Scrap Metal Explorer: Visual Exploration of Material Science Simulation Data

Tobias M. A. Nickel* Alexander Wiebel†

UX-Vis Research Group, Center for Research and Technology (ZFT)
Hochschule Worms University of Applied Sciences

Index Terms: Material science, Dimensionality reduction, t-SNE.

1 Introduction

The SciVis Contest 2025 tackles the challenge of accelerating materials research for a circular economy by applying the CALPHAD method to evaluate alloy compositions made from scrap metal [3].

CALPHAD (CALculation of PHAse Diagrams) is a computational method for predicting thermodynamic properties and phase equilibria in multicomponent systems. It combines experimental data with theoretical models, supported by thermodynamic databases, and is widely used in alloy design and microstructure simulation [1].

The dataset used in the contest, generated via CALPHAD, contains over 100,000 observations Ω_i, each with $n = 70$ variables X_j. Six input variables represent material composition (percentages summing to 100%), while the remaining 64 describe output properties such as density, yield strength (stress at which a material begins to deform plastically), and stable phases [3]. Due to its high dimensionality, the dataset requires specialized preprocessing and visualization techniques for effective exploration and analysis. The dataset includes several highly or perfectly correlated features.

In this contest contribution, we explore ways of reducing the number of dimensions to be processed (section 3) for analysis, ways of reducing the values to be explored (section 2 and section 4), and combinations (section 5) of dimensionality reduction and visualization techniques to support the discovery and evaluation of promising alloy compositions. In addition to widely used techniques like t-SNE, scatter plots and parallel coordinate plots, we suggest employing a user-adjustable scoring function and *sparse PCA*.

2 Scoring Function

In practical applications, certain material properties are critical, while others can vary without significantly affecting suitability of a material for a certain purpose. To accommodate this variability and provide users with flexibility to pursue their specific design goals, we implemented a customizable scoring function $S(\Omega_i)$. This function maps a user-defined subset of features to a single score dimension, allowing for the prioritization of relevant material characteristics according to specific use cases.

In the following defintion of the score $S(\Omega_i)$, we use x_{ij} to denote the value of feature variable X_j in observation Ω_i, the optimization direction of variable X_j (1 for maximization, -1 for minimization) is denoted as d_j, and w_j is the weight of variable X_j for the score. The overall score $S(\Omega_i)$ of an observation is calculated as:

$$S(\Omega_i) = \frac{\sum_{j=1}^{n} w_j \cdot M_d(d_j, x_{ij})}{\sum_{j=1}^{n} w_j}, \text{with}$$

$$M_d(d_j, x_{ij})) = \begin{cases} M(x_{ij}) & \text{if } d_j = 1 \\ 1 - M(x_{ij}) & \text{if } d_j = -1 \end{cases}$$

Min-max scaling M transforms each variable to a [0, 1] range, M_d is a flipped version of this variable in order to account for if a higher value or a lower value is better. Since this removes information about the absolute variance of each variable, dataset-specific adjustments to the weights are often necessary to optimize the scoring process and yield the most valuable insights. However, accurately calibrating these weights requires domain expertise to properly reflect each variable's relative importance.

We scored materials based on mechanical and thermophysical properties relevant for structural aircraft components [4], minimizing *Density*, *CSC*[1], and *delta_T*[2], while maximizing yield strength (*YS*) and *Therm. Conductivity*. Notably, no weighting factors were applied. This score provides a quantitative measure of material suitability to guide subsequent visualization.

3 Dimension Removal

The dimensions *Vf_MG2ZN3*, *T_MG2ZN3*, and *T_AL3X* were removed due to containing only `NaN` values. To further reduce dimensionality, we grouped semantically related and highly correlated features, selecting one representative from each group using a minimum correlation threshold of 0.95. The following correlated feature groups were identified, with retained features shown in **bold**:

- Group 1: ***Therm. conductivity***, *Therm. resistivity* and *Therm. diffusivity* exhibited a correlation of 0.99.
- Group 2: *Linear thermal expansion*, *Technical thermal expansion*, ***CTEvol*** and *heat capacity* with a correlation of 0.98.
- Group 3: *El. resistivity (Ωm)* and ***El. conductivity*** are reciprocal thus correlation is 1.0.
- Group 4: *Hardness (Vickers)* and ***YS (MPa)*** were highly correlated (1.0).

4 Sparse PCA

PCA was applied as a baseline linear dimensionality reduction method, but the resulting components showed little meaningful structure or clustering in scatterplots, suggesting limited utility.

To reduce dimensionality, Sparse PCA (SPCA) [5] was used. SPCA applies L_1 regularization to produce sparse loadings, enhancing interpretability by limiting the number of features influencing each sparse principal component (SPC). The regularization strength is controlled by hyperparameter λ [5]. We optimized λ (range $[1, 200]$) and the number of components m (1 to 18) to minimize reconstruction mean squared error (MSE), while balancing two additional criteria: minimizing component count and maximizing component separation (CSS) for interpretability. Weights for the objectives were 1 for MSE, 3 for CSS, as well as 0.5 for component count.

*e-mail: inf4481@hs-worms.de
†e-mail: wiebel@hs-worms.de

[1] Hot crack susceptibility coefficient
[2] Solidification interval (T(liqu) – T(sol), in °C)

Component separation was quantified via the mean pairwise Intersection-over-Union (IoU) across nonzero feature sets of components, extending the Jaccard index [2]. Given n components $C = \{c_1, c_2, \ldots, c_m\}$, where each c_k is the set of contributing features of component k, the Component Separation Score CSS is defined as:

$$CSS = 1 - \frac{2}{m(m-1)} \sum_{k=1}^{m-1} \sum_{l=k+1}^{m} \frac{|c_k \cap c_l|}{|c_k \cup c_l|}$$

This score ranges from 0 (complete overlap between components, normal PCA) to 1 (no overlap at all).

SparsePCA components were visualized via parallel coordinate plots (see subsection 5.2 and subsection 5.3), enabling overview and detailed exploration of the material property space.

4.1 t-SNE

Separate t-SNE visualizations were generated for the input and output parameters, using perplexity values of 30 and 50, respectively. The input t-SNE shows a 2D projection of a 6D convex diamond structure, indicating that the 6D search space was evenly sampled. The output displays multiple meandering river-like structures.

5 Tool: Scrap Metal Explorer

To facilitate practical and efficient data analysis, we developed an interactive visualization tool (see Figure 1). To ensure usability on machines with limited computational resources, the tool supports dataset subsampling, defaulting to 25% (see Figure 1, top left). Figures 4 and 5 demonstrate that subsampled layouts closely resemble those of the full dataset.

Users begin by specifying a custom objective (see Figure 1, top right) for the scoring function $S(\Omega_i)$. This involves selecting target variables, defining whether they should be maximized or minimized, and assigning a weight. Once these parameters are set, a reduced dataset is generated accordingly, allowing for exploration.

5.1 Linked t-SNE Views

The tool provides linked t-SNE views (see Figure 1, middle) of the dataset, colored by the user-defined score metric. Users can configure the direction of linking between views (Figure 3) and enable tooltips that reveal additional variable information on hover. This enables visual exploration of how specific input regions relate to output behaviors, while being able to oberserve simple patterns in the additional variables.

5.2 Explore Optimization results

Optimization results are visualized using a parallel coordinates plot (see Figure 1, bottom). The plot displays the optimization parameters alongside a user-defined number of sparse principal components (SPCs), which are sorted by their correlation with the scoring objective to guide interpretation. By default, the plot is colored by the score, though users may optionally select a different variable for coloring. In addition, the most correlated SPC per objective is shown. This visualization provides a compact summary of both raw input variables and latent structures, allowing users to quickly identify dimensions that impact optimization outcomes.

5.3 Drill Down

After identifying SPCs of interest, users can further examine their structure. The tool displays component loadings in tabular form, along with a parallel coordinates plot of the original variables that contribute to each SPC (see Figure 2). The default coloring remains the score, but this can be adjusted. Users may also add other variables to the plot to provide additional context. This detail-oriented plot enables targeted analysis of specific patterns, revealing interpretable connections between influential variables and the optimization objective or other variables.

6 Results

With the selected optimization target, the output t-SNE plot shows that geometric clusters do not directly correspond to high or low score regions. Instead, a gradient is visible within and across several river-like structures. Hovering reveals that within each "river," *KS1295[%]* remains constant, while *6082[%]* varies orthogonally, driving the score. Other input variables show no significant individual influence (see Figure 7).

This pattern is confirmed in the drill-down plot, where selecting the relevant input dimensions and coloring by score reveals lower scores for high *6082[%]* and *KS1295[%]* values, while the score remains relatively stable across the other variables (see Figure 8).

The optimization plot reveals conflicting objectives. Most variables align with their target directions, showing high scores at the desired ends of ranges, while *YS* (Yield Strength) is nearly inversely correlated with the score. This suggests a trade-off with *Therm. Conductivity*, as the two are inversely related (see Figure 13).

Coloring the plot by *6082[%]* shows that it increases both *Density* and *delta_1*, which are properties to be minimized (see Figure 11). *KS1295[%]* shows a similar, though weaker, effect (see Figure 12). Coloring the input dimensions by *YS* gives insight into the optimiziation conflict. High *YS* values result from increased *KS1295[%]*, which is otherwise unfavorable. *4032[%]* also raises *YS* but compromises thermal performance, making it a poor trade-off overall (see Figure 9, Figure 10).

Drilling down into *SPC_11*, the sparse component most correlated with the score, reveals that it captures shared variance across three optimization objectives. Its strong correlation with density indicates, that density is significantly correlated with multiple other relevant material properties (see Figure 17).

Given the strong correlation of *SPC_1* with both *YS* and *Therm. Conductivity*, this component can be interpreted as capturing a trade-off axis between these properties and related variables. It facilitates understanding of which underlying material features influence *YS* and *Therm. Conductivity* and their effect on the overall score (see Figure 16). Notably, the similarities between *KS1295[%]* and *4032[%]* in relation to *YS* also extend to the other related variables (see Figure 15, Figure 14).

7 Conclusion

The devised analysis tool enables us to reduce both the number of samples and dimensions, providing an intuitive overview as a starting point for exploration, while also supporting targeted, in-depth analysis of the data. In particular, the configurable score function allowed users to easily recognize parameter combinations which support their desired material properties and recodnise potential trade-offs. SPCA allowed us to significantly reduce the number of dimensions to be inspected while retaining an intuitive meaning attached to the variable. The expressiveness of the standard visualization techniques we used has been significantly boosted by the scoring and the topic-specific SPCA components we suggest.

References

[1] Modelling of phase diagrams and thermodynamic properties using calphad method – development of thermodynamic databases. *Computational Materials Science*, 66:3–13, 2013. Multiscale simulation of heterogeneous materials and coupling of thermodynamic models. 1

[2] P. Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 44:223–270, 1908. 2

[3] SciVis Contest 2025 Committee. Scivis contest 2025: Dataset description and download. Online, October 2024. 1

[4] A. Yılmaz and O. Civalek. A brief introduction to the properties of aerospace materials. *International Journal of Engineering and Applied Sciences*, 17:44–60, 05 2025. doi: 10.24107/ijeas.1640337 1

[5] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006. doi: 10.1198/106186006X113430 1

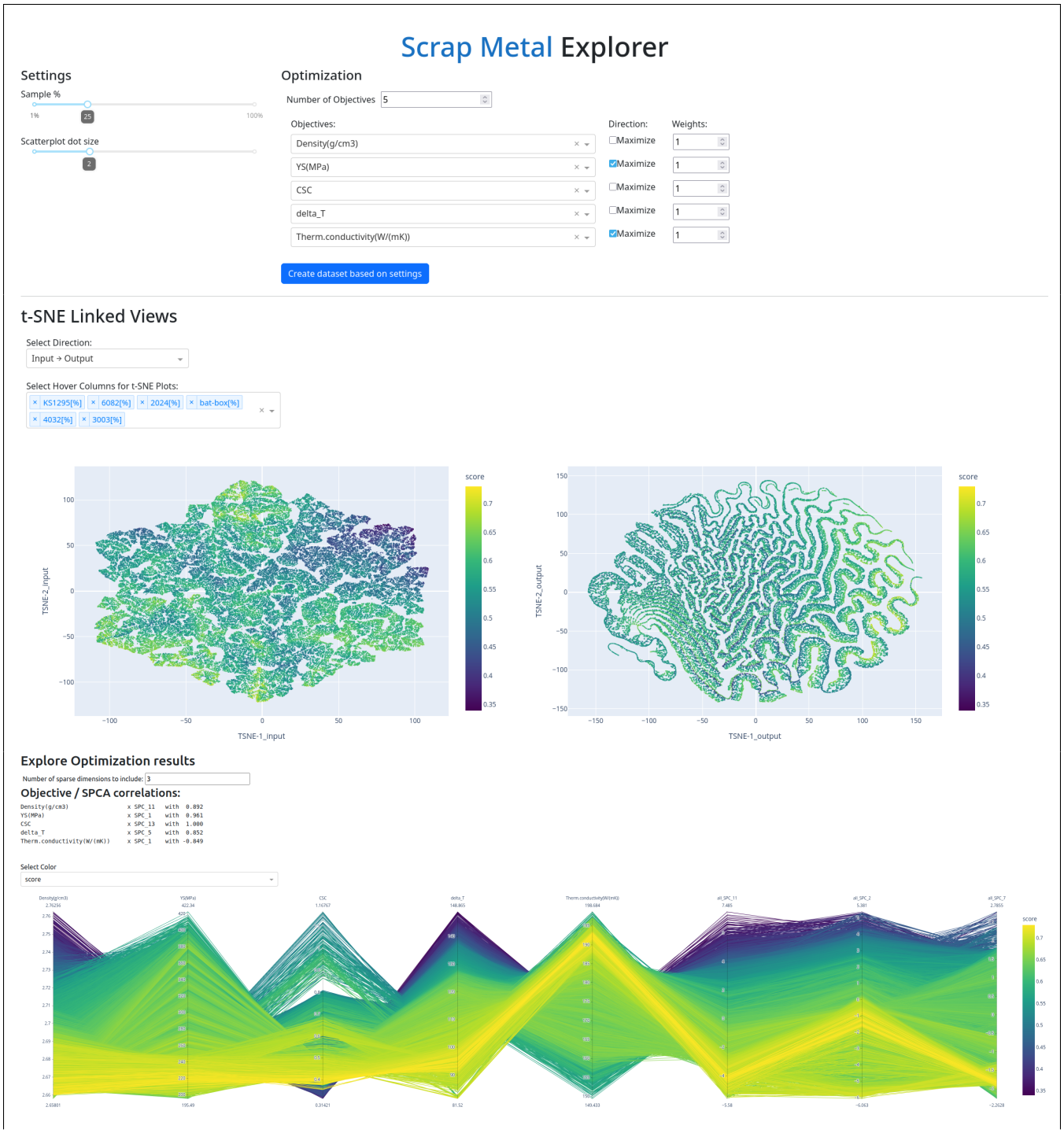

Figure 1: Top part ⊓ of the user interface □ of the tool after start-up. See Figure 2 for bottom part ⊔ .

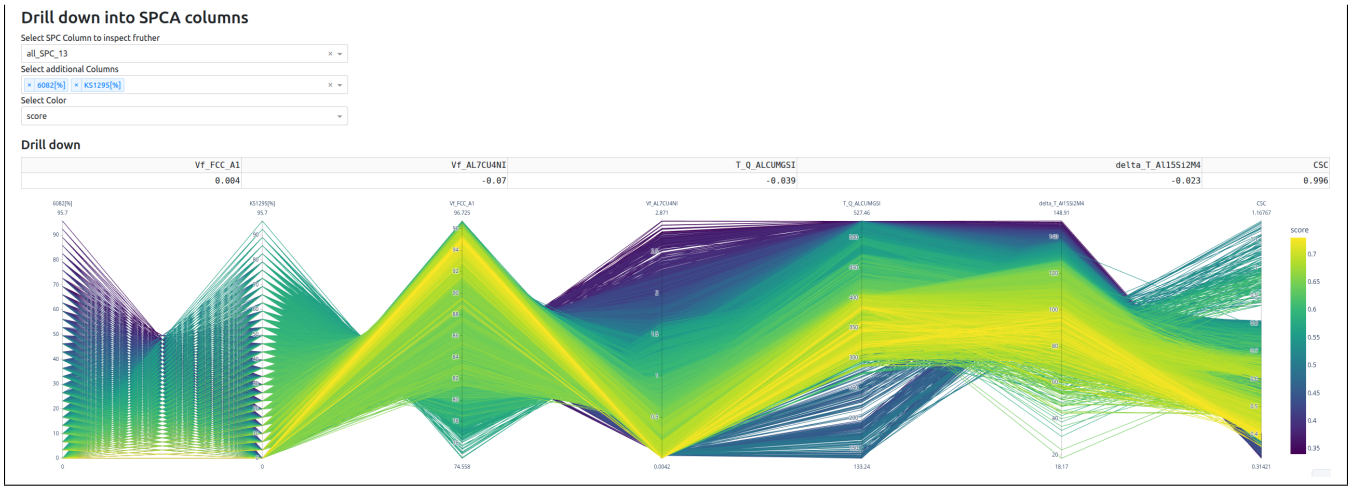

Figure 2: Bottom part ⊔ of the user interface ☐. See Figure 1 for top part ⊓.

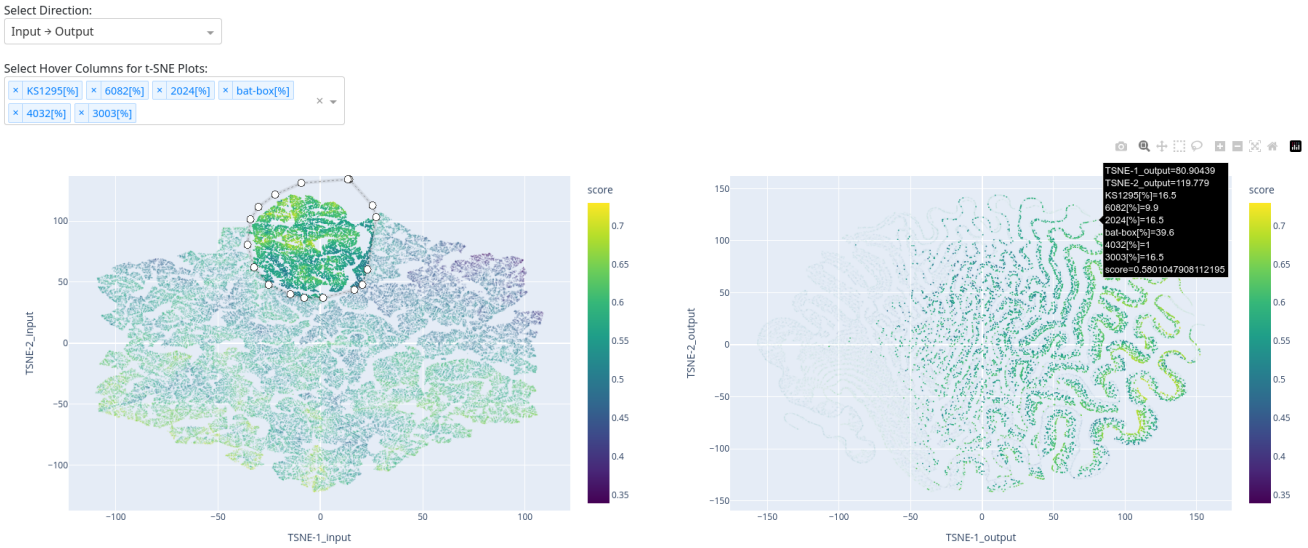

Figure 3: Selection of data samples in t-SNE of input parameters highlights the same samples in the t-SNE of all output parameters. In the right figure the mouse hovers over a sample in order to get detailed information of parameter values for this sample in an info-box.

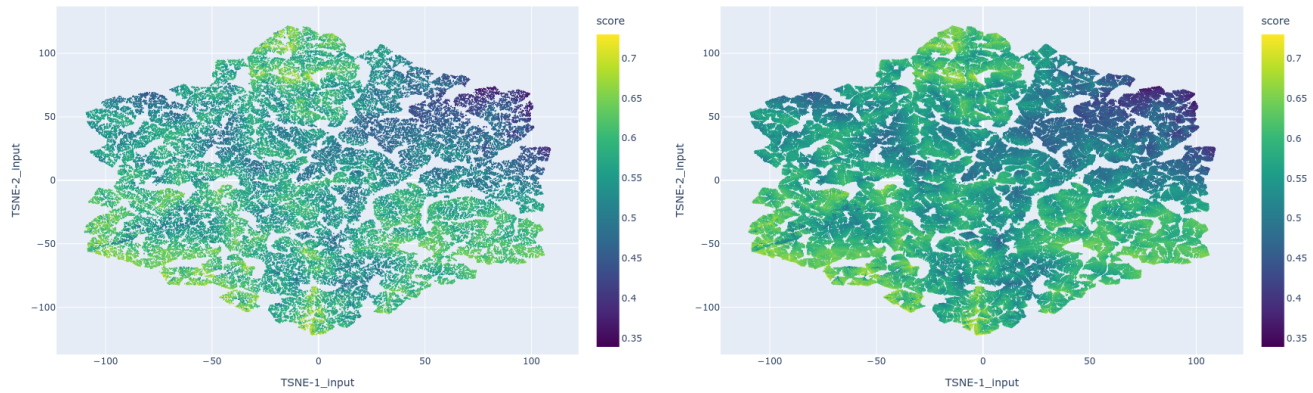

Figure 4: Comparison of t-SNE of **input** parameters for 25% of the data points vs. of all data points

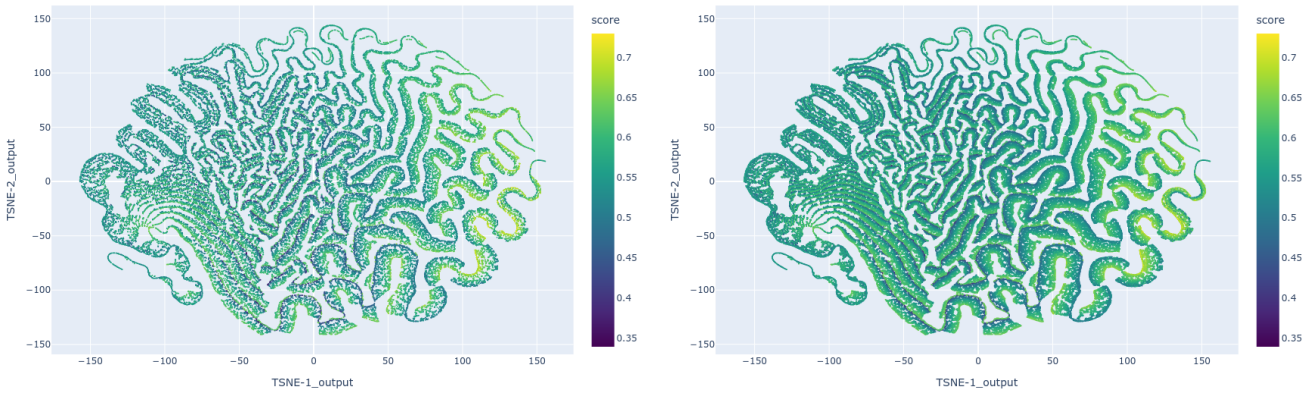

Figure 5: Comparison of t-SNE of **output** parameters for 25% of the data points vs. of all data points

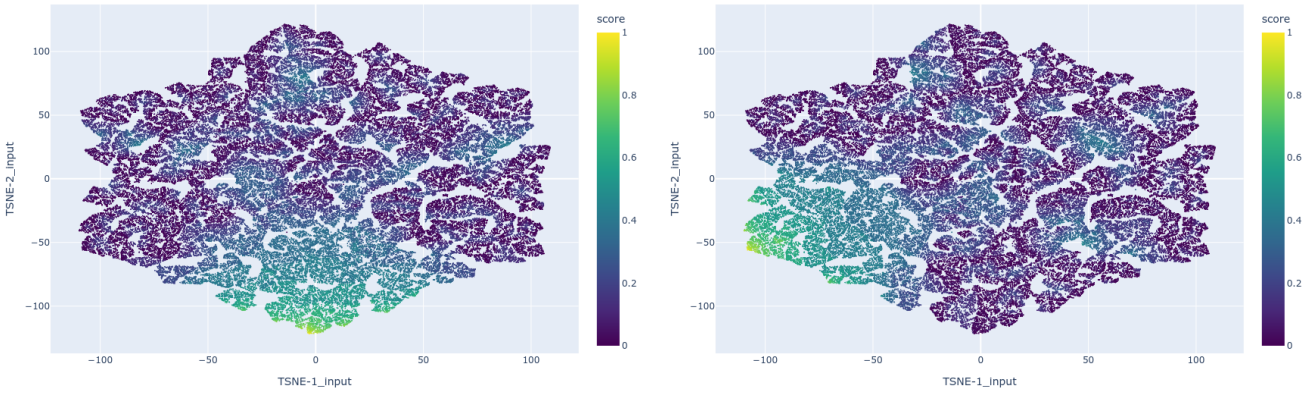

Figure 6: These two images show that the t-SNE of the six **input** parameters layouts datapoints such that the maximum for each variable is located in the corner of a hexagon. Because our t-SNE plots are always colored by score we are using **dummy scores** consisting of of only one input parameter for illustrative purposes in this figure. Left: Coloring for *2024*. Right: Coloring for *bat-box*.

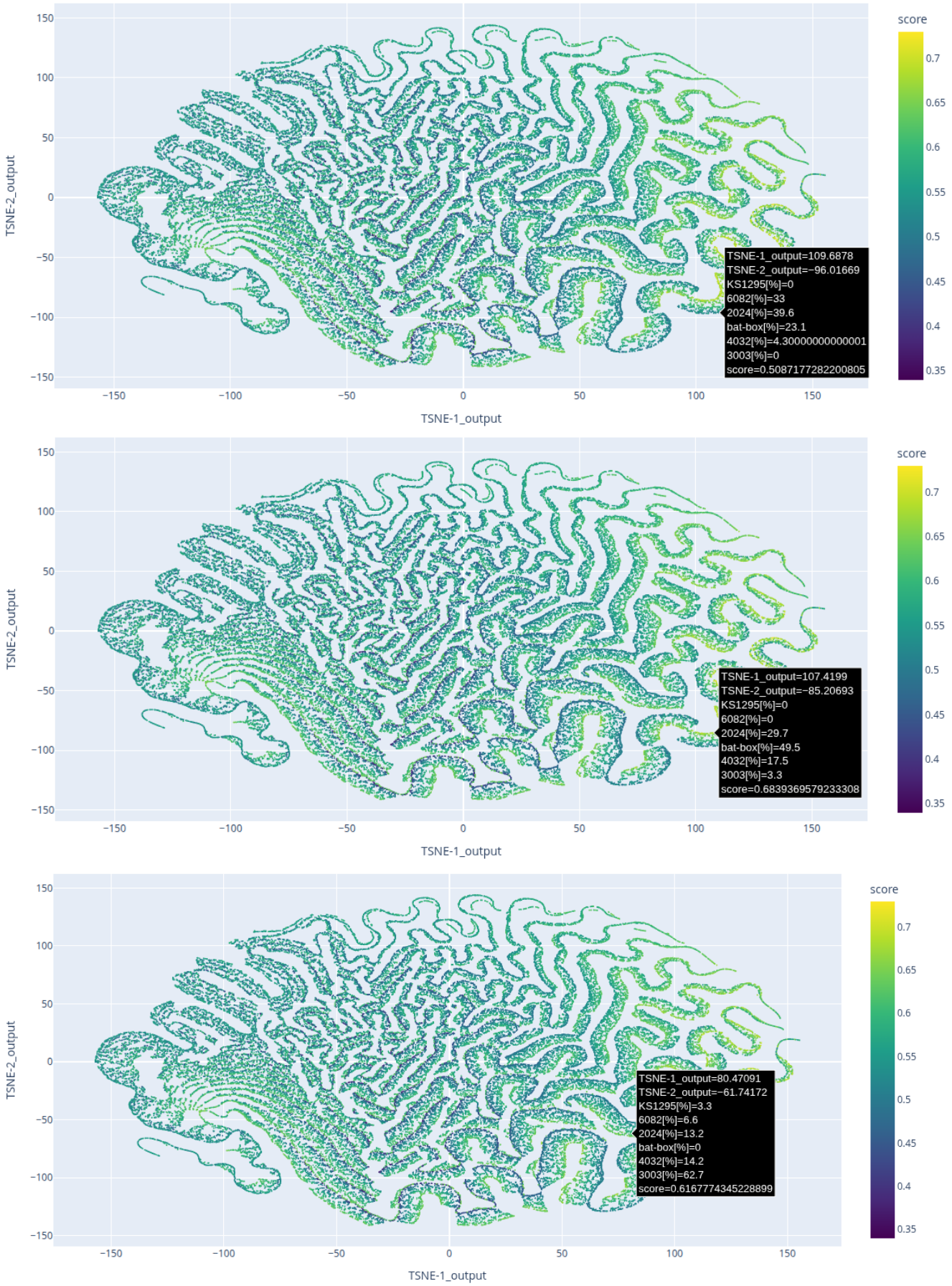

Figure 7: Showing the resulting river pattern and variable variation.

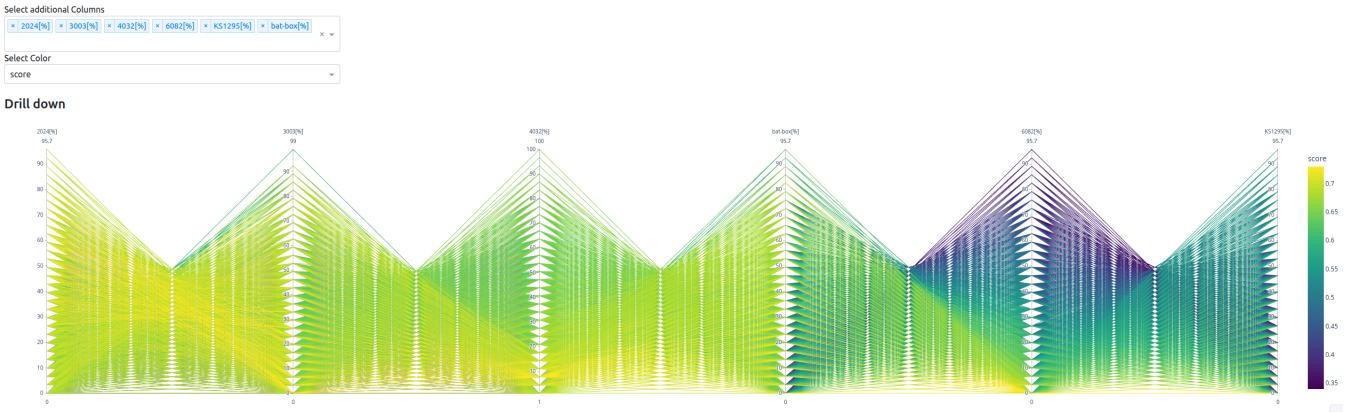

Figure 8: Showing the input dimensions colored by score.

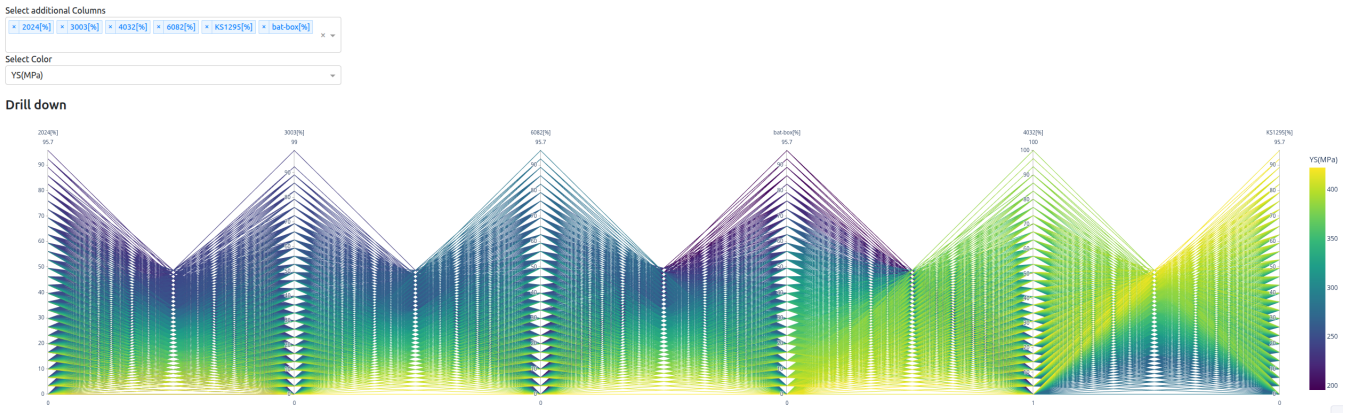

Figure 9: Showing the input dimensions colored by Yield. Strength

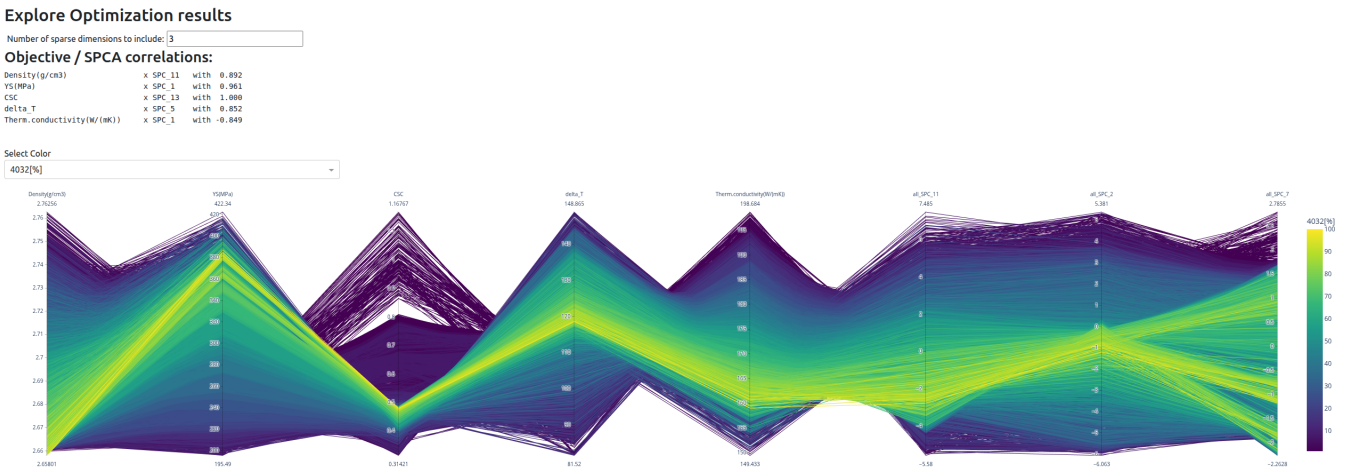

Figure 10: Showing the optimization targets colored by 4032[%].

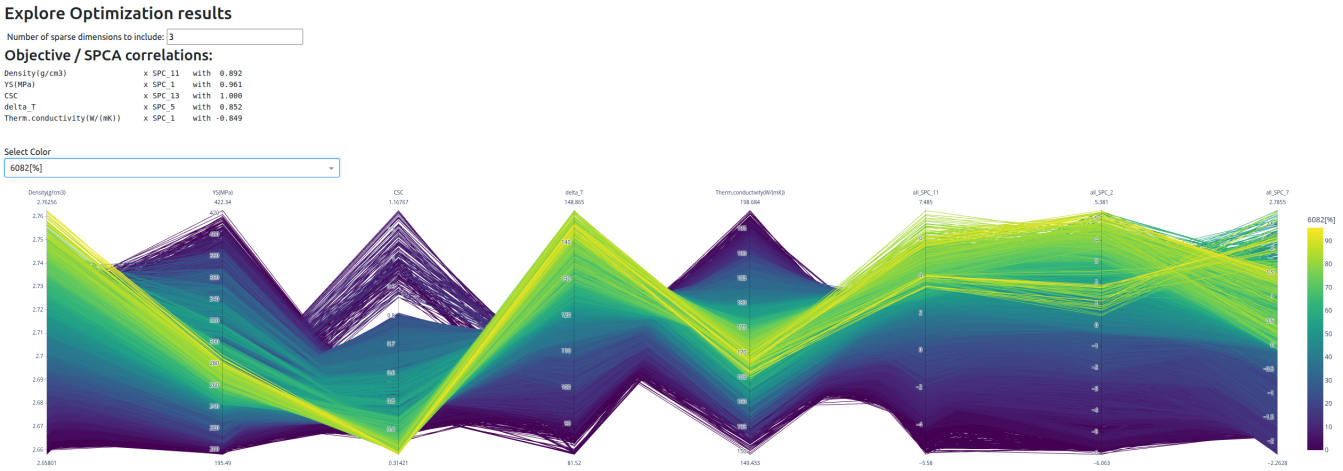

Figure 11: Showing the optimization targets colored by 6082[%].

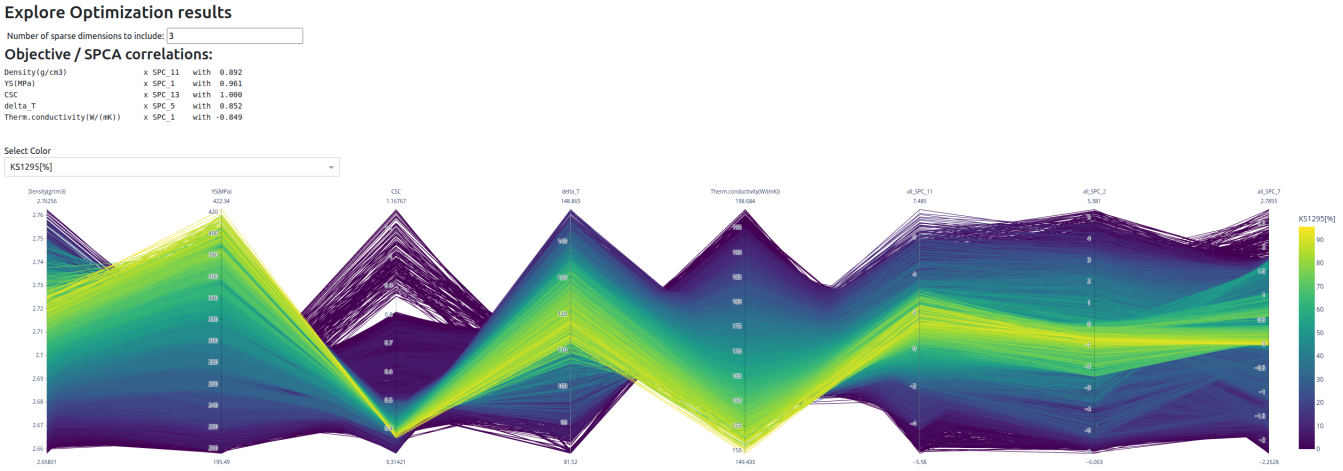

Figure 12: Showing the optimization targets colored by KS1295[%].

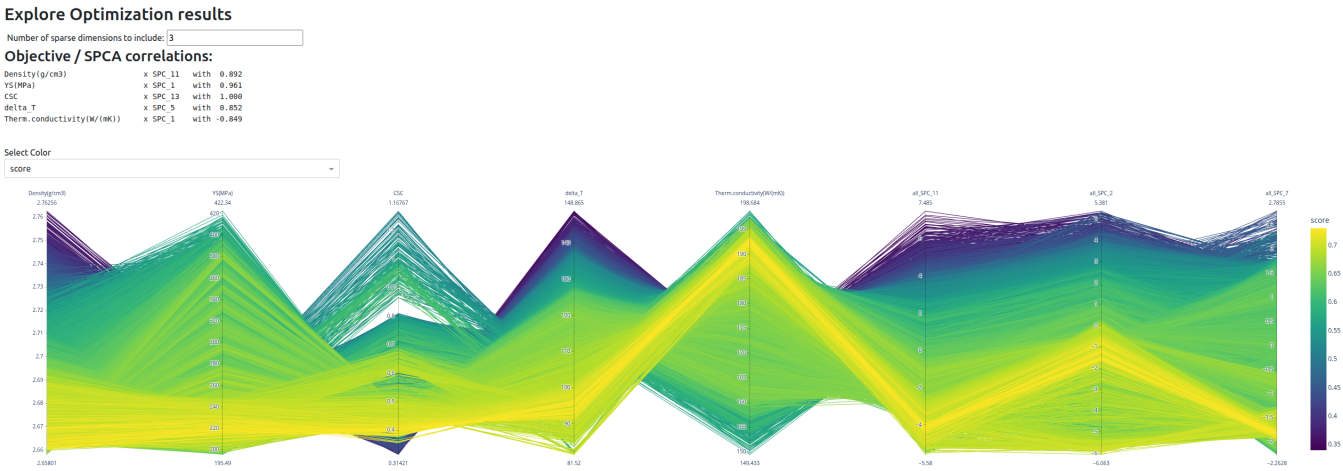

Figure 13: Showing the optimization targets colored by score.

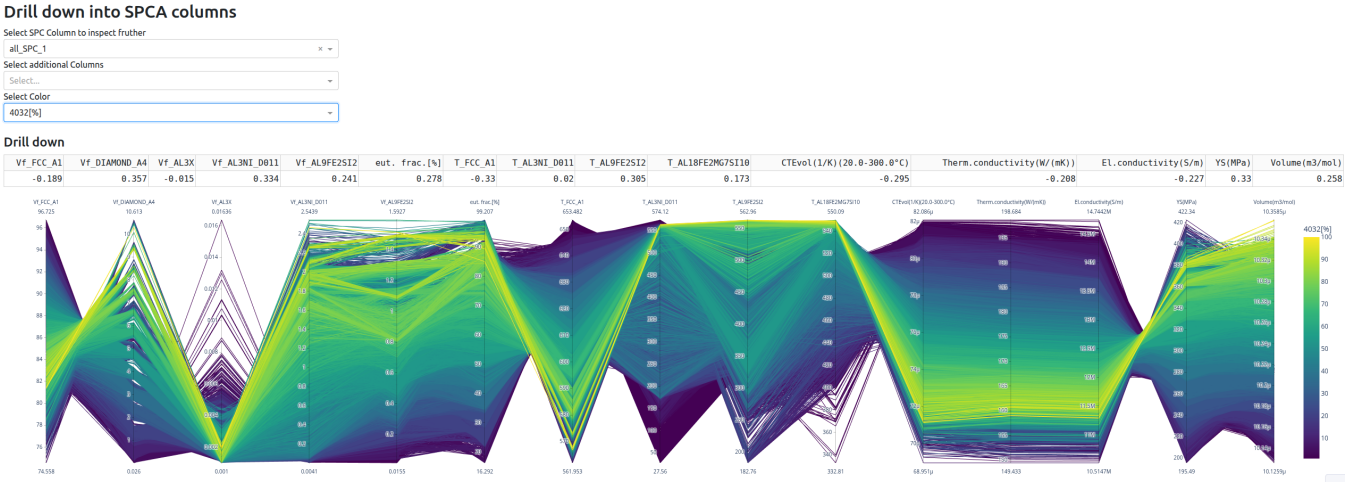

Figure 14: Showing loadings of SPC_1 colored by 4032[%].

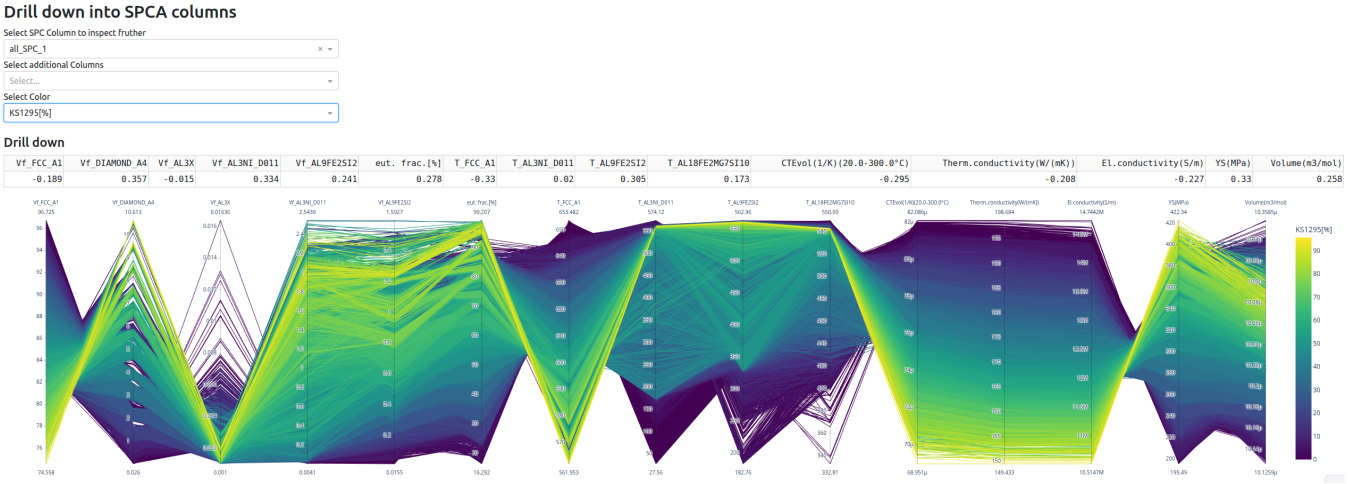

Figure 15: Showing loadings of SPC_1 colored by KS1295[%].

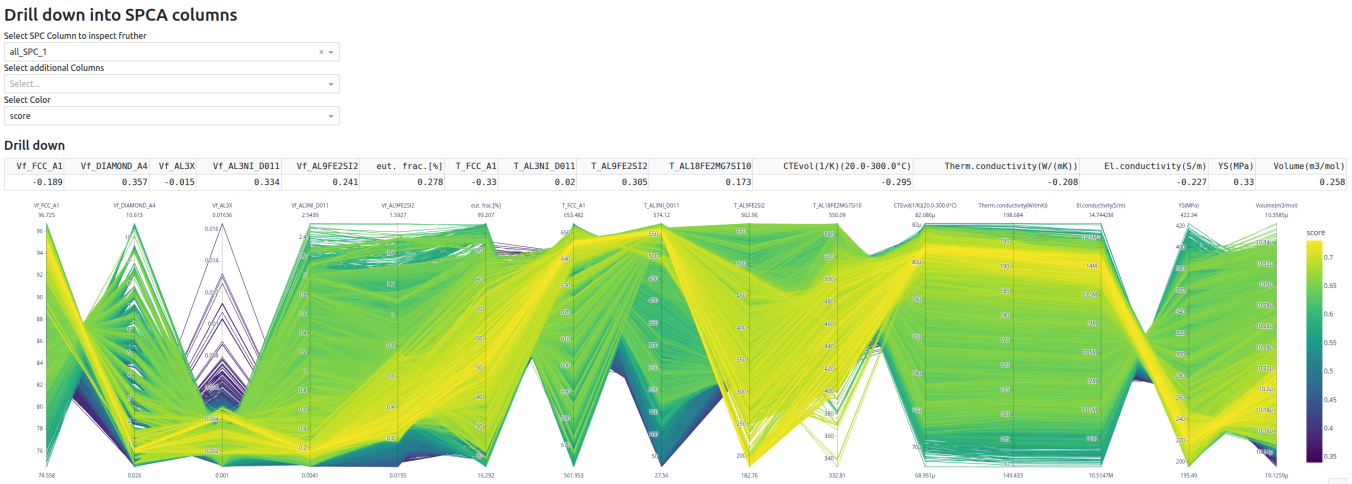

Figure 16: Showing loadings of SPC_1 colored by score.

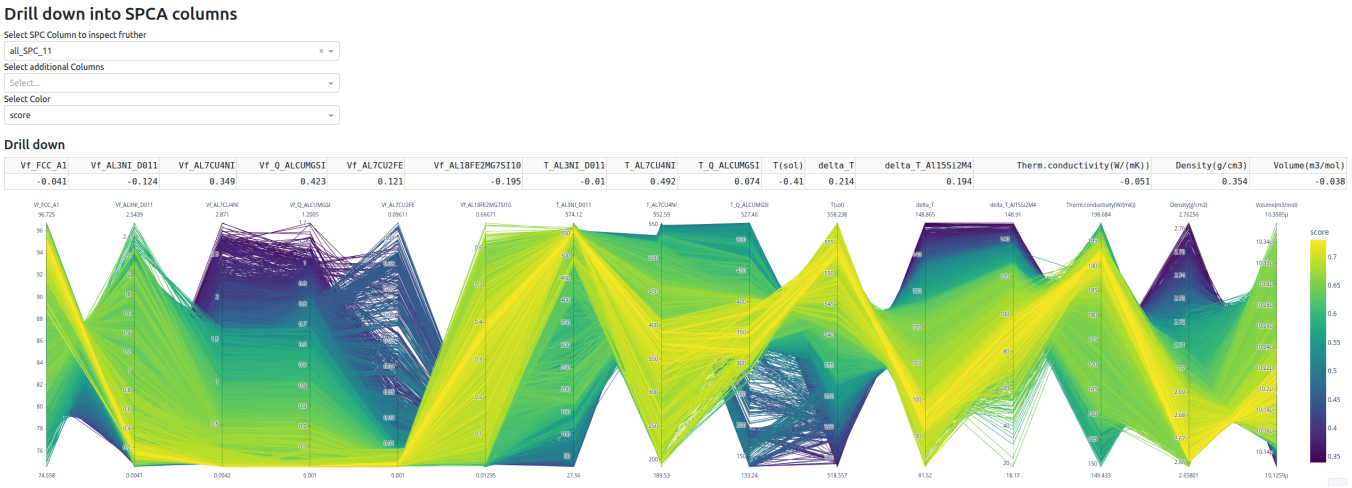

Figure 17: Showing loadings of SPC_11 colored by score.