Provenance of *ab initio* data for polymorphs of benzene, glycine, and succinic acid

Edgar A Engel*
TCM Group, Cavendish Laboratory, University of Cambridge, J. J. Thomson Avenue, Cambridge CB3 0HE, United Kingdom

Venkat Kapil
Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW,UK and Laboratory of Computational Science and Modeling, Institut des Matériaux, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
(Dated: March 18, 2021)

The following describes the provenance of the *ab initio* training, validation, and test data underlying the neural-network potentials and machine-learning models for crystalline benzene, succinic acid, and glycine discussed in the following publications:

(1) V. Kapil and E. A. Engel, "A complete description of thermodynamic stabilities of molecular crystals", arXiv:2102.13598 [cond-mat.mtrl-sci]

(2) R. K. Cersonsky, B. A. Helfrecht, E. A. Engel and M. Ceriotti, "Improving Sample and Feature Selection with Principal Covariates Regression", arXiv:2012.12253 [physics.chem-ph]

(3) E. A. Engel, A. Anelli, V. Kapil and M. Ceriotti, "Nuclear quantum effects on NMR chemical shieldings and their impact on NMR crystallography", in preparation

I. PROVENANCE

The data is based on the Cambridge Structural Database [1] entries for forms I [2] and II [3] and the hypothetical high-pressure forms I_{hp} and V' of benzene [4], the α [5], β [6], γ [5], and δ [5] polymorphs of glycine, and the α [7] and β [8] polymorphs of succinic acid.

Generation of reference configurations

In a first step, snapshot configurations of each polymorph were generated using density-functional tight-binding (DFTB) simulations, which provide an affordable means of sampling configurations representative of the target thermodynamic ensemble corresponding to a *ab initio* density-functional-theory (DFT) description. All DFTB simulations were performed using the i-Pi force engine [9] to drive the DFTB+ code [10] and used the DFTB3/3OB[11, 12] parametrisation and the D3BJ[13] dispersion correction. Sample i-Pi inputs for the simulations are provided. Specifically, for each polymorph a classical, Langevin-thermostatted temperature replica-exchange molecular dynamics (MD) [14] simulation in the NVT with 12 replicas at quadratically increasing temperatures between 300 K and 2955 K was performed using the experimental unit cell parameters. Additionally, PILE-L-thermostatted [15] path-integral (PI) MD simulation using the Suzuki-Chin fourth-order splitting [16] in the NVT ensemble at 300 K were performed for the experimental unit cell as well as a range of perturbed simulation cells. For each polymorph the experimental unit cell was *isotropically* scaled to change the cell volume by ± 0.5, ± 1, ± 1.5, ± 2, ± 2.5, ± 5, ± 5, ± 10, and $+20\%$, respectively. For each polymorph the unit cell was further *anisotropically* perturbed by independently rescaling the cell lengths by ± 2, ± 2.5, ± 5, and $\pm 10\%$, and the cell angles by ± 2, ± 2.5, ± 5, ± 10, and $\pm 20\%$. For each compound, a set of decorrelated configurations was collected by extracting snapshots at 25 fs intervals from the trajectories for all perturbed cells of all polymorphs. Separately for both validation and testing, 272 benzene, 336 glycine, and 240 succinic acid configurations were extracted at random from the decorrelated set corresponding to the respective experimental unit cells. The remaining configurations were ordered by structural diversity by means of farthest-point sampling (FPS) [17–19] and the 55,000 most distinct benzene and the 30,000 most distinct glycine and succinic acid configurations, respectively, were retained for training. For the purpose of FPS, Euclidean pairwise distances between configurations were measured using the smooth overlap of atomic positions (SOAP) description [20] in its radially-scaled variant [21] with the following hyperparameters: $n = 12$ radial basis functions, $l = 9$ angular spherical harmonic basis functions, a Gaussian width $\sigma = 0.275$ Å for all chemical species, a cut-off radius of $r_c = 8.0$ Å, a radial-scaling exponent of $r_{\text{exp}} = 4.5$, and an onset radius for radial scaling of $r_s = 2.5$.

Calculation of *ab initio* energies and forces

* Email: eae32@cam.ac.uk

For the training, test, and validation sets, *ab initio* DFT reference energies and forces were calculated using Quantum Espresso v6.3 [22]. The calculations were performed with the semi-local PBE exchange-correlation functional [23], the Tkatchenko-Scheffler (TS) dispersion correction [24], optimised norm-conserving Vanderbilt pseudopotentials from Ref. [25], a Monkhorst-Pack k-point grid [26] with a maximum spacing of $0.06 \times 2\pi$ Å^{-1}, and a plane-wave energy cut-off of 100 Rydberg for the wavefunction.

For the 1,800 most distinct benzene, the 3,600 most distinct glycine, and the 1,800 most distinct succinic acid configurations, as well as all test and validation sets, additional hybrid-functional DFT energies and forces were calculated using FHI-AIMS [27–29]. The calculations were performed with the hybrid PBE0 functional [30, 31] and the MBD dispersion correction [32, 33] (PBE0-MBD), using the same Monkhorst-Pack k-point grids and the standard FHI-AIMS "intermediate" basis sets.

II. AVAILABLE DATA AND FORMAT

The data are provided in lib-atom extended xyz format. Cell parameters and atomic positions are given in units of Angstrom. Energies and forces are provided in units of eV and eV/ Å. Stresses are provided in units of eV/ Å^3. The following provides a brief glossary of the acronyms used in naming the provided data files.

FPS	: farthest point sampling [17–19].
	DFT codes
QE	: Quantum Espresso [22].
AIMS	: FHI-AIMS code [27–29].
	exchange-correlation functionals
PBE	: semi-local PBE functional [23].
PBE0	: hybrid PBE0 functional [30, 31].
	dispersion corrections
TS	: Tkatchenko-Scheffler [24].
MBD	: many-body dispersion [32, 33].

The following data are available:

- **benzene_train_FPS_QE_PBE_TS.xyz**
  55,000 FPS-ordered configurations corresponding to benzene forms I, II, I_hp, and V' and their associated PBE-TS total energies, atomic forces, and stress tensors.

- **benzene_val_QE_PBE_TS.xyz**
  1000 configurations corresponding to benzene forms I, II, I_hp, and V' and their associated PBE-TS total energies, atomic forces, and stress tensors.

- **benzene_test_QE_PBE_TS.xyz**
  1000 configurations corresponding to benzene forms I, II, I_hp, and V' and their associated PBE-TS total energies, atomic forces, and stress tensors.

- **succinic_acid_train_FPS_QE_PBE_TS.xyz**
  30,000 FPS-ordered configurations corresponding to α and β-succinic acid and their associated PBE-TS total energies, atomic forces, and stress tensors.

- **succinic_acid_val_QE_PBE_TS.xyz**
  500 configurations corresponding to α and β-succinic acid and their associated PBE-TS total energies, atomic forces, and stress tensors.

- **succinic_acid_test_QE_PBE_TS.xyz**
  500 configurations corresponding to α and β-succinic acid and their associated PBE-TS total energies, atomic forces, and stress tensors.

- **glycine_train_FPS_QE_PBE_TS.xyz**
  30,000 FPS-ordered configurations corresponding to α, β, γ, and δ-glycine and their associated PBE-TS total energies, atomic forces, and stress tensors.

- **glycine_val_QE_PBE_TS.xyz**
  500 configurations corresponding to α, β, γ, and δ-glycine and their associated PBE-TS total energies, atomic forces, and stress tensors.

- **glycine_test_QE_PBE_TS.xyz**
  500 configurations corresponding to α, β, γ, and δ-glycine and their associated PBE-TS total energies, atomic forces, and stress tensors.

- **benzene_train_FPS_AIMS_PBE0_MBD.xyz**
  1,800 FPS-ordered configurations corresponding to benzene forms I, II, I_hp, and V' and their associated PBE-TS total energies and atomic forces. Stress tensors are available for 1,000 of the 1,800 configurations.

- **benzene_val_AIMS_PBE0_MBD.xyz**
  200 configurations corresponding to benzene forms I, II, I_hp, and V' and their associated PBE-TS total energies and atomic forces.

- **benzene_test_AIMS_PBE0_MBD.xyz**
  200 configurations corresponding to benzene forms I, II, I_hp, and V' and their associated PBE-TS total energies and atomic forces.

- **succinic_acid_train_FPS_AIMS_PBE0_MBD.xyz**
  1,800 FPS-ordered configurations corresponding to α and β-succinic acid and their associated PBE-TS total energies and atomic forces. Stress tensors are available for 1,000 of the 1,800 configurations.

- **succinic_acid_val_AIMS_PBE0_MBD.xyz**
  200 configurations corresponding to α and β-succinic acid and their associated PBE-TS total energies and atomic forces.

- **succinic_acid_test_AIMS_PBE0_MBD.xyz**
  200 configurations corresponding to α and β-succinic acid and their associated PBE-TS total energies and atomic forces.

- **glycine_train_FPS_AIMS_PBE0_MBD.xyz**
  3,600 FPS-ordered configurations corresponding to α, β, γ, and δ-glycine and their associated PBE-TS total energies and atomic forces. Stress tensors are available for 2800 of the 3600 configurations.

- **glycine_val_AIMS_PBE0_MBD.xyz**
  200 configurations corresponding to α, β, γ, and δ-glycine and their associated PBE-TS total energies and atomic forces.

- **glycine_test_AIMS_PBE0_MBD.xyz**
  200 configurations corresponding to α, β, γ, and δ-glycine and their associated PBE-TS total energies and atomic forces.

- **input_REMD.xml**
  Sample i-Pi input file for a classical temperature replica-exchange molecular dynamics simulation in the NVT ensemble.

- **input_PIMD.xyz**
  Sample i-Pi input file for a path integral molecular dynamics simulation in the NVT ensemble.

[1] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward, Acta Crystallographica B **72**, 171 (2016).
[2] A. Katrusiak, M. Podsiadło, and A. Budzianowski, Crystal Growth and Design **10**, 3461 (2010).
[3] A. Budzianowski and A. Katrusiak, Acta Crystallographica B **62**, 94 (2006).
[4] E. Schneider, L. Vogt, and M. E. Tuckerman, Acta Crystallographica B **72**, 542 (2016).
[5] A. Dawson, D. R. Allan, S. A. Belmonte, S. J. Clark, W. I. F. David, P. A. McGregor, S. Parsons, C. R. Pulham, and L. Sawyer, Crystal Growth and Design **5**, 1415 (2005).
[6] E. V. Boldyreva, T. N. Drebushchak, and E. S. Shutova, Zeitschrift fuer Kristallographie - Crystalline Materials **218**, 366 (2003).
[7] I. M. Dodd, S. J. Maginn, M. M. Harding, and R. J. Davey, (1998), cSD communication.
[8] J.-L. Leviel, G. Auvert, and J.-M. Savariault, Acta Crystallographica B **37**, 2185 (1981).
[9] V. Kapil, M. Rossi, O. Marsalek, R. Petraglia, Y. Litman, T. Spura, B. Cheng, A. Cuzzocrea, R. H. Meißner, D. M.Wilkins, B. A. Helfrecht, P. Juda, S. P. Bienvenue, W. Fang, J. Kessler, I. Poltavsky, S. Vandenbrande, J. Wieme, and M. Ceriotti, Computer Physics Communications **236**, 214 (2018).
[10] B. Aradi, B. Hourahine, and T. Frauenheim, Journal of Physical Chemistry A **111**, 5678 (2007).
[11] M. Gaus, A. Goez, and M. Elstner, Journal of Chemistry: Theory and Computation **9**, 338 (2013).
[12] M. Gaus, X. Lu, M. Elstner, and Q. Cui, Journal of Chemistry: Theory and Computation **10**, 1518 (2014).
[13] S. Grimme, S. Ehrlich, and L. Goerigk, Journal of Computational Chemistry **32**, 1456 (2011).
[14] R. Petraglia, A. Nicolaï, M. D. Wodrich, M. Ceriotti, and C. Corminboeuf, J. Comput. Chem. **37**, 83 (2015).
[15] M. Ceriotti, M. Parrinello, T. E. Markland, and D. E. Manolopoulos, Journal of Chemical Physics **133**, 124104 (2010).
[16] V. Kapil, J. Behler, and M. Ceriotti, Journal of Chemical Physics **145**, 234103 (2016).
[17] Y. Eldar, M. Lindenbaum, M. Porat, and Y. Y. Zeevi, IEEE Transactions on Image Processing **6**, 1305 (1997).
[18] M. Ceriotti, G. A. Tribello, and M. Parrinello, Journal of Chemical Theory and Computation **9**, 1521 (2013).
[19] R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander, ACM Trans. Knowl. Discov. Data **10**, 5 (2015).
[20] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).
[21] M. J. Willatt, F. Musil, and M. Ceriotti, Phys Chem Chem Phys **20**, 29661 (2018).
[22] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. Dal Corso, S. Fabris, G. Fratesi, S. de Gironcoli, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo,

G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch, Journal of Physics: Condensed Matter **21**, 395502 (2009).
[23] J. P. Perdew, K. Burke, and M. Ernzerhof, Physical Review Letters **77**, 3865 (1996).
[24] A. Tkatchenko and M. Scheffler, Physical Review Letters **102**, 073005 (2009).
[25] M. Schlipf and F. Gygi, Computer Physics Communications **196**, 36 (2015).
[26] H. J. Monkhorst and J. D. Pack, Physical Review B **13**, 5188 (1976).
[27] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, Computer Physics Communications **180**, 2175 (2009).
[28] X. Ren, P. Rinke, V. Blum, J. Wieferink, A. Tkatchenko, A. Sanfilippo, K. Reuter, and M. Scheffler, New Journal of Physics **14**, 053020 (2012).
[29] S. Levchenko, X. Ren, J. Wieferink, R. Johanni, P. Rinke, V. Blum, and M. Scheffler, Computer Physics Communications **192**, 60 (2015).
[30] J. P. Perdew, M. Ernzerhof, and K. Burke, Journal of Chemical Physics **105**, 9982 (1996).
[31] C. Adamo and V. Barone, Journal of Chemical Physics **110**, 6158 (1999).
[32] A. Tkatchenko, R. A. Di Stasio Jr, R. Car, and M. Scheffler, Physical Review Letters **108**, 236402 (2012).
[33] A. Ambrosetti, A. M. Reilly, R. A. Di Stasio Jr, and A. Tkatchenko, Journal of Chemical Physics **140**, 018A508 (2014).