Editors' Suggestion

Averaging Local Structure to Predict the Dynamic Propensity in Supercooled Liquids

Emanuele Boattini,[1] Frank Smallenburg,[2] and Laura Filion[1]

[1]*Soft Condensed Matter, Debye Institute of Nanomaterials Science, Utrecht University, 3584CC Utrecht, Netherlands*
[2]*Université Paris-Saclay, CNRS, Laboratoire de Physique des Solides, 91405 Orsay, France*

(Received 17 May 2021; accepted 19 July 2021; published 20 August 2021)

Predicting the local dynamics of supercooled liquids based purely on local structure is a key challenge in our quest for understanding glassy materials. Recent years have seen an explosion of methods for making such a prediction, often via the application of increasingly complex machine learning techniques. The best predictions so far have involved so-called Graph Neural Networks (GNNs) whose accuracy comes at a cost of models that involve on the order of 10^5 fit parameters. In this Letter, we propose that the key structural ingredient to the GNN method is its ability to consider not only the local structure around a central particle, but also averaged structural features centered around nearby particles. We demonstrate that this insight can be exploited to design a significantly more efficient model that provides essentially the same predictive power at a fraction of the computational complexity (approximately 1000 fit parameters), and demonstrate its success by fitting the dynamic propensity of Kob-Andersen and binary hard-sphere mixtures. We then use this to make predictions regarding the importance of radial and angular descriptors in the dynamics of both models.

DOI: 10.1103/PhysRevLett.127.088007

Unraveling the interplay between structure and dynamics in glassy materials is a major challenge in condensed matter science. When a liquid is rapidly cooled down or compressed to the point where it almost turns into a glass, its dynamics slow down by many orders of magnitude, while its structure typically stays largely unchanged. The dynamics of such glassy fluids are heterogeneous, with some regions of particles rearranging much more rapidly than others [1,2]. Key to understanding this phenomenon is identifying structural characteristics that are associated with these heterogeneities [3,4].

Traditionally, correlating structure and dynamics relied on physical intuition: one can look for local structural motifs [5–9]—such as icosahedra or tetrahedra—or for other (structure-dependent) local physical features of the system [10–13]—such as local density, or potential energy—that are expected to play a key role in determining the dynamics. A novel approach was pioneered in 2015 by Cubuk *et al.* [14], who demonstrated that machine learning (ML) techniques could be trained to identify slow and fast regions in a glassy liquid. Since then, a number of works have demonstrated the power of both supervised and unsupervised machine learning for correlating structure and dynamics in a variety of glassy systems [14–19].

In recent years, many studies attempting to unravel the link between structure and dynamics have tried to reveal structural quantities capable of predicting the dynamic propensity [4–6,11,15,16,20–24]. The dynamic propensity of a particle is the absolute [25] distance that the particle moves over a given time interval, averaged over many simulated trajectories starting from the same initial configuration. In essence, it provides a measure of the future mobility of a particle, as governed by the structure of its surroundings [26]. To date, the most accurate predictions of the dynamic propensity in glassy systems were achieved by Bapst *et al.* [23] using a highly advanced machine learning approach: Graph Neural Networks (GNNs). While the GNN is quite accurate, its design philosophy draws on physical insights as little as possible, resulting in a high computational complexity. Specifically, since the GNN designs its own structural descriptors, training it requires optimizing an enormous number of parameters (∼70 000 in Ref. [23]). As a result, the training requires both a large dataset (to avoid overfitting) and significant computational effort. Nonetheless, the predictive power of GNNs clearly indicates that their model architecture is able to capture the essential physics required to predict dynamics from local structure. This raises two important questions that we address in this Letter: Can we learn from the success of GNNs what structural features we need to consider to accurately predict local dynamics, and can we exploit these observations to design a significantly more efficient model that performs equally well?

To address the first question, we need to consider the architecture of a GNN and compare it to simpler machine learning approaches. In most ML approaches, such as the support vector machines (SVMs) used in Refs. [14,17], the environment of each particle is captured via a set of handcrafted structure functions centered around the particle under consideration, which provide information about, e.g., the radial density profile and bond angle distribution. Subsequently, a model is fitted that relates these structural descriptors to a dynamical descriptor of each particle. In contrast, in the GNN used in Ref. [23], the input is a graph, where each node represents a particle, and particles closer than a certain cutoff distance are connected by an edge which carries as information the vector connecting the two particles. After an encoding step, this information is then passed through a number of recursive iterations. In each iteration, the graph is mapped to a new graph with the same topology, but with updated information on the nodes and edges, where the mappings are nonlinear functions described by neural networks. Finally, a "decoder" step is used to produce a prediction for the desired dynamical descriptor—in the case of Ref. [23] the dynamic propensity.

The key trick of the GNN lies in the fact that in each successive iteration, information from further away is incorporated into the information for a given node or edge—in an average sense. Hence, in contrast to the more standard hand-crafted descriptors which are generally only centered on the particle under consideration, the GNN designs its own descriptors, that can in principle take into account averaged structural features at significant distances away. This strategy is reminiscent of the observation that a (single) local average of a structural descriptor often improves its ability to predict local dynamics [5,6,15,16,20]. As shown in the recent publication by Bapst *et al.* [23], this strategy works very well at fitting the dynamics of a supercooled Kob-Andersen mixture, outperforming all previous algorithms.

This raises the question: Would incorporating the shell-averaging concept of GNNs into handcrafted descriptors lead to similar predictive power? Here we design a set of descriptors that recursively incorporate information from multiple neighbour shells and fit them to dynamical information using simple linear regression.

We begin with a set of descriptors that encode the structure around each particle i, denoted as the vector $\mathbf{X}_i^{(0)}$. Note that many glassy systems consist of particles of two or more different species, and hence these descriptors take into account both the positions of the particles as well as their species. For $\mathbf{X}_i^{(0)}$, we design a set of structural descriptors consisting of both radial and angular structure functions. For the radial descriptors, we consider the same type of functions that were used in Refs. [17,23] in combination with SVMs. These functions essentially measure the density of particles at a distance r from a reference particle i in a shell of width 2δ, and are defined as follows:

$$G_i^{(0)}(r,\delta,s) = \sum_{j\neq i:s_j=s} e^{-\frac{(r_{ij}-r)^2}{2\delta^2}} \tag{1}$$

where i is the reference particle, r_{ij} is the distance between particle i and j, s_j is the species of particle j, and s is the species of particles whose density we wish to probe. For the angular descriptors, inspired by standard bond-orientational-order parameters [27], we use an expansion of the local density in terms of spherical harmonics. Similar descriptors were also explored in Ref. [17]. First, for any given particle i, we define the complex quantities

$$q_i^{(0)}(l,m,r,\delta) = \frac{1}{Z}\sum_{j\neq i} e^{-\frac{(r_{ij}-r)^2}{2\delta^2}} Y_l^m(\mathbf{r}_{ij}), \tag{2}$$

where $Y_l^m(\mathbf{r}_{ij})$ are the spherical harmonics of order l, with m an integer that runs from $m=-l$ to $m=+l$, and $Z=\sum_{j\neq i} e^{-((r_{ij}-r)^2/2\delta^2)}$ is a normalization constant. We then construct a rotationally invariant local descriptor

$$q_i^{(0)}(l,r,\delta) = \sqrt{\frac{4\pi}{2l+1}\sum_{m=-l}^{l}|q_i^{(0)}(l,m,r,\delta)|^2}. \tag{3}$$

The full vector $\mathbf{X}_i^{(0)}$ for a given particle i then consists of the values of $G_i^{(0)}(r,\delta,s)$ and $q_i^{(0)}(l,r,\delta)$, evaluated for a fixed set of r, δ, and l as specified in the Supplemental Material (SM) [28].

In order to incorporate the shell-averaging concept from GNNs, we then introduce higher-order descriptors $\mathbf{X}_i^{(n)}$, where each consecutive $\mathbf{X}_i^{(n)}$ is defined as a local average of the previous order $\mathbf{X}_i^{(n-1)}$. Specifically

$$\mathbf{X}_i^{(n)} = \frac{1}{C}\sum_{j:r_{ij}<r_c} e^{-r_{ij}/r_c}\mathbf{X}_j^{(n-1)}, \tag{4}$$

where r_c is a cutoff radius and $C=\sum_{j:r_{ij}<r_c} e^{-r_{ij}/r_c}$. Here, we choose r_c to approximately correspond to the second minimum in the radial distribution function, and we have confirmed that the results are only weakly dependent on the exact value (see SM).

The total descriptors for particle i, which we denote $\boldsymbol{\mathcal{X}}^{(n_{\max})}$, is then the combination of descriptors (angular and radial) from each shell up to a maximum level of $n_{\max}$.

Now that we have introduced a new set of descriptors, we explore how well they can be used to fit the dynamics, and whether including multiple generations (i.e., larger $n_{\max}$) improves the fitting quality. For our fitting approach, we use linear regression including a regularization term, also known as Ridge regression [29]. For this simple fitting algorithm, the number of fit parameters is simply equal to the number of descriptors (plus an offset), and the fit can be performed on a basic workstation in a matter of seconds.

As our dynamic quantity of interest, we use the dynamic propensity [6,11,26], a standard method for quantifying the component of the dynamical heterogeneity that is encoded in the structure. For a given configuration, it is found by performing M simulations of the same configuration, each initialized with a new set of velocities drawn from the Maxwell Boltzmann distribution. The dynamic propensity $d(t)$ of a chosen particle is then its average absolute displacement after time t. Note that in this Letter, all times are measured in units of the structural relaxation time τ_α.

As our model system, we first focus on precisely the same system as Ref. [23], i.e., the Kob-Andersen (KA) mixture [30], a well studied glass former that consists of an 80:20 mixture of nonadditive Lennard-Jones particles (see SM for further details). We construct a set of descriptors consisting of 200 radial and 192 angular descriptors per generation. We begin by exploring how the number of averages (i.e., $n_{\max}$) influences our fit quality, as measured by the Pearson correlation coefficient between the predicted and measured propensities (see SM). Clearly, as shown in Fig. 1, for both species in the KA mixture, there is a significant improvement by including the first averaging (LR$^{(1)}$), while adding a second averaging step (LR$^{(2)}$) only improves the correlations slightly. Adding a third averaging step does not lead to any further improvements.

The question is now: How does this prediction compare to other ML methods? As shown in Fig. 1(a), the zeroth order descriptors lead to a prediction that is similar in quality to the SVM fit taken from Ref. [23]. This is not surprising, as both the SVM method and our LR$^{(0)}$ prediction are linear models based on a similar set of descriptors. Much more interesting, however, is that our LR$^{(2)}$ results closely match the GNN results for all time scales. This indicates that we have indeed incorporated in our averaged descriptors the same relevant structural features captured by the GNN. Additionally, it should be noted that, as expected, our linear model requires significantly less training data than the GNN (see SM). Moreover, as also shown in the SM, the linear model trained at this low temperature can even give meaningful predictions for the dynamical behavior at lower degrees of supercooling.

One clear advantage of our approach compared to GNNs is the minimal cost required for training the model, allowing one to rapidly explore the importance of different classes of descriptors. While GNNs approximate the propensity as a complex nonlinear function of the particles' coordinates and species, our model is a linear combination of structural descriptors with a simple physical interpretation. As a result, we can exploit the interpretability of the input descriptors in order to unveil the structural features that are most relevant for predicting the dynamics. For example, we can ask the question: Is radial (or density) information sufficient to accurately predict the dynamics, or do we also need angular information? This question is intriguing in the context of the Kob-Andersen system, as previous studies similar to our LR$^{(0)}$ model have identified radial information to be the most important [16,17]. To answer this question, we separately fit the particles' dynamic propensity using only the radial and angular descriptors, and show the results in Fig. 1(c). When only nonaveraged descriptors (0th order) are included, we find the radial descriptors to be more informative on the dynamics (especially at long times), in agreement with previous works [16,17]. Interestingly, however, when averaged descriptors ($n_{\max} = 2$) are included, better predictions are obtained with the set of angular functions. One way to interpret this result is that when the region considered is small, the local density is the most important

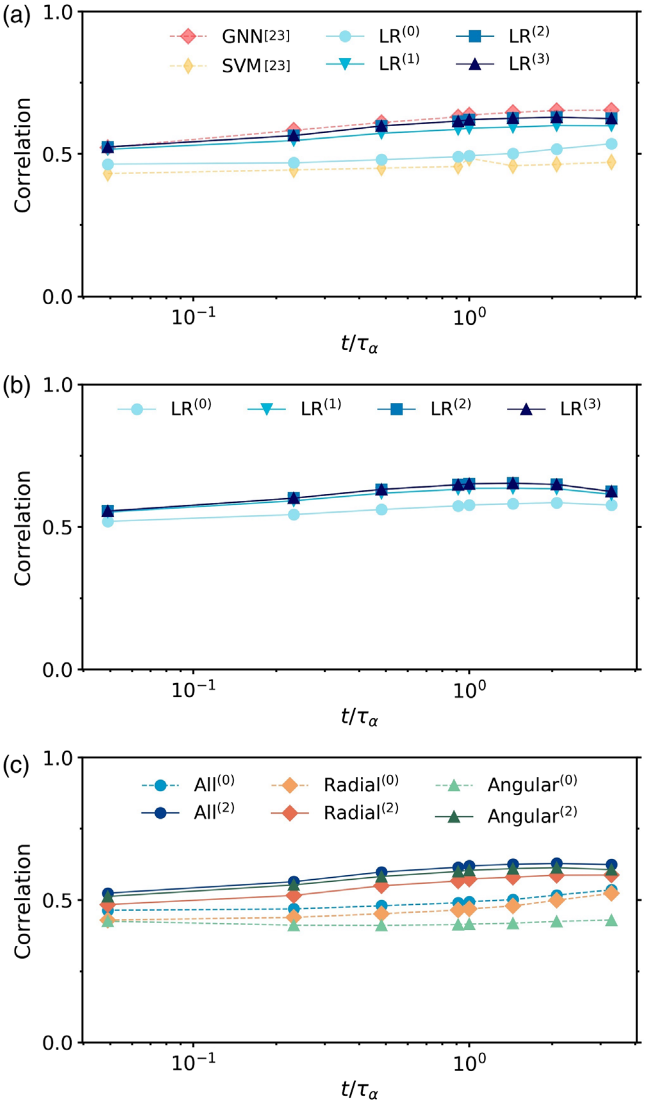

FIG. 1. (a) Pearson correlation coefficient between predicted and actual propensities for the A particles of the KA system at temperature $k_B T/\epsilon_{AA} = 0.44$ and pressure $P\sigma_{AA}^3/\epsilon_{AA} = 2.93$. LR$^{(n_{\max})}$ refers to linear regression using structural descriptor up to order $n_{\max}$. For comparisons, the results obtained in Ref. [23] using support vector machines (SVMs) and graph neural networks (GNNs) are also shown. (b) Pearson correlation coefficient between predicted and actual propensities for the B particles of the same system. (c) Comparison of the Pearson correlation coefficient between predicted and actual propensities of A particles obtained using only radial descriptors, only angular descriptors, or both.

feature, whereas when considering larger length scales, the anisotropy of the structural environment becomes more relevant. This is consistent with the observation that the local environment of particles can show angular ordering over impressively long ranges in KA mixtures [31].

In our current approach we have included approximately 10^3 descriptors. However, the linear nature of our model makes it easy to reduce the number significantly at a low cost in terms of accuracy—a highly useful feature for future extensive explorations of the relationship between structure and dynamics of glassy fluids. To this end, we employ the feature selection scheme introduced in Ref. [32] in the context of approximating many-body interactions. In this scheme, the most relevant descriptors are iteratively selected from a pool of candidates. At each step, the selected descriptor is the one that maximizes the linear correlation between the currently selected set and a target variable. The selection proceeds until the correlation stops increasing appreciably. Here, we consider the set of all descriptors up to order $n_{\max} = 2$ as the pool of candidates, and use this scheme to select an optimal subset of N_s descriptors for predicting the dynamic propensity at time t. In Fig. 2, we report the results of the predictions at $t = \tau_\alpha$ as a function of the number of selected descriptors. As shown in the figure, the best descriptor in the pool has a correlation of about 0.4 with the dynamic propensity. Moreover, using as few as $N_s = 6$ descriptors, the results already exceed those obtained with SVMs in Ref. [23] (that used 440 descriptors). After that, the results keep improving as N_s increases, and do not change appreciably after $N_s \approx 100$ descriptors have been selected. A table of the first 20 descriptors selected is given in the SM.

The final question we would like to address is how robust this algorithm is—e.g., how well does it perform on a different glass former? To this end, we consider a 30–70 binary mixture of large (A) and small (B) hard spheres with size ratio $\sigma_B/\sigma_A = 0.85$ and packing fraction $\eta = 0.58$. Previous works have shown that the local structure of this system correlates quite strongly with the dynamic propensity at times close to the relaxation time [5,15], with the strongest correlation reported associated with an unsupervised machine-learned (UML) order parameter based on averaged order parameters (somewhat comparable to our LR$^{(1)}$ averaging). For hard spheres, we use the same set of descriptors as for KA, but omit radial descriptors taken at distances where no pairs of particles can exist (resulting in 172 radial descriptors). Using these descriptors with different choices of the maximum order $n_{\max}$, we fit the propensities at different times, and report the results obtained for both species of particles in Fig. 3. Clearly, both the LR$^{(1)}$ and LR$^{(2)}$ outperform the UML over all time frames examined, with the smallest difference appearing near τ_α. Interestingly, only about 20 descriptors are necessary to reach approximately the optimum prediction accuracy (see SM, Fig. S3). Similar to the KA system, the radial LR$^{(0)}$ descriptors outperform the angular descriptors at long time scales (see SM, Fig. S4). Overall, as seen in Fig. 3, consistent with previous observations [15], the predictions are much more accurate than those obtained for the KA system, likely due to the simpler dynamics of this system, which lacks both attractions and nonadditivity. For the

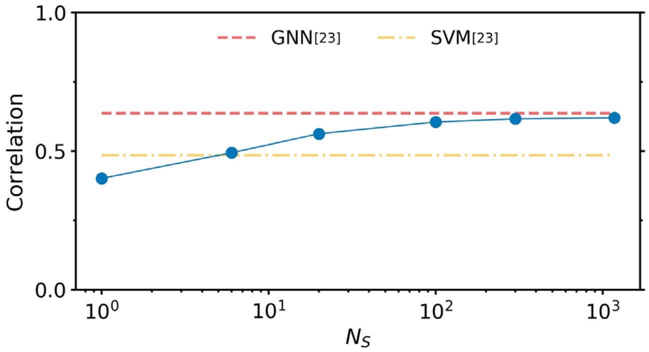

FIG. 2. Pearson correlation coefficient between predicted and actual propensities of the A particles of the Kob-Andersen mixture (same state point as Fig. 1) at $t = \tau_\alpha$ as a function of the number of selected descriptors. Lines represent the results obtained in Ref. [23] using SVMs and GNNs.

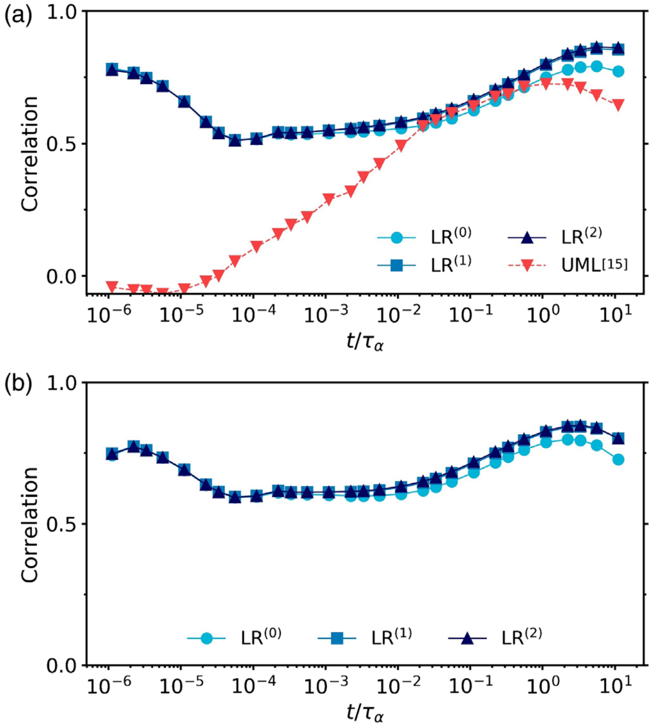

FIG. 3. Pearson correlation coefficient between predicted and actual propensities for a hard-sphere mixture at size ratio $q = 0.85$, packing fraction $\eta = 0.58$, and composition $x_A = 0.3$. The results are shown for both (a) large A particles and (b) small B particles. For comparison, in (a) we also report the results of the UML approach from Ref. [15]. Note that these results were obtained on a different dataset consisting of a single snapshot at the same state point considered here.

same reason, we also find that the inclusion of averaged descriptors has a weaker impact on the accuracy of the model, and the results essentially stop improving after including descriptors of order $n_{\max} = 1$. An intriguing observation is the fact that the peak in correlation for $\mathrm{LR}^{(1)}$ and $\mathrm{LR}^{(2)}$ occurs at significantly longer time scales than the peak in correlation associated with the UML approach, and indeed at time scales several times longer than the structural relaxation time τ_α.

In conclusion, we have introduced a fast, easy to implement, linear-regression-based model for fitting dynamic propensities in glassy fluids from local structural descriptors. Key to this model was the insight from GNNs that averaged structural features centered around nearby particles carry a significant amount of the necessary information required to predict the heterogeneous dynamics. This observation enabled us to design a significantly more efficient model that provides essentially the same predictive power at a fraction of the computational complexity—from the ~70 000 parameters of the GNN to approximately 1000 parameters in the linear regression model at $\mathrm{LR}^{(2)}$. Moreover, we show that by ranking the importance of the descriptors, we can further reduce the number of required descriptors by an order of magnitude. This result not only provides an efficient simple model for fitting the dynamic propensity of glassy fluids, but also suggests that similar local-average-based linear models should be considered in other situations where GNNs are applied to predict structural and dynamical properties of materials [33–35].

Perhaps the most intriguing observation in this work is that the linear model presented here and the GNNs predict the dynamic propensity to essentially the same accuracy. Given that the dynamic propensity must be completely encoded in the structure by its definition, the new linear model opens the door to asking—what structural information is missing to completely describe the dynamics? The linear model would appear to be missing information related to both anisotropic correlations within the averaged domains, as well as correlations between the averaged domains. This observation should lay the foundation for further extensions and improvements in fitting the dynamics propensity in the future.

We would like to thank Giuseppe Foffi for many useful discussions. L. F. and E. B. acknowledge funding from The Netherlands Organisation for Scientific Research (NWO) (Grant No. 16DDS004), and L. F. acknowledges funding from NWO for a Vidi grant (Grant No. VI.VIDI.192.102).

[1] M. D. Ediger, Annu. Rev. Phys. Chem. **51**, 99 (2000).

[2] L. Berthier, G. Biroli, J.-P. Bouchaud, L. Cipelletti, and W. van Saarloos, *Dynamical Heterogeneities in Glasses, Colloids, and Granular Media* (Oxford University Press, Oxford, 2011), Vol. 150.

[3] C. P. Royall and S. R. Williams, Phys. Rep. **560**, 1 (2015).

[4] H. Tanaka, H. Tong, R. Shi, and J. Russo, Nat. Rev. Phys. **1**, 333 (2019).

[5] S. Marín-Aguilar, H. H. Wensink, G. Foffi, and F. Smallenburg, Phys. Rev. Lett. **124**, 208005 (2020).

[6] H. Tong and H. Tanaka, Phys. Rev. X **8**, 011041 (2018).

[7] H. Tong and H. Tanaka, Nat. Commun. **10**, 5596 (2019).

[8] A. Malins, J. Eggers, C. P. Royall, S. R. Williams, and H. Tanaka, J. Chem. Phys. **138**, 12A535 (2013).

[9] M. Leocmach and H. Tanaka, Nat. Commun. **3**, 974 (2012).

[10] B. Doliwa and A. Heuer, Phys. Rev. Lett. **91**, 235501 (2003).

[11] A. Widmer-Cooper and P. Harrowell, Phys. Rev. Lett. **96**, 185701 (2006).

[12] A. Widmer-Cooper, H. Perry, P. Harrowell, and D. R. Reichman, Nat. Phys. **4**, 711 (2008).

[13] D. Richard, M. Ozawa, S. Patinet, E. Stanifer, B. Shang, S. Ridout, B. Xu, G. Zhang, P. Morse, J.-L. Barrat *et al.*, Phys. Rev. Mater. **4**, 113609 (2020).

[14] E. D. Cubuk, S. S. Schoenholz, J. M. Rieser, B. D. Malone, J. Rottler, D. J. Durian, E. Kaxiras, and A. J. Liu, Phys. Rev. Lett. **114**, 108001 (2015).

[15] E. Boattini, S. Marín-Aguilar, S. Mitra, G. Foffi, F. Smallenburg, and L. Filion, Nat. Commun. **11**, 5479 (2020).

[16] J. Paret, R. L. Jack, and D. Coslovich, J. Chem. Phys. **152**, 144502 (2020).

[17] S. S. Schoenholz, E. D. Cubuk, D. M. Sussman, E. Kaxiras, and A. J. Liu, Nat. Phys. **12**, 469 (2016).

[18] S. S. Schoenholz, E. D. Cubuk, E. Kaxiras, and A. J. Liu, Proc. Natl. Acad. Sci. U.S.A. **114**, 263 (2017).

[19] F. P. Landes, G. Biroli, O. Dauchot, A. J. Liu, and D. R. Reichman, Phys. Rev. E **101**, 010602(R) (2020).

[20] G. M. Hocky, D. Coslovich, A. Ikeda, and D. R. Reichman, Phys. Rev. Lett. **113**, 157801 (2014).

[21] A. J. Dunleavy, K. Wiesner, R. Yamamoto, and C. P. Royall, Nat. Commun. **6**, 6089 (2015).

[22] S. Pan, S. Feng, J. Qiao, W. Wang, and J. Qin, J. Alloys Compd. **664**, 65 (2016).

[23] V. Bapst, T. Keck, A. Grabska-Barwińska, C. Donner, E. D. Cubuk, S. Schoenholz, A. Obika, A. Nelson, T. Back, D. Hassabis *et al.*, Nat. Phys. **16**, 448 (2020).

[24] C. Balbuena, M. M. Gianetti, and E. R. Soulé, Soft Matter **17**, 3503 (2021).

[25] Note that some implementations of the dynamic propensity use the square distance, rather than the absolute one.

[26] L. Berthier and R. L. Jack, Phys. Rev. E **76**, 041509 (2007).

[27] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, Phys. Rev. B **28**, 784 (1983).

[28] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.127.088007 for additional details on the method and additional analysis of the results.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011).

[30] W. Kob and H. C. Andersen, Phys. Rev. E **51**, 4626 (1995).

[31] Z. Zhang and W. Kob, Proc. Natl. Acad. Sci. U.S.A. **117**, 14032 (2020).

[32] E. Boattini, N. Bezem, S. N. Punnathanam, F. Smallenburg, and L. Filion, J. Chem. Phys. **153**, 064902 (2020).

[33] T. Xie and J. C. Grossman, Phys. Rev. Lett. **120**, 145301 (2018).

[34] S.-Y. Louis, Y. Zhao, A. Nasiri, X. Wang, Y. Song, F. Liu, and J. Hu, Phys. Chem. Chem. Phys. **22**, 18141 (2020).

[35] M. Schwarzer, B. Rogan, Y. Ruan, Z. Song, D. Y. Lee, A. G. Percus, V. T. Chau, B. A. Moore, E. Rougier, H. S. Viswanathan *et al.*, Comput. Mater. Sci. **162**, 322 (2019).