The Journal of Chemical Physics

Response to "Comment on 'Manifolds of quasi-constant SOAP and ACSF fingerprints and the resulting failure to machine learn four-body interactions'" [J. Chem. Phys. 156, 034302 (2022)]

Cite as: J. Chem. Phys. **157**, 177102 (2022); https://doi.org/10.1063/5.0099525
Submitted: 17 May 2022 • Accepted: 06 October 2022 • Accepted Manuscript Online: 07 October 2022 • Published Online: 02 November 2022

Behnam Parsaeifard, Marco Krummenacher and Stefan Goedecker

The Journal of Chemical Physics RESPONSE scitation.org/journal/jcp

Response to "Comment on 'Manifolds of quasi-constant SOAP and ACSF fingerprints and the resulting failure to machine learn four-body interactions'" [J. Chem. Phys. 156, 034302 (2022)]

Cite as: J. Chem. Phys. 157, 177102 (2022); doi: 10.1063/5.0099525
Submitted: 17 May 2022 • Accepted: 6 October 2022 •
Published Online: 2 November 2022

Behnam Parsaeifard, Marco Krummenacher, and Stefan Goedecker[a)]

AFFILIATIONS

Department of Physics, University of Basel, Klingelbergstrasse 82, CH-4056 Basel, Switzerland

[a)]Author to whom correspondence should be addressed: stefan.goedecker@unibas.ch

https://doi.org/10.1063/5.0099525

It is uncontested that a machine learning scheme cannot correctly reproduce physical properties that vary on a manifold in configuration space if the fingerprint, used as an input for the machine learning scheme, is constant on this manifold. In our original paper (Ref. 1), we discovered manifolds of quasi-constant fingerprint for Smooth Overlap of Atomic Positions (SOAP) and Atom-Centered Symmetry Functions (ACSF) fingerprints for test environments that contain three or four atoms around the central atom. Standard parameters were used for both fingerprints. We defined a quasi-constant fingerprint as a fingerprint whose variation on the manifold is so small that the machine learning schemes behaves as if it was exactly constant. There is obviously a threshold for the variation of the fingerprint on the manifold above which machine learning will again become possible. In our original paper, we found several manifolds whose fingerprint variation is sufficiently small to prevent machine learning based on a standard training scheme even if some 40 configurations on the manifold are included in the training set.

The authors of the Comment[2] show that in combination with some kind of oversampling approach, structures on our original manifold of quasi-constant fingerprint can be machine learned. While we used a standard training scheme where the structures used for training are chosen in the usual way according to physical criteria (e.g., from a MD run) and then held fixed, the authors of the Comment[2] use a non-standard training scheme that has never been used before for machine learning of potential energy surfaces. In this scheme, the selection of the configurations to be fitted is dictated by the deficiencies of the SOAP fingerprint, requiring the calculation of a manifold of quasi-constant fingerprint. The configurational space on the manifold has then to be oversampled sufficiently to enable machine learning. This oversampling can be done by including a sufficiently large number of configurations on the manifold in the training set. Determining the manifold and the necessary degree of oversampling is a cumbersome process that is unlikely to be ever used in practice. Since this fitting scheme is constructed with the intention to compensate for the deficiencies of the SOAP fingerprint, we will refer in the following to this scheme as deficiency-adapted scheme.

Even though manifolds that were constructed from identical initial configurations but with different parameters look similar by eye, they differ considerably in the central property, namely, the variation of the fingerprint. In our publication, we constructed a manifold with SOAP parameters $n_{\max} = l_{\max} = 16, \sigma = 0.3$. The same value for σ and similar values for $n_{\max}$ and $l_{\max}$ ($n_{\max} = l_{\max} = 8$) were used in the Comment.[2] As shown in Fig. 1, the variation of the fingerprint becomes larger for smaller values of σ. It is, therefore, to be expected that a weaker oversampling will be required. However, reducing σ from the standard range of 0.3–0.5 to much smaller values makes the resulting fingerprint much slower since $n_{\max}$ and $l_{\max}$ have also to be increased for a balanced description. Hence, σ cannot be made arbitrarily small in practice. Even for a given set of SOAP parameters one can actually construct many manifolds with different fingerprint variation by starting from different initial structures.

The training of a neural network is a non-convex global optimization problem that does not have a unique solution. Relatively

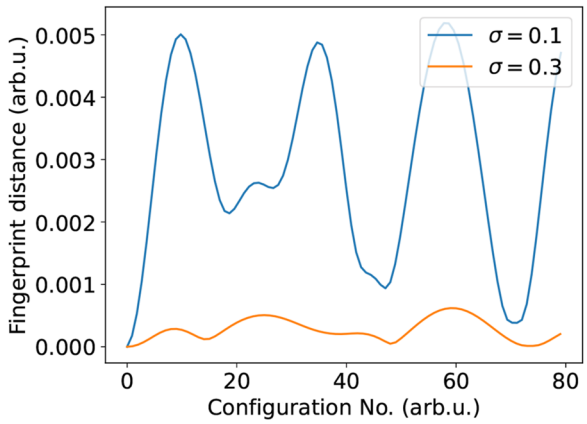

FIG. 1. Comparison of the variation of the fingerprint for our original manifold[1] (P & G) with $\sigma = 0.3$ compared to the variation of the fingerprint on another manifold with a smaller value of $\sigma = 0.1$.

small changes in the methodology can, therefore, lead to different results. To investigate the purely academic question of whether structures on a single quasi-constant manifold can be machine learned in a general machine learning context using a sufficiently strong oversampling, we therefore performed some additional tests using the Kernel Ridge Regression (KRR) method, which gives a unique solution by solving a linear system of equations. For the construction of a Gaussian kernel matrix only two parameters are necessary. So in contrast to neural network methods, the parameter space can be fully scanned to find the optimal parameter set. The results based on such a scan are summarized in Fig. 2. Putting more and more weight on the manifold structures by increasing the number of manifold structures in the training set to very large values clearly improves the learnability of manifold structures.

Unfortunately, the ability to machine learn a single manifold with the deficiency adapted scheme does not cure the problem of applying the method to the calculation of potential energy surfaces.

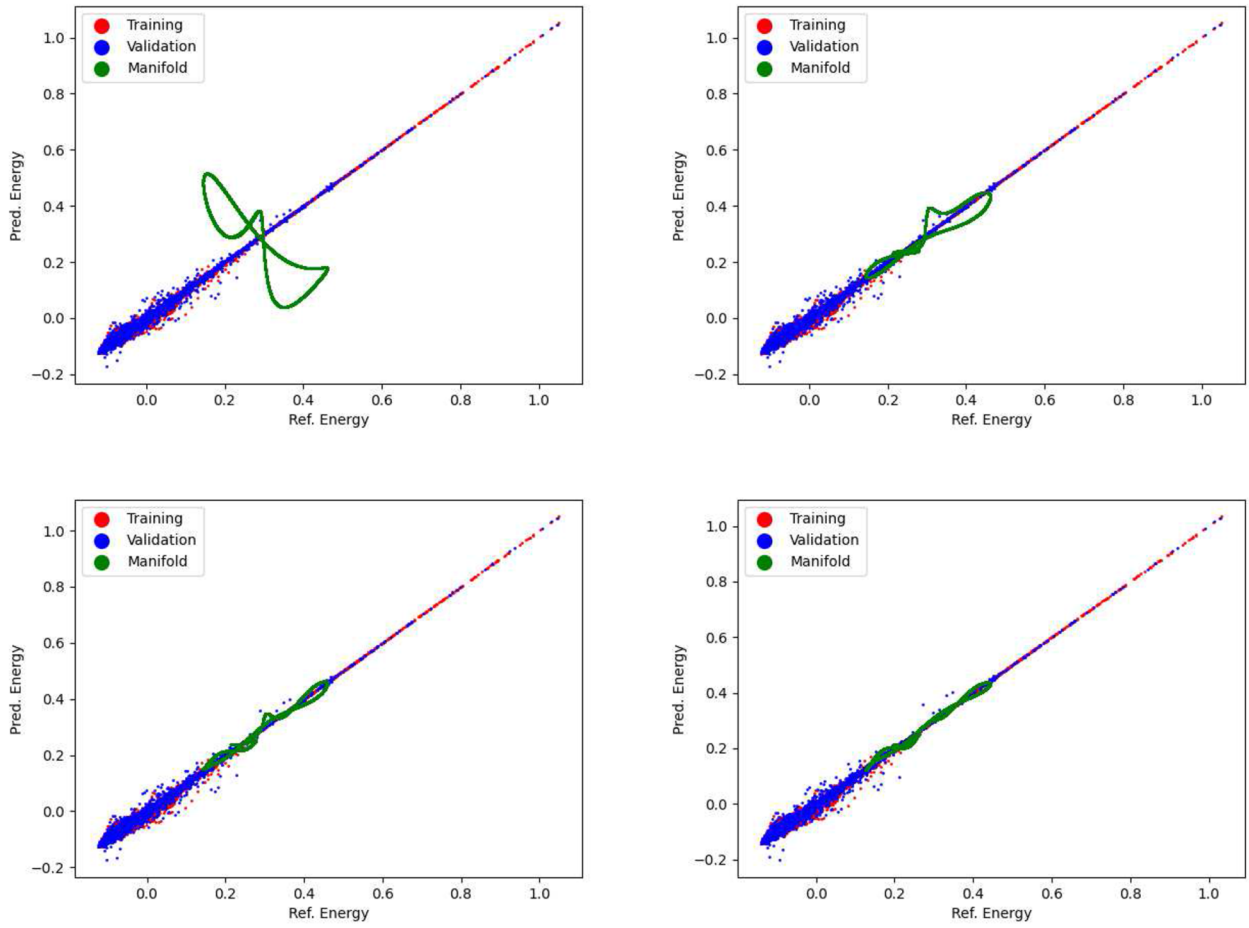

FIG. 2. Machine learning results obtained with the KRR method for the (P & G) manifold constructed with $\sigma = 0.3$. Going from the left to the right panel and from the top to the bottom line, no structures and 16, 160, and 1600 structures on the manifold were included in the training set.

As stressed in our publication and also acknowledged in the Comment,[2] there exist a huge number of manifolds and machine learning one manifold does not mean that any of the other manifolds can be described correctly. A detailed example of this is shown in Fig. 3, where manifold 1 is from the authors of the Comment.[2] Manifold 2 was constructed from a different initial structure with a larger four body energy. One can clearly see that including one manifold in the fitting data does not at all improve the quality of the fit for the other manifold.

Since in a physical application one needs a machine learning model that describes correctly all the manifolds, our statement that the manifolds of quasi-constant fingerprint prevent

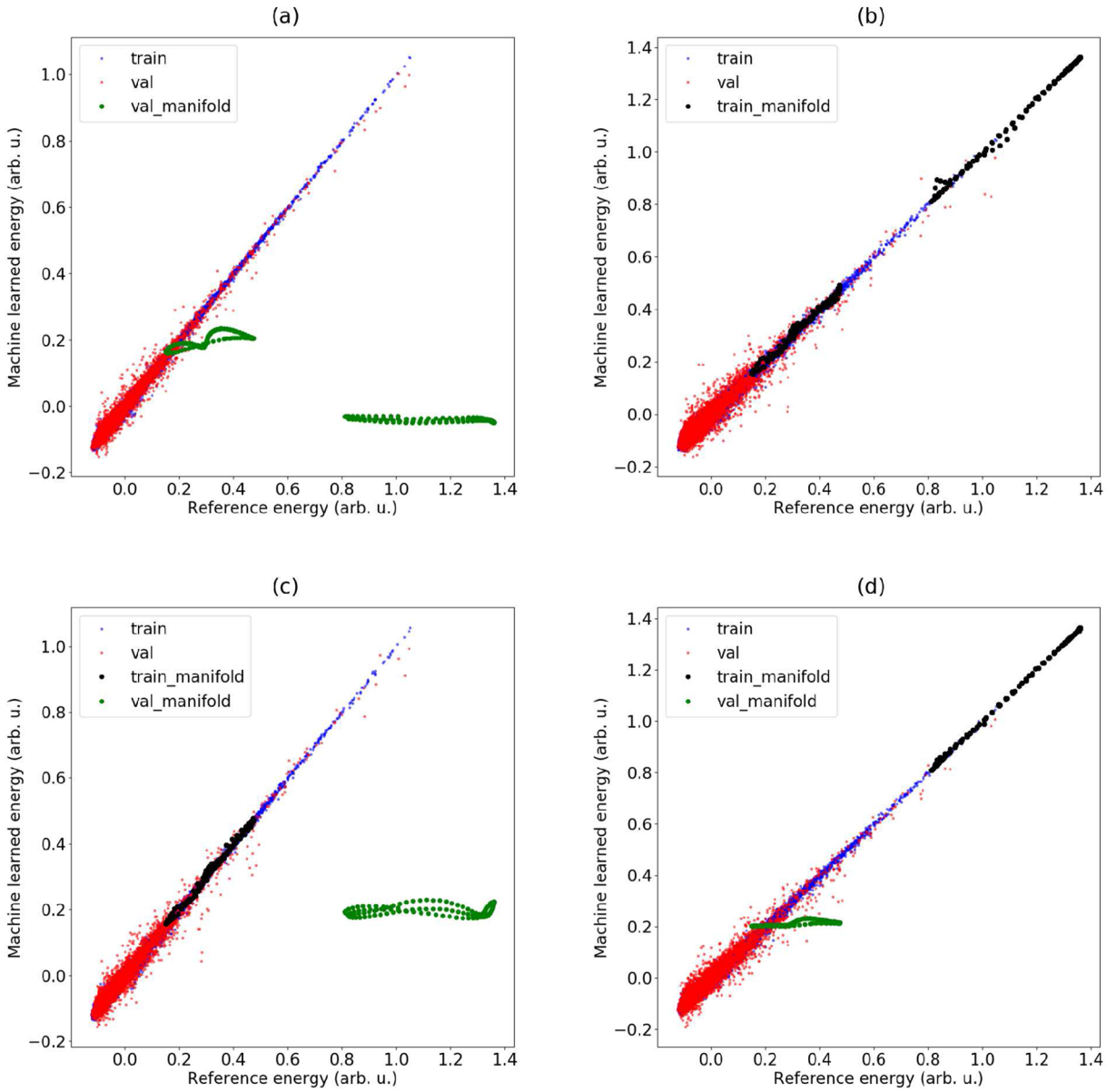

FIG. 3. Machine learning results using neural networks for two mild manifolds with $\sigma = 0.2$. The red and blue points show the results for the training and validation sets. In panel (a), no structures from either manifolds were added to the training set and the manifolds cannot be described. In panel (b), structures from both manifolds were included in the training set. They were machine learned according to the deficiency-adapted procedure by adding 24 000 manifold structures (120 evenly distributed structures repeated 200 times) to the training set. In panel (c), only structures from manifold 1 were included in the training. Manifold 1 can now be over-fitted, but there is no improvement for manifold 2. In panel (d), the opposite was done. Only the structures of manifold 2 were included in the training. In this case, manifold 1 is not learned. So even though a single mild manifold can be learned with the deficiency-adapted procedure, the description of other manifolds is not improved.

machine learning of four-body energies is true even in the case of mild manifolds with a relatively large variation of the fingerprint.

During the motion of the atoms on the manifold, the H–H distances vary considerably. Hence, the degeneracy will be lifted if the atomic environment fingerprints of all atoms are used to describe the molecule. So torsional energies might be represented as some long range pairwise interactions in machine learning schemes based on the SOAP fingerprint. It is at present not known how much such a compensation that is not in agreement with the standard chemical understanding of the origin of torsional energies will limit the accuracy of machine learned potential energy surfaces for systems where torsional motions are important. If torsional terms are expected to be important, fingerprints[3,4] can be used that allow to include in a systematic way higher order terms.

DATA AVAILABILITY

The data that support the findings of this study are available within the article.

REFERENCES

[1] B. Parsaeifard and S. Goedecker, "Manifolds of quasi-constant SOAP and ACSF fingerprints and the resulting failure to machine learn four-body interactions," J. Chem. Phys. **156**, 034302 (2022).
[2] S. N. Pozdnyadov, M. J. Willatt, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, "Comment on 'Manifolds of quasi-constant SOAP and ACSF fingerprints and the resulting failure to machine learn four-body interactions," J. Chem. Phys. **156**, 034302 (2022).
[3] A. V. Shapeev, "Moment tensor potentials: A class of systematically improvable interatomic potentials," Multiscale Model. Simul. **14**, 1153–1173 (2016).
[4] R. Drautz, "Atomic cluster expansion for accurate and transferable interatomic potentials," Phys. Rev. B **99**, 014104 (2019).