Iowa Initiative for Artificial Intelligence
Final Report

Project title:	Machine learning long-range interactions in metals for drug discovery and semiconductor physics
Principal Investigator:	Principal Investigator: James Shepherd, Assistant Professor, Department of Chemistry
Prepared by (IIAI):	Avinash Mudireddy
Other investigators:	Laura Weiler, Undergraduate Student, Department of Chemistry (research) & Departments of Physics and Computer Science (academics) Tina Mihm, Graduate Student, Department of Chemistry
Date:	11/10/2021

Were specific aims fulfilled:	Yes
Readiness for extramural proposal?	N/A
If yes … Planned submission date	N/A
Funding agency	N/A
Grant mechanism	N/A
If no … Why not? What went wrong?	Grant funding awarded after IIAI project award and prior to beginning of this IIAI research

Brief summary of accomplished results:

Research report:

Aims (provided by PI):

Energy calculations are the way we as computational chemists can examine the stability of a property of a material. For example, we understand from the world around us that silicon is a semiconductor (a property that enables its use in microelectronics). At a submicroscopic level this means a phase which allows electrons to flow is lower in energy (and thus more stable) than one which does not allow electrons to flow. Quantum mechanics is required to calculate which phase is more stable. This presents a dilemma: The high-accuracy calculations we would want to perform are too expensive and the lower-accuracy calculations which we can afford are not accurate enough. Our objective is to use machine learning (ML) to bridge the divide between the high-accuracy method and the low-accuracy method.

Quantum mechanical systems are said to be comprised of molecular orbitals which are functions that each contribute part of the energy of a system. Following an approach due to Wellborn, Cheng, and Miller (2018)1, our input variable (X) will be the low-accuracy method's energy per orbital, and the intrinsic energy of the orbital, which is easily calculable. Our output variable (Y) will be the prediction of the high-accuracy method's energy per orbital. Following our calculations, we will be able to combine these energy fragments into a total energy for the system to study the stability of material properties.

The long term targets of this work are in drug discovery and semiconductor physics. We have a new collaboration with Mike Schnieders (Biochemistry) and I am interested in inviting Fatima Toor (Electrical and Computer Engineering) to join this collaboration as experts who will support our work in these two fields.

Data:

The dataset contains 1803 data points from a model system (input/output pairs, see preliminary data). When we are applying these to real systems, the input variables take 2 minutes to generate. If the model system data is not sufficient, new training data points can be calculated in approximately 7 hours for 30 data points. These timing data are from 1 node on Argon.

The input variables X contain momentum transfer [G], transition structure factors from a low-level (MP2) algorithm [MP2 S(G)], two-electron integrals [MP2 V(G)], minimum momentum transfer resolution [minG], Wigner-steitz radius used to parameterize the density of the system [rs], energy associated with the momentum transfer [E], and the number of momentum transfers [n_G]. The output variable Y is the energy component associated with the high-level (CCSD) algorithm.

AI/ML Approach:

We propose to use an Artificial Neural Network (ANNs) to solve this regression-based prediction problem. The ANN model is an intelligent system of nodes similar to the network of neurons in the brain. It is a computational model inspired by the brain which is capable of learning and pattern recognition. In biological systems, the learning happens through adjustments to synaptic connections between the neurons, similarly, in the ANN model, this happens through connections between the nodes/neurons through weights. ANNs are used to solve complicated problems in many applications such as optimization, prediction, modeling, clustering, pattern recognition, simulation, and others.

The ANN consists of three layers: the input layer which collects the input data, an output layer that computes the desired outcome, and one or more hidden layers connecting input and output layers. These hidden layers construct the non-linearity relationship between inputs and outputs through their weights. Each layer constitutes neurons which are a basic processing unit of the network. A neuron takes the inputs, multiplies them with connection weights, adds the bais, and passes it through an activation function to produce the output. In this manner, each neuron produces an output which is passed as input to neurons in the next layer through a feed-forward mechanism to compute a final outcome. However, the learning happens through a mechanism called backpropagation. The error is computed by calculating the differences between the actual outcome and the computed outcome. This error is propagated back to the network to update the weights, such that in the next iterations the error is minimized. Fig 1. Illustrates this process.

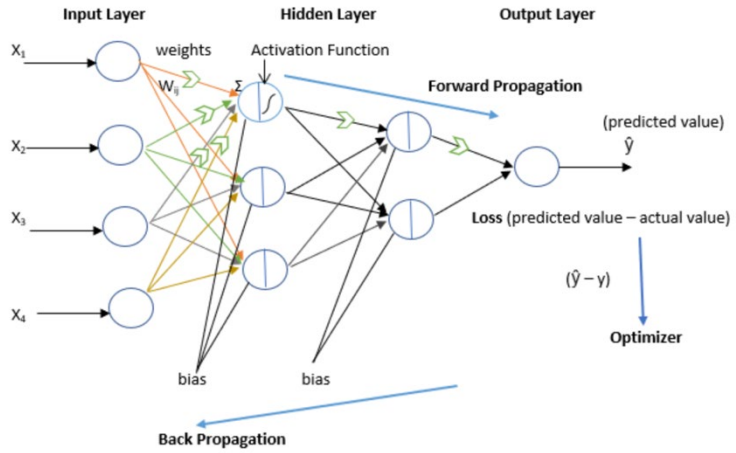

Fig 1. Artificial Neural Network with Backpropagation. [Image source: https://inblog.in/Part-1-Artificial-Neural-Network-Theory-ETDChQaKty]

In a regression problem, the measure of the performance of the network is calculated through indicators used as error assessment techniques like MSE, RMSE, MAPE, MABE, r, and R2.

Experimental methods, validation approach:

Model design:

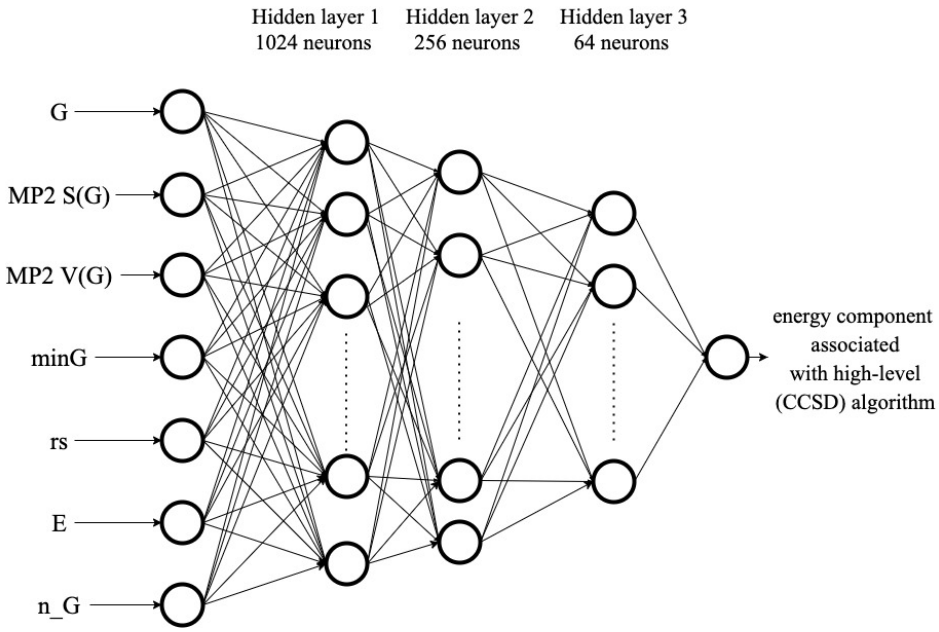

Fig 2. ANN model

As a part of model settings, we take 7-dimensional inputs to predict the 1-dimensional output. The dataset is power transformed for better model prediction with a train-validation-test split of 72:18:10.

Our neural network which is illustrated in Fig 2. has 3 hidden layers with 1024, 256, and 64 neurons fired by “relu” activation. For each layer, we added batch normalization and dropouts(0.1). The output layer has “linear” activation.

We used Adam optimizer and Huber loss as optimizer and loss function respectively. The model is trained for 500 epochs with 128 as batch size.

Ideas/aims for future extramural project:

The machine learning line of study in the Shepherd Group was funded by NSF CAREER between IIAI acceptance and the project’s start.

As a consequence, the results of this pilot project (while very positive and promising) are not reported here and wil be reported as part of the extramurally funded NSDF Career grant.