

A GAP potential for carbon from randomly selected, uncured structures: comparison with GAP-20

C. D. Chalkidis¹, G. Fanourgakis¹

¹Laboratory of Quantum and Computational Chemistry, School of Chemistry, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
email: chalchri@chem.auth.gr, fanourg@chem.auth.gr

Machine learning potentials such as Gaussian Approximation Potentials (GAP; Bartók et al., 2010) approach quantum-mechanical accuracy at a cost orders of magnitude lower, but they shift their limitation from the functional form to the data, whose iterative curation demands time and expertise. We examine what accuracy is obtained without any curation, and what limits it. Training sets of 50 to 2,000 structures were drawn by stratified random sampling from the same database of 17,525 carbon configurations from which the published GAP-20 potential was curated (Rowe et al., 2020). We trained 1,040 potentials with the SOAP descriptor (Bartók et al., 2013), in eight model settings with ten repetitions per size, and evaluated them on a common set of 203 structures.

The energy error falls by 43% up to about 700 structures and then rises by 6.6% up to 2,000, not through overfitting but because of rare repulsive structures with extreme forces, which enter larger sets more often and degrade the fit. A tenfold increase in capacity (from 300 to 3,000 representative points) lowers the error by 40.5% but not the failure rate on the repulsive structures, where GAP-20, too, has an error 33 times larger than on the rest. On the 116 evaluation structures that neither model had seen, the best potential reaches 0.0093 eV/atom versus 0.0103 for GAP-20, using 1,000 random structures instead of 6,088 curated ones (Figure 1); it is better on ordered structures and worse on disordered ones.

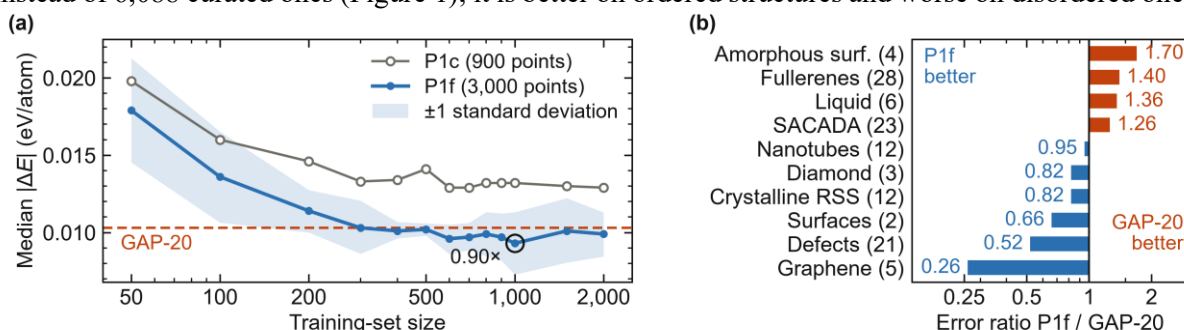


Figure 1. (a) Median $|\Delta E|$ on the 116 structures that no model had seen (shading: ± 1 standard deviation across repeats). (b) Error ratio P1f/GAP-20 per structural family, at 1,000 structures.

The binding constraint on accuracy therefore lies in the data, not in the model; curation is not a prerequisite but one of the available levers.

References

- Bartók, A. P., Payne, M. C., Kondor, R., & Csányi, G. (2010). Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters*, 104(13), 136403. <https://doi.org/10.1103/PhysRevLett.104.136403>
- Bartók, A. P., Kondor, R., & Csányi, G. (2013). On representing chemical environments. *Physical Review B*, 87(18), 184115. <https://doi.org/10.1103/PhysRevB.87.184115>
- Rowe, P., Deringer, V. L., Gasparotto, P., Csányi, G., & Michaelides, A. (2020). An accurate and transferable machine learning potential for carbon. *The Journal of Chemical Physics*, 153(3), 034702. <https://doi.org/10.1063/5.0005084>

Acknowledgements

The computations were performed on the “Aristotelis” infrastructure of the AUTH IT Center.