Machine Learning-enabled cartography of complex chemical spaces into distinct regimes

Yi Wei Lee[a], **Qian Xiao Li**[b], **Saif A. Khan**[c]

[a] *NUS Graduate School for Integrative Sciences & Engineering, Singapore 119077* e0823047@u.nus.edu

[b] *Department of Mathematics & I-FIM, National University of Singapore, Singapore 119076* qianxiao@nus.edu.sg,

[c] *Department of Chemical and Biomolecular Engineering, Singapore 117576* saifkhan@nus.edu.sg,

1. Introduction

Regime maps are handy tools in various fields of natural science, including fluid dynamics, plasma physics, and material science. For example, the Hertzsprung-Russell diagram in astrophysics is used for stellar classification [1], while two-phase flow regime maps are commonly employed in oil and gas pipeline design [2]. In material science, regime maps such as phase diagrams [3] and deformation mechanism maps [4] play a crucial role in material design, analysis, and optimization. Regime maps provide an intuitive representation of complex systems by identifying and classifying similar behavior classes and relating those classes to input variables. Once a regime map is created, scientists and engineers can use it for system prediction, process optimization, or simply as a visualization tool for further analysis of the underlying system. These maps are foundational for advancements in science and engineering.

Despite the numerous advantages of regime maps, their application in studying chemical/material systems is still somewhat limited due to two main challenges. First, creating a regime map requires grouping the output variables of a system into classes of similar behavior, which can be difficult when dealing with continuous and high-dimensional output variables. Typically, regime maps are based on discrete classes of behavior (such as phase states in phase diagrams or different deformation modes in deformation mechanism maps), which are easily identifiable and classifiable. However, in many cases, such discrete behavior classes are not available for the systems being studied, making it challenging to create a regime map. Secondly, regime maps generally focus on a limited number of input variables (typically two) because of the large volume of data points needed to create an accurate regime map in high dimensions. This issue is known as the "curse of dimensionality"; it restricts regime map implementation in various systems, as multiple input variables or parameters influence most systems. Although some regime plots, particularly in fluid dynamics, reduce the number of input variables by using dimensionless numbers, this approach requires prior knowledge that may not be easily accessible in many systems under investigation.

To address these challenges, we utilized the power of machine learning to develop a generalizable regime identification framework that enables the creation of regime maps based on continuous output variables within a multidimensional input space. Our approach combines unsupervised learning with active learning to 1) identify the natural behavior classes in the system under investigation and 2) perform efficient data sampling within the input space to create a reasonably accurate multidimensional regime map. We tested our framework experimentally on a silver nanoparticle (AgNP) synthesis platform and successfully created a regime map using four input variables without requiring prior knowledge of the system.

2. Regime identification framework

The proposed framework consists of three main components: initial sampling, clustering, and active learning with neural network ensembles, in that order. The final product is an ensemble of neural networks that encodes the relationships between the identified regimes and the input variables under investigation.

2.1 Initial sampling

For a system under investigation, let $\mathbf{x} \in X = \mathbb{R}^n$ and $\mathbf{y} \in Y \subseteq \mathbb{R}^m$ be the input and output vectors, respectively, where X and Y are Euclidean spaces. In the initial sampling step, we aim to sample enough $\mathbf{x}$ within X to provide sufficient coverage for the range of $\mathbf{y}$ in Y. To achieve this, we first employ a sequential space-filling design [5] coupled with a local outlier measurement metric [6] to select the initial dataset for the framework.

2.2 Clustering

Once the initial sampling process has been completed, we utilize spectral clustering coupled with a membership function to group the output vectors ($\mathbf{y}$) into similar behavior classes, where each class represents a regime. A soft clustering algorithm is chosen as the regimes typically do not have sharp boundaries; hence, such an approach can highlight the transitional regions between the regimes within X.

2.3 Active learning with neural network ensembles

With the regime identified, an ensemble of neural networks is then used as a surrogate function to relate the input space X to the identified regime. However, as the number of points used to train the ensemble is insufficient, an active learning loop is employed to sample additional points efficiently from the system by targeting locations within the input space where the information gain is predicted to be high.

3. Application on AgNP synthesis platform

The proposed framework is applied to an AgNP synthesis platform. The platform synthesizes AgNP through a flow-based milifluidic device, creating aqueous reactant mixture droplets in silicone oil. The reactants used to synthesize the AgNP were ascorbic acid (AA), silver nitrate ($AgNO_3$), trisodium citrate (TSC), and polyvinyl alcohol (PVA). The concentration of the reactants within each droplet is controlled by the flow rate of the individual reactants (Q_{AA}, Q_{AgNO3}, Q_{TSC}, and Q_{PVA}). The droplets generated were subjected to a hyperspectral imaging system [7] to obtain the absorbance map of the synthesis process. The absorbance map shows the evolution of the absorbance spectrum of the reactant mixture droplet

Machine Learning-enabled cartography of complex chemical spaces into distinct regimes

Yi Wei Lee[a], **Qian Xiao Li**[b], **Saif A. Khan**[c]

[a] *NUS Graduate School for Integrative Sciences & Engineering, Singapore 119077* e0823047@u.nus.edu

[b] *Department of Mathematics & I-FIM, National University of Singapore, Singapore 119076* qianxiao@nus.edu.sg,

[c] *Department of Chemical and Biomolecular Engineering, Singapore 117576* saifkhan@nus.edu.sg,

throughout the AgNP synthesis process. An example of the absorbance map for a specific set of a specific flowrate combination is shown in Fig 1.

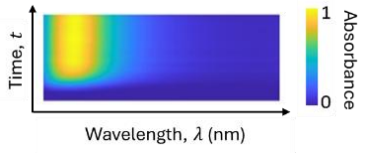

Fig 1: Example of an absorbance map.

3.2 Result

The four input variables investigated for the AgNP synthesis platform were the four reactants (Q_{AA}, Q_{AgNO3}, Q_{TSC}, and Q_{PVA}). The output vectors were derived from the absorbance map, where important features such as the induction time and the peak wavelength were extracted and concatenated. Through the application of the regime identification framework, six regimes were identified. A representative absorbance map for each regime is shown in Fig 2.

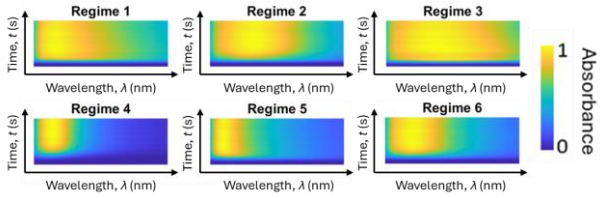

Fig 2. Representative absorbance map for each regime

The distinction between each representative absorbance map can be visually verified. To further quantify the distinction between each regime, the representative features - induction time (t_{ind}), maximum absorbance growth rate (G_{max}) and the peak wavelength (λ_{max}) for the absorbance map in each regime is presented in Table 1.

Table 1: Representative features of the absorbance maps in each regime

Regime	t_{ind} (s)	G_{max} (a.u./s)	λ_{max} (nm)
1	26.0	0.0294	456
2	24.4	0.0176	528
3	20.37	0.0356	557
4	43.9	0.0128	447
5	31.0	0.0246	426
6	29.5	0.0158	468

From Fig 2 and Table 1, some inferences regarding the growth profile can be generated. The λ_{max} values indicate that regime 5 exhibits the smallest final AgNP size, followed by regimes 4, 1, 6, 2, and 3, in increasing order. The t_{ind}, indicating the duration of slow and sustained nucleation, reveals that regime 4 had the most extended nucleation period, while regime 3 had the shortest. Based on the absorbance map and the feature extracted, we posit that the AgNP within regimes 4 and 5 undergoes diffusion growth due to their small final AgNP size. In contrast, the AgNP in regime 3 mainly undergoes agglomerative growth due to the large λ_{max} and large width of absorbance spectrum at the final time point. The other regimes mainly exhibit a mixture of both these growth profiles. Final confirmation of the growth profiles can be ascertained through TEM imaging of the final AgNP, which is still a work in progress.

Once the absorbance map regime has been established, the active learning loop is employed to perform cartography on the input space. After just one iteration of the active learning loop, the neural network ensemble achieves a 0.9 F1 score. In Fig 3, a 3D plot for each of the regimes within the input space is presented for different values of Q_{PVA}. The color for each regime is 1—red, 2—yellow, 3—green, 4—blue, 5—cyan, and 6—purple.

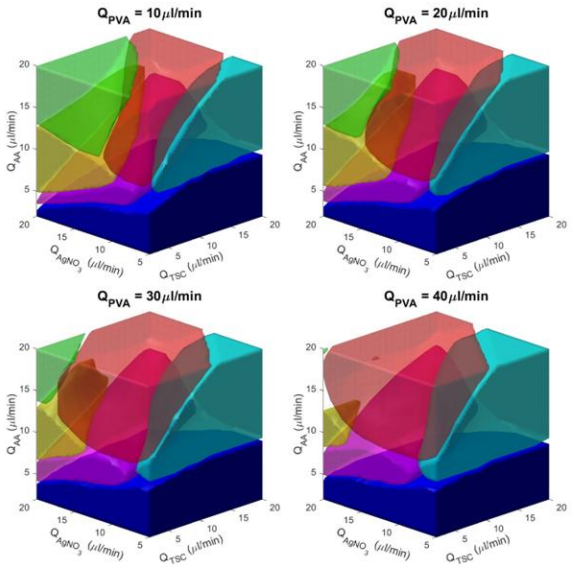

Fig 3: 3D regime map for different values of Q_{PVA}.

The 3D regime map illustrates the relationship between input flow rates and the observed AgNP synthesis regimes. A key observation is that increasing the flow rates of stabilizers (TSC and PVA) leads to a clear transition from agglomeration-dominated to growth-dominated regimes. Additionally, the flow rate of AA demonstrably affects the t_{ind} and G_{max} of the AgNP synthesis process. By analyzing these regime plots, we can draw a series of inferences based on the relationships defined by the regime transitions.

4. Conclusion

We proposed a regime identification framework that can identify the regimes available within the input space of a system under investigation. The regime map created by the framework can serve as a valuable tool for data analysis or as a groundwork for future research endeavors.

Machine Learning-enabled cartography of complex chemical spaces into distinct regimes

Yi Wei Lee[a], **Qian Xiao Li**[b], **Saif A. Khan**[c]

[a] *NUS Graduate School for Integrative Sciences & Engineering, Singapore 119077* e0823047@u.nus.edu

[b] *Department of Mathematics & I-FIM, National University of Singapore, Singapore 119076* qianxiao@nus.edu.sg,

[c] *Department of Chemical and Biomolecular Engineering, Singapore 117576* saifkhan@nus.edu.sg,

References

[1] A. V. Nielsen, "Contributions to the History of the Hertzsprung-Russell Diagram," *Centaurus*, vol. 9, no. 4, pp. 219-253, 1964.

[2] L. Cheng, G. Ribatski, and J. R. Thome, "Two-phase flow patterns and flow-pattern maps: fundamentals and applications," 2008.

[3] D. A. Young, *Phase diagrams of the elements*. Univ of California Press, 2023.

[4] M. F. Ashby, "A first report on deformation-mechanism maps," *Acta Metallurgica*, vol. 20, no. 7, pp. 887-897, 1972.

[5] B. Shang and D. W. Apley, "Fully-sequential space-filling design algorithms for computer experiments," *Journal of Quality Technology*, vol. 53, no. 2, pp. 173-196, 2021.

[6] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "LoOP: local outlier probabilities," in *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, pp. 1649-1652.

[7] F. Mekki-Berrada, J. Xie, and S. A. Khan, "High-throughput and High-speed Absorbance Measurements in Microfluidic Droplets using Hyperspectral Imaging," *Chemistry-Methods*, vol. 2, no. 5, p. e202100086, 2022.