Supplementary Information

Co-orchestration of Multiple Instruments to Uncover Structure-Property Relationships in Combinatorial Libraries

Boris N. Slautin, Utkarsh Pratiush, Ilia N. Ivanov, Yongtao Liu, Rohit Pant, Xiaohang Zhang, Ichiro Takeuchi, Maxim A. Ziatdinov, and Sergei V. Kalinin**

1. Multi-task GP with Linear Model of Coregionalization

The multi-task Gaussian Processes form the foundation for multi-task Bayesian optimization and therefore for multimodal automated experiments as well.[1–3] MTGP extends the concept of GP to handle multiple related tasks, representing distinct objectives or output variables, simultaneously. The model uncovers the dependencies and correlations between tasks, allowing them to leverage information from one task to improve predictions in another. The overall covariance between the outputs in the most general case can be defined as:[3]

$$K[f_a(x),f_b(x')] = K^f_{ab}k(x,x'), \tag{1}$$

where the K^f_{ab} specifies the inter-task similarities and $k(x,x')$ is the covariance over inputs. In practice, the Linear Model of Coregionalization (LMC) is often used to capture the correlation between multiple outputs in MTGP. LMC introduces the several (Q) shared latent processes with their coregularization matrixes, denoted as $B^{(q)}$, to establish a linear relationship between multiple tasks. In the LMC-based multi-task GP, the covariance between the i-th data point of z_a and the j-th data point of the z_b can be formulated as follows:

$$K\left[z_a(x_i),z_b(x_j)\right] = \sum_{q=1}^{Q} B^{(q)}_{ab}k_q(x_i,x_j) \tag{2}$$

.

In the expression above the Q is the number of the latent processes, $B^{(q)}_{ab}$ represents an element of the coregularization matrix for q-th latent process, $k_q(x_i,x_j)$ is the covariance function for q-th latent process. The non-diagonal elements of the $B^{(q)}$ are responsible for the multi-task correlations. To ensure the symmetric and positive semi-definiteness of $B^{(q)}$ it is parameterized as:

$$B^{(q)} = W^{(q)}(W^{(q)})^T + diag(v^{(q)}) \tag{3}$$

where $W^{(q)}$ is low-rank $D \times R$ matrix, D number of tasks (in our case $D = 2$), R is rank, $diag(v^{(q)})$ – encapsulating specific variances for each output. The elements of $W^{(q)}$, and $v^{(q)}$ are estimated directly throughout the learning of GP.

1 K. Swersky, J. Snoek, R. P. Adams, Multi-Task Bayesian Optimization, in *Advances in Neural Information Processing Systems*, **26** (Eds. C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger), Curran Associates, Inc., 2013**.**

2 X. Wang, Y. Jin, S. Schmitt, M. Olhofer, *ACM Computing Surveys*, 2023, **55**, 287.

3 E. V. Bonilla, K. Chai, K. Williams, Multi-Task Gaussian Process Prediction, in *Advances in Neural Information Processing Systems*, **20** (Eds. J. Platt, D. Koller, Y. Singer, S. Roweis), Curran Associates, Inc., 2007.

2. Radial Basis Function

The Radial Basis Function (RBF) is the covariance functions in GP regression defined as:

$$k(x_i,x_j) = exp(-\frac{1}{2l^2}\|x_i - x_j\|^2) \quad (4)$$

where x_i,x_j are the point in input space, $\|x_i - x_j\|$ represents the Euclidean distance between the x_i and x_j, and l is a kernel length, which controls the smoothness of the function. Larger values of l lead to smother function.

3. Raman spectra treatment

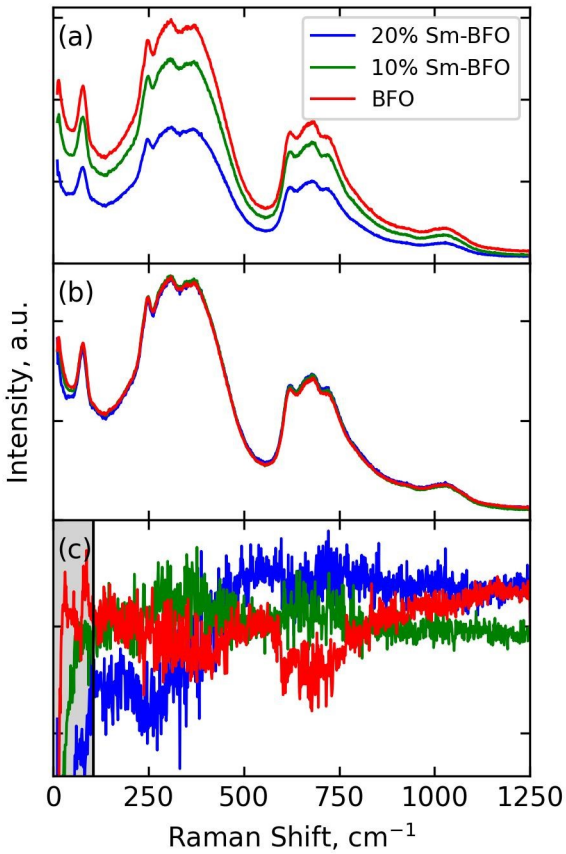

Fig. S1 The multi-step preprocessing of Raman spectra: (a) raw spectra, (b) spectra after normalization by the area under the curve, (c) footprint spectra, defining as a difference between local spectra and the mean spectrum across the entire dataset. The frequency range highlighted by the grey strip, was removed to diminish the influence of the residual Relay peak wing.

4. Reconstruction of the raw spectra using LVAE model

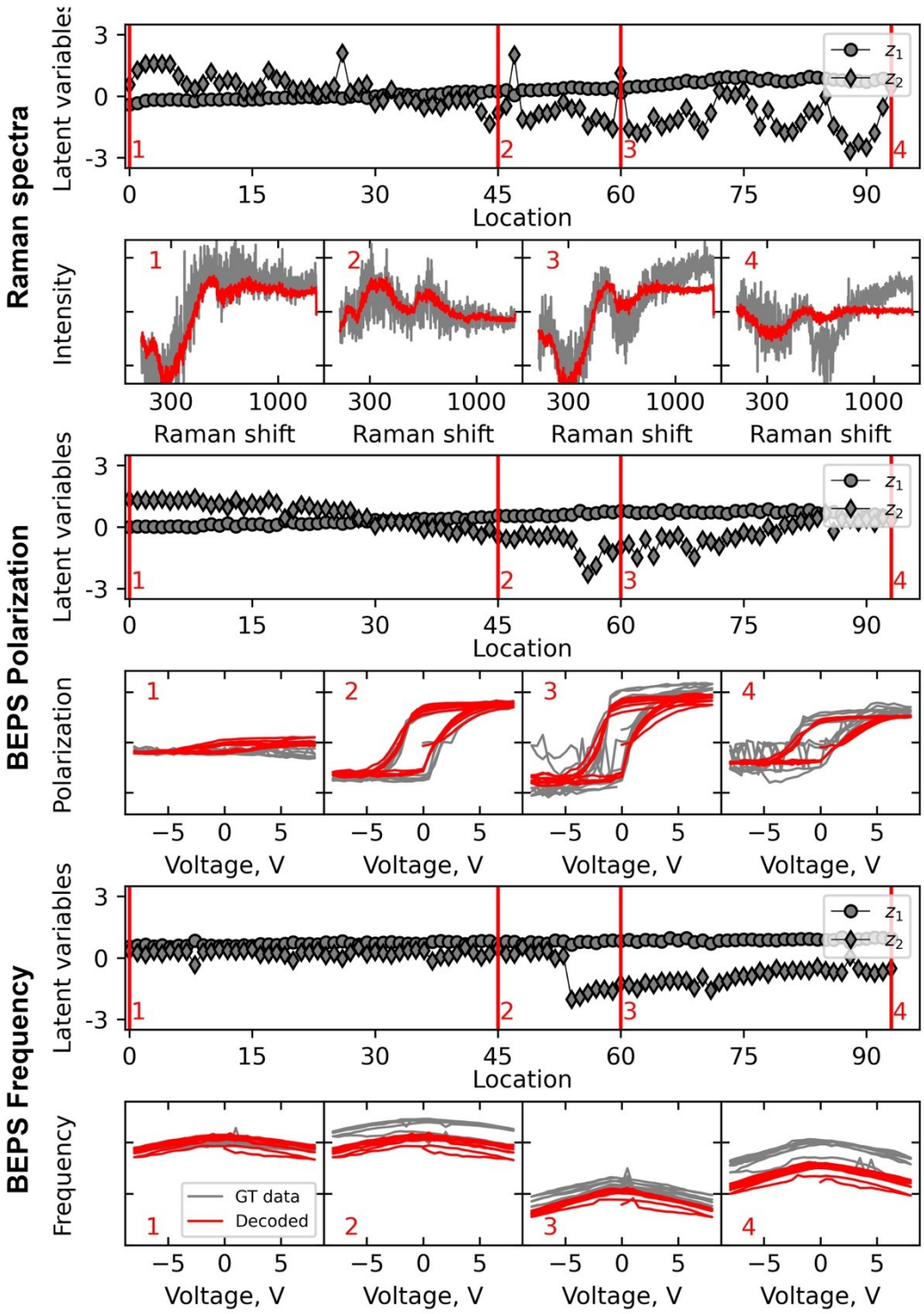

Fig S2 Reconstructions of the raw spectra using LVAE for various locations within a combinatorial library. The selected locations are emphasized by red vertical lines. Ground truth (GT) spectra are highlighted in a grey shade, while the LVAE-decoded spectra are depicted in red.

5. VAE latent distribution and latent variable compositional profiles at the different exploration steps

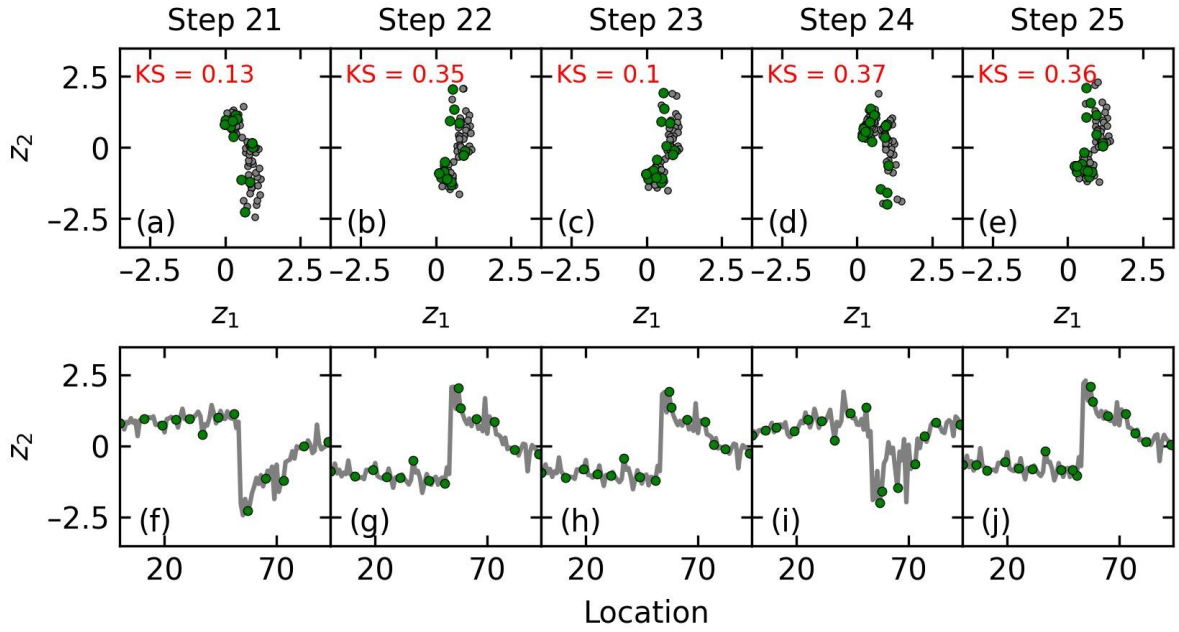

Figure S3. (a-e) VAE latent distributions and (f-j) corresponding compositional profiles of z_2 latent variables of the Modality 0 (BEPS frequency) for subsequent exploration steps in the second experiment. The KS criteria values in the (a-e) are calculated for pairs – latent distribution in this step and latent distribution in the previous step. The reflections of the distribution in steps 22, 23, and 25 yield high KS values, whereas in step 23, where the distribution orientation is preserved, the KS criterion is minimized.