



OPEN

DATA DESCRIPTOR

Reaction profiles for quantum chemistry-computed [3 + 2] cycloaddition reactions

Thijs Stuyver¹, Kjell Jorner^{2,3,4} & Connor W. Coley^{1,5} ✉

Bio-orthogonal click chemistry based on [3 + 2] dipolar cycloadditions has had a profound impact on the field of biochemistry and significant effort has been devoted to identify promising new candidate reactions for this purpose. To gauge whether a prospective reaction could be a suitable bio-orthogonal click reaction, information about both on- and off-target activation and reaction energies is highly valuable. Here, we use an automated workflow, based on the autodE program, to compute over 5000 reaction profiles for [3 + 2] cycloadditions involving both synthetic dipolarophiles and a set of biologically-inspired structural motifs. Based on a succinct benchmarking study, the B3LYP-D3(BJ)/def2-TZVP//B3LYP-D3(BJ)/def2-SVP level of theory was selected for the DFT calculations, and standard conditions and an (aqueous) SMD model were imposed to mimic physiological conditions. We believe that this data, as well as the presented workflow for high-throughput reaction profile computation, will be useful to screen for new bio-orthogonal reactions, as well as for the development of novel machine learning models for the prediction of chemical reactivity more broadly.

Background & Summary

Bio-orthogonal click reactions provide a straightforward method to tag and image certain biomolecules, enabling the study of their function *in vivo*^{1–3}. Besides realtime imaging, bio-orthogonal click reactions have also been considered for advanced applications in medicine/healthcare, e.g., targeted cancer treatments and *in situ* drug assembly^{4,5}. New candidate reactions—especially mutually orthogonal ones—are highly sought after^{6–8}.

A wide variety of criteria need to be fulfilled for a reaction to be considered suitable for bio-orthogonal click chemistry: (a) the reactants should be stable and the reaction selective, i.e., reactant degradation due to biological conditions (pH, solvent, etc.) should be limited and side reactions with biomolecules should not occur at appreciable rates; (b) the selective reaction should be sufficiently fast at room/body temperature so that ligation can take place before clearance of the reactants from the studied organism; and (c) it should be possible to incorporate the reactants into biomolecules via some form of metabolic or protein engineering. Ideally the compounds should also be small enough to not disturb the native behavior of the biological system under investigation⁹. As such, data regarding thermodynamic stability as well as on- and off-target kinetics could facilitate a principled screening campaign to discover novel candidate bio-orthogonal click reactions.

Here, we present a computational dataset of hypothetical click reactions evaluated for their potential bio-orthogonality constructed via a high-throughput reaction profile evaluation. Due to the enormous success—and computationally accessible mechanism—of [3 + 2] dipolar cycloaddition reactions in bio-orthogonal chemistry so far⁹, we decided to focus on this reaction class. A diverse chemical space spanning close to 5 M synthetic dipole-dipolarophile pairs was defined (Fig. 1a) and the reaction profiles, i.e., activation and reaction energies, were evaluated for a representative subset of over 3000 individual reactions. We also probe the inherent reactivity of various dipoles towards a variety of “native” dipolarophiles, i.e., unsaturated motifs within biomolecules

¹Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts, 02139, USA. ²Department of Computer Science, University of Toronto, 40 St George St, Toronto, Ontario, M5S 2E4, Canada. ³Department of Chemistry, Chemical Physics Theory Group, 80 St. George St., University of Toronto, Ontario, M5S 3H6, Canada. ⁴Department of Chemistry and Chemical Engineering, Chalmers University of Technology, Kemigården 4, SE-41258, Gothenburg, Sweden. ⁵Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts, 02139, USA. ✉e-mail: ccoley@mit.edu

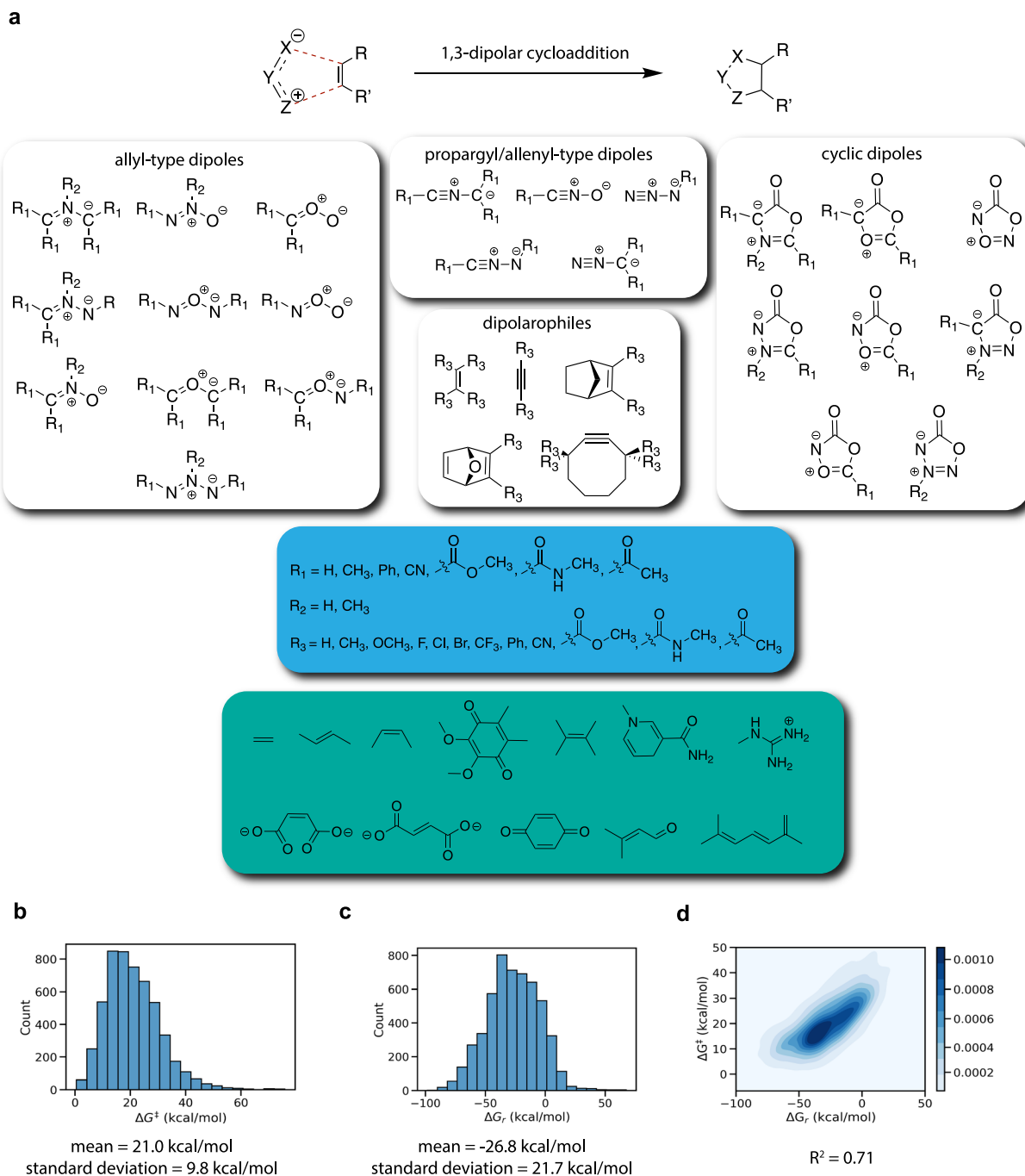


Fig. 1 (a) Schematic overview of the search space for the considered dipoles and dipolarophiles. In the green box, the biologically relevant motifs considered in this study are presented. These motifs are inspired respectively by fatty acids and prostaglandins²⁶, fumaric and maleic acid²⁷, ubiquinone or co-enzyme Q (a common coenzyme family that is ubiquitous in animals and most bacteria)²⁸, terpineols²⁹, NADH/NADPH (an important and common co-enzyme playing a pivotal role in the metabolism of most species)³⁰, retinal (a chromophore central to visual perception)³¹, β -carotene³², and the arginine amino-acid and argininosuccinic acid (an important intermediate in the urea cycle)³³. (b) Histogram representing the distribution of the computed activation energies (ΔG^\ddagger). (c) Histogram representing the distribution of the computed reaction energies (ΔG_r). (d) Correlation plot between activation and reaction energies.

that are present in varying concentrations in living organisms. To this end, more than 2500 off-target reactions involving 12 distinct “biofragments” (green box in Fig. 1a), selected from the literature, were computed as well.

A limited benchmarking, with respect to both computational reference data as well as experimental data, points to B3LYP-D3(BJ)/def2-TZVP//B3LYP-D3(BJ)/def2-SVP^{10–15} as the level of theory striking the optimal balance between performance and cost (*vide infra*). To mimic biological environments, water was taken as the solvent, which was modelled through the SMD polarizable continuum model, and standard temperature and

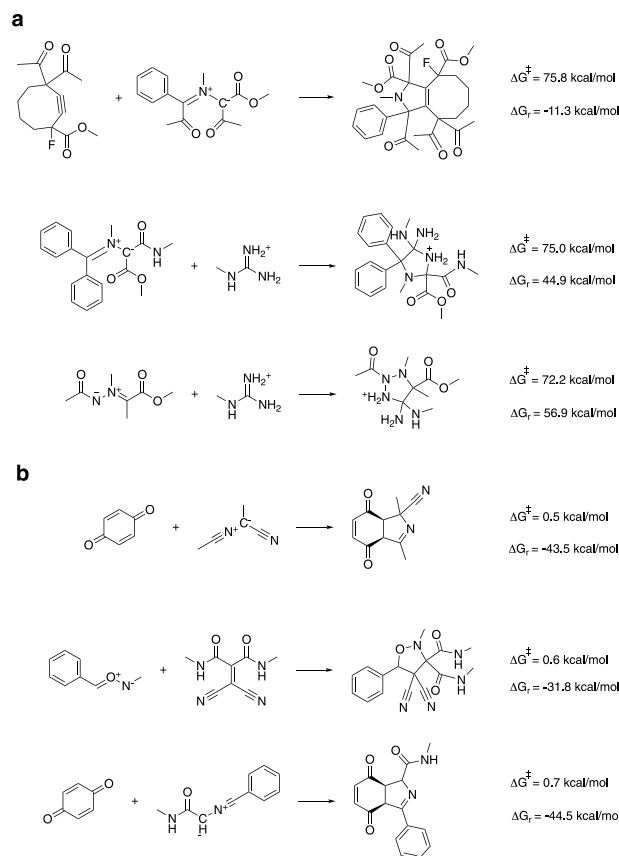


Fig. 2 (a) The three reactions with the highest computed barriers. (b) The three reactions with the lowest computed barriers.

pressure were assumed¹⁶. Histograms, depicting the distribution of the activation energies (ΔG^\ddagger) and reaction energies (ΔG_r) obtained with these settings for the 5269 successfully computed reaction profiles (out of a total of 5974 considered reaction SMILES), are provided in Fig. 1b–d. A significant spread in both the activation energies and the reaction energies is clearly obtained (the standard deviations amount to 9.8 and 21.7 kcal/mol respectively), indicating that chemically diverse reacting systems were sampled. In Fig. 2, the reactions with respectively the three highest and lowest activation energies found are presented. Even though the dataset contains some diversity in dipole/dipolarophile scaffolds, reaction and activation energies correlate quite well ($R^2 = 0.71$; see Fig. 1d) in accordance with the Bell-Evans-Polanyi principle¹⁷.

The produced dataset should prove useful for the construction of (surrogate) machine learning (ML) models for the discovery of promising dipole-dipolarophile combinations within the defined search space. Furthermore, the provided data will be useful for the development of methodologies for reactivity prediction in general, particularly due to the inclusion of high-quality 3D geometries for each species (see below).

Methods

To construct the dataset of reaction profiles for [3 + 2] cycloaddition reactions, an automated workflow based on the Python package autodE¹⁸, recently developed by the Duarte group, has been set up. autodE is a package that can be coupled to various popular electronic structure programs to automate the otherwise laborious task of computing full reaction profiles; here, we opted for Gaussian16¹⁹.

Concise overview of the autodE methodology. The starting point of an autodE computation can be a simple reaction SMILES string, together with a specification of the solvent, temperature and level of theory. In a first step, the provided SMILES string is converted into 3D structures for reactants and products with the help of RDKit²⁰. Subsequently, a conformer search is performed for each of the generated species with the help of the ETKDGV3 algorithm²¹. Each unique conformer is then optimized at the GFN2-xTB level of theory²². Next, the conformers are ranked based on their relative energy. If the `ade.Config.hmethod_conformers` keyword is specified, the final GFN2-xTB energies are used for this ranking, otherwise single-point low-level DFT—here B3LYP-D3(BJ)/def2-SVP—refinements are used (*vide infra*). The lowest energy conformer is then further optimized with the same low-level DFT functional and basis set combination. Subsequently, graphs are constructed for both reactants and products based on bond length criteria, and the bond rearrangements throughout the reaction are identified. A guess for the transition state (TS) connecting reactants and products is then generated based on the adaptive path method, after which the TS is optimized (with low-level DFT). In the next step, a conformer search is performed on the TS with the help of the randomize and relax (RR) algorithm, after which the lowest

energy conformer is selected, optimized with the low-level DFT method, and the validity of the TS is confirmed through a detailed analysis of the imaginary mode^{18,23}.

Once validated structures for reactants, products and TSs have been generated, the energy of each species is refined at a higher DFT level—here B3LYP-D3(BJ)/def2-TZVP—and frequency calculations are performed for the reactants and products at the lower DFT level.

Definition of the chemical space of dipoles and dipolarophiles. First, a combinatorial dataset of 1,3-dipoles and dipolarophiles was generated. For the dipoles, 15 allyl and propargyl-allenyl scaffolds²⁴, decorated with a variety of substituents (Fig. 1a), were selected. Additionally, (decorated) heterocyclic 5-membered rings based on 1,3-dipoles, e.g. sydnone- and münchnone-derivatives, were also included²⁴. Since dipoles used in bio-orthogonal click applications usually require an attachment site that can be linked to either the target biomolecule or the (fluorescent) probe, only substituent combinations containing at least one substituent ending with an (extendable) methyl or phenyl group were considered.

For the dipolarophiles, 3 strained scaffolds (cyclooctyne, norbornene and norbornadiene)⁹ as well as regular ethylene and acetylene scaffolds, each with a variety of substituents, were selected (see Fig. 1a). For the ethylene and acetylene scaffold, the same constraint on substituent combination diversity as for the dipoles was introduced (the remaining dipolarophile scaffolds have several alternative attachment points which can be connected to a target molecule/fluorescent probe). Additional constraints were placed on the substituent combinations for the ethylene and cyclooctyne scaffold to reign in the combinatorial explosion of possibilities. For ethylene, the total number of distinct substituent types was fixed to two at most; for cyclooctyne, the substituents on at least one of the functionalized carbon sites was forced to be the same.

In total, 3555 dipoles and 1339 dipolarophiles were generated in this manner with the help of RDKit²⁰, which results in a combinatorial chemical space of almost 4.8 M reacting systems and—taking regio-isomerism into account—9.6 M+ unique reactions. It should be noted that not every dipole and dipolarophile in the search space can be expected to be stable. However, the most extreme cases of instability, namely, those species that either do not exist at a local minimum in their respective potential energy surface (PES), or participate in dipole-dipolarophile combinations that give rise to no reaction barrier, will inherently be filtered out downstream in our workflow upon sampling during the electronic structure calculations and the associated isomorphism tests of the molecular graphs (*vide infra*)¹⁸.

From this vast chemical space, a computationally tractable subset was extracted. To ensure representation of each individual scaffold type, both dipoles and dipolarophiles were subdivided into individual buckets and random sub-samples were taken (with replacement) from each. For the dipoles, 1000 species were selected from the allyl-based scaffolds, 200 from the propargyl-allenyl scaffolds, and 300 from the heterocyclic 5-membered rings. For both the ethylene and acetylene dipolarophile scaffolds, 200 species were sampled, whereas 300 were sampled for both the norbornene and oxo-norbornadiene scaffolds. For the cyclooctyne scaffold, 500 species were selected.

Subsequently, the sampled list of dipolarophiles was shuffled, after which the entries of the dipole and dipolarophile lists were matched based on index values. The reactant systems generated in this manner were subsequently passed to the RunReactants function from RDChiral together with a set of SMARTS templates to generate full, atom-mapped reaction SMILES (*vide infra*)²⁵. The SMARTS templates used are shown in Table 1.

To probe potential off-target reactivity, 12 unsaturated biologically relevant motifs that are present/are produced as metabolites *in vivo* were selected^{26–33}. To construct a dataset of reactions involving these biofragments, a new sample of 1500 dipoles was taken in the same manner as before, as well as a sample (with replacement) of 1500 biofragments, see the green box in Fig. 1a for an overview of the motifs considered. Subsequently, the sampled biofragment list was again shuffled and the entries of the two lists were matched based on index value.

Solvent correction and thermal effects. To mimic biological environments, we model the reaction systems with water as a solvent through the SMD polarizable continuum model¹⁶. A recent computational study by Yang and co-workers investigated the role of explicit water molecules on 1,3-cycloaddition reactions using DFT-quality neural network potentials and compared to experimental data. They concluded that even in the most extreme cases considered, aqueous cycloaddition reactions tended to retain a (quasi-)concerted mechanism and most *qualitative* solvation trends could already be successfully recovered through implicit modelling³⁴. Our own benchmarking results also corroborate the adequacy of implicit solvent modelling to recover experimental trends (Fig. 6).

The choice for water as solvent has been informed by its ubiquity in biological settings, though this may not always be an appropriate approximation for the actual environment in which these click reactions will take place (e.g., in the vicinity of cell membranes or in protein condensates). It is, however, well-established that the rates of 1,3-cycloaddition reactions are fairly insensitive to solvent polarity, so that deviations from the idealized solvent environment is not an issue of major concern³⁵.

To retain the generality of our approach, we selected a standard state for the computation of thermal corrections as 1 mol/L. The temperature of each reaction was set to room temperature (298.15 K). It should be evident that in practice, significant deviations from these conditions can be expected, e.g., the body temperatures of most mammals hover around 310 K, which will make the reactions slightly more facile³⁶. Nevertheless, the presented values under standard conditions are suitable reference values that can be transferred to the specific application under consideration.

Selection of DFT functional and basis set and validation of workflow. The DFT level of theory was selected based on a (limited) benchmarking study. In total, 4 different functionals and dispersion correction combinations were considered (B3LYP-D3(BJ)^{10–13}, PBE0-D3(BJ)³⁷, M06-2X³⁸ and ω B97X-D^{39,40}, in combination

reactant system	SMARTS strings
propargyl-allenyl dipole scaffolds	[*:1][*+2][*:-3].[C,N:4]=[C:5]>[*:1]1=[*+0:2][*:-0:3][C,N:4][C:5]1
adding to double bonds	([*:1][*+2][*:-3].[C,N:4]=[C:5]>[*:1]1=[*+0:2][*:-0:3][C:5][C,N:4]1
propargyl-allenyl dipole scaffolds	([*:1][*+2][*:-3].[C:4][C:5]>[*:1]1=[*+0:2][*:-0:3][C:4]=[C:5]1
adding to triple bonds	([*:1][*+2][*:-3].[C:4][C:5]>[*:1]1=[*+0:2][*:-0:3][C:5]=[C:4]1
allyl dipole scaffolds	([*:1]=[*+2][*:-3].[C,N:4]=[C:5]>[*:1]1[*+0:2][*:-0:3][C,N:4][C:5]1
adding to double bonds	([*:1]=[*+2][*:-3].[C,N:4]=[C:5]>[*:1]1[*+0:2][*:-0:3][C:5][C,N:4]1
allyl dipole scaffolds	([*:1]=[*+2][*:-3].[C:4][C:5]>[*:1]1[*+0:2][*:-0:3][C:4]=[C:5]1
adding to triple bonds	([*:1]=[*+2][*:-3].[C:4][C:5]>[*:1]1[*+0:2][*:-0:3][C:5]=[C:4]1
cyclic dipole scaffolds	([*:1][*+2][*:-3].[C,N:4]=[C:5]>[*:1]1[*+0:2][*:-0:3][C,N:4][C:5]1
adding to double bonds	([*:1][*+2][*:-3].[C,N:4]=[C:5]>[*:1]1[*+0:2][*:-0:3][C:5][C,N:4]1
cyclic dipole scaffolds	([*:1][*+2][*:-3].[C:4][C:5]>[*:1]1[*+0:2][*:-0:3][C:4]=[C:5]1
adding to triple bonds	([*:1][*+2][*:-3].[C:4][C:5]>[*:1]1[*+0:2][*:-0:3][C:5]=[C:4]1

Table 1. Overview of the SMARTS templates used.

with the def2-SVP basis set¹⁴ for optimizations and frequency/thermal correction calculations, and def2-TZVP¹⁵ for single-point calculations.

First, we selected the nine 1,3-dipolar cycloaddition reactions within the (revised)⁴¹ BHPERI dataset (see Fig. 3)^{42,43} and compared the electronic energies to the Wn-F12 values from the literature⁴⁴. All of the level of theories tested resulted in almost perfect correlations ($R^2 \geq 0.99$), but B3LYP-D3(BJ) clearly outperformed the other functionals in terms of mean-absolute-error (MAE; 1.1 kcal/mol vs > 1.8 kcal/mol) and root-mean-square-error (RMSE; 1.5 kcal/mol vs > 2.0 kcal/mol), see Fig. 4. It should be noted that our findings here are also in line with the results presented in the very recent benchmarking study on pericyclic reactions by Bickelhaupt and co-workers: across the broader set of pericyclic reactions investigated, M06-2X was found to outperform B3LYP-D3(BJ), but specifically for the 1,3-dipolar cycloaddition barrier probed, B3LYP-D3(BJ) reproduced the CCSDT(Q)/CBS barrier within 1 kcal/mol, outperforming each of the other functionals by 1–3 kcal/mol⁴⁵.

Next, we curated an experimental benchmarking dataset based on rate constants found in a series of papers authored by Huisgen and co-workers on the kinetics (and mechanism) of 1,3-dipolar reactions^{35,46–48}. We transformed 11 reported experimental rate constants to Gibbs free energies of activation with the help of the Eyring equation assuming no recrossing (Fig. 5) and compared those to the corresponding values computed with our workflow (Fig. 6)⁴⁹. Correlations between experimental and computed values were still excellent ($R^2 > 0.92$), but the best result was now obtained for ω B97X-D ($R^2 = 0.94$). However, M06-2X and ω B97X-D now yielded significant systematic errors (MAEs of 4.1–4.2 kcal/mol), whereas the experimental and computational barriers agreed much better for B3LYP-D3(BJ) and PBE0-D3(BJ) (MAEs of 2.7–2.8 kcal/mol). Putting all of this together, we decided to settle on the B3LYP-D3(BJ)/def2-TZVP//B3LYP-D3(BJ)/def2-SVP level of theory.

Generation of reaction profiles with autodE. For every reaction SMILES generated with RDChiral, an autodE¹⁸ workflow to compute the associated reaction profile with the help of Gaussian16¹⁹ was executed in a fully automated, high-throughput manner. In first instance, the default settings of autodE were retained, and only the functionals and/or dispersion corrections were adjusted based on the benchmarking results (*vide supra*). More information about this workflow can be found in the autodE documentation, and will not be discussed here at length.

We directly select the lowest energy conformer for reactant and product from GFN2-xTB²² optimization results for the generated conformer set, rather than from (single-point) DFT computations (by setting the `ade.Config.hmethod_conformers` keyword to False). While this approximation/simplification may be problematic when dealing with highly flexible compounds, it has a fairly limited effect on the results for our reactions, which involve mostly small and rigid compounds (cf. the excellent accuracy obtained during benchmarking and the tests performed on the azide reaction set; *vide infra*). At the same time, this simplification resulted in a significant speed up of the reaction profile generation.

We also checked whether reactant and product complexes should be considered in the computed reaction profiles. By default, autodE does not compute complexes when thermal corrections (Gibbs free energies) are requested for the species along the profile, primarily because loose complexes with extremely low frequencies result in a high uncertainty for these quantities. To justify this choice, we compute reaction profiles for the (gas-phase) BHPERI dataset⁴¹ with both complexes and thermal corrections at room temperature. For each of the reactions for which the profile calculation terminated successfully, shallow minima for the complexes were found on the electronic energy surface, but thermal corrections erased these minima completely; the Gibbs free energy of the complexes consistently lay several kcal/mol above the isolated reactants (Table 2). Since electrostatic attraction—which constitutes the bulk of the complexation interactions—tends to diminish sharply when going from the gas-phase to a (polar) solvent, this result suggests that complex formation will not impact the reactivity for our cycloaddition reactions in a significant manner and can thus be safely neglected.

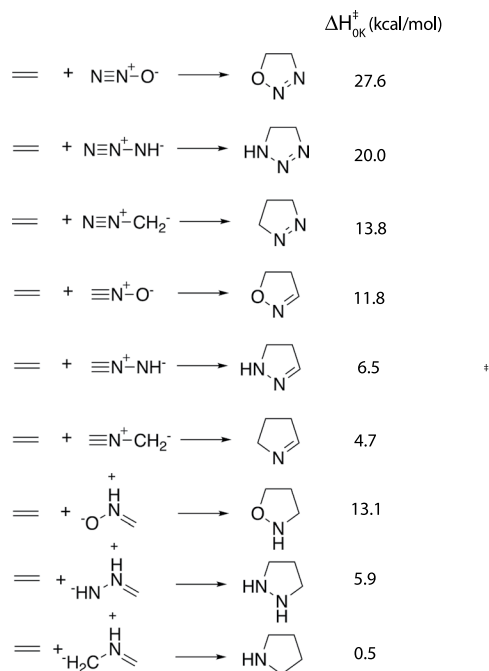


Fig. 3 Reference reaction data extracted from the (revised) BHPERI dataset⁴¹.

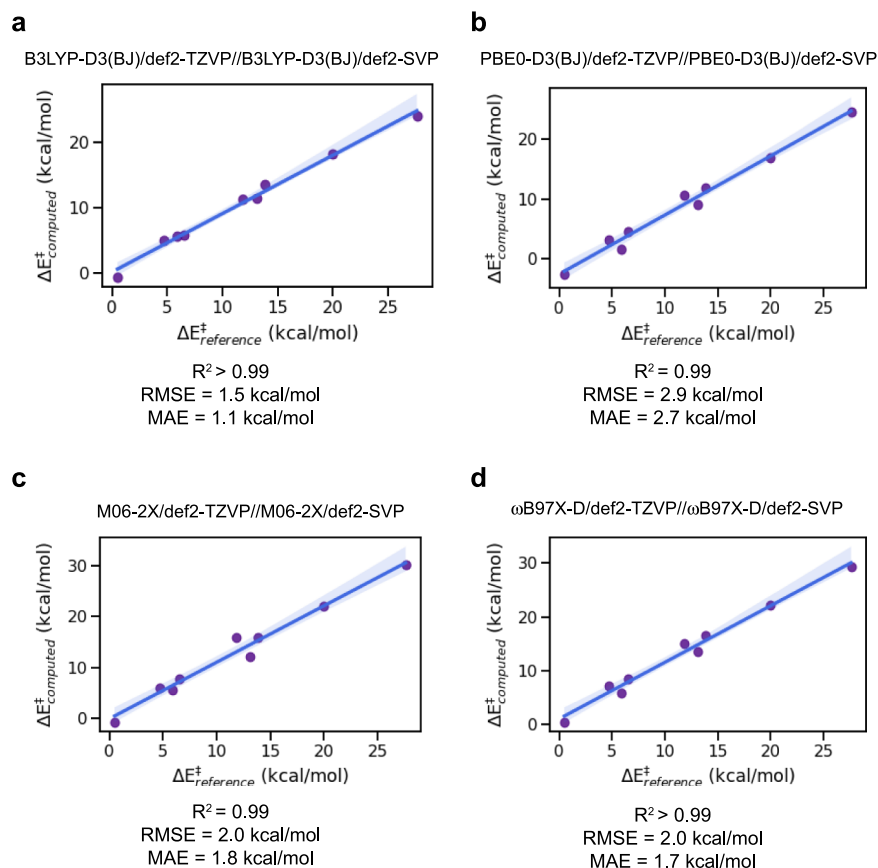


Fig. 4 Benchmarking functional + dispersion correction combinations against the (revised) BHPERI dataset⁴¹. The translucent bands around the regression line correspond to the 95% confidence interval determined through bootstrap resampling.

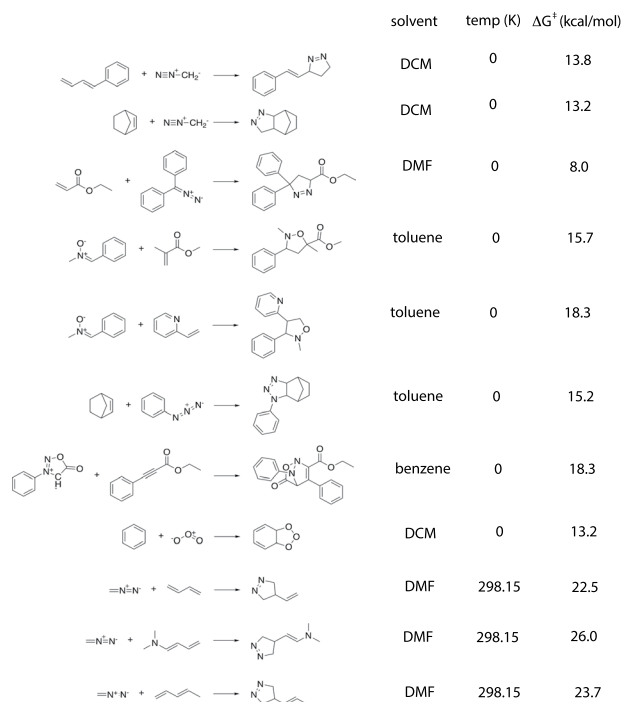


Fig. 5 Reference reaction data extracted from the literature^{35,46–48}. Temperatures of 0K indicate that the originally reported values are activation enthalpies ΔH_{0K}^\ddagger (obtained through extrapolation from measurements at different temperatures).

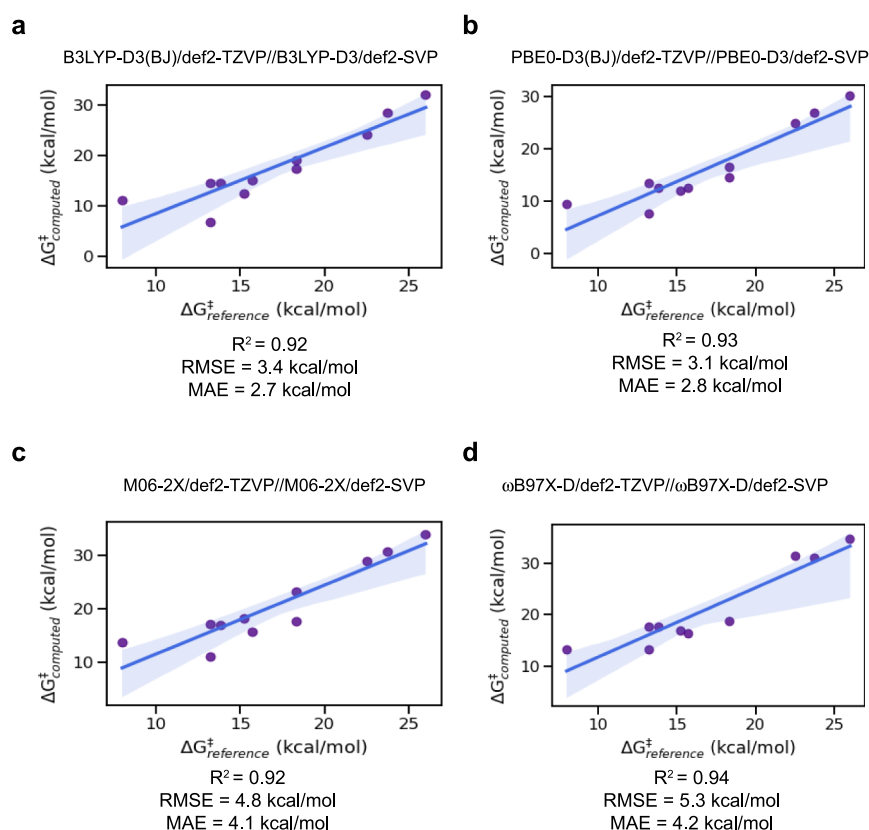


Fig. 6 Benchmarking functional + dispersion correction combinations against the experimental dataset extracted from the literature. The translucent bands around the regression line correspond to the 95% confidence interval determined through bootstrap resampling.

reaction SMILES	$\Delta E_{\text{complex}}$ (kcal/mol)	$\Delta G_{\text{complex}}$ (kcal/mol)
<chem>C=C.N#[N+][O-]>>C1N=NOC1</chem>	-3.2	1.9
<chem>C=C.N#[N+].[NH-]>>C1N=NNC1</chem>	-3.6	3.3
<chem>C=C.N#[N+].[CH2-]>>C1N=NCC1</chem>	-2.0	3.0
<chem>C=C.C#[N+].[O-]>>C1C=NOC1</chem>	-2.9	2.6
<chem>C=C.C#[N+].[NH-]>>C1C=NNC1</chem>	-3.2	4.0
<chem>C=C.C#[N+].[CH2-]>>C1C=NCC1</chem>	-2.1	4.6
<chem>C=C.C=[NH+][O-]>>C1CNOC1</chem>	-4.5	3.4
<chem>C=C.C=[NH+].[NH-]>>C1CNNC1</chem>	/	/
<chem>C=C.C=[NH+].[CH2-]>>C1CCNC1</chem>	/	/

Table 2. Complexation energies for the cycloaddition reactions of the BHPERI dataset⁴¹, where $\Delta E_{\text{complex}}$ stands for the electronic energy (+zero-point correction), and $\Delta G_{\text{complex}}$ stands for the Gibbs Free energy (at standard conditions, i.e., 298.15 K and 1 M).

Ensuring stereochemical consistency between reactants, transition states, and products. By default, autodE¹⁸ does not exchange stereochemical information between reactants and products in a reaction SMILES; it simply searches for the conformations of each species which have the lowest energy globally in an independent manner. The transition state (TS) therefore exclusively inherits the stereochemistry from one side. In the case of addition reactions, the product side stereochemistry is selected by default. Consequently, stereochemical compatibility between reactants, products and TSs is not guaranteed by default. As part of our workflow, a set of scripts was written to enforce such compatibility along the entire reaction profile.

For the reactive sites, i.e., the atoms undergoing a change in bonding situation throughout the reaction, stemming from the dipolarophiles, it is possible to readily enforce stereo-compatibility by setting the stereotags in the SMILES representations outputted from RDChiral²⁵ before the reaction profile computation is initiated, since doing so places constraints on the respective conformer search spaces in autodE. More specifically, we aimed to ensure that cis substituents in the reactant end up on the same side of the plane defined by the formed ring in the product (and *vice versa* for the trans substituents), see Fig. 7a.

To this end, we first verify that both centers undergoing addition are recognized as stereocenters in RDKit²⁰. Subsequently, the two potential product stereoisomers involving these centers are generated by setting the respective chiral tags, and guess structures are determined through a quick force field (MMFF94⁵⁰) optimization. A similar optimization is performed for the reactant dipolarophile. Next, the dihedral angle for both potential products are compared to the corresponding dihedral angle in the reactants (0° or 180°). The product SMILES resulting in the lowest deviation in dihedral angle from the reactant geometry is then retained as the stereochemically correct product.

For the reactive sites stemming from the dipoles, the situation is significantly more complex. First and foremost, it should be noted that for a significant number of scaffold types, there are no stereocenters at all (cf. the propargyl-allenyl ones), or these centers are so rigid that they are pre-set in practice (cf. the cyclic dipoles). For some allyl-type scaffolds however, particularly those involving two terminal C-centers, stereochemical considerations are relevant and compatibility needs to be enforced. Doing so in a similar manner as outlined for the dipolarophiles in the previous paragraph is not possible since the delocalization present in these dipoles causes the bonds connecting the individual centers to have bond orders in between 1 and 2, depending on the specific scaffold and the respective substituents. Consequently, either, both or neither of these bonds may be rotatable at room/body temperature⁵¹. To complicate the situation even further, it is impossible to know *a priori* which conformation around these partial double bonds will result in the lowest addition barrier to the specific dipolarophile considered. If no stereotags are specified for the dipole centers, autodE readily searches for the relative substituent orientation in the product that is most energetically favorable and the resulting arrangement will generally also be the lowest in the TS so that stereochemistry is generally retained in the second half of the reaction profile (*vide supra*). From the product and TS geometries, the compatible reactant dipole conformer can be determined, but this can only be done after an initial version of the profile has already been generated.

As such, the following pragmatic approach was taken throughout this study. Initially, we did not fix any stereotag associated with the dipole in the reaction SMILES fed to autodE in either reactant or product. Once the full reaction profile was obtained, a script to correct potential stereochemical incompatibilities for the dipoles was executed (Fig. 7b). The coordinates of the subset of atoms in the TS geometry corresponding to the reactant dipoles are first extracted. Subsequently, the obtained geometry is optimized with GFN2-xTB²². Next, a scan around the individual (partially double) dipole bonds is performed in 60 uniformly-spaced increments (GFN2-xTB level of theory; associated force constraint of 0.1 Hartree/Bohr²) to obtain a crude estimate of the barrier for rotation. If the spread in energies (or the bias energy due to the constraining potential) along the profile exceeds 20 kcal/mol, then the bond is assumed to be more or less rigid, i.e., non-rotatable, under physiological conditions. Finally, a *randomize and relax* (RR) conformer search is performed for each dipole within autodE¹⁸. 1000 additional conformers (in addition to the one extracted from the TS geometry) are generated while constraining the dihedral angles around non-rotatable bonds. After initial pruning, conformers are optimized at GFN2-xTB level of theory and the most stable one is compared with the originally selected conformer in the reaction profile. If the RR conformer is lower in energy, it is retained as a new reference for the reaction profile: DFT optimization and single-point frequency calculations are performed in a follow-up step, the final

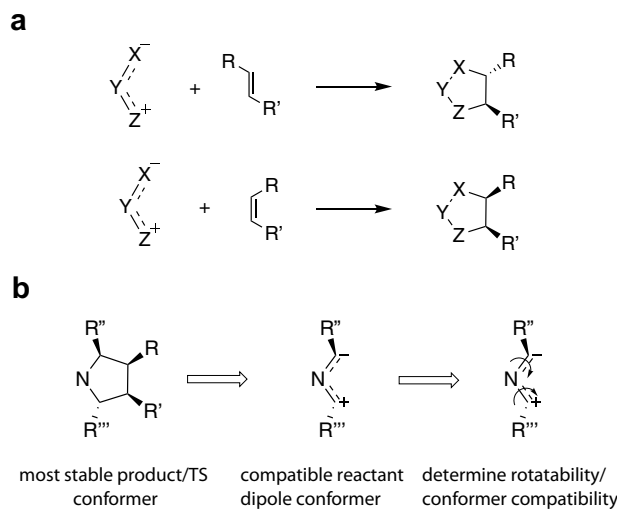


Fig. 7 (a) Illustration of stereo-retention around the unsaturated dipolarophile bond during a 1,3-dipolar cycloaddition reaction. (b) Schematic overview of the steps taken to ensure selection of a compatible dipole conformer: first, an approximate reactive dipole conformer is determined based on the selected (most stable) product/TS conformer geometry emerging from the reaction profile calculation); subsequently, the rotatability of the dipole bonds are assessed and the original reactant dipole conformer in the reaction profile is replaced by the (optimized) reactive one whenever rotation is determined to be too hindered.

species is included in the original output-folder for the reaction and updated activation and reaction energy values are computed. For dipoles with conformational restrictions (i.e., non-rotatable bonds), the lowest energy RR conformer is consistently selected as alternative reference for the reaction profile and the same procedure as described above is followed.

Data Records

All data files produced as part of this study are accessible through Figshare (https://figshare.com/articles/dataset/dipolar_cycloaddition_dataset/21707888)⁵². Reaction IDs and SMILES, activation energies (G^\ddagger ; in kcal/mol) and reaction energies (G_r ; in kcal/mol) for each computed reaction profile are provided in CSV format (`full_dataset.csv`). XYZ-files and LOG-files for both the final frequency and single-point calculation for each reactant (both the original and stereo-constrained versions), TS and product species, as well as a CSV file containing computed electronic energies and thermal corrections are available in a compressed archive file, `full_dataset_profiles.tar.gz`.

The files have been organized per reaction profile, identified through the reaction ID. Within each directory, reactant XYZ-files are of the form `r#####.xyz`, product XYZ-files are of the form `p#####.xyz`, and transition state XYZ-files are of the form `TS_#####.xyz`. If the reactant dipole conformer had to be corrected to enforce stereochemical compatibility, the latter XYZ-files are included under to form of `r#####_alt.xyz`. The frequency LOG-files can be found in a `frequency_logs` directory, and the single-point LOG-files can be found in a `single_point_logs` directory. The energies for all of these species are summarized per directory in `energies.csv`.

Additionally, all the benchmarking data are made available in the `benchmarking_data.tar.gz` directory (*vide supra*).

Technical Validation

The accuracy of the computed activation energies was assessed as part of the benchmarking study used to select the most appropriate DFT level of theory. As discussed in the Methods Section, errors on the (gas-phase) electronic energies were determined to be relatively small for the selected functional (MAE ~ 1 kcal/mol), and even when thermal and solvent corrections are included, the errors relative to experimental activation energies remain sufficiently low to extract chemically meaningful trends from the data (MAE ~ 2–3 kcal/mol).

The established workflow is also quite robust, with a failure rate of 3.5% during generation of the reaction SMILES from reactant combinations (mainly for the cyclic dipoles) and 12.3% during the ensuing autodE reaction profile computation and postprocessing. Unsurprisingly, the main cause of failure of the workflow is related to an inability to locate appropriate transition states, with Gaussian16 either not finding a saddle point on the PES altogether, finding a saddle point but not the right one, or not finding a converged solution of the TS search. Since it is hard to distinguish between outright TS search failures and genuinely barrierless reactions in an automated manner, we only retained reaction profiles for which an unequivocal saddle point could be determined on the PES. Another common source of failure is spontaneous rearrangement of the reactant(s) or product, indicating that the species encoded by the initial SMILES string does not correspond to a minimum on the PES, or, in other words, that this species is not a stable compound. It should be noted that for one specific biologically inspired motif (the guanidinium moiety), we did encounter a much higher failure rate (~66%). Our analysis

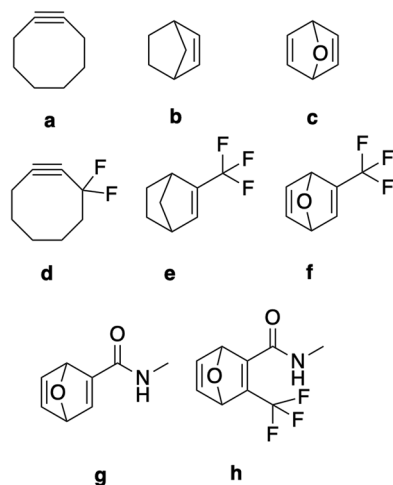


Fig. 8 The synthetic (strained) dipolarophiles included in the azide test reactions.

suggests that these failures are caused by the disappearance of the cycloaddition reaction mode for the involved species, not an actual failure of our methodology (see below).

To contextualize the numbers above, similar automated high-throughput reaction profile computation workflows achieve comparable or worse failure rates, e.g., a success rate of 85% was achieved by Friederich *et al.* in their study on dihydrogen activation of Vaska's complex⁵³ and Von Rudorff *et al.* were able to compute full reaction profiles for 25% of the E_2/S_N2 reaction systems considered in their search space⁵⁴. At the same time, even higher success rates have been reported as well, e.g., Jorner *et al.* were able to obtain 98% of the profiles in their study of nucleophilic aromatic substitution reactions. In the latter example, only reactions with a precedent in the experimental literature were considered, so “unrealistic” reactions were filtered out *a priori*, which is not the case for most other workflows developed⁵⁵.

Recovery of known bio-orthogonal click reactions from the presented workflow. To further validate our workflow, activation energies for two prototypical azide-based 1,3-dipoles (methyl and acyl azide respectively) with a couple of popular strained dipolarophiles, such as cyclooctyne and oxo-norbornadiene, as well as all the selected biofragments were computed (see Fig. 8 for an overview of the selected synthetic dipolarophiles).

In line with the results for the dataset as a whole, we find that our calculations fail for 3 out of 4 reactions with the tested guanidinium motif. This can be rationalized by considering the extremely unfavorable thermodynamics expected for these reactions: cycloaddition for these species requires the resonance in both the dipole and the guanidinium species to be broken, which imposes a significant delocalization penalty on any reaction mode involving these two species, and can potentially wipe out a barrier along the reaction pathway altogether (see Fig. 8a)⁵⁶. Support for this reasoning can be found in the fact that the reason for failure for each of these profiles was the inability by autodE to locate a stable molecule with the bonding pattern corresponding to the product, i.e., the expected product is not a minimum on the potential energy surface, and the observation that out of the—in total—61 reaction profiles involving guanidinium which we were, in fact, able to compute successfully throughout the entire dataset, close to 95% turned out to be endothermic, and over two-thirds are significantly so (endothermic by 20–65 kcal/mol). As an extra check, we attempted to compute reaction profiles for an alternative stepwise mechanism as well, but here the product tended to spontaneously decompose back into the reactants upon optimization.

Focusing on the successfully computed reaction profiles, we find that most reactions between acyl/methyl azide and strained dipolarophiles exhibit a relatively low activation energy (between 21 and 25 kcal/mol under standard conditions), suggesting that most will readily proceed under physiological conditions (Fig. 9). Furthermore, all of these reactions are highly exothermic, i.e., they are irreversible. The workflow readily identifies the reaction between (fluorinated) cyclooctyne and methyl/acyl azide as the most rapid transformations ($\Delta G^\ddagger = 18\text{--}22$ kcal/mol; see the upper part of Fig. 9b), underscoring the experimentally observed excellent click-potential of this popular dipole-dipolarophile combination⁹. On the other hand, reactions of the same azides with the biofragments tested tend to involve higher barriers (25 kcal/mol and up) on average, indicating that these will proceed at much slower/non-competitive rates than the ones involving synthetic/strained dipolarophiles.

We found a single reaction involving biofragments with a relatively low reaction barrier ($\Delta G^\ddagger = 23.6$ kcal/mol), which ought to be competitive with all but the most reactive ones involving strained dipolarophiles. This reaction involves a relatively exotic motif corresponding to (a fragment of) NADH/NADPH. While these compounds are intermediates of important biochemical processes⁵⁷, their concentrations tend to be relatively low; NADH can typically be found in mammalian cells at concentrations of approximately $10^{-4}\text{--}10^{-5}$ M, with most of it bound to proteins and hence only partially available for reaction^{58,59}. Consequently, even for synthetic reactions exhibiting similar activation energies, these native reactions will likely not disrupt the bio-orthogonal click

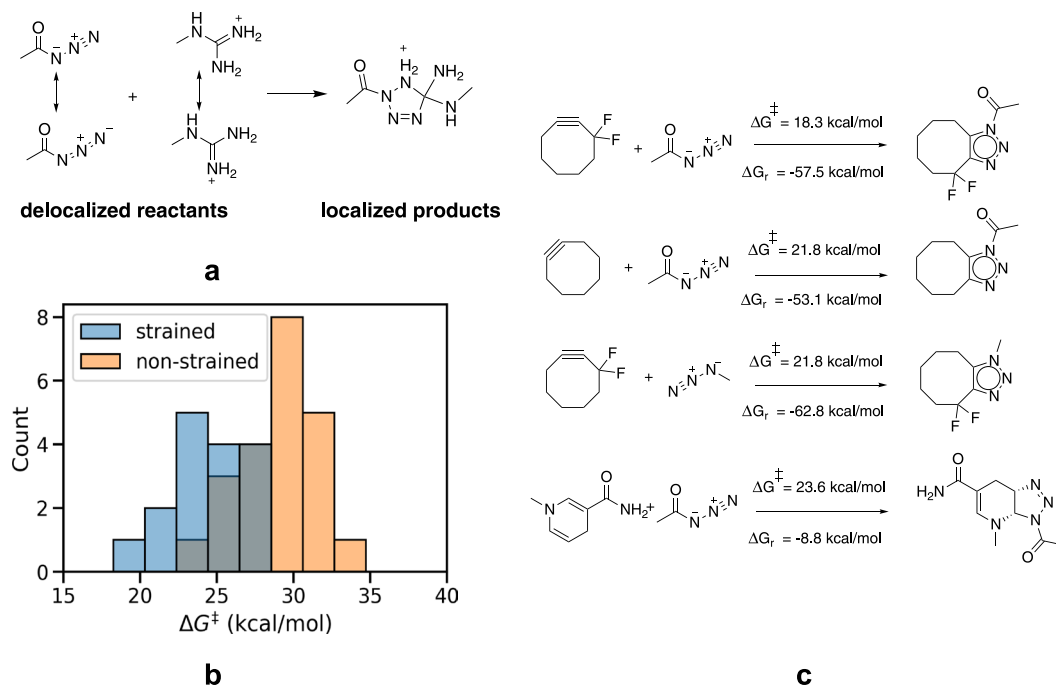


Fig. 9 (a) A schematic depiction of the resonance loss associated with the reaction between the guanidinium motif and acyl azide. (b) Histogram depicting the distribution of the activation energies (ΔG^\ddagger) for both strained and non-strained dipolarophiles with methyl and acyl azide dipoles. (c) The three lowest reaction barriers computed for the strained dipolarophiles (top) and for the lowest non-strained dipolarophiles (bottom).

function completely: unless the synthetic reaction is (significantly) slower and/or the synthetic dipolarophile is (much) less prevalent/available for the dipole partner, the latter will still dominate.

The test case above underscores the subtleties—and limits—associated with bio-orthogonal click chemistry: even for tried and tested synthetic reactions, the difference in reactivity compared to the fastest native reactions tends to be rather subtle. Nevertheless, our computational approach enables us to retrieve the main reactivity trends for these [3 + 2] cycloadditions to identify promising/suitable bio-orthogonal click reactions.

Accuracy and reproducibility of the data. Some additional tests to probe the accuracy and reproducibility of the data obtained through our workflow were performed on the azide reaction list described in the previous subsection. First, we considered the robustness and reproducibility of the cycloaddition reaction profiles generated with autodE in general. Since the conformer generation algorithm used in RDKit is stochastic in nature, two consecutive runs of autodE on the same reaction SMILES will not necessarily yield the exact same result since the sampled conformers can differ slightly.

To assess the magnitude of the resulting uncertainty, barrier heights and reaction energies for the azide reaction list were computed twice with the default autodE settings, i.e., conformers were generated with RDKit, after which the unique ones were optimized with DFT and the lowest-energy one was selected, and the resulting values were compared. As can be seen from Fig. 10a,b, activation and reaction energies for our test set are reproduced well (MAE \sim 0.7 kcal/mol and RMSE \sim 1.2 kcal/mol for the activation energies and MAE \sim 0.8 kcal/mol and RMSE \sim 1.3 kcal/mol for the reaction energies).

Next, we assessed the effect of the approximation in conformer selection applied during the dataset generation. More specifically, all the barriers for the azide reaction list were again computed twice: in one run, conformers were selected based on GFN2-xTB conformer energies, in the second run, conformer selection was done in the default manner, i.e., based on DFT energies. As can be observed from Fig. 10c,d, the correlation between the barriers obtained in both runs is reasonably close to the reproducibility error at the most accurate conformer selection criterion (MAE \sim 1.0 kcal/mol and RMSE \sim 1.7 kcal/mol for the activation energies and MAE \sim 0.8 kcal/mol and RMSE \sim 1.3 kcal/mol for the reaction energies).

Finally, the reproducibility of the workflow with conformer selection at GFN2-xTB level of theory was checked. As can be observed from Fig. 10e,f, the reproducibility errors with and without the approximation are perfectly in line (MAE \sim 0.7 kcal/mol and RMSE \sim 1.1 kcal/mol for the activation energies and MAE \sim 0.7 kcal/mol and RMSE \sim 1.2 kcal/mol for the reaction energies).

Taking everything together, one can conclude that selecting conformers based on GFN2-xTB energies introduces only a small additional error relative to selection based on DFT energies, and thus we decided to consistently apply this approximation since this accelerates the workflow significantly, facilitating a broader exploration of the defined chemical space.

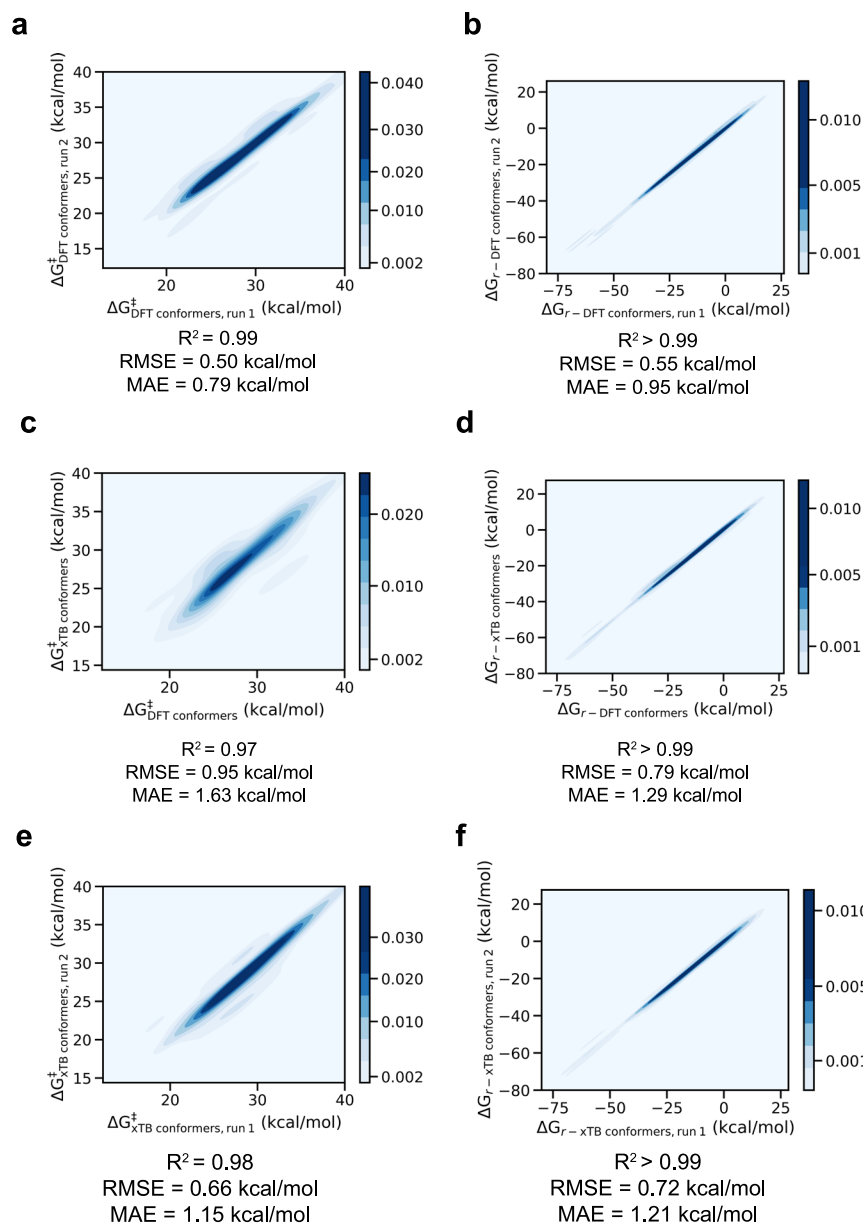


Fig. 10 (a) Correlation between the activation energies (ΔG^\ddagger) and (b) the reaction energies (ΔG_r) for two consecutive autodE runs with conformer selection at DFT level of theory for the azide test reactions. (c) Correlation between the activation energies and (d) the reaction energies for an autodE run with conformer selection at DFT level of theory and a consecutive run with conformer selection at GFN2-xTB level of theory. (e) Correlation between activation energies and (f) the reaction energies for two consecutive autodE runs with conformer selection at GFN2-xTB level of theory.

Usage Notes

The code used to generate the reaction SMILES and automate the autodE workflow is available on GitHub. The repository contains several Python scripts and Jupyter Notebooks:

- Notebooks to generate the respective search spaces and extract samples from them.
- A script to generate reaction SMILES with RDChiral and set the stereotags of the chiral centers associated with the dipolarophiles.
- A script and auxiliary modules to launch high-throughput reaction profile computation with autodE in a parallelized manner.
- A script to correct stereochemical incompatibilities in the selected reactant dipole conformers.
- A script and auxiliary modules to launch high-throughput DFT optimization and single-point fine-tuning of corrected dipole conformers.

Code availability

The code described in the previous section is freely available in GitHub under the MIT license (https://github.com/coleygroup/dipolar_cycloaddition_dataset). Further details on how to use it is provided in the associated README.md file.

Received: 21 September 2022; Accepted: 18 January 2023;

Published online: 01 February 2023

References

- Saxon, E. & Bertozzi, C. R. Cell surface engineering by a modified Staudinger reaction. *Science* **287**, 2007–2010 (2000).
- Agard, N. J., Prescher, J. A. & Bertozzi, C. R. A strain-promoted [3 + 2] azide-alkyne cycloaddition for covalent modification of biomolecules in living systems. *Journal of the American Chemical Society* **126**, 15046–15047 (2004).
- Sletten, E. M. & Bertozzi, C. R. Bioorthogonal chemistry: fishing for selectivity in a sea of functionality. *Angew. Chem., Int. Ed.* **48**, 6974–6998 (2009).
- Devaraj, N. K. The future of bioorthogonal chemistry. *ACS Central Science* **4**, 952–959 (2018).
- Agarwal, P. & Bertozzi, C. R. Site-specific antibody–drug conjugates: the nexus of bioorthogonal chemistry, protein engineering, and drug development. *Bioconjug. Chem.* **26**, 176–192 (2015).
- Kim, J. & Bertozzi, C. R. A bioorthogonal reaction of N-oxide and boron reagents. *Angew. Chem., Int. Ed.* **54**, 15777–15781 (2015).
- Sletten, E. M. & Bertozzi, C. R. A bioorthogonal quadricyclane ligation. *J. Am. Chem. Soc.* **133**, 17570–17573 (2011).
- Nguyen, S. S. & Prescher, J. A. Developing bioorthogonal probes to span a spectrum of reactivities. *Nat. Rev. Chem.* **4**, 476–489 (2020).
- Jewett, J. C. & Bertozzi, C. R. Cu-free click cycloaddition reactions in chemical biology. *Chem. Soc. Rev.* **39**, 1272–1279 (2010).
- Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **38**, 3098 (1988).
- Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37**, 785 (1988).
- Stephens, P. J., Devlin, F. J., Chabalowski, C. F. & Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **98**, 11623–11627 (1994).
- Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **27**, 1787–1799 (2006).
- Schäfer, A., Horn, H. & Ahlrichs, R. Fully optimized contracted Gaussian basis sets for atoms Li to Kr. *J. Chem. Phys.* **97**, 2571–2577 (1992).
- Schäfer, A., Huber, C. & Ahlrichs, R. Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *J. Chem. Phys.* **100**, 5829–5835 (1994).
- Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **113**, 6378–6396 (2009).
- Evans, M. & Polanyi, M. Further considerations on the thermodynamics of chemical equilibria and reaction rates. *J. Chem. Soc. Faraday Trans.* **32**, 1333–1360 (1936).
- Young, T. A., Silcock, J. J., Sterling, A. J. & Duarte, F. autodE: Automated calculation of reaction energy profiles—application to organic and organometallic reactions. *Angew. Chem., Int. Ed.* **60**, 4266–4274, <https://doi.org/10.1002/anie.202011941> (2021).
- Frisch, M. *et al.* Gaussian 16 (2016).
- Landrum, G. RDKit documentation. *Release* **1**, 4 (2013).
- Riniker, S. & Landrum, G. A. Better informed distance geometry: using what we know to improve conformation generation. *J. Chem. Inf. Model.* **55**, 2562–2574 (2015).
- Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **15**, 1652–1671 (2019).
- Goodman, J. M. & Silva, M. A. QRC: a rapid method for connecting transition structures to reactants in the computational analysis of organic reactivity. *Tetrahedron Lett.* **44**, 8233–8236 (2003).
- Breugst, M. & Reissig, H.-U. The Huisgen reaction: Milestones of the 1,3-dipolar cycloaddition. *Angew. Chem., Int. Ed.* **59**, 12293–12307 (2020).
- Coley, C. W., Green, W. H. & Jensen, K. F. RDChiral: An RDKit wrapper for handling stereochemistry in retrosynthetic template extraction and application. *J. Chem. Inf. Model.* **59**, 2529–2537 (2019).
- Bergström, S., Carlson, L. A. & Weeks, J. R. The prostaglandins: a family of biologically active lipids. *Pharmacol. Rev.* **20**, 1–48 (1968).
- Krebs, H. A., Salvin, E. & Johnson, W. A. The formation of citric and α -ketoglutaric acids in the mammalian body. *Biochem. J.* **32**, 113 (1938).
- Ernster, L. & Dallner, G. Biochemical, physiological and medical aspects of ubiquinone function. *Biochim. Biophys. Acta Mol. Basis Dis.* **1271**, 195–204 (1995).
- Khaleel, C., Tabanca, N. & Buchbauer, G. α -terpineol, a natural monoterpene: A review of its biological properties. *Open Chemistry* **16**, 349–361 (2018).
- Turrens, J. F. & Boveris, A. Generation of superoxide anion by the nadh dehydrogenase of bovine heart mitochondria. *Biochem. J.* **191**, 421–427 (1980).
- McIlwain, J. T. *An introduction to the biology of vision* (Cambridge University Press, 1996).
- Burton, G. W. & Ingold, K. β -carotene: an unusual type of lipid antioxidant. *Science* **224**, 569–573 (1984).
- Erez, A., Nagamani, S. C. S. & Lee, B. Argininosuccinate lyase deficiency—argininosuccinic aciduria and beyond. In *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 1, 45–53 (Wiley Online Library, 2011).
- Yang, X., Zou, J., Wang, Y., Xue, Y. & Yang, S. Role of water in the reaction mechanism and endo/exo selectivity of 1,3-dipolar cycloadditions elucidated by quantum chemistry and machine learning. *Chem. Eur. J.* **25**, 8289–8303 (2019).
- Huisgen, R. Kinetics and mechanism of 1,3-dipolar cycloadditions. *Angew. Chem., Int. Ed.* **2**, 633–645 (1963).
- Morrison, P. R. & Ryser, F. A. Weight and body temperature in mammals. *Science* **116**, 231–232 (1952).
- Adamo, C. & Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **110**, 6158–6170 (1999).
- Zhao, Y. & Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **120**, 215–241 (2008).
- Chai, J.-D. & Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Physical Chemistry Chemical Physics* **10**, 6615–6620 (2008).
- Chai, J.-D. & Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *J. Chem. Phys.* **128**, 084106 (2008).

41. Karton, A. & Goerigk, L. Accurate reaction barrier heights of pericyclic reactions: Surprisingly large deviations for the CBS-QB3 composite method and their consequences in DFT benchmark studies. *J. Comput. Chem.* **36**, 622–632 (2015).
42. Goerigk, L. & Grimme, S. A general database for main group thermochemistry, kinetics, and noncovalent interactions- assessment of common and reparameterized (meta-)GGA density functionals. *Journal of Chemical Theory and Computation* **6**, 107–126 (2010).
43. Ess, D. H. & Houk, K. Activation energies of pericyclic reactions: performance of DFT, MP2, and CBS-QB3 methods for the prediction of activation barriers and reaction energetics of 1,3-dipolar cycloadditions, and revised activation enthalpies for a standard set of hydrocarbon pericyclic reactions. *J. Phys. Chem. A* **109**, 9542–9553 (2005).
44. Karton, A. & Martin, J. M. Explicitly correlated W_n theory: W1-F12 and W2-F12. *J. Chem. Phys.* **136**, 124114 (2012).
45. Vermeeren, P. *et al.* Pericyclic reaction benchmarks: hierarchical computations targeting CCSDT(Q)/CBS and analysis of DFT performance. *Phys. Chem. Chem. Phys.* **24**, 18028–18042 (2022).
46. Huisgen, R. & Padwa, A. 1,3-dipolar cycloaddition chemistry. *Wiley, New York* **1**, 55–92 (1984).
47. Geittner, J., Huisgen, R. & Reissig, H. Solvent dependence of cyclo-addition rates of phenyldiazomethane and activation parameters. *Heterocycles* **11**, 109–112 (1978).
48. Huisgen, R., Ohta, A. & Geittner, J. 1,3-dipolar cycloadditions. LXXVI. Cycloadditions of diazomethane to substituted butadienes. *Chemical and Pharmaceutical Bulletin* **23**, 2735–2743 (1975).
49. Eyring, H. The activated complex in chemical reactions. *J. Chem. Phys.* **3**, 107–115 (1935).
50. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **17**, 490–519 (1996).
51. Huisgen, R., Scheer, W. & Huber, H. Stereospecific conversion of *cis-trans* isomeric aziridines to open-chain azomethine ylides. *Journal of the American Chemical Society* **89**, 1753–1755 (1967).
52. Stuyver, T., Jorner, K. & Coley, C. Dipolar cycloaddition dataset, *Figshare*, <https://doi.org/10.6084/m9.figshare.21707888.v5> (2023).
53. Friederich, P., dos Passos Gomes, G., De Bin, R., Aspuru-Guzik, A. & Balcells, D. Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex. *Chem. Sci.* **11**, 4584–4601 (2020).
54. von Rudorff, G. F., Heinen, S. N., Bragato, M. & von Lilienfeld, O. A. Thousands of reactants and transition states for competing E2 and S_N2 reactions. *Machine Learning: Science and Technology* **1**, 045026 (2020).
55. Jorner, K., Brinck, T., Norrby, P.-O. & Buttar, D. Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. *Chem. Sci.* **12**, 1163–1175 (2021).
56. Stuyver, T., De Proft, F., Geerlings, P. & Shaik, S. How do local reactivity descriptors shape the potential energy surface associated with chemical reactions? The valence bond delocalization perspective. *Journal of the American Chemical Society* **142**, 10102–10113 (2020).
57. Ying, W. NAD⁺/NADH and NADP⁺/NADPH in cellular functions and cell death: regulation and biological consequences. *Antioxidants & redox signaling* **10**, 179–206 (2008).
58. Yang, H. *et al.* Nutrient-sensitive mitochondrial NAD⁺ levels dictate cell survival. *Cell* **130**, 1095–1107 (2007).
59. Blinova, K. *et al.* Distribution of mitochondrial NADH fluorescence lifetimes: steady-state kinetics of matrix NADH interactions. *Biochem.* **44**, 2585–2594 (2005).

Acknowledgements

T.S. and C.W.C. acknowledge the Machine Learning for Pharmaceutical Discovery and Synthesis Consortium for funding this work. K.J. acknowledges funding through an International Postdoc grant from the Swedish Research Council (No. 2020-00314). The authors acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing HPC resources that have contributed to the research results reported within this paper.

Author contributions

T.S. and C.W.C. conceived the study, T.S. performed the computations and wrote the associated code, T.S. and K.J. developed the methodology and analyzed the data. All authors contributed to the writing of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.W.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023