SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry

Colin A. Grambow [ID], Lagnajit Pattanaik [ID] & William H. Green [ID] [✉]

Reaction times, activation energies, branching ratios, yields, and many other quantitative attributes are important for precise organic syntheses and generating detailed reaction mechanisms. Often, it would be useful to be able to classify proposed reactions as fast or slow. However, quantitative chemical reaction data, especially for atom-mapped reactions, are difficult to find in existing databases. Therefore, we used automated potential energy surface exploration to generate 12,000 organic reactions involving H, C, N, and O atoms calculated at the ωB97X-D3/def2-TZVP quantum chemistry level. We report the results of geometry optimizations and frequency calculations for reactants, products, and transition states of all reactions. Additionally, we extracted atom-mapped reaction SMILES, activation energies, and enthalpies of reaction. We believe that this data will accelerate progress in automated methods for organic synthesis and reaction mechanism generation—for example, by enabling the development of novel machine learning models for quantitative reaction prediction.

Background & Summary

Rapid advancements in computational methods for chemical synthesis planning and automated reaction mechanism generation, especially in the area of machine learning, are causing a significant shift in how such problems are tackled. Deep learning approaches are replacing conventional quantitative structure-activity relationships often based on support vector machines, decision trees, or linear methods like partial least squares[1,2]. These new systems are becoming widely available for computer-aided retrosynthesis[3], reaction outcome prediction[3], high-throughput virtual screening[4], and more general molecular property prediction[5,6]. Computational approaches are also increasingly common in reaction mechanism generation due to the large number of species and reactions that are generally required for accurate descriptions of phenomena like pyrolysis, combustion, and atmospheric oxidation[7–9]. Frequently, this involves characterizing chemical pathways with quantum chemistry[8], but deep learning methods have also recently been applied to estimate thermochemistry during mechanism generation[10,11].

While computers already outperform humans at qualitatively predicting reaction products[12,13] and successful yield predictions have been demonstrated for limited datasets[14,15], quantitative reaction information is still elusive in large databases like Reaxys[16], Pistachio[17], and the United States Patent and Trademark Office database[18]. Reaction yield, time, and some quantitative conditions like temperature are sometimes available, but there is usually no information on reaction kinetics. If such data were available, calculation of derived properties—such as minimum reaction times and branching ratios—would be possible. Our goal is to provide a quantitative dataset of reactions that enables the calculation of such data and can lead to more efficient drug design and help in deciding which reactions are important in mechanism generation.

Computationally generating a dataset of reactions is significantly more complex than only calculating stable equilibrium structures because transition states (TSs) of chemical reactions cannot be enumerated in the same manner as stable molecules. Even if the reactant and product structures are known, the exact TS geometry has to be found via a human-guided search or with expensive automated TS finding methods. Here, we use automated

Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA, 02139, United States. [✉]e-mail: whgreen@mit.edu
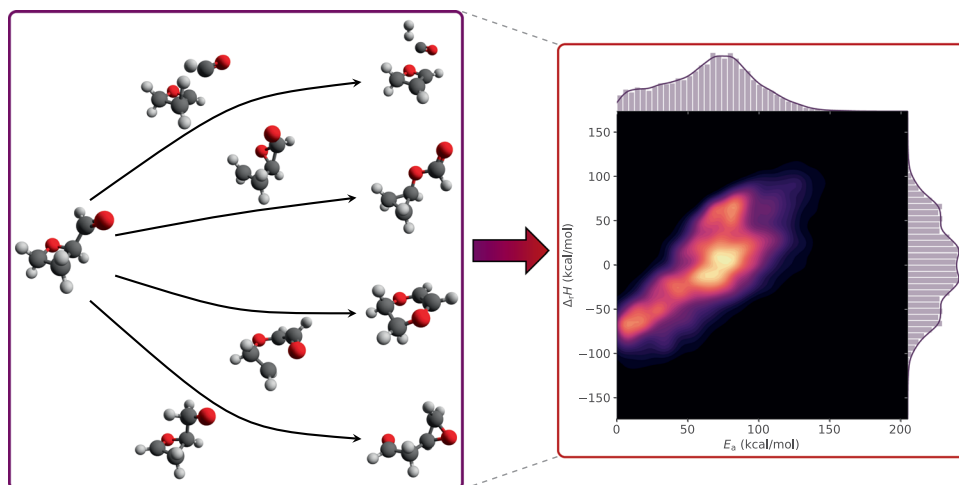
Fig. 1 Reaction data generation and visualization of reaction space. During data generation, many reactants are optimized, hundreds of reaction paths for each reactant are searched with an automated transition state finding method, and the resulting products are optimized. The reaction space spans a wide range of activation energies and is visualized with a bivariate kernel density estimate (using a Gaussian kernel) of the probability density of the activation energy and enthalpy of reaction. The visualization encompasses both forward and reverse reactions.

potential energy surface exploration to generate the dataset of reactions, which has been shown to be successful in cases when many reaction pathways have to be evaluated[19–21]. More specifically, we rely on the growing string method[22] to automatically optimize reaction paths and TSs.

We report quantum chemical data on more than 16,000 reactions in the form of reactants, products, and TSs at the B97-D3/def2-mSVP level of theory and 12,000 reactions at the ωB97X-D3/def2-TZVP level of theory. The data include the raw output from geometry optimizations and frequency calculations in addition to atom-mapped SMILES, activation energies, and enthalpies of reaction. All reactions are gas-phase calculations involving up to seven carbon, oxygen, or nitrogen atoms per molecule. The reactants are sampled from GDB-7, a subset of GDB-17[23], meaning that all reactions have a unimolecular reactant but potentially multi-molecular products. Figure 1 illustrates the dataset generation process and the resulting space of reactions in terms of their activation energies and enthalpies of reaction.

Methods

Overview.   The dataset generation procedure started by selecting molecules from GDB-7[23], generating conformers, and optimizing the lowest-energy conformer. An exhaustive set of driving coordinates subject to valence and connectivity constraints were generated for each reaction. Reaction paths were calculated with the growing string method[22], which searched along each of the driving coordinates. Products and TSs discovered in this way were reoptimized, duplicate reactions were removed, and checks were performed to verify the reactions. The generated reactions were then refined at a higher level of theory. Because of the large number of density functional theory (DFT) calculations required, the massively parallel nature of the calculations was exploited by running thousands of calculations in parallel on a supercomputer.

Reactant optimization.   Because of the unfavorable scaling of quantum chemical calculations, we only considered molecules with at most seven heavy atoms (C, N, O). *All* molecules with six or fewer heavy atoms were selected from GDB-7 (~770) and a random selection of ~430 molecules were selected from the set with seven heavy atoms. Starting from the SMILES strings, we embedded several hundred conformers for each molecule using the RDKit[24] with the ETKDG distance geometry method[25] and relaxed their geometries using the MMFF94 force field implemented in RDKit. The lowest energy structure was selected for each molecule and optimized at both the B97-D3/def2-mSVP with Becke-Johnson damping level of theory[26] and the ωB97X-D3/def2-TZVP[27] level of theory with Q-Chem 5.1[28]. We ascertained that none of the molecules contained imaginary frequencies. All calculations, including the subsequent string method calculations, were done in the singlet state and used a spin-unrestricted ansatz because the bond distortions occurring in the corresponding TSs might be better treated with an unrestricted formulation. The def2-mSVP basis set in the Karlsruhe *def2* basis set family[29] is a modified version of def2-SV(P), which corrects for an overestimation of bond lengths involving hydrogen[30]. All DFT calculations used the *SG-2* standard quadrature grid, which is of sufficient quality for B97-based functionals[31].

Potential energy surface exploration.   The most demanding and most time-intensive step of the reaction generation process is the optimization of reaction paths to the minimum energy paths (MEPs) containing the correct TS structures. We accomplished this in an automated fashion by using the single-ended growing string method (GSM)[22] at the B97-D3/def2-mSVP level of theory. GSM performs the reaction path optimization using a set of delocalized internal coordinates, which means that the resulting MEPs may be slightly different than those obtained via a reaction path following procedure in mass-weighted internal coordinates[32]. Single-ended methods

only require a reactant structure to find reactions whereas double-ended methods additionally require knowledge of the product[33,34]. *A priori* specification of the product can be problematic when there is no simple elementary step connecting reactant and product. Single-ended GSM solves this issue by only requiring a set of driving coordinates to initiate the reaction path search.

In our case, the driving coordinates are specified as bond transformations in terms of primitive internal coordinates. The direction given by the primitive internal coordinate vector is projected onto the nonredundant delocalized internal coordinates[35], which is the space in which the reaction path optimization occurs. This results in a single tangent vector that represents all of the driving coordinates simultaneously. Importantly, this allows all other coordinates to change without constraint during the optimization, thus allowing necessary angle, torsion, and even additional bond changes to occur. Once a path has been grown, the entire path is optimized towards the MEP while monitoring the number of TSs along the path and truncating it if more than one TS is detected—ensuring that the reaction is elementary. As a result of this, not all bond changes given in the driving coordinates are guaranteed to occur. Towards the end of the path optimization, an exact TS search takes place guided by curvature information from the string.

In order to obtain many reactions, we generated an exhaustive list of driving coordinate sets for each reactant subject to a few constraints. Because elementary reactions usually involve few bond changes, we specified that at most two bonds could be broken, at most two bonds could be formed, and a total of at most three bonds could be changed. A "bond" in this sense ignored bond orders and only considered whether two atoms were connected to each other. Note that these constraints were only selected to ensure a computationally tractable number of driving coordinates. As described in the previous paragraph, these limits did not apply during the actual path optimization, they were only used to specify the initial search direction. We also ignored driving coordinates involving only a single bond change as these would likely correspond to barrierless associations or dissociations. Driving coordinates involving equivalent hydrogens were not included. Equivalent hydrogens only differ in their atom indices, *e.g.*, a hydrogen atom that is part of a methyl group is considered to be equivalent to another hydrogen in the same methyl group. Lastly, the driving coordinates were further limited based on the valences of the expected product structures. Hydrogen atoms must have one bond, carbons can be connected to a minimum of two and a maximum of four atoms, oxygen to a minimum of one and a maximum of two, and nitrogen to a minimum of one and a maximum of three.

This process usually resulted in several hundred sets of driving coordinates per reactant. Following each GSM calculation, the endpoint of the paths were subjected to additional geometry optimizations to ensure that the product structures were at a minimum. For each reactant, there were many duplicate reactions. Instead of discarding all of them, up to four duplicates of the same reaction were retained for additional TS optimization in case some of the optimizations fail. While GSM already produces a mostly optimized TS structure, the additional optimization step ensured that the TSs were optimized to high accuracy.

Reaction verification and extraction.
After the additional TS optimizations, duplicate reactions were filtered out again. If duplicates were present, the lowest-barrier reaction was retained. Differences in barrier height may arise due to different TS conformers. Although GSM provided an optimized MEP for each reaction, it is possible that some reactions containing incorrect transition states remained. These were filtered out according to a normal mode analysis described in the Technical Validation section.

To convert from three-dimensional geometries to SMILES[36], connections and bond orders could be perceived with Open Babel[37]. However, there were cases where the derived bond orders were chemically unreasonable, for example, when the resulting SMILES contained adjacent radical atoms which most likely correspond to double bonds. To eliminate unreasonable structures, we converted the Open Babel molecule to InChI[38], which only treats bond orders implicitly and resolves the issue. A downside to using InChI is that tautomers are assigned the same string, but this can be circumvented by converting to a nonstandard InChI containing a fixed-hydrogen layer. Additionally, atom ordering was lost in the InChI conversion. We reconstructed the atom map by converting to an RDKit molecule and determining the graph isomorphism between the original molecule and the RDKit molecule without considering bond orders. In the future, an alternative procedure for perceiving SMILES could be implemented based on natural bond orbital analysis[39].

The activation energies were extracted by adding the zero-point energies from a harmonic vibrational analysis to reactant, product, and TS energies and computing the difference between resulting TS and reactant energies. Similarly, enthalpies of formation were determined based on the difference of product and reactant energies.

Refinement.
B97-D3/def2-mSVP strikes a reasonable balance between cost and accuracy for potential energy surface exploration, but does not provide particularly accurate energies. Therefore, we refined the discovered pathways using ωB97X-D3/def2-TZVP. As mentioned earlier, reactants were already optimized with ωB97X-D3/def2-TZVP. Reactions were extracted as described in the preceding subsection, but some duplicates were retained to increase the probability of successful reoptimization. Only the duplicate with the smallest activation energy was retained in the end. Products and TSs were then reoptimized with ωB97X-D3/def2-TZVP and the final high-level reactions were extracted as before.

Data Records
Q-Chem output files, extracted SMILES, activation energies, and enthalpies of formation are available for 16,365 B97-D3/def2-mSVP reactions and for 11,961 ωB97X-D3/def2-TZVP reactions[40]. The raw log files are stored in two compressed archive files, `b97d3.tar.gz` and `wb97xd3.tar.gz` for B97-D3/def2-mSVP and ωB97X-D3/def2-TZVP data, respectively. Each archive contains a separate folder for each reaction labelled `rxn######`, where `######` denotes the reaction number padded with zeros. Within each folder are the three log files for a reaction, `r######.log` for the reactant, `p######.log` for the product, and `ts######.log`

Column label	Description
idx	Reaction index
rsmi	Reactant SMILES
psmi	Product SMILES
ea	Activation energy (kcal mol^{-1})
dh	Enthalpy of reaction (kcal mol^{-1})

Table 1. A description of the columns in the comma-separated values files.

for the transition state. Each log file contains the output of a geometry optimization and harmonic vibrational analysis.

Atom-mapped SMILES, activation energies, and enthalpies of formation for each reaction are listed in the comma-separated values files `b97d3.csv` and `wb97xd3.csv` for the B97-D3/def2-mSVP and ωB97X-D3/def2-TZVP levels of theory, respectively. The reactions are listed in the same order as the corresponding folders in the archive files. The columns in the comma-separated values files are explained in Table 1.

During the potential energy surface exploration, many duplicate reactions were encountered which were filtered out. Additionally, reactions that did not pass the tests described in the Technical Validation section were removed from the final list. Nonetheless, all of these calculations also produced optimized transition states, although the reactants and products were not verified for many of them, and duplicate transition states exist. These data may still prove to be useful if only transition state structures are required or if additional calculations are done to obtain the corresponding reactants and products. Therefore, the log files for all successfully optimized transition states at both levels of theory are stored in `ts_with_dup_b97d3.tar.gz` and `ts_with_dup_wb97xd3.tar.gz`. There are 69,366 B97-D3/def2-mSVP transition states and 24,987 ωB97X-D3/def2-TZVP transition states.

Technical Validation

Although the growing string method produces an optimized minimum energy path that should contain the correct TS in most cases, insufficient path discretization and reoptimization of TS geometries can lead to convergence failures or result in incorrect transition states. We performed several checks to filter out incorrect reactions. We ensured that all TSs have exactly one imaginary frequency. Reactions were also removed if the energy during the TS optimization changed by more than 3 kcal mol^{-1} relative to the highest energy on the growing string path. The most important check that we performed was to verify that the atomic displacements for the imaginary frequency matched the bond changes that occurred going from the proposed reactant to product. For each proposed reaction, we determined which bonds were changing in the reaction and ensured that the imaginary frequency normal mode displacements along those bonds were larger than the displacements along all the other bonds. This indicated that movement along the reaction coordinate mostly involved atoms undergoing significant change in the reaction. After all these changes, there is still the possibility that some of the transition states are incorrect. As a final check, we removed all of the reactions where the imaginary frequency of the transition state was less than 100 cm^{-1} in magnitude, as these typically correspond to conformational changes.

To avoid excessive computational cost, DFT methods had to be used to generate the reaction dataset. The functional chosen for the string method calculations, B97-D3, does not provide accurate activation energies, but was selected due to its low computational cost. However, ωB97X-D3 has been shown to yield excellent quantitative barrier heights with a 2.28 kcal mol^{-1} root-mean-square deviation from reference data that is estimated to be more than ten times as accurate as the best density functionals[41], which makes this data very useful. Therefore, the following analysis was only completed for the ωB97X-D3 data.

In order to show that the dataset provides a reasonably diverse set of reactions spanning many different chemistries even though constraints were set on the number of atoms and driving coordinate generation parameters, it is necessary to characterize the types of reactions. Figure 1 already shows that the range of activation energies and enthalpies of formation is very large. Even high-energy reactions involving barriers of up to 200 kcal mol^{-1} are included in the dataset. If the data are used to learn reaction prediction models, including such high-energy paths is important in order to not bias models towards the low-energy regions. Figure 2 shows that even though the driving coordinates were limited to three bond changes, significantly more complex reactions involving more bond changes occur in the dataset. Nonetheless, most elementary reactions predominantly occur with only two or three bond changes. Furthermore, the median activation energy increases with an increasing number of bond changes, which is expected.

Instead of simply counting the number of bond changes, the reactions can be classified based on the types of bonds that are changed. Figure 3 shows that all combinations of bond changes between H, C, N, and O atoms occur in the dataset with many examples present for all reaction types. H–H changing reactions are the rarest because they only correspond to hydrogen molecule formation.

Lastly, we characterized the reaction diversity by automatically extracting a set of general templates. We only focused on the reactive center by using RDKit's *GetReactingAtoms* method to isolate atoms changing in the reaction. The molecular fragments in the reactants and products identified as the reactive center were then concatenated together to form the reaction template. In addition to the connectivity of the reacting atoms, the only features considered were atom identity, charge, aromaticity, and bond type. Figure 4 shows the results of this automated extraction. Many templates only have a single reaction example and only the eight most popular templates have more than 100 reaction examples, highlighting the diversity present in the dataset.
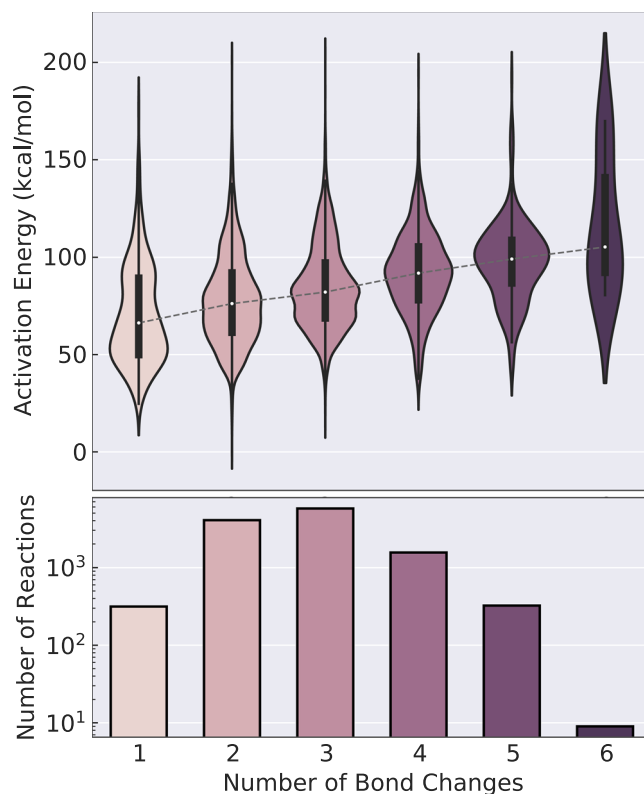
Fig. 2 Activation energy distributions. The distribution of activation energies split by the number of bond changes in the ωB97X-D3 reactions. Bond changes only consider changes in connectivity between atoms, irrespective of bond order. The distributions are scaled to have equal area.
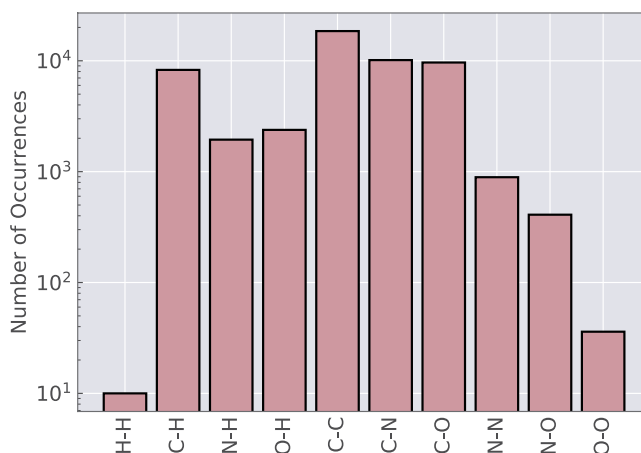

Fig. 3 Bond change types. The number of times each type of bond change occurs in the ωB97X-D3 reactions. For example, C–N denotes both forming a bond between C and N atoms and breaking a bond between the atoms. This also includes a change in the bond order between the two atoms.

Usage Notes

With the exception of the growing string method code, which is available from the developers of the method[42], and the Q-Chem quantum chemistry package, all code necessary to reproduce the generated data is available on GitHub[43]. The repository contains several scripts, which should be run in the following order:

- `parse_qm9.py`: Converts the QM9 data directory[44], which contains the GDB-9 SMILES along with quantum mechanically derived properties, to a pickled file containing a list of `MolData` objects, which store the information in QM9 as Python objects.
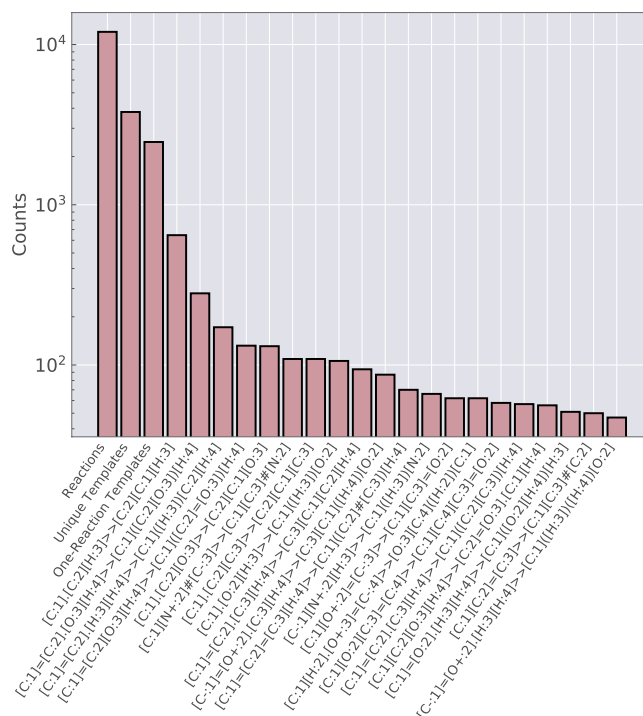
Fig. 4 Automatically extracted reaction templates. The reactions were grouped with very general reaction templates that only consider connectivity of atoms in the reactive center, atom identity, charge, aromaticity, and bond type. The top 20 templates are denoted with SMARTS strings[45].

- `make_opt_jobs.py:` Performs conformer searches and makes Q-Chem input files for optimization of reactant geometries based on the QM9 SMILES. The geometry optimizations themselves have to be performed with Q-Chem outside of the code, preferably in a massively parallel fashion on a supercomputer.
- `create_gsm_jobs.py:` Reads the geometry optimization outputs of the reactant optimizations, generates driving coordinates, and writes the files required for the GSM calculations. The GSM code has to be compiled separately[42]. The GSM calculations also have to be run separately and should produce output files with a `gsm#.out` format, where # corresponds to each reaction path.
- `create_prod_optfreq_jobs.py:` Reads the string endpoints from the successfully completed GSM calculations and writes the Q-Chem input files for the product optimizations.
- `create_ts_optfreq_jobs.py:` Extracts the TS geometries from the GSM output files, removes duplicate reactions using the output from the product optimizations, and writes the Q-Chem input files for additional TS optimizations.
- `extract_reactions.py:` Extracts the unique reactions using the reactant, product, and TS optimization outputs in the form of a comma-separated values file containing SMILES, activation energies, and enthalpies of reaction. Can also write the file path information of all relevant log files to the CSV output, which can be used to copy the log files for every reaction.
- `refine_reactants.py:` Writes Q-Chem input files for reoptimization of the reactants at the higher level of theory.
- `refine_products_and_ts.py:` Uses the same method as implemented in `extract_reactions.py` to extract reactions and write Q-Chem input files for the reoptimization of products and TSs at the higher level of theory. After running the Q-Chem jobs, `extract_reactions.py` can be run again to extract the high-level reactions.

If desired, the levels of theory and the reaction generation settings can be changed in the config folder.

Code availability
The code used to generate the data is freely available on GitHub under the MIT license[43]. Further details on how to use it to generate the data are given in the Usage Notes.

References
1. Sliwoski, G., Kothiwale, S., Meiler, J. & Lowe, E. W. Jr. Computational methods in drug discovery. *Pharmacol. Rev.* **66**, 334–395 (2014).
2. Cherkasov, A. *et al.* QSAR modeling: Where have you been? Where are you going to? *J. Med. Chem.* **57**, 4977–5010 (2014).

3. Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **51**, 1281–1289 (2018).
4. Pyzer-Knapp, E. O., Li, K. & Aspuru-Guzik, A. Learning from the Harvard Clean Energy Project: The use of neural networks to accelerate materials discovery. *Adv. Funct. Mater.* **25**, 6495–6502 (2015).
5. Wu, Z. *et al*. MoleculeNet: A benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
6. Yang, K. *et al*. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
7. Gao, C. W., Allen, J. W., Green, W. H. & West, R. H. Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms. *Comput. Phys. Commun.* **203**, 212–225 (2016).
8. Unsleber, J. P. & Reiher, M. The exploration of chemical reaction networks. *Annu. Rev. Phys. Chem.* **71**, 121–142 (2020).
9. Vereecken, L. *et al*. Perspective on mechanism development and structure-activity relationships for gas-phase atmospheric chemistry. *Int. J. Chem. Kinet.* **50**, 435–469 (2018).
10. Li, Y.-P., Han, K., Grambow, C. A. & Green, W. H. Self-evolving machine: A continuously improving model for molecular thermochemistry. *J. Phys. Chem. A* **123**, 2142–2152 (2019).
11. Grambow, C. A., Li, Y.-P. & Green, W. H. Accurate thermochemistry with small data sets: A bond additivity correction and transfer learning approach. *J. Phys. Chem. A* **123**, 5826–5835 (2019).
12. Coley, C. W. *et al*. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **10**, 370–377 (2019).
13. Schwaller, P. *et al*. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
14. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C-N cross-coupling using machine learning. *Science* **360**, 186–190 (2018).
15. Nielsen, M. K., Ahneman, D. T., Riera, O. & Doyle, A. G. Deoxyfluorination with sulfonyl fluorides: Navigating reaction space with machine learning. *J. Am. Chem. Soc.* **140**, 5004–5008 (2018).
16. Lawson, A. J., Swienty-Busch, J., Géoui, T. & Evans, D. The making of Reaxys—Towards unobstructed access to relevant chemistry information. In *The Future of the History of Chemical Information*, chap. 8, 127–148 (2014).
17. Mayfield, J., Lowe, D. & Sayle, R. Pistachio: Search and faceting of large reaction databases. Presentation at the *American Chemical Society National Meeting* (Washington, D.C., 2017).
18. Lowe, D. Chemical reactions from US patents (1976-Sep2016). *Figshare*, https://doi.org/10.6084/m9.figshare.5104873.v1 (2017).
19. Zádor, J. & Miller, J. A. Adventures on the C3H5O potential energy surface: OH + propyne, OH + allene and related reactions. *Proc. Combust. Inst.* **35**, 181–188 (2015).
20. Dewyer, A. L., Argüelles, A. J. & Zimmerman, P. M. Methods for exploring reaction space in molecular systems. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **8**, e1354 (2017).
21. Grambow, C. *et al*. Unimolecular reaction pathways of a γ-ketohydroperoxide from combined application of automated reaction discovery methods. *J. Am. Chem. Soc.* **140**, 1035–1048 (2018).
22. Zimmerman, P. M. Single-ended transition state finding with the growing string method. *J. Comput. Chem.* **36**, 601–611 (2015).
23. Ruddigkeit, L., Van Deursen, R., Blum, L. C. & Reymond, J. L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **52**, 2864–2875 (2012).
24. Landrum, G. RDKit: Open-source cheminformatics, http://rdkit.org (2006).
25. 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).
26. Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **32**, 1456–1465 (2011).
27. Lin, Y. S., Li, G. D., Mao, S. P. & Chai, J. D. Long-range corrected hybrid density functionals with improved dispersion corrections. *J. Chem. Theory Comput.* **9**, 263–272 (2013).
28. Shao, Y. *et al*. Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Mol. Phys.* **113**, 184–215 (2015).
29. Weigend, F. & Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **7**, 3297 (2005).
30. Grimme, S., Brandenburg, J. G., Bannwarth, C. & Hansen, A. Consistent structures and interactions by density functional theory with small atomic orbital basis sets. *J. Chem. Phys.* **143**, 054107 (2015).
31. Dasgupta, S. & Herbert, J. M. Standard grids for high-precision integration of modern density functionals: SG-2 and SG-3. *J. Comput. Chem.* **38**, 869–882 (2017).
32. Gonzalez, C. & Schlegel, H. B. Reaction path following in mass-weighted internal coordinates. *J. Phys. Chem.* **94**, 5523–5527 (1990).
33. Zimmerman, P. Reliable transition state searches integrated with the growing string method. *J. Chem. Theory Comput.* **9**, 3043–3050 (2013).
34. Henkelman, G. & Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **113**, 9978–9985 (2000).
35. Baker, J., Kessi, A. & Delley, B. The generation and use of delocalized internal coordinates in geometry optimization. *J. Chem. Phys.* **105**, 192–212 (1996).
36. Weininger, D. SMILES, a chemical language and information system: 1: Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
37. O'Boyle, N. M. *et al*. Open Babel: An open chemical toolbox. *J. Cheminform.* **3** (2011).
38. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.* **7** (2015).
39. Weinhold, F., Landis, C. R. & Glendening, E. D. What is NBO analysis and how is it useful? *Int. Rev. Phys. Chem.* **35**, 399–440 (2016).
40. Grambow, C. A., Pattanaik, L. & Green, W. H. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Zenodo*, https://doi.org/10.5281/zenodo.3581266 (2020).
41. Mardirossian, N. & Head-Gordon, M. Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals. *Mol. Phys.* **115**, 2315–2372 (2017).
42. Zimmerman, P. molecularGSM. *GitHub*, https://github.com/ZimmermanGroup/molecularGSM (2016).
43. Grambow, C. cgrambow/ard_gsm: Release version 1.0.0. *Zenodo* https://doi.org/10.5281/zenodo.3552859 (2019).
44. Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022 (2014).
45. Daylight Chemical Information Systems, Inc. SMARTS - A language for describing molecular patterns, https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (2019).

Acknowledgements

Author contributions

C.A.G. performed all potential energy surface exploration, quantum chemistry calculations, and SMILES and energy extraction. L.P. performed the reaction template analysis. All authors worked on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.H.G.

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.