Inside the Black Box: A Physical Basis for the Effectiveness of Deep Generative Models of Amorphous Materials

Michael Kilgour[1] *and Lena Simine*[1*]

[1]Department of Chemistry, McGill University, 801 Sherbrooke St. W, Montreal, Quebec, H3A 0B8, Canada.

We have recently demonstrated an effective protocol for the simulation of amorphous molecular configurations using the PixelCNN generative model (J. Phys. Chem. Lett. 2020, 11, 20, 8532). The morphological sampling of amorphous materials via such an autoregressive generation protocol sidesteps the high computational costs associated with simulating amorphous materials at scale, enabling practically unlimited structural sampling based on only small-scale experimental or computational training samples. An important question raised but not rigorously addressed in that report was whether this machine learning approach could be considered a physical simulation in the conventional sense. Here we answer this question by detailing the inner workings of the underlying algorithm that we refer to as the Morphological Autoregression Protocol or MAP. We identify the key object of physical interest for modeling of amorphous structures: an all-order correlation cluster expansion that fully captures structural information for amorphous substances, outline how it may be efficiently modelled by a neural network and study the convergence properties of a discrete, autoregressive sampling protocol guided by such a model. We find that such a MAP sequence constitutes a converging Markov process, guaranteed to realize a unique equilibrium distribution, and illustrate relevant concepts with abstract toy-model numerical experiments. This work lays the theoretical foundation for physically sound autoregressive sampling.

Keywords: Deep learning; amorphous materials; generative models; convolutional networks; Markov chains

I. Introduction

Currently, there is growing interest in machine learning approaches for large-scale sampling of morphologies of chemical systems, sidestepping traditional structural sampling methods such as molecular dynamics and Markov chain Monte Carlo to directly sample target physical distributions in novel ways. Such protocols come with several unique advantages over traditional simulations; recent studies, including our own work with PixelCNN-based autoregressive models[1] have demonstrated that appropriately trained generative models can draw accurate samples from the configuration spaces of proteins[2] and 2 and 3 dimensional amorphous aggregates of carbon and silicon[1,3,4], all without the complexities which may attend long-time and/or large-scale traditional simulation of the same system. At this time, we just at the beginning of our exploration of the potential of this class of methods for analysis and discovery of new materials and macromolecules.

The key insight postulated in our previous work[1] was that by exploiting the finite-range of structural correlations inherent to amorphous materials, one may train a generative, autoregressive model on samples with a maximum size on the order of that length and use it to accurately simulate samples of unlimited size. In general, we refer to this type of algorithm as a morphological autoregressive protocol (MAP), which was implemented in this case via the Gated PixelCNN architecture[5]. Beyond showing the empirical evidence that such a method generates convincing molecular configurations, there are important questions about *how* and *why* it is able to do so, and whether the generated configurations are simulated in the conventional sense of them being sampled from a physical distribution. What kind of physical information does such a model contain? What are the conditions within which the model will generate accurate samples? What kind of sampling process does this type of potentially multidimensional sequence embody? In this work, we explain the theoretical underpinnings of a morphological autoregressive protocol and present rigorous answers to these questions.

Our work to date has been focused on a particular MAP architecture known as PixelCNN. PixelCNN is an autoregressive convolutional generative model which can be used to sample a chemical morphology upon which it has been trained, iteratively, at a very low cost[6,7], up to practically arbitrary size. It is further architecturally flexible, easy to train, and naturally admits the use of conditioning variables[5]. A primary advantage of this, and other MAP approaches is that, via straightforward correlation modelling, it cheaply computes human interpretable sample probabilities for any size of sample. We consider PixelCNN as the prototypical MAP architecture, and will demonstrate its strengths in Section IV, though other suitable methods exist[7,8].

The paper is organized as follows: in Section II we outline the concept of a morphological autoregressive protocol and develop an expression which defines its propagation, in Section III we demonstrate its properties as a sequence generator on a discrete grid, in Section IV we illustrate key features of MAP sequences via numerical experiments, and we conclude in Section V.

II. The Morphological Autoregression Protocol (MAP).

A. Morphological Sampling as Sequence Propagation

We define a morphological autoregression protocol as an approach which casts morphological sampling as the generation of some sequence, $\{X\}$, encoding information about a corresponding physical system, most commonly in the form of particle positions. Such a sequence advances according to the general form,

Equation 1

$$X_i = \psi\left[c_b + F\left(\{X_{j<i}\}\right)\right],$$

with c_b a linear bias, $F(\{X\})$ accounting for the correlations between X_i and previously generated points $\{X_{j<i}\}$, and $\psi[x]$ a function which parses these correlations into outcomes for the atomic/molecular structure at point i. The precise form of $\psi[x]$ depends on the way this structure is encoded in X_i, and will be discussed in the next subsection, IIB.

Allowing the index i run over multiple dimensions, we may generate N-dimensional sequences corresponding to structures of real materials. We are particularly interested in materials which have a finite morphological correlation length, i.e., amorphous materials. For such materials, the length of the 'history' which must be considered may be truncated at some finite correlation length, L_c, without loss of accuracy,

Equation 2

$$\lim_{i\to\infty} F\left(\{X_{j<i}\}\right) = F\left(\{X_{j\in L_c}\}\right).$$

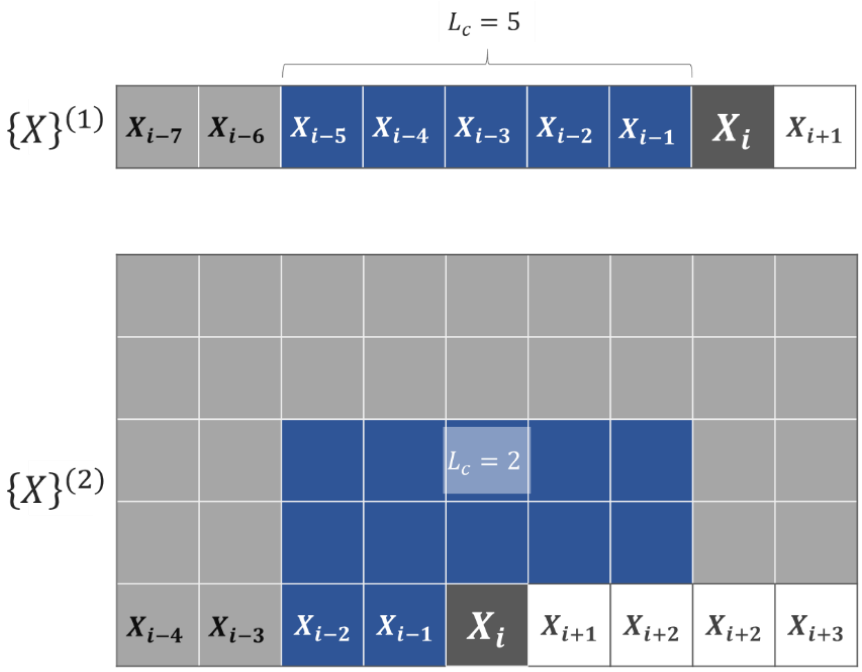

Figure 1: Illustration of the amount of context required to predict the sequence value at X_i, for sequences defined on a 1D or 2D grid. The maximum size of context required to accurately predict X_i is known as the correlation length, L_c.

To sample some target physical distribution, one has only to propagate this sequence, with the statistics of the physical system encoded in $F(\{X\})$ and c_b. As the sequence progresses, it will simulate the physical system, that is, it will draw samples from the target distribution. We will demonstrate in the next subsection that direct evaluation of $F(\{X\})$ is unfeasible for most practical purposes. Therefore, in practice, the ability of such a sequence generator to produce accurate, unbiased, and well-distributed samples depends on accurate *modelling* of $F(\{X\})$.

A popular family of approaches to model $F(\{X\})$ are generative autoregressive models, including PixelRNN, PixelCNN and their derivatives[5,9,10]. These models leverage the expressiveness and generality of deep neural networks to efficiently model the statistics of a target physical distribution with minimum human intervention. A key advantage of MAP architectures as we will see in the next section is that they produce explicit sample probabilities for samples of arbitrary size at no additional cost. Combining this with the physical intuitions also outlined in Section IIB makes simulation via MAP largely end-to-end human interpretable. Interpretability is less straightforward in non-autoregressive approaches to direct sampling such as autoencoders[11] or generative adversarial networks[12]. They fulfill similar overall functions and have certain advantages, particularly in the rate of sample generation[6,7], but they do not as easily admit computation of sample probabilities, or rationalization of particle positions based on local arguments. Since we are interested in providing the rationalization for the physicality of the generated molecular configurations, we work with models that are as interpretable as possible.

B. Encoding the Correlation Cluster Expansion in the MAP

The above definition for a MAP is very general and does not tell us e.g., the range of possible values or physical meaning of sequence elements X_i; many representations and interpretations are possible and potentially desirable in different cases. We will proceed using the following definitions, carried over from our previous work with PixelCNN[1] which provide practical advantages for keeping this method straightforward and applicable to a wide range of computational and experimental datasets:

1. The sequence $\{X\}$ maps to a discrete N-dimensional spatial grid, which is filled elementwise, typically via Raster scan, using contextual information up to a maximum range of L_c from the predicted element.
2. X_i represents the occupation of grid point i, with N_c different possible classes of occupants e.g., different atomic elements, nanoparticles, or molecular fragments, being represented via different integer values of X_i.

A MAP following these definitions sequentially outputs integers representing morphological features on a discrete spatial grid. To advance the sequence, we begin by computing the probability for the next sequence element to be of class θ, for all possible classes N_c,

Equation 3

$$p\big(X_i = \theta | \{X_{j<i}\}\big) = N\big[c_{b,\theta} + F_\theta\big(\{X_{j\in L_c}\}\big)\big],$$

with $\{\theta \in \mathbb{Z}^+ | \theta \leq N_c\}$, and $N[\{x\}]$ the softmax classwise normalization function,

Equation 4

$$N[x_i] = \frac{e^{\beta x_i}}{\sum_j e^{\beta x_j}},$$

where we take the effective inverse temperature, $\beta = 1$. This produces normalized, nonzero probabilities for each possible class, $\sum_{\omega=1}^{N_c} p\left(X_i = \theta_\omega \middle| \{X_{j\in L_c}\}\right) = 1$, which we sample to advance the sequence,

Equation 5

$$X_i \sim p\left(\{X_{(j<i\}}\}\right).$$

This element then becomes part of the context, $\{X_{j<i+1}\}$, used to predict the next sequence element, and so on.

These definitions yield a very general and easily interpretable sequence and allow us to explain the operation of a MAP in a clear and practical terms. One may, however, choose to work with different definitions, most pertinently in a continuous coordinate basis and/or with a continuous space of outputs. Though either of these could be approximated with a sufficiently dense grid of discrete values, it bears discussing how one might analyze a MAP sequence in a continuous basis. The discussion of correlation expansion which takes up the rest of this section is easily adapted to continuous space, with the correlation coefficients redefined as continuous functions in that space. However, the possibility of infinitely high-order correlations may make their modelling unwieldy. It should also be noted that the propagation of such a sequence in continuous space may require additional adjustments, and the validity of our arguments will depend on the choices made at that stage.

To develop the concepts required to describe the simulation of complex systems such as amorphous molecular aggregates using a MAP, we will consider a series of increasingly complex physical systems. Our goal is to improve the clarity and interpretability of the generative deep learning process, often regarded as a black-box modelling technique, by gradually increasing the complexity from analytically solvable models to one that requires a deep neural network, while keeping track of the physical features that are being captured. As a foundation, we begin with one of the simplest possible physical distributions, the single-component ideal gas. Since there are no inter-particle interactions, particle positions are completely uncorrelated; specific knowledge of any or all particle positions (prior sequence elements) is worthless in predicting the occupation of any unknown grid points. $F_\theta(\{X\})$, which incorporates information from previous sequence elements, is zero for all classes and all configurations. For a given sequence element X_i, we can trivially predict its probability of being occupied or unoccupied ($X_i = 1, 0$) with only a linear bias term. Also note that since there are only two classes in this system, it is sufficient to explicitly model the correlations for only one of them and infer the other, $p(X_i = 1) = 1 - p(X_i = 0)$,

Equation 6

$$p(X_i = 1) = N\left[c_{b,1}\right] = \rho$$
$$p(X_i = 0) = N\left[c_{b,0}\right] = N\left[-c_{b,1}\right] = 1 - \rho$$

with ρ as the grid occupation density for particles, and zero dependence on any prior element of the sequence $X_{j<i}$. Substituting the softmax normalization for $N[\{x\}]$, we have,

Equation 7

$$\frac{e^{c_{b,1}}}{e^{c_{b,1}} + e^{-c_{b,1}}} = \rho$$
$$c_{b,1} = \frac{1}{2}\ln\left(-\frac{\rho}{\rho - 1}\right).$$

Thus, with only knowledge of average particle density, one could simulate this system ad infinitum, gridpoint-by-gridpoint, simply by sampling Equation 6 against a uniform random number $x \sim [0,1)$.

Equation 6 suffices for non-interacting systems but fails for systems wherein particle positions are correlated in any way. The class of systems with interacting particles includes most of those of interest in chemistry and materials science including crystals, glasses, and other amorphous materials, and therefore an equation which accounts for inter-particle correlations is required to simulate them. We turn to statistical mechanics for inspiration on how to proceed in the case of interacting particles. Consider the energy of a 3D, dilute, isotropic fluid with weak pairwise interactions (a *nearly* ideal gas):

Equation 8

$$E = KE_{ideal} + \langle U \rangle = \frac{3}{2}Nk_BT + \frac{1}{Z_N}\int drr^2u(r)g(r).$$

The first term KE_{ideal} is the ideal gas kinetic energy with N as particle number, T as temperature and k_B as the Boltzmann constant. The second term $\langle U \rangle$ is the perturbation due to pairwise interactions, with $u(r)$ as the pair potential, $g(r)$ as the radial pair correlation function and Z_N as the N-particle partition function.

We can introduce prior sequence elements in a similar way to neighboring particles, and write an expression which incorporates pairwise correlations between prior sequence elements, and the proposed N_c elements for X_i,

Equation 9

$$p\left(X_i = \theta | \{X_{j<i}\}\right) = N\left[c_{\rho,\theta} + \sum_{j \in L_c}\sum_{\omega}^{N_c} c_{j,\omega,\theta}\delta_{X_j,\omega}\right],$$

where $c_{\rho,\theta}$ contains the influence of average density for sequence elements, (e.g., particles) of class θ, and the sums account for correlations between gridpoint X_i, with proposed class θ, and prior elements X_j of class ω, weighted by coefficients $c_{j,\omega,\theta}$, with $\delta_{i,j}$ as the Kronecker delta function. This type of correlation function is suitable for simple systems, such as a dilute fluid where > 2 particles are unlikely to aggregate or even interact simultaneously.

In some such systems, direct evaluation of the coefficients, $c_{j,\omega,\theta}$ for the equilibrated system may be possible through knowledge of the pair potential. As an example, consider a dilute single-component fluid at very low temperature which interacts via finite-range repulsive potential. Since this, like the earlier ideal gas example is a single component system, it is sufficient to explicitly

model the correlations of only one of the classes. At equilibrium, particle positions will be strongly anticorrelated within the interaction range, R_c, and uncorrelated beyond this range. With this knowledge we can infer the results of Equation 9 for all possible particle environments, given $\theta = 1,0$ for occupied and unoccupied gridpoints respectively,

$$p$$

Equation 10

$$\left(X_i = 1 \middle| \{X_{j \in R_c}\}\right) = \begin{cases} 0 \; for \sum_{j}^{R_c} \delta_{X_j,1} > 0 \\ \rho_{eff} \; for \sum_{j}^{R_c} \delta_{X_j,1} = 0, \end{cases}$$

$$p$$

Equation 11

$$\left(X_i = 0 \middle| \{X_{j \in R_c}\}\right) = \begin{cases} 1 \; for \sum_{j}^{R_c} \delta_{X_j,1} > 0 \\ 1 - \rho_{eff} \; for \sum_{j}^{R_c} \delta_{X_j,1} = 0, \end{cases}$$

where ρ_{eff} is greater than the actual particle density ρ due the excluded volume $\propto R_c^3$ around each particle. The values of inter-particle correlation and density coefficients, $c_{j,1,1}$ and $c_{\rho,1}$, which fulfill the above conditions can be computed given the average density, grid spacing, and interaction range. In the simplest case, with $R_c = 1$ and $\rho \ll 1$, we can evaluate a simplified Equation 9 accounting for particle-particle correlations,

Equation 12

$$p\left(X_i = 1 \middle| \{X_{j \in R_c}\}\right) = N\left[c_{\rho,1} + \sum_{j \in R_c} c_{j,1,1} \delta_{X_j,1}\right],$$

and isolate the coefficients which produce the outcomes inferred above,

Equation 13

$$p\left(X_i = 1 \middle| \{X_{j \in R_c}\}\right) = \begin{cases} N\left[c_{\rho,1} + n \cdot c_{j,1,1}\right] for \sum_{j}^{R_c} \delta_{X_j,1} > 0 \\ N\left[c_{\rho,1}\right] for \sum_{j}^{R_c} \delta_{X_j,1} = 0, \end{cases}$$

Equation 14

$$p\left(X_i = 0 \middle| \{X_{j\in R_c}\}\right) = \begin{cases} N\left[-c_{\rho,1} - n \cdot c_{j,1,1}\right] for \sum_{j}^{R_c} \delta_{X_j,1} > 0 \\ N\left[-c_{\rho,1}\right] for \sum_{j}^{R_c} \delta_{X_j,1} = 0, \end{cases}$$

Equation 15

$$c_{\rho,1} = \frac{1}{2}\ln\left(-\frac{\rho_{eff}}{\rho_{eff} - 1}\right),$$

Equation 16

$$c_{1,1} = -\infty,$$

where $n = \sum_{j}^{R_c} \delta_{X_j,1}$, the number of particles within R_c of X_i.

Pair correlations alone are generally insufficient to accurately describe amorphous molecular systems. The propensity in such materials for particles to aggregate, through processes as diverse as chemical bonding and long-time annealing, generally necessitate the consideration of correlations between more than two particles at a time. As an example, consider in Figure 2 a sample of mutually attractive particles, initially set at random positions on a grid, and allowed to aggregate over time. At the time the system was sampled, the aggregates share a narrow size distribution centered on nine particles. How could we assign an unknown pixel, X_i, on this grid as filled or empty using contextual information? If we considered isotropic pair correlations within $L_c = 4$, we would find only that there are 15 particles near X_i. To distinguish whether X_i is within, directly adjacent to, or simply nearby an existing aggregate, is impossible with pair correlations alone. To distinguish these environments, one must incorporate many-body correlations which specific give information on the size, orientation and proximity of any nearby aggregates.

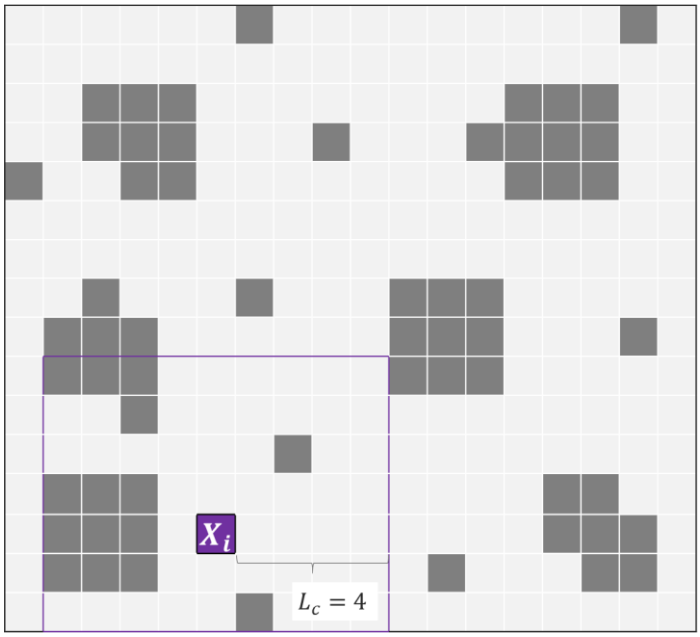

Figure 2: Sample of particles on a square grid in mid-aggregation, with gridpoint X_i highlighted for resampling.

To incorporate multi-particle correlations, we borrow inspiration from the statistical mechanics of dense fluids in the form of the cluster expansion, which corrects by orders Equation 9 with many-body interactions[13]. We develop a corresponding cluster expansion, collecting correlations order-by-order.

Equation 17

$$p\left(X_i = \theta | \{X_{j<i}\}\right) = N\left[c_{\rho,\theta} + \sum_{\Omega}^{\Omega_{max}} \sum_{\{x\} \in L_c} \sum_{\{\omega\}}^{N_c} c_{\{x,\omega\},\theta} \prod_{\mathrm{m}}^{\Omega} \delta_{X_{\{x\}_m},\{\omega\}_m}\right],$$

where Ω is the cluster size, $\{x\}$ the set of correlating gridpoints and $\{\omega\}$ the set of class identifiers for $\{x\}$. For example, $c_{\{x,\omega\},\theta}$, contains the influence of the set of Ω sequence elements $\{x\} \in L_c$, with classes $\{\omega\} \leq N_{class}$, on the probability of a particle of class θ being found on gridpoint i. Similarly to the case of the cluster expansion for dense fluids, evaluating this using a naive brute force approach is generally impractical for three reasons: 1) the number of terms explodes exponentially with correlation length, L_c, maximum cluster size, Ω_{max}, and number of output classes, N_c. 2) In general, the coefficients $c_{\{x,\omega\},\theta}$ cannot be computed directly from fundamental physics but must be fit using training examples. 3) Since the Kronecker delta function will only return nonzero when a specific cluster, $\{x, \omega\}$, appears, the coefficients in Equation 17 can only be fit using training examples which contain the exact many-body correlations assigned to each coefficient. Rather than learning generalizable heuristics or trends within the training distribution, Equation 17 can only accurately describe correlations identical or very similar to those it has already encountered. Therefore, fitting this equation accurately requires well-sampled configurations of every possible permutation of $\{x \in L_c, \omega \leq N_c\}$, up to order Ω.

Since naïve parameterization of this function is extremely difficult for nontrivial problems; our approach is to resort to deep learning models, known for their almost unreasonable effectiveness in fitting extremely complex, highly multidimensional functions with a tractable number of parameters[14]. In Section III we present the process of sampling by means of sequence generation using the MAP.

III. MAP Sequence Propagation and Properties

In this section we show that sequences generated with a MAP are physically sound simulations of materials. We approach this question by first making an analogy with the Markov Chain Monte Carlo (MCMC) simulation/sampling protocol. MCMC is considered a simulation tool because it generates ensembles of system's configurations sampled from a physical distribution. This is done by constructing aperiodic and irreducible Markov chains with the same equilibrium state as the physical distribution. These properties are guaranteed e.g., in the Metropolis algorithm by the sufficient but not necessary detailed balance condition,

Equation 18

$$p(x|x')p(x') = p(x'|x)p(x),$$

between states of the chain, x, x'. This relation determines the transition rates between states at equilibrium and requires all states of the Markov Chain to be connected. MAP sequences of the type described in Equation 3 also constitute an aperiodic, irreducible Markov chain, guaranteed to converge to a unique equilibrium distribution[15], though without strict detailed balance. We note that 'equilibrium' in this sense does not refer to the equilibrium state of the physical system being sampled, but rather the stationary state of the MAP Markov chain. Indeed, a hallmark of many amorphous or glassy systems is their nonequilibrium character, and it is therefore crucial that we can sample such out-of-equilibrium distributions. When Equation 17 is well approximated, by a neural network or other numerical model, the distribution of the MAP 'equilibrium' state will match that of the physical system being sampled. We liken this type of modelling to the parameterization of e.g., interatomic potentials by empirical functions, in that, when well-fit, both a MAP guided by a correlations model and molecular dynamics simulation under the influence of a force field will converge to a physical distribution.

We now prove that a MAP sequence constitutes an aperiodic, irreducible Markov chain beginning by rewriting Equation 3 as a probability distribution for the configuration of the next contextual field C_i conditional upon the current contextual field C_{i-1},

Equation 19

$$p\big(X_i|\{X_{j\in L_c}\}\big) = p(C_i|C_{i-1}),$$

where C_i and C_{i-1} are the contextual fields $\{X_{i-1\ldots L_c}\}, \{X_{i-2\ldots L_c-1}\}$, used to predict successive sequence elements (see Figure 3 for a visual explanation). We will begin for simplicity in 1D, where C_i and C_{i-1} share the same elements except for the newest element of C_i, X_i, and the oldest element of C_{i-1}, X_{i-L_c}.

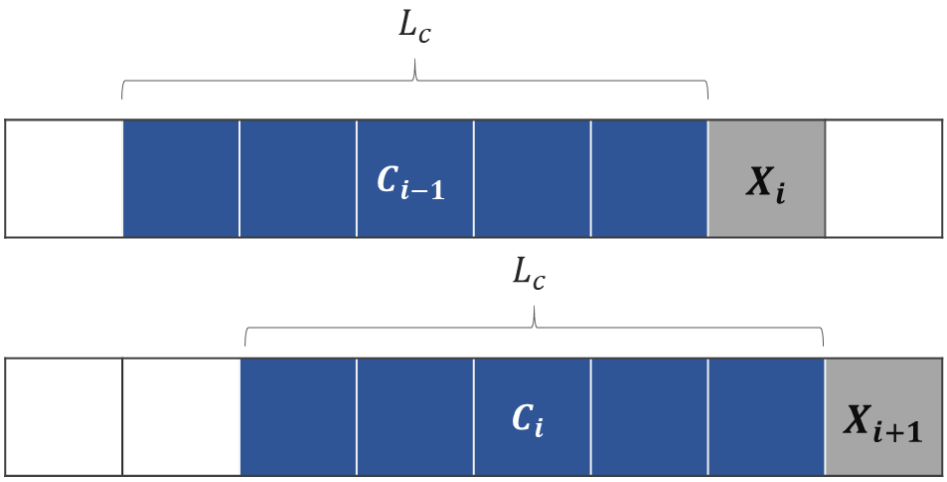

Figure 3: Recasting 1D sequence generation from predicting the occupation of the next pixel, to predicting next context state, given the current one, $C_{i-1} \rightarrow C_i$.

We may then consider the dynamics of a new configurational state space, defined as all $N_c^{L_c}$ permutations of the contextual field, C_i. Propagation in this space is mediated by transition probabilities $p(C_i|C_{i-1})$ computed via Equation 3, in practice modelled using a neural network.

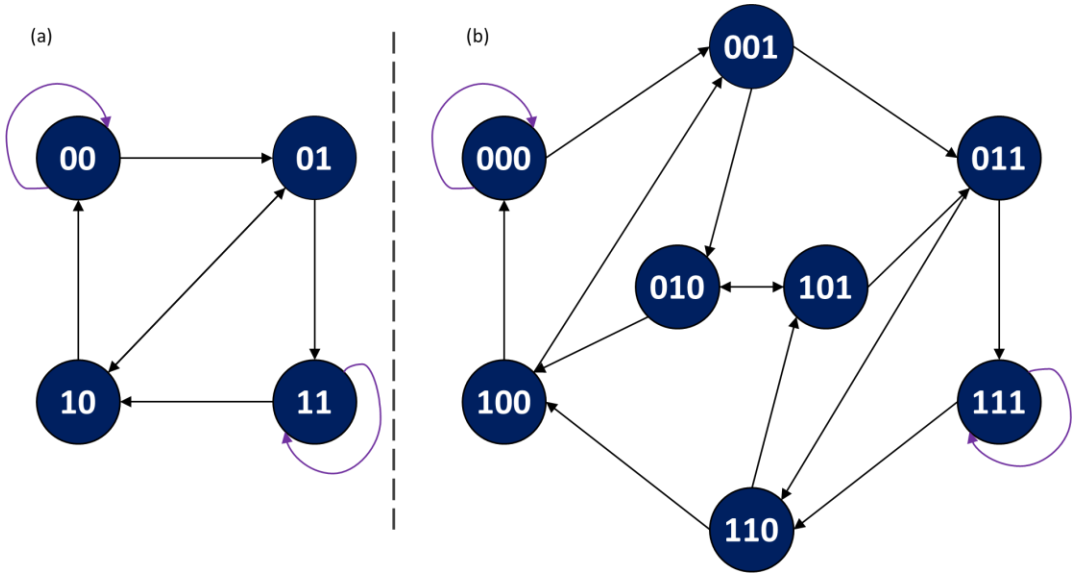

Figure 4: Example graph representations of the state space of 1D C_i's with $N_c = 2$, (a) $L_c = 2$ and (b) $L_c = 3$. The numbered nodes represent states, C_i, of the chain, with directed edges between states where $p(C_i|C_{i-1}) \neq 0$.

In **Figure** 4 we present minimal illustrative examples of the state spaces of MAP sequences in the form of directed graphs of the space of 1D context states, (C_i's). The graph connectivity is determined by our definitions for a MAP sequence. Each node is connected via directed edge to exactly N_c nodes, with exactly N_c nodes connected to themselves. This arises from the permutation structure of the sequence; at each step we compute the normalized probabilities for all N_c possible classes of the next element using Equation 3. From Equation 19 we can see that these probabilities

are exactly the transition probabilities between context states of the chain or equivalently nodes on the graph. Since we use a softmax function (

Equation 4) to normalize classwise probabilities, all the allowed transitions will have non-vanishing probabilities for any finite $\beta < \infty$. This means that the most distant states of the chain are connected within a maximum of L_c steps, since, as we exemplify in **Figure** 4 and demonstrate explicitly in **Figure** 5, the sequence is free to choose any of N_c options for the next element, X_i, with probability $p\left(X_i = \theta \middle| \{X_{i \in L_c}\}\right)$.

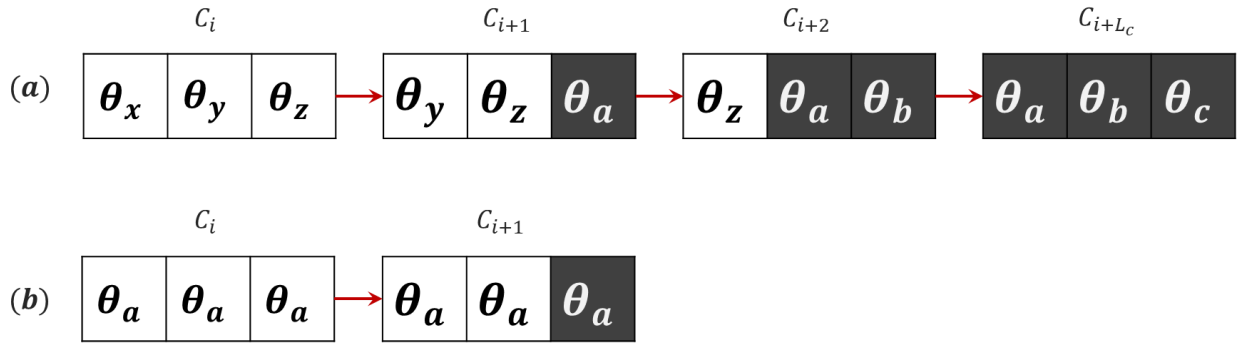

Figure 5: (a) Demonstrating the permutation structure of the chain. Starting from any arbitrary state with elements $\{\theta_x \theta_y \theta_z\}$, within a maximum of L_c steps the sequence can reach any other arbitrary state of the chain $\{\theta_a \theta_b \theta_c\}$. Panel (b) shows that a state with elements of all one class is self-connected. Since each state may always choose from all N_c possible classes for its next element, a state which is composed only of elements of a single class always has finite probability for its next element to also be of the same class. This visually exemplified in the minimal examples in **Figure** 4.

We also see in **Figure** 5 how N_c states of the chain which have only elements from one class are self-connected. For example, given $N_c = 3$, $C_i = \{1,1,1\}, \{2,2,2\}, \{3,3,3\}$ are the three self-connected states. The fact that all states of the chain are connected means that it is irreducible, every state is in the same communicating class. Further, the chain can be easily seen to be aperiodic due to the presence of N_{class} states which are self-connected. The irreducibility of the full chain means that the chain itself is also aperiodic[15], and since the transition probabilities do not change during propagation, the full MAP sequence is guaranteed to converge to a unique limiting or equilibrium distribution.

While desirable, guaranteeing convergence in only one dimension is of limited value when the physical systems we are interested in describing are generally at least 2-dimensional. To demonstrate that this property holds in higher dimensions, this we will recast the propagation in 2D or 3D to an effective 1D sequence. In $N > 1$ dimensions, one must first select a pattern by which to propagate the sequence. We will follow the ubiquitous Raster pattern and accompanying masking, though others have been explored in this context[16]. From **Figure 6**, it appears naively that a MAP sequence may not be as well behaved in 2D, due to the presence of rows above X_i within C_i. We can see that determining the transition probability from C_i to C_{i-1} involves predicting the values not just of X_i, but of rows above, labelled B_i.

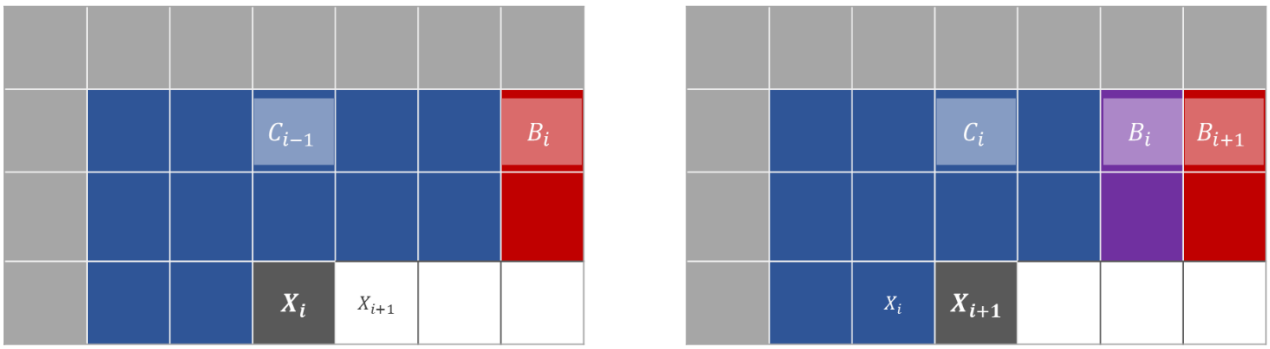

Figure 6: A Markov state model in 2D is complicated by the introduction of boundary terms B_i as the sequence propagates from left to right, which are not accounted for in the conditional probability distribution for X_i.

The situation is simplified by recasting a higher dimensional sequence as a 1D sequence of larger objects. As we show in **Figure 7**, this can be done in two steps. First, we switch the propagation pattern from elementwise from left to right, to row wise from the top down. This new sequence predicts the value of all the elements of a row i simultaneously, using previous rows of width W within L_c $\{X_{i\ldots i-L_c}^{1\ldots W}\}$ as context. This is now effectively a 1D sequence of 1D objects $\{X_i^{1\ldots W}\}$, which we can make more explicit by re-expressing the permutations of row elements as single values in a larger set $Y \leq N_c^W$.

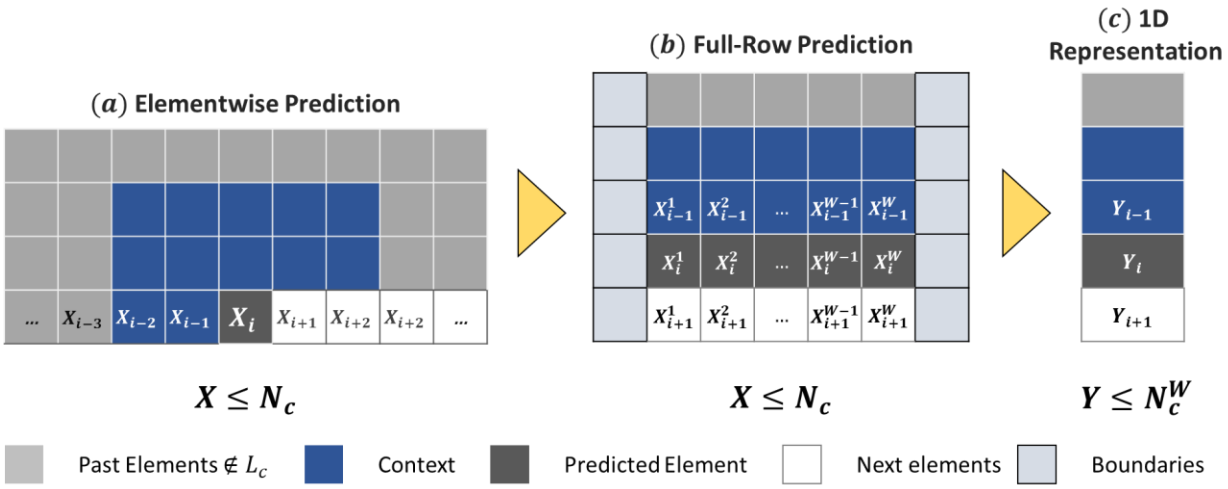

Figure 7: Recasting a 2D MAP propagation from elementwise prediction in 2D (a) to full-row prediction in 2D (b), to elementwise prediction in 1D (c), with the new elements $Y_i \leq N_c^W$ labelling all possible permutations of $\{X_i^{1\ldots W}\}$.

This new 1D sequence exists in an irreducible, aperiodic state space in the same way as the original 1D, except that each state $\{Y_{1\ldots L_c}\}$ is connected to N_c^W other states, and N_c^W states are now self-connected. As we outline in **Figure 8**, these states of the contextual state space, R_i, have a size of $L_c \times W$ in the 2D basis of X_i's, or L_c in the 1D Y_i basis.

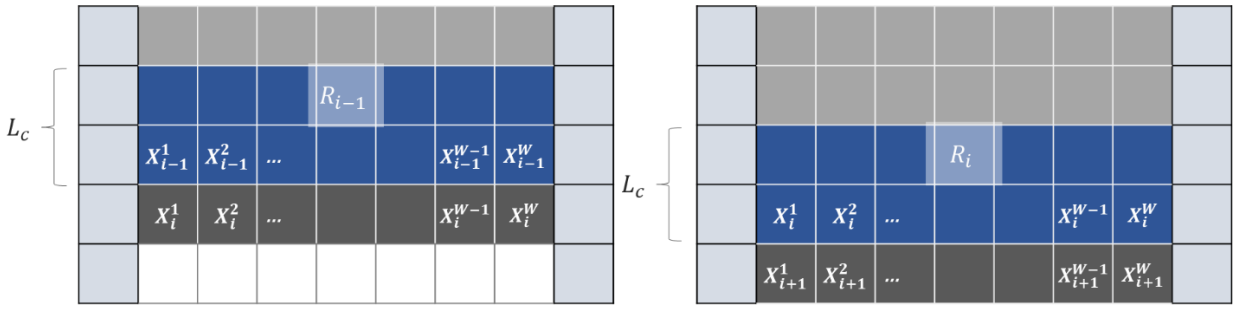

Figure 8: Illustration of row-wise propagation, using full-row-width contextual fields, R_{i-1}, to predict the next row of elements, or equivalently, to predict the next contextual field, R_i.

We can quantify the transition probabilities between states R_{i-1}, R_i by decomposing the joint probability for the full row $\{X_i^{1\cdots W}\}$ into elementwise probabilities using the chain rule.

Equation 20

$$p(R_i|R_{i-1}) = p(X_i^1 \cap X_i^2 \cap \ldots \cap X_i^W | R_{i-1}) = \prod_j^W p\left(X_i^j | \{X_{k\in L_c}\}\right),$$

In Equation 20, we see that this can be computed according to the familiar elementwise conditional probability from Equation 3. In practice, this is intractable to compute due to the extremely large state space of R_i's, but it can be easily sampled by repeatedly propagating the sequence. This is not necessary in practice, since the conditional probabilities in Equation 20 are constants, the transition probabilities between any R_i and R_{i-1} are also constant, and the full chain is Markovian.

The above arguments for reducing a 2D sequence to 1D by recasting the sequence from predicting elements to full rows function identically in higher dimensions. For example, in 3D one would propagate a 1D sequence in the state space of $L_c \times H \times W$ slabs, with elements $Y_i \leq N_c^{H\times W}$. As in 1D, 2 or 3D chains can be seen in this way to be Markovian, irreducible, and aperiodic. We can therefore see that any N-dimensional MAP should converge to a unique equilibrium distribution, conditional only on the sample boundaries and initial condition, which as we discuss in Section IVB, is typically taken as blank.

It is not ab-initio obvious what the mixing or equilibration time for such a sequence would be, but we can set lower bounds based on connectivity, and we provide a numerical example in Section IVB. All possible states in a MAP sequence are directionally connected within a maximum number of jumps L_c. If we presume that the sequence must explore a significant portion of the state space before equilibrating, it is reasonable to set a lower bound on the mixing time on the order of L_c. One can rigorously derive counting and diameter bounds following this reasoning[15]. When using a finite-temperature softmax normalization as we have done, a MAP contains no connectivity bottlenecks, and every state is connected to N_{class} states. This does not mean that mixing will always be fast or barrierless since, while not bottlenecked, paths between e.g., initial conditions and high-probability regions may be gated by low-probability regions, acting just the same as high free energy barriers in a molecular dynamics simulation.

IV. Numerical Experiments

In this section, we fit MAP sequence generators to a series of physically derived two-dimensional systems, starting with dilute fluids, and finishing with a complex and highly correlated nanoparticle aggregate. The purpose of these experiments is not to simply show that a PixelCNN-based MAP can accurately learn and re-express the static correlations in each system. Rather, we make use of these examples specifically to illustrate the properties derived in the prior sections.

A. Direct Correlation Fitting

We begin by fitting the all-order correlation expansion Equation 17 via a brute force numerical approach. To keep the problem tractable, we consider very simple physical distributions, as shown in Figure 9, and truncate the maximum range and order of the considered correlations for each system, as shown in

Table **1**. Further, since there are only two possible pixel classes (occupied / unoccupied, 1 / 0), we may safely ignore cross-correlations between them, and examine the statistics of only one class. Since the model systems are dilute, we base our analysis on occupied pixels to minimize the required maximum correlation order. These simplifications yield the following correlation expansion to be fit,

Equation 21

$$p\left(X_i = 1 | \{X_{j \in L_c}\}\right) = N\left[c_\rho + \sum_{\Omega}^{\Omega_{\max}} \sum_{\{x\} \in L_c} c_{\{x\}} \prod_j X_{x_j}\right].$$

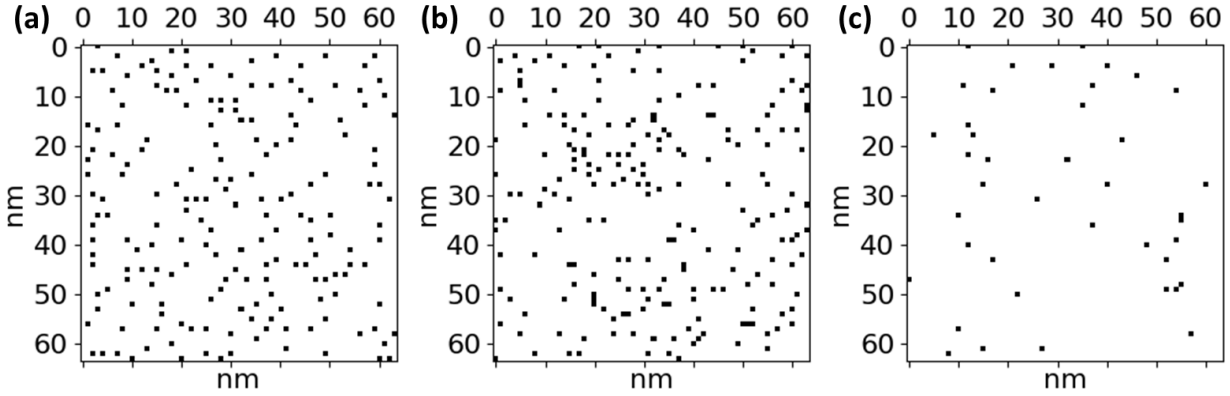

Figure 9: Examples from the training sets for the 3 dilute fluids, with panels (a-c) corresponding to systems (1-3), cold with repulsive interactions, hot with repulsive interactions and hot with attractive interactions, respectively, with all interactions ranges set to 1.

Table 1: Properties of physical systems numerically fit to the truncated correlation expansion. The correlation length and order are emergent properties of the simulation parameters. Detailed simulation parameters given in Appendix A.

System	Relative Temperature	Inter-Particle Interactions	Grid Occupation Density	Correlation Length L_c	Maximum Correlation Order	# of Fitting Coefficients
1	Cold	Repulsive	5%	1	2	5
2	Hot	Repulsive	5%	2	3	79
3	Hot	Attractive	1%	3	3	301

The simplicity of these distributions highlights the extreme difficulty of the all-order problem. For context, fitting even the simplified Equation 21 scales exponentially poorly with the range and complexity of considered correlations, to say nothing of the required training data. For example, with correlation order and length both equal to 5, a naïve fitting would require 5.9 million fitting coefficients. For $\Omega_{max}, L_c = 6$, there are 439 million, and so on.

Numerical fitting was accomplished via stochastic gradient descent in a custom PyTorch model, using PixelCNN-style masking and autoregression to hide information about 'future' pixels during training, and to generate new samples, respectively[9].

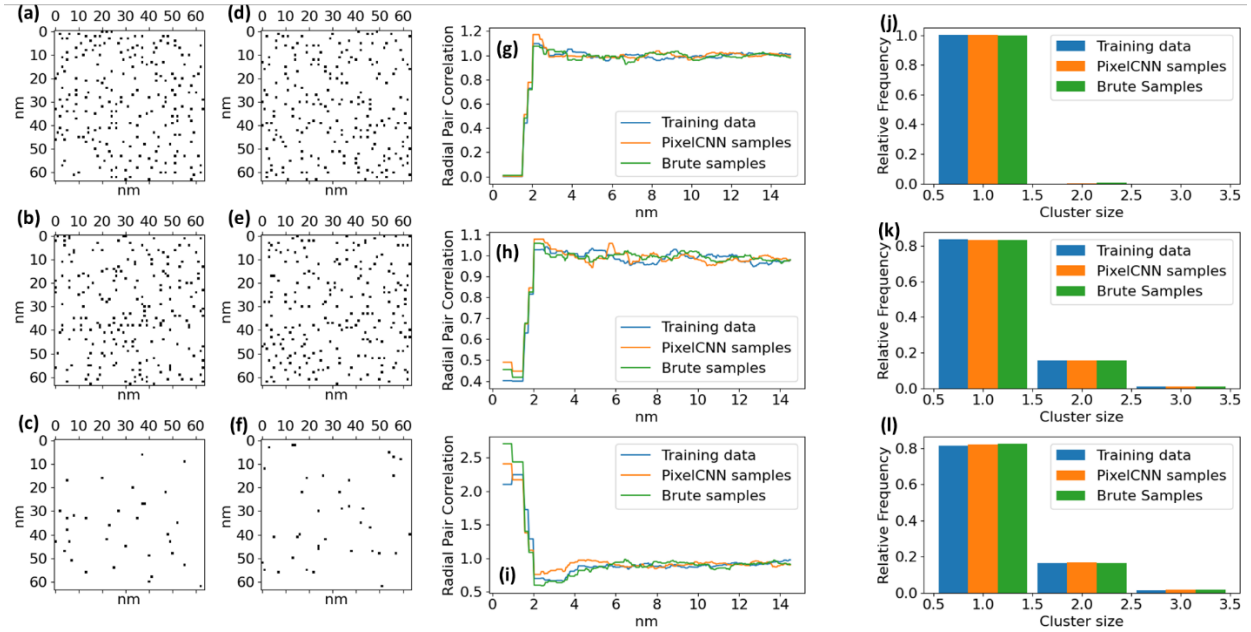

Figure 10: Simulation results comparing training data, brute force correlation fitting and PixelCNN. Panels (a-c) and (d-f) show examples generated via MAP, using Equation 21 and PixelCNN, respectively, for the 3 different physical systems. Panels (g-i) and (j-l) show the pair-correlation functions and particle neighborhood distributions respectively for particles the training data, and samples generated by the two fitting approaches. Details of the PixelCNN models used are given in Appendix A.

We can see in Figure 10 that our simplifying assumptions Equation 21were well-justified, as the statistics of distributions generated by this truncated all-order model agree very well with the target distributions, and further with the higher-capacity convolutional neural network. While this is a good result for the brute force approach, the poor scaling noted above highlights the need for an alternative. Consider the system sampled in **Figure 2**; it is possible that a clever researcher could thoughtfully and painstakingly prune an all-order correlation expansion and possibly fit a tractable correlation function. Alternatively, they could train an off-the-shelf convolutional neural network to learn and re-express the salient features without human intervention. In any case, interesting physical systems such as e.g., organic heterojunctions and graphene derivatives go far beyond the ability of even specialists to model by hand, and so the case for a neural network approach is easy to make.

B. Sequence Convergence with PixelCNN

We now move to a significantly more complex physical system to illustrate the concepts of MAP convergence developed in Section III. Convergence analysis was omitted from Section IVA because, with such short correlation lengths and uncomplicated configurational motifs, they converge essentially immediately upon departing the sample boundary.

In our prior experiments using a PixelCNN MAP to generate the structures of amorphous nanoparticle and 2D carbon aggregates, we have typically found that the MAP converges

(according to the human eye) within 1-2 multiples of the convolutional receptive field, L_c, from the sample boundary. This may sensitively depend on the choice of what exists outside the image boundary; we typically assign this boundary region a unique class which denotes 'outside the sample'. One may also seed the boundary with samples from the training set, noise, or a related known structure, such as a known crystalline structure, to speed up convergence or direct generation to a particular phase or structural motif. At present we believe an empty boundary results in the most general, unbiased samples, and since the MAP converges quickly, we do not appear to pay a significant cost by having the sequence bootstrap itself from an empty initial condition in this way.

To go beyond our visual intuition on this point, we demonstrate in **Figure 11** the convergence properties of a PixelCNN model on a nanoparticle aggregate generated via Markov chain Monte Carlo (MCMC) simulations of solvent drying dynamics[17]. This system combines strong short-range pair correlations with many-body correlations on the length scale of the solvent voids. In this system, the particles repel one another, but are strongly attracted to the solvent (see MCMC simulation and PixelCNN details in Appendix A). Particles are pushed together by drying of the solvent, resulting in the observed voids, with the remaining solvent (omitted) serving as a 'glue', keeping them close but not directly adjacent to one another.

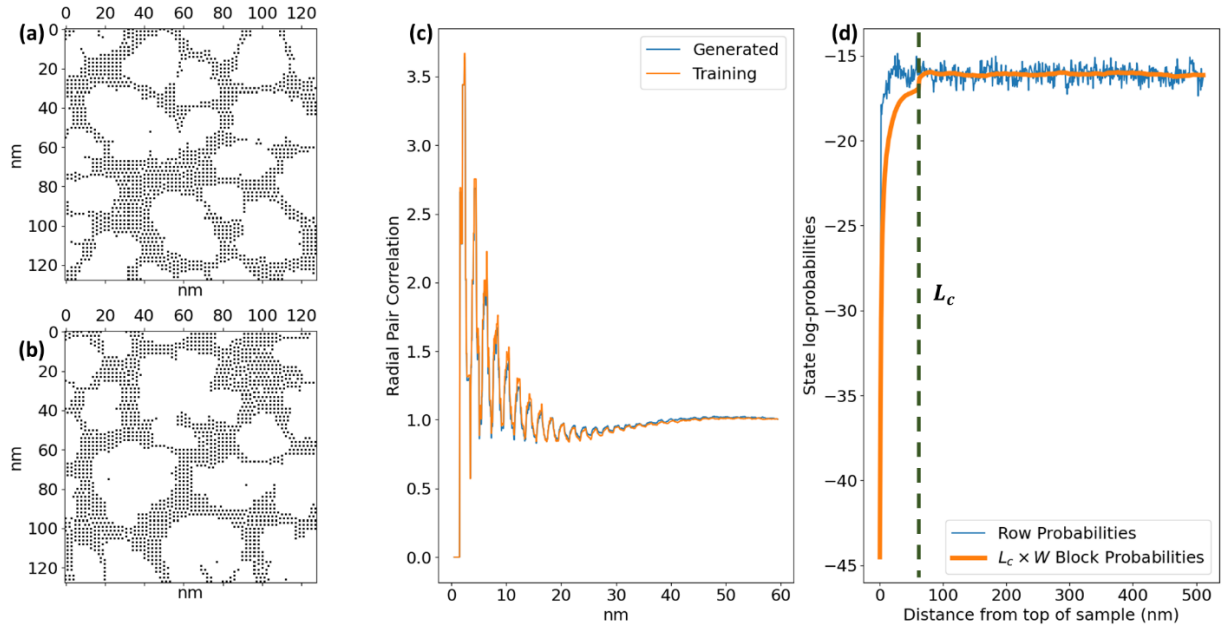

Figure 11: Panels (a) and (b) are example structures from the training data and generated by a PixelCNN model, respectively. We omit solvent to show particle positions only, both here and in training. Panel (c) compares the pair correlation functions between samples from the training data and generated by the PixelCNN MAP. Panel (d) shows the log of the product of pixel probabilities for each row, starting from the top of the sample, and for a context block, R_i, of size $L_c \times W$, averaged over 100 generated samples. The probabilities have been normalized for easier visibility on the same axis.

Figure 11 is very interesting, as it shows that the row-wise pixel probabilities saturate the maximum bound very quickly and, on average, vary relatively little once converged. This suggests that the sequence can reach high probability configurations *before* exploring a significant portion of the configuration space, here determined by the correlation length, $L_c = 61$. Though the corresponding probabilities for whole contextual blocks, R_i appear to take longer to converge, we see by the nearly vertical step around the correlation length, that this apparent lag in convergence is in very large part determined by the first few, very low probability rows. Once these early rows leave the context window, the sequence saturates its maximum probability, apparently converging to its equilibrium distribution. Since we cannot map the occupations of all possible states of the Markov chain ($2^{61\times128}$) this relatively straightforward analysis cannot tell us unambiguously that the sequence fully equilibrates by the 70^{th} row. However, the rapid saturation of the average row probability is compelling evidence that the sequence at least reaches high-probability configurations very quickly.

It should be highlighted that this effective convergence is quite sensitive to the choice of what sits outside the system boundary, i.e., the context used to predict the earliest rows. If we had seeded the boundary with a sample from the training set, for example, the probabilities of the first few rows would likely have been significantly higher, as even the first predictions made by the network would be in familiar morphological territory. Also, while the sequence appears to converge easily on this relatively complex system, that does not preclude the existence of systems which may converge more slowly, perhaps including for example systems with multiple metastable phases.

While we have found a Gated PixelCNN[5] architecture to perform very well with minimal tuning, there are classes of structures which have proven more challenging to model. Particularly, our models perform well but not perfectly on systems which combine sparsely populated samples with very long-range correlations. This includes systems such as monolayer amorphous carbon, modelled in our previous work[1], and nanoparticle aggregates similar to those modelled in **Figure 11**, though with pixel occupation density reduced to 1.5%. We consider these challenges to be technical in origin, potentially correctable using upgraded architectures or conditioning variables[4,18–21], and not a major present concern.

V. Conclusion

We explored the theoretical underpinning which justifies the use of MAP models such as PixelCNN for simulation of amorphous chemical systems. Such a simulation tool could be used for a variety of sampling tasks of chemical relevance, most obviously cheaply simulating arbitrarily large samples of amorphous materials.

We explained that the role of deep autoregressive models such as PixelRNN or PixelCNN is to efficiently approximate the exact all-order correlation cluster expansion on a physical system. This is extremely important since direct evaluation of the all-order expression via brute force is essentially impossible beyond trivial models. Deep learning algorithms save the user the effort of identifying and modelling by-hand the correlations which determine the structures of amorphous

systems, instead allowing the model to learn them efficiently and automatically with minimal intervention. We also showed that, when the autoregressive sequence is defined on a discrete grid with discrete classes of outputs, and with probabilities normalized via softmax function, such a sequence constitutes a converging Markov chain with a unique equilibrium state and well-defined lower bounds on its mixing time. When trained properly, propagation of an equilibrated sequence constitutes sampling from the target distribution.

In the space of generative models for structure prediction, we may now add to the list of advantages for MAP approaches: beyond flexibility and ease of use, we have here elucidated the process by which such an approach parses structural data to model the structural correlations of amorphous materials. Specifically, given the discreteness assumptions in Section IIB, we can see that MAP models in-practice approximate the exact all-order result given in Equation 17. MAP inputs, outputs and basic function are all human interpretable. Further, it is remarkable that such a sequence is guaranteed to converge to a unique equilibrium distribution in any dimension, conditional only on the initial / boundary condition. One has only to appropriately train the model such that it appropriately apportions probability density according to the target physical distribution.

Future work on autoregressive structural sampling may proceed along three paths. First, algorithmic improvements such as equivariant convolutions[21] may continue to boost the speed and accuracy of MAP generators on atomistic systems, wherein sparsity and long-range correlation have presented a challenge for Gated PixelCNN. Second, extensions to more involved systems such as 3D aggregates of molecules present a rich environment for structural modelling and exploration of functional materials. Finally, and most relevant for this study are comparisons between generative models which can be used for structure prediction. Within autoregression, one may consider alternatives to PixelCNN, such as PixelRNN[7] or PixelSNAIL[16], and weigh their advantages and disadvantages. Looking further to normalizing flows, autoencoders and generative adversarial networks[2,3,22] which have been developed for similar tasks, one may consider the mathematical similarities and differences between the activity of these models. Is there a connection between how a MAP models many-body structural correlations to, for example, the way that a generative adversarial network or normalizing flow reshape an input distribution to a physical one? In some sense the answer must be yes, but the details of how and why may inform future developments in this space.

Acknowledgements:

Funding from NSERC Discovery grant RGPIN-2019-04734 is gratefully acknowledged. Computations were made on the supercomputer Béluga, managed by Calcul Québec (https://www.calculquebec.ca/) and Compute Canada (www.computecanada.ca). The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI).

Data Availability:

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Appendix A: Computational Details

1. **Nanoparticle systems: parameters and simulation details**

The nanoparticle systems modelled in Section IV were generated by a basic Markov chain Monte Carlo algorithm incorporating optional solvent and drying dynamics on a 2D grid[17]. The Markov Chain advances via Metropolis algorithm under the influence of the model Hamiltonian,

Equation 22

$$\mathrm{H} = \epsilon_{nn} \sum_{i,j} \mathrm{n_i n_j} + \epsilon_{\mathrm{ll}} \sum_{\mathrm{ij}} \mathrm{l_i l_j} + \epsilon_{\mathrm{nl}} \sum_{\mathrm{ij}} \mathrm{n_i l_j} + \mu \sum_{\mathrm{i}} \mathrm{l_i},$$

with particles denoted by n and solvent by l, their respective interactions by $\epsilon_{nn}, \epsilon_{ll}$ and ϵ_{nl}, and the chemical potential of the solvent by μ. In our simulations all interactions were limited ty a range of 1.

The simpler systems modelled in Section IVA considered only nearest-neighbor particle-particle interactions and equilibrated extremely quickly.

Table 2: Parameters for MCMC simulations of dilute fluids.

System	1	2	3
$\boldsymbol{\epsilon_{nn}}$	1	1	-1
$\boldsymbol{T}$	~ 0	1	1
$\boldsymbol{\rho}$	5%	5%	1%
Age	100 steps	100 steps	100 steps

The more complex system modelled in Section IVB incorporated solvent interactions and drying dynamics, as well as annealing, though with interactions still limited to nearest neighbors.

Table 3: Parameters for drying MCMC simulations of mutually repulsive nanoparticles in an attractive solvent.

Param.	Value
$\boldsymbol{\epsilon_{ll}}$	-2

ϵ_{nl}	-20
ϵ_{nn}	-4
μ	11
T	1
ρ	10%
Age	2000 steps

2. PixelCNN parameters and implementation details

The PixelCNN model used in this study is nearly identical to that used in our previous work[1]. It consists of horizontal and vertical convolutional stacks with gated activations, two fully-connected layers and a softmax normalization at the end, all set-up identically to the original Gated PixelCNN[5]. The only architectural difference is the addition of skip connections after each horizontal stack activation directly to the output layers, inspired by WaveNet[23].

For the modelling in Section IVA, for each system a small PixelCNN model with 48 convolutional filters per layer and a number of convolutional layers equal to the correlation length, L_c, was trained on 10000 64×64 samples of each of the distributions. The more complex aggregate in Section IVB was modelled using a PixelCNN with 60 layers, 20 filters per layer and trained on 18,000 256×256 samples.

References

[1] M. Kilgour, N. Gastellu, D.Y.T. Hui, Y. Bengio, L. Simine, Generating Multiscale Amorphous Molecular Structures Using Deep Learning: A Study in 2D, J. Phys. Chem. Lett. (2020) 8532–8537. https://doi.org/10.1021/acs.jpclett.0c02535.

[2] F. Noé, S. Olsson, J. Köhler, H. Wu, Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning, Science (80-.). 365 (2019). https://doi.org/10.1126/science.aaw1147.

[3] M. Comin, L.J. Lewis, Deep-learning approach to the structure of amorphous silicon, Phys. Rev. B. 100 (2019) 94107. https://doi.org/10.1103/PhysRevB.100.094107.

[4] C. Casert, K. Mills, T. Vieijra, J. Ryckebusch, I. Tamblyn, Optical lattice experiments at

unobserved conditions and scales through generative adversarial deep learning, (2020) 1–11. http://arxiv.org/abs/2002.07055.

[5] A. Van Den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, K. Kavukcuoglu, Conditional image generation with PixelCNN decoders, in: Adv. Neural Inf. Process. Syst., 2016: pp. 4797–4805. http://arxiv.org/abs/1606.05328.

[6] P. Ramachandran, T. Le Paine, P. Khorrami, M. Babaeizadeh, S. Chang, Y. Zhang, M. Hasegawa-Johnson, R. Campbell, T. Huang, Fast generation for convolutional autoregressive models, 5th Int. Conf. Learn. Represent. ICLR 2017 - Work. Track Proc. (2019) 1–5.

[7] A. Van Den Oord, N. Kalchbrenner, K. Kavukcuoglu, Pixel recurrent neural networks, 33rd Int. Conf. Mach. Learn. ICML 2016. 4 (2016) 2611–2620.

[8] X. Chen, N. Mishra, M. Rohaninejad, P. Abbeel, PixelsNail: An improved autoregressive generative model, in: 6th Int. Conf. Learn. Represent. ICLR 2018 - Work. Track Proc., 2018.

[9] A. van den Oord, N. Kalchbrenner, K. Kavukcuoglu, Pixel Recurrent Neural Networks, (2016). http://arxiv.org/abs/1601.06759.

[10] T. Salimans, A. Karpathy, X. Chen, D.P. Kingma, PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications, 5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc. (2017) 1–10. http://arxiv.org/abs/1701.05517.

[11] M. Comin, L.J. Lewis, Deep-learning approach to the structure of amorphous silicon, Phys. Rev. B. 100 (2019) 94107. https://doi.org/10.1103/PhysRevB.100.094107.

[12] K. Mills, C. Casert, I. Tamblyn, Adversarial generation of mesoscale surfaces from small scale chemical motifs, Under Rev. (2020).

[13] F. Schreiber, F. Zanini, F. Roosen-runge, Virial Expansion – A Brief Introduction, (2011) 1–16.

[14] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature. 521 (2015) 436–444. https://doi.org/10.1038/nature14539.

[15] D.A. Bini, G. Latouche, B. Meini, D.A. Bini, G. Latouche, B. Meini, Introduction To Markov Chains, Numer. Methods Struct. Markov Chain. (2007) 3–22. https://doi.org/10.1093/acprof:oso/9780198527688.003.0001.

[16] X. Chen, N. Mishra, M. Rohaninejad, P. Abbeel, PixelsNail: An improved autoregressive generative model, 6th Int. Conf. Learn. Represent. ICLR 2018 - Work. Track Proc. (2018).

[17] E. Rabani, D.R. Reichman, P.L. Geissler, L.E. Brus, Drying-mediated self-assembly of nanoparticles, Nature. 213901 (2003) 271–274. https://doi.org/10.1038/nature02087.

[18] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks, Adv. Neural Inf. Process. Syst. 2015-Janua (2015) 2017–2025.

[19] W. Luo, Y. Li, R. Urtasun, R. Zemel, Understanding the effective receptive field in deep convolutional neural networks, in: Adv. Neural Inf. Process. Syst., 2016: pp. 4905–4913.

[20] C. Ye, M. Evanusa, H. He, A. Mitrokhin, T. Goldstein, J.A. Yorke, C. Fermüller, Y. Aloimonos, Network Deconvolution, (2019). http://arxiv.org/abs/1905.11926.

[21] B.K. Miller, M. Geiger, T.E. Smidt, F. Noé, Relevance of Rotationally Equivariant

Convolutions for Predicting Molecular Properties, (2020) 1–12. http://arxiv.org/abs/2008.08461.

[22] K. Mills, C. Casert, I. Tamblyn, Adversarial generation of mesoscale surfaces from small scale chemical motifs, J. Phys. Chem. C. (2020) 8–12. https://doi.org/10.1021/acs.jpcc.0c06673.

[23] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, WaveNet: A Generative Model for Raw Audio, (2016) 1–15. http://arxiv.org/abs/1609.03499.