VisGap'24 – The Gap between Visualization Research and Visualization Software (2024)
C. Gillmann, M. Krone, G. Reina, T. Wischgoll (Editors)

Visual Scalar Matrix Evaluation: An Application to Thermodynamics

Jan-Tobias Sohns[1], Dominik Gond[2], Fabian Jirasek[2], Hans Hasse[2], Heike Leitte[1]

[1]University of Kaiserslautern-Landau, Visual Information Analysis Group, Germany
[2]University of Kaiserslautern-Landau, Laboratory of Engineering Thermodynamics, Germany

Figure 1: *Exploring Mixture Data: A sorted heatmap visually groups blocks of similar chemical substances (rows and columns). Pattern strength is analyzed by variation in internal and external data. Linked widgets connect the discovered groups to additional domain knowledge.*

Abstract

Modeling and predicting thermodynamic properties of binary mixtures is crucial in chemical engineering. Understanding how the mixture behavior, represented as a scalar matrix, depends on properties of pure substances offers valuable insights into substance interactions. While there is robust support for pattern-based sorting of matrices in general, limited support exists for evaluating patterns against external attributes available in many fields. In this paper, we introduce an interactive software to detect and analyze block patterns in scalar matrices using annotated domain knowledge. Therefore, we revisit canonical matrix patterns, explore their translation to this application, and describe a workflow to fit the matrix ordering. Our interactive software allows users to explore hierarchical aggregation levels, rating them based on additional domain-specific data properties of various type. Using our tool, chemical engineers are able to identify and interpret cluster structures in their mixture data. These insights contribute to the development of improved prediction methods for thermodynamic properties, forming the foundation for modeling and simulation in chemical engineering.

CCS Concepts

• ***Human-centered computing*** → *Heat maps; Dendrograms;* • ***Applied computing*** → *Chemistry; Engineering;*

DOI: 10.2312/visgap.20241121

1. Introduction

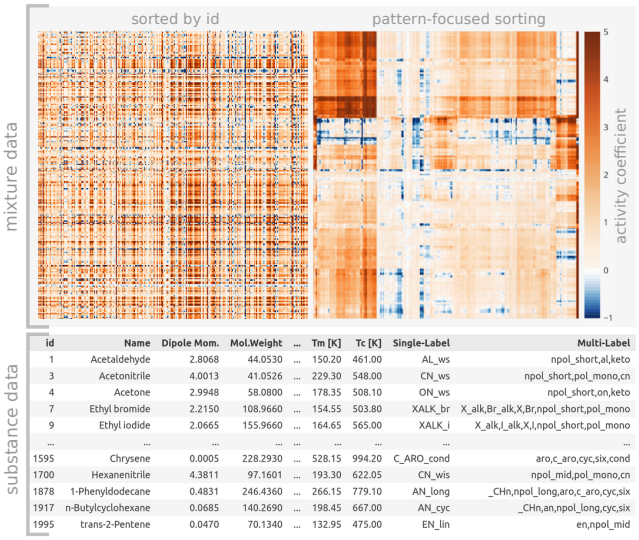
The analysis of matrix data is of paramount importance in application domains like graph theory, biology, engineering, sociology, or archaeology [Lii10]. Heatmaps are a central inspection tool in application domains as they directly visualize scalar matrix data itself without the need for abstractions. Patterns evident in the heatmaps offer data-driven understanding of the interrelation between row and column elements. As the observable patterns are inseparably dependent on the ordering, we describe a workflow for pattern-focused assessment of scalar matrix orderings on thermodynamic application data.

Application domain data is usually not constrained to only the matrix elements, but each row or column can further be described by features differing in data type or scale from the matrix elements. Contrasting the matrix elements with these additional feature values is called enrichment analysis and can provide crucial insights about the quality or relationships within or between matrix patterns. While extensive support for general heatmap visualization exists through domain specific tools [FGR*17] and common plotting libraries, the support for enrichment analysis is often limited to the application case. The enrichment is either too domain-specific [LSKS10] or supports only minimal data types [FGR*17]. As the sole purpose of enrichment is to compare it to patterns in the matrix, the focus should be on matching the two data sets to judge the bidirectional quality of identified patterns. We therefore present an interactive software that links matrix patterns to views of common enrichment data types as well as propose validation measures to guide the analysis.

2. Application Background

Modeling and predicting thermodynamic properties of mixtures is of paramount importance in chemical engineering, since their knowledge is the basis for design and optimization of processes in many industries, like chemical, pharmaceutical, and biotechnological industry. The vast majority of prediction methods rely on features of the pure substances that make up the mixtures [JH21], e.g., their composition with regard to structural groups as in so-called group-contribution methods [GCS15]. A better understanding of the relationship between properties of pure substances and mixture properties holds the potential of significantly improving present and future prediction methods, with a direct impact on process design and optimization in chemical engineering. With our software, we address this challenge by analyzing the activity coefficient of binary mixtures, a measure that describes the deviation from ideal mixture behavior. While the software is currently visually slightly specialized, it is applicable to any other domain with similar data types.

The data considered for the analysis falls into two categories: a scalar property of substance mixtures, i.e. the heatmap, and properties of pure substances, the enrichment. The data for binary *mixtures* can conveniently be arranged in matrices. The heatmaps in Figure 2 (top) represent one such thermodynamic property of binary mixtures, namely the activity coefficient of a solute (row) at infinite dilution in a solvent (column). The left heatmap is sorted by substance IDs—for comparison, the right one is sorted to show



	id	Name	Dipole Mom.	Mol.Weight	...	Tm [K]	Tc [K]	Single-Label	Multi-Label
substance data	1	Acetaldehyde	2.8068	44.0530	...	150.20	461.00	AL_ws	npol_short,al,keto
	3	Acetonitrile	4.0013	41.0526	...	229.30	548.00	CN_ws	npol_short,pol_mono,cn
	4	Acetone	2.9948	58.0800	...	178.35	508.10	ON_ws	npol_short,on,keto
	7	Ethyl bromide	2.2150	108.9660	...	154.55	503.80	XALK_br	X_alk,Br_alk,X,Br,npol_short,pol_mono
	9	Ethyl iodide	2.0665	155.9660	...	164.65	565.00	XALK_i	X_alk,I_alk,X,I,npol_short,pol_mono
	...	...	...	...	...	...	...	...	...
	1595	Chrysene	0.0005	228.2930	...	528.15	994.20	C_ARO_cond	aro,c_aro,cyc,six,cond
	1700	Hexanenitrile	4.3811	97.1601	...	193.30	622.05	CN_wis	npol_mid,pol_mono,cn
	1878	1-Phenyldodecane	0.4831	246.4360	...	266.15	779.10	AN_long	_CHn,npol_long,aro,c_aro,cyc,six
	1917	n-Butylcyclohexane	0.0685	140.2690	...	198.45	667.00	AN_cyc	_CHn,an,npol_long,cyc,six
	1995	trans-2-Pentene	0.0470	70.1340	...	132.95	475.00	EN_lin	en,npol_mid

Figure 2: *Input data: Each row and column in the heatmap represents the mixture properties of a substance that has additional properties shown in the bottom table. (top left) A heatmap sorted by substance ID does not permit pattern analysis. (top right) The same data sorted by row and column similarity reveals patterns in mixtures that shall be related to pure substance properties.*

visible patterns. Uniform regions in the matrix indicate that the respective solutes (solvents) are similar with regard to the activity coefficients.

Properties of pure substances as represented in the table in Figure 2 (bottom) enrich each row/column of the heatmap. We call them *substance features* and distinguish into two types here: (1) rigorous properties, which are measurable or unambiguously deducible from the molecular structure of the substance, and (2) more indistinct descriptors, which are defined based on experience or chemical intuition. As properties of type (1), we consider dipole moment, molar weight, polarizability, anisotropy, normalized anisotropy, relative number of H-bond acceptors, relative number of H-bond donors, normal melting temperature (Tm), and critical temperature (Tc), which are all of numerical type. As (rather subjective) descriptors of type (2), we consider the affiliation of a substance to a chemical class, like 'branched alkanes', 'cyclic alkanes', or 'heteroaromatics', which are single-label classes, as well as attributes defined based on the molecular structure, like 'cyclic', 'aromatic', or 'long-chained', which are multi-label classes.

The activity coefficient is essential for chemical process engineering in practice, but was, thus far, hard to predict precisely. Due to recent advances [JBM20], the data was expanded to more than 50k binary mixtures, 234 solutes and 214 solvents. This novel approach generates mixture data from other mixtures rather than relying on substance-driven methods [GCS15]. Hence, it enables unbiased comparison between mixture and enriched substance properties. Correlation with measurable properties like dipole moment may provide guidance for future prediction methods. Moreover, aligning mixture groups with established class labels like heteroaromatics can assess their suitability for prediction techniques.

3. Requirement Analysis

The present work results from a cooperation between the Information Visualization Group and the Laboratory of Engineering Thermodynamics from the University of Kaiserslautern-Landau and includes experts in data analysis, visualization, thermodynamics, chemical engineering, and chemistry. We loosely followed the design study methodology proposed by Sedlmair et al. [SMM12]. In the *discovery phase*, the groups exchanged necessary domain knowledge and collected the data. A central pillar of the *design phase* were continually improved prototypes that were build using accessible visualization tools like seaborn [Was21] and holoviews [Ruda], which include necessary visualization and data aggregation tools like clustermap [Was21], i.e., heatmap visualization with associated cluster trees, and statistical aggregates of the data like grouped violin plots. These early prototypes helped us build a common ground for communication, explore shortcomings of existing solutions, and refine the 'wish list' of the domain experts.

During the joint discussions of these prototypes, we made the following design observations: Central features are *access to raw data and use of established chart types*. Our work centers around pattern mining in complex data. Using chart types that are familiar to the user and represent the raw data makes it easier to judge the effect of the automatic analysis routine. For example, we stick to a clustermap as the central plot. *Flexibility of the software* is another important aspect. During the progressive data analysis, we realized that the data basis is not fixed. New insights may require the integration of new substance properties. Hence, a flexible design is necessary that can adapt to arbitrary numbers of descriptors and new data types. *Linked interactive filtering* in multiple/all directions discloses relationships between views. In each prototype iteration, the users intuitively tried to experiment with linked selections first. Thus, we made interaction and linking central components.

The main challenge in mixture prediction is that most methods rely on implicit *knowledge of experts*. With today's rising availability of data problem-driven visualizations should move the information location towards *data*- and therefore *computer*-driven approaches [SMM12]. Thus, the central goal of our work is finding data-driven patterns in mixture data and linking them to properties of pure substances. Likewise, we are also interested in breaks in expected patterns, i.e., if substances that belong, according to chemical intuition, to the same chemical class show few similarities. In a nutshell: we aim at understanding what similarity among substances actually means, but with regard to their behavior in mixtures. Up to now, no software is available in this field to answer these questions and current research in thermodynamics basically depends on manual work of experienced physical chemists.

4. Related Work

Research directions on pattern analysis in matrices are threefold: Patterns are either defined within the matrix, on the tree that constitutes an ordering, or the distribution of annotated attributes in the ordered matrix.

Matrix Patterns For observable patterns in symmetric binary matrices, Behrisch et al. provide a comprehensive analysis in Magnostics [BBH*17]. However, they state that the defined patterns are specific to symmetric matrices, which are commonly sparse and binary, and do not generalize to data tables [BBH*16]. Wilkinson [Wil05] describes canonical data patterns observable in general heatmaps — asymmetric scalar matrices. For patterns found in real applications, we argue in subsection 5.1 that Lekschas et al. [LBK*18] present a more suitable description of patterns, although their approach is focusing on small recurring motifs in symmetric scalar matrices.

Tree Patterns To visualize patterns in a tree, Parthl et al. [PLS*13] classify four options to visualize attributes: directly on the node; small-multiples of the graph; linked views of graph and attributes; or adaption of graph layout. Small-multiples work for comparing attributes in graphs [BMGK08], but do not match with the strict order of a matrix. Degree-of-Interest trees aggregate the tree-layout depending on a function of interest [CN02]. While initially interest was defined over interaction with a node, Lineage [NGCL19] extends this idea to attributes in genealogy. They propose several strategies for a binary Degree-of-Interest, which we extend to a continuous measure of node variation. Chen et al. [CMP10] show that the interactive exploration of matrices over a dendrogram provides insight. Combining the benefits of Lineage [NGCL19] for on-node mapping and the linked views of GAP [WTC10] and GUIRO [BSP20], we derive our design for annotating external attributes in section 6.

Enrichment Patterns Heatmap literature typically allocates minimal effort to enriching heatmaps with external attributes. Notable exceptions include VIS-STAMP [GCML06], incorporating linked parallel coordinates and a map for context and filtering, and Lex et al.'s genealogy-specific system [LSKS10], which links domain-specific views to a sorted 2.5D heatmap. Clustergrammer [FGR*17] stands out as a recent, well-implemented paradigm for heatmap plotting, adding color-coded columns for categorical features. Additional data is accessible via hyperlinks to open databases. In HiPiler [LBK*18] individual matrix snippets are enriched via border colors, though this does not transfer well to many categories or continuous distributions. Since our annotated data drastically exceeds the number of distinguishable class colors or is continuous, we opt for aggregating label variation and visualizing variable distributions rather than directly relying on color-coded attribute values.

5. Method

To answer the questions described in section 3, we propose a three-step workflow. We start with an overview of practically observable patterns in scalar matrices in subsection 5.1 and continue with a block-focused assessment of ordering techniques in subsection 5.2. We close with augmentation strategies for validating these patterns with external data in subsection 5.3. Subsequently, we discuss the design of suitable supporting plots to identify relationships with enrichment data in section 6.

5.1. Matrix Patterns in Scalar Asymmetric Matrices

The solute-to-solvent ratio in mixtures is generally not interchangeable, e.g., you would not solve water in salt. Hence, there is a one-way relationship between each pair of solute and solvent and the

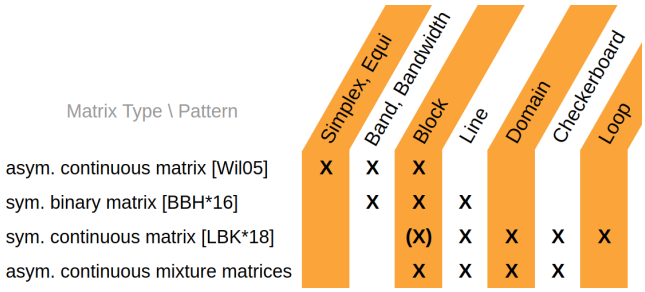


Figure 3: *The observable patterns depend on the type of matrix (rows). The columns mark which patterns have been (indirectly) described for each type.*

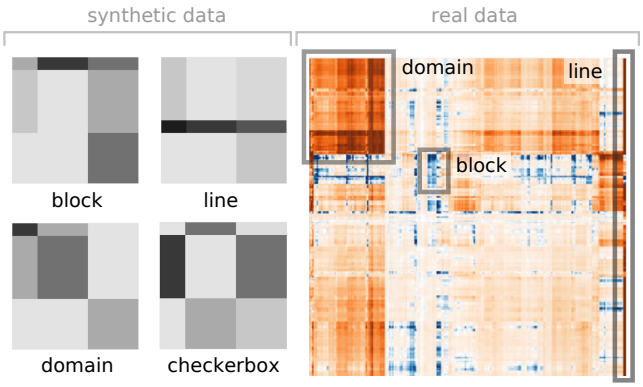


Figure 4: *Patterns in asymmetric matrices: (left) Four patterns are identified in asymmetric matrices. (right) In real data, the patterns feature various degrees of expressiveness.*

corresponding matrix is inherently asymmetric. From the established matrix patterns summarized in Figure 3, we therefore deduce the ones that are applicable to our asymmetric practical data.

We start with established patterns that do not transfer to our use case. Simple and Equi patterns [Wil05] are assuming a uniform global correlation unlikely to be found in real data. Similarly, the Band/Circumplex and Bandwidth patterns [Wil05; BBH*16] are only relevant if the diagonal holds meaningful information, i.e. for symmetric matrices. Loops [LBK*18] are only sensible in adjacency matrices.

Block The most common pattern throughout applied literature is the block pattern, where coherent rectangular areas appear in the sorted matrix. A block can be described by a range of row- and column-IDs and a corresponding scalar value. Due to noise in real data, the block area is commonly not as uniform as in the synthetic example in Figure 4 (top-left). A block denotes that a number of entities (set of rows) share similar values in a number of features (set of columns). For binary mixtures, this pattern denotes that the related set of solutes exhibits similar activity coefficients in any solvent from the respective set, i.e. they form a common class.

Line The line pattern is a special variant of the block pattern. Here, a single row or column features extremal values that set the data points apart. Multiple similar entities/features can exist. If ordered accordingly, they will form a long narrow block spanning multiple rows/columns. Lines are entities that are highly dissimilar from any other data point therefore indicating outliers. In binary mixtures, water plays such a special role as it leads to extreme activity coefficients if mixed with different components as can be seen in Figure 4 (right).

Domain In a domain, a larger block region contains additional sub-blocks where the color is darker or lighter, as represented in Figure 4. Like blocks, domains can be characterized by their respective row and column IDs. The nesting structure requires a hierarchical description model. We found cluster trees suitable to describe the nesting structure and the distinctness of sub-blocks. From a chemical-engineering perspective, we can interpret this type of pattern as a rather large group of solutes and solvents that mix similarly with another group, while there are subgroups that are even more similar.

Checkerboard Checkerboard patterns are an arrangement of globally alternating blocks of high and low values. While they are a common phenomenon in gene expression data, they are barely visible in mixture data, indicating groups of high intra- but low inter-activity.

We conclude that the occurring patterns can all be described as (hierarchical) blocks, which have been shown to be sufficient for heatmap interpretation [Che02; WTC08] and are the most common in practical data [FGR*17; EWJP17]. However, for any pattern to be visible, the order of a matrix is of tremendous importance.

Matrix Reordering Algorithms As manual reordering is too tedious for real datasets, we resort to a choice of reordering algorithms. In Behrisch et al.'s [BBH*16] extensive evaluation of available algorithms focused on block-diagonal patterns, they conclude that hierarchical clustering, specifically optimal leaf-ordering [Bra07], excels at producing local patterns. That is even though hierarchical clustering is intended to cluster, not to induce a global linear order on matrix rows [Lii10]. Recent studies on effectiveness of ordering algorithms in continuous matrices also suggest that Robinsonian and machine learning techniques are best to detect block patterns [BdS22]. Since we further aim to capture hierarchical domain patterns, we opt for hierarchical clustering with optimal leaf-ordering — a Robinsonian technique that is long-established in biological contexts [ESBB98; SS02; CP04; WTC10].

5.2. Evaluation of Blocks in Matrices

Hierarchical clustering has two parameters: a metric that defines the distance between two rows/columns, and a linkage type that defines the distance between two clusters. The choice of these parameters is crucial for the clustering result and therefore the matrix ordering. While all common combinations can be explored in our tool, we describe a generally applicable workflow to compare hierarchical clusterings regarding their ability to uncover block patterns.

The strength of a block pattern can be quantified over its uniformity. As a well-established and therefore readily understood measure of uniformity in a dataset, we use the standard deviation of values within a block. Alternative choices with similar results are mean square error or mean absolute error. The two dendrograms defined by hierarchically clustering rows and columns can be pruned

at any point, partitioning the matrix into blocks. We express the quality of such a partitioning by the average uniformity of each individual block. To account for the relative importance of each block, we weight the score of each block with the number of contained elements, then average over all blocks. Since we perform separate clusterings on each of the axis, separating into 1 to n clusters per axis leads to n^2 numbers of possibilities for block layouts. We assume that a matrix ordering that forms more distinct block patterns with the same number of blocks is preferable.

To find the default parameter setting for our tool, we analyzed the induced matrix orderings for all combinations of common distance measures (Euclidean, Manhattan, cosine) and linkage parameters. Single, average, and complete linkage minimize the shortest, average, and longest distances between any two points in a cluster, respectively. In Figure 5 (top) we present three charts for Euclidean distance with different linkage strategies. The x- and y-axis denote the number of clusters, and the color-value indicates the quality score for this partitioning. The standard deviations for single linkage are consistently higher than those for complete- and average linkage.

We note that the standard deviations decrease along a path from bottom-left (single block) to top-right (finest granularity). To ease comparability, we reduce the heatmap to a line in Figure 5 (center). Therefore, we choose a suitable path through the heatmap that captures the fastest decline in standard deviation. In our experiments, a suitable path occurred along the diagonal, though the path could be shifted or bent for more asymmetric data. We observe that complete linkage results in consistently the lowest standard deviations for all three metrics. Average linkage contains characteristic drops (marked in pink) that are also visible in the 2D plot, which indicate strong local changes in the block quality. Euclidean and Manhattan distance showed equally low scores in our tests. We chose Euclidean(-complete) as our default, as it is the most common. To verify our result, we search for the 'elbow' in the line for Euclidean-complete (solid orange), which gives us the Pareto optimum, the best trade-off between minimized number of clusters and low error rates. We find it at approx. 7–10 clusters. The partitioned matrix with 10 clusters each is shown in Figure 5 (bottom). Coloring the blocks by their mean (left) partitions the heatmap into regions of predominantly high or low activity coefficients. The distribution of the standard deviation (right) provides insight into the error rates within each block.

Note that it is crucial to keep the variation comparable across rows and columns, i.e., the matrix data needs to be standardized. In our application case, that was already achieved by having the same measure for all data points.

5.3. Validating Patterns using Domain Knowledge Variation

Finally, we want to guide the interpretation of patterns based on enriched substance properties; among others, we thereby want to identify the subset of properties that characterizes the substances (rows and columns) the best with regard to their mixture behavior (matrix entries). Clusters where domain knowledge matches with the similarity in the matrix signal a correlation, which is usually considered interesting [FGR*17; RSW*19]. On the other hand,

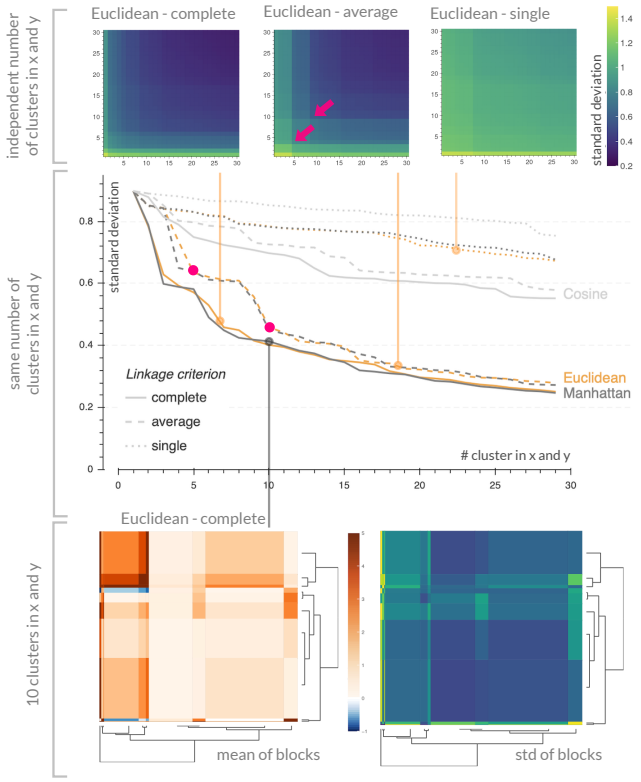


Figure 5: *Effects of clustering parameters: (top) Comparison of linkage criteria for arbitrary combinations of cluster granularity in x- and y-direction. (center) Comparison of clustering parameters for same granularity in x and y. (bottom) Effects for division into 10 clusters in x- and y-direction.*

clusters that do not match with domain knowledge are even more interesting, since they can spark new ideas for undiscovered relationships. A clustering is considered to be matching the domain knowledge, if both approaches group the same annotated elements together. Hence, we consider the variation of associated domain knowledge within a cluster as its validation score. This concept is illustrated in Figure 6, where the inner nodes of the dendrograms are colored based on the validation score of user-selected substance descriptors.

As we have seen in section 3, substance descriptors come in three data types: single-label (chemical class), multi-label (composition with regard to functional groups), and numerical (measurable and deducible quantities). This also covers most of the datatypes potentially occurring in other application domains. The direct color-coding of descriptors is limited by available colors and screen space [FGR*17; RSW*19]. Hence, we recommend dedicated scalar measures for these three types of data.

For the **single-label** case, we suggest *entropy*:

$$e_{\text{single}} = -\sum_{i=1}^{n} p_i * \log_2(p_i) \tag{1}$$

where p_i is the probability of label i in a given cluster. Entropy is maximal for uniformly distributed data and increases with in-

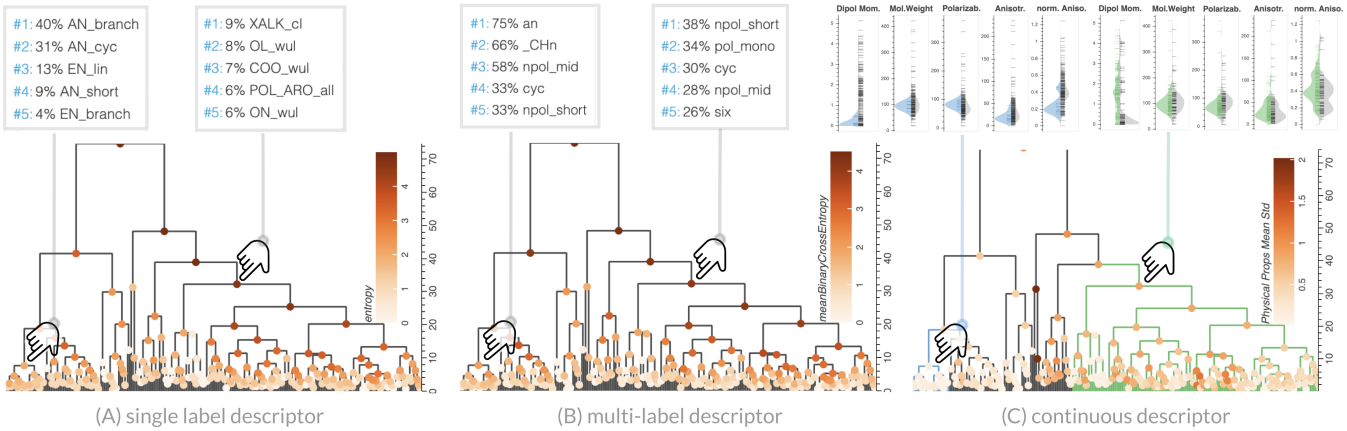


Figure 6: *Interactive dendrograms for cluster analysis and refinement: Each dendrogram is color-coded with a dedicated measure for a different data type. Dark colors indicate nodes with high variation. A dedicated interaction tool provides detailed information about the respective descriptors.*

creasing numbers of elements, which is what we expect for our application. Alternative information theoretic measures [ALF21] would work, but e.g. purity considers only the biggest class and not the full distribution within a cluster. Other measures, such as pair-counting and set-matching [ALF21], rely on ground truth labels, which are typically not available in exploration workflows. Figure 6 (A) shows the entropy for the single-class labels, which were manually assigned based on chemical intuition (a total of 41 labels, e.g., cyclic alkanes, water-soluble alcohols, etc. were considered). Hovering over the dendrogram nodes shows a tooltip with the most frequent labels. For the left tooltip, two labels make up 71% of the classes, which results in low entropy. The right tooltip shows a cluster with many different substances and, accordingly, high entropy.

For **multi-label** assignments, we use a measure from machine learning, namely *binary cross entropy*, which is commonly used as a loss function to quantify how well a predicted multi-labeling approximates the ground truth:

$$e_{\text{multi}} = \frac{1}{N}\sum_{i=1}^{N}((1-y_i)*\log_2(1-p_i)) - (y_i*\log_2 p_i) \overset{y=1}{=} -\frac{1}{N}\sum_{i=1}^{N}\log_2(p_i) \tag{2}$$

For all labels y in a cluster, $y_i = 1$ if the label is correctly predicted to be in the cluster and $y_i = 0$ if it is falsely predicted and p_i is the fraction of elements with this label in the cluster. Since we again lack the ground truth necessary for cluster validation, we assume the ideal case in that clusters are supposed to be pure, i.e., all labels N occurring in a cluster are expected to be present in all elements of a cluster. With $y = 1$, the formula simplifies significantly. Figure 6 (B) shows the binary cross entropy for a set of structural attributes characterizing the substances (e.g., cyclic, aromatic, long-chained); the set of structural attributes was defined manually here, but any categorical multi-labeling, like the well-established group-contribution method UNIFAC [FJP75], could be used. The tooltip again shows the most frequent labels and denotes how many substances share a label.

For **numerical data**, we chose the *mean standard deviation of standardized values* σ_{mean}, which means we compute an average standard deviation over all included continuous measures:

$$\sigma_{\text{mean}} = \frac{1}{|V|}\sum_{v=1}^{|V|}\sigma_v \quad \text{with} \quad \sigma_v = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(z_{v,i}-\overline{z_v})^2} \quad \text{and} \quad z_{v,i} = \sum_{i=1}^{m}\frac{x_{v,i}-\overline{x_v}}{\sigma_v} \tag{3}$$

In a first step, we make variables comparable by standardizing each variable individually based on their distribution in the full dataset m. We then determine the variation within a cluster by computing the individual standard deviation σ_v for each variable restricted to the values of the cluster n. To make the measure independent of the number of variables, we output the mean of the standard deviations over all variables V. We chose this formula, since standardization of features will be necessary in almost all application scenarios, and it indicates directly how the standard deviation within the cluster compares to the global one. Figure 6 (C) shows σ_{mean} for five continuous descriptors. Selecting a node in the tree shows violin plots for the numerical descriptors, contrasting the selected cluster (colored-coded in the tree and the violin) with the rest of the data.

6. System Design and Implementation

Using the substance feature variation, the user can now manually traverse the tree and search for clusters with a semantic meaning. For their interpretation, they need detailed information about the substance features, which we provide in interactively linked widgets. The interface is deduced from the previously compiled design goals: Hover and explicit plots offer *access to raw data*; enrichment plots scalable with regard to number of features and data type give *flexibility*; and *linked interaction* provides intuitive exploration.

Figure 1 shows the entire GUI. The collapsible parameter sidebar on the left covers algorithmic and style settings that can be controlled by the user. The visualization section on the right contains multiple linked views on the data with interaction capabilities. We have already discussed the design of the clustermap. For the

enrichment substance features, we provide four additional visualizations. (1) A table containing the names and molecular formulas of the substances, (2) a 2D projection of matrix rows or columns to reveal relative distances, (3) an extension matrix plot showing multi-labels, e.g. the composition of the molecules with regard to structural groups, and (4) violin plots to analyze distributions of numerical features. We thereby extend previous enrichment support [FGR*17] to more and potentially continuous variables, independent of application domain.

All widgets are interactively linked. Selecting a substance or substance group in one of the widgets triggers a highlighting operation in all the other ones, supporting the user in finding relevant features. Individual data point selections are drawn with bigger lines and bright orange color. In Figure 1 two high-level nodes were selected subsequently. The respective subtrees are color-coded blue and red. The table only shows currently selected substances. As kernel density estimates have been proven to work for cluster attribute comparison [SDMT16], the violin plots contrast the current selection against the remaining data or a previously saved selections. We combine violin plots with hover-able rug plots to ease the reading of outliers and to provide interaction capabilities, e.g. selection of individual substances. In case a user is uncomfortable with violin plots, they can change to equally arranged histograms. Even if carefully chosen, the matrix imposes a linear ordering that cannot capture the potentially complex neighborhood relationships between rows/columns. We include a non-linear projection that mirrors the colors of user selections. We chose MDS over other non-linear projections for its simplicity to explain to domain scientists. The hover tool provides the exact values and substance names of the glyphs in every view. A demonstration of all interactions is given in the accompanying video. The implementation is built in Python with Bokeh and Panel [Rudb] as the charting and interaction libraries. The tool is available at *https://github.com/Jan-To/EnrichMatrix*.

7. Case Studies

From our analysis in subsection 5.2, we know that we work on a suitable ordering. Therefore, we demonstrate how the software can be used to first find pattern-correlating numerical features and then confirm or question reference classifications.

7.1. Matrix Pattern Correlation with Continuous Features

Finding informative substance properties is crucial for the development of prediction methods for thermodynamic properties. To date, this task is based on intuition of human experts and, generally, by considering properties of *pure* substances. The software developed in this work facilitates an unbiased analysis based on *mixture* data, namely by studying which substance properties are particularly homogeneous in matrix clusters. We notice two prevalent groups in the matrix, which correspond to the highest ranked nodes in the dendrogram. We selected and saved them with different colors in Figure 1. In an initial review of the substance table, we already observe a high degree of homogeneity among the substances. The blue cluster mainly contains non-polar hydrocarbons, whereas the red cluster includes highly polar compounds. The distributions of the substance properties in the violin plots, confirm our observation: we find rather small dipole moments and relatively high polarizabilities, a characteristic of non-polar molecules, in the blue cluster compared to the red cluster, while the red cluster exhibits greater heterogeneity. Hence, we conclude that the polarity of the molecules is one of the most important properties with regard to activity coefficients. While this agrees well with chemical intuition, polarity is usually integrated in the input of previous hand-crafted mixture prediction, so it is interesting to quantitatively find this in our unbiased *mixture* data. Thus, we can interactively analyze data-driven relationships between matrix patterns and multiple external feature distributions.

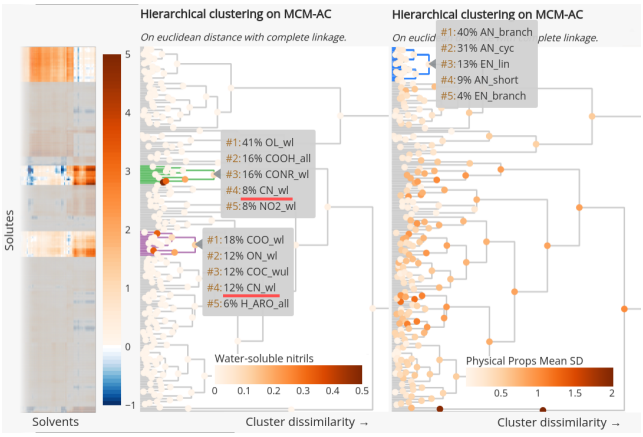


Figure 7: *(Center): Water-soluble nitriles (CN_wl) can be found in two clusters, which indicates different mixture behaviors. (Right): Example for a cluster with high homogeneity regarding single-labels, which, however, includes multiple expert-labeled classes.*

7.2. Transparent Evaluation of Reference Classifications

Classifying substances is fundamental for the development of predictive thermodynamic models, but a non-trivial task that is usually done by a human expert in a subjective manner. The software developed in this work enables a data-based evaluation of such classifications. We therefore change the color coding of the tree nodes to the occurrence of a specific class label as defined by an expert, e.g., water-soluble nitriles (CN_wl), cf. Figure 7 (center). In this mode, we notice that the water-soluble nitriles are part of two clusters as found based on the mixture data. Taking into account that the other substances in the two clusters are strongly different with regard to their class membership, we can conclude that 'water-soluble nitriles' is not a very characteristic group label regarding the behavior of substances in mixtures and should not be used for this purpose. A reverse procedure is also conceivable: if we look for a cluster that is very homogeneous in its feature values, we quickly find the blue cluster, cf. Figure 7 (right), which is surprisingly quite heterogeneous according to its expert group labels. We find branched, cyclic, and short alkanes together with linear and branched alkenes, which apparently all show a very similar mixture behavior. Based on this observation, we can deduce that a separate consideration of these tagged groups is not pertinent and that a group classification should rather be based on the (relevant) substance properties, which are very homogeneous in this cluster, namely, dipole

moment, polarizability, and anisotropy (not shown). Through the multidirectional analysis between matrix patterns, class labels and feature values we enable domain scientists to validate, invalidate and suggest annotated matrix classifications.

8. User Evaluation

In addition to the case studies by our domain scientist authors, we conducted a qualitative user study with six PhD students to evaluate the accessibility and effectiveness of our tool in practical use. All six are pursuing a degree in chemical engineering, albeit most do not specialize in the analysis of mixture data. Only one of them was familiar with the matrix data before the study, but has only seen numerical representations without annotations.

After a brief introduction to the data, the tool, and its functionality, we asked three introductory questions, after which they could explore the data on their own. We encouraged the participants to think aloud, took notes of their comments, and additionally recorded the audio during each session. In the end of the session, we conducted a short interview to capture a summary feedback on usability and productiveness. Each session took between 30 and 60 minutes. All participants started their analysis on the screen ordered as determined in subsection 5.2 and were given the same tasks: *Can you find patterns in the matrix and characterize them with domain knowledge? Can you determine substance properties that significantly influence the mixture behavior? Which patterns coincide with domain knowledge?*

Most participants were able to immediately identify, explore and correlate patterns in the matrix. Selections in all plots were made abundantly. One person was overwhelmed with the abundance of simultaneous views and another felt unfamiliar with the concept of dendrograms, though both reservations resolved quickly and without intervention. The filtered table was equally used as the violin plots and checked against each other for reliability. The implications of the various node colorings in the dendrogram were extensively explored. The participants used them to recognize both consistent and inconsistent clusters and confirmed them within the violins.

The users particularly praised the wide range of consistent interaction and selection possibilities (4 users), the visual clarity (3 users), and the accessibility of complementary information (3 users). The participants commented on the beginner-friendly design through abstractions to color and violins, enabling analysis without knowledge of statistical methods (3 users), though the person familiar to the dataset liked that raw data is accessible by hovering at all times. Improvement suggestions were aimed to enhance the user experience. Participants were missing a 'reset' button (3 users), images of the structural formula of substances (2 users), and resizing violin plots (2 users).

Overall, the positive feedback from our user study showed that application domains can greatly benefit from visual analytic interfaces compared to current workflows. Interactive linked views as well as visual abstractions are quick to learn and use as long as design and interactions are intuitive. With regard to our specific application, users were able to successfully check various levels of cluster validation (data, visible pattern, dendrogram, violin) against each other. The participants were so interested in the interaction possibilities that, even though most of them had no relationship with the data, everybody continued exploring after the official session ended. That participants were excited about the tool's interaction possibilities and wanted to apply their own data, indicates (1) that relating attributes to matrices is a common problem and (2) that the need for visual analytics software in application domains is still huge.

9. Conclusion and Future Work

In this paper, we introduced an analysis software for asymmetric scalar matrices that are complemented with meta-data for row and column entities. Central building blocks are a pattern-focused sorting of the heatmap and the guided variation analysis of the meta-data across multiple linked views. While the workflow and most parts of the software are independent of the application domain, we focus on chemical engineering, particularly, the exploration of the relationship between the activity coefficients in binary mixtures and properties of the pure substances. We demonstrate that the software can be used in practice to find informative descriptors for modeling mixture properties based on a cluster analysis of the mixture data, and contrast it with existing domain knowledge. The analysis with our software provided data-driven directions towards suitable substance descriptors and classification schemes for binary mixtures, advancing the field from manual analysis by physicochemical intuition.

Our analysis software currently has limitations. Firstly, we rely on hierarchical clustering for its interpretability and inherent hierarchy for domain patterns, though alternative ordering techniques may be viable. Secondly, displaying all data points simultaneously limits the matrix size to ∼500x500 due to pixel resolution, but solutions like scrolling or agglomeration have been successfully employed before [CMP10]. Thirdly, enrichment data must be available; however, in our experience with other engineering sciences, this data is typically accessible, and engineers are often eager to consolidate it. Hence, an apparent direction for future research is the application to new domains. The infrastructure of the system is designed to be generic by handling multiple data types and lends itself directly to the integration with data from other domains.

10. Acknowledgements

This research was funded by the priority program *SPP 2363 Molecular Machine Learning* (No. 460865652) and the *FOR 5359 Deep Learning on sparse chemical process data* (No. 459419731) of the German Research Foundation (DFG).

References

[ALF21] ARINIK, NEJAT, LABATUT, VINCENT, and FIGUEIREDO, ROSA. "Characterizing and Comparing External Measures for the Assessment of Cluster Analysis and Community Detection". *IEEE Access* 9 (2021), 20255–20276. DOI: 10.1109/ACCESS.2021.3054621 6.

[BBH*16] BEHRISCH, MICHAEL, BACH, BENJAMIN, HENRY RICHE, NATHALIE, et al. "Matrix Reordering Methods for Table and Network Visualization". *Comput. Graph. Forum* 35.3 (June 2016), 693–716. ISSN: 0167-7055. DOI: 10.1111/cgf.12935 3, 4.

[BBH*17] BEHRISCH, MICHAEL, BACH, BENJAMIN, HUND, MICHAEL, et al. "Magnostics: Image-Based Search of Interesting Matrix Views for Guided Network Exploration". *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), 31–40. DOI: 10.1109/TVCG.2016.2598467 3.

[BdS22] BARONI, MATHEUS PERON and da SILVA, CELMAR GUIMARÃES. "A comparative analysis of matrix reordering algorithms regarding canonical data patterns". *Information Visualization* 21.3 (2022), 321–332. DOI: 10.1177/14738716221091487 4.

[BMGK08] BARSKY, AARON, MUNZNER, TAMARA, GARDY, JENNIFER, and KINCAID, ROBERT. "Cerebral: Visualizing Multiple Experimental Conditions on a Graph with Biological Context". *IEEE Transactions on Visualization and Computer Graphics* 14.6 (2008), 1253–1260. DOI: 10.1109/TVCG.2008.117 3.

[Bra07] BRANDES, ULRIK. "Optimal leaf ordering of complete binary trees". *Journal of Discrete Algorithms* 5.3 (2007). Selected papers from Ad Hoc Now 2005, 546–552. ISSN: 1570-8667. DOI: https://doi.org/10.1016/j.jda.2006.09.003 4.

[BSP20] BEHRISCH, M., SCHRECK, T., and PFISTER, H.-P. "GUIRO: User-Guided Matrix Reordering". English. *IEEE Transactions on Visualization and Computer Graphics* 26.1 (Jan. 2020), 184–194. ISSN: 1077-2626. DOI: 10.1109/TVCG.2019.2934300 3.

[Che02] CHEN, CHUN-HOUH. "Generalized Association Plots: information visualization via iteratively generated correlation matrices". *Statistica Sinica* 12 (Jan. 2002), 7–29 4.

[CMP10] CHEN, JIN, MACEACHREN, ALAN, and PEUQUET, DONNA. "Constructing Overview + Detail Dendrogram-Matrix Views". *IEEE transactions on visualization and computer graphics* 15 (Jan. 2010), 889–96. DOI: 10.1109/TVCG.2009.130 3, 8.

[CN02] CARD, STUART K. and NATION, DAVID. "Degree-of-Interest Trees: A Component of an Attention-Reactive User Interface". *Proceedings of the Working Conference on Advanced Visual Interfaces*. AVI '02. Trento, Italy: Association for Computing Machinery, 2002, 231–245. ISBN: 1581135378. DOI: 10.1145/1556262.1556300 3.

[CP04] CARAUX, GILLES and PINLOCHE, SYLVIE. "PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order". *Bioinformatics* 21.7 (Nov. 2004), 1280–1281. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti141 4.

[ESBB98] EISEN, MICHAEL B., SPELLMAN, PAUL T., BROWN, PATRICK O., and BOTSTEIN, DAVID. "Cluster analysis and display of genome-wide expression patterns". *Proceedings of the National Academy of Sciences* 95.25 (1998), 14863–14868. ISSN: 0027-8424. URL: https://www.pnas.org/content/95/25/14863 4.

[EWJP17] ENGLE, SOPHIE, WHALEN, SEAN, JOSHI, ALARK, and POLLARD, KATHERINE S. "Unboxing cluster heatmaps". *BMC Bioinformatics* 18.2 (2017), 63. DOI: 10.1186/s12859-016-1442-6 4.

[FGR*17] FERNANDEZ, NICOLAS F., GUNDERSEN, GREGORY W., RAHMAN, ADEEB, et al. "Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data". *Scientific Data* 4.1 (Oct. 2017), 170151. ISSN: 2052-4463. DOI: 10.1038/sdata.2017.151 2–5, 7.

[FJP75] FREDENSLUND, AAGE, JONES, RUSSELL L, and PRAUSNITZ, JOHN M. "Group-contribution estimation of activity coefficients in nonideal liquid mixtures". *AIChE Journal* 21.6 (1975), 1086–1099. DOI: 10.1002/aic.690210607 6.

[GCML06] GUO, DIANSHENG, CHEN, JIN, MACEACHREN, ALAN, and LIAO, KE. "A visualization system for space-time and multivariate patterns (VIS-STAMP)". *IEEE transactions on visualization and computer graphics* 12 (2006), 1461–74. DOI: 10.1109/TVCG.2006.84 3.

[GCS15] GMEHLING, JÜRGEN, CONSTANTINESCU, DANA, and SCHMID, BASTIAN. "Group contribution methods for phase equilibrium calculations". *Annual review of chemical and biomolecular engineering* 6 (2015), 267–292. DOI: 10.1146/annurev-chembioeng-061114-123424 2.

[JBM20] JIRASEK, FABIAN, BAMLER, ROBERT, and MANDT, STEPHAN. "Hybridizing physical and data-driven prediction methods for physicochemical properties". *Chemical Communications* 56.82 (2020), 12407–12410. DOI: 10.1039/D0CC05258B 2.

[JH21] JIRASEK, FABIAN and HASSE, HANS. "Perspective: Machine Learning of Thermophysical Properties". *Fluid Phase Equilibria* 549 (2021), 113206. DOI: 10.1016/j.fluid.2021.113206 2.

[LBK*18] LEKSCHAS, FRITZ, BACH, BENJAMIN, KERPEDJIEV, PETER, et al. "HiPiler: Visual Exploration Of Large Genome Interaction Matrices With Interactive Small Multiples". *IEEE Transactions on Visualization and Computer Graphics*. InfoVis '17 (2018). DOI: 10.1109/TVCG.2017.2745978 3, 4.

[Lii10] LIIV, INNAR. "Seriation and matrix reordering methods: An historical overview". *Statistical Analysis and Data Mining: The ASA Data Science Journal* 3.2 (2010), 70–91. DOI: https://doi.org/10.1002/sam.10071 2, 4.

[LSKS10] LEX, ALEXANDER, STREIT, MARC, KRUIJFF, ERNST, and SCHMALSTIEG, DIETER. "Caleydo: Design and evaluation of a visual analysis framework for gene expression data in its biological context". *2010 IEEE Pacific Visualization Symposium (PacificVis)*. 2010, 57–64. DOI: 10.1109/PACIFICVIS.2010.5429609 2, 3.

[NGCL19] NOBRE, CAROLINA, GEHLENBORG, NILS, COON, HILARY, and LEX, ALEXANDER. "Lineage: Visualizing Multivariate Clinical Data in Genealogy Graphs". *IEEE Transactions on Visualization and Computer Graphics* 25.3 (2019), 1543–1558. DOI: 10.1109/TVCG.2018.2811488 3.

[PLS*13] PARTL, CHRISTIAN, LEX, ALEXANDER, STREIT, MARC, et al. "enRoute: Dynamic Path Extraction from Biological Pathway Maps for Exploring Heterogeneous Experimental Datasets". *BMC Bioinformatics* 14 (2013), S3. DOI: 10.1186/1471-2105-14-S19-S3 3.

[RSW*19] RYAN, MICHAEL C., STUCKY, MARK, WAKEFIELD, CHRIS, et al. "Interactive Clustered Heat Map Builder: An easy web-based tool for creating sophisticated clustered heat maps". *F1000Research* 8 (2019), ISCB Comm J–1750. DOI: 10.12688/f1000research.20590.2 5.

[Ruda] RUDIGER, P. *Holoviews*. https://holoviews.org 3.

[Rudb] RUDIGER, P. *Panel*. https://panel.holoviz.org 7.

[SDMT16] STAHNKE, JULIAN, DÖRK, MARIAN, MÜLLER, BORIS, and THOM, ANDREAS. "Probing Projections: Interaction Techniques for Interpreting Arrangements and Errors of Dimensionality Reductions". *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2016), 629–638. DOI: 10.1109/TVCG.2015.2467717 7.

[SMM12] SEDLMAIR, MICHAEL, MEYER, MIRIAH, and MUNZNER, TAMARA. "Design study methodology: Reflections from the trenches and the stacks". *IEEE transactions on visualization and computer graphics* 18.12 (2012), 2431–2440. DOI: 10.1109/TVCG.2012.213 3.

[SS02] SEO, JINWOOK and SHNEIDERMAN, BEN. "Interactively Exploring Hierarchical Clustering Results". *Computer* 35 (Aug. 2002), 80–86. DOI: 10.1109/MC.2002.1016905 4.

[Was21] WASKOM, MICHAEL L. "seaborn: statistical data visualization". *Journal of Open Source Software* 6.60 (2021), 3021. DOI: 10.21105/joss.03021. URL: https://doi.org/10.21105/joss.03021 3.

[Wil05] WILKINSON, LELAND. *The Grammar of Graphics (Statistics and Computing)*. Heidelberg: Springer-Verlag, 2005. ISBN: 0387245448. DOI: 10.1007/0-387-28695-0 3, 4.

[WTC08] WU, HAN-MING, TZENG, SHENGLI, and CHEN, CHUN-HUOH. "Handbook of Data Visualization". 2008, 681–708. DOI: 10.1007/978-3-540-33037-0 4.

[WTC10] WU, HAN-MING, TIEN, YIN-JING, and CHEN, CHUN-HOUH. "GAP: A Graphical Environment for Matrix Visualization and Cluster Analysis". *Computational Statistics & Data Analysis* 54 (Mar. 2010), 767–778. DOI: 10.1016/j.csda.2008.09.029 3, 4.