CoSi: Visual Comparison of Similarities in High-Dimensional Data Ensembles

A. Heim[1,2] ID, E. Gröller[2,3] ID and C. Heinzl[1] ID

[1] University of Applied Sciences Upper Austria, Austria
[2] TU Wien, Austria
[3] VRVis, Austria

Abstract
Comparative analysis of multivariate datasets, e.g. of advanced materials regarding the characteristics of internal structures (fibers, pores, etc.), is of crucial importance in various scientific disciplines. Currently domain experts in materials science mostly rely on sequential comparison of data using juxtaposition. Our work assists domain experts to perform detailed comparative analyses of large ensemble data in materials science applications. For this purpose, we developed a comparative visualization framework, that includes a tabular overview and three detailed visualization techniques to provide a holistic view on the similarities in the ensemble. We demonstrate the applicability of our framework on two specific usage scenarios and verify its techniques using a qualitative user study with 12 material experts. The insights gained from our work represent a significant advancement in the field of comparative material analysis of high-dimensional data. Our framework provides experts with a novel perspective on the data and eliminates the need for time-consuming sequential exploration of numerical data.

CCS Concepts
• Human-centered computing → Visual analytics; • Applied computing → Physical sciences and engineering;

1. Introduction

For complex material systems such as fiber-reinforced composites (FRP), which are used in safety-critical industries, like automotive or aeronautics, the analysis of the material's performance in terms of durability or strength is essential for quality assurance [NKUC20]. To facilitate the discovery and optimization of novel material systems, detailed knowledge about the internal structure is of utmost importance. FRPs typically consist of a base matrix material and various reinforcements, i.e., the fibers. Among other characteristics of these reinforcements, the placement, length and orientation have a significant influence on the FRP material's properties. Consequently, domain experts are interested in detailed analyses of the respective features in the material, as well as how these features are distributed in terms of their various characteristics [FHG*09]. To achieve comprehensive conclusions about how the various manufacturing and optimization processes affect the materials' properties, domain scientists need to compare the internal structure of different specimens with each other or execute in situ tests. During these tests, a composite is exposed to stress. The changing of its microstructures is recorded in discrete time steps, resulting in time-varying datasets of a specimen (see Figure 1) [NKUC20].

In this work, we refer to the weight, strength, etc. of a specimen, i.e., the sample, as its *properties*. We refer to its inner structures,

such as pores or fibers, as *features* and to the attributes describing the features, like length or orientation, as *characteristics*. In our work, a dataset represents a single specimen, or a part of it, at a particular time step. The quantity of datasets to be compared is considered as an ensemble, where an individual dataset is called an ensemble member (see Figure 1).

The comparison of various table-based datasets is an inherent part of the material scientist's workflow. So far domain experts rely on juxtapositions such as side-by-side views or superpositions of basic charts, as histograms, bar charts, scatterplot matrices (SPLOM), parallel coordinate plots (PCP) etc., to analyze individual characteristics of a specimen. Analyzing several datasets based on these representations can become quite complex for materials scientists. Especially if charts of many specimens have to be explored, the workflow is imposing high cognitive loads to experts.

To support experts in analyzing materials and following the open research challenges in visual computing concerning materials science as outlined by Heinzl and Stappen [HS17], the **goal of this work is to make a comparison of hundreds to thousands of features from dozens of datasets possible by providing CoSi, an interactive visualization framework**. CoSi enables experts to perform a visual **Co**mparison of **Si**milarities of individual features (i.e., fibers or pores) according to their characteristics within an individual dataset. More importantly a comparison across several

specimens and all features at once is possible. We developed CoSi in close collaboration with material experts and we see our main contributions in the following points:

C1: Design study of a visual analysis framework to compare ensembles and ensemble members by using feature and characteristic based similarity, with several key aspects. We provide an overview visualization, the ensemble similarity explorer, that provides a holistic summary of the individual features of the entire ensemble. Interaction is provided to the user via: a linear zooming function, a non-linear zooming interaction, and a ranking operation. For a detailed numerical investigation, a similarity widget offers the users similarity scorings for characteristics via bar charts, as well as their summarized distributions via a box plot. Additionally, we visualize potential linear correlations between the individual characteristics in the correlation widget.

C2: Evaluation of the visual analysis framework is done through two use cases from the X-ray computed tomography (XCT) domain showing the result of an in situ test of a fiber specimen and an in-depth comparison of two pore specimens. Furthermore, a qualitative user study with 12 material experts was performed.

Our paper is organized as follows: Section 2 describes the data structure and tasks. In Section 3 we review the related work. In Section 4 we address our framework CoSi. In Section 5 we describe two usage scenarios and our user study. In Section 6 we provide a discussion and illustrate future work in Section 7.

2. Data Characterization and Task Abstraction

Material ensemble data is generated as follows: XCT data, also referred to as primary data, is acquired from a sample of interest. When the sample is analyzed in an in situ tensile test, XCT scans are triggered consecutively while stepwise increasing the force on the specimen. Each individual XCT image, i.e., time-step, consists of a three-dimensional volume that stores intensities. By applying segmentation and extraction, individual features can be identified and quantification can be used to calculate different numerical characteristics for each feature, resulting in a multivariate, tabular dataset, also referred to as secondary data [WAL*14].The structure of ensemble datasets, generated in an in situ tensile test experiment is shown in Figure 1. Each ensemble, *e1* and *e2*, integrates M members, each consisting of secondary datasets recorded for T time steps. With our framework it is possible to compare different compilations of ensembles, consisting of a specimen over a defined period of time (see Figure 1 A) and of different specimens (see Figure 1 B). Also the combination of both is possible.

Three tasks were identified after several discussions with three material experts about their daily workflow and analysis goals:

T1: Comparative visualization of the similarity among individual ensemble members (feature based similarity) – Domain experts need to evaluate the similarity of the ensemble members in terms of internal structures. For example, they should be able to determine whether the members contain groups of the same type of fibers, such as very short or very long fibers (see Figure 1 **T1**).

T2: Comparative visualization of the similarity among the ensemble member's characteristics (characteristic based similarity) – For groups of similar features it is important to visualize why they are similar. Domain experts require to accurately identify in
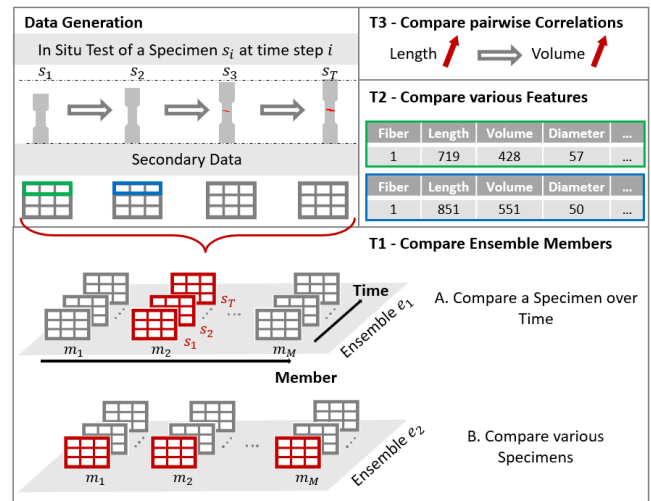

Figure 1: *Data generation and tasks: During an in situ test of a specimen m_j, several time-varying secondary datasets s_i can be computed. Various specimens, with all their time steps, form an ensemble e_k. Our framework is capable of: (T1) comparing different compilations of ensemble members consisting of (A) a specimen over time and (B) various specimens, (T2) the comparison of multiple features, and (T3) the analysis of pairwise correlations.*

which characteristics, like length, orientation, etc., the features are similar or dissimilar (see Figure 1 **T2**).

T3: Visualization of pairwise correlations based on the ensemble members' characteristics – A simple representation is required by material scientists to determine whether the similarity in different characteristics can be inferred, if, for example, features are similar in one specific characteristic (see Figure 1 **T3**).

3. Related Work

In this section, we provide an overview of visualization techniques used in the field of materials science. Few approaches address the visualization of many characteristics or the comparison of different or changing materials. Nevertheless, we reviewed these methods to learn from the experience and advantages of the techniques used. We also examined techniques published in the field of ensemble visualization and comparative visualization to learn from the experiences in different domains.

3.1. Visualization and Analysis for Materials Science Applications

As discussed in the study by Heinzl and Stappen [HS17], a body of research already exists in the field of visualization that is dedicated to improve representations for materials science data. Zhang et al. [ZFS*19] presented an approach to analyze pores in rock formations by applying a segmentation on XCT scans, followed by the classification of porous structures based on their morphology and geometry. In the survey by Hergl et al. [HBK*21], the authors summarized various methods of how to combine tensor information with the specimen's spatial representation. Chiverton

et al. [CIBP17], visualized the arrangement of fibers in concrete by using multiscale entropy to aggregate the orientation and spatial distribution of the structures in the volume. Weissenböck et al. [WFG*19] introduced an analysis tool that allows material scientists to do a voxel-based comparison of many different XCT-datasets. By linearizing the scans with a Hilbert line curve, the differences in the voxel intensities are visualized with Hilbert line plots.

The methods and techniques presented so far have concentrated on the visualization of spatial data or the presentation of a few characteristics in the spatial data. The visualization of many, such as tens to hundreds of characteristics, has not been the primary focus of research to date. FiberScout [WAL*14] is an approach dealing with the visualization of secondary data for a single specimen. To get an overview of all characteristics, a SPLOM is used, and for the orientation distribution a polar plot is shown. The concepts developed for visualizing a few characteristics in a single material cannot be easily adopted to compare multiple samples with many characteristics. This is because the inherent structure of the ensemble data induces an additional member dimension [WHLS19]. To develop suitable visualization techniques, we investigated approaches used in ensemble visualization.

3.2. Ensemble Data Analysis and Visualization

Wang et al. [WHLS19] gave a detailed definition of ensemble data. Although such data has naturally very different meanings in the different domains, the overall workflows of respective approaches follow similar paths: either an aggregation technique is applied on the data prior to the actual data visualization or a visual composition approach is used when the data visualization is performed. The combination of both procedures is also found in a various techniques [WHLS19].

Aggregation. In ensemble visualization, various aggregation methods are used to convert high-dimensional data into a form that can be transformed into visual encodings [WHLS19]. Statistical methods, e.g., from descriptive statistics, can be utilized to achieve a summary of the data. Another method of aggregation is to subdivide the data into groups using cluster-based techniques. When the data does not allow a well-defined subdivision, ambiguous solutions are the result [LMW*17]. To reduce the high dimensionality of the data to that of the visual channels, dimensionality reduction methods can be used. These are classified into linear and non-linear methods. The first type uses linear functions to project high-dimensional data into lower space, while the second methods use non-linear approximations for the projection. Currently popular non-linear techniques are t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP). Both compute clusters of similar features, but do not guarantee that inter-cluster distances are correctly preserved [LMW*17]. In contrast to t-SNE and UMAP, Multidimensional Scaling (MDS) is a method that preserves the global distances of pairs of data points [AHT20].

Comparative Visualization. In ensemble visualization, as well as in the domain of visual comparison of time series data, the comparison of multivariate non-spatial data is typically achieved by line charts, PCPs, SPLOMs, or heatmaps, as well as variants thereof [WHLS19, AAJX19]. PCP and SPLOM based representations of-

ten suffer from overplotting or from scalability issues as they become harder to read the more attributes are visualized. Heatmaps adapt better to large datasets and can be used very well as alignment visualizations, as Albers et al. [ADG11] show in their work. The ease of integrating human perceptual concepts into this type of visualization makes it particularly simple to find similar areas in the data and to identify patterns. Therefore, the design of our analysis framework is based on such visualizations.

4. CoSi

For CoSi, we followed the design study methodology proposed by Sedlmair et al. [SMM12] separated in a preconditioning, a core, and an analysis phase. All methods and techniques have been designed and implemented along these phases and in close collaboration with our domain specialists. To efficiently explore ensemble members regarding specimens of interest, our analytic framework (see Figure 2) is designed to enable a holistic comparison of all ensemble members. The ensemble similarity explorer (see Figure 2 A) presents all ensemble members in a high-level abstraction to allow experts to determine at one glance which members exhibit similar features (**T1**). Through the abstraction applied on the data in this overview visualization, details about the individual features and their characteristics are lost. Our similarity widget (see Figure 2 B) presents this information by providing a similarity rating of the characteristics and exact specifications of groups of features (**T2**). Finally, the correlation widget (see Figure 2 C) visualizes pairwise correlations of the characteristics (**T3**). All three widgets are interactive and linked with each other.

4.1. Ensemble Similarity Explorer

The ensemble similarity explorer is a 2D overview visualization and illustrates a summarized representation of the individual ensemble members. The members are defined by their n-dimensional features (e.g., fibers which are described by n different characteristics) and have to be reduced into a 2D space to apply a visual encoding. Since we want to avoid overlapping of features that do not belong to the same member, we chose the vertical axis (y-dimension) to represent the affiliation to a certain ensemble member and therefore to aggregate the high-dimensional features into the horizontal axis (x-dimension). To generate the ensemble similarity explorer, we first apply an aggregation step. As an unambiguous classification of features such as fibers in subgroups is usually not possible, we decided against a cluster-based approach. Furthermore, we did not use linear dimensionality reduction methods, as we cannot generally assume linearity in material datasets. For our overview visualization, it was important that the aggregation method represents the similarity of features based on the distance between them. Features, that have similar characteristics, like the same length, should be mapped close to each other, while dissimilar features should have a greater distance in 1D space. We renounced from using t-SNE and UMAP, since the distances between the resulting clusters do not encode any similarity information. We chose the non-metric MDS technique, because this method fulfills our requirement that similar features are positioned closer together than dissimilar ones. The MDS computation results in a 1D similarity
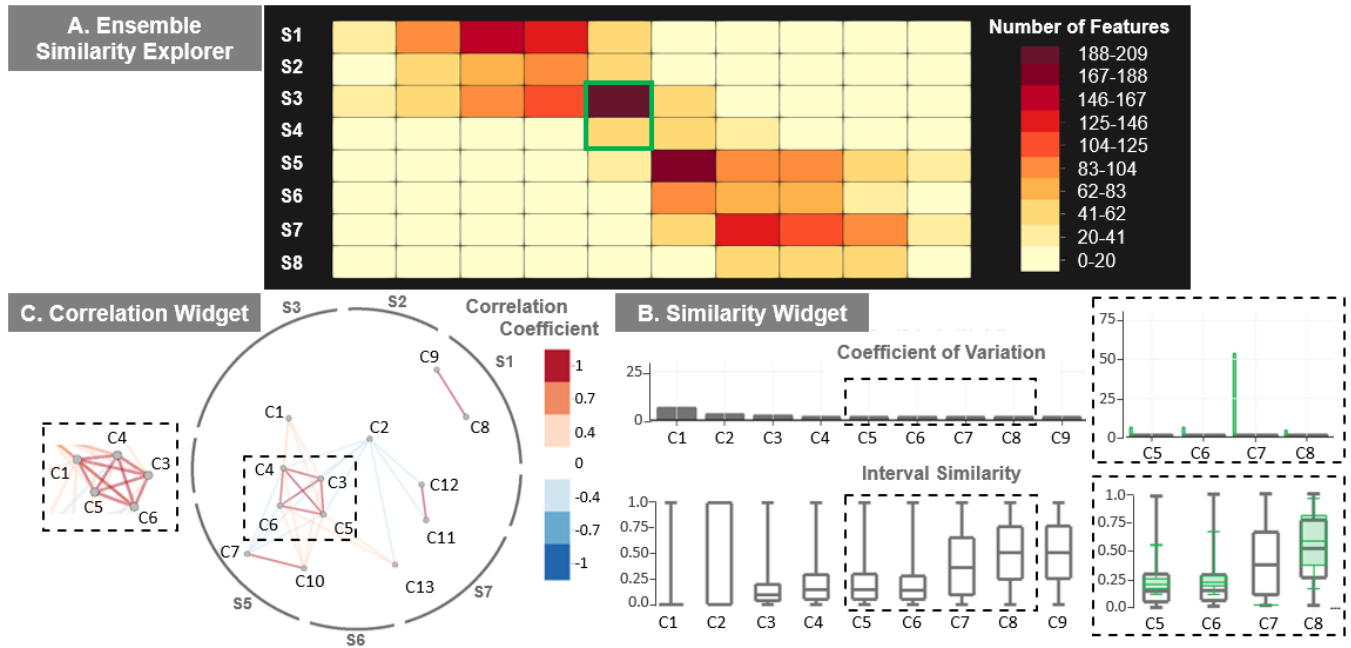
Figure 2: *CoSi Framework: (A) The Ensemble Similarity Explorer shows the similarity of the ensemble members S1 to S8, (B) the Similarity Widget presents the similarity of the characteristics C1 to C9 via a bar chart and box plot, (C) the Correlation Widget reveals pairwise correlations between the characteristics. The black boxes show the changes happening in (B) and (C) when a selection in (A) is performed.*

measure, storing for each feature in the ensemble a single similarity value. All features are ordered according to their attribution to a certain member (vertical axis) and are mapped as points along a 1D horizontal line (horizontal axis), positioned according to their similarity value (Figure 3 B). The different datasets of an ensemble are drawn below each other. This procedure yields a point based representation as overview of the similarities in the complete ensemble. As all datasets are taken into account for the MDS computation at once, features lying close to each other on the same horizontal position are similar within the ensemble member (representing intra-member similarity). Features lying on the same vertical position but in different ensemble members are similar as well (representing inter-member similarity). So, features positioned close to each other are similar (Figure 3 B (light green)), while features at great distance are dissimilar (Figure 3 B (light red)). The numerical values, calculated by the MDS, cannot be interpreted as absolute values, merely the differences between them are meaningful for the analysis. Therefore, we have omitted concrete horizontal axis labels in the visualization to not distract the viewer during the analysis. Due to the distance representing the similarity between the features, the efficient use of available space depends on the current scenario. If members with similar features are compared, the screen space is used efficiently. If a member is very different from the others, the similar members have many features on one side and the dissimilar one on the other side. Since there are thousands of features to be visualized, assigning each item to a specific position leads to visual clutter in the pointbased representation. To avoid overplotting, each line is therefore subdivided into areas of equal size and the number of features inside each area is color-coded. By

Figure 3: *Aggregation Procedure: (A) Each nD feature (red) of the ensemble is aggregated to a 1D similarity value. (B) Features are positioned according to their similarity values and dataset attribution. Light green indicates similar features, while light red illustrates dissimilar ones. (C) Through binning the point representation is composed to a histogram-table to avoid overplotting.*

partitioning the lines into adjacent bins, similar features are connected using colors encoding their frequency. We summarize by binning as this discretization technique ensures that all bins in each row have exactly the same boundaries. This allows the viewer to make a consistent and easy comparison. After binning has been performed, the ensemble similarity explorer represents a histogram heat map, where areas that lie in close proximity are similar while areas located far apart are dissimilar. A sequential color map from yellow to red has been chosen to encode the number of features in

a bin, since it is best suited to identify differences in neighboring areas. Each line representing a dataset in the line based representation is transferred to a row in the histogram heat map. Similar color patterns in different rows reveal that in these members the distribution of features is similar to each other. Figure 2 A displays an ensemble with eight members in the ensemble similarity explorer. The ensemble members *S1*, *S2* and *S3* are similar to each other, while they are dissimilar to the others.

By providing various interaction possibilities, we ensure that the users can inspect different levels of abstraction in the ensemble. In the overview visualization, each row is partitioned into ten individual bins. Even though this coarse subdivision provides a compact overview of the feature distribution, important details are lost. To make an identification of smaller collections of similar features possible, we integrate a linear zooming function with three different levels, where the number of partitions is increased to 20, 40, and finally to 80 bins. Since a comparison across zoom levels should be possible, the same color scheme is retained for each level. In linear zoom mode, all datasets are enlarged or reduced simultaneously. To inspect areas of interest individually, a non-linear zoom mode has been added. Users can select any bin at any zoom level, no matter whether they want to inspect a single area or several areas at once, in the same dataset or in multiple datasets (see Figure 4 C). Individual features located in the selected bins are visualized in a separate area, drawn below the original row. Hence, single bins as well as groups of bins or even complete datasets can be compared in detail. As we are still dealing with hundreds to thousands of selected features, the partitioning into individual areas is also applied to the zoom rows and can be adjusted. In addition to switching between the three zoom levels, users can zoom in on the underlying point representation in the zoom rows, where each individual feature is depicted as a point. As binning causes blockiness and position inaccuracies, these introduced errors should be compensated by the point based representation [ADG11]. In the case of highly similar datasets in an ensemble, it is difficult for the users to visually determine, which of the datasets are most similar to each other. To address this issue, we integrated a statistical computation of the similarity in the datasets. Since the ensemble similarity explorer is a depiction of several histograms, each shown as colored bars, we chose the chi-squared distance metric to measure the similarity between the individual datasets [Cha08]. This statistic measures the difference between the frequencies of a reference dataset and a test dataset which results in a similarity score. As observed by Naik et al. [NPJ09], the chi-square measure does provide very accurate results when comparing very similar multimodal histograms. Since our analysis tool is intended to help determine the similarity of materials that may share a very similar distribution of features, we chose this measure. To start the ranking procedure, users only need to select the dataset according to which the other datasets should be ranked. As a result, the datasets are rearranged, first the reference dataset is drawn, then the datasets follow in descending order of similarity to the reference one.

4.2. Similarity Widget

Strong aggregation by assigning a 1D similarity value to a multidimensional feature, leads to a considerable loss of information. In the ensemble similarity explorer the users can recognize similar groups of features, but they can no longer infer, in which characteristics these groups are similar, or which range of values similar characteristics share (**T2**). Therefore, we added the similarity widget to the CoSi framework. It provides the experts with a similarity score for each characteristic and allows to examine their interval ranges in more detail.

Similarity Score. It is important for experts to be able to perceive at a glance, in which characteristics the similarity is highest. This specification should be evident from a single score per characteristic and its calculation should be easily comprehensible for experts. Therefore, we decided to use the empirical coefficient of variation. This statistic is a relative measure of variation. A bar chart has been chosen to visualize this information, as this chart is considered as one of the most efficient ways to compare multiple 1D values. Moreover, experts are familiar with bar charts, so the learning curve for using our framework is low. The bar chart visualizes the similarity within each characteristic (Figure 2 B). While each grey bar represents an individual characteristic, the heights of the bars represent the similarities in percent, between 0% (dissimilar) and 100% (identical). Once the user makes a selection of one or more bins in the ensemble similarity explorer, the coefficient of variation is recalculated solely on the characteristic's values of the selected features. The result of each characteristic is then superimposed as green bar on the original grey bar (Figure 2 B). Thus, a comparison between the similarity of the selected features and all features in the ensemble can be performed. In addition, the characteristics are rearranged from the most similar to the most dissimilar one. The width of the green bars is variable. It is defined by the ratio between the number of selected features and the number of all features in the ensemble. The fewer features are selected, compared to the total number of features, the thinner the bars are drawn with respect to the grey bars.

Interval Similarity. Furthermore, it is important for experts to get an overview of how the interval ranges of the characteristics are distributed. Methods from descriptive statistics were chosen to compute a statistical summary for each characteristic's distribution, since these are known to the experts. The statistical summary consists of the minimum, the median, the first and third quartile, and the maximum value. These measures give enough information to get an idea of the dispersion and skewness of the distributions. For the visualization of these measures we provided a box plot representation, since this chart is again familiar to most material scientists. We discussed more detailed charts, like violin plots or bean plots with the experts, but they argued that the level of detail is enough and that they would favour the simpler representation of the box plots [TGU20]. The box plots show the statistical summary for every characteristic in the ensemble (Figure 2 B). Since the distributions of the characteristics can have very diverse units, the values of each characteristic are mapped to the interval $[0, 1]$. Thus, all box plots can be displayed side by side in one chart. The order of the box plots is linked to the similarity order of the bar chart, so the position of the characteristics in both charts is the same. The box plots are also recomputed based on selections made in the ensemble similarity explorer. Green box plots, representing the selected features, are superimposed on the original grey box plots, allowing a comparison of the selected features with all features.

4.3. Correlation Widget

Based on the information provided by the similarity widget, experts may address the question of whether features that share one characteristic are also similar in another one (**T3**). In our application, we compute pairwise linear correlations based on the Pearson product-moment correlation coefficient [Coh13]. Since m characteristics result in $O(m^2)$ correlation pairs, visualization techniques such as correlation matrices, can quickly become large and confusing. Although the perception of correlations is improved by encoding the correlation coefficients with color and brightness, Zhang et al. [ZMZM15] argue that position and size are preferable for interpreting quantitative information. Hence, we base our visualization on their work and represent the correlation information in a graph-based layout, called correlation map (Figure 2 C). Each vertex in the correlation map represents a characteristic. It is positioned according to a force-directed layout algorithm. Characteristics that have a strong correlation are positioned close to each other, while characteristics that have no correlation are positioned further away from each other. The edges are color coded according to the type of correlation using a discrete color scheme running from red (positive correlation), to white (no correlation), to blue (negative correlation). This color scheme causes edges with weak or no correlation to become invisible, minimizing the problem of overlapping lines and bringing important correlations into focus. To make all edges visible, the user is able to alter the color scheme by replacing white with grey. The length of the grey circular segments, arranged around the correlation map, represent the ratio of the number of features of each member. Since the correlation coefficient can be significantly affected by the number of features selected for the calculation, the circular segments represent the number of features used for the computation. This visual representation can assist users in selecting an appropriate subpopulation of datasets for a balanced correlation calculation. The correlation map is recalculated after selecting specific bins in the ensemble similarity explorer. Then, the correlation calculation is based solely on the selected features, and the circular segments depict only the selected ensemble members.

5. Results

CoSi is tested on two different ensemble compilations: one for comparing various specimens (Scenario 1, see also Figure 1 B) and one for comparing the changes of a specimen over time (Scenario 2, see Figure 1 A). In Scenario 1, two different samples are compared based on their internal pore structure to determine whether they are similar, and if so, in which region. The ensemble consists of six members. The first member describes the pores of a small material sample. The other five members represent adjacent subregions of one big pore specimen. The small material scan contains roughly 1.700 pores, while each region of the big material scan contains around 1.500 pores. The pores are described by 23 different characteristics.

In Scenario 2 the ensemble describes a fiber reinforced composite specimen which is modified through loading during an in situ test. The material was scanned after a subjection to a shear force for 10 minutes and again after 60 minutes. To facilitate a detailed analysis for the user, the fiber specimen was divided into four areas in each of the two scans. Each of the individual regions contains about 2.500 fibers, which are described by 13 different characteristics.

5.1. Usage Scenarios

Scenario 1 - Compare Various Specimens. At first glance, the two materials *smallMat* and *bigMat* are very similar due to the distribution of their pores (Figure 4 A). Therefore, the ranking function is invoked in the ensemble similarity explorer to find the most similar region of the large material with respect to the small one. It turns out that *bigMat_1* is the most similar one (Figure 4 B). Now, we are interested in finding the group of pores that is predominantly present in *smallMat*. Therefore, we zoom in to the most detailed zoom level and see that there is a larger number of pores located on the right side of the center (Figure 4 C). We select the four dark yellow bins, the nonlinear zoom is activated, and we zoom down to the point representation to check if the pores are clustered in a particular location. However, we find that the pores are fairly evenly distributed across the bins. Now, we want to investigate in which features the selected pores are similar. We therefore take a look at the green bars in the bar chart (Figure 4 D). The pores are most similar in the direction tensors a_{33}, a_{13}, a_{22}, phi, flatness, and volume. Next, we examine the exact value ranges of the characteristics, thus we look at the green box plots. Here, we notice that the direction tensors all take on very small values, additionally these types of pores are very flat and very small in volume (Figure 4 E).
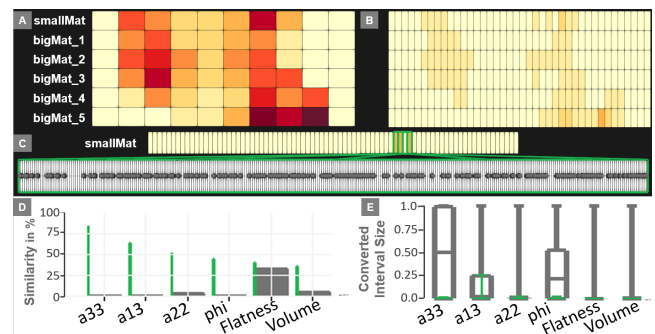

Figure 4: *Scenario 1 - Comparison of two pore materials. An overview of the pore distributions is given in (A) and (B) with different zoom levels. (C) shows the most detailed zoom level of smallMat and the point representation of the pores selected by the user. (D) and (E) give information about the characteristics' similarities.*

Scenario 2 - Compare the Changes of a Specimen over Time. The initial order of the datasets in the ensemble similarity explorer shows all regions of the scan taken after 10 minutes and then all regions of the scan taken after 60 minutes of shearing (Figure 5 A). Each region contains a very specific group of fibers, as only two bins in each member are darker in color. We are interested in region _4 since it contains the most fibers in a bin. We activate the ranking function to find the dataset that is most similar to *10min_4*. As suspected, the most similar region is *60min_4*, which represents the same region at a later time (Figure 5 B). We can see a change in the distribution of the fibers, as many of them have changed their position by one bin to the left. We now want to investigate, which char-

acteristics have changed to produce this shift in position. Therefore, we first select the darkest bin in *10min_4*, and then the darkest bin in *60min_4* (Figure 5 C). The comparison of the box plots shows the characteristics volume, surface area, and straight length (Figure 5 D). According to the maximum and median marks, these characteristics have become shorter over time. Finally, we look at the correlation map (Figure 5 E). There, these three features are positively correlated, as they are very close to each other and are connected by red edges.
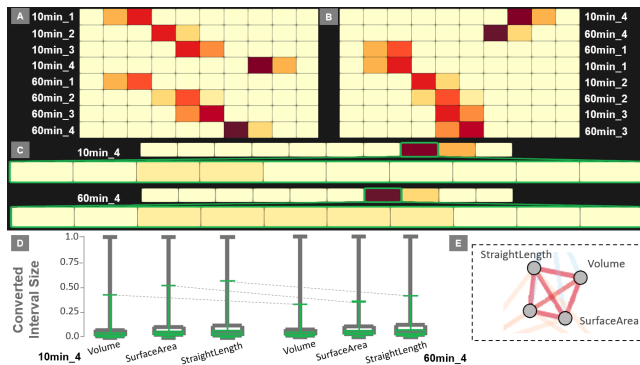

Figure 5: *Scenario 2 - Comparison of four regions of a fiber material after being exposed to a force for* 10 *(10min_X) and* 60 *(60min_X) minutes. (A) gives an overview of the fiber distributions. (B) shows the ensemble after a ranking was performed according to 10min_4. (C) presents the bins that were selected successively by the user. (D) shows the box plots illustrating that the fibers in 60min_4 have smaller maxima in volume, surface area, and straight length. (E) shows the correlation map.*

5.2. Evaluation

Procedure. To evaluate the comprehensibility and applicability of our analytical framework, we conducted a qualitative user study with 12 material experts, who study the microstructures of polymer materials. We began the study by introducing CoSi to each participant in a 10-minute demonstration, explaining how to interpret the visualizations. The participants were given 20 minutes to explore the ensemble from Section 5.1. Meanwhile, the participants were asked to explain how they interpreted each representation. We prepared qualitative tasks in advance, which are based on the tasks **T1-T3**, to ensure that all interactions and visual encodings were observed. Figure 6 shows the defined tasks and whether they were answered correctly, partially correctly, or wrongly.

Results. The ensemble similarity explorer was found to be an intuitive tool to get an overview on complex ensembles. The valuation is supported by the number of participants who correctly understood this visualization technique (Figure 6 (1)). Merely two participants could not recognize which datasets were similar to each other. They had difficulties in understanding the strong abstraction of the data. The bar chart was rated as helpful, since it was possible to see at a glance, in which characteristic the fibers were most similar. This positive feedback is reflected in the results of Task 2 (Figure 6 (2)) and Task 3 (Figure 6 (3)). Two participants only partially solved Task 2, since they claimed to "prefer to look at the box
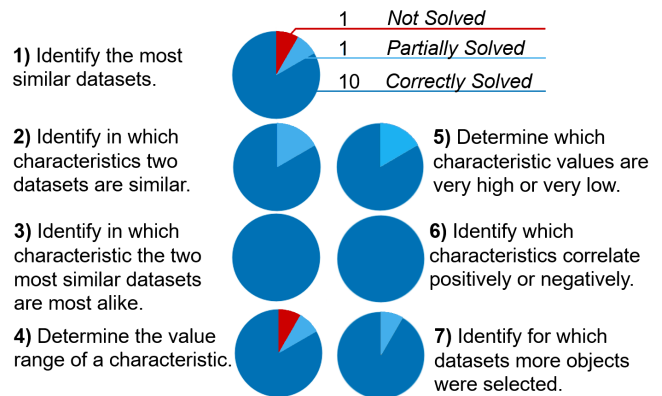

Figure 6: *Defined tasks and their correctness.*

plot rather than the bar chart, as it shows way more information". Besides the ensemble similarity explorer, the box plot was the visualization that participants were most interested in. The readability of the box plot was strongly dependent on the previous experience of the test persons. Task 4 (Figure 6 (4)) and Task 5 (Figure 6 (5)) could be solved without problems. The correlation map received the best feedback concerning its visual encoding. According to the participants, the color coding of the linear correlations was intuitive, which is confirmed by the good results concerning Task 6 (Figure 6 (6)) and Task 7 (Figure 6 (7)).

Performance Measurements. CoSi was developed as module for the open-source application open_iA [FWS*19]. Both usage scenarios were evaluated on the same test setup, i.e., a laptop equipped with an Intel i7-6820HK CPU with 16 GB RAM and a screen size of 17 inches. A total of about 20.000 features were compared in each case. The calculation of the representations took about 15 minutes for each scenario. A bottleneck with respect to memory consumption and runtime arises from the use of the dimension reduction method MDS, which is based on the SMACOF algorithm [dLM09]. To calculate the pairwise distances between s features, a matrix of dimension s^2 is required, which is why the number of features to be compared is currently limited by the size of the RAM. The computation time is bound to the computation time of the SMACOF algorithm, which is $O(s^3 \times k)$, where k is the number of iterations. These performance issues could be solved, by using a more effective implementation of the MDS [Bae08].

6. Discussion

Reflection of the Method. Throughout the development of CoSi, we regularly reviewed the mock-ups of each widget with the material experts. We discussed with all participants whether the strong data reduction would make an exploration problematic. All agreed that CoSi would mainly be used to give an overview of the data, so for them the data reduction was appropriate. One respondent noted that "This tool is a great work relief because I no longer have to look at the datasets individually to group the materials. It helps enormously to make a pre-selection of the data. Instead of two hours I can now perform this task within a quarter of an hour". During the evaluation, we noticed that the experts were using the tool not

only to compare features across members, as it was intended by our tasks, but also to examine groups of features within a single dataset in more detail.

Scalability. The ensemble similarity explorer can handle datasets consisting of a large number of characteristics and features, since all are aggregated into one similarity value. The only limiting factor is the size of the memory. Visualizing a large number (~50) of members is possible, as the ensemble similarity explorer is a space-filling table visualization, where each member is assigned to a single row. Since the colored patterns of the rows indicate the similarity of the members, the height of the rows can be reduced to a certain extent, so only the height of the screen limits the number of ensemble members.

Generalizability. Our framework can handle high-dimensional data ensembles from all kinds of disciplines, since our computational methods are not based on specific domain-related information. Any data ensemble containing features with a set of numerical attributes can be loaded into CoSi (see supplemental material).

7. Conclusion and Future Work

We presented CoSi, a visual analysis framework for the comparison of material data ensembles. We focus on the visualization of similarities between various ensemble members at different levels of detail. We evaluated the functionality with two usage scenarios and conducted a quantitative evaluation with domain experts. In our approach, the focus was on the comparison of selected features with all features of the ensemble. But a comparison among the selected groups of features would also be of importance and is planned as future work.

References

[AAJX19] ALI M., ALQAHTANI A., JONES M. W., XIE X.: Clustering and Classification for Time Series Data in Visual Analytics: A Survey. *IEEE Access 7* (2019), 181314–181338. 3

[ADG11] ALBERS D., DEWEY C., GLEICHER M.: Sequence Surveyor: Leveraging Overview for Scalable Genomic Alignment Visualization. *IEEE Transactions on Visualization and Computer Graphics* (2011), 3, 5

[AHT20] AYESHA S., HANIF M. K., TALIB R.: Overview and Comparative Study of Dimensionality Reduction Techniques for High Dimensional Data. *Information Fusion 59* (jan 2020), 44–58. 3

[Bae08] BAE S.-H.: Parallel Multidimensional Scaling Performance on Multicore Systems. In *2008 IEEE Fourth International Conference on eScience* (dec 2008), IEEE. 7

[Cha08] CHA S.-H.: Taxonomy of Nominal Type Histogram Distance Measures. In *Proceedings of the American Conference on Applied Mathematics* (Stevens Point, Wisconsin, USA, 2008), MATH'08, World Scientific and Engineering Academy and Society (WSEAS), p. 325–330. 5

[CIBP17] CHIVERTON J. P., IGE O., BARNETT S. J., PARRY T.: Multiscale Shannon's Entropy Modeling of Orientation and Distance in Steel Fiber Micro-Tomography Data. *IEEE Transactions on Image Processing 26*, 11 (nov 2017), 5284–5297. 3

[Coh13] COHEN J.: *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, 2013. 6

[dLM09] DE LEEUW J., MAIR P.: Multidimensional Scaling using Majorization: SMACOF in R. *Journal of Statistical Software 31*, 3 (2009). 7

[FHG*09] FRITZ L., HADWIGER M., GEIER G., PITTINO G., GRÖLLER E.: A Visual Approach to Efficient Analysis and Quantification of Ductile Iron and Reinforced Sprayed Concrete. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (nov 2009), 1343–1350. 1

[FWS*19] FRÖHLER B., WEISSENBÖCK J., SCHIWARTH M., KASTNER J., HEINZL C.: open_iA: A Tool for Processing and Visual Analysis of Industrial Computed Tomography Datasets. *Journal of Open Source Software 4*, 35 (mar 2019), 1185. 7

[HBK*21] HERGL C., BLECHA C., KRETZSCHMAR V., RAITH F., GÜNTHER F., STOMMEL M., JANKOWAI J., HOTZ I., NAGEL T., SCHEUERMANN G.: Visualization of Tensor Fields in Mechanics. *Computer Graphics Forum* (mar 2021). 2

[HS17] HEINZL C., STAPPEN S.: STAR: Visual Computing in Materials Science. *Computer Graphics Forum 36*, 3 (jun 2017), 647–666. 1, 2

[LMW*17] LIU S., MALJOVEC D., WANG B., BREMER P.-T., PASCUCCI V.: Visualizing High-Dimensional Data: Advances in the Past Decade. *IEEE Transactions on Visualization and Computer Graphics 23*, 3 (mar 2017), 1249–1268. 3

[NKUC20] NARESH K., KHAN K., UMER R., CANTWELL W. J.: The Use of X-ray Computed Tomography for Design and Process Modeling of Aerospace Composites: A Review. *Materials & Design 190* (may 2020), 108553. 1

[NPJ09] NAIK N., PATIL S., JOSHI M.: A Scale Adaptive Tracker Using Hybrid Color Histogram Matching Scheme. In *Second International Conference on Emerging Trends in Engineering and Technology, ICETET-09* (2009), IEEE. 5

[SMM12] SEDLMAIR M., MEYER M., MUNZNER T.: Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (dec 2012), 2431–2440. 3

[TGU20] THRUN M. C., GEHLERT T., ULTSCH A.: Analyzing the Fine Structure of Distributions. *PLOS ONE 15*, 10 (oct 2020). 5

[WAL*14] WEISSENBÖCK J., AMIRKHANOV A., LI W., REH A., AMIRKHANOV A., GRÖLLER E., KASTNER J., HEINZL C.: FiberScout: An Interactive Tool for Exploring and Analyzing Fiber Reinforced Polymers. In *2014 IEEE Pacific Visualization Symposium* (mar 2014), IEEE. 2, 3

[WFG*19] WEISSENBÖCK J., FRÖHLER B., GRÖLLER E., KASTNER J., HEINZL C.: Dynamic Volume Lines: Visual Comparison of 3D Volumes through Space-Filling Curves. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (jan 2019), 1040–1049. 3

[WHLS19] WANG J., HAZARIKA S., LI C., SHEN H.-W.: Visualization and Visual Analysis of Ensemble Data: A Survey. *IEEE Transactions on Visualization and Computer Graphics 25*, 9 (sep 2019), 2853–2872. 3

[ZFS*19] ZHANG H., FREY S., STEEB H., URIBE D., ERTL T., WANG W.: Visualization of Bubble Formation in Porous Media. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (jan 2019), 1060–1069. 2

[ZMZM15] ZHANG Z., MCDONNELL K. T., ZADOK E., MUELLER K.: Visual Correlation Analysis of Numerical and Categorical Data on the Correlation Map. *IEEE Transactions on Visualization and Computer Graphics 21*, 2 (feb 2015), 289–303. 6