Towards data-driven quantum-accurate Neural Network Interatomic Force Fields

Ji Wei Yoon

UC Berkeley

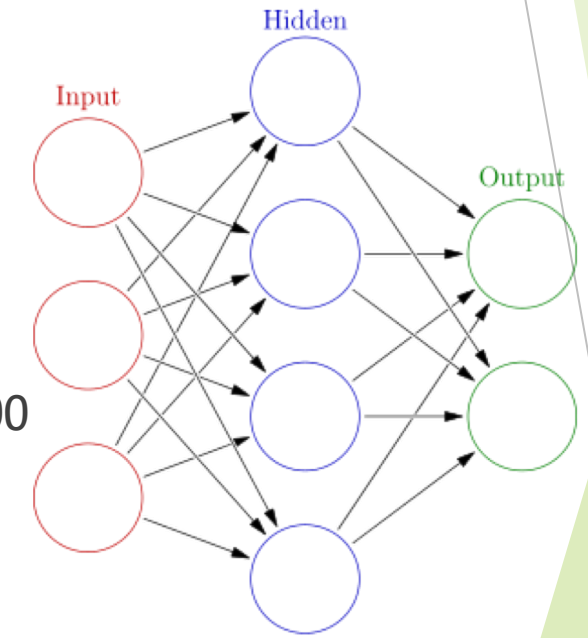
Specifics

- **Methodology**

1. Flexible Neural Network functional form
2. Behler-like descriptors to describe Atomic Environments
3. Pytorch for dynamic computation graphs
   (as opposed to tensorflow that deals with static graphs)

- **Data**

1. Mo dataset from UCSD Ong's group (MSE) in the form of ~100 mb of JSON files.



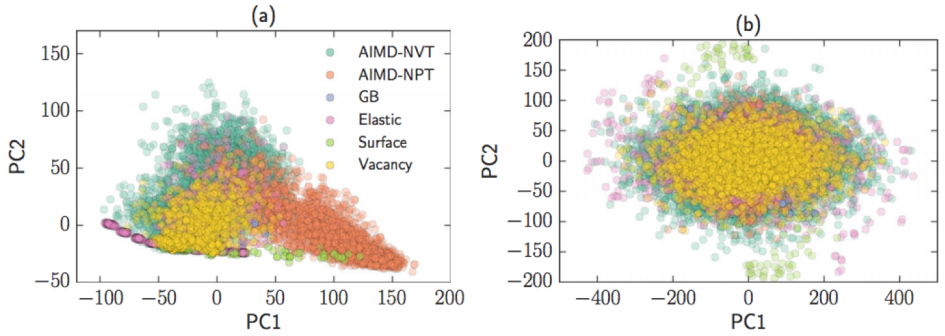


FIG. 2. Two-dimensional projection of the principal components of (a) the atomic bispectrum coefficients and (b) their first derivatives.

Resources

- Berkeley Research Computing Savio Cluster
- Jupyter notebooks on Jupyterhub installation on the cluster

Partition	Nodes	Node List	CPU Model	# Cores/Node	Memory/Node	Infiniband	Speciality	Scheduler Allocation
savio	164	n0[000-095].savio1 n0[100-167].savio1	Intel Xeon E5-2670 v2	20	64 GB	FDR	-	By Node
savio_bigmem	4	n0[096-099].savio1	Intel Xeon E5-2670 v2	20	512 GB	FDR	BIGMEM	By Node
savio2	136	n0[027-162].savio2	Intel Xeon E5-2670 v3	24	64 GB	FDR	-	By Node
savio2	4	n0[183-186].savio2	Intel Xeon E5-2680 v3	24	64 GB	FDR	-	By Node
savio2	4	n0[290-293].savio2 n0[230-240].savio2	Intel Xeon E5-2650 v4	24	64 GB	FDR	-	By Node
savio2	35	n0[187-210].savio2 n0[230-240].savio2	Intel Xeon E5-2680 v4	28	64 GB	FDR	-	By Node
savio2_bigmem	20	n0[163-182].savio2	Intel Xeon E5-2670 v3	24	128 GB	FDR	-	By Node
savio2_bigmem	8	n0[282-289].savio2	Intel Xeon E5-2650 v3	24	128 GB	FDR	-	By Node
savio2_htc	20	n0[000-011].savio2 n0[215-222].savio2	Intel Xeon E5-2643 v3	12	128 GB	FDR	HTC	By Core
savio2_gpu	17	n0[012-026].savio2 n0[223-224].savio2	Intel Xeon E5-2623 v3	8	64 GB	FDR	4x Nvidia K80	By Core
savio2_1080ti	3	n0[227-229].savio2	Intel Xeon E5-2623 v3	8	64 GB	FDR	4x Nvidia 1080ti	By Core
savio2_knl	28	n0[254-281].savio2	Intel Xeon Phi 7210	64	188 GB	FDR	Intel Phi	By Node

Challenges

- Generated training data(~150 Gb) fills up the memory of a node and cause the execution to crash:

  Find ways to implement multiple node generation of training data and training

- Matrices are small in size so does not achieve good parallelism during training:

  Rewrite code to introduce batching. Then, use GPU (get allocation).

- If GPU be used, need to generate training data on CPU and then run on GPU. Lack of space of home directory storage (10 Gb allocated) so cant write it out. Could explore the use of global scratch but need to know scratch policy.