AI Hackathon in Molecular Dynamics

Organizers

Juan de Pablo Eliu Huerta Ian Foster
University of Chicago and Argonne National Laboratory

Mike Papka
Argonne Leadership Computer Facility

Tom Gibbs
NVIDIA

Volodymyr Kindratenko
National Center for Supercomputing Applications
University of Illinois at Urbana-Chamapaign

Coordinators

Ludwig Schneider Joshua Mysona Pablo Zubieta
University of Chicago

Prediction of Randomly Sequenced Copolymer Properties

1 Introduction

Polymers are macromolecules comprised of many connected repeat units known as unimers. Both the chemical structure of unimer, composition, and sequence of unimers may be varied, resulting in a virtually infinite array of macromolecular structures with properties tuned by changing these three variables. An example of this diversity of function here are diblock copolymers which consist of a linear series of unimers, that can be thought of as two distinct blocks. The first block of this macromolcule consists of a linear chain of a single unimer type. The second block, in contrast consists of a linear series of unimers that may have many different chemical identities in a particular sequence. Systems of these aggregates form ordered lamellar structures with a fixed periodicity determined by the sequence and chemical identity.

While this periodicity is well described for simple diblock cases consisting of where the secondary block is of a single bead type (thus eliminating sequence and composition as a variable), for blocks with more than one bead type with myriad sequences, polymer physics struggles to provide a cohesive model. We thus turn to machine learning in an attempt to develop models capable of describing the behavior of such diverse systems.

An ideal ML model should accomplish two goals.

- The ability to accurately predict the periodicity of the lamellar system with given molecular sequence, and chemical nature
- Insights into what combination of beads and chemical identity are associated with particular degenerate configurations

Training Data Provided We provide several data sets for training. Each point in each dataset consists of a lamellar period, thirty two sequential monomer identities, and a parameter characterizing chemical interaction. Datasets A, B, and C each isolate the effect of bead sequence, and consist of 768 data points each, though share an interaction parameter. Dataset D contains datapoints with both varied sequence and chemical interaction. Figure 1 visualizes several of these structures, but the exact structural parameters and individual bead positions in space, are not provided.

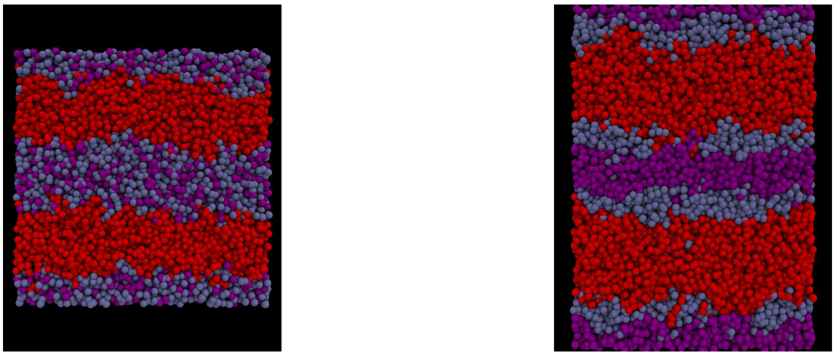

Figure 1: Example two different lamellar configurations for different sequences. The lamellae are clearly visible, but in one case only two layers are formed, whereas in the second three distinct layers are visible. The lamellar period of the three layer case is larger than that of the two layer case

What an AI model should do Predict the lamellar period as a function of sequence and chemical parameter, and output which features or combination of features in the data most contributes to the output lamellar period. Existing models of polymer physics utilize collective variables based on the sequence and blockiness. It is desirable that the ML model in addition making predictions help provide insights into which patterns in the data lead to which behavior. In general mixing different bead types leads to a depression in the lamellar period. However for certain rare combinations, mixing the beads leads to a larger period. It is desirable to determine what combination of parameters leads to these deviant structures.

1.1 A winning solution

We want you to come up with creative solutions for this challenge. This section can serve you as a measure of your success.

- physical correctness
  We are happy to help you to understand some of the physics. Having a solution that correctly predicts the physics is important. "Correctly predicts the physics" generally is a phrase which means the model is generalizable on other data outside the training data. However there exist known physical parameters that are a reduction of the dimensionality that are known in other systems to predict polymer behavior
  - volume fraction $\hat{\varphi}$.
    For these systems, the fraction of A, B, and C beads is used to predict polymerphase behavior, but alone is not enough to predict the lamellar period in this complex system. We leave A fraction fixed in all cases, whereas the B and C fractions are variable

  - blockiness λ.
    Sequences in which multiple beads of the same type follow each other, called blockiness, are known to exhibit cooperative effects leading to anomalous results. It may be beneficial to add this parameter to the model as the quantity $p_{BB}p_{CC} - p_{BC}p_{CB}$ where p_{ij} is the fraction of beads j that follow bead i in the system

  Examples of the data represented with these parameters are shown in Fig 2

- presentation quality.
  Results should be summarized in a slide deck (.pptx or .pdf), and with code (Jupyter notebooks) documented in a GitHub repository.

- computational efficiency.
  Training and inference efficiency should be reported.

- Reduced dimensionality: In the spirit of polymer physics, it is desirable to reduce the data, which is 33 dimensional, to a smaller subset of relevant dimensions or functions of the data which each play a larger role than the individual identities. This requirement will also aid in presentation of the data.

2 Technical Description

The data for the polymer sequence can be encoded as a simple binary vector, while the associated value for the BC bead type interactions in simulation is a simple scalar which ranges from zero to 3.5. The data set is packaged as a dataset and will be available for users. Alternatively the data may be obtained in raw text format via git. Four distinct datasets are available. Three isolate sequence effects for three sample chemistries, while the fourth incorporates different chemical identities. Prior efforts have enlisted a simple integer scheme for the sequence (ie 1 for species A, 0 for species B) while treating the scalar independently. Each data point consists of 32 integer bead identities, a period length, and the chemistry dependent scalar.

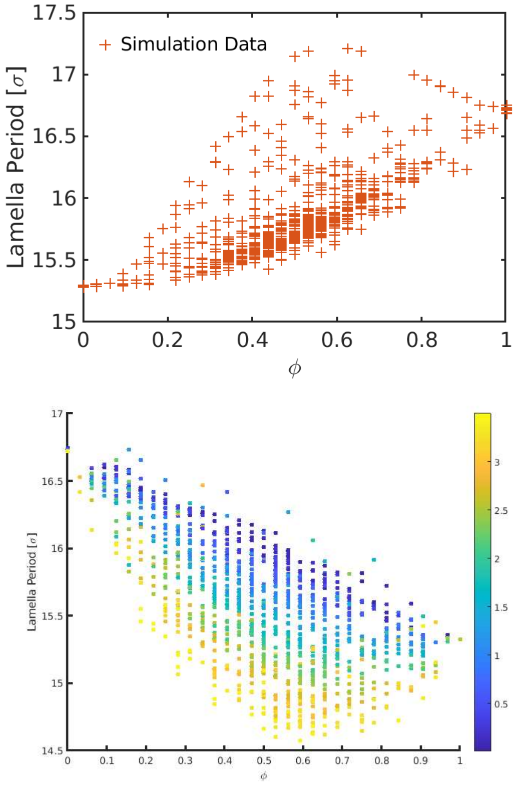

Figure 2: Examples of two datasets, compressed on the dimensions of lamellar period, ϕ, which is the fraction of B beads in the random block, and in the case of the second image, α which measures chemical compatibility.

Prediction of Diblock Copolymer Morphology Dynamics

1 Introduction

Self-assembly of soft matter provides a practical and scalable route towards the production of nanostructured materials. Symmetric diblock copolymers, which self-assemble into a lamellar phase, can be considered as a prototype for this class of materials.

A range of sophisticated simulation techniques have been developed over the past two decades to describe microphase separation in polymeric materials. It is now possible to predict their behavior with high fidelity over short to intermediate length scales, ranging from nanometers to microns. Their underlying free-energy landscape, however, is riddled with metastable states that are separated by large barriers. Overcoming such barriers through the natural temporal relaxation of these materials, however, continues to represent a formidable computational challenge.

In this project, we aim to create a machine learning model to predict long-time kinetics of a symmetric diblock copolymer morphology based on density configurations on a grid (a one-channel image). We had success with deep convolutional networks and Fourier inverted networks to predict long-time trajectories for large 2D systems. [L. Schneider and J. J. de Pablo, Macromolecules, submitted (2021), preprint attached]. However, the deep-convolutional network nature is computational and memory intensive for 3D systems. So challenging that the memory demand exceeds the memory capacity of NVIDIA A100 GPUs, which makes training unfeasible.

Challenge Optimize 3D models to make it fit into an A100 GPU, or to distribute the training to multiple GPUs. Novel AI architectures may also be explored so gain new insights into the mechanisms of defect kinetics and/or make the model transferable.

Training Data Provided We provide 1248 independent samples.Each sample consists of 9 time steps from total disorder towards an ordered state and is described by a normalized density $\varphi \in [0, 1]$ on a $64 \times 64 \times 64$ grid. The underlying model of diblock copolymers results in a lamellar morphology. However, the free energy barriers are too high to easily overcome by traditional simulations. Instead, the system remains stuck in a local minimum where different orientations of lamellae are separated by grain boundaries and morphological defects. Figure 1 visualizes one of these samples from a later time.

What an AI model should do Predict morphological kinetics. We had success in 2D with the approach of learning the transition from $t_i \rightarrow t_{i+1}$ with a deep convolutional

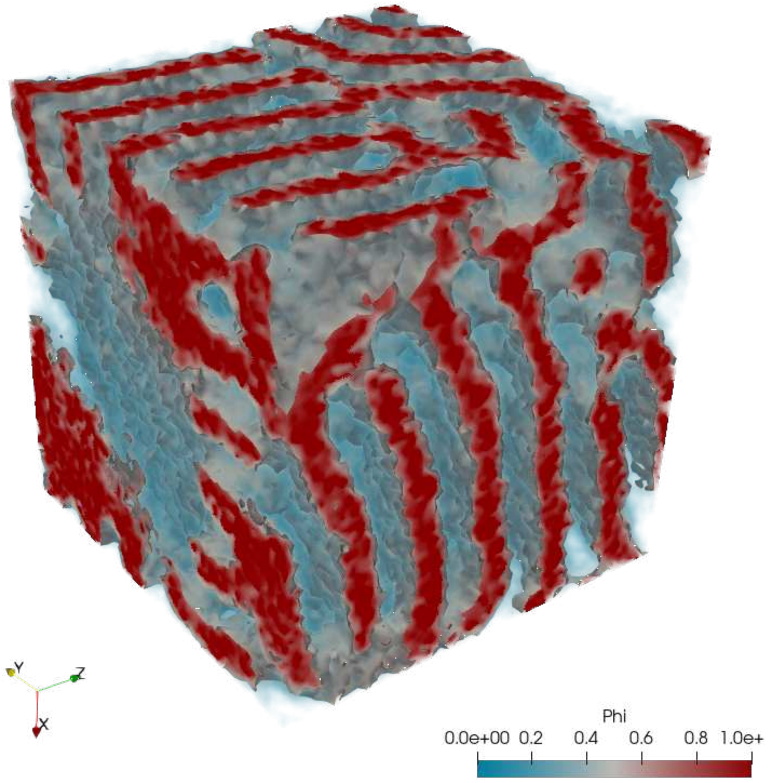

Figure 1: Example of a lamellar configuration in a local minimum. The lamellae are clearly visible, but the simulation box does not show a global orientation of the lamellae. Instead multiple orientations are separated by grain boundaries and defects.

neural network. Repeated application of this transition renders a trajectory and thus complete time evolution. We exploit that the transition is time independent, so we can learn from the transition $t = 8 \to t = 9$ all subsequent transitions $t = n \to t = n + 1$. Additionally, we assume the transition to the Markovian, such that only the last state $\phi(t = n)$ is necessary to transition to the next state $\phi(t = n + 1)$.

The attached preprint explains more detailed some of the mentioned relations and how the assumptions are justified. It also describes a machine learning model for 2 dimensions and can serve as inspiration.

1.1 A winning solution

We want you to come up with creative solutions for this challenge. This section can serve you as a measure of your success.

- physical correctness
  We are happy to help you to understand some of the physics. Having a solution that correctly predicts the physics is important. However, we provide you with a few simple tools implemented in python to quickly assess some of the most important properties.
  - volume fraction $\hat{\varphi}$.
    We have conservation of mass, so the ratio between A and B remains constant at 1/2.
  - structure factor $S(q)$.
    The structure factor can be used to quantify the morphology – more details in the preprint. The structure factor is always strongly peaked and the peak position should stay fixed.
  - correlation length ξ.
    From the structure factor, we can determine the correlation length of the lamellar order. The correlation should in create as a power law in time $\xi(t) \propto t^{\eta}$. More details are in the preprint.
- length scale independence.
  The training data comes in a fixed $64 \times 64 \times 64$ matrix format, but the trained model should be able to predict on larger inputs (and outputs). The physical correctness should still be satisfied for larger inputs.
- trajectory predictions.
  It should be possible to apply the model on its output to obtain a trajectory in time. The result should be stable and physical correctness ensured.
- presentation quality.
  Results should be summarized in a slide deck (.pptx or .pdf), and with code (Jupyter notebooks) documented in a GitHub repository.

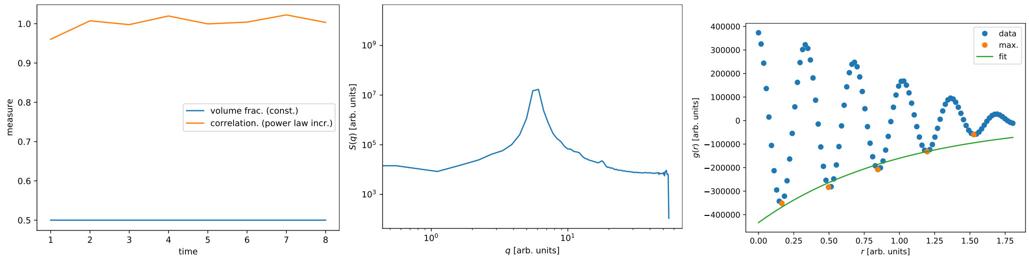

Figure 2: Demonstration of physical characteristics of a single simulated trajectory. Averaging over an ensemble of trajectories is highly encouraged. From left to right: i) Simulated time evolution of the volume fraction φ and correlation length ξ. ii) Structure factor in log-log plot with a dominant peak for time $t = 8$. iii) Corresponding spatial correlation and fit for the correlation length ξ.

- computational efficiency.
  Training and inference efficiency should be reported.

- bonus: transferability: The provided data is for a fixed volume fraction $\varphi = 1/2$ and incompatibility parameter. The resulting morphologies change with both of these parameters. If you want to make the model trainable with these two parameters changing contact us and we can provide more details and data.

2 Technical Description

The data is compiled in an HDF5 binary data container. `https://hdfgroup.org`, which has a straight forward python API: `h5py` `https://docs.h5py.org/en/stable/` Bindings for other programming languages are available notably C. The data is sorted in datasets with the handle `timeT`, where `T` is $0 - 9$ and represents time steps. Each dataset has the dimensions of $M \times N \times N \times N$, where $M = 1248$ is the sample index and $N = 64$ stands for the spatial dimensions in X, Y, and Z. The spatial boundary conditions are periodic. The value of each image is in the range $\varphi \in [0, 1]$ and represents the ratio of A material at this point in space. The m sample index is synchronized between time steps T such that the simulated trajectory can be reconstructed.

The python script `tools.py` offers helper functions that can calculate volume fraction, structure factor, and correlation length. The Figure 2 shows the result of the helper function with an example of a simulated trajectory.