}_{atoms}\}) = \sum_{i<j}U_{AB}(r_{ij}) \tag{10}$$

where i and j both run over the atoms in the system and AB indicates the type of atom pair ij.

The parametrization process optimizes these $U_{AB}(r)$ pair potentials to cover the difference between the reference energies of certain fit systems and the corresponding electronic DFTB energy. The reference energies may be taken from experimental data or *ab initio* calculations. We prefer the latter, as it allows versatile reference data generation. The best parametrization can be viewed as the one where the set of pair potentials minimizes the error:

$$R = \sum (E_{\text{ref}} - E_{\text{DFTB}})^2 = \sum \Big(\sum_{i<j} U_{AB}(r_{ij}) - (E_{\text{ref}} - E_{\text{el}}) \Big)^2 = \min \tag{11}$$

Due to the approximative nature of DFTB, parametrizations lack universal transferability, but as the cases of successful parametrizations show, the validity of a good parameter set can extend to a wide range of problems.

2.3. Hand-Made Repulsive Potentials. In its usual course, parametrization for a bond type (e.g., the carbon−carbon bond) begins with stretching one bond of that kind in an appropriate molecule, as the simplest case, and creating a $U_{\text{CC}}(r)$ curve based on the energy difference between the DFT reference and DFTB:

$$U_{\text{CC}}(r) = E_{\text{DFT}}(r) - E_{\text{el}}(r) + \text{const} \tag{12}$$

with r being the length of the stretched bond. The constant term covers the limit of $E_{\text{DFT}}(r) - E_{\text{el}}(r)$ at $r \to \infty$ (this limit contains, e.g., the repulsive contributions from nonvarying bonds, that may not be known in detail at all) in order to ensure a zero limit for $U_{\text{CC}}(r)$, $r \to \infty$. Of course, one chooses stretched molecules so that stretching affects only one bond (or maybe several bonds, but in a totally equivalent way); all of the other pairs of atoms with changing distances remain outside the ranges of their respective repulsives.

One can construct a reasonable curve for a given interaction by merging curve sections created for different molecules which represent different chemical bonds between the considered elements. For example, a carbon−carbon pair potential can be constructed by taking the sections near 1.2 Å, 1.34 Å, and 1.54 Å from ethyne, ethene, and ethane, respectively, in order to take single, double, and triple carbon bonds into account. The resulting compound curves can then be heuristically improved by comparing DFTB results on some test systems to DFT data and fine-tuning them by hand. Unfortunately, the fine-tuning involves a tremendous amount of human work, making the fast extension of a given set or creating a new set from scratch rather difficult.

3. AUTOMATIC PARAMETRIZATION SCHEME

In order to reduce the work involved in creating repulsive potentials, we propose an automatic algorithm based on least-squares fitting of repulsive potentials to reference energy values. During our early automatic fitting attempts, we experimented also with genetic algorithms, but the simpler least-squares fits turned out to be easier to handle and far less resource-hungry while delivering results of the same or even better quality. The process to be described below is not limited to the bare fitting of the repulsive potentials $U_{AB}(r)$, but it also helps in selecting and producing fit systems and fit data, tuning the priorities of different systems or properties, etc., making the whole parametrization process largely automatic.

3.1. Least-Squares Fitting of Repulsive Potentials. In order to make a least-squares fitting for the pairwise repulsive potentials possible, we express them in terms of some arbitrary basis functions as

$$U_{AB}(r_{ij}) = \sum_{\nu} \alpha_{AB,\nu} f_{AB,\nu}(r_{ij}) \tag{13}$$

where AB is the type of atomic pair ij. Substituting this into the pair potential structure of E_{rep} from eq 10, the total repulsive energy for a given system becomes a linear combination

$$E_{\text{rep}} = \sum_{AB,\nu} \alpha_{AB,\nu} X_{AB,\nu} \tag{14}$$

of the structure-dependent quantities

$$X_{AB,\nu} = \sum_{\substack{\text{type}(ij)=AB \\ i<j}} f_{AB,\nu}(r_{ij}) \tag{15}$$

The sum runs over all possible atom pairs where the pair ij belongs to pair type AB.

Using the above, the best $\alpha_{AB,\nu}$ coefficients may be easily approximated by a least-squares fit to energy values of several different distortions of a chemical system as a function of the changing X values. Due to the linearity of the energy as a function of $X_{AB,\nu}$'s, this fitting is a multidimensional linear regression.

Running over a sequence of distortions denoted by s of the same system (we will call this sequence a *fit path* and the distortions *fit steps*), the least-squares fit minimizes the overall error

$$R = \sum_{s}^{\text{all steps}} (E_{\text{rep}}^{(s)} - (E_{\text{ref}}^{(s)} - E_{\text{el}}^{(s)}))^2 \tag{16}$$

With expression 14 of the total repulsive energy, the stationary condition

$$\frac{\partial}{\partial \alpha_{AB,\nu}} R(\alpha_{AB,\nu}) = 0 \tag{17}$$

of the above error leads to a matrix expression of the coefficients $\alpha_{\text{AB},\nu}$:

$$\mathbf{A} = (\mathbf{X}\mathbf{X}^{\text{T}})^{-1}\mathbf{X}\mathbf{E} \tag{18}$$

The matrices $\mathbf{E}$, $\mathbf{X}$, and $\mathbf{A}$ are constructed from the above energies, X structural constants, and α's in the following way:

$$\mathbf{E} = \begin{pmatrix} E_{\text{ref}}^{(1)} - E_{\text{el}}^{(1)} \\ E_{\text{ref}}^{(2)} - E_{\text{el}}^{(2)} \\ \vdots \end{pmatrix} \tag{19a}$$

$$\mathbf{X} = \begin{pmatrix} X_{\text{HH},1}^{(1)} & X_{\text{HH},1}^{(2)} & \cdots \\ X_{\text{HH},2}^{(1)} & X_{\text{HH},2}^{(2)} & \cdots \\ \vdots & & \\ X_{\text{CH},1}^{(1)} & X_{\text{CH},1}^{(2)} & \cdots \\ X_{\text{CH},2}^{(1)} & \cdots & \cdots \\ \vdots & & \end{pmatrix} \tag{19b}$$

$$\mathbf{A} = \begin{pmatrix} \alpha_{\mathrm{HH},1} \\ \alpha_{\mathrm{HH},2} \\ \vdots \\ \alpha_{\mathrm{CH},1} \\ \alpha_{\mathrm{CH},2} \\ \vdots \end{pmatrix} \tag{19c}$$

where, as an example, we assumed that the enumeration of the investigated atomic pairs begins with HH and contains CH.

As an example, a fit path could be built from a propane molecule with its middle carbon atom being shifted by 40 small random displacements around its equilibrium position. Each movement as well as the original configuration is a different fit step. The energy and structure data of these 41 steps would then give enough input to fit $U_{\mathrm{CC}}(r)$ and $U_{\mathrm{CH}}(r)$[10] provided the number of independent fitting parameters $\alpha_{\mathrm{AB},\nu}$ is well below 40, i.e., in this specific case, the number of basis functions used to describe one pairwise repulsive is well below 20. This criterion is normally fulfilled, but if not, increasing the amount of steps is always a straightforward remedy.

3.2. Fitting to Multiple Fit System Types and Objectives. An important expectation toward repulsive potentials is their transferability to a broad range of different systems. Usually, this requires compromises; transferability can be reached via a tradeoff between individual systems. Our automatic parametrization scheme enables the optimization of this tradeoff by enabling the fit on multiple test systems (multiple fit paths) at the same time. Staying with the example of the C−C and C−H repulsive fitting, by taking several different carbohydrogen molecules and distorting them, one can generate several molecular fit paths for the fit. Additionally, taking bulk diamond (with various deformations) as an additional fit path, one can tune the transferability toward the description of crystalline systems as well.

The goal of the fit becomes the minimization of overall error along all fit paths, modifying eq 16 to

$$R = \sum_{p}^{\text{all paths}} \sum_{s \in p} \left(E_{\mathrm{rep}}^{(ps)} - \left(E_{\mathrm{ref}}^{(ps)} - E_{\mathrm{el}}^{(ps)}\right)\right)^2 \tag{20}$$

with p enumerating the paths and E_{rep} written as a function of α's and X's within each path in the same way as the one-path case. It should be noted here that the number of $\alpha_{AB,\nu}$ parameters *does not depend* on the number of fit paths (nor on the number of steps in the individual fit paths). Its value is determined only by the choice of the basis functions to describe the repulsive interactions in question.

The $E_{\mathrm{rep}}^{(ps)}$ column vector of the repulsive energies for the multiple-path fitting is created by putting the $\mathbf{E}^{(p)}$ vectors for each path on top of each other:

$$\mathbf{E} = \begin{pmatrix} \mathbf{E}^{(1)} \\ \mathbf{E}^{(2)} \\ \vdots \end{pmatrix} = \begin{pmatrix} E_{\mathrm{ref}}^{(11)} - E_{\mathrm{el}}^{(11)} \\ E_{\mathrm{ref}}^{(12)} - E_{\mathrm{el}}^{(12)} \\ \vdots \\ E_{\mathrm{ref}}^{(21)} - E_{\mathrm{el}}^{(21)} \\ E_{\mathrm{ref}}^{(22)} - E_{\mathrm{el}}^{(22)} \\ \vdots \end{pmatrix} \tag{21a}$$

while the geometry matrix is created by putting single-path geometry matrices together in a similar way:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}^{(1)} \\ \hline \mathbf{X}^{(2)} \\ \vdots \end{pmatrix} \tag{21b}$$

For a multisystem fit, this all gives back the matrix eq 18 on **A**.

The scheme proposed here is not restricted to obtaining repulsive potentials by fitting on energy differences between *ab initio* DFT and DFTB calculations. One can naturally extend it to interatomic forces or even Hessians as targets. This way, one gains the possibility of not just choosing the transferability range by selecting various systems for the fitting procedure but also of being able to select the properties which are required to be transferable to the maximum possible amount over those systems. Furthermore, by using energy differences between successive steps as a target instead of the absolute energies, fitting on molecular dynamics (MD) trajectories is also made efficient. Details for these three target extensions (force, Hessian, and energy difference) are given in the Appendix.

3.3. Weighting of Fit Targets. In the formalism described until now, every fit step contributes to the R overall error with the same weight. As this may not always be the desired behavior, we allow each step in each path to have an individual weight for its contribution to the total error. If the fit is done for multiple physical properties (e.g., energies and forces), each property can also be weighted differently.

The weighting issues come to play mainly in two areas. First, one typically would overweight near-equilibrium geometries to ensure a higher precision at near-equilibrium bond lengths at the cost of less precise description for strongly distorted geometries. Furthermore, weighting becomes a key issue when multiple physical properties are invoked into the fit, since the numerical values of the differences in the various properties (energy, force, etc.) must be converted to the same scale. This requires some experimenting, but it offers the possibility of balancing the performance of repulsives for various physical quantities. For example, heavy weights for forces are usually necessary when the fitted repulsives give poor results with geometry optimization otherwise.

3.4. Basis Function Shapes. We have experimented with several different $f_\nu(r)$ basis functions for the repulsive fitting. The splines used in most of the current DFTB implementations turned out to be inappropriate for a fitting procedure, as they tend to give very oscillatory behavior. As a straightforward alternative, we decided for the eq 22 cut-off polynomials first. They were used in the earliest DFTB-implementation[11] and still retain popularity with doing parametrization by hand, since it has been possible to do most of the parametrizations up to now with them. The zeroth and first-degree terms are omitted from such a polynomial to ensure a smooth decay at its cutoff distance r_0:

$$f_\nu(r) = \begin{cases} (r - r_0)^\nu & \text{if } r < r_0 \text{ and } \nu \geq 2 \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

This representation was shown to be successful at the relatively easy hydrocarbon parametrization.

To emulate spline-like behavior in our scheme, we also tested bases containing the above cut-off polynomials, but having no universal cutoff value (these bases can be regarded as sums of multiple single-cutoff bases). Bases with two cutoff values are

very efficient at improving polynomial repulsives, while more cutoff values bring up the oscillatory nature of splines. Less successful but still noteworthy examples of spline-like bases are wavelet bases, which we also probed.

Another important basis was the family of exponential functions. $e^{-a_\nu r^\nu}$ ($\nu = 1, 2, 3, ...$) and their linear combinations, which seem to be a very natural choice for a repulsive function basis. These exponential functions proved to be a successful basis for our fittings with Ti and Zn. In these cases, a fairly tiny set of exponential basis functions (one to three of them) was quite enough to fit remarkably good parameter sets.

3.5. Further Automation in the Parametrization Workflow. Besides the automatized fitting process itself, there are three subprocesses of the parametrization workflow in which our program substantially lowers the human contribution.

- The path-building methods mentioned so far and some others are implemented to be executed automatically. They include bond stretchings, displacing atoms, uniform volume changes, linear interpolations between two configurations, and using predefined paths (e.g., MD trajectories or reaction paths).
- Instead of using fixed sets of metaparameters (input parameters determining the parametrization itself) for the fitting process, batch fits can use intervals of them. Scanning over all of these values in all of these intervals in every combination spares a lot of try-and-fail cycles for the user. At the end of the batch run, the set with the lowest total error on the targets (as defined in eq 20) is picked as the fittest solution.
- A module for defining test systems is built into our program too. It tests the energetical and geometrical performance of the fitted repulsives on the specified test systems. This way it can give a first-glance feedback about the performance of the fitted repulsive set on systems that were not necessarily fit systems.

4. RESULTS

4.1. Computational Tools. Since the molecular reference *ab initio* calculations in the handmade sets were mostly done using the Gaussian[12] code, we also used it as a reference for molecular systems. For the periodic systems, however, we have found the Siesta[13] code far more stable (less prone to convergence failures) in our automatic fitting environment, where distorted systems far from the equilibrium must be calculated very often. Apart from stability issues, this choice is also a good cross-calculator and cross-methodology (e.g., between different xc functionals in DFT references) consistency check of our algorithm and in general for the DFTB parametrization philosophy. As will be seen from the results, this mixing of DFT references did not pose any problem. The DFTB calculations were carried out using the DFTB+ package.[14]

4.2. Carbohydrogen Systems. The carbohydrogen case is a relatively easy case of parametrization in the sense that quite useful parameter sets can be fitted to it even with a small effort. Fitting to DFT references with the PBE exchange-correlation functional, the resulting parameter sets produce, according to our experiences, geometrical errors typically within a few 10^{-2} Å and atomization energy errors in the range of a few 10^{-2} au. This quality, which is almost comparable with the handmade mio set,[9] is pretty easy to reach at an automatic fit with a nontrivial handful of fit systems and a couple of hours working with them.

Table 1. Molecular and Crystalline Data Calculated with the Three Parameter Sets (the mio Set[9] and the Two Automatically Fitted Ones) Compared to Reference Values[a]

property	reference	mio	hom	inhom
		methane		
ΔE	0	7.3	−2.5	−0.1 (52.9)
C−H	1.093	1.089	1.094	1.080
		ethane		
ΔE	0	17.7	−1.0	0.1 (94.8)
C−C	1.531	1.501	1.535	1.516
C−H	1.096	1.098	1.102	1.088
		ethene		
ΔE	0	14.6	−2.4	−3.6 (68.6)
C═C	1.331	1.327	1.327	1.326
C−H	1.087	1.094	1.099	1.084
		ethyne		
ΔE	0	21.7	10.2	−3.5 (53.1)
C≡C	1.205	1.203	1.200	1.204
C−H	1.067	1.075	1.080	1.066
		benzene		
ΔE	0	52.9	−1.7	0.8 (170.8)
C−C	1.397	1.396	1.405	1.397
C−H	1.087	1.098	1.104	1.090
		butane		
ΔE	0	38.7	2.8	1.6 (179.8)
(1,2) C−C	1.547	1.519	1.555	1.537
(2,3) C−C	1.536	1.518	1.552	1.534
(1) C−H	1.097	1.097	1.102	1.088
		isobutane		
ΔE	0	38.0	1.9	1.5 (178.7)
C−C	1.535	1.518	1.552	1.534
C−H	1.097	1.098	1.102	1.088
		diamond		
C−C	1.555	1.540	1.575	1.558
		cyclobutane		
ΔE	0	40.0	7.3	13.4 (169.2)
C−C	1.557	1.539	1.569	1.534
C−H	1.095	1.102	1.107	1.094
		isobutene		
ΔE	0	36.2	0.5	−2.7 (153.0)
C−C	1.509	1.493	1.524	1.505
C═C	1.337	1.341	1.34	1.339
C−H (in CH_3)	1.099	1.100	1.104	1.090
C−H (in CH_2)	1.087	1.093	1.099	1.084
		bicyclobutane		
ΔE	0	26.9	−2.1	10.3 (143.4)
C−C (edge)	1.510	1.464	1.549	1.486
C−C (middle)	1.900	2.003	2.112	1.980
C−H (in CH_2)	1.112	1.195	1.161	1.158
C−H (in CH)	1.095	1.066	1.021	1.065
		cyclobutene		
ΔE	0	29.5	−1.8	7.4 (140.6)
C−C	1.573	1.569	1.597	1.538

Table 1. Continued

property	reference	mio	hom	inhom
C═C	1.519	1.524	1.548	1.493
C−H (in CH_2)	1.097	1.104	1.109	1.097
C−H (in CH)	1.087	1.097	1.103	1.089
	cyclopropane			
Δ*E*	0	18.9	−9.6	−0.2 (113.8)
C−C	1.509	1.489	1.523	1.502
C−H	1.087	1.096	1.100	1.087
	propane			
Δ*E*	0	27.7	0.2	0.0 (136.6)
C−C	1.532	1.509	1.544	1.525
C−H (end)	1.097	1.098	1.102	1.088
C−H (middle)	1.099	1.107	1.110	1.097
	cyclopropene			
Δ*E*	0	12.2	−13.8	−7.7 (83.7)
C−C	1.508	1.495	1.528	1.508
C═C	1.295	1.319	1.319	1.318
C−H (opposite to C═C)	1.095	1.107	1.109	1.096
C−H (neighbor to C═C)	1.080	1.090	1.095	1.081
	spiropentane			
Δ*E*	0	29.4	−18.9	−2.0 (173)
C−C ("radial")	1.485	1.479	1.508	1.488
C−C (outer)	1.530	1.508	1.547	1.524
C−H	1.088	1.097	1.102	1.088
	methylene-cyclopropane			
Δ*E*	0	25.9	−10.6	−4.5 (128.7)
C═C	1.322	1.328	1.327	1.327
C−C ("radial")	1.470	1.465	1.491	1.472
C−C (outer)	1.540	1.512	1.551	1.529
C−H (in CH_2)	1.088	1.095	1.101	1.086
C−H (on ring)	1.089	1.098	1.102	1.089
	propadiene			
Δ*E*	0	21.2	−2.6	−7.8 (83.6)
C═C	1.307	1.312	1.312	1.312
C−H	1.088	1.096	1.102	1.087
	1,3-butadiene			
Δ*E*	0	49.9	16.2	10.0 (143.2)
C−C	1.439	1.436	1.457	1.441
C═C	1.392	1.372	1.373	1.370
C−H (middle)	1.089	1.098	1.103	1.089
C−H (end)	1.086	1.104	1.085	1.095
	2-butyne			
Δ*E*	0	38.8	6.5	−1.2 (132.0)
C−C	1.462	1.455	1.477	1.461
C≡C	1.209	1.209	1.205	1.209
C−H	1.097	1.100	1.105	1.091
	propyne			
Δ*E*	0	30.3	8.3	1.2 (92.6)
C−C	1.460	1.453	1.475	1.459
C≡C	1.207	1.206	1.203	1.207
C−H (in CH_3)	1.097	1.100	1.104	1.090
C−H (in CH)	1.066	1.074	1.079	1.066

Table 1. Continued

property	reference	mio	hom	inhom
	propene			
Δ*E*	0	24.9	−1.5	−3.7 (110.3)
C−C	1.502	1.485	1.517	1.497
C═C	1.333	1.334	1.334	1.333
C−H (in CH_3)	1.098	1.100	1.105	1.091
C−H (in CH)	1.091	1.102	1.106	1.092
C−H (in CH_2)	1.087	1.093	1.098	1.084

[a] Δ*E* means atomization energy error relative to the reference in kcal/mol, and *A*−*B* atom pairs denote distances of the appropriate neighboring atoms in Å. The column "hom" contains a fit without dissociation energy correction, "inhom" contains a fit with it. Values in parentheses indicate errors for the set with dissociation energy correction when used in a DFTB implementation without this correction scheme. Italicized names denote systems that were fit systems too; the other molecules are the rest of the carbohydrogen part of the G2[18] test set.

Adding more configurations, the results can be further improved. In order to demonstrate the automatism in our procedure, we give the instructions used to generate those configurations:

- a methane molecule with its central carbon atom randomly displaced on five shells within a sphere[15] of diameter 0.75 Å
- an ethane molecule with one carbon atom displaced on 10 equidistant shells within a 0.75 Å sphere
- a butane molecule with its 1-2 carbon−carbon bond stretched in 15 0.1 Å steps, from a shortening of 0.6 Å to a lengthening of 0.9 Å
- a benzene ring with one of its carbon atoms displaced on five equidistant shells within a 0.75 Å diameter sphere
- an ethene molecule with one carbon atom displaced on five shells in a 0.75 Å diameter sphere
- a series of random displacements similar to the above with an ethyne molecule
- a hydrogen molecule with its only bond shortened in eight and lengthened in 12 0.025 Å steps
- an isobutane molecule with its central carbon atom displaced in a 1 Å diameter sphere

As the mio set, the basis of comparison, was fitted to calculations with the B3LYP xc functional and the 6-31G* basis, we also used this as a reference. The force objective had a weight of three while energy had a weight of one, and each path had its near-equilibrium steps (at most three steps away from equilibrium) overweighted by five. For the diamond test system, we used the CRYSTAL2003 code[16] (because of the problems with Gaussian mentioned above) with a 6-21G*[17] basis set and a *k*-space mesh of an 8 × 8 × 8 Monkhorst−Pack scheme.

During the fitting process, the automaton was allowed to sweep over the following metaparameters to search for the best fit:

- The cutoff of C−C: 2.0−2.3 Å
- The cutoff of C−H: 1.3−2.1 Å
- The cutoff of H−H: 1.3−2.1 Å
- The highest degree of polynomials: 10−12

The best fit was achieved with values of 2.3 Å, 2.1 Å, 1.3 Å, and 11 for the above metaparameters, respectively. For the sake of smoothness, the polynomials contained a minimal power of 4. Table 1 shows the performance of the resulting repulsive in comparison with the mio set (columns "mio" and "hom") on the respective equilibrium structures. As our method aims at not only

Table 2. Reference Data and Its Comparison with Previous Handmade Parametrization[19] ("znorg") and the Automatically Created One ("auto") for Zn and ZnO Crystals[a]

property	reference	znorg	auto
	Zn hcp		
ΔE (per Zn_2)	0	115.5	94.1
Zn−Zn (#1)	2.523	2.796	2.433
Zn−Zn (#2)	2.886	2.864	2.931
Zn−Zn (#3)	3.831	4.051	3.788
Zn−Zn (#4)	4.591	4.872	4.524
	ZnO zincblende		
ΔE (per ZnO)	0	22.3	−1.1
Zn−O	2.005	2.015	2.011
Zn−Zn	3.274	3.290	3.281
	ZnO wurtzite		
ΔE (per Zn_2O_2)	0	46.5	−0.7
Zn−O	2.017	2.015	2.018
Zn−O′	2.037	2.014	2.004

[a] The values given here refer to the equilibrium structure of each system. ΔE denotes the atomization energy difference with respect to the reference in kcal/mol. Atom pairs denote distances in Å.

describing equilibrium properties as close to *ab initio* results as possible, but also to provide a reasonable accuracy when dealing with structures out of equilibrium, we calculated also the energy errors over all nonequilibrium configurations in the fit paths. They remained generally within the error of 10^{-2} Hartree compared to the DFT reference except some of the extremely distorted geometries.

4.2.1. Using One-Body Repulsive Terms. With this carbohydrogen fit, we also experimented with using one-body terms in the repulsive energy

$$E_{\mathrm{rep}}(\{R_{\mathrm{nuclei}}\}) = \sum_{i<j} U_{AB}(r_{ij}) + \sum_{i} U_A \qquad (23)$$

One-body terms are a special case of inhomogeneous or dissociation energy terms: they represent a fixed, geometry-independent energy part as a sum of atomwise parts that do not come from the linear combinations of pairwise basis functions and that maintain the asymptotic value of E_{rep} at the dissociation limit. One-body energies are the only mathematically correct means of putting any correction to dissociation energy because only a sum of one-atomic dissociation energy terms behaves like an extensive quantity, i.e., is an additive function of stoichiometry. This fact strongly encourages investigating their use.

As the results in Table 1 illustrate (column "inhom"), one-body terms can slightly improve geometry results via eliminating the need of trying to set absolute atomization energy levels using the pair potential profiles. The resulting one-body terms were U_C = 0.030633H and U_H = 0.017967H for C and H atoms, respectively. The optimal cutoff distances were (determined by a similar batch) equal to those of the homogeneous case.

To maintain compatibility with the current DFTB implementations lacking one-body repulsive parts, we also took another way of improving results by one-body terms into account. Using them only at the fitting process but dropping them after it retains improved geometries and reaction energies calculated with the produced set yet leaves the pair potential structure of the repulsive energy built in DFTB intact (deteriorated bare atomization energy values are shown in parentheses in the appropriate column of Table 1).

Table 3. Titanium−Oxygen Compound Reference Data and Its Comparison with Previous Handmade Parametrization[20] ("tiorg") and the Automatically Created One ("auto")[a]

property	reference	tiorg	auto
	TiO		
ΔE	0	55.0	40.2
Ti−O	1.586	1.592	1.586
	Ti_2O_2 planar		
ΔE	0	87.7	69.4
Ti−Ti	2.198	2.355	2.092
Ti−O	1.857	1.891	1.866
	Ti_2O_2 nonplanar		
ΔE	0	68.0	57.6
Ti−Ti	2.127	2.249	2.133
Ti−O	1.838	1.888	1.826
	Ti_2O_4 #1 (dibridged with end O atoms in cis position)		
ΔE	0	49.4	81.3
Ti−Ti	2.716	2.800	2.635
bridging Ti−O	1.848	1.887	1.812
end Ti−O	1.622	1.606	1.589
O−Ti−Ti (ending O)	126.1	124.7	123.7
	Ti_2O_4 #2 (dibridged with end O atoms in trans position)		
ΔE	0	145.7	82.8
Ti−Ti	2.709	2.726	2.709
bridging Ti−O	1.840	1.831	1.806
end Ti−O	1.625	1.608	1.590
O−Ti−Ti (ending O)	123.7	122.3	122.2
	Ti_2O_4 #3 (tribridged with an O atom at one end)		
ΔE	0	123.3	66.4
Ti−Ti	2.394	2.540	2.399
bridging Ti−O (opposite to end O)	1.763	1.801	1.742
end Ti−O	1.628	1.606	1.586
	Ti hcp		
ΔE (per Ti)	0	−40.6	5.4
Ti−Ti	2.900	2.993	2.915
	TiO_2 anatase		
ΔE (per TiO_2)	0	22.5	−4.5
shortest Ti−Ti	3.028	2.996	3.082
shortest Ti−O	1.933	1.921	1.957
		1.995	1.958
	TiO_2 rutile		
ΔE (per TiO_2)	0	19.4	1.1
shortest Ti−Ti	3.559	3.613	3.605
shortest Ti−O	1.974	1.914	1.976
		1.992	1.995

[a] The values given here refer to the equilibrium structure of the various systems. ΔE means atomization energy difference with respect to the reference in kcal/mol. Atom pairs denote neighbor distances in Å; triples denote angles in degrees (double distance values show artificially broken symmetries of DFTB-optimized lattices). Italicized system names denote fit systems.

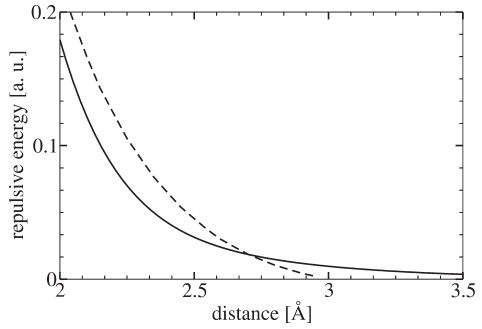

Figure 1. Comparison of the tiorg (dashed) and the automatically generated (solid line) Ti−Ti repulsives in the area of the sharp cutoff of the former.

4.3. Zinc−Oxygen Compounds. As a further demonstration for our fitting procedure, we attempted to create a parametrization for the Zn−O interaction. A high-quality and well-tested parameter set had been recently created manually for the zinc−organic chemistry by Moreira et al.,[19] which should serve as an etalon for our Zn−O repulsive. For the DFT references, the same settings had been used as for the handmade parametrization (PBE functional, double-ζ polarized basis, norm-conserving Troullier−Martins pseudopotentials, 8 × 8 × 8 Monkhorst−Pack scheme for *k* sampling). The fit paths were made with distortions applied to the test systems (see Table 2) in addition to Zn−Zn and Zn−O dimers with very low weights. The distortions applied to crystalline paths were uniform volume scaling and moving a Zn atom around. We show a comparison between the performance of the two Zn−O sets in Table 2. As fit targets, we used the two energy targets (energy and energy differences between steps weighted by 1:10); step weighting was by 10 and 2 in the immediate and in a wider neighborhood of equilibria. Here, the basis of repulsives consisted of exponential functions of type $e^{-a_2 r^2}$ and $e^{-a_3 r^3}$, as these shapes offered good results quickly in situations where absolute energy targets were not heavily weighted. As can be seen in Table 2, the resulting set is superior in the crystalline properties to the handmade one.

4.4. Titanium−Organic Repulsives. After the hydrocarbon fits, our next test of the fitting automaton was producing a titanium−oxygen set and extending it to a titanium−organic set. For this parametrization, a good-quality handmade set (tiorg) had been recently created by Dolgonos et al.[20] We used the same reference structures and *ab initio* reference data (various molecular systems calculated with the B3LYP functional and with mixed SDD+ basis set) augmented with crystalline reference systems. For the reference calculations of the periodic systems, the PBE functional, double-ζ plus polarized basis functions, and norm-conserving Troullier−Martins pseudopotentials had been used. K-point sampling was set to an 8 × 8 × 8 Monkhorst−Pack scheme with both Siesta and DFTB in this fit session.

In order to fit repulsive functions for the Ti−Ti and Ti−O interactions, we used a fit set including a titanium dimer (with a very low weight), a TiO_2 molecule, a planar Ti_2O_2 molecule, a tribridged Ti_2O_4, the bulk hcp titanium, and the bulk anatase and rutile forms of TiO_2. The molecular fit paths were created by stretching bonds and displacing titanium atoms while the crystalline paths were created by uniformly changing the volume of the crystal lattices and by using crystals with displaced titanium atoms. We used both energy and force targets (generally weighted 1:2) in the fit.

Table 4. Reference Data and Its Comparison with Previous Hand Made Parametrization[20] ("tiorg") and the Automatically Created One ("auto") for Various Titanium Compounds[a]

property	reference	tiorg	auto
	$Ti(CH_3)_4$		
ΔE_b	0	64.6	180.2
Ti−C	2.072	2.096	2.025
	$Ti(CH_3)_2$		
ΔE_b	0	−38.8	93.4
Ti−C	2.038	2.096	2.025
C−Ti−C	113.7	110.2	109.9
	crystalline TiC		
ΔE	0	111	91.7
Ti−C	2.141	2.159	2.170
Ti−i	3.024	3.047	3.067
	$Ti(NH_2)_4$		
ΔE_b	0	30.6	287.4
Ti−N	1.899	1.902	1.853
	$H_3Ti(NH_2)$		
ΔE_b	0	12.0	76.2
Ti−N	1.846	1.898	1.837
	HN═Ti═NH		
ΔE_b	0	15.5	156.1
Ti−N	1.707	1.703	1.671
N−Ti−N	114.8	114.7	113.7
	crystalline TiN		
ΔE	0	196.6	192.2
Ti−N	2.094	2.159	2.115
Ti−Ti	2.958	3.043	2.982
	Ti_2H_2 (dibridged planar)		
ΔE	0	123.5	131
Ti−Ti	1.985	1.967	2.011
Ti−H	1.868	1.827	1.899
Ti−H−Ti	64.2	65.2	63.9

[a] The values given here refer to the equilibrium structure of each system. ΔE denotes the atomization energy difference with respect to the reference in kcal/mol. ΔE_b indicates the binding energy between the central Ti atom and the ligands compared to the reference value in kcal/mol. Atom pairs denote neighbor distances in Å; triples denote angles in degrees. Italicized system names refer to fit systems.

In a fit session of a few days, we were able to produce a set of Ti−Ti and Ti−O repulsive potentials which reproduce energy and geometrical data in the same quality as the reference handmade set. A detailed comparison is given in Table 3. These results were obtained using $e^{-a_1 r}$- and $e^{-a_2 r^2}$-type exponential functions as basis functions for the fit because this analytical basis gave very good results quickly with the Ti and Ti−O chemistry.

After creating the repulsives for the Ti−Ti and Ti−O interactions, we extended the set to a complete Ti−organic set, still using exponential basis functions for expressing the repulsive potentials. The extension turned out to be more difficult than expected, mainly due to the sudden cutoff in the handmade Ti−Ti repulsive giving a very stable 3 Å Ti−Ti distance in hcp

titanium, titanium nitride, and titanium carbide. This feature (shown in Figure 1) can hardly be reconstructed with analytical sets. Although this peculiar shape is the numerically most convenient way to confine the range of the Ti−Ti repulsive well below the second-neighbor distance and gives good results for various systems, it may be an interesting question for future investigations whether it is a precise representation of the underlying physics.

Similar to the case of the titanium−oxygen fit, we used the same molecules (TiH_4, $Ti(CH_3)_4$, $Ti(NH_2)_4$) as during the handmade parametrization[20] extended by crystalline fit systems (TiN and TiC). The fit paths were similar constructions to the titanium−oxygen case. Every molecule had two fit paths with the relevant bonds stretched and the titanium atom displaced, while the crystalline systems had a volume change path and a path with a titanium atom displaced. We handled the relative difficulty of Ti−C and Ti−N fitting compared to Ti−Ti and Ti−O by lowering the relative weights of the energy targets in the Ti−C and Ti−N case. As can be seen from the results (Table 4), this resulted in fairly good geometries at the expense of accuracy in energies (one-body terms as a tool for resolving the conflict between energy and geometry accuracy was not used here).

5. CONCLUSION

We are suggesting a new fitting mechanism to create repulsive potentials for the DFTB method in an automatic way using least-squares fitting on automatically generated reference data. Using the proposed scheme, we fitted new repulsives for carbohydrogen systems, zinc and zinc oxide crystal structures, and titanium-containing organic compounds. Due to its efficiency and high degree of automatization, the fitting took in each case at most a couple of days' work of a researcher with the new fitting scheme. Comparing the new fits against existing handmade fits showed that we were able to create a general-purpose parametrization engine for the DFTB method. The engine enables us to optimize new parameters from scratch for any group of systems where the DFTB formalism with the pairwise repulsive potentials gives a reasonable description of the underlying physics. While these new parameter sets are very close in accuracy to handmade sets, they require considerably less time and human effort to be created.

The fitting procedure was planned to be as easy to handle, as comprehensive and as interactive as possible. As a demonstration of its easy integrability into current quantum chemical tools, it had been included into the Material Studio program package, using a graphical user interface to control the parametrization process.

APPENDIX

In the following sections, we give a detailed derivation about how the energy fitting procedure described above can be extended to objectives other than the basic energy objective (of them, the first two are fully implemented and tested in our program). This enables the extension of the fitting procedure to further physical properties (forces and frequencies) and the effective use of energy data from existing MD trajectories in the fitting process.

Fitting to Forces. The force objective from the repulsive interaction is the repulsive force $\mathbf{F}_i$ acting on atom i projected onto a unit vector (a direction) $\mathbf{u}$

$$\mathbf{F}_{i,\mathbf{u}} = \mathbf{F}_i \cdot \mathbf{u} = \sum_{j \neq i,\nu} \alpha_{\mathrm{type}(ij),\nu} f'_{\mathrm{type}(ij),\nu}(r_{ij}) \frac{\mathbf{r}_{ij}}{r_{ij}} \cdot \mathbf{u} \tag{24}$$

This can be, similar to the energy expression 14, decomposed into a linear combination

$$F_{i,\mathbf{u}} = \sum_{AB,\nu} \alpha_{AB,\nu} X^{(i,\mathbf{u})}_{AB,\nu} \tag{25}$$

with coefficients $\alpha_{AB,\nu}$, built of geometry-dependent factors

$$X^{(i,\mathbf{u})}_{AB,\nu} = \sum_{\substack{\mathrm{type}(ij)=AB \\ j \neq i}} f'_{AB,\nu}(r_{ij}) \frac{\mathbf{r}_{ij}}{r_{ij}} \cdot \mathbf{u} \tag{26}$$

containing the first derivatives of the basis functions $f_{AB,\nu}'(r)$. Because the α coefficients in these force components are the same as the ones used for the energy fitting, fitting to energies and forces can be unified when both are required. If $(F_{\mathrm{ref}} - F_{\mathrm{el}})_{i,\mathbf{u}}$ takes the place of $E_{\mathrm{ref}} - E_{\mathrm{el}}$ and the above new X's are used as independent variables, fitting to force components can be simply regarded as additional new fit paths. The matrices $\mathbf{E}$ and $\mathbf{X}$ can then be extended in the same way as in eqs 21a and 21b.

Fitting to MD Trajectory Energies. A problem that often compromises fitting to MD trajectories (or to large molecules where only a tiny part is distorted) is the fact that equilibrium bond lengths are heavily overweighted by their overwhelming presence in the sample fit paths. This can make efficient fitting to ranges of bond lengths other than the covalent equilibrium impossible with the original energy target described above.

A remedy of this problem can be found by fitting to *energy differences between subsequent fit steps* instead of energies of each fit step. As

$$\begin{aligned} \Delta E^{(s)}_{\mathrm{rep}} &= E^{(s+1)}_{\mathrm{rep}} - E^{(s)}_{\mathrm{rep}} \\ &= \sum_{AB,\nu} \alpha_{AB,\nu} (X^{(s+1)}_{AB,\nu} - X^{(s)}_{AB,\nu}) \end{aligned} \tag{27}$$

is a linear combination of structural quantities of type $X^{(s+1)} - X^{(s)}$, it is a valid target in our least-squares fit scheme. This modified energy target, however, contains virtually nothing arising from those bonds which do not change over the fit path; thus the overweighting of unchanged bonds is avoided. Of course, if fitting to absolute energy values at molecular equilibrium bond lengths is required, it can be brought back by an appropriate weighting between the original energy objective and the current one, or by defining additional molecular fit paths.

As an alternative use, the fit target based on the energy differences can also be used in cases where retrieving force data from a DFT reference is for some reason problematic or meaningless (e.g., with symmetric distortions of symmetric systems, atomwise total forces are constant zero). Using small distortion steps and the energy difference fit target, one automatically obtains a fit mimicking the fit on certain force or stress tensor components.

Fitting to Hessians. Similar to the forces, the repulsive contribution to the Hessian matrix of a chemical system can also be projected onto unit vectors $\mathbf{u}$ and $\mathbf{v}$ (these unit vectors can be regarded as virtual displacements of atoms). When both $\mathbf{u}$ and $\mathbf{v}$

are on the same ith atom (i.e., we examine the ith "on-site" 3×3 hyperdiagonal block of the $3N \times 3N$ collective molecular Hessian),

$$H_{i,\mathbf{uv}} = \mathbf{uHv} = \sum_{j\neq i,\,mn} u_m \frac{\partial^2 U_{\mathrm{type}(ij)}(r_{ij})}{\partial x_{i,m}\partial x_{i,n}} v_n$$

$$= \sum_{j\neq i,\,mn,\,\nu} \alpha_{\mathrm{type}(ij),\nu} u_m \frac{\partial^2 f_{\mathrm{type}(ij),\nu}(r_{ij})}{\partial x_{i,m}\partial x_{i,n}} v_n$$

$$= \sum_{AB} \sum_{\substack{\mathrm{type}(ij)=AB \\ j\neq i,\,\nu}} \alpha_{\mathrm{type}(ij),\nu} \left(\frac{1}{r^2}\frac{\partial^2 f_\nu}{\partial r^2} - \frac{1}{r^3}\frac{\partial f_\nu}{\partial r}\right)(\mathbf{u}\cdot\mathbf{r})(\mathbf{v}\cdot\mathbf{r})$$

$$+ \frac{1}{r}\frac{\partial f_\nu}{\partial r}(\mathbf{u}\cdot\mathbf{v}) \tag{28}$$

(with $\mathbf{r}_{ij} = (x_{j,1} - x_{i,1}, x_{j,2} - x_{i,2}, x_{j,3} - x_{i,3})$ at the beginning and $f_\nu = f_{\mathrm{type}(ij),\nu}(r_{ij})$, $r = r_{ij}$, and $\mathbf{r} = \mathbf{r}_{ij}$ in the last step). So, with

$$X^{(i,\mathbf{uv})}_{\mathrm{AB},\nu} = \sum_{\substack{\mathrm{type}(ij)=AB \\ j\neq i}} \left(\frac{1}{r^2}\frac{\partial^2 f_\nu}{\partial r^2} - \frac{1}{r^3}\frac{\partial f_\nu}{\partial r}\right)(\mathbf{u}\cdot\mathbf{r})(v\cdot r)$$

$$+ \frac{1}{r}\frac{\partial f_\nu}{\partial r}(\mathbf{u}\cdot\mathbf{v}) \tag{29}$$

the usual linear combinations can be written again

$$H_{i,\mathbf{uv}} = \sum_{AB,\nu} \alpha_{AB,\nu} X^{(i,\mathbf{uv})}_{AB,\nu} \tag{30}$$

With the $\mathbf{E}$ vector composed of $(H_{\mathrm{ref}} - H_{\mathrm{el}})_{i,\mathbf{uv}}$'s and the $\mathbf{X}$ matrix composed of the above X's, the fitting of the Hessian can be included as an additional path into the fitting scheme.

When $\mathbf{u}$ and $\mathbf{v}$ are on the ith and jth atoms, respectively,

$$H_{ij,\mathbf{uv}} = \mathbf{uHv} = u_m \frac{\partial^2 U_{\mathrm{type}(ij)}(r_{ij})}{\partial x_{i,m}\partial x_{j,n}} v_n$$

$$= -u_m \frac{\partial^2 U_{\mathrm{type}(ij)}(r_{ij})}{\partial x_{i,m}\partial x_{i,n}} v_n \tag{31}$$

Therefore, a similar construction applies to "off-site" Hessian parts, but with the opposite sign and without the summation over j.

As a linear combination of the above $\mathbf{u}$ and $\mathbf{v}$ atomic virtual displacements, every collective distortion of a molecule can be constructed. This knowledge can be used to fit to Hessians of DFT reference algorithms that give no detailed Hessian matrix but only vibrational modes and frequencies in their output. If $\mathbf{e}$ is a ($3N$-component) collective eigenmode of the molecular Hessian with ω frequency,

$$\omega^2 \mathbf{Me} = \mathbf{He} = (\mathbf{H}_{\mathrm{el}} + \mathbf{H}_{\mathrm{rep}})\mathbf{e} \tag{32}$$

where $\mathbf{M}$ is the diagonal mass matrix. The vector of equations contained in

$$\omega^2 \mathbf{Me} - \mathbf{H}_{\mathrm{el}}\mathbf{e} = \mathbf{H}_{\mathrm{rep}}\mathbf{e} \tag{33}$$

can then be used as a new fit path with the left hand side as a vector of $\mathbf{E}$ values and the right hand side as the usual linear combinations coming from the repulsives and using α's as coefficients.[21] Note that the last equation contains explicit Hessian data from DFTB only.

Fitting to the Stress Tensor. The repulsive part of the stress tensor in periodical systems is calculated as

$$\sigma_{mn} = -\frac{1}{\tilde{V}}\frac{\partial \tilde{E}_{\mathrm{rep}}}{\partial \varepsilon_{mn}} = \frac{1}{\tilde{V}} \sum_{\substack{i\in \mathrm{a\ cell} \\ j}} F_{ij,m} r_{ij,n}$$

$$= -\frac{1}{\tilde{V}} \sum_{\substack{i\in \mathrm{a\ cell} \\ j}} \frac{1}{r_{ij}}\frac{\partial U(r_{ij})}{\partial r_{ij}} r_{ij,m} r_{ij,n}$$

$$= -\frac{1}{\tilde{V}} \sum_{\substack{i\in \mathrm{a\ cell} \\ j}} \alpha_{\mathrm{type}(ij),\nu} f'(r_{ij}) \frac{r_{ij,m} r_{ij,n}}{r_{ij}} \tag{34}$$

where ε_{mn} is the strain tensor, $\tilde{V}$ is the unit cell volume, $\tilde{E}_{\mathrm{rep}}$ is the cellwise repulsive energy, $r_{ij,m}$ is a component of the relative position vector $\mathbf{r}_{ij}$ from the ith atom to the jth, and r_{ij} is the length of it. A double projection of σ_{mn} onto unit vectors $\mathbf{u}$ and $\mathbf{v}$ can be written

$$\sigma_{\mathbf{uv}} = \sum_{mn} \sigma_{mn} u_m v_n = \sum_{AB,\nu} \alpha_{AB,\nu} X^{(\mathbf{uv})}_{AB,\nu} \tag{35}$$

if our structural quantities are

$$X^{(\mathbf{uv})}_{AB,\nu} = -\frac{1}{\tilde{V}} \sum_{\substack{i\in \mathrm{a\ cell} \\ j}} f'(r_{ij}) \frac{(\mathbf{r}_{ij}\mathbf{u})(\mathbf{r}_{ij}\mathbf{v})}{r_{ij}} \tag{36}$$

So with the above X's, $\sigma_{\mathbf{uv}}$ can be another valid target of our repulsive fitting algorithm.

AUTHOR INFORMATION

Corresponding Author

*E-mail: bodrog.zoltan@bccms.uni-bremen.de.

ACKNOWLEDGMENT

The authors thank Gotthard Seifert, Martin Persson, and Grygoriy Dolgonos for fruitful discussions. Z.B. acknowledges support from the *Scientific Computing in Engineering* doctoral school at the University of Bremen.

REFERENCES

(1) Seifert, G. *J. Phys. Chem. A* **2007**, *111*, 5609–5613.
(2) Elstner, M.; Porezag, D.; Jungnickel, G.; Elstner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260–7268.
(3) Köhler, C.; Seifert, G.; Frauenheim, T. *Chem. Phys.* **2005**, *309*, 23–31.
(4) Niehaus, T. *THEOCHEM* **2009**, *914*, 38–49.
(5) Niehaus, T.; Rohlfing, M.; Sala, F. D.; Carlo, A. D.; Frauenheim, T. *Phys. Rev. A* **2005**, *71*, 022508.
(6) Pecchia, A.; Carlo, A. D. *Rep. Prog. Phys.* **2004**, *67*, 1497.

(7) Knaup, J. M.; Hourahine, B.; Frauenheim, T. *J. Phys. Chem. A* **2007**, *111*, 5637–5641.
(8) Gaus, M.; Chou, C.-P.; Witek, H.; Elstner, M. *J. Phys. Chem. A* **2009**, *113*, 11866–11881.
(9) Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 5614–5621.
(10) Displacements of a carbon atom in a carbohydrogen molecule alter the C−C and C−H bonds (but not the H−H bonds) and thus enable fitting to these repulsive potentials only.
(11) Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. *Phys. Rev. B* **1995**, *51*, 12947.
(12) Frisch, M. J. et al. *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2003.
(13) Soler, J. M.; Artacho, E.; Gale, J. D.; Garca, A.; Junquera, J.; Ordejón, P.; Sánchez-Portal, D. *J. Phys. (Paris)* **2002**, *14*, 2745.
(14) Aradi, B.; Hourahine, B.; Frauenheim, T. *J. Phys. Chem. A* **2007**, *111*, 5678–5684.
(15) "Displaced on *n* shells in a sphere" means that the atom is dislocated with a random vector on *n* spherical shells around its original position; the *n* equidistant shells are defined within the largest sphere, from radius 0 up to the largest radius. The random vectors are generated isotropically, one with length zero and at least four on each nontrivial shell. This way, a path with an atom jumping around *n* times contains at least 4*n* steps, plus one for the original configuration.
(16) Saunders, V.; Dovesi, R.; Roetti, C.; Orlando, R.; Zicovich-Wilson, C. M.; Harrson, N.; Doll, K.; Civalleri, B.; Bush, I.; D'Arco, P.; Llunell, M. *CRYSTAL2003 User's Manual*; University of Torino: Torino, Italy, 2003.
(17) Dovesi, R.; Causa, M.; Orlando, R.; Roetti, C.; Saunders, V. *J. Chem. Phys.* **1990**, *92*, 7402–7411.
(18) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **1997**, *106*, 1063–1079.
(19) Moreira, N. H.; Dolgonos, G.; Aradi, B.; da Rosa, A. L.; Frauenheim, T. *J. Chem. Theory Comput.* **2009**, *5*, 605–614.
(20) Dolgonos, G.; Aradi, B.; Moreira, N. H.; Frauenheim, T. *J. Chem. Theory Comput.* **2010**, *6*, 266–278.
(21) An important issue is whether we must compare DFT-equilibrium Hessians to DFTB-equilibrium Hessians or we must compare Hessians of the very same geometry (practically, the DFT equilibrium geometry, as DFT calculators tend to compute (real) eigenvalues and eigenmodes instead of outputting raw Hessian matrices). From a theoretical point of view, the latter comparison is more valid, while from a semiempirical point of view, the former comparison is much more justified.