The Journal of Chemical Physics

Driving torsion scans with wavefront propagation

Cite as: J. Chem. Phys. **152**, 244116 (2020); https://doi.org/10.1063/5.0009232
Submitted: 28 March 2020 • Accepted: 26 May 2020 • Published Online: 25 June 2020

Yudong Qiu, Daniel G. A. Smith, Chaya D. Stern, et al.

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Driving torsion scans with wavefront propagation

Cite as: J. Chem. Phys. 152, 244116 (2020); doi: 10.1063/5.0009232
Submitted: 28 March 2020 • Accepted: 26 May 2020 •
Published Online: 25 June 2020

Yudong Qiu,[1] Daniel G. A. Smith,[2] Chaya D. Stern,[3] Mudong Feng,[4] Hyesu Jang,[1] and Lee-Ping Wang[1,a)]

AFFILIATIONS

[1] Department of Chemistry, UC Davis, Davis, California 95616, USA
[2] The Molecular Sciences Software Institute, Blacksburg, Virginia 24060, USA
[3] Computational and Systems Biology Program, Sloan-Kettering Institute, New York, New York 10065, USA
[4] Department of Chemistry and Biochemistry, UC San Diego, La Jolla, California 92093, USA

[a)] **Author to whom correspondence should be addressed:** leeping@ucdavis.edu

ABSTRACT

The parameterization of torsional/dihedral angle potential energy terms is a crucial part of developing molecular mechanics force fields. Quantum mechanical (QM) methods are often used to provide samples of the potential energy surface (PES) for fitting the empirical parameters in these force field terms. To ensure that the sampled molecular configurations are thermodynamically feasible, constrained QM geometry optimizations are typically carried out, which relax the orthogonal degrees of freedom while fixing the target torsion angle(s) on a grid of values. However, the quality of results and computational cost are affected by various factors on a non-trivial PES, such as dependence on the chosen scan direction and the lack of efficient approaches to integrate results started from multiple initial guesses. In this paper, we propose a systematic and versatile workflow called *TorsionDrive* to generate energy-minimized structures on a grid of torsion constraints by means of a recursive wavefront propagation algorithm, which resolves the deficiencies of conventional scanning approaches and generates higher quality QM data for force field development. The capabilities of our method are presented for multi-dimensional scans and multiple initial guess structures, and an integration with the MolSSI QCArchive distributed computing ecosystem is described. The method is implemented in an open-source software package that is compatible with many QM software packages and energy minimization codes.

I. INTRODUCTION

The potential energy surface (PES) along the torsional dihedral angle degrees of freedom is a crucial part of model potentials for computer simulations of bio/organic molecules and polymers, including commonly used molecular mechanics force fields. "Proper" torsion angles (i.e., those involving four consecutively bonded atoms a—b—c—d and labeled as ϕ_{abcd}) can be highly flexible due to the periodic nature and relatively small range of the free energy profile (often less than 5 kcal mol^{-1}), which leads to broadly diverse conformations and accessible barrier crossings in ambient temperature experiments and simulations. Because the torsional angle is a principal descriptor of molecular conformation, the torsional potential energy is important for determining the thermodynamic distribution of molecular conformations and kinetics of conformational changes. Therefore, accurate empirical potentials, or molecular mechanics (MM) force fields are needed to predict properties of interest such as biomolecular structure and function, receptor-ligand binding free energies, and timescales of protein folding.[1–8]

The four-body energy term for proper torsion in most force fields uses a periodic functional form of the dihedral angle ϕ_{abcd} represented as a truncated Fourier series, i.e.,

$$E(\phi_{abcd}) = \sum_{n=1}^{N_k} k_{abcd}^{(n)}(1 + \cos(n\phi_{abcd} - \phi_{abcd;0}^{(n)})), \quad (1)$$

where the sum is over periodicity n and $N_k \leq 6$. The potential parameters for barrier height and phase shift $k_{abcd}^{(n)}, \phi_{abcd;0}^{(n)}$ may be assigned from parameter libraries based on the chemical environment, or they may be specifically fitted (i.e., bespoke) for an individual torsion angle of a specific molecule. The non-covalent interaction between the terminal atoms of the torsion angle may also be modeled using pairwise Coulomb and Lennard-Jones interactions on atoms separated by exactly three bonds (i.e., "1–4 interactions"), which

may be modified from conventional non-bonded terms using scaling factors or alternative parameter values.[9] Because the 1–4 distance depends strongly (but not exclusively) on the torsion angle, it may be considered as another contribution to the torsional potential energy.

Proper torsions have characteristics of both valence (i.e., bonded) and nonbonded regimes because the total energy includes contributions from the quantum nature of covalent bonding such as resonance and conjugation, as well as non-covalent interactions such as electrostatic and steric effects on vicinal functional groups. As the torsion angle is varied in a molecule, several important properties of the molecule are affected including the electronic character of the central bond as well as steric and other nonbonded interactions between groups on opposite sides of the bond.[10–12] Importantly, the torsion angle dependence of these properties can induce relaxations in the orthogonal degrees of freedom as the torsion angle is varied. Such relaxations include bond stretching that accompanies disruption of conjugation, the bending of angles to minimize steric hindrance, and changes in distance between nonbonded functional groups in order to avoid clashes or make intramolecular contacts.

Force fields must accurately account for torsion-induced structural relaxations in order to produce accurate free energy profiles; thus, the standard practice of generating quantum mechanical (QM) data for MM force field parameterization involves minimizing the QM potential energy with the torsion angle of interest constrained to various values, e.g., on a regularly spaced grid.[13–16] The result of this calculation is a set of QM constrained optimized structures and energies that includes relaxation effects from orthogonal degrees of freedom, which can be used to develop more accurate torsion parameters in the context of other energy contributions in the force field. In addition, two or more dihedral angles can be varied independently on a multi-dimensional grid to sample the conformational space more broadly and/or to generate data for parameterizing torsion–torsion coupling (also called CMAP) energy terms used in some force fields.[17,18]

For relevant molecular systems, the feasible geometry optimization methods involve local energy minimization starting from an initial guess structure. The optimized structure and energy, as well as the probability of the optimization algorithm successfully converging to a minimum, both depend strongly on the initial guess. The straightforward approach to this problem involves carrying out a series of constrained minimizations where the constraint value is scanned along the grid, and each minimization is initiated from the optimized structure of the previous one.[19] This calculation, which we term a "serial relaxed scan," is a standard feature in several widely used quantum chemistry and geometry optimization codes.[20–28]

The serial relaxed scanning approach has some major drawbacks.[29] For one, the resulting optimized structures are dependent on the chosen sequence of calculations, such as the direction of the one-dimensional scan. This is because a series of constrained minimizations often stays in the same qualitative local minimum as determined by the orthogonal degrees of freedom even if another local minimum with a lower energy is reachable by scanning in the opposite direction; the other local minimum is found only if the energy barrier vanishes, which is not guaranteed. The result thus has a risk of including structures with unnecessarily high potential energies, which are not appropriate for fitting force field parameters because they introduce a bias toward thermodynamically unlikely conformations. This problem becomes more serious for multi-dimensional scanning as a greater number of choices need to be made for the scanning direction, and the results may depend on the ordering of dimensions. Another drawback is the lack of an efficient way to use multiple initial guesses, such as those resulting from a conformer generation method;[30,31] intuitively, it should be possible to perform one scan using several initial guesses and keep the lowest energy structure from each but at a lower total cost compared to running each scan independently.

In this manuscript, we describe a new systematic workflow for generating optimized geometries along grids of torsion constraints by wavefront propagation, which addresses the drawbacks of serial relaxed scanning described above. The method, called *TorsionDrive*, generates results that are independent of scan direction and naturally incorporates multiple initial guesses into a single grid of constrained minimized structures. A predecessor to the present method was used to scan the two-dimensional torsion angles of blocked amino acid dipeptides in Ref. 8, but due to limitations of implementation, the method was limited to 2D grids and was not easily applicable to other molecules. The present method can be applied to any molecule subject only to the limits of the underlying energy and gradient method. Furthermore, it is capable of driving an arbitrary number of torsions to generate N-dimensional grids of optimized structures and energies where $N \geq 1$. The current workflow is implemented as a Python package that interfaces with energy minimization routines in a modular way, including the open-source geomeTRIC optimization package[28] that uses externally obtained energy and gradient information, as well as "native" optimization methods implemented in many quantum chemistry codes. The method is released as an open source package[32] and includes a number of useful features such as including energy upper limits, extra constraints, and limited scan ranges. In addition to the standalone operation mode, TorsionDrive is also implemented as a service in the MolSSI QCArchive ecosystem,[33] and it is available to compute results for any implemented gradient method in QCArchive; this includes not only quantum chemistry methods but also some MM force fields and recently developed neural network potentials parameterized by machine learning methods.[34]

II. METHOD

The main idea of TorsionDrive can be described conceptually as scanning the torsion angles with wavefront propagation. The details are illustrated by walking through a complete scan procedure, shown in Fig. 1. Before the start of the scan, we specify the dihedral angles in the molecule of interest using quartets of atomic indices and the spacing (resolution) of the scan. A grid of constraint values is created, which has the same dimension as the number of dihedral angles provided. In the illustration, we perform a 1D scan with a 60° spacing. The data associated with each grid point are represented by a circle, which consist of one or more "optimization datasets" (i.e., the Cartesian coordinates of a constrained local minimum and corresponding QM energy and gradient). Importantly, each grid point is able to contain multiple constrained optimization datasets that may correspond to local minima with different energies and values of the orthogonal degrees of freedom.

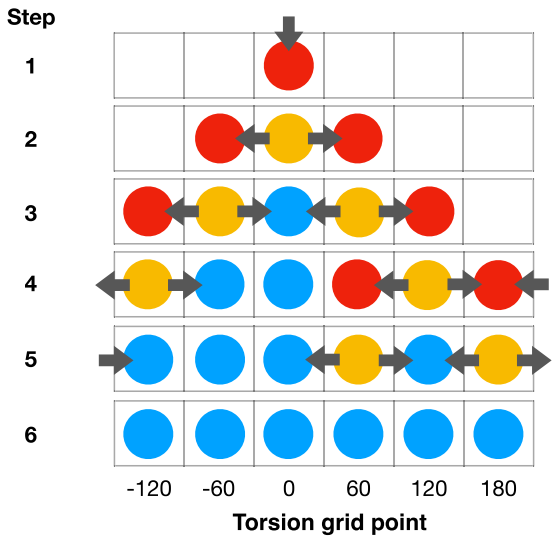

FIG. 1. TorsionDrive method illustration. Steps proceed from top to bottom. Red: new active point; orange: active point from last step; blue: inactive point; and arrow: constrained optimizations that were carried out in the current step. See the text for details.

Given the above input, TorsionDrive starts to iteratively fill and refine the data of all grid points using constrained energy minimizations, denoted by the arrows in Fig. 1. Each constrained optimization is specified by an initial molecular structure and a target set of dihedral angle constraints, and the result is an optimized structure that matches the constraints with the other degrees of freedom relaxed. Within this workflow, TorsionDrive specifies the optimizations and processes the output data, and the actual optimizations are carried out via interfaces to other software packages. Multiple constrained optimizations that are specified at the same step in the workflow can be carried out in parallel. In the standalone operation mode, TorsionDrive uses the Work Queue distributed computing framework[35,36] to take advantage of parallel resources. When TorsionDrive is used as part of the QCArchive ecosystem,[33] it works as an application programming interface (API) to specify the constrained optimization inputs while QCArchive is responsible for job management; this is described in detail in Sec. IV.

The steps of the example scan (i.e., rows in Fig. 1) are described in the following example. For clarity of presentation, it is necessary to define the basic procedures within a step and the separation between steps. At the start of a step, constrained optimization calculations are started based on the results of the previous step. Upon completion of these calculations, some grid points are set as "active points" as described below, and then, the step is concluded. The result of one step is independent of the order of completion of the individual optimizations within a step.

Step 1: An initial constrained optimization is performed starting from the user-provided initial geometry of the molecule, with constraints set equal to the closest dihedral grid point (0° in the example). After the optimization is completed, the optimization data (structure and energy) are assigned to the grid point, and it is set as an "active" point, denoted by the red color.

Step 2: New constrained optimizations are launched from each active point of Step 1 toward each of its neighboring points. The number of neighboring points is equal to 2× the dimension of the scan. In this example, there is one active point at 0° in step 1, and two constrained optimizations are started at the two neighboring points (−60° and 60°) in Step 2. The active points from the last step, which are used to launch the optimizations in the previous step, are colored orange. Upon completion of the two constrained optimizations, the two neighboring points gain their initial set of optimization data, and they are set as active points, colored in red.

Step 3: The two active points from Step 2 spawn new optimizations toward each of their neighbors. Two such constrained optimizations expand to the left and right, resulting in new active grid points at −120° and 120°. The other two constrained optimization are both targeted at the grid point at 0°; thus, the grid point at 0° gains two new sets of geometries and energies, which are potentially better (lower in energy), equal, or worse (higher in energy) compared to the existing data. To determine which data to keep, we compare the energy of each new result with the current lowest energy at this grid point. In this example, we assume that both new optimization results are equal to or higher than the energy obtained from the original optimization in Step 1. In such cases, the grid point is marked as inactive (blue).

Step 4: The two active points from Step 3, located at −120° and 120°, spawn four new constrained optimizations. Since the dihedral grid is periodic, the "leftward" optimization from −120° wraps around to the "right-most" grid point at 180°. The "rightward" optimization from +120° also targets this grid point. The result of the two new optimizations is compared, and the one with the lowest energy is assigned as the new data for the grid point at 180°, which is assigned as active (red) in the current step. In this example, the optimization from −120° to −60° results in an equal- or higher-energy geometry similar as before. However, the optimization from 120° to 60° results in a lower-energy geometry due to finding a lower-energy local minimum. To explore the potential energy surface around this new lowest-energy local minimum, the 60° grid point is set as an active point.

Step 5: The active points from Step 4 at 60° and 180° spawn four constrained optimizations. In this example, we assume that all four new optimizations result in equal or higher energy structures compared to stored data, so all four points are set to inactive.

Step 6: There are no active points from Step 5. The TorsionDrive procedure is complete, and the data for the lowest-energy structure at each grid point are compiled and saved. The data from other constrained optimizations at equal or higher energies are retained in the scratch space of the calculation but are not considered to be part of the final result.

To summarize, the TorsionDrive scan follows these rules: (a) Any grid point that gains its initial set of optimization data, or new optimization data with lower energy than its current lowest energy, is set to "active". (b) All active points from the previous step spawns new constrained optimizations, starting from the lowest-energy structure, targeting all neighboring grid points. (c) If no active point is left, the scan converges.

The above example only illustrates a simple 1D scan. It should be noted that TorsionDrive supports dihedral scans of arbitrary dimensions, with the minimum cost scaling as by $O(2d \times N^d)$, where N is the number of grid points on each dimension and d is the number of dimensions. In addition, multiple initial geometries can be provided to improve the coverage of the PES. Over the course of applying this software package in ongoing research projects, we have also created additional features that we found useful, which are stated in Sec. III D.

III. RESULTS AND DISCUSSION

A. One-dimensional scanning example

A comparison between a 1D scan in TorsionDrive and a conventional serial relaxed scan is shown in Fig. 2. The dihedral angle to be scanned is indicated by the four highlighted atoms. In both cases, the calculation is initiated from a single structure with a dihedral angle of 0°, and scans were performed with a 15° resolution. Geometry optimizations were carried out using the geomeTRIC software,[28] and energies and gradients were calculated using density functional theory (DFT), as implemented in the TeraChem software package.[26,27] A restricted Kohn–Sham wavefunction and the B3LYP hybrid functional[37] with the corresponding D3(BJ) empirical dispersion correction[38–40] were used.

The serial relaxed scan is carried out in the $+\phi$ direction, and the result is clearly asymmetric as there are two regions in the plot around −90° and +90° where the energy rises gradually and then sharply drops making a sawtooth pattern. This occurs because as the torsion angle deviates from planar, both atoms of the central bond start to adopt pyramidal geometries. The energy barrier to pyramidal inversion causes the optimizations to yield increasingly high-energy structures and eventually breaks down causing the large energy drop. Although the shape of this particular PES might be due to the lack of multireference effects in the wavefunction,[41] it is sufficient to illustrate the general tendency of serial relaxed scans to get stuck in local minima in the orthogonal degrees of freedom. By contrast, the energy profile generated by TorsionDrive is more symmetrical, as expected from the twofold symmetry of the molecule. The wavefront propagation procedure initiates constrained optimizations in both directions, and although the central atoms still adopt pyramidal geometries, both "branches" of the potential energy surface are treated equally. Moreover, the final energy profile generated by TorsionDrive has significantly lower energy barriers compared to the serial scan though the barrier is still quite high at around 30 kcal/mol.

To compare the quality of these data for force field fitting, we computed MM single point energies at the optimized structures using the recently developed Open Force Field "Parsley" small molecule force field version 1.1.0,[42,43] which did not include this molecule in its training set. The results show that the highest-energy conformations in the sequential scans have QM−MM energy differences that are more than twice as large as the wavefront propagation scan (Fig. 2, red lines). These data would introduce unwanted biases during force field fitting as the parameters would tend to minimize the energy errors in the highly strained structures at the expense of accuracy in the lower-energy regions. Therefore, we think that the QM data from wavefront propagation can improve the "ingredients" for force field fitting and ultimately lead to more accurate parameters.

In terms of computational cost, the serial relaxed scan involved a total of 24 constrained optimizations (601 gradient evaluations), whereas the TorsionDrive calculation involved 19 wavefront propagation steps with a total of 91 constrained optimizations (2073 gradient evaluations). Although the total computational cost of

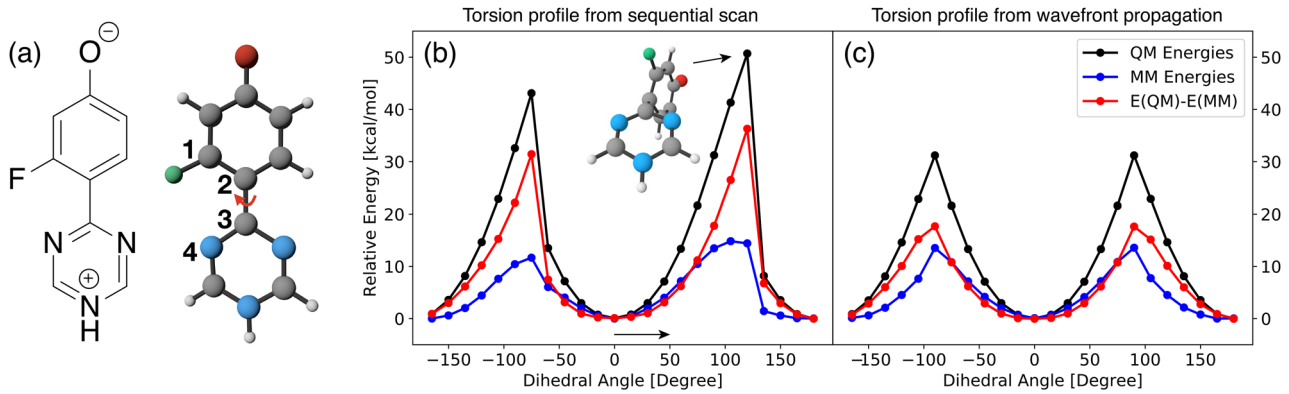

FIG. 2. Comparison of one-dimensional torsion scans for zwitterionic 3-fluoro-4-(1,3,5-triazin-2-yl)phenol carried out at the B3LYP/6-31G* level of theory. Black: QM constrained optimized energies; Blue: MM single-point energies at same geometries from OpenFF "Parsley" force field; Red: QM−MM energy difference. (a) Molecular structure with labeled indices for the torsion being scanned. (b) Result of the serial relaxed scan with the scan direction indicated and 3D rendering of the highest energy structure. (c) Result of the wavefront propagation scan using TorsionDrive. The geomeTRIC package was used to carry out the constrained optimizations in both cases.

TorsionDrive is higher, we note that the wall time to job completion may actually be faster if sufficient parallel resources are made available (for 1D scans, four parallel jobs is mostly sufficient).

B. Two-dimensional scanning example

Multi-dimensional torsion scans provide greater insight into the conformational degrees of freedom of many biologically relevant molecules. For example, the backbone torsion angles of proteins occur in (φ, ψ) pairs, and amino acid side chains and glycosidic bonds contain flexible chains with two or more connected torsions. When scanning the torsion angle in two or higher dimensions, the conventional serial scanning approach suffers from similar problems as in the 1D case, but the problems may be more serious. In addition to the two choices of scan direction for each dimension, the *ordering* of dimensions may also affect the results because only one torsion angle may be varied between contiguous calculations, while the others are held fixed. Molecular systems with two or more coupled torsions tend to exhibit a high degree of flexibility, which also increases the chances of multiple local minima that are easily missed by a sequential scan.

Figure 3 compares the results for glutamine dipeptide for the wavefront propagation scan with TorsionDrive and two sequential scans with different choices of dimensional ordering. The grid spacing, level of theory, and software used were the same as Sec. III A, and there are now 576 total points on the torsion grid due to the increased dimensionality. The scan is initiated from a single structure near the energy minimum where (φ, ψ) = (−83°, 62°). While the results appear similar in the low-energy region near the starting structure, there are major differences in the more distant regions of the energy profiles. Namely, the serial scan results include several high-energy "islands" in excess of +25 kcal/mol, whereas the TorsionDrive scan results have much lower energies in these regions. The sequential scan results also contain more sharp differences in the energy between adjacent grid points, for example, near (+90, +90)° in Fig. 3(c), whereas the TorsionDrive energy profile is smoother. Some of the most significant differences between the potential energy surfaces are highlighted by the starred regions, indicating that both sequential scans visited higher-energy local minima compared to the wavefront propagation scan. These results show that serial scanning poses an increased risk of providing incorrect or insufficient data compared to wavefront propagation scanning for parameterizing force fields in simulations.

In terms of computational cost, the sequential relaxed scans involved running 576 geometry optimizations and a total of 21 208/21 658 gradient evaluations, depending on the ordering of dimensions. The TorsionDrive calculation ran for 33 wavefront propagation steps, involving a total of 4953 energy minimizations and 166 714 gradient evaluations. The number of gradient evaluations in the TorsionDrive calculation is about 7.5 times greater than the sequential relaxed scans, but the wall time may be greatly reduced if parallel resources are available as each wavefront

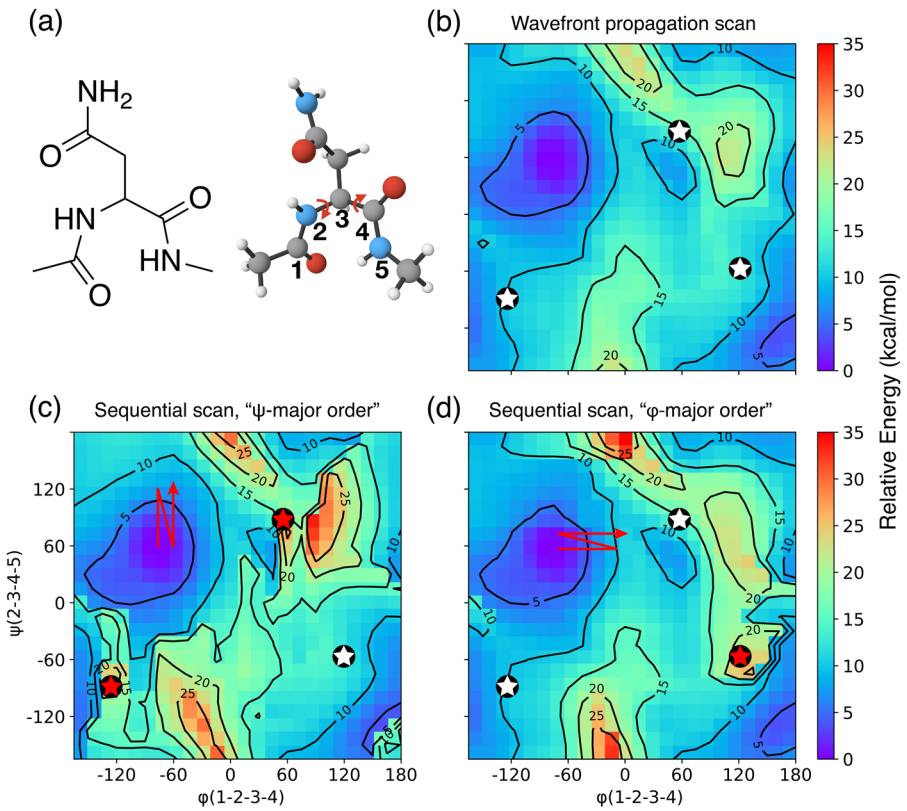

FIG. 3. Comparison of two dimensional torsion scans of glutamine dipeptide at the B3LYP/6-31G* level of theory. Contour lines are drawn at 5 kcal/mol intervals. (a) Line drawing and initial 3D structure of the scanned molecule. The two coupled torsion angles being scanned are denoted by red curved arrows and, more specifically, by indexed atoms φ(1-2-3-4) and ψ(2-3-4-5). (b) Wavefront propagation scan with TorsionDrive. (c) Sequential scan results in ψ-major order (consecutive elements of ψ are next to each other, and φ is incremented upon completion of scanning ψ). (d) Sequential scan results in φ-major order (consecutive elements of φ are next to each other). Red arrows conceptually illustrate ordering of dimensions and scan directions. Starred regions indicate where the potential energy surfaces differ significantly (red = higher energy).

propagation step could launch up to 300 energy minimizations in parallel. In the ideal case that all calculations are able to run in parallel, the TorsionDrive calculation wall time would be equivalent to around 33 sequential geometry optimizations.

One can also take advantage of parallelism in other ways, such as by slightly modifying the sequential scanning approach to use the results of a 1D scan along one dimension to start an array of 1-D scans along the other dimension to create the 2D PES. In this case, the number of sequential geometry optimizations is reduced to as low as 26 (if one goes in both directions simultaneously). Figure S1 of the Supplementary material shows an example where a unidirectional 1D scan along φ is used to start an array of 24 1D scans along ψ. The shape of the PES and general locations of high-energy minima are largely consistent with Fig. 3(c), in line with expectations.

C. Multiple initial structures

The wavefront propagation procedure of TorsionDrive is naturally able to incorporate multiple starting conformations. The initial constrained optimizations are performed on all starting structures with the torsion angle constrained to the closest grid point. If more than one initial structure is mapped to the same grid point, the lowest energy optimized conformer is used to launch new constrained optimization for neighboring points. This feature is beneficial because a grid of torsion angles can be covered in a smaller number of wavefront propagation steps when starting from multiple structures, and it also provides a natural way to consistently include the lowest-energy local minimum from multiple initial guesses.

In many molecules, the potential energy surface includes coupling across multiple torsion angles due to intramolecular nonbonded interactions, with protein backbone torsion angles (φ, ψ) being a well-known example. Ethylene glycol is an example of a molecule with strong intramolecular interactions between the hydroxyl functional groups. Figure 4 compares the results of a 1D torsion drive started from one initial conformation (indicated with +) and multiple initial conformations (indicated with *) together with a 2D torsional PES. These calculations were performed within the QCArchive infrastructure that provides TorsionDrive calculations as a service, as described in Sec. IV B. Energies and gradients were calculated using the B3LYP-D3(BJ) functional and DZVP basis optimized for DFT,[44] as implemented in the Psi4 software package,[23] and optimizations were carried out using the geomeTRIC software.[28]

The results show that a 1D torsion scan started from a conformation where the hydroxyl groups are facing in opposite directions (4b upper) will fail to find the lowest energy conformer. However, when the 1D scan is started from multiple conformations with different starting values of the O–C–C–O and H–O–C–C torsion angles, the resulting scan includes some structures that contain intramolecular hydrogen bonding character and lower overall energies (4b lower). Most conformers found by this scan are lower in energy than the structures found in the other scan. The 2D scan is shown in Fig. 4(c) with the two 1D scans mapped onto the heatmap. While the H–O–C–C dihedral angle does not change a lot in the scan started with one initial conformation (indicated with +), the H–O–C–C dihedral angle of the scan started with multiple conformations (indicated with *) changes to follow the lowest energy path on the potential energy surface.

It is well known that intramolecular electrostatic contacts (IECs), such as intramolecular hydrogen bonds, are much stronger in the gas phase than in solvent.[45,46] This can be attributed to the dielectric solvent's attenuation of electric fields and competing hydrogen bonding effects from the solvent, and is consistent with the observation that non-polarizable force fields for condensed phase simulation tend to underestimate gas-phase hydrogen bond energies.[47] Therefore, searching for the lowest-energy structure in systems with strong gas-phase IECs could cause undesirable biases when fitting parameters of non-polarizable force fields; polarizable force fields are not as susceptible to this problem.[48] One approach to avoiding forming IECs in geometry optimizations is to modify the potential surface by adding artificial repulsive potentials between groups expected to interact electrostatically[49,50] or by using an implicit solvent model. Another promising approach is to carry out all steps of QM data generation and force field fitting with an implicit solvent model.[51] Current implicit solvent implementations make it difficult to use the same solvent model for MM and QM during force field fitting, though this effect appears to be minor; we look forward to seeing more advances in this area in the future.

D. Generalized dihedral scanning and restricted grid

TorsionDrive has several additional features for flexibility in setting scanning coordinates for different molecular systems. First, the scanned coordinate(s) are not required to be strict definitions of proper torsion angles as four atoms in three sequential covalent bonds; any dihedral angle defined by four atom indices can be scanned, such as generalized torsion angles defined by four nonconsecutive atoms or improper torsion angles describing pyramidalization. Second, the scan range can be restricted to focus computations on regions of interest, in case certain ranges of the dihedral angle are not physically reasonable.

The usefulness of these non-conventional torsion degrees of freedom is further enhanced by TorsionDrive's robustness in building a smooth PES even for difficult systems. As an example, we conducted dihedral scans on a molecular motor that works by rotating around its torsion angle, as shown in Fig. 5. The subject molecule is a crowded and strained alkene where rotation of central double-bond torsion has to overcome a considerable energy barrier.[52] Between the two sides of the rotation barrier, there are also large structural differences such as ring pucker flips. In this example, rotation around the central double bond is characterized by a torsion angle defined by four non-consecutive atoms because using consecutive atoms to define the torsion angle would lead to poor projection of the barrier onto the scanned coordinates.

The performance of TorsionDrive and conventional serial scanning is compared in Fig. 5 where a generalized proper torsion and improper torsion are scanned over. The serial scan started from the (−30, 0) grid point, and the dimensions are ordered such that consecutive optimizations involved changing the improper torsion angle. A comparison of the two calculations shows that TorsionDrive and serial scanning produce markedly different potential surfaces, with TorsionDrive giving a superior result in terms of finding much lower energy conformations. The serial scan succeeded in locating the leftmost minimum near the start point of the scan (−30, 0) but failed to obtain the other two local minima found by

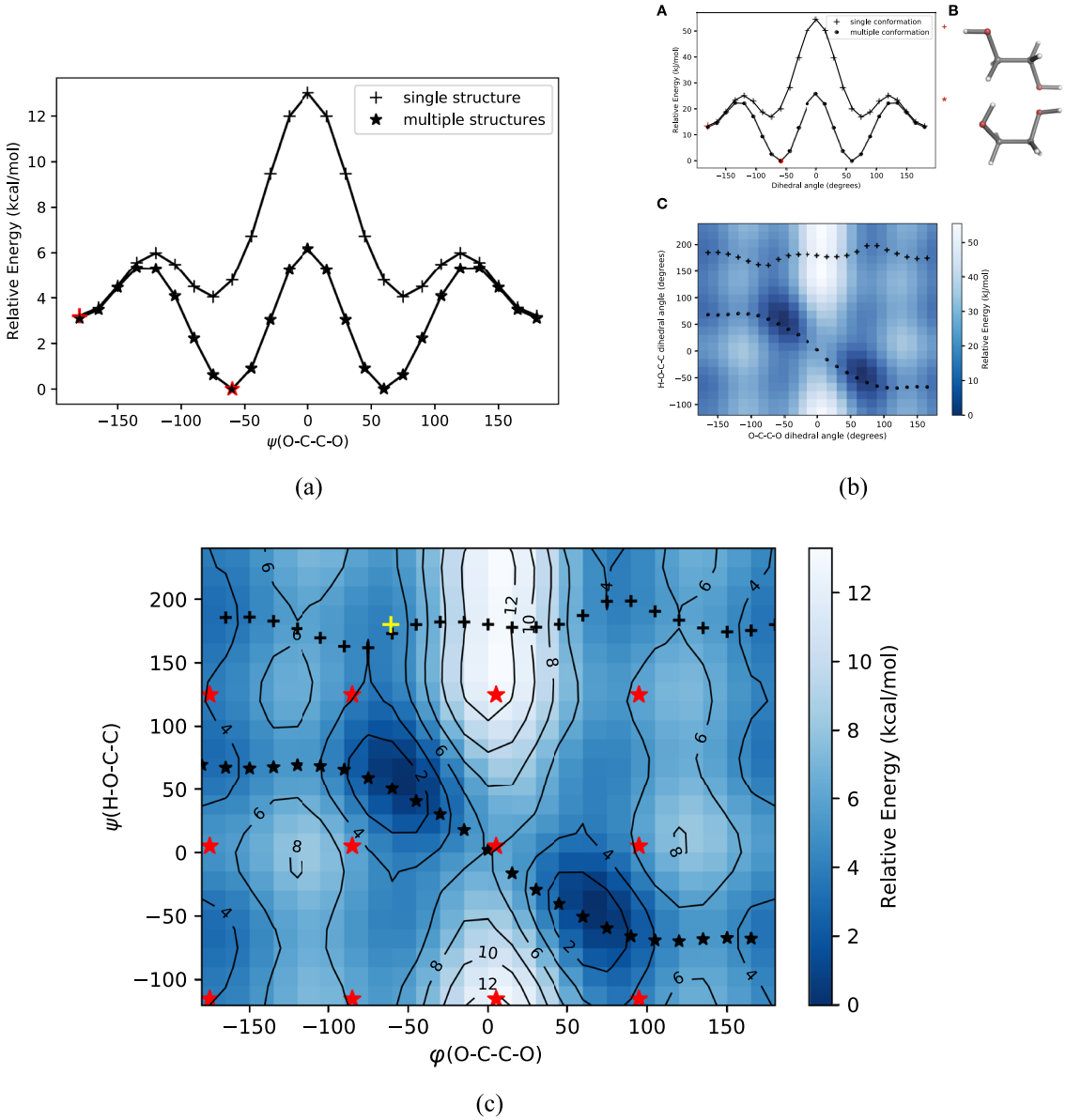

FIG. 4. Comparison of torsional potential surfaces computed using TorsionDrive with single vs multiple starting conformations. (a) One-dimensional scans of the torsion angle formed by atoms O–C–C–O started from one conformation (+) vs multiple conformations (⋆). Red color indicates the lowest energy structure. (b) 3D renderings of lowest energy structures found in (a). (c) 2D torsion scan along O–C–C–O and H–O–C–C torsion angles. 1D scan results for single and multiple starting conformations are mapped onto the heat-map as (+, ⋆), with colored symbols indicating starting structures. The 1D scan using multiple starting conformations finds the lowest energy conformer on the 2D scan.

TorsionDrive. Moreover, the serial scan always yielded equal or higher energy than TorsionDrive at each grid point and reached many structures in excess of 40 kcal/mol on the right side of the PES. The choice of dimensional ordering also affects the outcome of a serial scan as the serial scan crashed at the (180, 0) grid point when the opposite ordering was used, likely due to reaching an excessively strained structure.

In summary, we think that the superior performance of TorsionDrive is because it optimizes multiple structures at the same grid point from different propagating directions. This procedure makes the TorsionDrive result less sensitive to the sometimes unpredictable convergence of geometry optimization methods to different local minima depending on starting conformation. This procedure, optionally augmented by using multiple starting conformations,

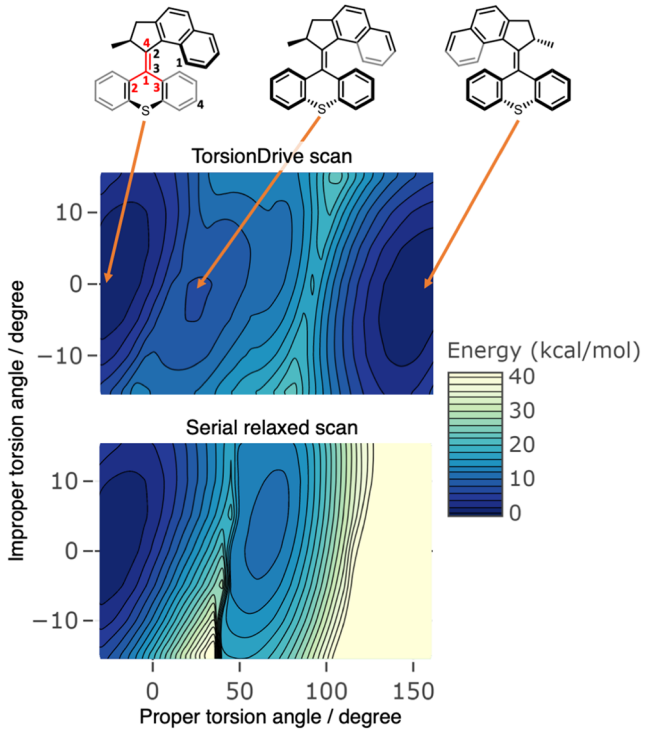

FIG. 5. Comparison of 2D scans (generalized proper torsion—four atoms numbered in the chemical structure and improper torsion—marked by red in the above chemical structure) for characterizing the PES of a molecular motor at the B3LYP-D3/3-21G level of theory. Chemical structures of three local minima are drawn on the top, with gray and bold, indicating "behind" and "in front of" the plane, respectively; result of the wavefront propagation scan using TorsionDrive is shown in the middle; result of the serial relaxed scan is shown at the bottom. The results were generated using geomeTRIC as the geometry optimization method, interfaced with Psi4 for gradient calculations.

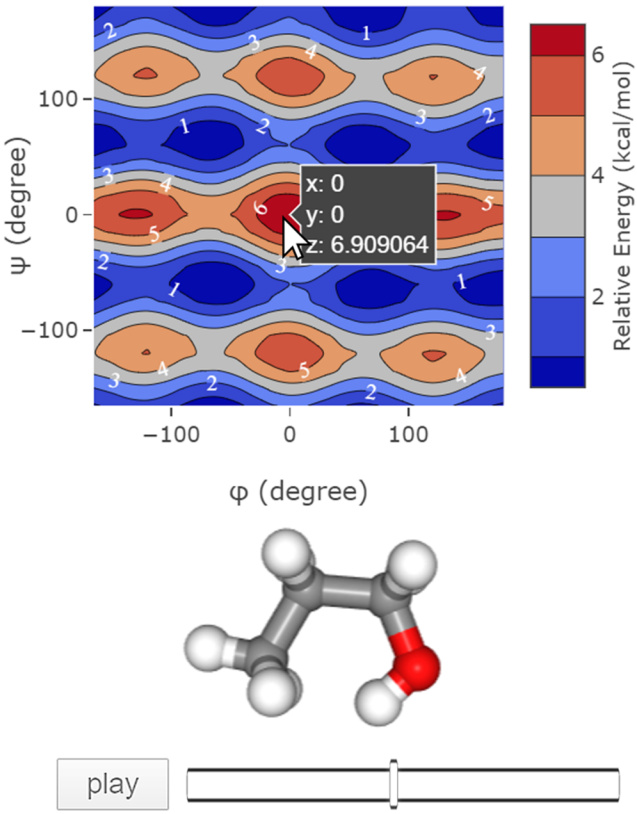

FIG. 6. Image for the typical usage of the visualization notebook.

results in a higher quality PES and much higher chances of finding relevant local minima. In addition, TorsionDrive saves human effort and wall time in troubleshooting and restarting crashed calculations due to its robustness against the sometimes unpredictable convergence failure of geometry optimizations.

E. Data analysis and visualization

The TorsionDrive software package includes Python scripts to parse output files into formatted files containing energies, gradients, coordinates, and associated metadata. These file formats are designed to facilitate automated fitting of force field parameters, which we will describe elsewhere. Also provided are Jupyter Notebooks[53] for interactive visualization of the resulting energy surfaces (powered by Plotly) and inspection of corresponding molecular structures (powered by NGLview[54]). Figure 6 shows a typical usage of the visualization notebooks. Upon execution of the code cells, a contour is plotted to visualize the energies at each grid point, and hovering the mouse pointer on the plot shows the dihedral and energy values of the nearest grid point. Left clicking on the plot displays the optimized structure of the nearest grid point. The displayed structure is interactive and can be rotated, translated, zoomed, etc. Such synchronized visualization of energies and structures allow the user to efficiently examine critical points and other features of the PES.

IV. SOFTWARE INFRASTRUCTURE

The described method naturally lends itself to several layers of interoperability and parallelization strategies, which are detailed below. The simplest invocation of TorsionDrive at the API layer takes in a list of dihedral angles to scan over, the granularity of each scan, and necessary information on the initial molecules (atomic symbols and Cartesian coordinates). At every step, TorsionDrive emits a series of dihedral angles to perform a constrained optimization as well as the starting geometry, which can be evaluated by many downstream programs. The next iteration is then started by supplying the TorsionDrive procedure with the Cartesian coordinates and final molecular energy of each constrained geometry optimization.

This API design abstracts away the details of which program is used to evaluate the constrained geometry optimization, allowing many different quantum chemistry, semi-empirical, force field, or machine learning (ML) potential programs to be used to generate

the necessary values. This strategy is robust to new methodology and program development and avoids being "locked into" a particular software package. Several examples of this are using Python-based geometry optimizers that are agnostic to the backend program to evaluate these constrained geometry optimizations such as geomeTRIC, PyBerny,[55] and PyOptKing.[56] In addition, Python-based suites of tools that attempt to abstract back ends exist, such as the Atomic Simulation Environment (ASE)[57] and QCEngine, allowing for many additional programs to be used with a simplified interface.

A. Task execution systems

On top of allowing flexibility in the evaluated program, this structure also provides integration with task execution system parallelization tools.[58] Task execution systems are typically software programs that can acquire computational resources on supercomputers through standard resource programs (e.g., SLURM[59]) and automate the computation of tasks (constrained geometry optimizations) on these resources. Tasks are typically computed via the following procedure:

1. A central task scheduler is created on a head node.
2. The central task scheduler acquires compute nodes through the local resource scheduler.
3. The compute nodes are harnessed by spawning a "worker" daemon process, which can communicate tasks to and from the scheduler via the local intranet.
4. Tasks are shipped from the central scheduler to a worker process, and the results of the task are shipped back to the central task scheduler.

Task execution systems allow the TorsionDrive calculation to be parallelized not only across cores of a single node but also across computational nodes even if the underlying quantum chemistry program is not able to do so. There are many such task execution systems available such as Work Queue,[35,36] Dask,[60] Parsl,[61] RADICAL Pilot,[62] and Fireworks[63] in the academic computing space, which provide this service and can be trivially integrated with TorsionDrive. At present, TorsionDrive supports Work Queue when running in the standalone operation mode and, through the QCArchive integration, supports several other task execution systems such as Dask, Parsl, RADICAL, and Fireworks. Figure 7(a) shows how TorsionDrive interacts with task execution systems when running in the standalone execution mode, i.e., outside of QCArchive.

B. QCArchive integration

The MolSSI QCArchive project is a platform for computing, storing, analyzing, and sharing quantum chemistry data. The QCArchive software infrastructure uses a client–server model; the server (QCFractal) is a permanent Python-based server, which stores quantum chemistry computations, runs "services" such as TorsionDrive to generate new quantum chemistry tasks, and provides an API to search and organize previous computations, and QCPortal is a Python-based API for interacting with the server suitable for Jupyter Notebooks.[53]

Fundamentally, QCFractal is a tool to compute a large number of quantum chemistry primitives such as an energy or gradient computation or procedures such as a geometry optimization with a variety of different community packages. Building on top of this core of primitives, it is easy to add workflows such as TorsionDrive to the software stack due to its API layers, which are agnostic to how the geometry optimization is evaluated. In addition, tools such as QCFractal allow many TorsionDrives to be evaluated concurrently on one or more physical resources such as a campus cluster or supercomputer to improve the possible parallelization of these computations further. A general workflow with QCFractal would be as follows:

1. A user submits one or more TorsionDrive computations to QCFractal from the QCPortal front-end client.

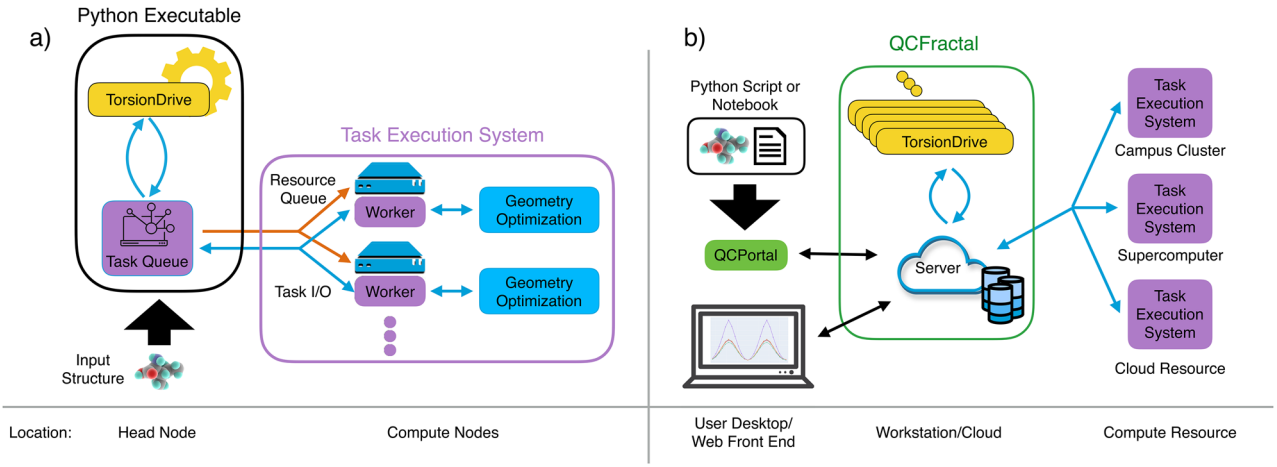

FIG. 7. Diagrams showing two different modes of TorsionDrive operation. (a) In the standalone execution mode, the TorsionDrive algorithm will generate new geometry optimizations that the task execution system ship to compute nodes and back to iterate over the procedure described in Sec. II. (b) Within the QCArchive ecosystem, the user can submit a new TorsionDrive via QCPortal to interact with a QCFractal server, which can run many concurrent TorsionDrives.

```python
import qcportal as ptl

client = ptl.FractalClient("https://myserver")

HOOH = ptl.models.Molecule.from_data("""
H 0 0 1.1
O 0 0 0
O 0 1.5 0
H 0 1.5 1.1
""")

torsiondrive_input = ptl.models.TorsionDriveInput(
    initial_molecule=[HOOH],
    keywords={"dihedrals": [[0, 1, 2, 3]], "grid_spacing": [18]},
    optimization_spec={"program": "geometric"},
    qc_spec={"program": "mopac", "method": "PM6", "driver": "gradient"},
)

identifier = client.add_service([torsiondrive_input])
```

```
<ComputeResponse(nsubmitted=1 nexisting=0)>
```

```python
torsiondrive_output = client.query_procedures(identifier.ids)
proc.get_final_energies(18)
```

```
-151.4746215927454
```

FIG. 8. An example usage of the QCArchive infrastructure stack where QCPortal is used to build a hydrogen peroxide molecule from an XYZ file. TorsionDrive is then submitted for the H–O–O–H dihedral angle use the geomeTRIC geometry optimizer and the PM6 level of theory using MOPAC. The computation is then retrieved from the server, and the energy at 18° is displayed.

2. The QCFractal server uses TorsionDrive to generate new geometry optimizations to be computed.
3. The geometry optimizations are computed on one or more physical resources where a physical resource can be a single core to a large supercomputer through a task execution system.
4. Items 2–3 are iterated until convergence.

Figure 7(b) shows how the user is able to use TorsionDrive as a service within the QCArchive infrastructure. An example usage of the QCPortal client is show in Fig. 8. The object returned in this image has API-based access to every geometry optimization and gradient evaluation.

V. CONCLUSIONS

The reformulation of torsion angle scanning as wavefront propagation comes with several important benefits that we think are worth the increased computational cost. These benefits include improved symmetry of the potential surface, which is related to the calculation results being independent of any chosen scan direction or dimensional ordering. The resulting potential energy surfaces have fewer discontinuities compared to sequential scanning, and in typical cases, lower-energy structures and potential minima can often be found. Multiple initial guesses can be naturally incorporated, allowing the calculations to finish in reduced wall time, given sufficient computing resources. This procedure also has a reduced tendency to get trapped in high-energy local minima, resulting in improved performance and reliability for challenging systems and generalized choices of dihedral angles.

In terms of software, TorsionDrive is a flexible package that can utilize either Python-based geometry optimization codes or quantum chemistry packages with integrated geometry optimization routines. It can run either in standalone mode and take advantage of parallel resources using the Work Queue task execution system or as a service in the QCArchive ecosystem that features centralized management of data and computer resources. Overall, we believe that this component-based approach to software development allows TorsionDrive to be a flexible piece of middleware that can be harnessed by a large number of external programs and incorporated into existing software ecosystems in a straightforward manner. This approach also helps users by enabling consistent approaches to geometry optimizations or higher level workflows when using different quantum chemistry software packages that may differ in terms of the methods that are implemented in each.

SUPPLEMENTARY MATERIAL

The supplementary material contains input files for TorsionDrive calculations including quantum chemistry inputs and output molecular structures and associated energies for Fig. 2–5. Structures are provided as Cartesian coordinates in XYZ format.

ACKNOWLEDGMENTS

Y.Q. was supported by a fellowship from the Open Force Field Consortium and ACS-PRF Award No. 58158-DNI6. D.G.A.S. was supported by U. S. National Science Foundation (NSF) Grant No. ACI-1547580 and from the Open Force Field Consortium. C.D.S. was supported by a fellowship from the Molecular Science Software Institute under NSF Grant No. ACI-1547580. M.F. acknowledges funding from the National Institute of General Medical Sciences of the National Institutes of Health under Award No. R01GM61300 to Michael K. Gilson. H.J. was supported by a fellowship from the Open Force Field Consortium. L.P.W. acknowledges funding from the National Institute of General Medical Sciences of the National Institutes of Health under Award No. R01GM132386. We acknowledge

John Chodera, John Stoppelman, Alberto Gobbi, Joshua Horton, and Xinjun Hou for helpful discussions. We also thank the Molecular Sciences Software Institute (MolSSI) for its support of the Open Force Field Consortium and Initiative. The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the NIH or the commercial partners of the Open Force Field Consortium.

The Open Force Field Consortium is composed of academic investigators from the Open Force Field Initiative and sponsoring Industry Partners collaborating to advance open force field science, toolkits, and standards for biomolecular drug discovery. The full list of funding partners is available online at https://openforcefield.org/consortium/.

DATA AVAILABILITY

The data that support the findings of this study are available within the article (and its supplementary material).

REFERENCES

[1]W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, "Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids," J. Am. Chem. Soc. **118**, 11225–11236 (1996).

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

[3]E. J. Sorin and V. S. Pande, "Exploring the helix-coil transition via all-atom equilibrium ensemble simulations," Biophys. J. **88**, 2472–2493 (2005).

[4]A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Laughton, and M. Orozco, "Refinement of the AMBER force field for nucleic acids: Improving the description of/conformers," Biophys. J. **92**, 3817–3829 (2007).

[5]K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell, Jr., "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields," J. Comput. Chem. **31**, 671–690 (2010).

[6]K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the Amber ff99SB protein force field," Proteins: Struct., Funct., Bioinf. **78**, 1950–1958 (2010).

[7]E. Harder, W. Damm, J. Maple, C. Wu, M. Reboul, J. Y. Xiang, L. Wang, D. Lupyan, M. K. Dahlgren, J. L. Knight, J. W. Kaus, D. S. Cerutti, G. Krilov, W. L. Jorgensen, R. Abel, and R. A. Friesner, "OPLS3: A force field providing broad coverage of drug-like small molecules and proteins," J. Chem. Theory Comput. **12**, 281–296 (2016).

[8]L.-P. Wang, K. A. McKiernan, J. Gomes, K. A. Beauchamp, T. Head-Gordon, J. E. Rice, W. C. Swope, T. J. Martínez, and V. S. Pande, "Building a more predictive protein force field: A systematic and reproducible route to AMBER-FB15," J. Phys. Chem. B **121**, 4023–4039 (2017).

[9]W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules," J. Am. Chem. Soc. **117**, 5179–5197 (1995).

[10]F. M. Bickelhaupt and E. J. Baerends, "The case for steric repulsion causing the staggered conformation of ethane," Angew. Chem., Int. Ed. **42**, 4183–4188 (2003).

[11]F. Weinhold, "Rebuttal to the Bickelhaupt–Baerends case for steric repulsion causing the staggered conformation of ethane," Angew. Chem., Int. Ed. **42**, 4188–4194 (2003).

[12]B. Mertz, M. Lu, M. Brown, and S. Feller, "Steric and electronic influences on the torsional energy landscape of retinal," Biophys. J. **101**, L17–L19 (2011).

[13]N. L. Allinger, "Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms," J. Am. Chem. Soc. **99**, 8127–8134 (1977).

[14]V. Tran, A. Buleon, A. Imberty, and S. Perez, "Relaxed potential energy surfaces of maltose," Biopolymers **28**, 679–690 (1989).

[15]W. L. Jorgensen and J. Tirado-Rives, "Molecular modeling of organic and biomolecular systems using BOSS and MCPRO," J. Comput. Chem. **26**, 1689–1700 (2005).

[16]H. Fujitani, A. Matsuura, S. Sakai, H. Sato, and Y. Tanida, "High-level *ab initio* calculations to improve protein backbone dihedral parameters," J. Chem. Theory Comput. **5**, 1155–1165 (2009).

[17]M. Buck, S. Bouguet-Bonnet, R. W. Pastor, and A. D. MacKerell, "Importance of the CMAP correction to the CHARMM22 protein force field: Dynamics of hen lysozyme," Biophys. J. **90**, L36–L38 (2006).

[18]Y. Shi, Z. Xia, J. Zhang, R. Best, C. Wu, J. W. Ponder, and P. Ren, "Polarizable atomic multipole-based AMOEBA force field for proteins," J. Chem. Theory Comput. **9**, 4046–4063 (2013).

[19]J. Foresman and A. Frisch, *Exploring Chemistry with Electronic Structure Methods* (Gaussian, Inc., Wallingford, CT, 2015).

[20]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 C.01 (Gaussian Inc, Wallingford CT, 2016).

[21]Y. Shao, Z. Gan, E. Epifanovsky, A. T. Gilbert, M. Wormit, J. Kussmann, A. W. Lange, A. Behn, J. Deng, X. Feng, D. Ghosh, M. Goldey, P. R. Horn, L. D. Jacobson, I. Kaliman, R. Z. Khaliullin, T. Kuś, A. Landau, J. Liu, E. I. Proynov, Y. M. Rhee, R. M. Richard, M. A. Rohrdanz, R. P. Steele, E. J. Sundstrom, H. L. Woodcock III, P. M. Zimmerman, D. Zuev, B. Albrecht, E. Alguire, B. Austin, G. J. O. Beran, Y. A. Bernard, E. Berquist, K. Brandhorst, K. B. Bravaya, S. T. Brown, D. Casanova, C.-M. Chang, Y. Chen, S. H. Chien, K. D. Closser, D. L. Crittenden, M. Diedenhofen, R. A. DiStasio, Jr., H. Do, A. D. Dutoi, R. G. Edgar, S. Fatehi, L. Fusti-Molnar, A. Ghysels, A. Golubeva-Zadorozhnaya, J. Gomes, M. W. Hanson-Heine, P. H. Harbach, A. W. Hauser, E. G. Hohenstein, Z. C. Holden, T.-C. Jagau, H. Ji, B. Kaduk, K. Khistyaev, J. Kim, J. Kim, R. A. King, P. Klunzinger, D. Kosenkov, T. Kowalczyk, C. M. Krauter, K. U. Lao, A. D. Laurent, K. V. Lawler, S. V. Levchenko, C. Y. Lin, F. Liu, E. Livshits, R. C. Lochan, A. Luenser, P. Manohar, S. F. Manzer, S.-P. Mao, N. Mardirossian, A. V. Marenich, S. A. Maurer, N. J. Mayhall, E. Neuscamman, C. M. Oana, R. Olivares-Amaya, D. P. O'Neill, J. A. Parkhill, T. M. Perrine, R. Peverati, A. Prociuk, D. R. Rehn, E. Rosta, N. J. Russ, S. M. Sharada, S. Sharma, D. W. Small, A. Sodt, T. Stein, D. Stück, Y.-C. Su, A. J. Thom, T. Tsuchimochi, V. Vanovschi, L. Vogt, O. Vydrov, T. Wang, M. A. Watson, J. Wenzel, A. White, C. F. Williams, J. Yang, S. Yeganeh, S. R. Yost, Z.-Q. You, I. Y. Zhang, X. Zhang, Y. Zhao, B. R. Brooks, G. K. Chan, D. M. Chipman, C. J. Cramer, W. A. Goddard III, M. S. Gordon, W. J. Hehre, A. Klamt, H. F. Schaefer III, M. W. Schmidt, C. D. Sherrill, D. G. Truhlar, A. Warshel, X. Xu, A. Aspuru-Guzik, R. Baer, A. T. Bell, N. A. Besley, J.-D. Chai, A. Dreuw, B. D. Dunietz, T. R. Furlani, S. R. Gwaltney, C.-P. Hsu, Y. Jung, J. Kong, D. S. Lambrecht, W. Liang, C. Ochsenfeld, V. A. Rassolov, L. V. Slipchenko, J. E. Subotnik, T. V. Voorhis, J. M. Herbert, A. I. Krylov, P. M. Gill, and M. Head-Gordon, "Advances in molecular quantum chemistry contained in the Q-Chem 4 program package," Mol. Phys. **113**, 184–215 (2015).

[22]R. M. Parrish, L. A. Burns, D. G. A. Smith, A. C. Simmonett, A. E. DePrince, E. G. Hohenstein, U. Bozkaya, A. Y. Sokolov, R. Di Remigio, R. M. Richard, J. F. Gonthier, A. M. James, H. R. McAlexander, A. Kumar, M. Saitow, X. Wang, B. P. Pritchard, P. Verma, H. F. Schaefer, K. Patkowski, R. A. King, E. F. Valeev,

F. A. Evangelista, J. M. Turney, T. D. Crawford, and C. D. Sherrill, "Psi4 1.1: An open-source electronic structure program emphasizing automation, advanced libraries, and interoperability," J. Chem. Theory Comput. **13**, 3185–3197 (2017).

[23]D. Smith, L. Burns, A. Simmonett, R. Parrish, M. Schieber, R. Galvelis, P. Kraus, H. Kruse, R. Di Remigio, A. Alenaizan, A. James, S. Lehtola, J. Misiewicz, M. Scheurer, R. Shaw, J. Schriber, Y. Xie, Z. Glick, D. Sirianni, J. O'Brien, J. Waldrop, A. Kumar, E. G. Hohenstein, B. Pritchard, B. Brooks, H. Schaefer, A. Sokolov, K. Patkowski, E. DePrince, U. Bozkaya, R. King, F. Evangelista, J. Turney, T. Crawford, and D. Sherrill, Psi4 1.4: Open-Source Software for High-Throughput Quantum Chemistry, 2020, Publisher: ChemRxiv.

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

[25]See http://www.turbomole.com for TURBOMOLE V7.0 2015, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007.

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

[27]A. V. Titov, I. S. Ufimtsev, N. Luehr, and T. J. Martinez, "Generating efficient quantum chemistry codes for novel architectures," J. Chem. Theory Comput. **9**, 213–221 (2013).

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

[29]J. Cioslowski, A. P. Scott, and L. Radom, "Catastrophes, bifurcations and hysteretic loops in torsional potentials of internal rotations in molecules," Mol. Phys. **91**, 413–420 (1997).

[30]P. C. D. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson, and M. T. Stahl, "Conformer generation with OMEGA: Algorithm and validation using high quality structures from the protein databank and cambridge structural database," J. Chem. Inf. Model. **50**, 572–584 (2010).

[31]N.-O. Friedrich, C. de Bruyn Kops, F. Flachsenberg, K. Sommer, M. Rarey, and J. Kirchmair, "Benchmarking commercial conformer ensemble generators," J. Chem. Inf. Model. **57**, 2719–2728 (2017).

[32]Y. Qiu, L.-P. Wang, D. G. A. Smith, J. Horton, H. Jang, and M. Feng, lpwgroup/torsiondrive: Release 1.0.0, 2020, https://zenodo.org/record/3686014#.Xn-k3NNKhSM.

[33]D. G. A. Smith, L. A. Burns, L. Naden, and M. Welborn, See https://qcarchive.molssi.org for QCArchive: A central source to compile, aggregate, query, and share quantum chemistry data; accessed January 2020.

[34]J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost," Chem. Sci. **8**, 3192–3203 (2017).

[35]M. Albrecht, D. Rajan, and D. Thain, "Making work queue cluster-friendly for data intensive scientific applications," *2013 IEEE International Conference on Cluster Computing (CLUSTER)* (IEEE, 2013), pp. 1–8, ISSN: 1552-5244, 2168-9253.

[36]N. Kremer-Herman, B. Tovar, and D. Thain, "A lightweight model for right-sizing master-worker applications," *SC18: International Conference for High Performance Computing, Networking, Storage and Analysis* (IEEE, 2018), pp. 504–516.

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

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

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

[40]D. G. A. Smith, L. A. Burns, K. Patkowski, and C. D. Sherrill, "Revised damping parameters for the D3 dispersion correction to density functional theory," J. Phys. Chem. Lett. **7**, 2197–2203 (2016).

[41]A. I. Krylov, C. D. Sherrill, E. F. C. Byrd, and M. Head-Gordon, "Size-consistent wave functions for nondynamical correlation energy: The valence active space optimized orbital coupled-cluster doubles model," J. Chem. Phys. **109**, 10669–10678 (1998).

[42]H. Jang (2020). "Update on Parsley minor releases (openff-1.1.0, 1.2.0)," Zenodo. https://doi.org/10.5281/zenodo.3781313.

[43]J. Wagner and H. Jang (2020). "Openforcefield/openforcefields: Version 1.1.1 "Parsley"," Zenodo. https://doi.org/10.5281/zenodo.3751818.

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

[45]P. I. Nagy, W. J. Dunn, G. Alagona, and C. Ghio, "Theoretical calculations on 1,2-ethanediol. Gauche-trans equilibrium in gas-phase and aqueous solution," J. Am. Chem. Soc. **113**, 6719–6729 (1991).

[46]A. Lonardi, P. Oborský, and P. H. Hünenberger, "Solvent-modulated influence of intramolecular hydrogen-bonding on the conformational properties of the hydroxymethyl group in glucose and galactose: A molecular dynamics simulation study," Helv. Chim. Acta **100**, e1600158 (2017).

[47]R. S. Paton and J. M. Goodman, "Hydrogen bonding and -stacking: How reliable are force fields? A critical evaluation of force field descriptions of nonbonded interactions," J. Chem. Inf. Model. **49**, 944–955 (2009).

[48]C. Liu, J.-P. Piquemal, and P. Ren, "Implementation of geometry-dependent charge flux into the polarizable AMOEBA+ potential," J. Phys. Chem. Lett. **11**, 419–426 (2020).

[49]C. Bayly, see https://docs.eyesopen.com/applications/quacpac/theory/molcharge_theory.html for MolCharge Theory—Applications, v2019.Nov.2.

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

[51]C. Tian, K. Kasavajhala, K. A. A. Belfon, L. Raguette, H. Huang, A. N. Migues, J. Bickel, Y. Wang, J. Pincay, Q. Wu, and C. Simmerling, "ff19SB: Amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution," J. Chem. Theory Comput. **16**, 528–552 (2020).

[52]M. Klok, N. Boyle, M. T. Pryce, A. Meetsma, W. R. Browne, and B. L. Feringa, "MHz unidirectional rotation of molecular rotary motors," J. Am. Chem. Soc. **130**, 10484–10485 (2008).

[53]T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, and C. Willing, *Jupyter Notebooks*, A Publishing Format For Reproducible Computational Workflows (Positioning and Power in Academic Publishing: Players, Agents and Agendas, 2016), pp. 87–90.

[54]H. Nguyen, D. A. Case, and A. S. Rose, "NGLview–interactive molecular graphics for Jupyter notebooks," Bioinformatics **34**, 1241–1242 (2017).

[55]J. Hermann, see https://github.com/jhrmnn/pyberny for berny: Molecular structure optimizer. For the current version; accessed January 2020.

[56]See https://github.com/psi-rking/optking for OptKing: optking: A Python version of the PSI4 geometry optimization program. For the current version; accessed January 2020.

[57]A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułstrokak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, "The atomic simulation environment—A Python library for working with atoms," J. Phys.: Condens. Matter **29**, 273002 (2017).

[58]M. Turilli, M. Santcroos, and S. Jha, "A comprehensive perspective on pilot-job systems," ACM Comput. Surv. **51**, 23 (2018).

[59]A. B. Yoo, M. A. Jette, and M. Grondona, "SLURM: Simple linux utility for resource Management," *Job Scheduling Strategies for Parallel Processing* (Springer-Verlag, Berlin, Heidelberg, 2003), pp. 44–60.

[60]M. Rocklin, "Dask: Parallel computation with blocked algorithms and task scheduling," in *Proceedings of the 14th Python in Science Conference* (SciPy, 2015), pp. 130–136.

[61]Y. Babuji, A. Woodard, Z. Li, D. S. Katz, B. Clifford, R. Kumar, L. Lacinski, R. Chard, J. Wozniak, I. Foster, M. Wilde, and K. Chard, "Parsl: Pervasive parallel programming in Python," in 28th ACM International Symposium on High-Performance Parallel and Distributed Computing (HPDC), 2019.

[62]M. Turilli, A. Merzky, V. Balasubramanian, S. Jha, "Building blocks for workflow system middleware," in *Proceedings of the 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing* (IEEE, 2018), p. 348–349.

[63]A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G.-M. Rignanese, G. Hautier, D. Gunter, and K. A. Persson, "FireWorks: A dynamic workflow system designed for high-throughput applications," Concurrency Comput.: Pract. Exper. **27**, 5037–5059 (2015), CPE-14-0307.R2.