

Extracting Crystal Chemistry from Amorphous Carbon Structures

Volker L. Deringer,^{*,[a, b]} Gábor Csányi,^[a] and Davide M. Proserpio^{*,[c, d]}

Dedicated to Professor Roald Hoffmann on the Occasion of his 80th Birthday

Carbon allotropes have been explored intensively by ab initio crystal structure prediction, but such methods are limited by the large computational cost of the underlying density functional theory (DFT). Here we show that a novel class of machine-learning-based interatomic potentials can be used for random structure searching and readily predicts several hitherto unknown carbon allotropes. Remarkably, our model draws structural information from liquid and amorphous carbon exclusively, and so does not have any prior knowledge of crystalline phases: it therefore demonstrates true transferability, which is a crucial prerequisite for applications in chemistry. The method is orders of magnitude faster than DFT and can, in principle, be coupled with any algorithm for structure prediction. Machine-learning models therefore seem promising to enable large-scale structure searches in the future.

Exploring structural space—of allotropes, polymorphs, materials—is today not only done experimentally but also by computational techniques and in an increasingly automated fashion.^[1] Indeed, with advanced algorithms and high-performance computing centres available, ab initio crystal structure prediction methods revealed novel and intriguing structures of elements and compounds, including stoichiometric compositions and coordination modes that would not have been expected from

textbook knowledge. Many of these predictions were subsequently validated by experiments.^[2]

Among the elements, carbon is one of the structurally most diverse, and naturally has long been the target of crystal-chemical considerations^[3] and later of structure-searching algorithms. New carbon allotropes have been predicted using practically every computational method available, including ab initio random structure searching (AIRSS),^[4] genetic algorithms,^[5] particle swarm optimization,^[6] metadynamics,^[7] and minima hopping.^[8] A recent, critical survey of the field is in Ref. [9].

Despite their predictive power, ab initio structure searches are inherently limited by the underlying computational workhorse, most commonly density-functional theory (DFT), which becomes prohibitively expensive for larger system sizes. To overcome the latter, more general problem, a novel class of interatomic potentials based on machine learning (ML) is currently emerging in the solid-state theory communities.^[10] Such ML potentials are trained on DFT or other quantum-mechanical data, and in doing so provide a high-dimensional fit of the potential-energy surface. These potentials enable simulations that can come close to DFT accuracy, but are faster by many orders of magnitude; they are still much slower than established empirical force fields, but the trade-off is often worthwhile. For example, an ML-based neural-network potential enabled realistic, atomic-scale insight into the graphite-diamond transition, learning from DFT computations on structural snapshots taken from graphite, diamond, and intermediates.^[11] It was recently suggested that ML potentials might be beneficial for structure searches.^[12]

Herein, we show that structural information from liquid and amorphous forms of carbon can be harnessed, via machine learning, to guide searches for crystalline phases (Figure 1). This serves as proof-of-concept that ML models, if properly trained, can indeed be used for applications in solid-state chemistry, including the exploration of (previously unknown) structural space.

To validate our approach, we performed a numerical experiment, starting with a set of fully DFT-driven structure searches^[1d] using 1,000 randomized unit cells that each contained eight carbon atoms. DFT relaxation of these cells readily identified diamond and graphite, and also several less stable structures with mixed coordination numbers, all as expected (Figure 2a). We then performed Gaussian approximation potential (GAP)-driven searches, starting from the same initial configurations and probing how close the results would come to DFT. Initially, this led to a set of structures slightly higher in energies than the reference data, but this can be easily remedied by

[a] Dr. V. L. Deringer, Prof. G. Csányi
Engineering Laboratory
University of Cambridge
Trumpington Street, Cambridge CB2 1PZ (United Kingdom)
E-mail: vld24@cam.ac.uk

[b] Dr. V. L. Deringer
Department of Chemistry
University of Cambridge
Lensfield Road, Cambridge CB2 1EW (United Kingdom)

[c] Prof. D. M. Proserpio
Università degli Studi di Milano
Dipartimento di Chimica, Milano (Italy)
E-mail: davide.proserpio@unimi.it

[d] Prof. D. M. Proserpio
Samara Center for Theoretical Materials Science (SCTMS)
Samara University, Samara (Russia)

Supporting Information and the ORCID identification number(s) for the author(s) of this article can be found under:
<http://dx.doi.org/10.1002/cphc.201700151>.

© 2017 The Authors. Published by Wiley-VCH Verlag GmbH & Co. KGaA. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

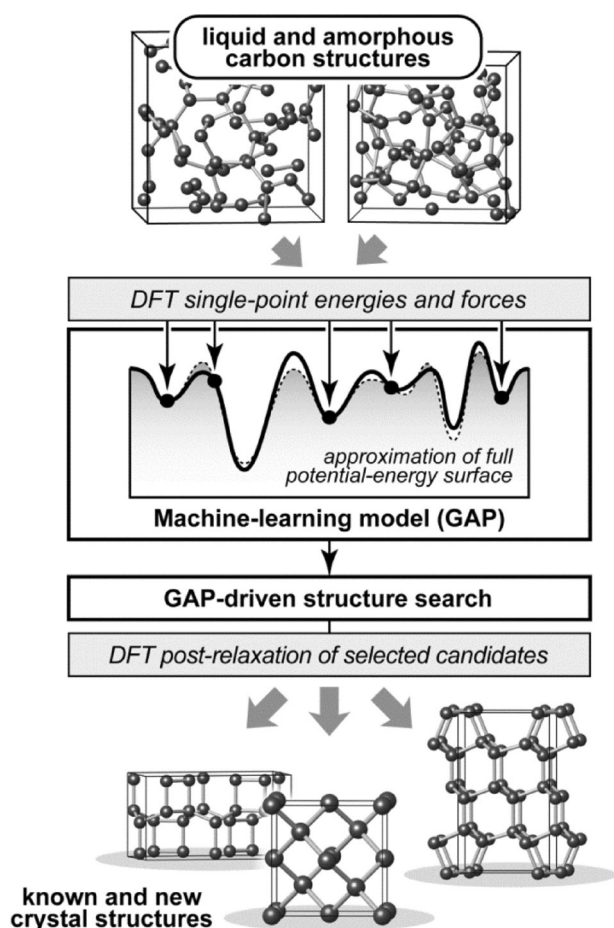


Figure 1. Flowchart of the strategy employed in the present work. Starting from liquid and amorphous carbon structures, we generate a Gaussian approximation potential (GAP) similar to that in Ref. [13] but here excluding any crystalline training data on purpose. We then use this for random structure searching,^[14] and subsequently re-relax suitable candidate structures with DFT (see text).

a subsequent DFT relaxation (Figure 2a, left). Likewise, the distribution of optimized volumes is similar for the DFT- and GAP-based procedure (right). This justifies our strategy: we perform large numbers of GAP-driven searches, and then re-relax only the most promising candidates using DFT.

The main result of this work is summarized in Figure 2b. We have here performed a large-scale search for allotropes with fourfold coordination exclusively, but we stress that our GAP model can describe mixed coordination environments just as well.^[13] Our search yielded 197 distinct carbon networks which were classified according to their topology,^[14] they were checked against the Samara Carbon Allotrope Database (SACADA; Ref. [9]) and furthermore against other topological nets as collected in ToposPro TTD;^[14] some of these were seen in zeolites or metal-organic frameworks but not in carbon allotropes. These structures are considered known and here referred to as such.

In addition, our search returned 150 possible allotropes that are neither known to SACADA nor from other topology databases; of these, 52 are no more than 0.3 eV per atom ($\approx 30 \text{ kJ mol}^{-1}$) above diamond in their DFT-computed energy.

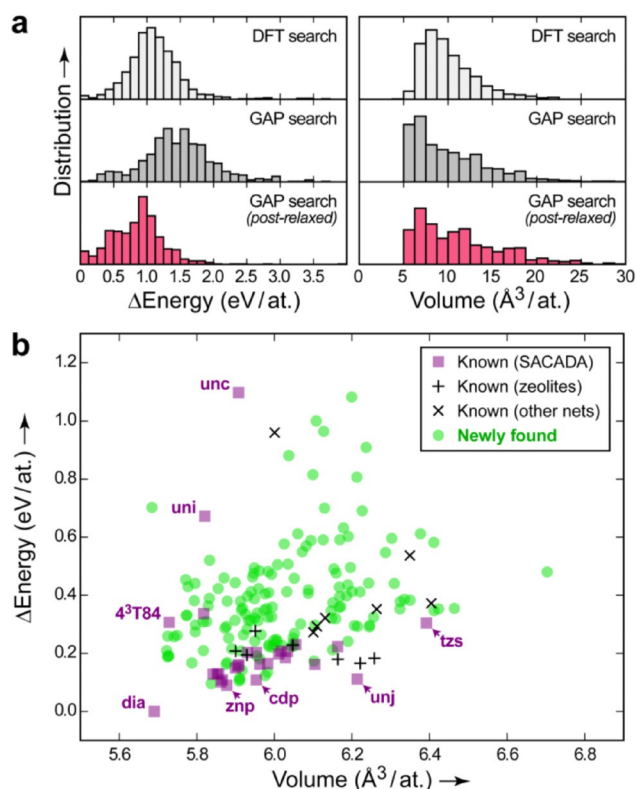


Figure 2. a) Tests of energy and volume distributions from a trial set of DFT- and GAP-driven structure searches. To enable comparison, all energies given have been re-computed using DFT. b) Energy–volume plot for the results of our main, much larger structure search, including known (purple, black) and new (green) carbon allotropes. Topology symbols^[14] such as „*dia*“ are given for a number of representative known structures. Our search also found lonsdaleite (*lon*) and a mixed *dia/lon* stacking sequence that are omitted for clarity; detailed results are given in the Supporting Information.

Many of these structures are best understood by dissecting them into characteristic, topological building blocks.^[15,16] For example, carbon atoms in diamond form six-membered rings exclusively, and four of these combine into an adamantane-like cage, such that the tiling symbol for *dia* is written as $[6^4]$ (details may be found in Ref. [17]). It was previously pointed out how other structural motifs can be combined with *dia* (or lonsdaleite, *lon*) cages to form more complex allotropes.^[8b,15] Figure 3a illustrates this using an example: combining *dia*, *lon*, and the characteristic five- and seven-membered ring fragments of *cbn* (M-carbon; Ref. [5a-b]) leads to a new structure, **G95**, that is found by our search. (We label all new structures with a G for GAP, and the number is simply a running index).

We also find several new 5+5+8 allotropes^[8b] that contain, as the name suggests, sets of five- and eight-membered ring fragments (Figure 3b). In **G12**, layers of such motifs (blue/yellow) are interwoven with two consecutive layers of *dia* spacers (empty). The stacking sequence of the building units can be written as *AabAab*, where capital letters denote the stacking of the defining ring structures, and lowercase italics refer to the spacers. Reducing the concentration of the latter, we have **G21** (and **G6**, which is similar but with a less favourable stacking sequence). We also find a corresponding structure

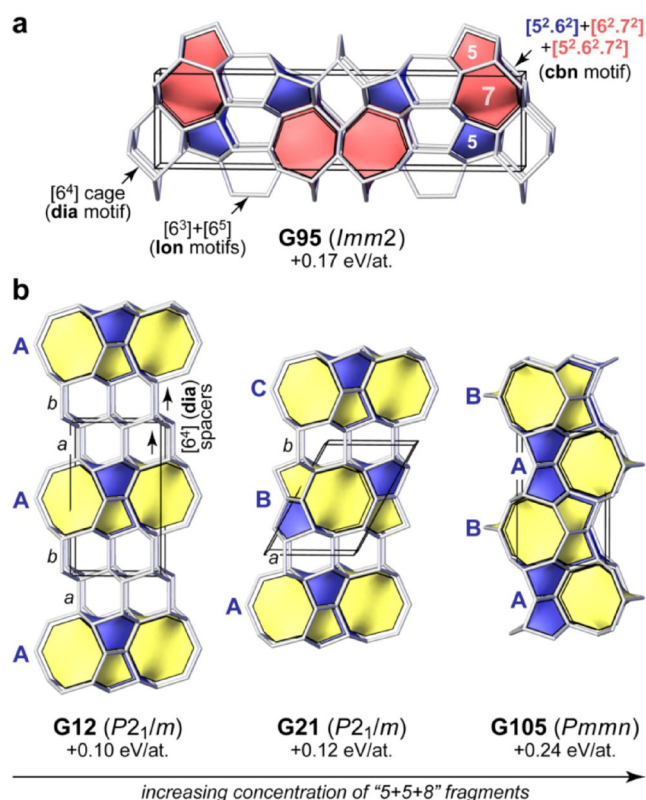


Figure 3. a) **G95**, a predicted carbon allotrope that combines structural motifs from diamond (**dia**), lonsdaleite (**lon**), and „M-carbon“ (**cbn**), as revealed by topology analysis. b) Predicted carbon allotropes with „5 + 5 + 8“ building blocks. Cages that involve an eight-membered ring, typically $[5^2.6^2.8^2]$, are highlighted in yellow; $[5^2.6^2]$ cages are blue. Interestingly, the apparent end-member of this series, **G105**, is different in cage topology from the others, and filling space exclusively with $[5^2.6^2.8^2]$ and $[5^2.6^2]$ cages would yield the **bik** network instead (Supporting Information). All energies are given relative to diamond.

without any diamond-like cages, **G105**, which is higher in energy as there is no dilution by **dia** spacers. In general, no straightforward correlation exists between the simplicity of the structures and their stability (Supporting Information, Figure S8).

As pointed out by Botti et al.,^[8b] one may freely add more and more **dia** (or **lon**) like spacers to such structures, and therefore create infinite numbers of topologically unique networks. Where is the limit? Figure 3b suggests a possible answer: we believe that truly distinct carbon allotropes should be restricted to cases with clear and simple stacking sequences both for the mixed-ring units and the spacers. For the same reason, we have excluded a combination of only **dia** and **lon** cages from the plot in Figure 2b, as an infinite number of similar polytypes can be trivially defined.

The critical reader will now ask whether one needs high-throughput computations to devise such simple stacking sequences. Indeed, the true strength of GAP-driven structure searching is that due to its speed (our search comprised over 290,000 runs), it is likely to unveil more complex cases that depart from previously established structural principles but are still energetically viable.

In the latter category fall carbon allotropes with what we call pseudo-tiling patterns (Figure 4). Drawing 2D projections of such structures gives the impression of very small, three- and four-membered rings—but in fact the relevant atoms lie atop each other along the viewing direction, and so only create the illusion of touching. Figure 4a illustrates this for **4³T143**, a variant of the chiral **unj** net^[18] that has been observed in a database of hypothetical zeolites,^[15,19] we note that the same topology was very recently described for Si allotropes.^[16] In **unj**, chiral tubes of fivefold rings form what looks like a honeycomb structure when viewed down the tube axis (Figure 4b).^[18] Similar tubes exist in **4³T143** but there they form 2D sheets (in the *ab* plane), and these are then stacked perpendicularly along *c*. Hence, the fivefold rings seen in Figure 4c are the actual structural motif, observed when a tube is

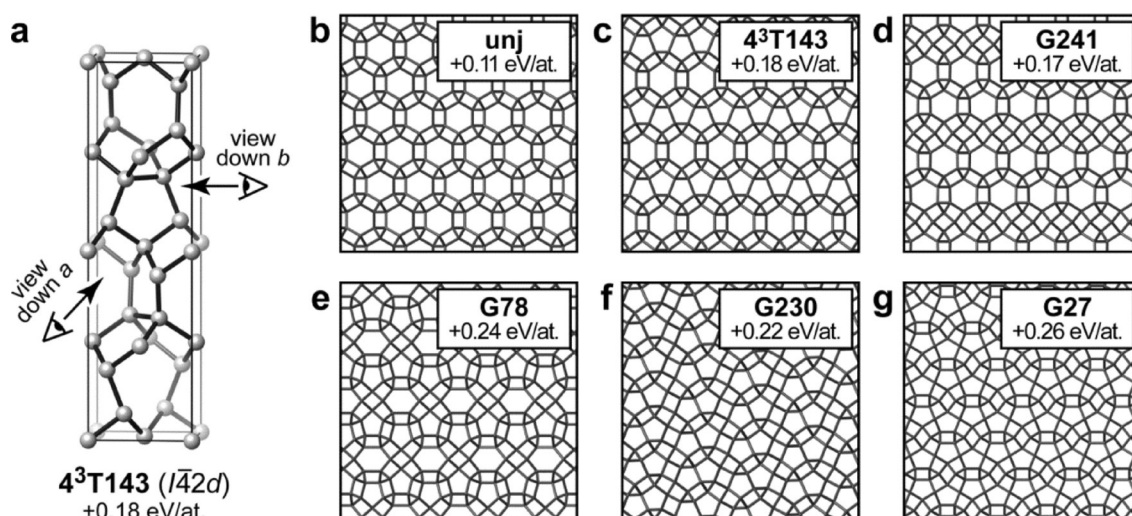


Figure 4. a) **4³T143**, a variant of the chiral **unj** framework: the structure is composed of five-membered rings, but viewing it down the *a* or the *b* axis as indicated creates the impression of three-, four-, and six-membered ones. b–g) Projection views of known and new structures found in our search.

looked upon in side view. Both structures are generated by filling space with the same topologically unique cage, [5².8²].

Just like combinations of small polygons can create diverse patterns, the putative family of pseudo-tiling allotropes is here found to span a wider range. Figure 4d shows **G241**, a related motif in which **unj**-like [5².8²] cages are mixed with **dia**-like [6⁴] ones; Figure 4e shows **G78** which adds cages with seven-membered rings instead; both structures are chiral in space group C222₁. There are further pseudo-tiling patterns without apparent six-rings, as we find for **G230** and **G27** (Figure 4f,g); both are loosely reminiscent of the P4₁2₁2 spiral structure for group-14 allotropes predicted recently from AIRSS (4³T130 in SACADA).^[4b] All these structures need not only be hypothetical constructs: recent experiments showed that complex Si allotropes of such type may indeed be formed in “microexplosions”, locally induced in a crystalline matrix by ultrashort laser pulses.^[20]

We stress that AIRSS, like all ab initio methods, can only span a certain subspace: that of structures with a small number of atoms in the primitive unit cell (here, ≤ 16). By contrast, novel carbon allotropes have been predicted based on chemical knowledge, deriving them from zeolites^[15,21] or clathrate structures;^[22] these are often highly competitive in energy, but inaccessible to DFT-based searches. In the future, ML-based techniques might enable the ab initio prediction of networks with hundreds of atoms in the primitive cell. And in the end, the crucial task will be not only to generate large numbers of structures out of the infinitely many possible ones (which a machine can do), but to derive new chemical insight and guidelines for experiments (which a machine cannot).

In conclusion, we have explored the structural space of carbon allotropes by combining random structure searching with an efficient machine-learning based interatomic potential. Our GAP model readily enables predictions of crystalline phases, despite having been trained on liquid and amorphous structures alone. This represents a hard test in terms of transferability, and it opens up the road for further applications of ML models in solid-state chemistry—where the ability to assemble and correctly describe new structures is paramount. We focused on one particular structure-prediction method, but the ML model might just as well be coupled to others (such as genetic algorithms), or even to the nested-sampling technique to assess temperature-pressure phase diagrams fully from first principles.^[23] Likewise, the field of organic crystal-structure prediction might benefit from similar techniques,^[24] albeit in that case the focus is on long-range dispersion interactions rather than on the making and breaking of covalent bonds.^[25] Further work will extend our present findings to carbon allotropes at (very) high pressure, to networks with mixed coordination numbers, and to other materials for which similar approaches seem promising.

Computational Methods

A GAP model^[10b] was fitted to DFT energies and forces using the same protocols and parameters as outlined in our preceding, more technical work in Ref. [13]. The input for this was a database of

3,070 liquid and amorphous carbon configurations taken from Ref. [13]. Using this GAP, a total of 290,885 relaxations were performed for randomized cells containing 3–16 atoms, at ambient and elevated pressure. No symmetry operations were applied during the search itself, to allow for maximal degrees of freedom; instead, space-group symmetry was determined a posteriori. Only structures with fourfold coordinated atoms are reported (determined using a cutoff of 1.70 Å), and candidates with three-membered carbon rings were discarded due to the associated large strains. Post-processing of remaining candidate structures was done by full DFT-GGA^[26] relaxation of lattice parameters and atomic positions to zero external pressure, using CASTEP;^[27] the final structures are provided as Supporting Information in CIF format. Symmetry analyses were done as implemented in PHONOPY,^[28] PLATON,^[29] and SYSTRE;^[30] structures were visualized using GAVROG (www.gavrog.org) and VESTA.^[31]

Acknowledgements

We thank Professors R. Dronskowski, S. R. Elliott, and C. J. Pickard for valuable discussions. V.L.D. gratefully acknowledges a Feodor Lynen fellowship from the Alexander von Humboldt Foundation and support from the Isaac Newton Trust (Trinity College Cambridge). D.M.P. thanks the Russian Government (Grant 14.B25.31.0005). This work used the ARCHER UK National Supercomputing Service (<http://www.archer.ac.uk>) via EPSRC Grant EP/K014560/1. Data Access Statement: Data supporting this publication are provided as online Supporting Information and at <https://doi.org/10.17863/CAM.7944>.

Keywords: ab initio calculations · carbon allotropes · high-throughput screening · machine learning · solid-state structures

- [1] a) J. C. Schön, M. Jansen, *Angew. Chem. Int. Ed. Engl.* **1996**, *35*, 1286–1304; *Angew. Chem.* **1996**, *108*, 1358–1377; b) S. M. Woodley, R. Catlow, *Nat. Mater.* **2008**, *7*, 937–946; c) A. Fadda, G. Fadda, *Phys. Rev. B* **2010**, *82*, 104105; d) A. R. Oganov, A. O. Lyakhov, M. Valle, *Acc. Chem. Res.* **2011**, *44*, 227–237; e) C. J. Pickard, R. J. Needs, *J. Phys. Condens. Matter* **2011**, *23*, 053201; f) *Modern Methods of Crystal Structure Prediction* (Ed.: A. R. Oganov), Wiley-VCH, Weinheim, **2011**; g) M. Jansen, *Adv. Mater.* **2015**, *27*, 3229–3242.
- [2] a) C. J. Pickard, R. J. Needs, *Nat. Mater.* **2008**, *7*, 775–779; b) Y. Ma, M. Eremets, A. R. Oganov, Y. Xie, I. Trojan, S. Medvedev, A. O. Lyakhov, M. Valle, V. Prakapenka, *Nature* **2009**, *458*, 182–185; c) L. Zhu, Z. Wang, Y. Wang, G. Zou, H. Mao, Y. Ma, *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 751–753; d) W. Zhang, A. R. Oganov, A. F. Goncharov, Q. Zhu, S. E. Boulfelfel, A. O. Lyakhov, E. Stavrou, M. Somayazulu, V. B. Prakapenka, Z. Konôpková, *Science* **2013**, *342*, 1502–1505; e) L. Zhu, H. Liu, C. J. Pickard, G. Zou, Y. Ma, *Nat. Chem.* **2014**, *6*, 644–648.
- [3] For some important early examples, see: a) A. T. Balaban, C. C. Rentia, E. Ciupitu, *Rev. Roum. Chim.* **1968**, *13*, 231–247; Corrigendum: A. T. Balaban, C. C. Rentia, E. Ciupitu, *Rev. Roum. Chim.* **1968**, *13*, 1233; b) M. V. Nikerov, D. A. Bochvar, I. V. Stankevich, *Izv. Akad. Nauk SSSR Ser. Khim.* **1981**, 1177–1178 (in Russian); c) R. Hoffmann, T. Hughbanks, M. Kertesz, P. H. Bird, *J. Am. Chem. Soc.* **1983**, *105*, 4831–4832; d) A recent, interesting read is L. Öhrström, M. O’Keeffe, *Z. Kristallogr.* **2013**, *228*, 343–346.
- [4] a) R. T. Strong, C. J. Pickard, V. Milman, G. Thimm, B. Winkler, *Phys. Rev. B* **2004**, *70*, 045101; b) A. Mujica, C. J. Pickard, R. J. Needs, *Phys. Rev. B* **2015**, *91*, 214104.
- [5] a) A. R. Oganov, C. W. Glass, *J. Chem. Phys.* **2006**, *124*, 224704; b) Q. Li, Y. Ma, A. R. Oganov, H. Wang, H. Wang, Y. Xu, T. Cui, H.-K. Mao, G. Zou, *Phys. Rev. Lett.* **2009**, *102*, 175506; c) Q. Zhu, A. R. Oganov, M. A. Salvadó,

- P. Pertierra, A. O. Lyakhov, *Phys. Rev. B* **2011**, *83*, 193410; d) Q. Zhu, Q. Zeng, A. R. Oganov, *Phys. Rev. B* **2012**, *85*, 201407.
- [6] Z. Zhao, F. Tian, X. Dong, Q. Li, Q. Wang, H. Wang, X. Zhong, B. Xu, D. Yu, J. He, H. T. Wang, Y. Ma, Y. Tian, *J. Am. Chem. Soc.* **2012**, *134*, 12362–12365.
- [7] D. Selli, I. A. Baburin, R. Martoňák, S. Leoni, *Phys. Rev. B* **2011**, *84*, 161411.
- [8] a) M. Amsler, J. A. Flores-Livas, L. Lehtovaara, F. Balima, S. A. Ghasemi, D. Machon, S. Pailhès, A. Willand, D. Caliste, S. Botti, A. San Miguel, . Goedecker, M. A. L. Marques, *Phys. Rev. Lett.* **2012**, *108*, 065501; b) S. Botti, M. Amsler, J. A. Flores-Livas, P. Ceria, S. Goedecker, M. A. L. Marques, *Phys. Rev. B* **2013**, *88*, 014102.
- [9] R. Hoffmann, A. A. Kabanov, A. A. Golov, D. M. Proserpio, *Angew. Chem. Int. Ed.* **2016**, *55*, 10962–10976; *Angew. Chem.* **2016**, *128*, 11122–11139.
- [10] a) J. Behler, M. Parrinello, *Phys. Rev. Lett.* **2007**, *98*, 146401; b) A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, *Phys. Rev. Lett.* **2010**, *104*, 136403; c) N. Artrith, A. Urban, *Comput. Mater. Sci.* **2016**, *114*, 135–150; d) A. V. Shapeev, *Multiscale Model. Simul.* **2016**, *14*, 1153–1173; e) V. Botu, R. Batra, J. Chapman, R. Ramprasad, *J. Phys. Chem. C* **2017**, *121*, 511–522; f) For a recent overview of this growing field, see: J. Behler, *J. Chem. Phys.* **2016**, *145*, 170901.
- [11] a) R. Z. Khaliullin, H. Eshet, T. D. Kühne, J. Behler, M. Parrinello, *Phys. Rev. B* **2010**, *81*, 100103; b) R. Z. Khaliullin, H. Eshet, T. D. Kühne, J. Behler, M. Parrinello, *Nat. Mater.* **2011**, *10*, 693–697.
- [12] P. E. Dolgirev, I. A. Kruglov, A. R. Oganov, *AIP Adv.* **2016**, *6*, 085318.
- [13] V. L. Deringer, G. Csányi, *Phys. Rev. B* **2017**, *95*, 094203.
- [14] V. A. Blatov, A. P. Shevchenko, D. M. Proserpio, *Cryst. Growth Des.* **2014**, *14*, 3576–3586.
- [15] I. A. Baburin, D. M. Proserpio, V. A. Saleev, A. V. Shipilova, *Phys. Chem. Chem. Phys.* **2015**, *17*, 1332–1338.
- [16] Many of these approaches and techniques can likewise be applied to silicon allotropes. Very recent studies are in a) L.-A. Jantke, S. Stegmaier, A. J. Karttunen, T. F. Fässler, *Chem. Eur. J.* **2017**, *23*, 2734–2747; b) A. J. Karttunen, D. Usvyat, M. Schütz, L. Maschio, *Phys. Chem. Chem. Phys.* **2017**, in press, DOI:10.1039/C6CP08873B.
- [17] V. A. Blatov, M. O’Keeffe, D. M. Proserpio, *CrystEngComm* **2010**, *12*, 44–48.
- [18] a) M. O’Keeffe, N. E. Brese, *Acta Crystallogr. Sect. A* **1992**, *48*, 663–669; b) This very structure has been earlier identified from DFT-driven AIRSS and dubbed „chiral framework structure“ (CFS): C. J. Pickard, R. J. Needs, *Phys. Rev. B* **2010**, *81*, 014106.
- [19] M. W. Deem, R. Pophale, P. A. Cheeseman, D. J. Earl, *J. Phys. Chem. C* **2009**, *113*, 21353–21360.
- [20] L. Rapp, B. Haberl, C. J. Pickard, J. E. Bradby, E. G. Gamaly, J. S. Williams, A. V. Rode, *Nat. Commun.* **2015**, *6*, 7555.
- [21] R. Nesper, K. Vogel, P. E. Blöchl, *Angew. Chem. Int. Ed. Engl.* **1993**, *32*, 701–703; *Angew. Chem.* **1993**, *105*, 786–788.
- [22] a) A. J. Karttunen, T. F. Fässler, M. Linnolahti, T. A. Pakkanen, *Inorg. Chem.* **2011**, *50*, 1733–1742; b) A. J. Karttunen, T. F. Fässler, *ChemPhysChem* **2013**, *14*, 1807–1817.
- [23] R. J. N. Baldock, L. B. Pártay, A. P. Bartók, M. C. Payne, G. Csányi, *Phys. Rev. B* **2016**, *93*, 174108.
- [24] a) M. A. Neumann, F. J. J. Leusen, J. Kendrick, *Angew. Chem. Int. Ed.* **2008**, *47*, 2427–2430; *Angew. Chem.* **2008**, *120*, 2461–2464; b) S. L. Price, *Chem. Soc. Rev.* **2014**, *43*, 2098–2111; c) G. M. Day, C. H. Görbitz, *Acta Crystallogr. Sect. B* **2016**, *72*, 435–436.
- [25] While such long-range electrostatic and dispersive interactions are not included in the present GAP, as they are less relevant for diamond and related structures, we note that their inclusion in ML-based potentials has been pioneered before—this aspect therefore does not affect the generality of our findings. For a recent study including van der Waals interactions see, e.g., T. Morawietz, A. Singraber, C. Dellago, J. Behler, *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 8368–8373.
- [26] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- [27] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. J. Probert, K. Refson, M. C. Payne, *Z. Kristallogr.* **2005**, *220*, 567–570.
- [28] A. Togo, I. Tanaka, *Scr. Mater.* **2015**, *108*, 1–5.
- [29] A. L. Spek, *Acta Crystallogr. Sect. D* **2009**, *65*, 148–155.
- [30] O. Delgado-Friedrichs, M. O’Keeffe, *Acta Crystallogr. Sect. A* **2003**, *59*, 351–360.
- [31] K. Momma, F. Izumi, *J. Appl. Crystallogr.* **2011**, *44*, 1272–1276.

Manuscript received: February 14, 2017

Final Article published: March 8, 2017