Towards Structural Reconstruction from X-Ray Spectra: Supplementary Information

Anton Vladyka,[1,*] Christoph J. Sahle,[2,†] and Johannes Niskanen[1,‡]

[1] *University of Turku, Department of Physics and Astronomy, 20014 Turun yliopisto, Finland*
[2] *European Synchrotron Radiation Source, 71 Avenue des Martyrs, 38000 Grenoble, France*

(Dated: January 19, 2023)

Optimal features configuration

To find the optimal number of Ge and O atoms for the best prediction of the spectral moments from the Coulomb matrices, we performed the grid search for each set containing of 3–15 Ge atoms and 3–20 O atoms, and for each MLP configuration with 2-3 hidden layers and 64 and 128 neurons in each layer, as presented in Figure S1. Panels a) and b) show mean MSE and mean RMSE for the predicted spectral moments. Panel c) shows the geometric mean of the pairwise correlations between known and predicted spectral moments. Panel d) depicts the configuration of the neural network with the best prediction for the given Ge+O set.

For our system, the deduced optimal model contains 10 closest Ge atoms and 7 closest O atoms.

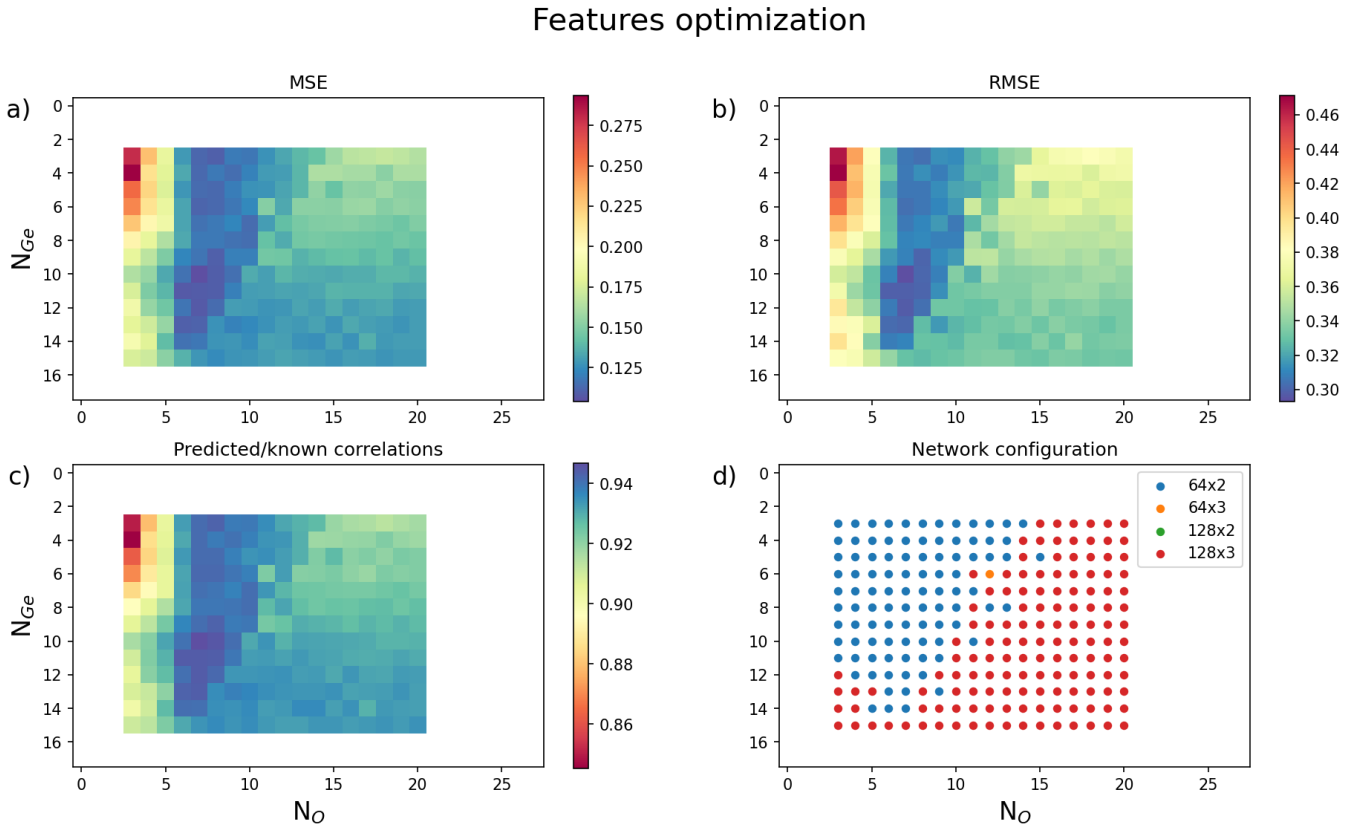

FIG. S1: Results of the features search.

Comparison of spectral moments prediction and full spectrum prediction

We compared the prediction of the spectral moments and full spectrum for the same configuration of hidden layers in the neural network, the same structural features set configuration, and the same train/evaluation data splitting, using MSE of standardized spectral features as a metric. Additionally, we performed the comparison with a prediction of the spectrum within Kβ''–Kβ_2 peaks energy range.

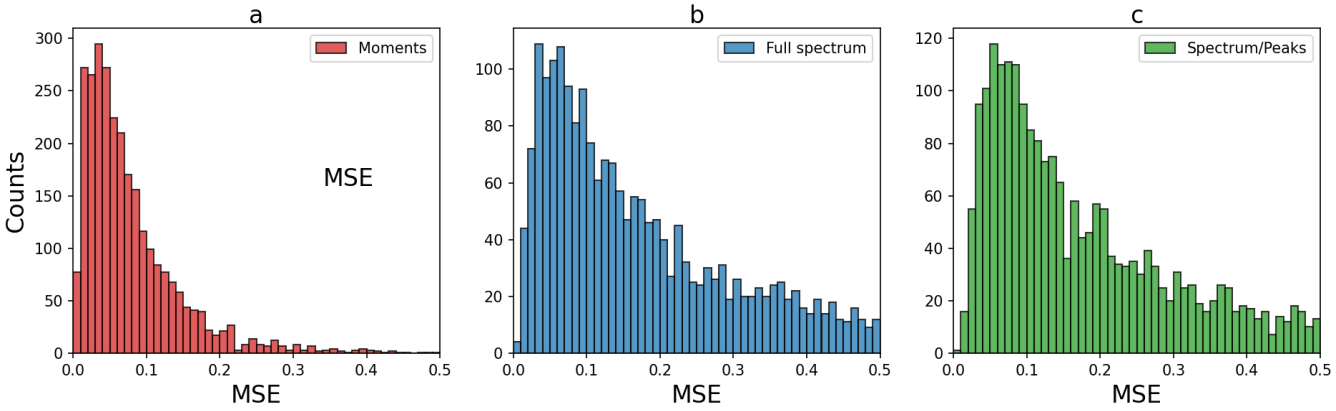

FIG. S2: Distribution of MSE values for a) 8 spectral moments prediction; b) full spectrum prediction (500 pts); c) spectrum within Kβ''–Kβ_2 peaks range prediction (3.5 eV–28 eV, 245 pts).

Sample ECA components

Every ECA component $\tilde{\mathbf{v}}$ is 153-dimensional vector in a space of normalized Coulomb matrices. Each component of ECA vector can be interpreted in terms of the changes in Coulomb matrix, which are attributed to the changes in the structure.

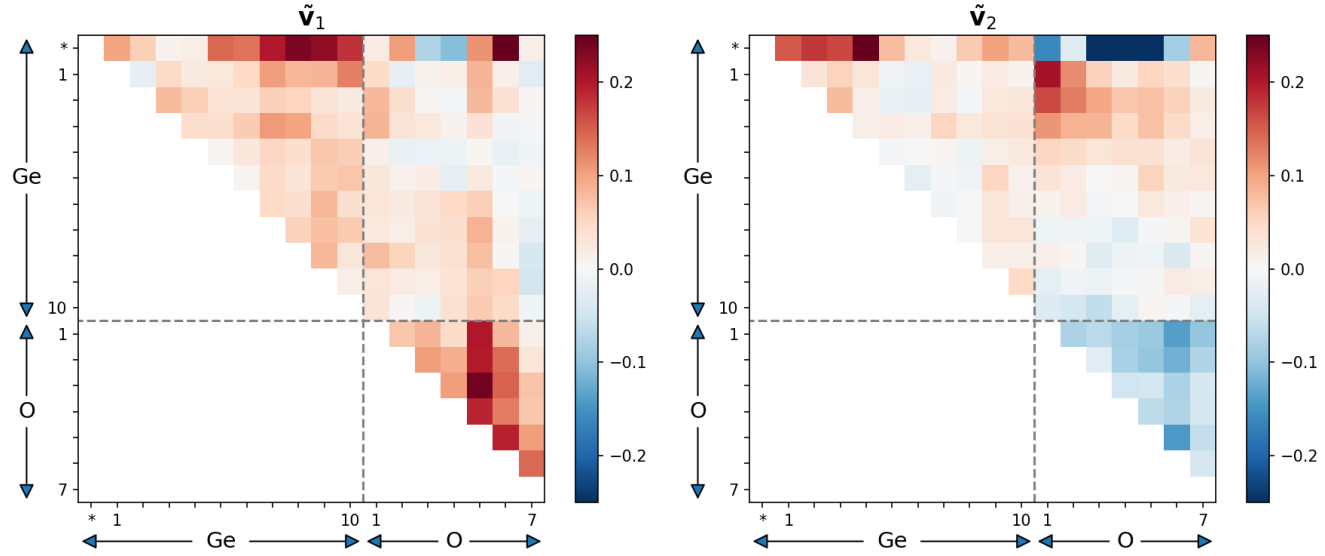

FIG. S3: First two ECA component vectors $\tilde{\mathbf{v}}_1$ and $\tilde{\mathbf{v}}_2$.

Spectral moments along first component vectors

We evaluated the prediction of the spectral moments along the first ECA/PLS component vectors. For each t_1 value between -7 and $+7$, we calculated the normalized parameter space coordinates as $\tilde{\mathrm{p}} = t_1 \tilde{\mathbf{v}}_1$ and $\tilde{\mathrm{p}} = t_1 \tilde{\mathbf{v}}_1^{(\mathrm{PLS})}$. The corresponding moments are deduced using model prediction and denormalization. These curves are shown as blue and red lines in Figure S4. For each pressure, we also calculated the projection of the average parameter coordinate on ECA/PLS component vector, and average spectral moment. These (t_1, m_i) pairs are shown as colored markers.

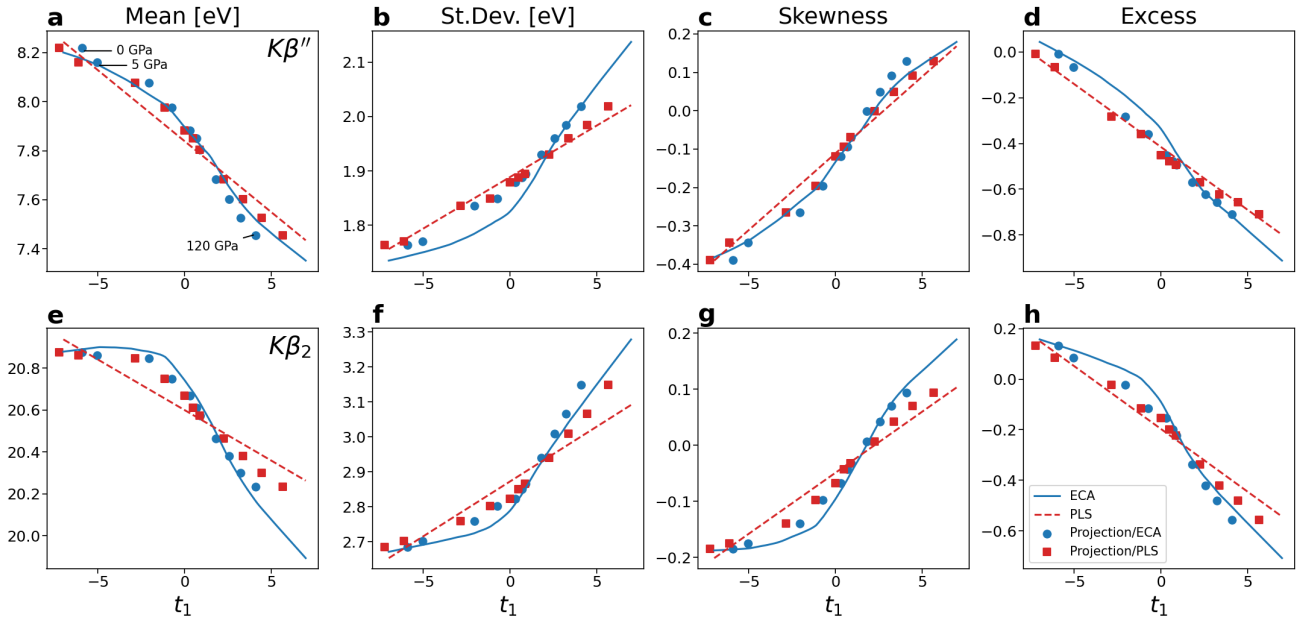

FIG. S4: Spectral moment predictions along the first decomposition components with the score t_1 range $[-7, 7]$ to cover the known test data spread in the parameter space. Nonlinearity of the emulator makes ECA a nonlinear approach accounting for more spectral variance than PLSSVD. Blue and red markers represent the projections of every pressure mean on ECA and PLS first component vectors, respectively. The directions of the component vectors have been selected for the highest pressure to yield the highest score; in this case the pressure increases monotonously with the projection score.

Projected distances from the central Ge atom to the nearest Ge and O atoms along first component vectors

We analyzed how interatomic distances are changing along the component vectors for ECA and PLS approaches. Such analysis includes the following steps:

1) For each t_1 value between -6 and 6 we calculated the corresponding configuration in the normalized parameter space: $\tilde{p} = t_1 \cdot \tilde{\mathbf{v}}_1$ (or $\tilde{\mathbf{v}}_1^{(\mathrm{PLS})}$ for PLS).

2) Obtained 153-dimensional vector was inverse-standardized to deduce unrolled Coulomb matrix elements.

3) Coulomb matrix is converted into distance matrix, and its first row represents the distances from the central Ge atom (Ge$*$) to all other atoms used for Coulomb matrix calculation.

Obtained $d(t_1)$ dependencies are presented in Figure S5.

Also for each pressure, the projection of the mean scaled coulomb matrix on first component ECA/PLS vectors were calculated and presented as scatter plot in the same figure.

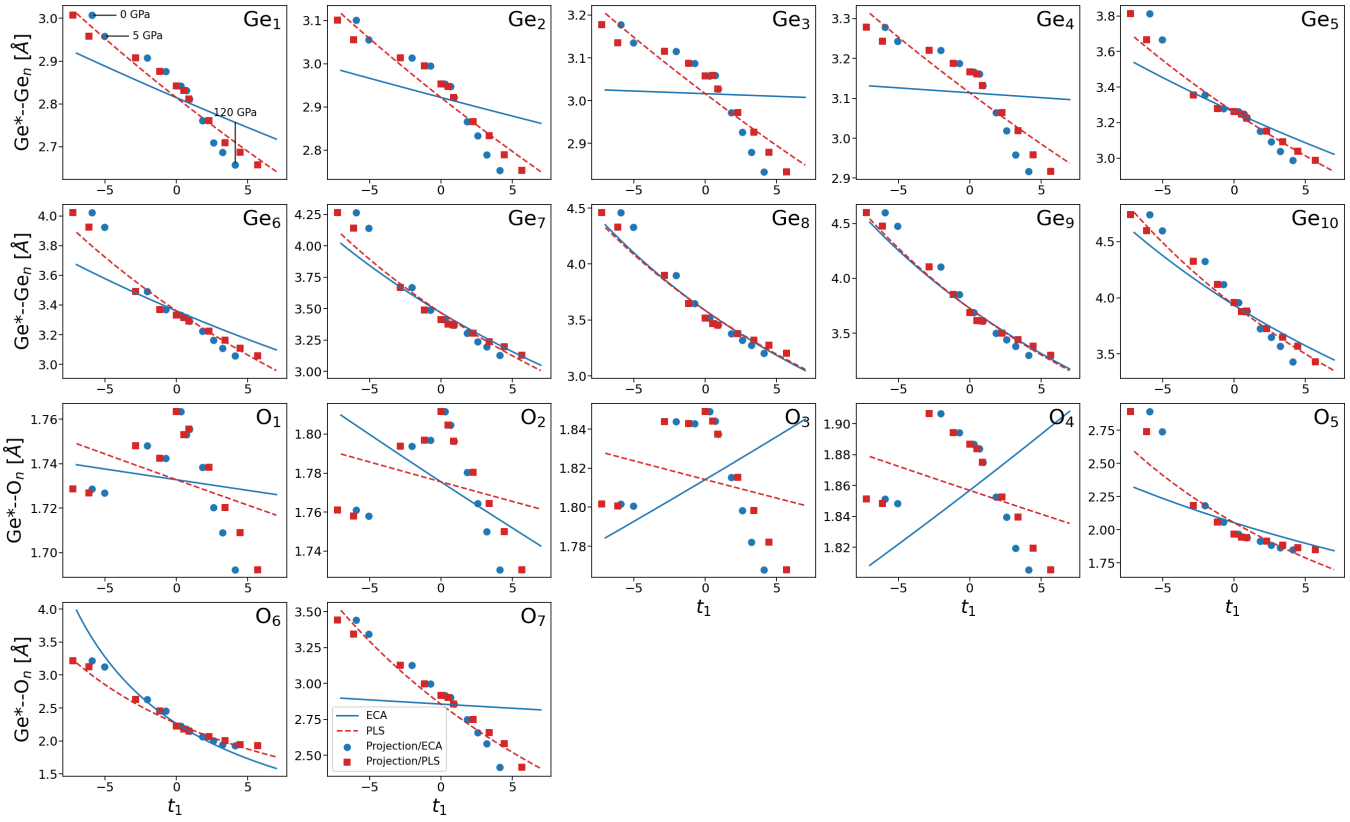

FIG. S5: Distances from the central Ge atom to the closest 10 Ge and 7 O atoms used for the calculation of the Coulomb matrix, deduced along the first ECA and first PLS component directions. Blue lines depict the 'predicted' positions of the atoms projected on the first ECA component, red dashed lines indicated the first PLS component. The projections of actual average distances for each pressure are shown as blue and red markers for ECA and PLS, respectively.

Reconstruction of the spectral moments

Reconstruction of the spectral moments from (t_1, t_2) projections in ECA space

Figure S6 shows the predicted spectral moments as a function in two-dimensional ECA parameter space. Every point (t_1, t_2) on each of the 8 panels depicts the predicted spectral moment for the structure with $\tilde{\mathbf{p}} = t_1\tilde{\mathbf{v}}_1 + t_2\tilde{\mathbf{v}}_2$.

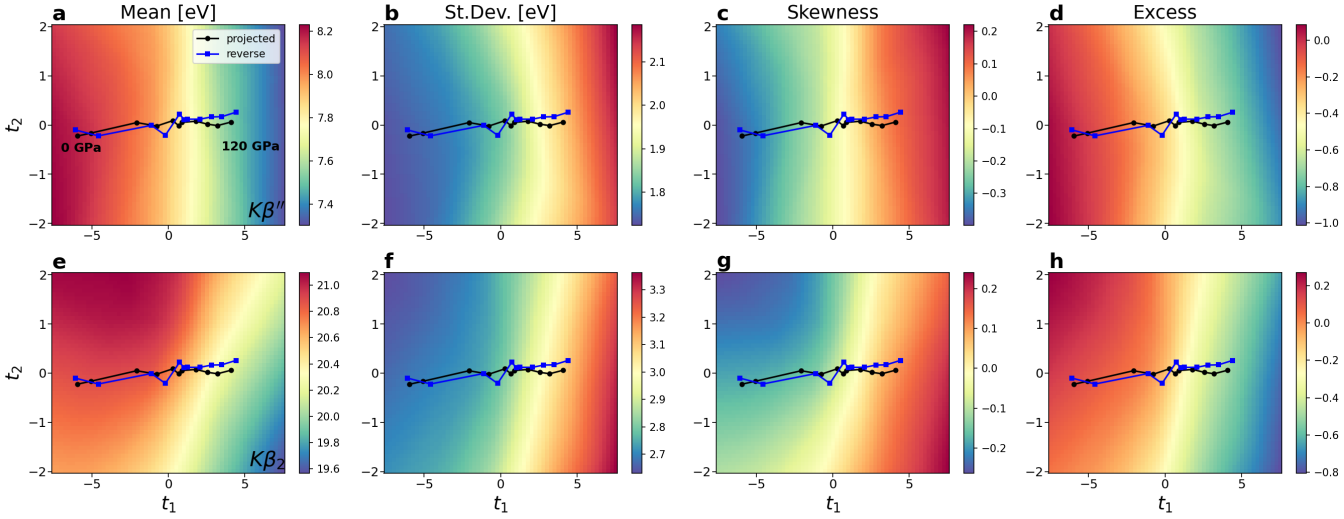

FIG. S6: Predicted spectral moments in 2-dimensional ECA space. Each point on the plot represents a moment predicted by the model for the $\tilde{\mathbf{p}} = t_1\tilde{\mathbf{v}}_1 + t_2\tilde{\mathbf{v}}_2$. Black markers represent the projected means parameters for each pressure, and blue markers depict the reconstructed projections from the mean moments.

Reconstruction of the spectral moments from (t_1, t_2) PLS coordinates

Figure S7 shows the predicted spectral moments as a function in two-dimensional PLS parameter space. Every point (t_1, t_2) on each of the 8 panels depicts the predicted spectral moment for the structure with $\tilde{\mathbf{p}} = t_1\tilde{\mathbf{v}}_1^{(\mathrm{PLS})} + t_2\tilde{\mathbf{v}}_2^{(\mathrm{PLS})}$.

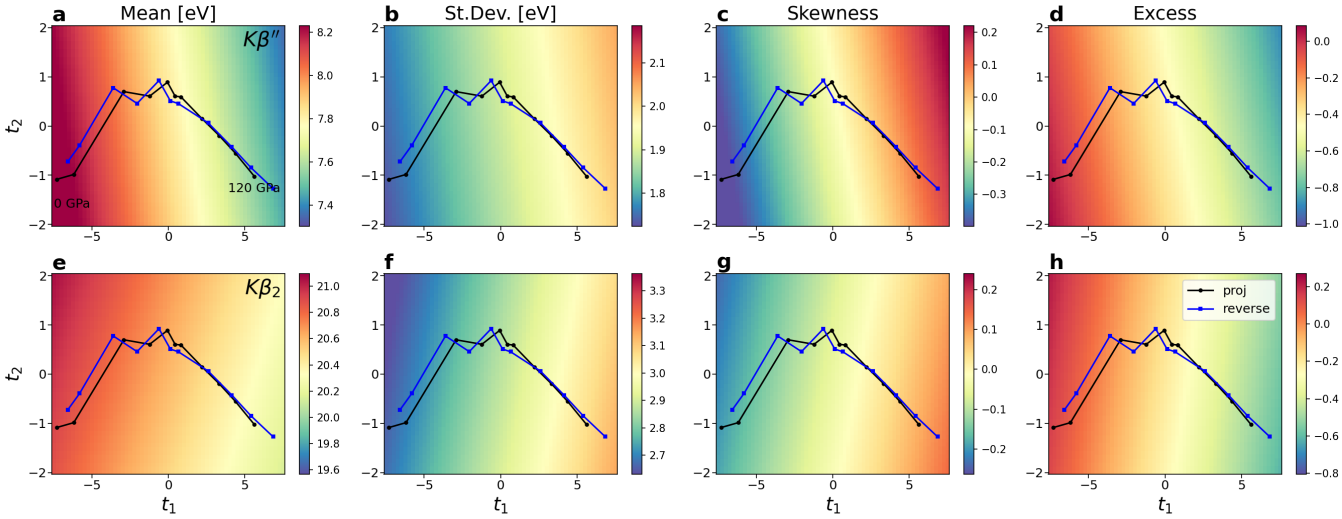

FIG. S7: Predicted spectral moments in 2-dimensional PLS space. Each point on the plot represents a moment predicted by the model for the $\tilde{\mathbf{p}} = t_1\tilde{\mathbf{v}}_1^{(\mathrm{PLS})} + t_2\tilde{\mathbf{v}}_2^{(\mathrm{PLS})}$. Black markers represent the projected means parameters for each pressure, and blue markers depict the inversely calculated projections from the mean moments.

Figure S8 shows the reconstructed (t_1, t_2) coordinates from spectral moments using PLS approach, for each point from evaluation data set.

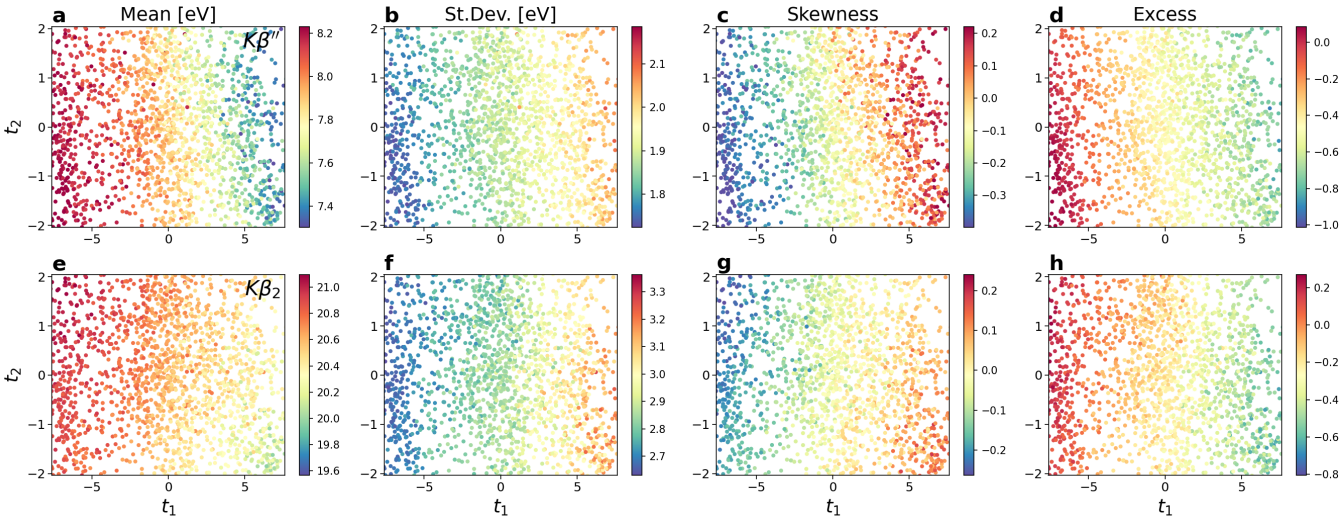

FIG. S8: Reconstructed t_1, t_2 coordinates for all evaluation data points using PLS approach. Color scales are identical for each corresponding panel in the Figure S7.

Reconstruction of the interatomic distances as a function of pressure using PLS projections

Analysis of interatomic distances for the PLSSVD was done using the same approach as for Figure 5 im the main text.

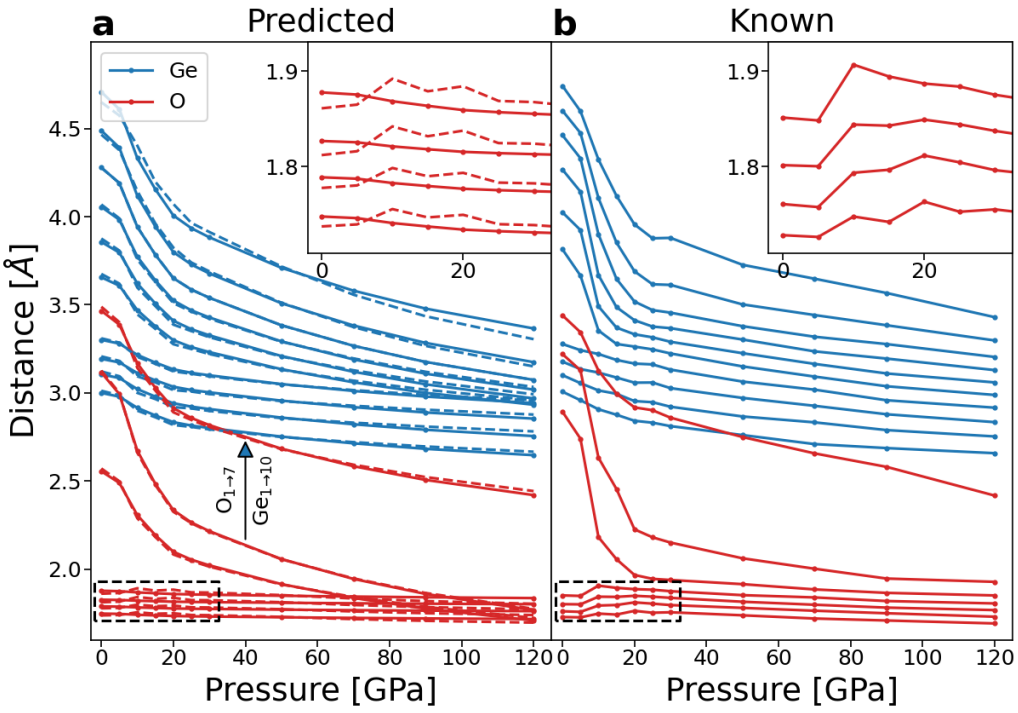

FIG. S9: (a) Reconstructed mean-structural-parameter-based distances from the central atom for $t_1^{(\mathrm{PLS})}$ component. The inset shows distances for the 4 closest O atoms (dashed area). Dashed lines depict corresponding distances for sum of two projections (t_1, t_2). (b) Known mean-ensemble distances.

Predicted atomic distances from ECA projections

To evaluate the validity of predicted interatomic distances, we compared predicted curves with known values.

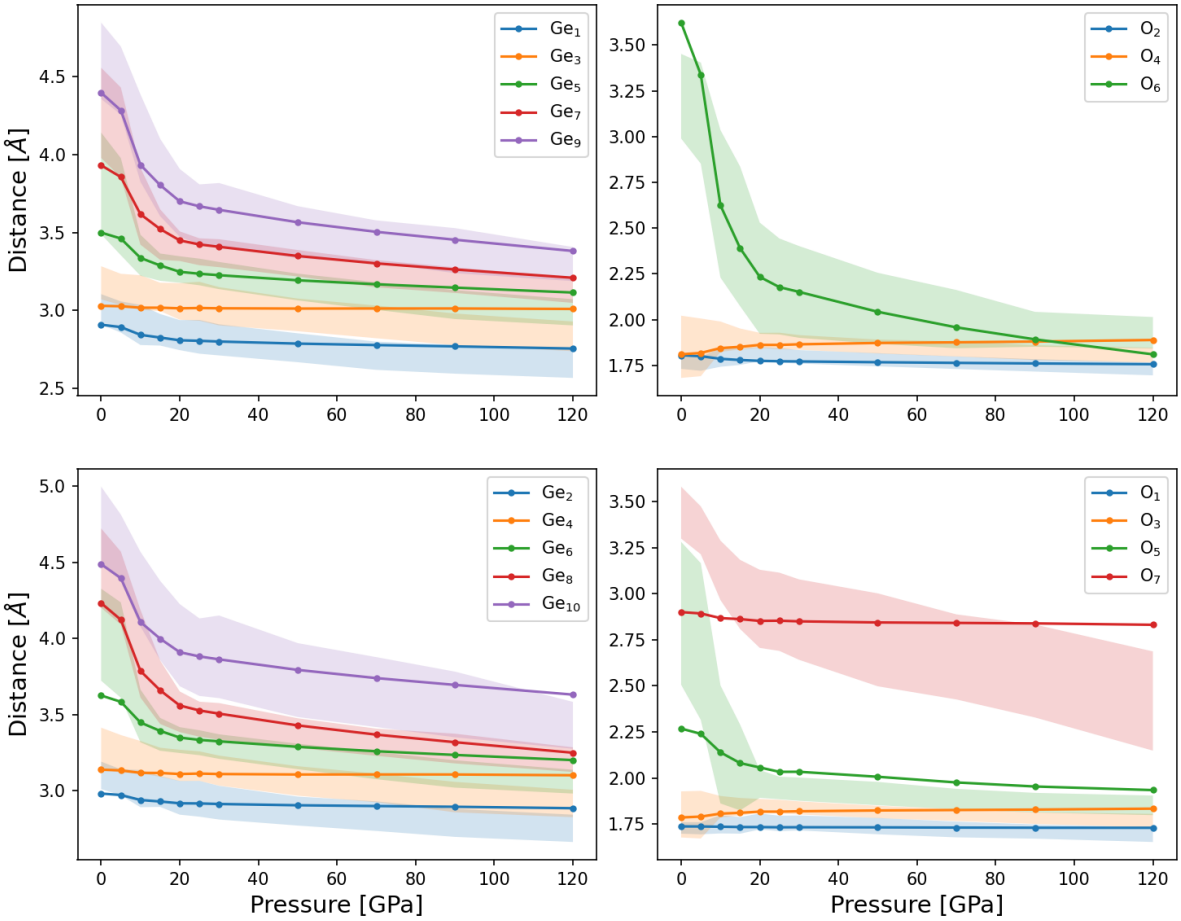

FIG. S10: Distances from the central Ge atom to the closest 10 germanium atoms and 7 oxygen atoms. Shaded areas: mean±std of the known distances for each pressure. Lines: projected distance for each mean spectral moment per pressure.

Reconstruction of structure from the Coulomb matrix

We start indexing of atoms from 0 and denote the cartesian coordinates of atomic site n by (x_n, y_n, z_n). The original snapshot can be recovered from the distance matrix $\mathbf{R}$ (and from the Coulomb matrix $\mathbf{C}$) by using the following algorithm

1. Place the active site 0 at origin
2. Place site 1 at $(R_{01}, 0, 0)$
3. Place site 2 at $(R_{02}\cos\gamma, R_{02}\sin\gamma, 0)$ (in xy-plane in positive direction) where
$$\gamma \;=\; \arccos\frac{R_{01}^2 + R_{02}^2 - R_{12}^2}{2R_{01}R_{02}} \tag{1}$$
4. Place site 3 in positive z (negative z for the alternate handedness) so that R_{03}, R_{13}, R_{23} are fulfilled. This corresponds to solving
$$-2\begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{bmatrix}\begin{bmatrix} x_3 \\ y_3 \\ z_3 \end{bmatrix} = \begin{bmatrix} (R_{13}^2 - R_{03}^2 - R_{01}^2) \\ (R_{23}^2 - R_{03}^2 - R_{02}^2) \end{bmatrix} \tag{2}$$
by the Moore-Penrose pseudoinverse to obtain $(x_3, y_3, 0)$. The component z_3 is then obtained as
$$z_3 = \sqrt{R_{03}^2 - x_3^2 - y_3^2}, \tag{3}$$
where the requirement for positive z_3 (negative z_3 for the alternate handedness) has been used to obtain the unique solution.
5. Place each remaining site n so that $R_{0n}, R_{1n}, R_{2n}, R_{3n}$ are fulfilled. This corresponds to solving
$$-2\begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{bmatrix}\begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix} = \begin{bmatrix} (R_{1n}^2 - R_{0n}^2 - R_{01}^2) \\ (R_{2n}^2 - R_{0n}^2 - R_{02}^2) \\ (R_{3n}^2 - R_{0n}^2 - R_{03}^2) \end{bmatrix} \tag{4}$$
with respect to (x_n, y_n, z_n) which are the coordinates of atom n.

Steps 1-4 fix the orientation and handedness of the coordinate system. An analysis of all structural data confirms, that apart from this handedness, each constructed structure matches with the original one after a suitable rotation (maximal location deviation of the order 10^{-8} Å). This also means that only the first 4 rows of the upper triangle of the symmetric Coulomb matrix describe the geometry completely apart from the aforementioned handedness-based symmetry. A code of the reconstruction algorithm is given with the data and other scripts.

* Electronic address: anton.vladyka@utu.fi
† Electronic address: christoph.sahle@esrf.fr
‡ Electronic address: johannes.niskanen@utu.fi