www.nature.com/scientificdata

scientific data

OPEN

DATA DESCRIPTOR

# Zeo-1, a computational data set of zeolite structures

Leonid Komissarov & Toon Verstraelen

Fast, empirical potentials are gaining increased popularity in the computational fields of materials science, physics and chemistry. With it, there is a rising demand for high-quality reference data for the training and validation of such models. In contrast to research that is mainly focused on small organic molecules, this work presents a data set of geometry-optimized bulk phase zeolite structures. Covering a majority of framework types from the Database of Zeolite Structures, this set includes over thirty thousand geometries. Calculated properties include system energies, nuclear gradients and stress tensors at each point, making the data suitable for model development, validation or referencing applications focused on periodic silica systems.

## Background & Summary

Atomistic models are an essential tool for the prediction of thermodynamic, mechanical or biochemical properties of a substance. More recently, the use of pre-trained models has become increasingly popular due to their comparably low complexity and high accuracy on modern hardware[1–6]. In order for such models to perform well, their empirical parameters require fitting to high-quality reference data. Depending on the application, reference data are either experimental, or come from computationally more expensive *ab initio* calculations. Although there are already a handful of large computational data sets covering small organic molecules[7–9], such data is still scarce for larger periodic systems (*cf.* Materials Cloud Archive[10,11] or the NOMAD database[12,13]). Motivated by this fact, we present a quantum-chemical data set for zeolites. Zeolites are porous materials comprised of interconnected $SiO_4$ or $AlO_4$ tetrahedra. Their properties can be fine-tuned through synthesis of materials with specific pore size, or the inclusion of additional metal cation sites[14–17]. Because of their topology and synthetic flexibility, zeolites have various applications as adsorbents[18–20] and catalysts[17,21–23]. To this day, a myriad of different zeolite framework types is available experimentally, and many more hypothetical structures can be derived[24–26]. The documentation of fundamental zeolite framework types and derived materials has led to the publication of the well-known *Atlas of Zeolite Structures*[27] in several editions. The atlas lists each unique framework type by its three-letter-code, as assigned by the by the Structure Commission of the International Zeolite Association (IZA). Today, its contents are available online at the *Database of Zeolite Structures*[28], which we use as a source of initial structures for our data set. In this first installment, we include properties for 204 out of the currently available 256 zeolite framework types in the database (a total of 226 unique geometries when also considering derived materials). Our descriptor provides the complete optimization trajectories for each system with atomic positions, lattice vectors, atomic gradients and stress tensors at each step. We envision future extensions of the data set to focus on derived geometries, covering structural defects and host-guest interactions.

## Methods

Initial zeolite structures are collected from the public *Database of Zeolite Structures*[28] in the *Crystallographic Information File* (CIF) format, before conversion to the XYZ format with the Atomic Simulation Environment[29] (ASE) package. After selection of all systems with less than 301 atoms, each is manually filtered by removing redundant atom positions in case of fractional occupancies and adding missing hydrogen atoms where needed. Each structure's coordinates and cell parameters are energy-minimized with the periodic density functional code BAND[30], as implemented in the Amsterdam Modeling Suite[31] (AMS). The calculations are performed with the revPBE functional[32,33], a 'Small' frozen core and the double-$\zeta$ polarized (DZP) basis set. Grimme's D3(BJ) dispersion correction[34] is applied to all calculations. Previous research has shown that the selected level of theory can accurately reproduce zeolite geometries, albeit slightly overestimating the Si-O bond length (in the range of 2 pm) and smaller Si-O-X angles (in the range of 5 degrees) when compared to experimental results[35,36]. At the same time, dispersion-corrected functionals are generally more accurate when describing adsorption

Center for Molecular Modeling (CMM), Ghent University, Technologiepark-Zwijnaarde 46, B-9052, Ghent, Belgium. e-mail: leonid.komissarov@ugent.be; toon.verstraelen@ugent.be

| Data | Unit | Key | Array Shape |
|---|---|---|---|
| Atomic Numbers | — | numbers | $(R,)$ |
| Atomic Coordinates | Å | xyz | $(N, R, 3)$ |
| $x$-, $y$- and, $z$-Components of the Lattice Vectors | Å | lattice | $(N, 3, 3)$ |
| Energy | hartree | energy | $(N,)$ |
| Nuclear Gradients | hartree/bohr | gradients | $(N, R, 3)$ |
| Stress Tensors | atomic units | stress | $(N, 3, 3)$ |
| Hirshfeld Charges | atomic units | charges | $(R,)$ |

**Table 1.** Overview of the data structures stored in a .npz file. Each array can be accessed through the respective key. The variables $N$ and $R$ denote the number of geometry optimization steps and the system size respectively. Partial charges are only computed for the last geometry.

| Element | Occurrence |
|---|---|
| Si | 226 |
| O | 226 |
| H | 21 |
| Al | 12 |
| N | 4 |
| Ca | 4 |
| Ge | 3 |
| Li | 2 |
| Na | 2 |
| K | 2 |
| C | 2 |
| F | 1 |
| Be | 1 |
| Cs | 1 |
| Ba | 1 |

**Table 2.** Elemental occurrences in the complete data set. Counting all structures containing at least one atom of the listed element. Each element's isolated atomic energy is listed in hartree.

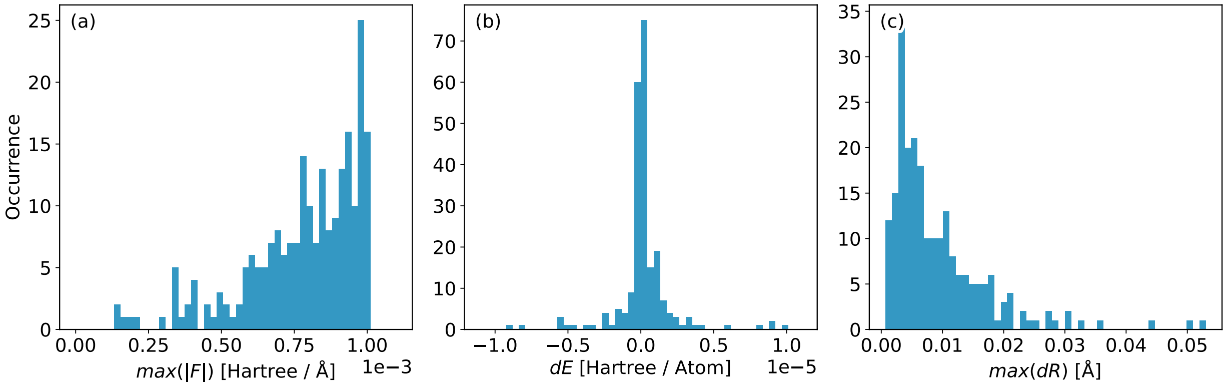

**Fig. 1** Distribution of convergence criteria at the last optimization step for all calculated systems in the data set. Showing (**a**) the highest absolute component of all nuclear gradients, (**b**) change in system energy and (**c**) highest relative atomic displacement.

processes[37–39]. For the optimization of the initial structures, geometry convergence criteria are left at their default values, namely 0.001 Hartree/Å, 0.00001 Hartree/Atom and 0.1 Å for atomic gradients, energy and atomic displacements respectively. We use a Quasi-Newton optimizer[40] in the delocalized coordinates space for the initial optimizations. Cases of problematic convergence are restarted with the FIRE[41] optimizer.

## Data Records

The data is made available at the Materials Cloud Archive[42]. Each system's trajectory is stored in an individual NumPy[43]. *npz* file. We describe the data types held in each file in Table 1, storing the complete geometry optimization trajectory, including atomic coordinates, system energies, nuclear gradients, lattice vectors and stress

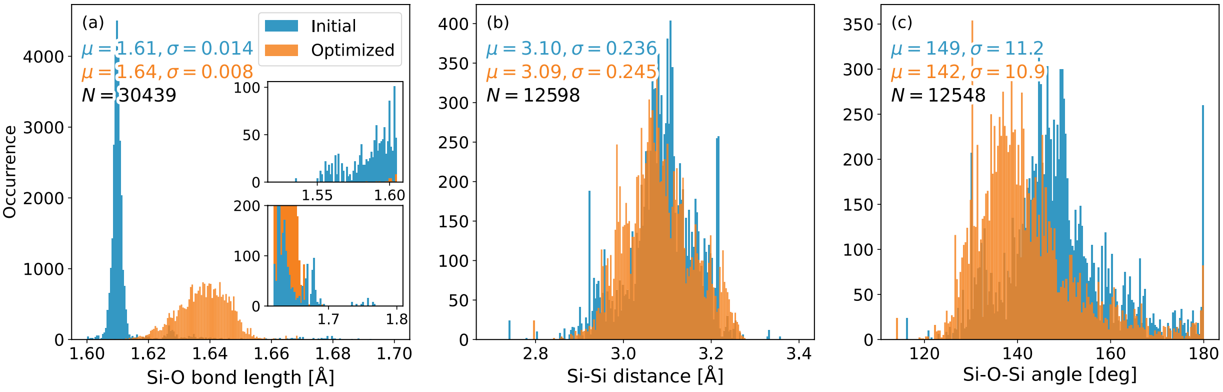

**Fig. 2** Distributions of (**a**) Si-O bond lengths, (**b**) Si-Si distances in the second coordination sphere and (**b**) Si-O-Si angles as calculated from all geometries in the data set. Blue and orange bars denote data from initial and optimized geometries, respectively. Mean $\mu$ and standard deviation $\sigma$ printed in the same color as the underlying data. $N$ denotes the total sample size.

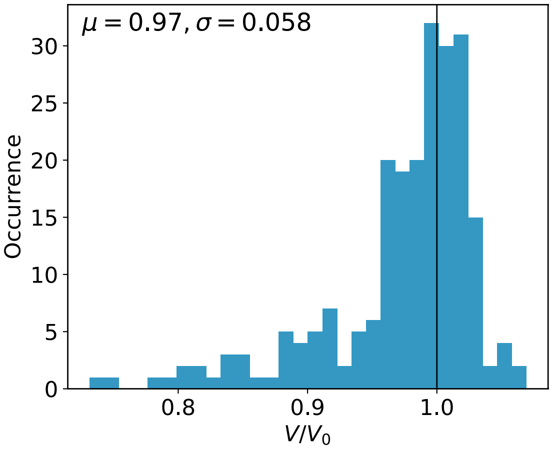

**Fig. 3** Distribution of relative cell volumes per system as the quotient of optimized-to-initial cell volumes. Values below 1 describe a shrinking cell as the optimization progresses. Black line marks $V/V_0 = 1$. Sample size is 226.

| Bond | Mean | Std. Dev. | Number of points |
|---|---|---|---|
| Si-O | 1.638 | 0.0085 | 30439 |
| H-O | 0.999 | 0.1174 | 266 |
| Al-O | 1.763 | 0.0133 | 234 |
| Ge-O | 1.795 | 0.0239 | 202 |
| Na-O | 2.473 | 0.0841 | 104 |
| C-C | 1.540 | 0.0049 | 100 |
| C-H | 1.100 | 0.0027 | 98 |
| K-O | 3.175 | 0.4809 | 61 |
| Ca-O | 2.469 | 0.0925 | 57 |
| N-H | 1.055 | 0.0913 | 50 |
| Si-K | 3.945 | 0.3196 | 41 |
| Cs-O | 3.429 | 0.2820 | 28 |
| Li-O | 1.970 | 0.0263 | 21 |
| Be-O | 1.669 | 0.0152 | 16 |
| Al-K | 3.625 | 0.1650 | 14 |
| C-N | 1.472 | 0.0037 | 10 |
| Ba-O | 2.903 | 0.1261 | 10 |

**Table 3.** Mean atomic bond length distributions and their standard deviations (std. dev.) in in ångström. Averaged over all geometry-optimized structures.

| Angle | Mean | Std. Dev. | Number of points |
|---|---|---|---|
| Si-O-Si | 148.7 | 11.2 | 12548 |
| Si-O-Al | 140.6 | 8.9 | 170 |
| Si-O-K | 106.8 | 8.8 | 81 |
| Si-O-Na | 112.8 | 15.1 | 64 |
| Si-O-Ge | 143.2 | 12.0 | 52 |
| Si-O-H | 110.7 | 7.9 | 40 |
| Si-O-Cs | 101.6 | 6.9 | 36 |
| Si-O-Ca | 118.5 | 16.6 | 19 |
| Si-O-Be | 129.9 | 0.2 | 16 |
| Si-O-Li | 112.7 | 4.2 | 8 |
| Si-O-Ba | 112.6 | 14.1 | 5 |

**Table 4.** Mean Si-O-R angle distributions and their standard deviations (std. dev.) in degrees. Averaged over all geometry-optimized structures.

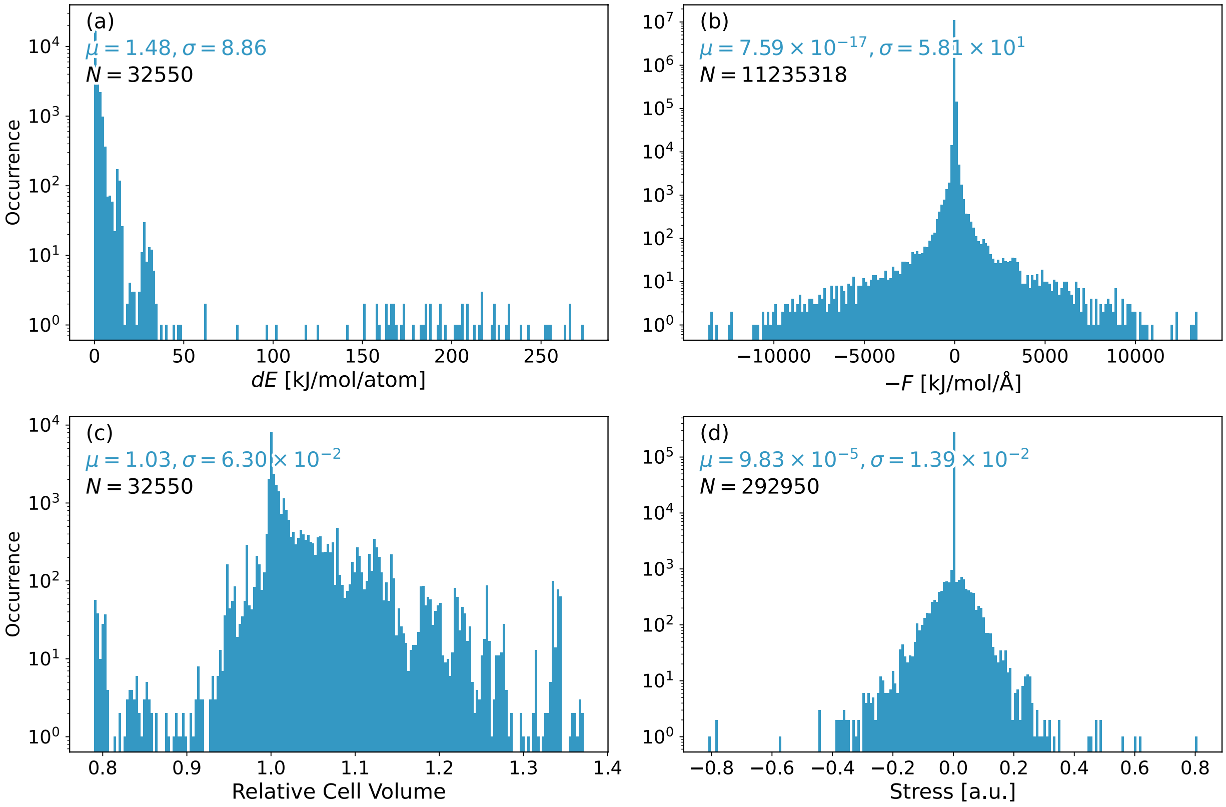

**Fig. 4** Distributions of physical quantities in the data set. Showing (**a**) energy differences per atom, relative to the respective energy of the optimized system; (**b**) atomic gradient components; (**c**) unit cell volumes, relative to the optimized system's volume; (**d**) stress tensor components. Data is printed on a logarithmic y-scale for a clear display of the distribution. Mean $\mu$ and standard deviation $\sigma$ printed in the same units as the underlying data. $N$ denotes the total sample size.

tensors for each geometry optimization step. Entries at the first position correspond to the input structure; the last position holds the data for the final, optimized structure. Hirshfeld partial charges[44] are provided for the final (optimized) geometries. Atomic coordinates and lattice vectors are stored in ångström, all other properties are stored in atomic units.

## Technical Validation

The complete data set includes geometry optimizations of 226 systems, resulting in a total of 32550 geometries. System sizes range between 15 and 334 atoms (mean: 126). We illustrate the convergence of all reference calculations in Fig. 1, showing that all optimized systems are well within the defined convergence criteria. Elemental occurrences in the data set are listed in Table 2. Si-O, Si-Si distances as well as Si-O-Si angles are presented in

Fig. 2 as the most prominent geometrical descriptors. As most of the initial structures from the IZA database are idealized geometries[45], a sharp mean for the Si-O bond distance can be observed at roughly 161 pm (Fig. 2a, blue histogram). Long tails in the distribution vanish and the mean is shifted towards approximately 164 pm when considering geometry-optimized structures (Fig. 2a, orange histogram). Considering the Si-O-Si angles, a slight shift towards smaller values is observed (mean of 149 vs. 142 degrees, Fig. 2c). Both effects have been previously reported by Fischer *et al.*[35,36] and are inherent to the selected level of theory. Distributions of the Si-Si distances in the second coordination sphere do not shift significantly when comparing initial and optimized geometries (Fig. 2b). Relative changes in the cell volumes are presented in Fig. 3 as the ratio of each system's optimized-to-initial volume. Values below 1 translate to a shrinking unit cell as the optimization progresses. Overall, the geometrical descriptors are in good agreement with experimental data[46–51]. Additional averages for bond distances and angles are summarized in Tables 3, 4 respectively. Distributions of energies, atomic gradients, cell volumes and stress tensors are depicted in Fig. 4. As expected from geometry optimization trajectories, all properties have – with the exception of relative cell volumes – a distinct mean close to zero. Structures close to the initial input geometries contribute to the relatively high standard deviations. Evaluation of the relative cell volumes shows a shifted distribution, with roughly 76% of all structures having a larger volume than their respective optimized geometry. A detailed overview of all calculated structures, sorted by their IZA three-letter-code, the system size and number of iterations is provided in Online Table 1.

## Usage Notes

No data points were filtered as outliers with regards to the distributions of chemical properties (see. Figure 4). Consecutive structures from the same optimization trajectory will be autocorrelated. The data repository provides an interactive plotting script, displaying the system energy, maximum absolute component of the nuclear gradients and the cell volume at every iteration step for each structure. This requires the Bokeh[52] (v. 2.3.1) package for Python to be installed. SHA-1 hash sums are provided for each file to guarantee data integrity, as well as an example input script for a calculation with BAND. Naming conventions: Derived materials are referred to by their IZA three-letter-code, *e.g.* H-EU-12 is tabulated as ETL_0. Leading non-alphabetical characters have been removed, *e.g.* *-ITN is tabulated as ITN.

## Code availability

Downloads of the Atomic Simulation Environment[29] (v. 3.21.1) and NumPy[43] (v. 1.20.1) packages for Python are freely available. Amsterdam Modeling Suite[31] (v. 2020.203, r92091) is a commercial software, for which a free trial may be requested at www.scm.com.

Received: 2 September 2021; Accepted: 14 January 2022;
Published online: 22 February 2022

## References

1. Smith, J. S. *et al.* Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature Communications* **10** (2019).
2. 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).
3. Shao, Y., Hellström, M., Mitev, P. D., Knijff, L. & Zhang, C. PiNN: A python library for building atomic neural networks of molecules and materials. *Journal of Chemical Information and Modeling* **60**, 1184–1193 (2020).
4. Satorras, V. G., Hoogeboom, E. & Welling, M. E(n) equivariant graph neural networks. Preprint at https://arxiv.org/abs/2102.09844 (2021).
5. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
6. Kondratyuk, N. *et al.* Performance and scalability of materials science and machine learning codes on the state-of-art hybrid supercomputer architecture. In Voevodin, V. & Sobolev, S. (eds.) *Supercomputing*, 597–609 (Springer International Publishing, Cham, 2019).
7. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules. *Scientific Data* **4**, 170193 (2017).
8. Smith, J. S. *et al.* The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Scientific Data* **7**, 134 (2020).
9. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **1**, 140022 (2014).
10. *Materials Cloud Archive*. https://archive.materialscloud.org/ (2021).
11. Talirz, L. *et al.* Materials cloud, a platform for open computational science. *Scientific Data* **7**, 299 (2020).
12. *NOMAD Laboratory*. https://nomad-lab.eu/ (2021).
13. Draxl, C. & Scheffler, M. Nomad: The fair concept for big data-driven materials science. *MRS Bulletin* **43**, 676–682 (2018).
14. Davis, M. E. & Lobo, R. F. Zeolite and molecular sieve synthesis. *Chemistry of Materials* **4**, 756–768 (1992).
15. Cundy, C. S. Microwave techniques in the synthesis and modification of zeolite catalysts. a review. *Collection of Czechoslovak Chemical Communications* **63**, 1699–1723 (1998).
16. Chen, L.-H. *et al.* Hierarchically structured zeolites: synthesis, mass transport properties and applications. *Journal of Materials Chemistry* **22**, 17381 (2012).
17. Moliner, M., Martnez, C. & Corma, A. Multipore zeolites: Synthesis and catalytic applications. *Angewandte Chemie International Edition* **54**, 3560–3579 (2015).
18. Ozekmekci, M., Salkic, G. & Fellah, M. F. Use of zeolites for the removal of H2S: a mini-review. *Fuel Processing Technology* **139**, 49–60 (2015).
19. Papaioannou, D., Katsoulos, P., Panousis, N. & Karatzias, H. The role of natural and synthetic zeolites as feed additives on the prevention and/or the treatment of certain farm animal diseases: a review. *Microporous and Mesoporous Materials* **84**, 161–170 (2005).
20. Dehghan, R. & Anbia, M. Zeolites for adsorptive desulfurization from fuels: a review. *Fuel Processing Technology* **167**, 99–116 (2017).

21. Derouane, E. *et al.* The acidity of zeolites: concepts, measurements and relation to catalysis: A review on experimental and theoretical methods for the study of zeolite acidity. *Catalysis Reviews* **55**, 454–515 (2013).
22. Weitkamp, J. Zeolites and catalysis. *Solid State Ionics* **131**, 175–188 (2000).
23. Corma, A. State of the art and future challenges of zeolites as catalysts. *Journal of Catalysis* **216**, 298–312 (2003).
24. Treacy, M. M. J., Randall, K. H., Rao, S., Perry, J. A. & Chadi, D. J. Enumeration of periodic tetrahedral frameworks. *Zeitschrift für Kristallographie - Crystalline Materials* **212**, 768–791 (1997).
25. Treacy, M. M. J. & Foster, M. *Atlas of Prospective Zeolite Structures*. http://www.hypotheticalzeolites.net/ (2021).
26. Pophale, R., Cheeseman, P. A. & Deem, M. W. A database of new zeolite-like materials. *Phys. Chem. Chem. Phys.* **13**, 12407–12412 (2011).
27. Baerlocher, C., McCusker, L. & Olson, D. *Atlas of Zeolite Framework Types* (Published on behalf of the Structure Commission of the International Zeolite Association by Elsevier, 2007).
28. Baerlocher, C. & McCusker, L. *Database of Zeolite Structures*. http://www.iza-structure.org/databases/.
29. Larsen, A. H. *et al.* The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter* **29**, 273002 (2017).
30. te Velde, G. & Baerends, E. J. Precise density-functional method for periodic structures. *Phys. Rev. B* **44**, 7888–7903 (1991).
31. Rüger *et al. Amsterdam Modeling Suite*. https://scm.com (2019).
32. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Physical Review Letters* **77**, 3865–3868 (1996).
33. Zhang, Y. & Yang, W. Comment on "generalized gradient approximation made simple". *Physical Review Letters* **80**, 890–890 (1998).
34. Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry* **32**, 1456–1465 (2011).
35. Fischer, M., Evers, F. O., Formalik, F. & Olejniczak, A. Benchmarking dft-gga calculations for the structure optimisation of neutral-framework zeotypes. *Theoretical Chemistry Accounts* **135** (2016).
36. Fischer, M. & Angel, R. J. Accurate structures and energetics of neutral-framework zeotypes from dispersion-corrected dft calculations. *The Journal of Chemical Physics* **146**, 174111 (2017).
37. Göltl, F., Grüneis, A., Bučko, T. & Hafner, J. Van der waals interactions between hydrocarbon molecules and zeolites: periodic calculations at different levels of theory, from density functional theory to the random phase approximation and mÃ¸ller-plesset perturbation theory. *The Journal of Chemical Physics* **137**, 114111 (2012).
38. Rehak, F. R., Piccini, G., Alessio, M. & Sauer, J. Including dispersion in density functional theory for adsorption on flat oxide surfaces, in metal—organic frameworks and in acidic zeolites. *Physical Chemistry Chemical Physics* **22**, 7577–7585 (2020).
39. Stanciakova, K., Louwen, J. N., Weckhuysen, B. M., Bulo, R. E. & Göltl, F. Understanding water—zeolite interactions: on the accuracy of density functionals. *The Journal of Physical Chemistry C* **125**, 20261–20274 (2021).
40. Swart, M. & Bickelhaupt, F. M. Optimization of strong and weak coordinates. *International Journal of Quantum Chemistry* **106**, 2536–2544 (2006).
41. Bitzek, E., Koskinen, P., Gähler, F., Moseler, M. & Gumbsch, P. Structural relaxation made simple. *Physical Review Letters* **97** (2006).
42. Komissarov, L. & Verstraelen, T. *Zeo-1: a computational data set of zeolite structures*. *Materials Cloud Archive* https://doi.org/10.24435/materialscloud:cv-zd (2021).
43. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
44. Hirshfeld, F. L. Bonded-atom fragments for describing molecular charge densities. *Theoret. Chim. Acta* **44**, 129–138 (1977).
45. Baerlocher, C., Hepp, A. & Meier, W. Dls-76, a fortran program for the simulation of crystal structures by geometric refinement. *Institut fur Kristallographie und Petrographie, ETH, Zurich, Switzerland* (1978).
46. Pettifer, R., Dupree, R., Farnan, I. & Sternberg, U. NMR determinations of Si–O–Si bond angle distributions in silica. *Journal of Non-Crystalline Solids* **106**, 408–412 (1988).
47. Mauri, F., Pasquarello, A., Pfrommer, B. G., Yoon, Y.-G. & Louie, S. G. Si-O-Si bond-angle distribution in vitreous silica from first-principles 29 Si NMR analysis. *Physical Review B* **62**, R4786 (2000).
48. Wragg, D. S., Morris, R. E. & Burton, A. W. Pure silica zeolite-type frameworks: A structural analysis. *Chemistry of Materials* **20**, 1561–1570 (2008).
49. Ramdas, S. & Klinowski, J. A simple correlation between isotropic 29 si-nmr chemical shifts and t–o–t angles in zeolite frameworks. *Nature* **308**, 521–523 (1984).
50. Antao, S. M. Quartz: structural and thermodynamic analyses across the $\alpha \leftrightarrow \beta$ transition with origin of negative thermal expansion (NTE) in $\beta$ quartz and calcite. *Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials* **72**, 249–262 (2016).
51. OKeeffe, M. & Hyde, B. G. On Si–O –Si configurations in silicates. *Acta Crystallographica Section B* **34**, 27–32 (1978).
52. Bokeh Development Team. *Bokeh: Python library for interactive visualization*. https://bokeh.pydata.org/en/latest/ (2021).

## Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 814143. T.V. acknowledges funding of the research board of Ghent University. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government–department EWI.

## Author contributions

L.K. designed and performed the study. Both authors wrote the manuscript. T.V. oversaw the project.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to L.K. or T.V.

**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.