Supplementary information for: Extraction of local structure difference in Silica based on unsupervised learning

Anh Khoa Augustin Lu,[1,2,*] Jianbo Lin,[3] Yasunori Futamura,[4,5,6]
Tetsuya Sakurai,[4,5,6] Ryo Tamura,[3,7,†] and Tsuyoshi Miyazaki[1,6,8,‡]

[1] *Research Center for Materials Nanoarchitectonics,*
National Institute for Materials Science, Tsukuba 305-8568, Japan

[2] *Mathematics for Advances Materials Open Innovation Laboratory,*
National Institute of Advanced Industrial Science and Technology, Sendai 980–8577, Japan

[3] *Center for Basic Research on Materials,*
National Institute for Materials Science, Tsukuba 305-0047, Japan

[4] *Department of Computer Science, University of Tsukuba, Tsukuba 305-8573, Japan*

[5] *Center for Artificial Intelligence, University of Tsukuba, Tsukuba 305-8573, Japan*

[6] *Master's/Doctoral Program in Life Science Innovation,*
University of Tsukuba, Tsukuba 305-8577, Japan

[7] *Graduate School of Frontier Sciences,*
The University of Tokyo, Chiba 277-8568, Japan

[8] *Graduate School of Engineering, Nagoya university, Nagoya 464-8603, Japan*

(Dated: March 5, 2024)

[*] LU.Augustin@nims.go.jp, augustinlu@gmail.com

[†] TAMURA.Ryo@nims.go.jp

[‡] MIYAZAKI.Tsuyoshi@nims.go.jp

I. SIZE OF DATA SETS

	α-quartz	β-quartz	β-cristobalite	β-tridymite	Coesite	Stishovite	Liquid	Glass	Total
α-quartz and β-quartz (section II A)	1,296	1,296	0	0	0	0	0	0	2,592
β-cristobalite and β-tridymite (section II B)	0	0	768	1,536	0	0	0	0	2,304
Liquid and glass (section II C)	0	0	0	0	0	0	1,296	2,592	3,888
All eight phases (main text)	1,080	1,080	640	1,280	640	900	1,080	2,160	8,860

TABLE S1: Size of the data sets for each study case presented in this supplementary information file

The size of the data sets are listed in Table S1.

II. DIFFERENTIATION OF PARTICULAR PHASES

A. α-quartz and β-quartz

The α-β transition in quartz has been studied in the past [1]. It consists of the transformation from a tetragonal symmetry to a hexagonal symmetry (Figure S1), which occurs at 848 K under atmospheric pressure. The radial distribution function (RDF) of both quartz phases are shown in Figure S2. They can be distinguished from around 3 Å of distance.

The impact of the cutoff for the descriptor R_d is evaluated with projections to a two-dimensional space. As shown by Figure S3, low cutoff values such as 2 Å or 3 Å do not produce spaces where the two phases can be distinguished. This is in agreement with the RDF (Figure S2), which cannot be distinguished up to around 2.5 Å. A slight difference appears near 3 Å but setting R_d failed to generate a proper low-dimensional space. However, from 4 Å, a separation of the data points is observed and the two set of data points are properly divided at $R_d = 5$ Å and beyond. Therefore, in this case a cutoff value for the descriptor of 5 Å is considered appropriate.

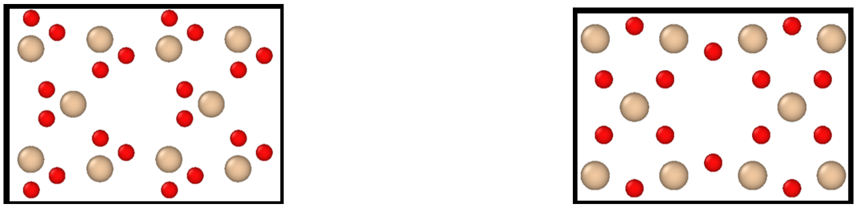


FIG. S1: Visual representation of (left) α-quartz and (right) β-quartz

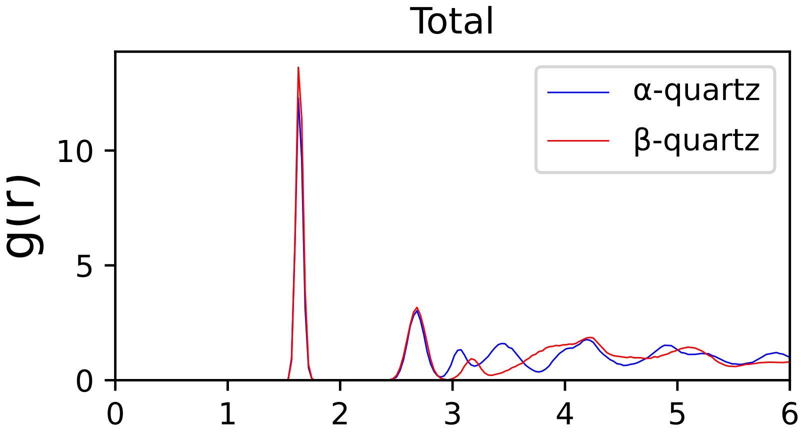


FIG. S2: RDF of α- and β-quartz

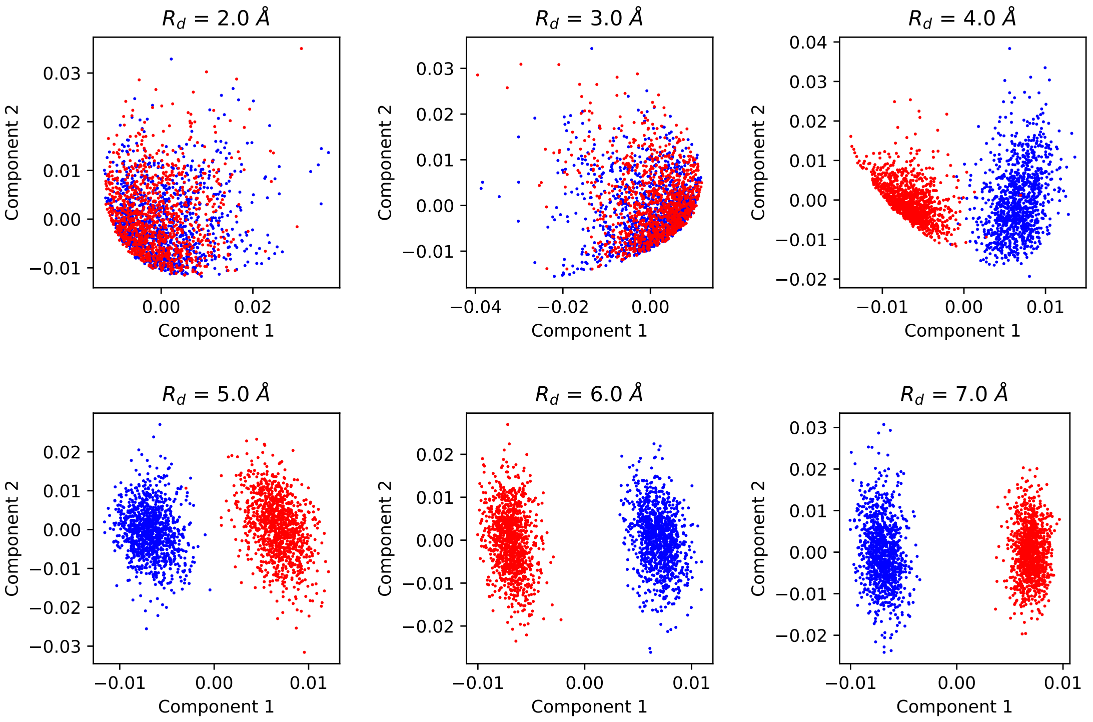


FIG. S3: Distribution of the projected data points in the low-dimensional space with respect to the cutoff distance R_d. Projected data from α(resp. β)-quartz are shown in blue (resp. red).

1. *PCA, LPP and TS-LPP on quartz phases*

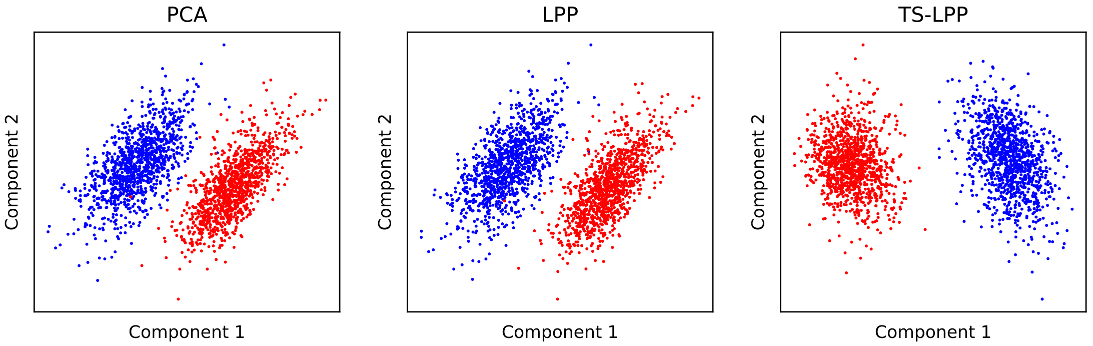


FIG. S4: PCA, LPP and TS-LPP compared for α-quartz and β-quartz

Figure S4 shows the comparison between PCA, LPP and TS-LPP ($R_d = 5$ Å). Note that for TPP, a large value of σ was found to be optimal, leading to a space nearly identical to PCA.

B. β-cristobalite and β-tridymite

β-cristobalite and β-tridymite are found at high temperature and are relatively difficult to be distinguished from each other according to the radial distribution function, as shown in Figure S5. Compared to the case of quartz (Figure S3), it appears that for β-cristobalite and β-tridymite, a higher cutoff is needed to make a proper distinction (Figure S6). Nevertheless, many outliers data still remain, even at R_d = 6 Å and R_d = 7 Å. In this case, locally averaging the atomic fingerprint (Equation 8) of the Si atoms allows proper differentiation. For this pair of phases, a reasonable cutoff for averaging is 4 Å with a cutoff for descriptor of 6 Å.

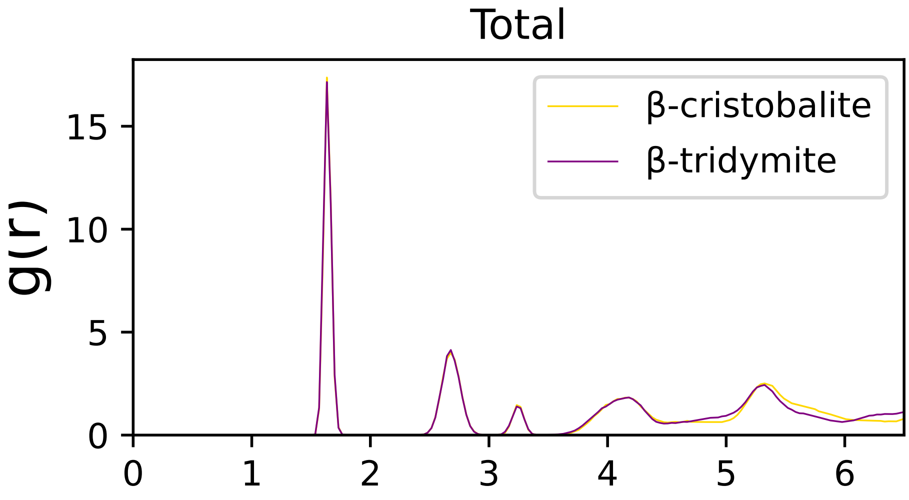


FIG. S5: RDF of beta-cristobalite and beta-tridymite (300 K)

Interestingly, only TS-LPP can properly differentiate the two phases, while PCA and LPP are unable to properly captures the difference between the two phases, as shown in Figure S7.

C. Liquid and glass

Aside from crystals, it is important to be able to differentiate disordered phases. Here, we considered two trajectories for the liquid phase (3,000 K and 4,000 K) and four trajectories for the glass phase (300 K, with structure obtained by melt and quench with a cooling rate of 10^{10}, 10^{11}, 10^{12} and 10^{13} K/s). The system size for these disordered structures is small due to the fact that a first principles accuracy was desired when generating the training

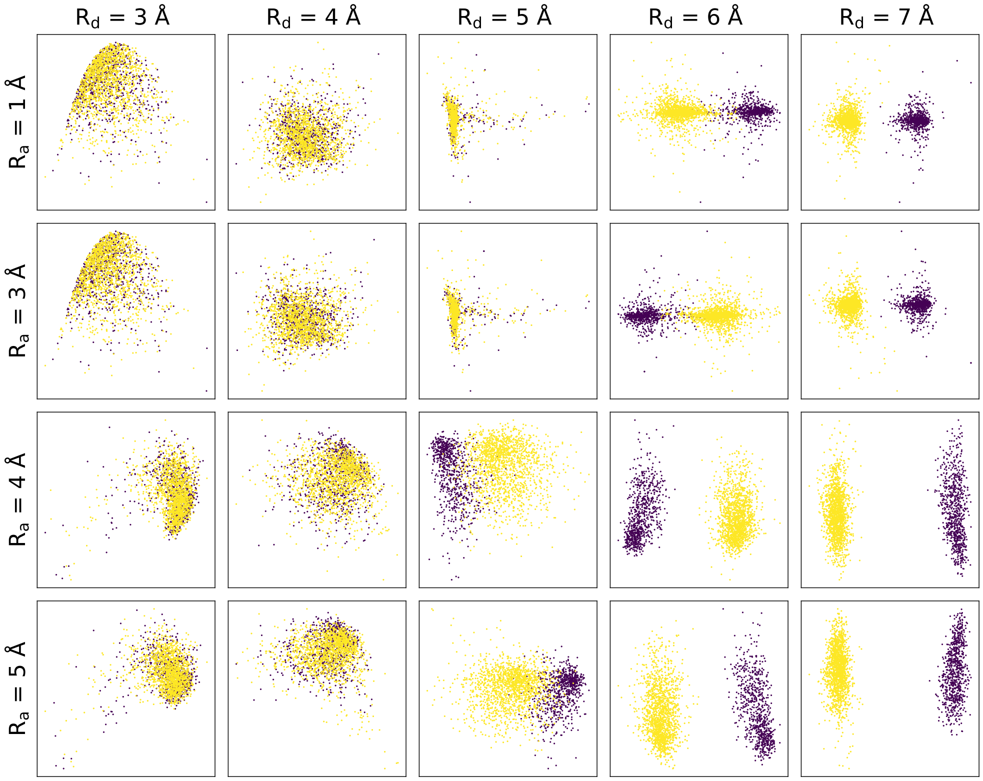


FIG. S6: Distribution of the projected data points in the low-dimensional space with respect to the cutoff distances R_d and R_a. Projected data from descriptors of β-cristobalite (resp. β-tridymite) are shown in dark purple (resp. yellow).

sets. In later studies, larger cells should be studied, either with linear-scale first-principles calculations [2] or machine-learning force fields [3].

For the disordered phases, as was the case for the high-temperature crystal phases, local averaging of the descriptors is needed in order to properly distinguish liquids from glasses. The disordered nature of liquids and glasses means that there is more variability in local environment than in crystal phases. Having multiple different trajectories for each phase avoid developing models too specialized on a single configuration. Here, setting both cutoffs for descriptors and averaging to 5 Å and 5 Å is necessary to obtain a satisfactory differentiation between the liquid and glass phases. An averaging cutoff of 6 Å enables a better separation, as shown in Figure S8.

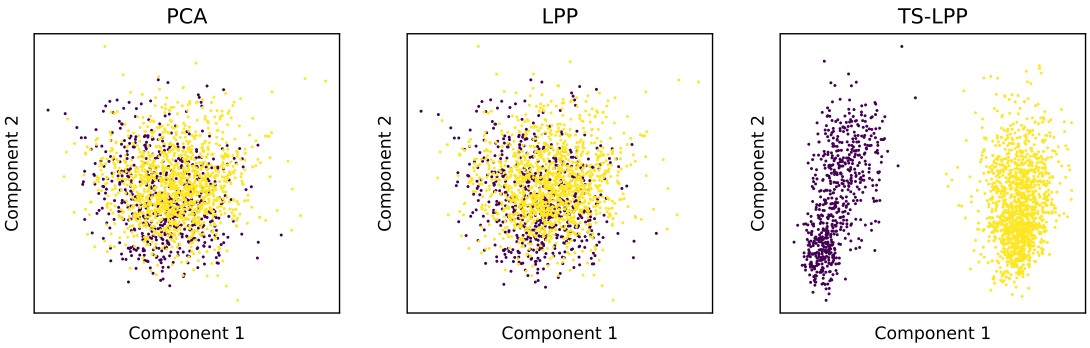


FIG. S7: PCA, LPP and TS-LPP compared for distinction between β-cristobalite and β-tridymite

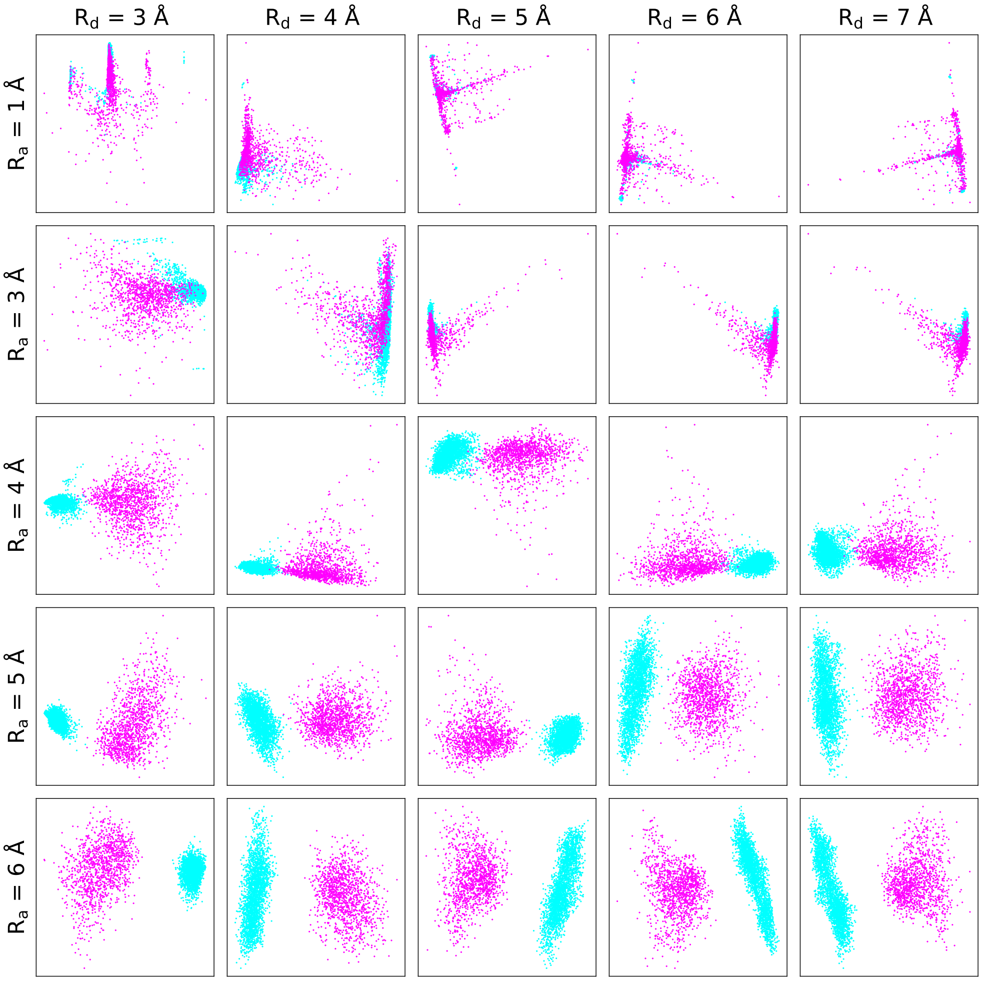


FIG. S8: Distribution of the projected data points in the low-dimensional space with respect to the cutoff distances R_d and R_a. Projected data from liquid (resp. glass) are shown in pink (resp. light blue).

III. DISCUSSION ON THE PARAMETERS

A. Distance cutoffs

The cutoffs for the descriptor and for local averaging were determined by fist verifying the ability of a smaller model to distinguish pairs of phases, namely:

1. α-quartz and β-quartz : $R_d \geq 5$ Å; R_a not needed

2. β-cristobalite and β-tridymite: $R_d \geq 6$Å; $R_a \geq 4$ Å

3. liquid and glass: $R_d \geq 5$ Å; $R_a \geq 6$ Å

To ensure a proper differentiation of these phases, the cutoffs were both set to $R_d = R_a = 6$ Å. Details can be found in the supplementary information (Figure S3, Figure S6, Figure S8).

B. Feature selection and standardization

For some of the LAAF features with small η_m, the variance is vanishingly small, as shown in Figure S9. These features are removed by setting a variance threshold, illustrated by the dash blue line in Figure S9. Different values for the variance threshold were tested and it was found that a variance threshold superior to $1e-5$ was required to properly differentiate all phases. By using a variance threshold of 10^{-4}, 63 of the 100 features are kept. The remaining features are standardized to unit variance.

C. Number of dimensions

In previous works, a two-dimensional space was generally used, which facilitates the visualization. In the case of silica, due to the high number of possible phases, no satisfactory two-dimensional space could be constructed. At least 7 dimensions are needed to capture relevant information about the local structures, as can be seen from Table S2. Except for stishovite in which Si atoms have six neighboring O, each Si atom has four O neighbors.

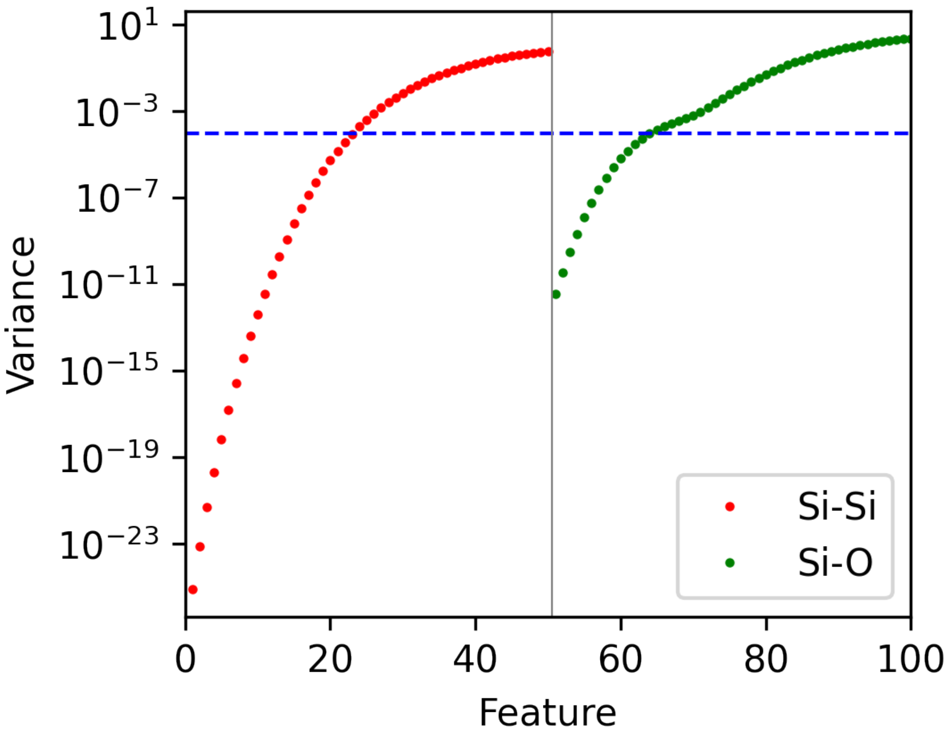


FIG. S9: Variance of the LAAF features. The blue dashed line indicates the variance threshold used in this work.

# dim.	d_m	σ	Stishovite	$\alpha-$quartz and $\beta-$quartz	Liquid and glass	$\beta-$tridymite and $\beta-$cristobalite
1	30	1	**Yes**	No	No	No
2	20	1	**Yes**	No	No	No
3	20	1	**Yes**	**Yes**	No	No
4	30	20	**Yes**	**Yes**	No	No
5	30	20	**Yes**	**Yes**	No	**Yes**
6	10	1	**Yes**	**Yes**	**Yes**	No
7	20	1	**Yes**	**Yes**	**Yes**	**Yes**
8	20	1	**Yes**	**Yes**	**Yes**	**Yes**

TABLE S2: Number of dimensions and proper differentiation of specific phases

IV. CLUSTERING ON LAAF DESCRIPTORS

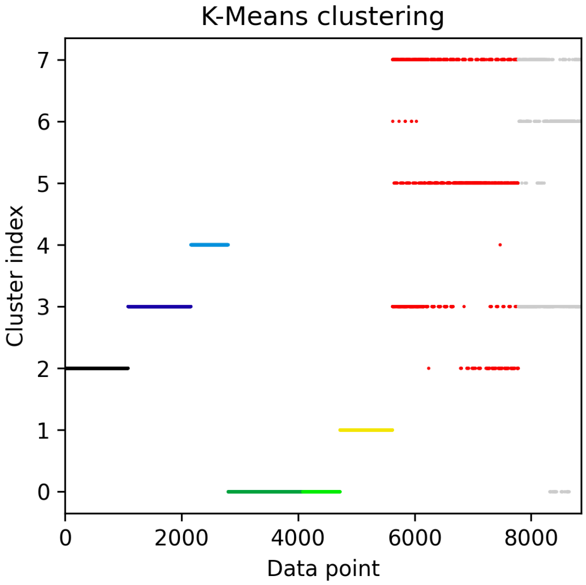


FIG. S10: Labels attributed by k-means using the original descriptors

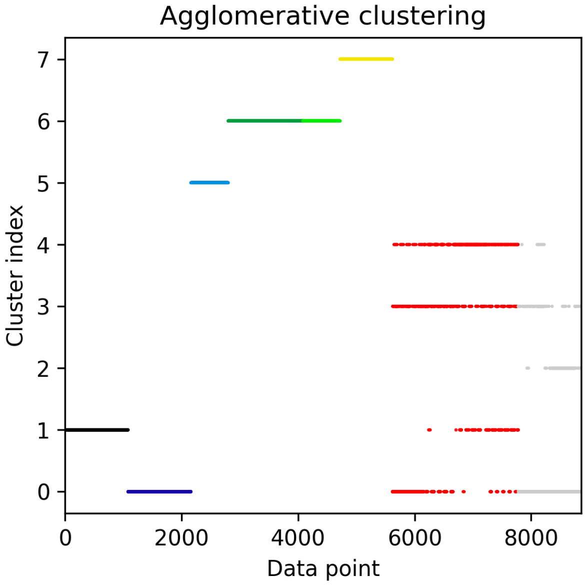


FIG. S11: Labels attributed by agglomerative clustering using the original descriptors

Figure S10 and Figure S11 show the results of k-means clustering and agglomerative clustering used on the original descriptor in the high-dimensional space.

V. LPP WITH HIGH DIMENSIONALITY

LPP was performed to reduce the space to 15 dimensions Figure S12. In this space, the phases can be distinguished, for instance by k-means clustering Figure S13.

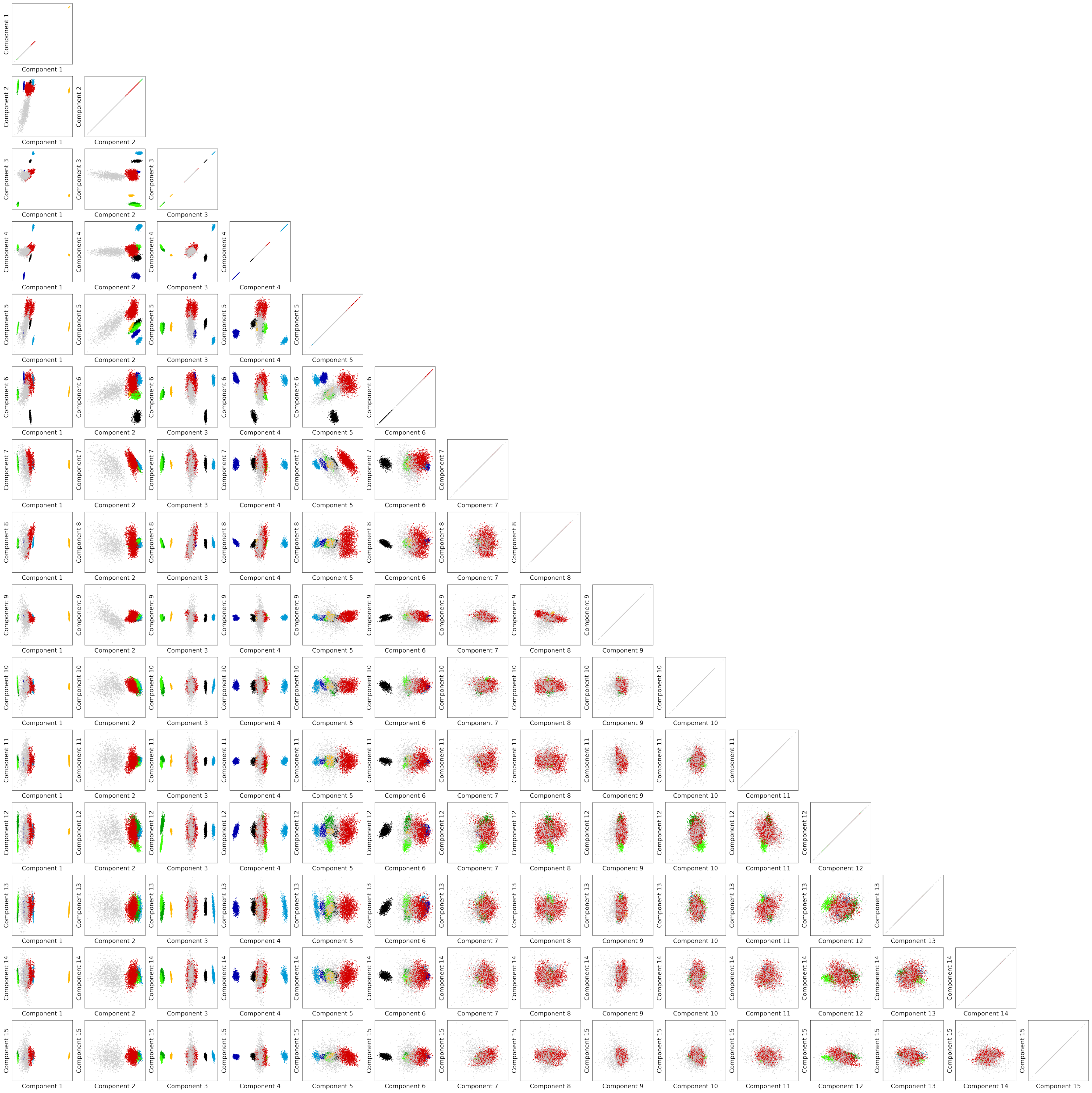


FIG. S12: Visual representation of the 15-dimensional space generated by LPP with $\sigma = 20$ by pair of components

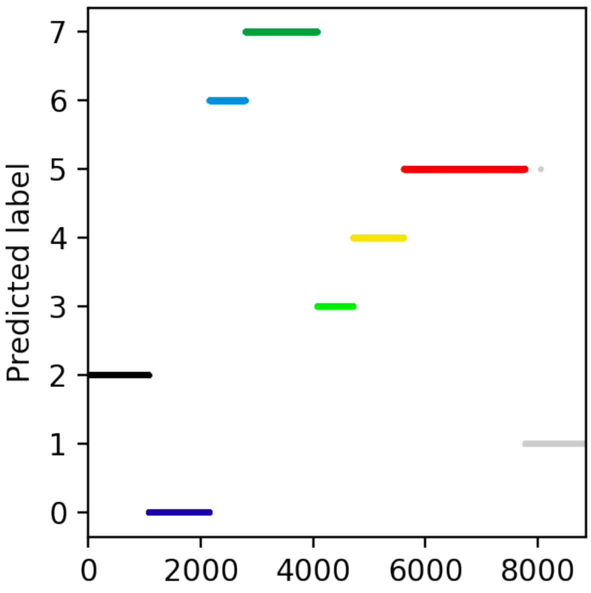


FIG. S13: Labels attributed by k-means using the space shown in Figure S12

VI. CLUSTERING WITH DBSCAN

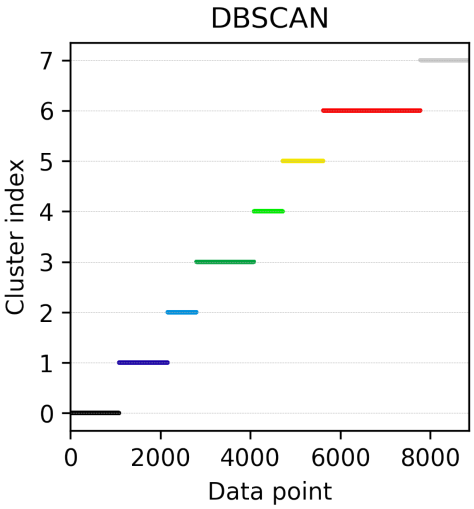


FIG. S14: Labels by DBSCAN on the projected data in the low-dimensional space. The color reflect the phase of the data point.

VII. PHASE TRANSITION USING THE TRAJECTORY AS THE TRAINING SET

Although we have shown that the 7-dimensional TS-LPP model can properly track the phase transition between β-quartz to α-quartz, an alternative way to study this specific transition is to use the trajectory itself as the training set and train the unsupervised model based on its snapshots. For instance, the difference between initial and final structures can be found with ease by using an early and a late snapshot in the training set, as shown in Figure S15.

Another method is to use part of the trajectory as the training set, for instance the first 50 fs, 100 fs, etc., as shown in Figure S16. If can be seen that in the very early stage (50-200 fs), the structure evolves and moves away from its initial configuration, then remains in a relatively stable region from 500 fs to 1,000 fs, and finally evolves to its final state after 3,000 fs.

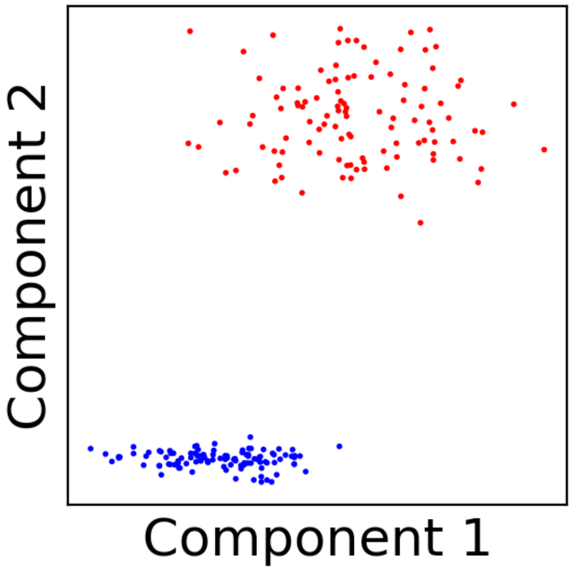
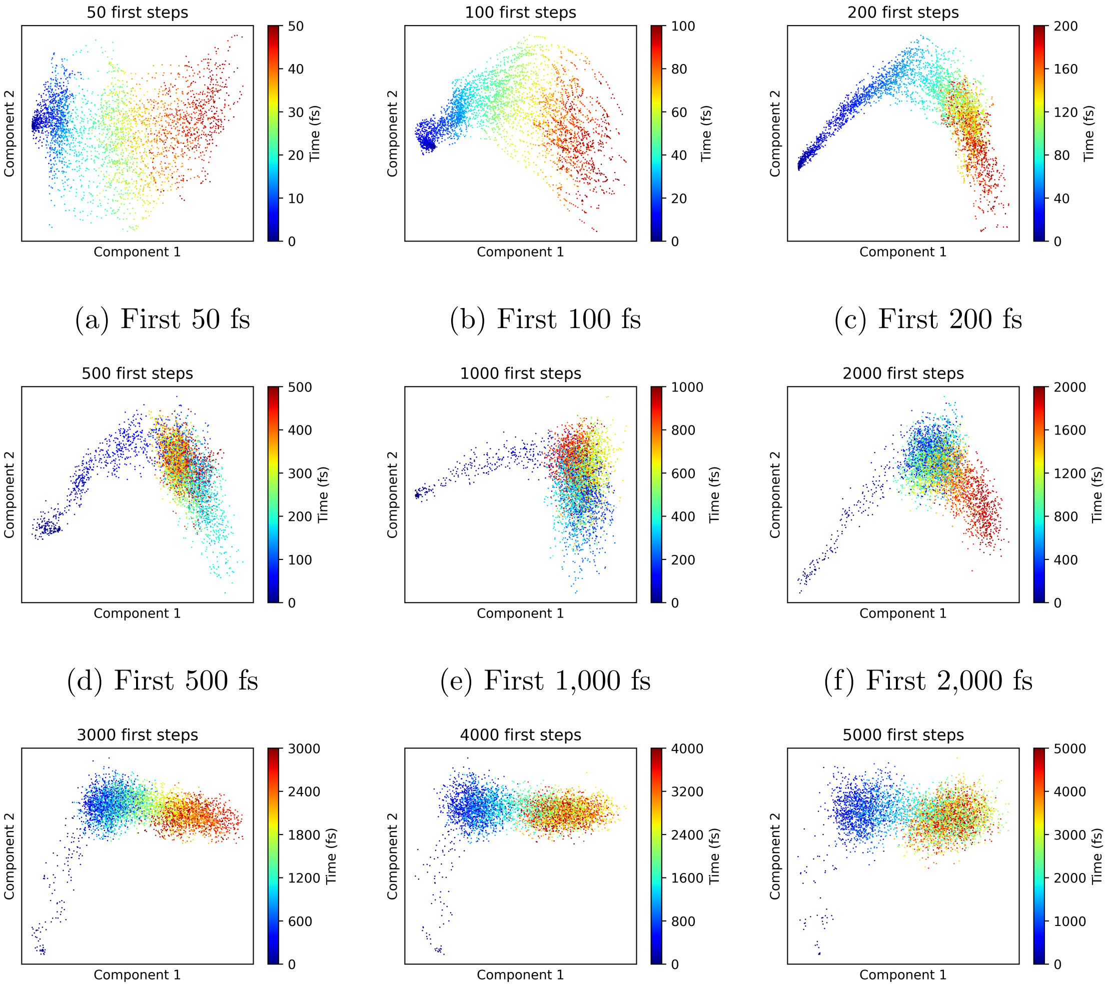


FIG. S15: Projected data of an early snapshot (50 fs) and late snapshot (5,000 fs) used as the training set.



(a) First 50 fs

(b) First 100 fs

(c) First 200 fs

(d) First 500 fs

(e) First 1,000 fs

(f) First 2,000 fs

(g) First 3,000 fs

(h) First 4,000 fs

(i) First 5,000 fs

FIG. S16: Transition from β-quartz to α-quartz at 600 K followed by a TS-LPP space trained on a single trajectory.

[1] C. V. Raman and T. M. K. Nedungadi, *Nature*, 1940, **145**, 147.

[2] A. Nakata, J. S. Baker, S. Y. Mujahed, J. T. L. Poulton, S. Arapan, J. Lin, Z. Raza, S. Yadav, L. Truflandier, T. Miyazaki and D. R. Bowler, *The Journal of Chemical Physics*, 2020, **152**, 164112.

[3] R. Tamura, J. Lin and T. Miyazaki, *J. Phys. Soc. Jpn.*, 2019, **88**, 044601.