Qeios

Peer Review

Review of: "Machine Learning of Slow Collective Variables and Enhanced Sampling via Spatial Techniques"

J. Galen Wang[1]

1. University of Missouri - Columbia, United States

The authors thoroughly reviewed various machine-learning methods to learn slow collective variables (CVs) from spatial data of molecular dynamics (MD) simulations. These methods center on building the Markovian transition matrix from spatial coordinates, typically involving learning CVs with anisotropic kernels to determine the transition matrix and reweighting the transition matrix to correct the bias from enhanced sampling MD. The learned CVs can then be used to advance the dynamics, which in turn can improve the quality of CVs in an iterative manner. This work is significant to the chemical physics community, in the sense that it can potentially give insights into the mechanisms governing the relevant physical processes. I have a few comments/questions that are hopefully taken in a constructive manner.

1. I like how pedagogical the Introduction and the Background sections are, and I appreciate the clarity. One thing I wish the authors could comment on a bit more is the physical meaning coming out of the learned CVs. The authors mention “interpretability” several times. To me, using kernels makes the CVs sit in a very high-dimensional space of the original dataset, which gives less physical intuition, although I get the idea that the Markov transition matrix built using these CVs can converge to something physical such as diffusion processes. To put it differently, I am wondering whether the CVs learned can be as universal as quantities like free energy, which can be applied to any thermodynamic system or process; i.e., are learned CVs system-dependent? Does running the algorithm across multiple systems (e.g., protein folding and glass transition) yield anything meaningful?
2. Another question I have for this type of method is how large of a dataset from MD is required to generate high-quality results. The authors mention in the Introduction that the sampling problem in MD can be challenging, due to metastability in the free energy landscape and the rare energy-

 doi.org/10.32388/PF2A28

barrier-crossing events. So if the MD data does not run sufficiently long and the system stays in one metastable state, do the CVs learned represent only this local minimum of the energy landscape, rather than representing the entire energy landscape? Do these CV-learning methods require the original MD data to present some transitions between the metastable states at first place?

3. I am also curious about the computational cost of conducting the spatial learning algorithms reviewed in this manuscript compared to the temporal learning algorithms. It may also be useful to compare the computational cost of enhanced sampling of CVs to the conventional MD simulations.
4. Here are some minor points about some typos:
    a. In the paragraph below Eq. (7), things seem to be mixed up. "The dominant eigenvalues of the Markov generator decay exponentially and are linked to the *slowest* relaxation timescales in the system." I think the dominant eigenvalues should be associated with the **fastest** relaxation timescales, according to the equation right below, $t_k=1 / \mu_k$. The last sentence puts "This implies that the eigenvalues much *lower* than μ_{k+1} can be neglected ...", which I think means that the eigenvalues much **higher** than μ_{k+1} should be neglected, because higher μ_k represents smaller t_k corresponding to faster processes.
    b. In the paragraph above Eq. (22), it should be "neural network-based" instead of "neutral network-based".
    c. In the paragraph above Eq. (26), there is a typo for "ehnanced", which should be "enhanced".

Declarations

Potential competing interests: No potential competing interests to declare.