Eurographics Workshop on Visual Computing for Biology and Medicine (2025)
L. Garrison and R. Krueger (Editors)

Short Paper

Clusters in Focus: A Simple and Robust Detail-On-Demand Dashboard for Patient Data

L. Schilcher[1†], P. Waldert[1†‡], B. Kantz[1] and T. Schreck[1]

[1] Institute of Visual Computing, Graz University of Technology

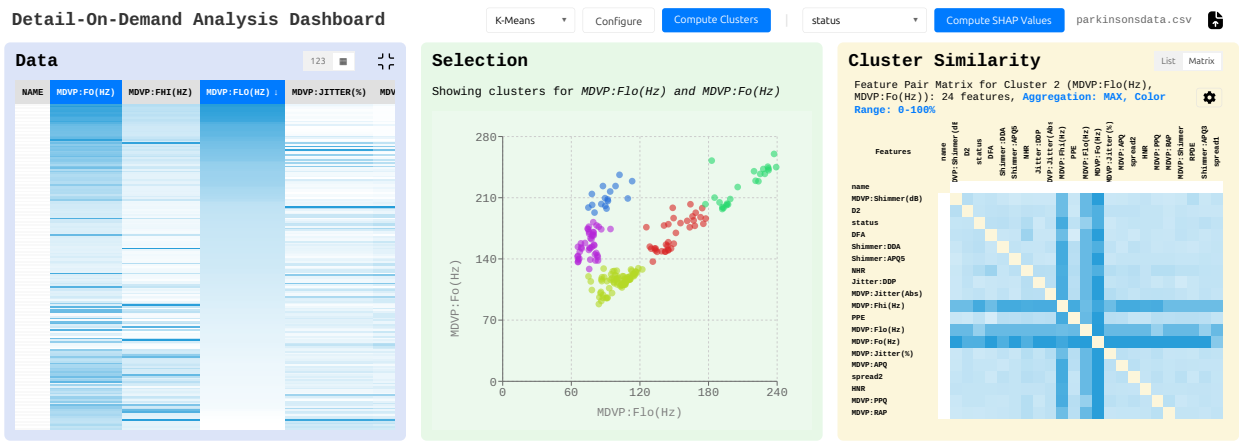


Figure 1: *Overview of Clusters in Focus: Data Panel (left), Selection Panel (center), and Cluster Similarity Panel (right).*

Abstract

*Exploring tabular datasets to understand how different feature pairs partition data into meaningful cohorts is crucial in domains such as biomarker discovery, yet comparing clusters across multiple feature pair projections is challenging. We introduce Clusters in Focus, an interactive visual analytics dashboard designed to address this gap. Clusters in Focus employs a three-panel coordinated view: a **Data Panel** offers multiple perspectives (tabular, heatmap, condensed with histograms / SHAP values) for initial data exploration; a **Selection Panel** displays the 2D clustering (K-Means/DBSCAN) for a user-selected feature pair; and a novel **Cluster Similarity Panel** featuring two switchable views for comparing clusters. A ranked list enables the identification of top-matching feature pairs, while an interactive similarity matrix with reordering capabilities allows for the discovery of global structural patterns and groups of related features. This dual-view design supports both focused querying and broad visual exploration. A use case on a Parkinson's disease speech dataset demonstrates the tool's effectiveness in revealing relationships between different feature pairs characterizing the same patient subgroup.*

CCS Concepts

• ***Human-centered computing*** → ***Visual analytics;*** • ***Applied computing*** → *Health informatics;*

1. Introduction

In many applications, data items need to be understood in terms of groups of similar items. When data is described by many attributes (or features), the question arises in which features the grouping is present. While often, *all* features are used to compare data simultaneously, it is also interesting to look for *subsets* of features in which data can be grouped. This can for example be relevant in multidimensional biomedical datasets, where researchers want to discover potential biomarkers (objective measures indicating a particular biological state or condition), and understand disease subtypes described in different data features [RHL*15]. A key challenge in this process involves understanding how different combinations of features, especially pairs, relate to underlying patterns or partition the data into meaningful cohorts or clusters. Even when established biomarker pairs are known, discovering alternative or related feature pairs that characterize a similar cohort of subjects (e.g., patients) can lead to novel hypotheses or reveal previously overlooked correlations. Traditional exploratory data analysis often relies on examining feature pairs individually, for example through scatterplots, as these provide directly interpretable views based on original data dimensions. A sys-

† The first two authors contributed equally to this work.
‡ Corresponding Author. Email: peter.waldert@tugraz.at.

DOI: 10.2312/vcbm.20251250

tematic comparison of all possible subgroups or clusters becomes time-intensive and quickly infeasible as the number of features grows (and the number of pairs, accordingly). Furthermore, clustering based on the full feature set or using dimensionality reduction techniques may obscure patterns prominent only within specific low-dimensional subspaces, introduce hard-to-explain abstract features, and is sensitive to the curse of dimensionality [Bel66]. *Therefore, our approach specifically focuses on analyzing feature pairs*, which maintains interpretability by operating directly on original feature combinations. Existing visual analytics systems, such as Caleydo [LSKS10] and LineUp [GLG*13], provide coordinated views and interactive ranking mechanisms for multivariate data exploration. However, they do not specifically address the systematic comparison of cluster compositions generated from different feature subspaces, especially from feature pair projections. Tools like VICTOR [KGH*21], Clustrophile [Dem17], and Clustrophile 2 [CD19] primarily support comparing different clustering algorithms or tuning parameters on a fixed feature space.

Our design is motivated by a core analytical task: once users identify a meaningful cohort using one feature pair, they need a systematic way to discover other pairs that characterize a similar subgroup. This process is essential for validating findings and comparing related biomarkers within an interpretable, low-dimensional context. To address this gap, we present Clusters in Focus, an interactive visual analytics dashboard designed for the exploration of tabular data, aimed at identifying related feature pairs by comparing the similarity of their induced clusters. Clusters in Focus implements a coordinated three-panel workflow:

1. The **Data Panel** provides multiple views (tabular, heatmap, condensed with histograms and optional SHAP-based feature importance [LL17]) for initial data overview and feature selection.
2. The **Selection Panel** displays the 2D clustering (using K-Means or DBSCAN) of data points based on two user-selected features.
3. The **Cluster Similarity Panel**, our core contribution, activates when a user selects a data point (and thus its cluster) in the Selection Panel. It quantitatively compares this source cluster against clusters from all other feature pairs using the Jaccard Index. The results are presented in two complementary views: a **ranked list** for targeted identification of top-matching pairs, and a **reorderable similarity matrix** to reveal global patterns and relationships between feature groups.

This workflow allows domain experts, such as biomedical researchers, to start with a known feature pair of interest, identify a relevant cluster, and efficiently discover other feature pairs that might serve as alternative or complementary biomarkers characterizing a similar subgroup.

2. Related Work

Our work intersects with several areas in visual analytics, primarily with interactive clustering analysis and visual cluster comparison. Due to the vast number of related works in visual cluster analysis, we can only refer to a limited selection; it is important to note that this field has been extensively researched for many years, and our references represent just a small sample of the existing literature.

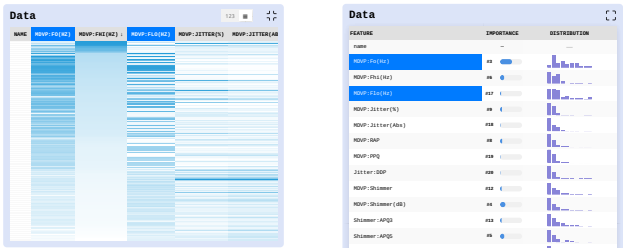
Interactive Clustering Tools: A rich body of work exists on systems that allow users to guide the clustering process. These tools support this goal through various means, such as iterative parameter tuning (Clustrophile 2 [CD19]), direct manipulation of cluster labels and their corresponding 2D projections (Cluster Sculptor [BPBO15]), or visual supervision of results via quality metrics (Clustervision [KEV*18]). While these systems excel at helping users optimize or validate a clustering configuration within a single, fixed feature space, Clusters in Focus's primary contribution is different. Our focus is on exploring the feature space itself by systematically comparing the resulting cluster structures across many different feature pair subspaces.

Visual Cluster Comparison Tools: Tools like VICTOR [KGH*21] enable users to compare pre-computed clustering partitions generated by different algorithms or parameters on the same dataset. VICTOR uses metrics like the Jaccard Index and Adjusted Rand Index, visualized through methods like heatmaps or Sankey diagrams, to assess the overall similarity between different partitioning strategies. While VICTOR excels at comparing partitions based on the same feature set, it does not directly address clusters formed from different feature combinations. Clusters in Focus fills this gap by systematically clustering across all feature pairs and providing ranked similarity comparisons centered on a user-selected cluster, thus comparing the *results* of projecting data onto different feature pairs.

Visual Analytics for Multivariate Data: Existing visual analytics systems for multivariate data form the basis for our design. Caleydo [LSKS10] integrates tabular data (e.g., gene expression) with contextual information (e.g., biological pathways) using multiple coordinated views, with a focus on biological interpretation. While Clusters in Focus also utilizes coordinated views, it specifically supports workflows for comparing cluster membership similarity across different feature pair views within a single table, without integrating external context. LineUp [GLG*13] enables interactive ranking of items based on user-defined combinations of attributes. Clusters in Focus's Panel 3 presents a ranked list that ranks (feature pair, cluster ID) tuples by their Jaccard similarity to a reference cluster to identify feature pairs yielding similar data cohorts rather than ranking individual items by attribute scores.

Subspace Analysis and Visualization: Our work relates to the visual analysis of feature subspaces. Several systems support this exploration with different goals. For instance, some tools visually compare 2D subspaces by ranking them according to various quality metrics [TMF*12]. The Patterntrails approach [JHB*17] traces data points across an ordered sequence of subspaces to reveal their behavior. A related approach is "SmartStripes" by May et al. [MBD*11], which guides feature subset selection for predictive modeling. Their system decomposes global statistical quality measures across data partitions, visualizing local importance for specific cohorts and interactively steering a selection algorithm.

Based on the indicated problems, Clusters in Focus differentiates itself from these methods by focusing on the direct comparison of data cohorts. Instead of ranking subspaces by abstract statistical metrics for model building [TMF*12, MBD*11] or tracing individual points [JHB*17], our approach is centered on a user-selected cluster. It then systematically discovers other feature pairs which



(a) ***Heatmap View****: Condensed overview of normalized feature correlations in Panel 1, similarly arranged as in LineUp [GLG*13].*

(b) ***Condensed View****: Miniature histograms per feature, with an optional SHAP-based feature ranking [LL17] in Panel 1.*

Figure 2: *Panel 1 views in Clusters in Focus: (a) Heatmap and (b) Condensed feature summary.*

partition the data similarly, directly addressing the task of finding related features that characterize the same underlying group.

3. The Clusters in Focus-Tool

Clusters in Focus provides an interactive web-based environment for analyzing feature pair relationships in tabular data via clustering. Its three coordinated panels support a structured workflow from initial overview to detailed cluster comparison (Figure 1).

Implementation Architecture: Clusters in Focus follows a client-server architecture to outsource computational demands while keeping interactive performance. The frontend is built with React and TypeScript and provides a responsive user interface running entirely in the user's web browser. A lightweight Python backend using FastAPI handles computational tasks. With this separation, the frontend focuses exclusively on visualization and user interaction, while the backend manages analytical processing and result caching. The full source code is available at github.com/hereditary-eu/ClustersInFocus and the application has been dockerized. To build and start the application, run `docker compose build` and `docker compose up`.

3.1. Panel 1: Data & Feature Exploration

Once a CSV dataset has been uploaded via the upload control in the application's header, Panel 1 provides the primary interface for feature exploration and selection. This panel offers three distinct views, switchable via icons:

- **Tabular View:** Presents the data in a standard spreadsheet-like table. Users can sort the table rows based on the values in any column and hide columns to focus the view.
- **Heatmap View:** Provides a condensed overview of numerical features. Each cell's value is mapped to a color gradient (normalized within its column), and rows are shrunk vertically to maximize the number of visible items (Figure 2a). Hovering over a row expands it to reveal the actual numerical values which allows quick inspection within the overview context. Similar to the Tabular View, columns can be sorted or hidden.
- **Condensed View:** Offers a feature-centric summary (Figure 2b). It lists all features, showing a miniature histogram of the respective distribution. Optionally, if SHAP values have been computed (by selecting a target variable and clicking "Compute" in the header), this view displays the rank of each feature's importance. Clicking on a mini-histogram opens a larger view with configurable bin count for detailed distribution analysis.

Regardless of the active view within Panel 1, users select features for analysis in Panel 2 by clicking directly on the respective column headers. Clusters in Focus enforces a limit of exactly two active features for the subsequent clustering visualization. If a user selects a third feature header, the chronologically first selected feature is automatically deselected to ensure that only the two most recently selected features are passed to the Selection Panel.

3.2. Panel 2: Focused 2D Clustering

If two features (F_A, F_B) are selected in Panel 1, the central Selection Panel displays the corresponding 2D scatterplot of the data points (Figure 1, center panel). The points in this scatterplot are colored to reflect their cluster assignments. These assignments are derived from clustering results computed using the algorithm and parameters specified via controls in the application header. Specifically, the header controls allow the user to:

- Select the clustering algorithm (K-Means or DBSCAN) via a dropdown menu.
- Configure the algorithm's hyperparameters (e.g., 'k' for K-Means; 'eps' and 'min_samples' for DBSCAN).
- Trigger the backend clustering (re-)computation (or re-computation if parameters change) for the currently selected feature pair (F_A, F_B).

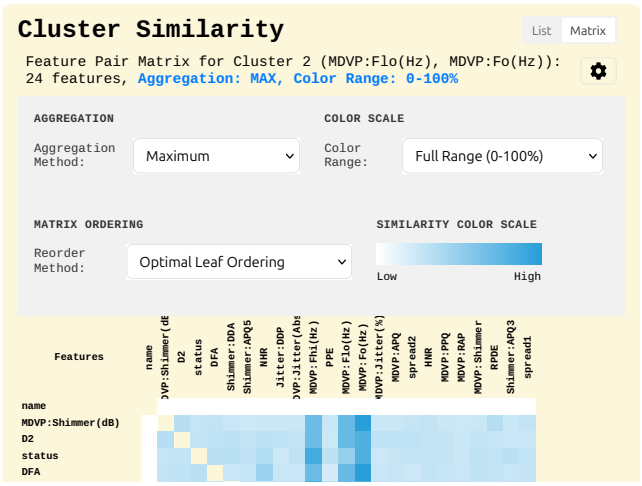
Clustering results are retrieved from the backend after computation, and points are colored according to their assigned cluster IDs in the Panel 2 scatterplot. In principle, the number of clusters (and therefore, colors) is not limited from a technical viewpoint, but of course after a certain number (~ 10) they become hard to distinguish visually. The primary interaction *within* Panel 2 itself is **selecting a data point.** Clicking on any point selects the cluster it belongs to and populates Panel 3 with the cluster similarity analysis.

3.3. Panel 3: Cross-Pair Cluster Similarity

This rightmost panel is the core analytical component of Clusters in Focus, activating upon selection of a data point (and thus its cluster) in Panel 2. Let the features selected in Panel 1 be F_A and F_B, and let the user-selected cluster be C_{AB} (representing a specific set of data points). Panel 3 then performs a systematic comparative analysis before presenting the results in one of two switchable views. The underlying analysis proceeds as follows:

- For *every other unique pair* (F_X, F_Y) in the dataset, the tool retrieves or computes the clustering result obtained using the same algorithm and parameters specified in the header. This yields a set of clusters $\{C_{XY:1}, C_{XY:2}, \cdots, C_{XY:k}\}$ for each pair (F_X, F_Y).
- The Jaccard similarity $J(C_{AB}, C_{XY:k})$ is calculated between the initially selected cluster C_{AB} and **every individual cluster** $C_{XY:k}$ from **every other feature pair** (F_X, F_Y). The Jaccard Index $J \in [0,1]$ measures the overlap in data point membership

$$J = \frac{|C_{AB} \cap C_{XY:k}|}{|C_{AB} \cup C_{XY:k}|}.$$



[cut off; a full view is available in Figure 1.]

Figure 3: *The Cluster Similarity Panel in Matrix View configuration. The panel provides two switchable views (top right): a List View for targeted ranking and this Matrix View for global pattern discovery. Each cell's color represents the aggregated Jaccard similarity, with the aggregation method (Maximum) and matrix ordering (Optimal Leaf Ordering) selected by the user. The reordering algorithm groups features with similar profiles, revealing a distinct high-similarity block around the source features and exposing, through the bright horizontal and vertical lines, their strong relationship with other feature groups related to vocal stability (e.g., HNR, DFA) and jitter (Jitter:DDP).*

These calculated similarity scores form the basis for the two complementary views offered by the panel:

- **List View**: This view directly utilizes the individual similarity scores. It presents a table that ranks all (Feature Pair, Cluster ID) combinations based on their calculated Jaccard similarity with the source cluster C_{AB}, from highest to lowest. This view is ideal for focused queries to quickly identify the single best-matching clusters from other feature projections.
- **Matrix View**: This view (Panel 3) provides a global, structural overview. Instead of listing individual cluster scores, it aggregates them. For each feature pair (F_X, F_Y), it combines the set of similarity scores $\{J(C_{AB}, C_{XY:1}), J(C_{AB}, C_{XY:2}), \ldots\}$ into a single value using a user-selected aggregation method (e.g., Maximum, Average). The result is a feature-by-feature heatmap where cell (F_X, F_Y) is colored by this aggregated score. With interactive matrix reordering (e.g., Optimal Leaf Ordering), this view aims at revealing high-level patterns, such as blocks of related features that consistently define similar cohorts.

By offering both views, the system supports both focused, query-driven analysis and broader, discovery-oriented exploration of the dataset's cluster relationships.

Workflow Summary: **1.** Upload a dataset (CSV). **2.** Explore features in Panel 1 (Table, Heatmap, Condensed views) to assess distributions and SHAP-ranked importance (optional). **3.** Select two features (F_A, F_B). **4.** Configure parameters and compute clusters in Panel 2. **5.** Select a cluster C_{AB} by clicking a data point. **6.** Analyze the results in Panel 3. Use the **List View** for a ranked overview of top matches, or switch to the **Matrix View** and apply reordering to discover higher-level patterns and relationships between feature groups.

4. Use Case: Parkinson's Disease Speech Biomarkers

We demonstrate Clusters in Focus using the Parkinson's Disease speech dataset from the UCI Machine Learning Repository [Lit07], containing voice measurements and a 'status' attribute (healthy/PD). The goal is to identify speech patterns that differentiate the groups.

The analyst loads the dataset and selects `MDVP:Flo(Hz)` (minimum fundamental frequency) and `MDVP:Fo(Hz)` (average fundamental frequency) in Panel 1. Using the header controls, they choose K-Means clustering, set $k = 5$, and trigger the cluster computation. Panel 2 then displays the 2D scatterplot, coloring points based on their cluster assignments (Figure 1). The analyst observes the clusters and decides to investigate 'Cluster 2' further by clicking on a point within it, hypothesizing that it represents a specific patient subgroup.

Panel 3 updates to display the similarity analysis for 'Cluster 2'. The analyst first inspects the **List View**, which confirms that other feature pairs, especially those involving vocal perturbation, can also identify highly similar cohorts. For a more comprehensive overview, they switch to the **Matrix View** and apply *Optimal Leaf Ordering* (cf. Figure 1). The reordering reveals a distinct block of high-similarity cells, grouping the fundamental frequency measures with features related to non-linear dynamics (`DFA`, `PPE`, `spread1`) and vocal stability (`HNR`). This indicates that 'Cluster 2' is not an artifact of a single feature pair, but a stable cohort identifiable across a family of different acoustic measurements.

Interpretation: The analysis reveals that the cohort identified by 'Cluster 2', based purely on fundamental frequency measures, is consistently captured by a broader family of acoustic features. The high Jaccard similarities across this block, made evident by the Matrix View, suggest that 'Cluster 2' represents a robust subgroup characterized by a distinct and multi-faceted acoustic profile. The substantial overlap observed between this cluster and the cohort defined by the clinical `status` label supports the clinical relevance of this discovered subgroup. This workflow demonstrates how Clusters in Focus facilitates the validation of an initial observation, by highlighting that an identified pattern extends across multiple features and corresponds to a clinically significant population.

5. Discussion

Clusters in Focus is an easy-to-use dashboard for analysis of clusters in data. Its focus is on the comparison of clusters in multiple 2-dimensional subspaces. Thereby, the results of Clusters in Focus are easily interpretable and can be visually comprehended, as 2-dimensional subspaces can be directly visualized in scatter plots. A main application, and motivation for its development, are analysis tasks in biomedical applications. For example, grouping patients, or medical compounds into groups according to interpretable features, is often important for making conclusions. The results can be presented in a straightforward way. Future work should explore the

benefits and limitations of Clusters in Focus, considering visual analytics approaches for subspace analysis in high-dimensional subspaces exist. Note that the latter often have to employ feature selection or dimensionality reduction to be interpretable, and may hence incur a presentational bias.

Clusters in Focus allows comparing a selected cluster across different subspaces by offering two views. Alongside a **tabular list view**, it provides an **interactive heatmap matrix** to overview the Jaccard index across all pairwise dimension combinations. This matrix incorporates sorting capabilities to help identify block patterns, allowing for conclusions on influential and similar features. This approach provides a high-level view related to the comparison of cluster properties across subspaces, as discussed in [JHB*17], by showing the stability of a cohort's composition as the underlying feature pair changes.

While effective for pairwise exploration, the current approach is limited by the combinatorial growth in feature pairs for very high-dimensional datasets and its focus on 2D projections. Furthermore, practical tests indicate a scalability limit, as the backend's in-memory data processing and the frontend's direct rendering of all data points (which can be easily addressed in future versions of the tool) can lead to unresponsiveness with datasets exceeding several megabytes. Future work will explore methods for handling or selecting higher-dimensional feature combinations, likely guided by either dimensionality reduction techniques or feature importance measures. We also plan to extend the cluster comparison functionality by integrating other similarity metrics (e.g., Adjusted Rand Index) and exploring asymmetric or subset-aware measures to better capture relationships between clusters of differing sizes. To improve usability and lower the entry barrier, we include a pre-loaded default dataset. While the tool's three-panel design provides an implicit workflow following Shneiderman's *Overview first, zoom and filter, details-on-demand* mantra [Shn96], we acknowledge that more explicit user guidance is an important next step.

As demonstrated in the use case discussion, we have indication Clusters in Focus is able to provide interesting analysis across feature spaces. Future work should refine the requirement analysis, and apply and evaluate the approach in different domain settings.

6. Conclusion

We presented Clusters in Focus, an interactive visual analytics tool for exploring tabular data through cluster similarity comparisons across different 2D feature pair projections. By integrating data overview techniques, interactive 2D clustering, and a cross-pair comparison panel ranked by Jaccard similarity, Clusters in Focus enables users to efficiently identify feature pairs that capture similar data cohorts. Our approach provides a simple, fast exploration. Future work will expand and validate the approach.

Acknowledgements

This work was supported by the HEREDITARY Project, as part of the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074. Part of this work has already been outlined in a technical report, Deliverable 5.1 [SLW*24].

References

[Bel66] BELLMAN R.: Dynamic Programming. *Science 153*, 3731 (July 1966), 34–37. doi:10.1126/science.153.3731.34. 2

[BPBO15] BRUNEAU P., PINHEIRO P., BROEKSEMA B., OTJACQUES B.: Cluster sculptor, an interactive visual clustering system. *Neurocomputing 150* (2015), 627–644. doi:10.1016/j.neucom.2014.09.062. 2

[CD19] CAVALLO M., DEMIRALP Ç.: Clustrophile 2: Guided visual clustering analysis. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2019), 267–276. doi:10.1109/TVCG.2018.2864477. 2

[Dem17] DEMIRALP Ç.: Clustrophile: A tool for visual clustering analysis, 2017. arXiv:1710.02173. 2

[GLG*13] GRATZL S., LEX A., GEHLENBORG N., PFISTER H., STREIT M.: Lineup: Visual analysis of multi-attribute rankings. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '13) 19*, 12 (2013), 2277–2286. doi:10.1109/TVCG.2013.173. 2, 3

[JHB*17] JÄCKLE D., HUND M., BEHRISCH M., KEIM D. A., SCHRECK T.: Pattern trails: Visual analysis of pattern transitions in subspaces. In *12th IEEE Conference on Visual Analytics Science and Technology (VAST)* (2017), pp. 1–12. doi:10.1109/VAST.2017.8585613. 2, 5

[KEV*18] KWON B. C., EYSENBACH B., VERMA J., NG K., DE FILIPPI C., STEWART W. F., PERER A.: Clustervision: Visual supervision of unsupervised clustering. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2018), 142–151. doi:10.1109/TVCG.2017.2745085. 2

[KGH*21] KARATZAS E., GKONTA M., HOTOVA J., BALTOUMAS F. A., KONTOU P. I., BOBOTSIS C. J., BAGOS P. G., PAVLOPOULOS G. A.: Victor: A visual analytics web application for comparing cluster sets. *bioRxiv* (2021). doi:10.1101/2021.03.22.436502. 2

[Lit07] LITTLE M.: Parkinsons. UCI Machine Learning Repository, 2007. doi:10.24432/C59C74. 4

[LL17] LUNDBERG S. M., LEE S.-I.: A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, pp. 4765–4774. 2, 3

[LSKS10] LEX A., STREIT M., KRUIJFF E., SCHMALSTIEG D.: Caleydo: Design and evaluation of a visual analysis framework for gene expression data in its biological context. In *2010 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 2010, pp. 02–05. doi:10.1109/PACIFICVIS.2010.5429609. 2

[MBD*11] MAY T., BANNACH A., DAVEY J., RUPPERT T., KOHLHAMMER J.: Guiding feature subset selection with an interactive visualization. In *6th IEEE Conference on Visual Analytics Science and Technology (VAST)*. 2011, pp. 23–28. doi:10.1109/VAST.2011.6102448. 2

[RHL*15] RITCHIE M. D., HOLZINGER E. R., LI R., PENDERGRASS S. A., KIM D.: Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics 16*, 2 (Feb. 2015), 85–97. doi:10.1038/nrg3868. 1

[Shn96] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages* (1996), IEEE, pp. 336–343. doi:10.1109/VL.1996.545307. 5

[SLW*24] SCHRECK T., LENGAUER S., WALDERT P., KANTZ B., GRABNER M., SCHILCHER L., PERCIC D., VAN LEEUWEN C., LISSANDRINI M., ROMANOVYCH A.: Deliverable 5.1: Visualization components for sequences, networks, text, and high- dimensional data, Dec. 2024. doi:10.5281/zenodo.14628086. 5

[TMF*12] TATU A., MAASS F., FÄRBER I., BERTINI E., SCHRECK T., SEIDL T., KEIM D. A.: Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In *7th IEEE Conference on Visual Analytics Science and Technology (VAST)* (2012), pp. 63–72. doi:10.1109/VAST.2012.6400488. 2