Mapping diverse structures of liquid water and ice using variational autoencoders: A vector quantization approach to discover structural motifs in model latent spaces

Yifei Yue [1 2 3] **Lechun Xing** [2 4] **Saad A. Mohammed** [3] **Jianwen Jiang** [3 1] **N. Duane Loh** [1 2 4]

[1]*Integrative Sciences and Engineering Programme, National University of Singapore* [2]*Center for Bio-Imaging Sciences, National University of Singapore* [3]*Dept. of Chemical and Biomolecular Engineering, National University of Singapore* [4]*Dept. of Physics, National University of Singapore*. Correspondence to: N. Duane Loh duaneloh@nus.edu.sg.

1. Introduction

Ice formation from liquid water is a ubiquitous phenomenon in nature. Yet, a detailed mechanistic understanding of how ice nucleates and grows remains elusive. Particularly, it remains unclear how short-lived, metastable local structures (or structural motifs) form over the process of ice nucleation and growth. This motivates experimental[1] and computational[2] efforts to characterize such structural motifs in liquid and ice, which will be of great significance in understanding pertinent processes in natural sciences (e.g., precipitation models[1]).

In this work, we aim to characterize structural motifs in liquid water and ice from molecular dynamic (MD) simulations by applying vector quantization to interpret the latent space of a variational encoder (VAE). First, MD simulations were performed using the ML-BOP model[2] to simulate homogeneous ice nucleation from supercooled water at 210 K. Thereafter, structural descriptors of local atomic environments (e.g., rotationally invariant SOAP features[3]) were extracted from the MD trajectories. Next, we trained a VAE model to learn the latent embedding of structural descriptors, and applied vector quantization (VQ) to analyze the latent space. Analysis of the latent embedding reveals the emergence of locally favored structural motifs in supercooled water prior to nuclei formation. Furthermore, we observe the gradual formation of various ice polymorphs, such as stacking disorder ice (I_{sd}), that persists at different stages of nucleation and ice crystallization. Besides providing a comprehensive mapping of the structural motifs in ice polymorphs and liquid water, our VAE approach provides a flexible framework for exploring structural polymorphs in other systems of interest.

2. Results and Discussion

2.1 VAE model and vector quantization approach

Fig. 1 summarizes the workflow of our study. To characterize structural motifs, we first performed MD simulations (using the ML-BOP potential[2]) to investigate homogeneous ice nucleation of supercooled water (**Fig. 1a**). As shown in **Fig. 2**, we can identify various liquid and ice structures (namely cubic I_c and hexagonal I_h ice[4]) across different stages of ice nucleation and growth. Subsequently, frames extracted from the MD trajectory, and structural descriptors that represent the local atomic environments were computed (**Fig. 1a**). The structural descriptors consist of rotationally invariant Steinhardt[5] and SOAP descriptors[3], which constitute the input to the vAE model (**Fig. 1b**). Next, we trained VAE model to learn the latent embedding Z of the structural descriptors (in **Fig. 1b**). The UMAP algorithm[6] were applied to visualize the latent space (**Fig. 3**). Finally, we applied vector quantization (via HDBSCAN[7]) to detect clusters representing similar local atomic environments in the UMAP projection (**Fig. 3**).

With our approach, we detect 18 clusters corresponding to structurally distinct local atomic environments from the trajectory (in **Fig. 3a**). These motifs exhibit various energy signatures (**Fig. 3b**), and persist at different stages of ice nucleation and crystallization (**Fig. 3c**). From the local atomic environments corresponding to the cluster medoids (inset in **Fig. 3a**), these structural motifs exhibit varying compositions of liquid, interfacial, and I_{sd} structures in the neighborhood of the central atoms. These observations affirm that the latent embedding of the VAE can distinguish ice-like from liquid-like environments.

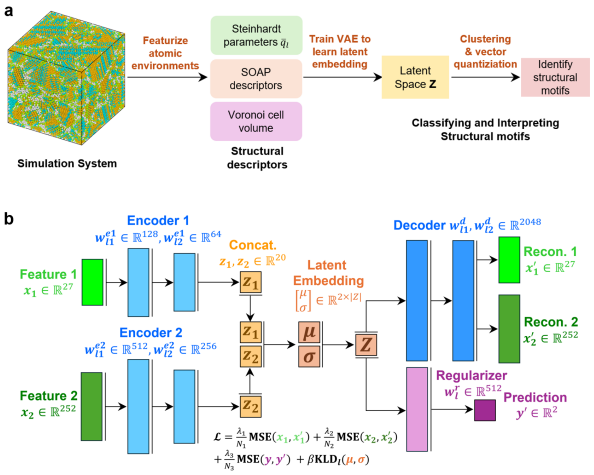

Fig. 1: **Workflow and VAE model.** (a) Schematic illustration of workflow employed, including the featurization of atomic environments to generate structural descriptors, which were used to train the VAE to learn a compressed latent representation Z. (b) Block diagram illustrating VAE architecture. The dimensionality of the input features and latent space of each component and loss function are displayed.

2.2 Clustering and Analysis of structural motifs

We now turn to analyze the structure and physical properties of the distinct clusters corresponding to distinct local atomic environments (which we call motifs) identified in **Fig. 3**. The local environ-

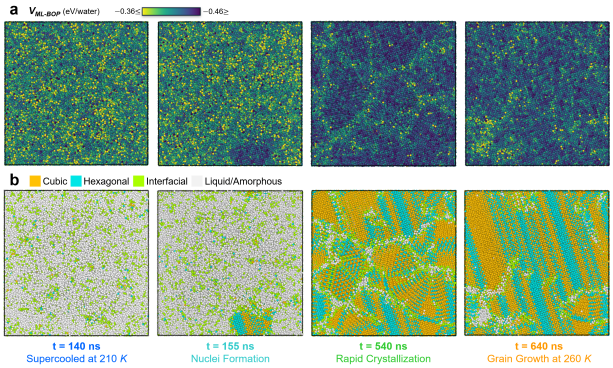

Fig. 2: **Stages of nucleation and crystal growth from nucleation trajectory**. (a) Atomic potential energy ($V_{ML,BOP}$) within the range of -0.36 eV (-34.7 kJ mol^{-1}) to -0.46 eV (-44.4 KJ mol^{-1}), and (e) ice structure types classified by CHILL+[4] with the color code: cubic (orange), hexagonal (blue), interfacial (green), liquid (white). Hydrates and interfacial hydrate structure types identified are negligible.

ments of predominantly ice- (motifs *1-5*) and liquid-like motifs (*8*,*11-13*) can be distinguished using CHILL+ algorithm[4] (**Fig. 3a**). Indeed, the map of atomic potentials energies over the UMAP projection ($V_{ML,BOP}$; **Fig. 3b**) illustrates regions of higher $V_{ML,BOP}$ exhibited by liquid-like motifs (see **Table 1**). The intermediate region in the UMAP projection (dashed rectangle in **Fig. 3a**) highlights the presence of transient motifs, whose local environments contain disordered ice (or stacking fault ice I_{sd}) formed at early stages of nuclei formation (e.g., from *16* → *10*; **Fig. 3d**). Based on this, we classify three distinct categories of motifs as summarized in **Table 1**: (1) disordered liquid motifs (*8*, *11-13*); (2) transient motifs (*6,7,9,10*, *14-18*) that temporarily form during nuclei growth; and (3) ice motifs (*1-5*) that comprises different compositions of I_h, I_c and I_{sd} ices.

The structural and physical properties exhibited by these various motifs can be characterized. As shown in **Table 1**, liquid motifs (e.g., *12,13*) exhibits a higher local atomic density (ρ_{voro} of 0.97~0.99 g/cm^3) compared to transient motifs and ice motifs (V_{voro} of 0.93 g/cm^3). Indeed, this affirms that liquid motifs exhibit densities close to liquid water (0.997 g/cm^3), whereas ice motifs exhibit densities comparable to ice (0.917 g/cm^3) under ambient conditions. Following the clusters of nuclei formation in UMAP projection, we observe that the median ρ_{voro} of motifs decreases in the order of *13* → *11* → *16* → *17* that corresponds to the gradual transition from liquid-like to transient motifs (**Fig. 3a**). This aligns with the trends of increasing tetrahedral order (decrease in A) at the onset of nuclei formation.[2] These trends also point to the development of ordered tetrahedral regions in supercooled water that resemble LDA ice[8], which constitute locally favorable structures that precede nuclei formation[9]. As such, we attribute motif *16* - which exhibits a moderate ρ_{voro} of 0.94 g/cm^3 and a low A of 0.265- to the presence of structurally ordered LD environments in supercooled water (cyan cluster in **Fig. 3**)[8, 10]. Conversely, motifs *8* and *12*, which exhibit lower tetrahedral order (A of 0.6~0.67) and higher local density (ρ_{voro} of 0.99 g/cm^3), can be attributed to bulk water prior to supercooling. Transient motifs *11* and *13* possess intermediate values of ρ_{voro} of 0.96 g/cm^3 and A of 0.265, which can be attributed to supercooled water.

In summary, our vector quantization approach enables explainable interpretation of structural motifs in the latent space of autoencoders. We expect that our model and workflow can be applied to elucidate structural motifs in other materials of interest.

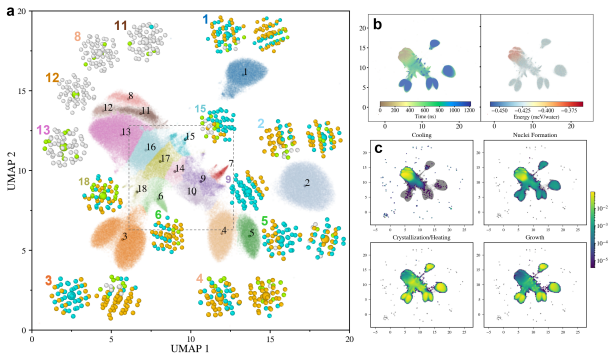

Fig. 3: **Clustering analysis of structural motifs during ice nucleation**. (a) UMAP projection of the VAE latent embedding (Z) for the nucleation trajectory labeled by HDBSCAN clusters. Local environments corresponding to the cluster medoids are visualized. (b) UMAP projection labeled by simulation time (ns) and $V_{ML,BOP}$ (meV/H_2O). (c) Density distribution of data at various stages of the nucleation trajectory. (d) Region on the UMAP illustrating transient motifs within the dashed rectangle in (a).

Motif Category	Cluster Labels	ρ_{voro} (g/cm^3)	Tetrahedral order (A)
Liquid	*8*, *11-13*	0.97~0.99	0.48~0.67
Transient	*6,7,9,10*, *14-18*	0.93~0.94	0.15~0.27
Ice	*1-5*	0.93~0.93	0.14~0.14

Table 1: **Categories and properties of structural motifs**. Summary of the different categories and of key physical properties of identified motifs in **Fig. 3**. The range of mean values of the physical properties across the different categories are shown, including the local atomic density ρ_{voro} (g/cm^3) calculated from the Voronoi cell volume, tetrahedral order parameter A. Note that a lower value of A indicates a greater tetrahedral order.

Acknowledgments

The authors acknowledge financial support provided by NUS HPC (NUSREC-HPC-00001 and CFP01-CF-077) and the Center of Biological Imaging Sciences (CBIS). We thank Bai Chang and Wang Junhong for HPC support, and Ervin Chia for fruitful discussions.

References

[1] Armin Kalita, Maximillian Mrozek-McCourt, Thomas F Kaldawi, Philip R Willmott, N Duane Loh, Sebastian Marte, Raymond G Sierra, Hartawan Laksmono, Jason E Koglin, Matt J Hayes, and et al. Microstructure and crystal order during freezing of supercooled water drops. *Nature*, 620(7974):557–561, 2023.

[2] Henry Chan, Mathew J Cherukara, Badri Narayanan, Troy D Loeffler, Chris Benmore, Stephen K Gray, and Subramanian KRS Sankaranarayanan. Machine learning coarse grained models for water. *Nat. Commun.*, 10(1):379, 2019.

[3] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B.*, 87(18):184115, 2013.

[4] Andrew H Nguyen and Valeria Molinero. Identification of clathrate hydrates, hexagonal ice, cubic ice, and liquid water in simulations: The chill+ algorithm. *J. Phys. Chem. B*, 119(29):9369–9376, 2015.

[5] Paul J Steinhardt, David R Nelson, and Marco Ronchetti. Bond-orientational order in liquids and glasses. *Phys. Rev. B.*, 28(2):784, 1983.

[6] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[7] Leland McInnes, John Healy, Steve Astels, and et al. Hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

[8] Jonas A Sellberg, C Huang, Trevor A McQueen, ND Loh, Hartawan Laksmono, Daniel Schlesinger, RG Sierra, Dennis Nordlund, CY Hampton, Dmitri Starodub, and et al. Ultrafast x-ray probing of water structure below the homogeneous ice nucleation temperature. *Nature*, 510(7505):381–384, 2014.

[9] Emily B Moore and Valeria Molinero. Structural transformation in supercooled water controls the crystallization rate of ice. *Nature*, 479(7374):506–508, 2011.

[10] Niloofar Esmaeildoost, Harshad Pathak, Alexander Späh, Thomas J Lane, Kyung Hwan Kim, Cheolhee Yang, Katrin Amann-Winkel, Marjorie Ladd-Parada, Fivos Perakis, Jayanath Koliyadu, and et al. Anomalous temperature dependence of the experimental x-ray structure factor of supercooled water. *J. Chem. Phys.*, 155(21), 2021.