“Maturing the production standards of ultraporous structures for high density hydrogen storage bank operating on swinging temperatures and low compression” – MAST3RBoost

MAST3RBoost

D1.6. Interactive material maps (unsupervised ML), fabrication-structure (supervised ML) models

Due date of submission: 30/05/2023

Actual submission date: 25/05/2023

Funded by the European Union

TABLE OF CONTENTS

TABLE OF CONTENTS 2
PROJECT INFORMATION 3
DELIVERABLE DETAILS 4
1 Introduction 5
2 t-Stochastic Neighbour Embedding 5
3 Implementation Details 6
4 Test Data 6
5 Nanoporous Materials Data 8
6 Supervised Learning Methods 10
7 Conclusions & Future Work 11
8 Bibliography 11

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.

PROJECT INFORMATION

Project full title: Maturing the production standards of ultraporous structures for high density hydrogen storage bank operating on swinging temperatures and low compression

Acronym: MAST3RBoost

Call: HORIZON-CL4-2021-RESILIENCE-01

Topic: HORIZON-CL4-2021-RESILIENCE-01-17

Start date: 1st June 2022

Duration: 48 months

List of participants:

Number	Name of beneficiary	Acronym of beneficiary	Country
1	ENVIROHEMP	ENV	Spain
2	CONTACTICA	CTA	Spain
3	Consejo Superior de Investigaciones Científicas	CSIC	Spain
4	Spike Renewables Srl	SPIKE	Italy
5	EDAG Engineering GmbH	EDAG	Germany
6	Nanolayers	NANO	Estonia
7	FUNDACIÓN CIDETEC	CIDETEC	Spain
8	Leichtmetallkompetenzzentrum Ranshofen GmbH	LKR	Austria
9	University of Pretoria	UP	South Africa
10	Council for Scientific and Industrial Research	CSIR	South Africa
11	PSA	PSA	Portugal
12	TWI Ltd	TWI	UK
13	University of Nottingham	UoN	UK

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.

DELIVERABLE DETAILS

Document Number:	D1.6
Document Title:	Interactive material maps (unsupervised ML), fabrication-structure (supervised ML) models
Dissemination level	PU – Public
Period:	PR1
WP:	1
Task:	T1.3
Author:	Nanolayers OÜ
Abstract:	Nanolayers has developed the infrastructure needed to perform unsupervised and supervised machine-learning (ML) studies on nanoporous materials. The t-distributed stochastic neighbour embedding (TSNE) method was implemented and fully integrated into LabCore. This enables researchers to organise high-dimensional sample descriptors into convenient 2D or 3D maps, where similar materials appear grouped together. Maps were calculated for the database of pre-exisitng carbon-based material, revealing useful patterns that could be exploited in the search for higher performance. Given the small dataset size, it remains unclear whether the input descriptor containing the synthesis parameters is truly complete, but the appearance of patterns in the maps and the success of previous supervised ML, indicate it might be enough. The Nanolayers computational server now hosts a variety of standard libraries for supervised machine-learning applications, as well as custom tools to create more flexible models.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.

1 INTRODUCTION

The LabCore Digital Notebook platform has been further extended with an implementation of the t-stochastic neighbour embedding (tSNE) for unsupervised machine learning. In short, tSNE transform a set of high-dimensional points into a low-dimensional representation, typically in two or three dimensions, where it can be visualized. Using this non-linear transformation, the algorithm organises data points in simpler maps that can reveal patterns and structures too difficult to see in the high-dimensional space, suggesting the underlying relationships in the data.

This method can be a powerful tool to construct materials maps and guide experimental studies towards synthesis conditions likely to yield samples with desirable properties, without directly modelling the complex relationship between synthesis inputs and sample performance.

2 T-STOCHASTIC NEIGHBOUR EMBEDDING

The general idea of the method is to *project* high-dimensional data points onto a space of manageable dimensionality for visualisation purposes, for example two- or three-dimensional, and arrange them by similarity [1]. The similarity measure between pairs of data points, x_i and x_j, is calculated using a distance metric. Any distance metric can be chosen, but the Euclidean distance is often used for simplicity:

$$d(x_i, x_j) = |x_i - x_j| \quad .$$

Based on this measure, the algorithm calculates the conditional probability $p_{j|i}$ that x_i would choose x_j as one of its neigbours using the Gaussian distribution:

$$p_{j|i} = A_i \, exp\left(-\frac{d(x_i, x_j)^2}{2\sigma_i^2}\right)$$

$$A_i = \left[\sum_{k \neq i}^{N} exp\left(-\frac{d(x_i, x_k)^2}{2\sigma_i^2}\right)\right]^{-1}$$

where A_i are normalisation factors, and the standard deviation of the probability distributions σ_i , remain to be determined. The corresponding points y_i will be placed in the low-dimensional space, and their similarity probability is calculated using the student-t distribution:

$$q_{j|i} = B_i \left[1 + |y_i - y_j|^2\right]^{-1}$$

$$B_i = \left[\sum_{k \neq l}^{N} [1 + |y_k - y_l|^2]^{-1}\right]^{-1}$$

This distribution is chosen over Gaussian, because its fast decay will push similar points close to each other, but the long tail will prevent them from collapsing into a single location. The algorithm then distributes the points y_i randomly in the low-dimensional space and iteratively adjusts their positions in order to minimise the

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.

differences in similarity probabilities in the two spaces, $q_{j|i}$ and $p_{j|i}$.

The standard deviation parameter σ_i cannot be calculated and has to be injected in the procedure with the help of additional heuristics. The user has to provide a value for perplexity π_i, related to the Shannon entropy of the similarity probability:

$$\pi_i = 2^{-\sum_j p_{j|i} log_2(p_{j|i})}$$

During the optimisation, the algorithm performs a binary search of the σ_i so that the perplexity if the resulting $p_{j|i}$ has the fixed value provided by the user. Typically, all data points are assigned the same perplexity to avoid steering artificially the results. Loosely speaking, the perplexity determines the number of neighbours expected around each data point, thus defining the level of coarse-graining in the resulting maps. There is no optimal perplexity, and reasonable values depends on the size of the dataset.

Interpreting tSNE maps is not trivial since some of their features may be artificially induced by the choice of perplexity. For example, small perplexity may cause a set of completely random datapoints to form small clusters, hinting at patterns and relationships that are not real. Too large perplexity will push data points together, hiding any pattern in their similarities. The actual position of the points and distance between clusters of points are not meaningful, since there is no linear scaling between the original data space and the low-dimensional representation, thus it is not possible to sort the clusters based on their positions. Another issue is the choice of distance metric. The Euclidean norm is easy to understand and motivate, but it is not really meaningful in high-dimensional space, since it assigns the same importance to all components of the data, while in reality some may be more important in determining similarities.

3 Implementation Details

The LabCore integration is implemented in JavaScript and works entirely on the client web browser as a configurable notebook element (similar to a regular plot). The user is required to specify the data source among active tables and data matrices present in the notebook. Optionally, it is possible to specify a data array or table column as labelling data, used to colour output points. Through the interface, the user can also control the learning rate controlling the optimisation algorithm, and the perplexity. Despite being in principle possible to project the original data onto an any-dimensional space, the choice is here limited to only 2D or 3D maps which can be visualised and understood. The coordinates of the transformed points are automatically saved in the notebook, following the LabCore data standards, and can be retrieved in CSV format, along with labelling information (if applied in the calculation). tSNE calculations can be saved in the notebook as either static or dynamic. In static mode, the map is calculated by the user in the configuration panel of the tSNE element, and saved as an image to visualise in the notebook. In dynamic mode only the parameters necessary to calculate the map are saved, and the notebook will recalculate it each time it is refreshed. As a result, dynamic tSNE maps will change appearance each time, due to their stochastic initialisation, but, if converged properly, will always show the same patterns.

4 Test Data

In order to test the implementation, a well characterized known dataset is needed. The Mast3rboost datasets are exploratory by nature and are currently incomplete so the algorithm was applied to a public database of

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.

wine. The dataset (in matrix form) consists of 178 samples (rows), each characterised by 13 features (columns) related to their chemical and physical properties. An additional data column contains the quality of each wine sample, assessed by experts as one of three classes, is used for labelling. If there truly is relationship between the chemical properties of the wine and the classification assigned by human experts, the tSNE map should group the data points into, at least, three separate clusters, with matching labelling colour.

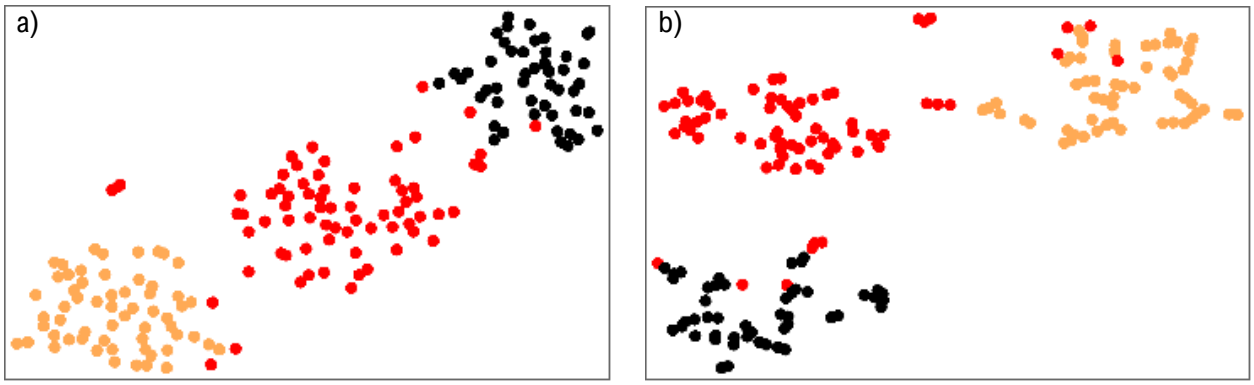

Figure 1: t-SNE maps of the wine dataset calculated with perplexity 20 (a) and 5 (b). The colours are assigned based on the quality classification labels.

The resulting tSNE maps calculated with perplexity 20 and 5 are shown in Figure **1**. As expected, the data points are organised into three main clusters, and the spatial arrangement matches the classification labels. The labels of few data points do not match those of the majority found in the same cluster, however, these points are located at the boundary. This indicates that while they are mostly similar to other samples in the same cluster, they are also not typical, and some differences are present. Likely the issue is in the human classification.

With low perplexity, the map still shows the main three clusters, although there seems to be a finer structure. This is likely an artifact of the computation, attempting to separate points that already have too many neighbours; however, it is not possible to determine for sure whether the finer patterns are truly there, since they might correspond to similarities in the data that could prove useful to observe with different labelling scheme.

These results confirm that the algorithm is working as expected.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.

5 NANOPOROUS MATERIALS DATA

The method was applied to the database of pre-existing carbon-based nanoporous materials compiled in D1.2 and used to train supervised ML models in D1.5. Three different datasets were composed from the features in the full database.

Table 1: datasets construction details.

Feature	# of features	dataset
atomic ratios [O,H,N,K]/C, K/O [input]	5	1,2
relative C/O content [input]	2	1,2
porogen (PO or KOH)	1	1,2
porogen weight ratio	1	1,2
compactivation (true/false)	1	1,2
activation temperature	1	1
ramp rate	1	1
relative C/O content [output]	2	3
atomic ratios [O,H,N]/C [output]	3	3

Table 1 lists the available features of each sample, and the datasets in which they appear. The first and second datasets includes features that characterise the precursor mixture and the synthesis process; the second one does not include the activation temperature and ramp rate, following the recommendation of experimentalists. The last dataset only includes features of elemental analysis on the sample resulting from synthesis of the precursor mixture. The tSNE maps calculated with the first two datasets can be used to guide the synthesis process, while maps from the third datasets can help identify relationships between the actual sample characteristics and its performance.

The Z-Score transform was applied to each dataset in order to avoid numerical artifacts: some features such as temperatures have very large numerical values, while some atomic ratios are 5 orders of magnitude smaller. The numerically large features would dictate the Euclidean similarity measure, possibly hiding patterns originating from the small-value features. All maps presented here were fully converged with perplexity set to 10. This choice is somewhat dictated by the small dataset size: larger values clustered the points too much, and lower values produced way too scattered maps, making it difficult to see patterns in both cases.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.

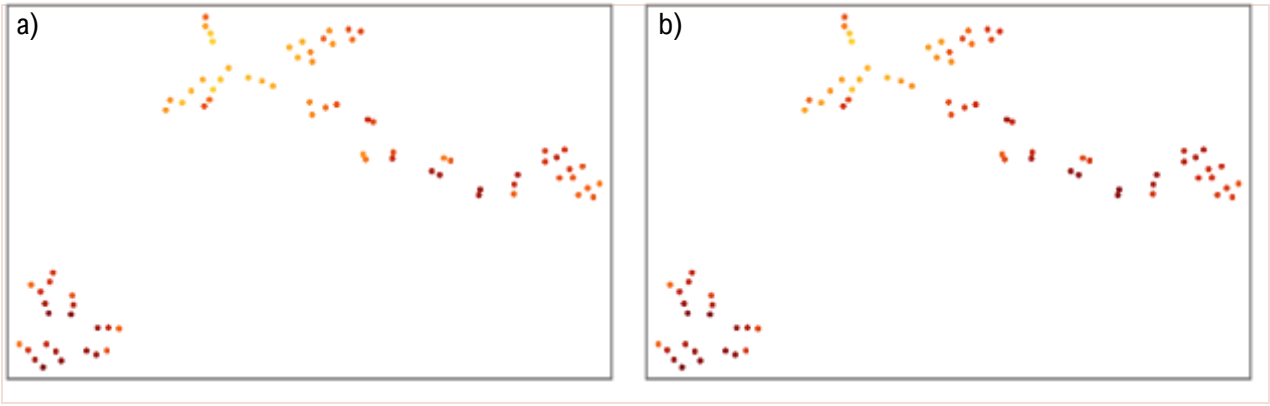

Figure 2: tSNE map of dataset 1 (synthesis inputs) labelled by A_{BET} (a) and V_{sp} (b). The colour brightness is proportional to the surface or volume of the pores.

Figure **2** shows the tSNE map of the synthesis inputs of the samples (dataset 1), labelled by A_{BET} and V_{sp}. The first thing to notice is that the colour labels are very similar in the two plots, indicating a correlation between A_{BET} and V_{sp}. The data is organised into two main clusters: a small one in the lower-left side and the bigger one in the top-right side. The smaller cluster mostly contains samples with low area or volume. The larger cluster appears elongated and structured. Most of the samples with high A_{BET} and V_{sp} are found in the upper-left portion of this cluster. The elongated arrangement of data points seems to indicate a smooth transition between the samples. It is possible that with more data, these structures could grow and become proper clusters, more clearly separated from the others.

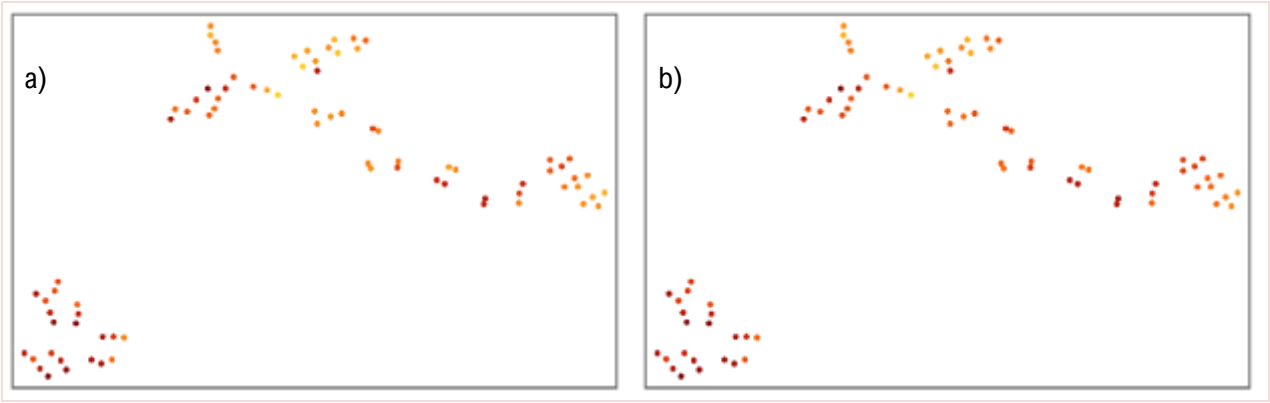

Figure 3: tSNE map of dataset 1 (synthesis inputs) labelled by A_{micro} (a) and V_{micro} (b). The colour brightness is proportional to the pore area or volume.

The same maps were labelled with A_{micro} and V_{micro} are shown in Figure **3**. The colour labels still match, so that surface and volume correlate, and samples with high values can still be found in the top-left portions of the larger cluster. However, the distribution is not as homogeneous as in Figure **2**, and the more high-valued samples are located in the right portion of the cluster. Some data points with A_{BET} and V_{sp} at the top of the scale, are found at the bottom of the scale of A_{micro} and V_{micro}. This is expected because A_{BET} measures the whole surface area of the sample (micro-, meso-, and small macro-pores) while A_{micro} only accounts for micropores, and the same reasoning applies to the volume measurements.

According to the experimentalists who measured these samples, the activation temperature and ramp rate should not contain useful, additional information not already included in the other features. In fact, the activation temperature is the threshold where the reaction is observed, and the ramp rate is determined by the instrument capabilities and operator choice. However, their presence in the dataset can introduce additional dissimilarities

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.

among samples, that should not be accounted. For this reason, the tSNE maps were recalculated for dataset 2, where these two features are missing.

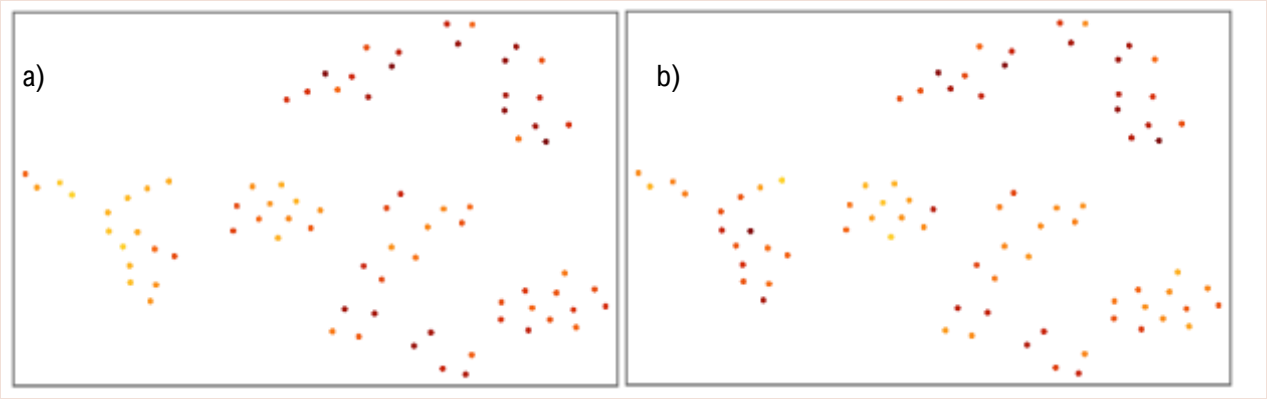

Figure 4: tSNE map of dataset 2 (synthesis inputs without activation temperature and ramp rate) labelled by A_{BET} (a) and A_{micro} (b). The colour brightness is proportional to the pore area.

The resulting maps in Figure **4** appear significantly less scattered than before, likely due to the missing temperature treatment features in the dataset, but still preserve their meaning. Samples with high A_{BET} samples are mostly located in one cluster on the left side, and some can be found in the nearby cluster at its right side. The situation is reversed when labelling the points after A_{micro}: the left-most cluster has only few high-area samples, while most are in the adjacent group.

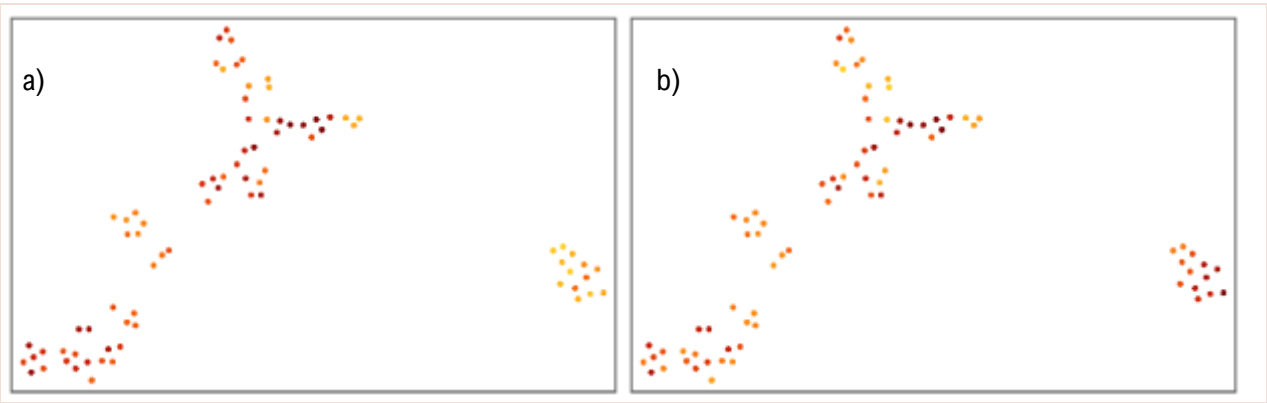

Figure 5: tSNE map of dataset 3 (sample composition outputs) labelled by A_{BET} (a) and A_{micro} (b). The colour brightness is proportional to the pore area.

The maps obtained with dataset 3 organise the samples according to similarities in their actual composition (Figure **5**). As seen in all other maps, samples with high A_{BET} tends to group together into one cluster, but few are found elsewhere: these scattered samples also have large A_{micro}.

6 Supervised Learning Methods

Even though there is currently not enough data to attempt training a more sophisticated supervised ML model, such as deep neural networks, Nanolayers has prepared a software suite dedicated to this task, ready to operate when larger datasets will be available. The standard python packages (Keras, TensorFlow) commonly used in ML applications have been installed on Nanolayer's server, alongside implementations of custom, proprietary methods and tools to train models. These are not integrated within the LabCore web platform since their computational cost would be prohibitive for a simple web-browser, but they can interface with notebooks and

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.

their data using the LabCore python API. This makes it possible to fetch data directly from notebooks already in a format suitable for the ML software.

7 Conclusions & Future Work

The tSNE calculator is now fully integrated in LabCore: tSNE maps can be calculated from data tables, following the directives of the user. The tool is quite flexible and allows the user to easily select the data columns to include as features, and to use as colour labels. The maps are recalculated each time the notebook is refreshed, and new data is added to the source tables.

Despite the small size of the current dataset, tSNE has proved to be useful in making easy-to-read maps of nanoporous materials arranged by similarity. Moreso, labelling the data points by their sample properties (pores area/volume) allows us to use this unsupervised ML method almost as a supervised predictor model. This is possible when the spatial patterns correlate to the labels (colour), as is the case in all our calculations. Ideally, samples with both high A_{BET} and A_{micro} (and volumes) are desired, and few of these could be found in our maps.

Future work will be focused on improving usability of the tSNE element within notebooks, so that it can be directly used in the laboratory to guide investigation. To this end, it will be necessary to add support for mixed labelled/unlabelled data, which will make it possible to visualise maps even if some points have missing label values. With this feature, experimentalists could infer the final properties of a sample from its location in the map, just by knowing its composition, or even the composition of the precursor mixture, before synthesis.

When more data will be available, the tSNE patterns may become clearer and even more insightful.

8 Bibliography

[1] L. van der Maaten und G. Hinton, „Visualizing Data Using t-SNE," *Journal of Machine Learning Research,* Nr. 9, p. 2579, 2008.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or [name of the granting authority]. Neither the European Union nor the granting authority can be held responsible for them.