Supporting information for: "Active Learning The Potential Energy Landscape of Water Clusters from Sparse Training Data"

Troy D. Loeffler[1], Tarak K. Patra[1], Henry Chan[1], Stephen Gray[1], and Subramanian K.R.S. Sankaranarayanan[1,2*]

1. Center for Nanoscale Materials, Argonne National Lab, Lemont, Illinois 60439, United States

2. Department of Mechanical and Industrial Engineering, Illinois 60607, United States

E-mail: tloeffler@anl.gov,skrssank@anl.gov,skrssank@uic.edu

Randomly Generated Training Data

A second network with the same network structure was trained on a set of 426 configurations generated via Metropolis MC and Nested Ensemble sampling. This was done to match the final number of structures that the AL-ANN was trained on in order to show what a network trained without the AL scheme would perform. The results for this network can be found in Fig. S1. While according the energy plot the network performs reasonably well for the 20-50 and 51+ range, it is unable to predict the RDF of a cluster size of 50 with any level of accuracy. It is not sufficient to simply generate the same number of configurations. The configurations must be representative of a wide variety of unique configurations in order to give the network enough information to correctly sample .

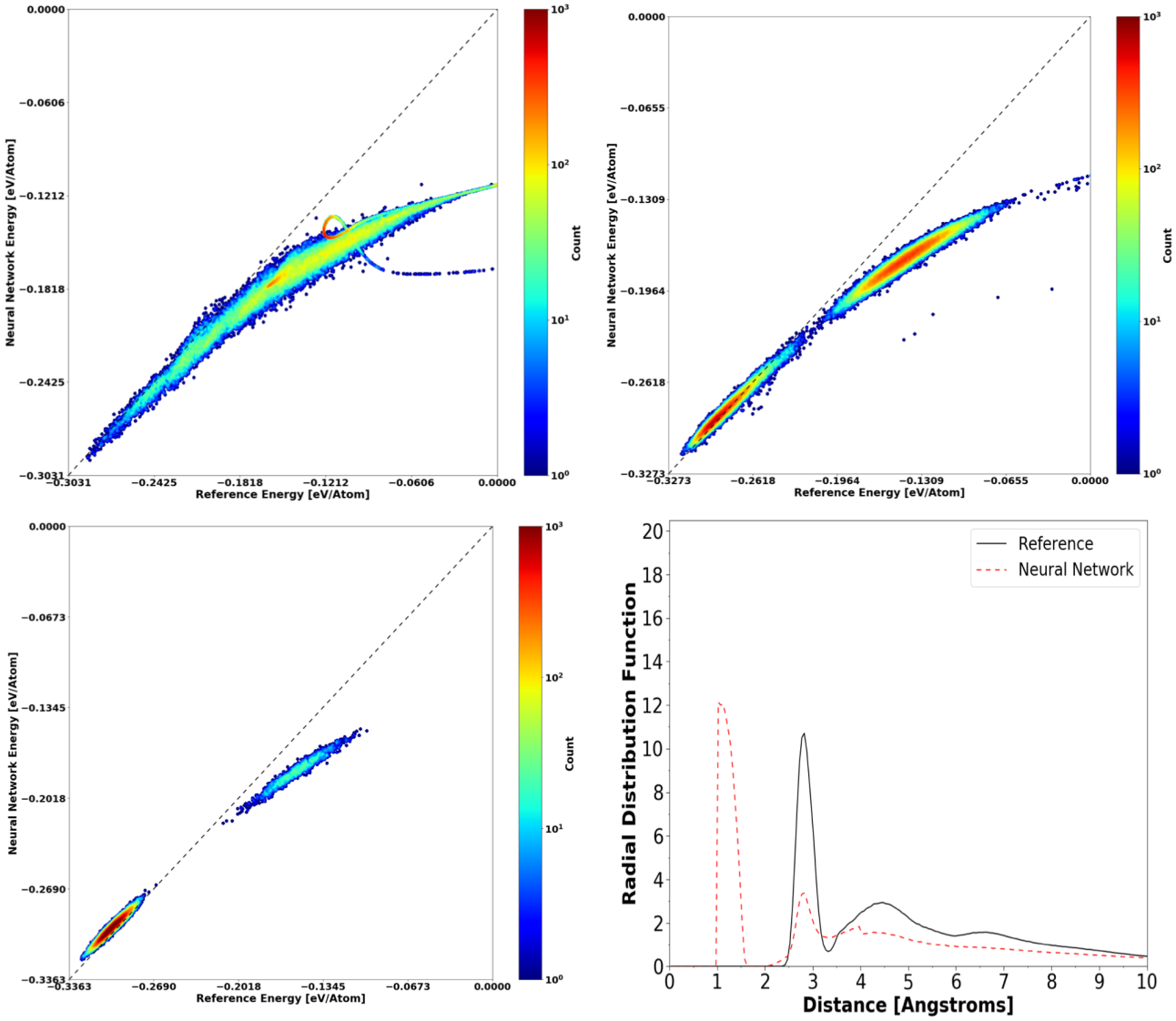

Figure S1: The results from a neural network trained on a randomly generated set of 426 structures is compared against the same 140,000 structure test set as the AL-ANN potential. The prediction vs reference energies are plotted for the 1-20, 21-50, and 51+ cluster size ranges in the top-left, top-right, and bottom-left panels respectively. The RDF for a simulation of a fixed cluster size of 50 (dashed red curve) is plotted on top of the reference RDF (solid black line) in the bottom-right panel.