Machine learning of committor functions in the Ising model

Dr D. Quigley
Physics Theory Group

D.Quigley@warwick.ac.uk

The rates at which crystals nucleate and grow from a supersaturated solution are essential inputs to solidification models in a variety of contexts. These include materials synthesis and processing, and understanding the formation of harmful biological crystals such as kidney stones. In principle, these rates can be obtained from atomistic computer modelling. Unfortunately the timescales involved are generally inaccessible to "brute force" simulation. Instead one resorts to biased simulation techniques (umbrella sampling, metadynamics and others), or the rate can be computed via path sampling approaches (transition interface sampling, milestoning and more).

All of these methods rely to some extent on dimensionality reduction - identifying a scalar function (sometimes called a reaction coordinate) of all degrees of freedom which quantifies progress along the nucleation and growth pathway. It is argued that the optimal choice for this function is the *committor* p_B - the probability that a given configuration of the system will evolve to the final crystal state under some appropriate choice of stochastic dynamics, i.e. that it will reach state B (the crystal) before state A, the parent phase from which the crystal nucleates.

Calculation of the committor is hugely expensive. For each configuration of the system, one must simulate a large number of trajectories to reduce the statistical uncertainty associated with the corresponding estimate of p_B. Routine calculation of p_B is hence impossible and one normally resorts to somewhat arbitrary choices for the reaction coordinate. If this committor could be computed rapidly from a given input configuration, predictions of nucleation rates (and kinetics of rare events in general) would be greatly improved.

There has been long-standing interest in using machine learning approaches to calculate p_B. This has been tested on small problems involving conformational changes in small molecules [1], but never in the context of crystallisation. The proposed project will explore this possibility using the simplest possible nucleation problem - the formation of "crystals" in the Ising lattice gas model. Here the parent phase is represented by a sparsely populated cubic lattice under thermodynamic conditions (temperature and chemical potential) at which the "crystal" phase (fully occupied lattice - fig 1 right) is most stable.

Such calculations are certainly feasible (at least for small lattices) due to two developments. Firstly one can rapidly calculate the committor directly for such systems using GPUs. The Ising model is sufficiently simple that one can evolve the stochastic dynamics of the lattice using a single CUDA core, allowing thousands of trajectories to be generated concurrently on a single GPU device [2]. Training/validation data for the neural network can therefore be generated much more rapidly than for any detailed chemical model.

Secondly, software to train neural networks from high-dimensional input data (such as Keras + TensorFlow) is now readily available.

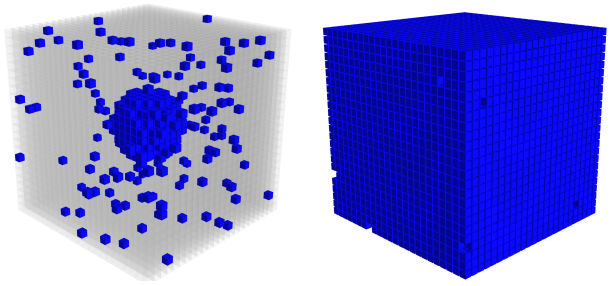

Figure 1: Snapshots of a "crystal" nucleus within a saturated lattice gas (phase A - left) and a crystalline configuration (phase B - right). The commitor p_B is the probability that a configuration such as that on the left will reach a crystalline configuration.

Inputs to the network could be the complete Ising configuration (i.e. list of occupied lattice sites), or one could explore the most important feature inputs. For example the number and size/shape distribution of clusters. Other machine learning techniques may be more expedient (e.g. Gaussian processes) and the student will have freedom to choose the most appropriate.

The miniproject will follow the usual approach of generating data, using the GPU-based lattice gas implementation. This will be separated into training and validation sets. Of particular interest is an understanding of how many trajectories are required for each sample of the committor, how many samples are required for a given accuracy, and how to optimally sample input configurations. For example, training data involving multiple large nuclei is unlikely to be relevant as such configurations are statistically insignificant along the most probable nucleation trajectories. It will also be important to ensure that the statistical uncertainty in the training data is correctly propagated into the estimate of the committor.

A potential follow-on PhD project will link into an active EPSRC programme grant on realistic crystallisation. Here we would use the calculated committor to accelerate minimal lattice-based simulations which explore how competition between multiple nuclei is modified (and potentially controlled) by the presence of spacial and temporal inhomogeneities in temperature and concentration. We are also interested in how introduction of additional lattice species (e.g. templating surfaces or impurities) can select one crystal polymorph over another. Ultimately we aim to use such simulations as proxy models within a predictive modelling framework for polymorph selection.

The project would suit a student interested in computational statistics, machine learning and high performance computing. A strong background in computer programming (in any language) and statistics would be beneficial.

References

[1] A. Ma and A. R. Dinner, *The Journal of Physical Chemistry B* **109**, 6769 (2005).

[2] J. Gross, J. Zierenberg, M. Weigel, and W. Janke, *Computer Physics Communications* **224**, 387 (2018).