Manifolds in commonly used atomic fingerprints lead to failure in machine-learning four-body interactions
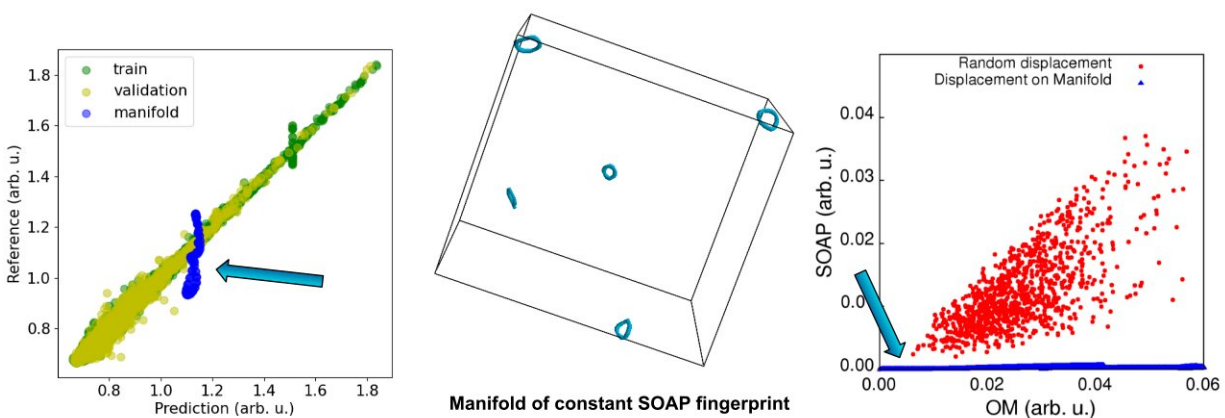
January 10 2022, by Carey Sargent

Manifold of constant SOAP fingerprint

In Figure 1, the central panel shows trajectories of a rotation and translation-free movement of the methane molecule on a manifold of quasi constant SOAP fingerprint. The blue points in the right panel show the fingerprint distances for configurations on this manifold. The fact that they are quasi zero indicates that the SOAP fingerprint is quasi constant. An alternative Overlap Matrix (OM) fingerprint recognizes however the structural differences. The red points show the much larger SOAP fingerprint distances for other movements with smaller or similar atomic displacements. The left panel shows the machine learning capabilities for a torsional energy. The torsional angles change on the manifold, but since the fingerprint is quasi constant the torsional energies on the manifold are also quasi constant (blue points) and do therefore not capture the correct variation of the torsional energy. For the points off the manifold, shown in green/yellow, the variation of the torsional energy can however be well tracked. Credit: @UniBasel Stefan Goedecker

Atomic environment fingerprints, or structural descriptors, are used to describe the chemical environment around a reference atom. Encoding information such as bond-lengths to neighboring atoms or coordination numbers, these fingerprints are used, for example, as inputs in machine learning approaches or to eliminate redundant structures in structural searches

To serve their purpose well, such fingerprints should be invariant under uniform translations, rotations, and permutation of identical atoms in the system. They should also be unique in the sense that two environments are necessarily identical if their fingerprints are identical. If this isn't the case, two different non-degenerate structures will be assigned the same energy in any machine learning scheme that uses such a fingerprint as an input. Numerically, and for use in ML schemes, the difference in fingerprint between two atoms must correlate with the amount of dissimilarity between the environments: two structures whose fingerprint difference is very small should be very similar.

In the paper "Manifolds of quasi-constant SOAP and ACSF fingerprints and the resulting failure to machine learn four body interactions," recently published in *The Journal of Chemical Physics*, researchers Stefan Goedecker and Behnam Parsaeifard of the University of Basel showed that these conditions are not always fulfilled in two commonly used fingerprints, the so-called atom-centered symmetry functions (ACSF) and the smooth overlap of atomic positions (SOAP).

ACSF and SOAP are two fingerprints commonly used in machine learning of the potential energy surface. ACSFs consist of radial symmetry functions that are sums of two-body terms and describe the radial environment of an atom and angular symmetry functions that contain the summation of three-body terms and describe an atom's

angular environment.

In the SOAP fingerprint, a Gaussian is centered on each atom within a cutoff distance around the reference atom. The resulting density of atoms multiplied with a cutoff function, which goes smoothly to zero at a cutoff radius over some characteristic width, is then expanded in terms of orthogonal radial functions and spherical harmonics.

In the study, the researchers introduced a sensitivity matrix to study the behavior of atomic fingerprints under infinitesimal changes on the atomic coordinates. They then applied the framework to three small molecules, H_2O, NH_3, and CH_4, to study, respectively, scenarios where a central atom has environments containing 2, 3 or 4 neighboring atoms.

Their investigation revealed that SOAP and ACSF fingerprints are insensitive to certain movements, meaning that the researchers were able to construct manifolds of quasi-constant fingerprint, that is, fingerprints with variations so small that they behave numerically like constant fingerprints, for NH_3 and CH_4. This leads to the fingerprints considering different configurations with different torsional energies to be quasi-identical, preventing, in turn, the machine learning of these torsional energies. Tortional energies of this order are responsible for the folding of proteins and other biomolecules and as such these four-body interactions are considered an essential part in all classical force fields. The limitations of these fingerprints mean that they can only be reproduced with limited accuracy in any ML scheme that uses them. Contrary to widespread belief, the SOAP fingerprint was no better than ACSF in resolving such four-body terms, the researchers said.

The scientists weren't able to find any such manifold for any of the studied molecules or configurations with the so-called Overlap Matrix (OM) fingerprint approach. This fingerprint, based on a diagonalization

in which a minimal basis set of Gaussian orbitals is placed on all atoms within a cutoff radius before the overlap matrix between all orbitals is calculated, always recognizes structural differences, they said, making it an attractive option for resolving structural differences with high fidelity as well as for use in various other applications.