Contributions and Challenges of Machine Learning in Materials Science

Felipe V. Calderan[1], Gabriel A. Pinheiro[1], Juarez L. F. Da Silva[2], Ronaldo C. Prati[3], Marcos G. Quiles[1]

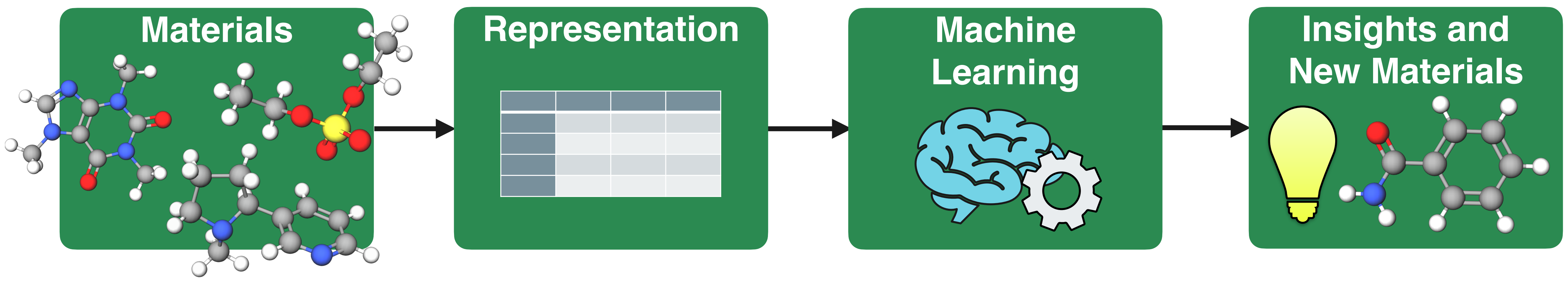
[1]Institute of Science and Technology, Federal University of São Paulo, São José dos Campos, SP, Brazil
[2]São Carlos Institute of Chemistry, University of São Paulo, São Carlos, SP, Brazil
[3]Center of Mathematics, Computation and Cognition, Federal University of ABC, Santo André, SP, Brazil

ABSTRACT

Material science plays an important role in society by developing novel materials across multiple sectors, such as biodegradable, medical, and renewable energy materials, to name a few. However, the ongoing race for new materials with specific properties, under time and cost constraints, requires tools that help solve these demands. In this sense, machine learning has driven advancements in several scientific fields, showcasing itself as a potential tool for these challenges. Here, we present our recent works aimed at improving the accuracy of material property predictions via such approach and general physicochemical analyses. Specifically, we have explored methods for data generation and collection, alongside the development of new machine learning frameworks.

INTRODUCTION

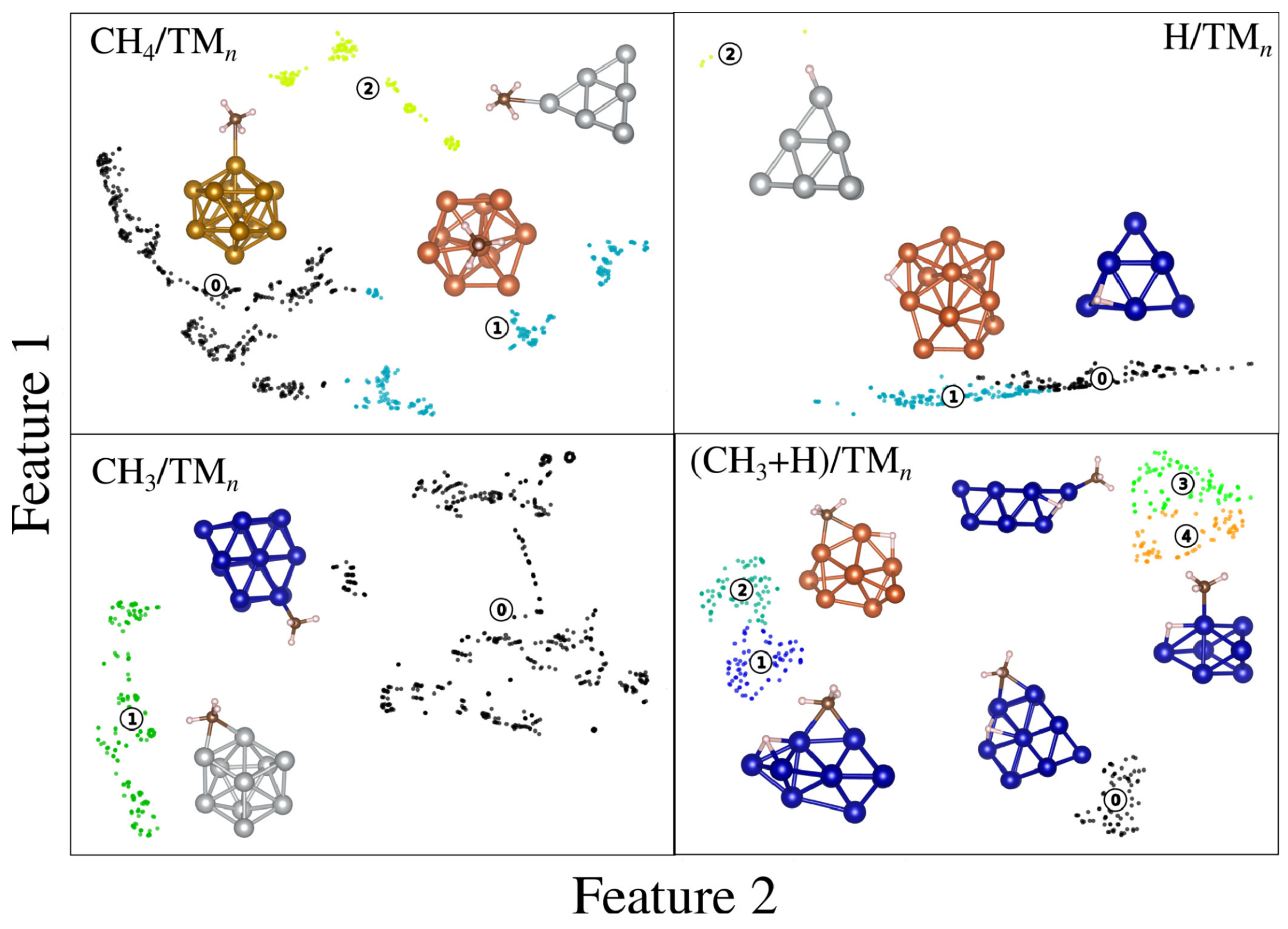
Over the years, materials science has evolved through several research paradigms. Initially, it relied on trial and error, then progressed to controlled laboratory experiments, and later to complex in silico simulations like Density Functional Theory (DFT). The latest paradigm leverages the extensive simulation data accumulated over time to train robust, accurate, and efficient machine learning models. Though still in its early stages, this approach has already led to significant advancements, demonstrating the potential of machine learning in materials science. [1]



Materials → Representation → Machine Learning → Insights and New Materials

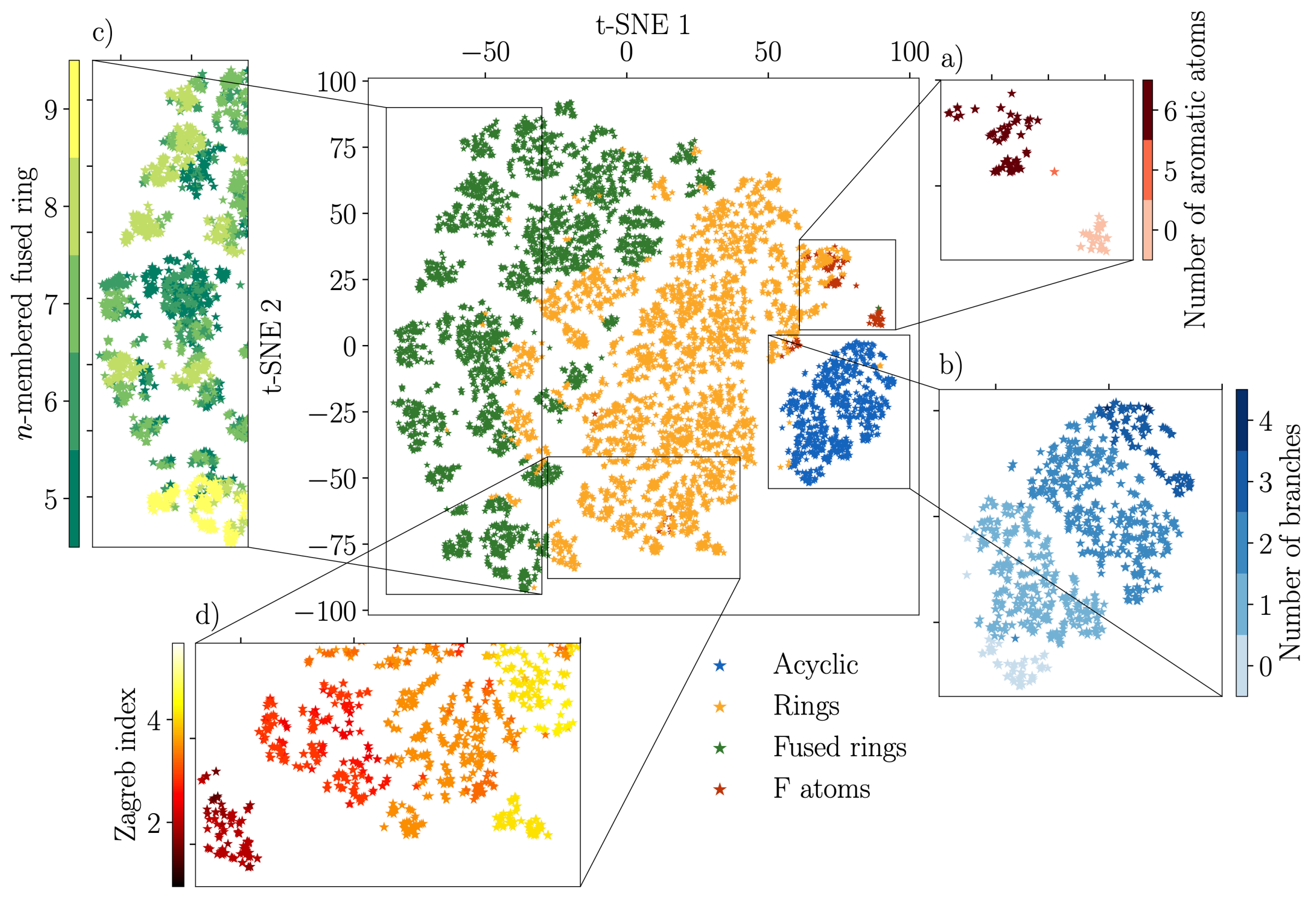
CLUSTERING ANALYSES

Clustering is a machine learning technique for grouping similar elements, based on a given set of characteristics. Our team has successfully used this method for analyses of chemical databases described, such as identifying phase transitions in finite-size particles [2]. Now, through similar techniques, we are also investigating the adsorption of H, CH3, CH3+H, and CH4 molecules on transition metal clusters (Fe, Co, Ni, Cu, and others) using data from a previous study [3]. The Figure on the right shows general modes of adsorption for Fe (yellow), Co (blue), Ni (gray), and Cu (orange) substrates with their adsorbing molecules



REPRESENTATION LEARNING

We also investigated machine learning approaches that learn representations from unlabeled data. Here, our goal is to develop a model that generates representations capable of generalizing across various categories of materials. Our initial results show that the model clusters similar molecules in the latent space using the learned representation, as illustrated in the figure on the right. Now, we are focusing on expanding this approach to handle datasets that contain a wide range of molecules. This will allow us to produce a unified model that can be fine-tuned for several molecular tasks.

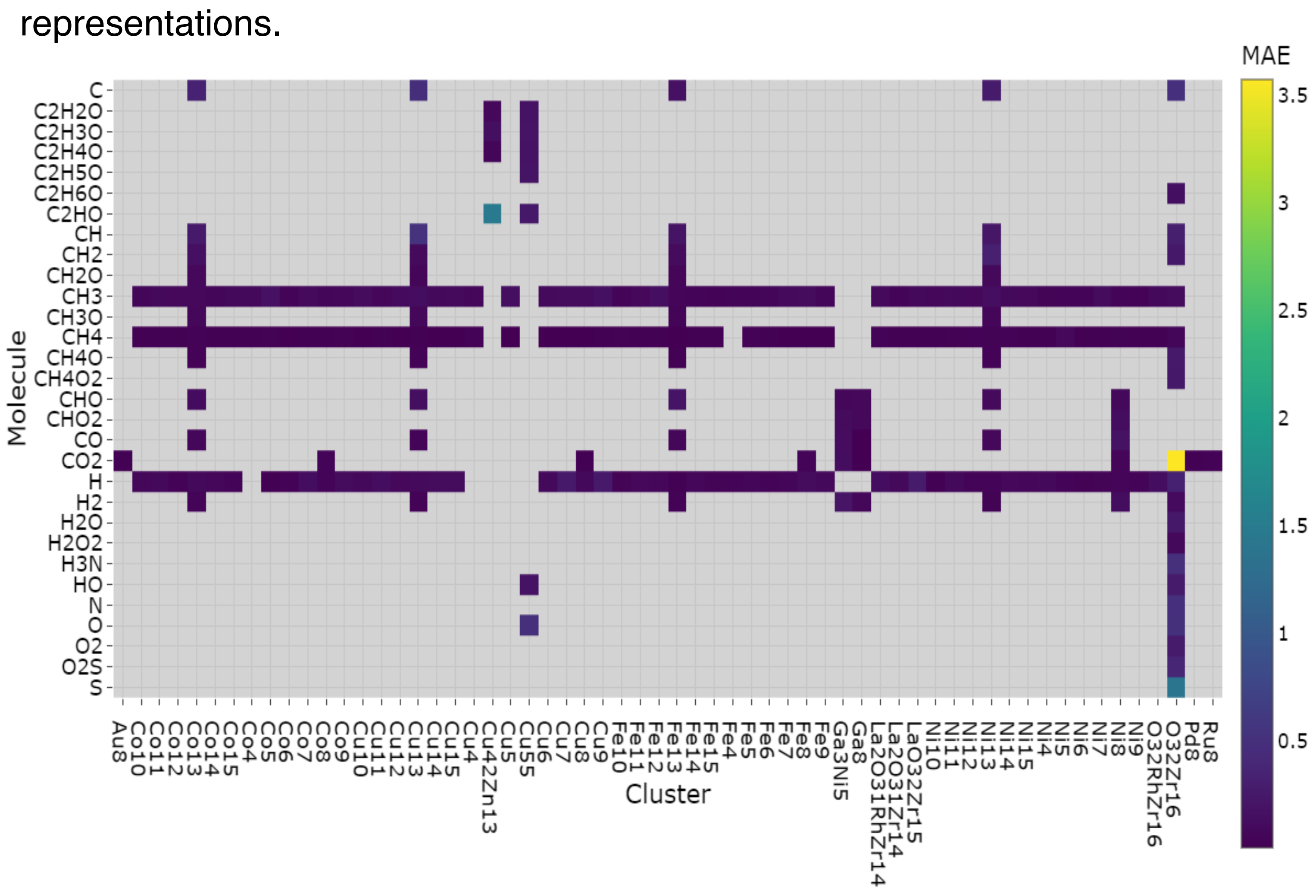


CONCLUSIONS & PERSPECTIVE

In conclusion, by leveraging the power of Machine Learning, we are addressing critical challenges in Materials Science. Our clustering analyses have successfully aided phase transitions and adsorption modes identification in chemical databases, while our property prediction frameworks have shown to be very accurate and promise to overcome data scarcity through multi-task learning techniques. Representation learning from unlabeled data shows promise in generalizing across diverse material categories, facilitating more robust and versatile models. Additionally, the application of NLP and large language models enhances our ability to automate literature analysis and data extraction.

PROPERTY PREDICTION

Another part of our work involves studying chemical descriptors to best represent data [4]. Through these studies, we are developing a framework that predicts interaction energy with minimal error using real-world data from our group, as shown in the Figure below. Our next step is to enhance these predictors through more sophisticated techniques, such as transformer-based models, which can leverage information from unlabeled databases to learn representations.



Another facet we are dealing with is the problem of data scarcity, which can be a reality for different material groups. Our team has been addressing this problem in our work with polymers by collecting data from literature and proposing a multi-task technique that benefits from multiple data sources.

NATURAL LANGUAGE PROCESSING

Our team also utilizes Natural Language Processing (NLP) and Large Language Models (LLMs) to support materials design. Specifically, we focus on two main lines of application:

Automatic Literature Analysis: By employing advanced NLP techniques, we automate the analysis of scientific literature. This approach allows us to efficiently sift through vast amounts of published research to identify trends, key findings, and emerging areas of interest.

Automatic Data Extraction from Literature: NLP and LLMs to automatically extract data from scientific papers. This includes experimental results, material properties, and methodological details, which are then integrated into our databases, allowing the development of machine learning models.

REFERENCES

[1] Schleder, Gabriel R., et al. "From DFT to machine learning: recent approaches to materials science–a review." Journal of Physics: Materials 2.3 (2019): 032001.

[2] de Mendonça, João Paulo A., et al. "Theoretical framework based on molecular dynamics and data mining analyses for the study of potential energy surfaces of finite-size particles." Journal of Chemical Information and Modeling 62.22 (2022): 5503-5512.

[3] Andriani, Karla F., et al. "Role of quantum-size effects in the dehydrogenation of CH 4 on 3d TM n clusters: DFT calculations combined with data mining." Catalysis Science & Technology 12.3 (2022): 916-926.

[4] Pinheiro, Gabriel A., et al. "Machine learning prediction of nine molecular properties based on the SMILES representation of the QM9 quantum-chemistry dataset." *The Journal of Physical Chemistry A* 124.47 (2020): 9854-9866.

ACKNOWLEDGEMENTS

CINE CENTER FOR INNOVATION ON NEW ENERGIES
CAPES
FAPESP