

HEREDITARY

HetERogeneous sEmantic Data integration for the guT-bRain interplaY

Deliverable 5.3

Visualization components for spatial, image, and simulation data

Version 1.0, December 20, 2024

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.



**Funded by
the European Union**

EXECUTIVE SUMMARY

Research in HetERogeneous sEmantic Data integration for the guT-bRain interplaY (HEREDITARY) involves understanding properties, relationships and patterns in large and complex, heterogeneous data. Interactive data visualization provides techniques which can help to overview, and make sense of large and complex data sets, by finding appropriate visual representations and user interactions for the data. WP5 in HEREDITARY focuses on research and development of effective tools for interactive visual data analysis, supporting the HEREDITARY use cases.

This deliverable presents the results achieved in WP5 in developing visualization components for spatial, image, and simulation data in the first project year. The components have been informed by data and use cases from HEREDITARY partners and available open research data. The components are the basis for building standalone applications to address specific HEREDITARY use cases from M13, including functionality for interactive search and user guidance.

DOCUMENT INFORMATION

Deliverable ID	5.3
Deliverable Title	Visualization components for spatial, image, and simulation data
Work Package	WP 5
Lead Partner	SURF
Due Date	31 December 2024
Date of submission	December 20, 2024
Deliverable Type	R – Report
Dissemination level	PU – Public

AUTHORS

Name	Organization
Tobias Schreck (Author)	TUGRAZ
Stefan Lengauer (Author)	TUGRAZ
Peter Waldert (Contributor)	TUGRAZ
Benedikt Kantz (Contributor)	TUGRAZ
Michael Grabner (Contributor)	TUGRAZ
Lukas Schilcher (Contributor)	TUGRAZ
Dorian Percic (Contributor)	TUGRAZ
Casper van Leeuwen (Author)	SURF
Stancho Yordanov (Reviewer)	ONTO
Anna Romanovych (Contributor)	UNIPD

REVISION HISTORY

Version	Date	Author	Description
0.10	2024-02-05	Tobias Schreck	Initial skeleton
0.20	2024-11-13	Casper van Leeuwen	Draft content
0.30	2024-12-06	Casper van Leeuwen	Resolved reviewers comments
1.00	2024-12-12	Anna Romanovych	Resolved reviewers comments
1.00	2024-12-20	Gianmaria Silvello	Final check and PDF production

Contents

1	Introduction	8
2	Visualisation Components	9
2.1	Gut-brain visualization	9
2.1.1	Considered data	9
2.1.2	Visual Analytics Approach	10
2.1.3	Implementation Description (Backend, Frontend, Development Environment)	12
2.1.4	Development Roadmap: What to Develop Next?	12
2.2	Droplets: Bio+MedVis Challenge - Redesign (IMAGE, HIGH-DIM)	13
2.2.1	Considered Data / Test Dataset / Task Description	14
2.2.2	Visual Analytics Approach	14
2.2.3	Implementation Description (Backend, Frontend, Development Environment)	15
2.2.4	Demonstration of Preliminary Use Case(s)	15
2.2.5	Development Roadmap: What to Develop Next?	15
3	Conclusion	16
	References	16

List of Figures

1	Gut-brain visualization web-application	9
2	Gut micro-biota matrix visualization	11
3	Modality contribution visualization	11
4	Conventional brain visualization	11
5	Visual challenge outline of the Bio+MedVis Challenge @ IEEE VIS 2024	13
6	Alternative visualizations of part-to-whole relationships: barcharts, stacked barcharts, and treemaps.	14
7	Side by side comparison of the given baseline visualization (left) with our Droplets design, applied to the tissue sample provided by the redesign challenge.	16

Acronyms

DMN Default Mode Network. 10

ECN Executive Control Network. 9, 10

HEREDITARY HetERogeneous sEmantic Data integration for the guT-bRain interplaY. 3, 8, 13

LICA Linked Independent Component Analysis. 9–12

NIfTI Neuroimaging Informatics Technology Initiative. 10, 12, 13

SN Salience Network. 9, 10

SVG Scalable Vector Graphics. 15

1 Introduction

Research in the HEREDITARY project involves understanding properties, relationships and patterns in large and complex, heterogeneous data. Interactive data visualization provides techniques which can help to overview, and make sense of large and complex data sets, by finding appropriate visual representations and user interactions for the data. WP5's objective in the HEREDITARY project is the research and development of effective tools for interactive visual data analysis, supporting the HEREDITARY use cases, the aggregated data results from the polystore of WP3 and the results of the analytics and learning workflows developed in WP4.

This deliverable presents the results achieved in WP5 in designing and developing visualization components for spatial, image and simulation data in the first project year. Both exemplar use cases 4 and 5 were used as a starting point on which the visualization components were considered and designed. The considered visualization components are described in the next section.

2 Visualisation Components

The following visualisation components have been developed.

2.1 Gut-brain visualization

This gut-brain visualization component/application (see figure 1) can be sub-divided into 3 visualization components:

- Gut micro-biota matrix visualization
- Modality contribution visualization
- Conventional brain visualization

These 3 visualizations component, linked and combined, give a exploratory overview of the Linked Independent Component Analysis (LICA) model output data. The next sections will either be explained per individual component or together depending on the subject.

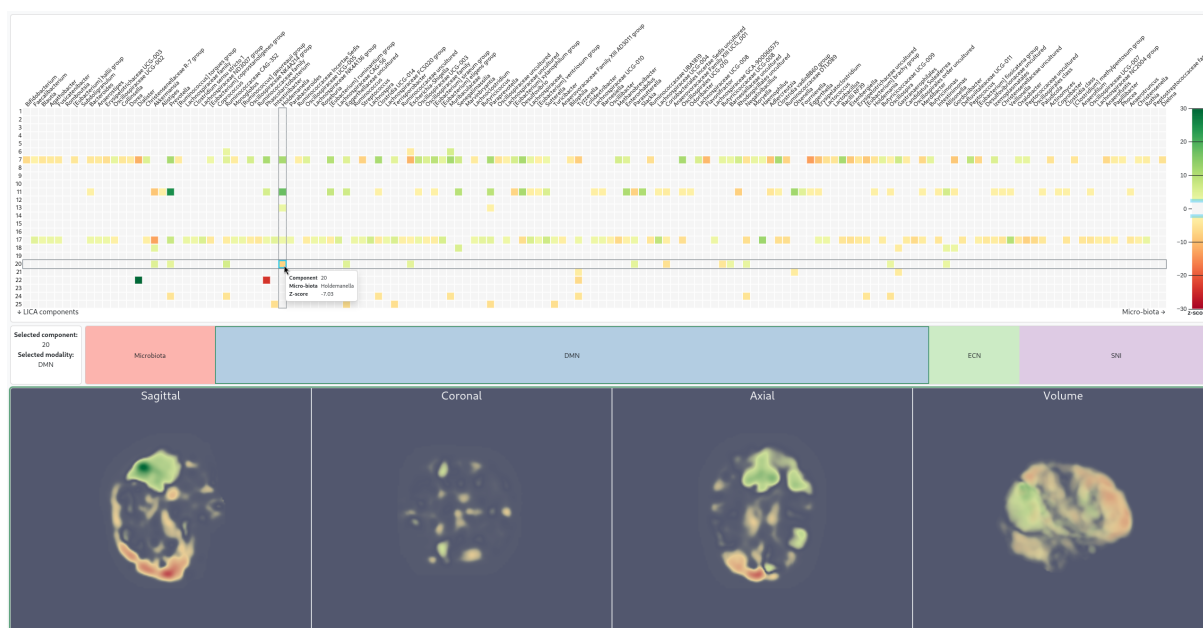


Figure 1: Gut-brain visualization web-application

2.1.1 Considered data

The considered data for this web-application is the result of the LICA method [Groves et al., 2011] used in the study by Nils Kohn et al. [Kohn et al., 2021]. The input data for the LICA method were multiple datasets of 189 individuals over 4 different modalities, micro-biota data (active micro-biota relative abundance table) and 3 different types of resting state networks of the brain (the Salience Network (SN); the Executive Control

Network (ECN); and the Default Mode Network (DMN)). The LICA method resulted in 25 individual components indicating how much every data-point per modality contributes to that component. Each data-point's contribution was indicated by a Z-scores i.e. for micro-biota, a Z-score per micro-organisms and for resting state networks a Z-score per voxel). Finally, the contribution of the 4 individual modalities as a whole was also provided in a percentage distribution. This resulting LICA data is fertile ground to start designing and developing our visualizations for this work package.

The LICA method resulted in the following data formats, one for each visualization sub-component:

Micro-biota Z-scores The micro-biota LICA Z-scores are stored in a tabular space separated file, containing a row per micro-organism, a column per component and a Z-score per cell.

Modality contributions The modality contributions are stored in a tabular space separated file, containing a row per modality, a column per component and a percentage per cell.

Resting state network Neuroimaging Informatics Technology Initiative (NIfTI) file Z-scores

For each resting state network type (SN, ECN and DMN) there is a 4D NIfTI, where 1 dimension represents the LICA components and the other 3, the x, y and z coordinates of the voxels. The voxel's Z-score represents the contribution of the small cubic volume of the brain on the x, y, z coordinates.

2.1.2 Visual Analytics Approach

The visual and interactive flow of the web-application takes into account the exploratory nature of the LICA models output, i.e. finding underlying covariance and patterns across the different input modalities, and the order in the causal relationship of the modalities where the gut micro-biota generally influences the brain activity. It starts with an extensive exploratory gut micro-biota matrix visualization at the top, to the modality contribution visualization in the middle to see which of the modalities contribute the most to the selected component and finally at the bottom the brain visualization.

Gut micro-biota matrix visualization. The gut micro-biota visualization (figure 2), is a simple matrix view where every square represents a Z-score for that micro-organism and LICA component combination. This gives a quick overview on which micro-organisms is the most co-active per component. Adjustable threshold values are provided at the right to map the lower positive and negative Z-score values to a grey color to reduce the visual clutter. Clicking one of the cells (or row) will select that specific component which will reveal that component's Z-score brain visualizations of the currently select brain modality (which can be selected through the modality contribution visualization). Currently, there is no way of sorting or filtering the rows and columns to find clusters of micro-organisms, this will be added in later stages of development (see sub section 2.1.4).

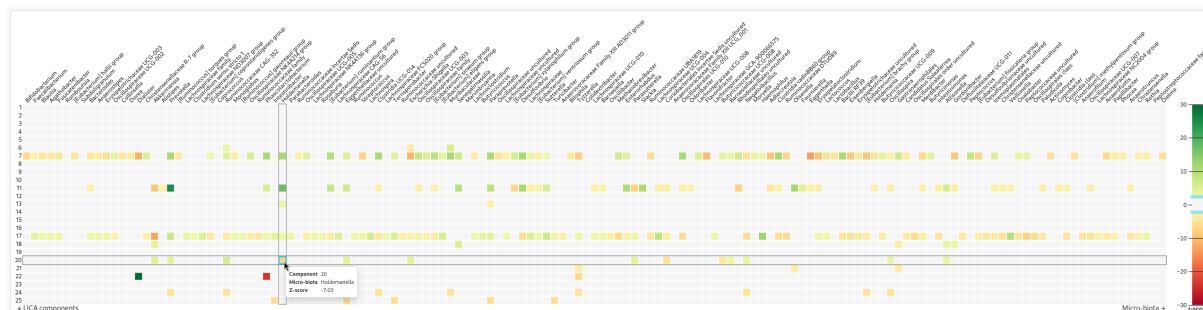


Figure 2: Gut micro-biota matrix visualization

Modality contribution visualization. The simple modality contribution visualization (figure 3) in the middle of the interface, shows the contribution of each modality for the selected LICA component in a stacked bar chart. Each modality in the stacked bar chart can be selected by clicking on them, this will outline in green where in the interface this modality is represented and if it is a brain modality it will show its corresponding brain visualization.



Figure 3: Modality contribution visualization

Conventional brain visualization. The brain visualization (figure 4) shows the anatomical planes (sagittal, coronal and axial) through axial slicing and a volumetric representation of the Z-scores of the selected brain data modality. Each of these visualizations are interactable the same as any standard 3D visualization program like Paraview¹, scrolling through the slices of the anatomical planes and orbiting around and zooming into the volumetric representation. The choice to represent these brain modalities in a conventional way was to give a familiar view to the visualizations for the partners that regularly work with brain data visualization tools like FSLEyes².

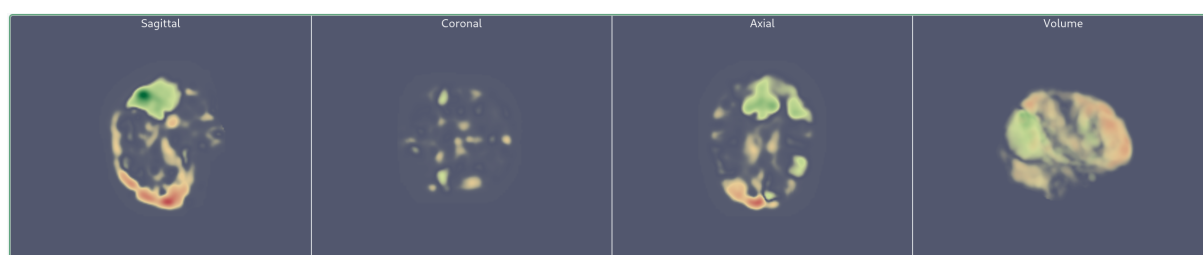


Figure 4: Conventional brain visualization

¹<https://www.paraview.org/>

²<https://fsl.fmrib.ox.ac.uk/fsl/docs/>

2.1.3 Implementation Description (Backend, Frontend, Development Environment)

The web-application was developed in Typescript³ with the following tools and frameworks:

- Vite⁴: For boilerplate code, transpiling, building and running the applications.
- Vue.js⁵: A framework for building web interfaces
- d3⁶: For the non-spatial visualization micro-biota matrix visualization.
- VTK.js⁷: For the spatial brain visualization.
- ITK-wasm⁸: For reading the NIfTI files.

2.1.4 Development Roadmap: What to Develop Next?

The development of this visualization web-application was initiated taking consideration the exemplar use-cases 4 and 5 detailed in the proposal. After finding the right candidate study/dataset an iterative development approach was set in motion. This approach leaves room for discussion and future improvements which led to a set of items for on the road-map.

Combining visualizations with WP5.1. During our WP5 meetings we found that there is a lot of synergy and potential of combining visualizations, either in a single interface or referenced to each other. A singular platform or catalog for the visualizations based on the input data formats of the partners in the project would be highly beneficial.

Flexibility of the input data. During one of our meetings with the partners that were providing the data asked questions about the flexibility of the visualization tool. Currently it is geared towards tabular and NIfTI LICA data, but this can be setup to be more flexible since these are common data formats/types. To safeguard this flexibility we need to have more partner meetings that provide spatial data. This point also synergizes with the previous point, the more flexible the visualizations tool is the less visualization tools the partners need to use.

Visualization component specific feature request and improvements

- Micro-biota matrix visualization:
 - Reducing visual clutter. Currently the matrix has a lot of empty regions where the micro-organism have a low or zero positive/negative contribution to the components and these regions have no added value in the visualizations. Filtering and sorting based on the micro-organisms contributions for a certain component could create a more concise visualization and makes it easier to draw conclusions.
- Brain visualization:

³<https://www.typescriptlang.org/>

⁴<https://vite.dev/>

⁵<https://vuejs.org/>

⁶<https://d3js.org/>

⁷<https://kitware.github.io/vtk-js/>

⁸<https://github.com/InsightSoftwareConsortium/ITK-Wasm>

- Add an atlas reference image to the brain visualizations.
- Make the brain visualizations more conform frequently used NIfTI visualization tools to make it visually more familiar (examples are yet to be received by the research partner).
- Improved comparison of Z-score regions between the different resting state networks. Currently only one resting state network can be selected to be visualized and cross referenced with the micro-biota visualization.

2.2 Droplets: Bio+MedVis Challenge - Redesign (IMAGE, HIGH-DIM)

A novel visualization technique for linked *Image* and *high-dimensional* data was developed in the course of a submission to the Bio+MedVis Challenge⁹ @ IEEE VIS 2024 as part of the research at TU Graz in HEREDITARY. The challenge addressed the important and often-occurring problem of understanding biomedical imaging data with belonging additional high-dimensional data. Specifically, the objective of the redesign track was to improve upon an existing visualization of multi-cell gene expressions of tissue samples. As the cell-type contributions to a sample area constitute proportional relations, a Pie Chart design was used as a baseline (Figure 5(b)). Each of these “spot-wise” distributions is linked to the area in the tissue sample it was computed from (Figure 5(a)), which allows us to superimpose the visualizations to the colored tissue sample in a small-multiple manner (Figure 5(c)).

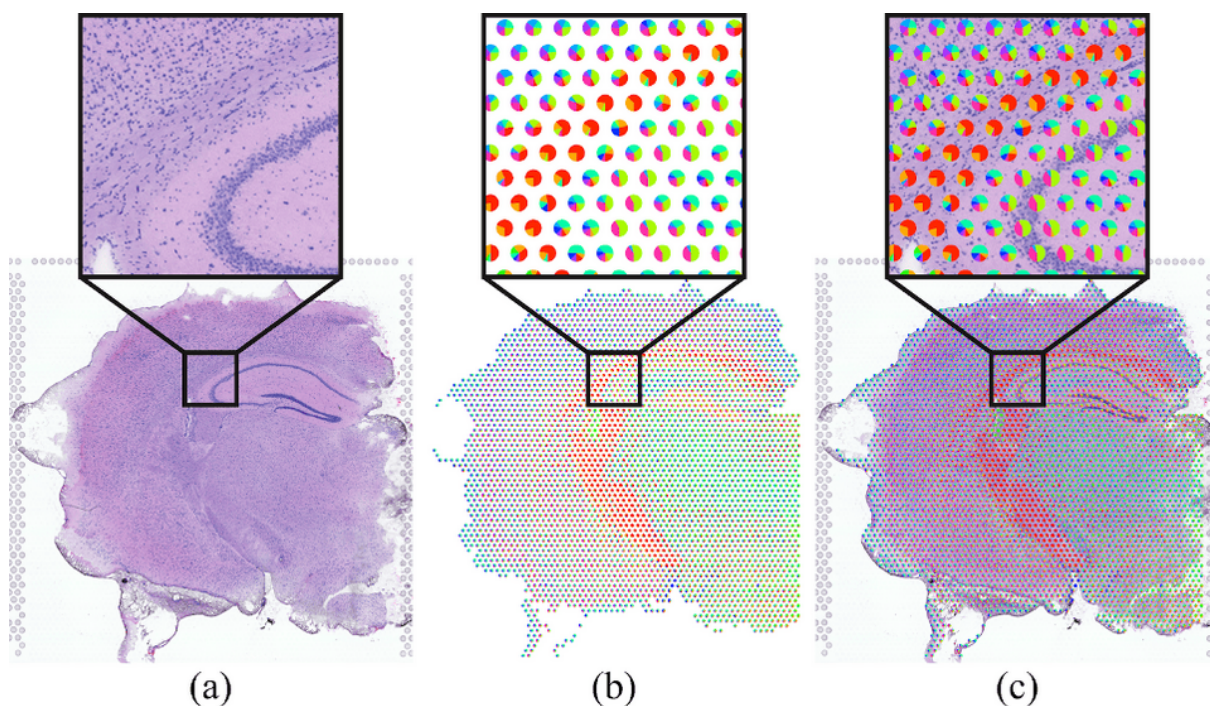


Figure 5: The cell-type proportions of the areas of a tissue sample (a), can be visualized with spot-wise Pie Charts (b), which can be even superimposed to the tissue (c).

⁹http://biovis.net/2024/biovisChallenges_vis/

2.2.1 Considered Data / Test Dataset / Task Description

The dataset, provided by the respective challenge track, contains a single hematoxylin and eosin (H&E) stain of a tissue sample. A deconvolution method [Miller et al., 2022] was employed to obtain proportional cell-type memberships on a per-spot basis – e.g., 30% Cell Type 1, 50% Cell Type 2, etc. A “spot” thereby represents multiple cells. In total, the given dataset comprises 3,499 spots in a regular grid, with 9 distinct cell types. In addition to locations and proportions, the dataset features a list of genes, gene expressions, and a reference image.

2.2.2 Visual Analytics Approach

Principal visualization and interaction techniques The most common visualization for part-to-whole relationships are pie charts. However, they are inherently bad in a small-multiple visualization as the individual contributions are in spatial disarray, making it very hard to perceive intensity clusters (c.f. Figure 5(b)). Other well known designs are bar charts, stacked bar/column charts or treemaps. Those have also been evaluated on the problem at hand (Figure 6), but similarly exhibit various downsides. More advanced approaches for displaying small multiples in their spatial context exist [Mörth et al., 2020; Stoppel et al., 2016], but they are not applicable for showing proportional relations.

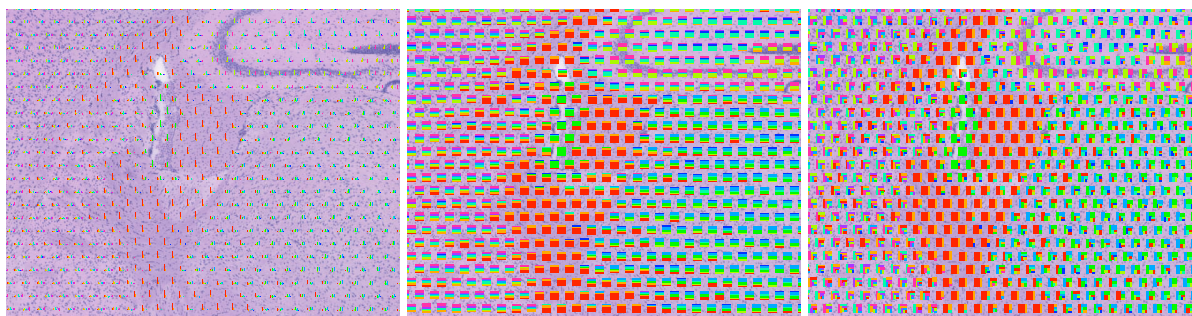


Figure 6: Alternative visualizations of part-to-whole relationships: barcharts, stacked barcharts, and treemaps.

Preliminary requirement / task description One of the fundamental tasks of visual analytics is to recognize latent structures in given data, such as coherent regions. For the problem at hand, we reason that discovering regions of homogeneous cell-type compositions is inherently relevant for domain experts.

Visual analytics design We propose a design which optimally uses the available space, being as translucent as possible (so that the underlying H&E image can be perceived), while still conveying the proportional information. Additionally the component markers should visually reflect their association to local clusters.

The essential idea of our proposed design is to visualize the different cell-type contributions as separated, yet tightly-packed, smooth shapes which reflect both a component’s contribution and its belonging to local accumulations. To this end, a circular base shape is equipped with a protruding spike (tail) pointing towards the center of such a local cluster, with the length of this spike reflecting the size of the respective cluster. This is in accordance with the Gestalt principles of *Continuity* and *Common Fate*. Continuity states that viewers group elements that seem to follow a continuous path – in our case towards a local accumulation. Common Fate refers to viewers’ tendency to group elements which seem to move in a common direction. This marker design emphasizes latent borders of regions with similar cell-type contributions and facilitates the recognition

of – otherwise invisible – coherent regions. The idea is inspired by the winglets analogy by Lu et al. [Lu et al., 2020] for scatter plots, where each point is fitted with dual-sided trails, reflecting the affiliation confidence to a cluster and protruding perpendicular to the direction of the respective cluster. However, to allow for a tight packing we opt for said protruding spike as opposed to such dual-sided trails.

Regarding the positioning of the individual components within a pixel, we intend to arrange them such that they appear to be “drawn towards” their belonging local cluster centers (if any are present). This ensures both that components’ markers are closer to their respective clusters and that their tails are unlikely to obfuscate other markers, while still maintaining a tight packing. I.e., for every pixel location, we need to know the prevalence and directivity of clusters for each of the contributions. To this end, we employ a localized version of the L_1 median [Lipman et al., 2007]. This information is the basis for a physics-based force layouting which tries to arrange all markers in an optimal manner. In more recent iterations of our design, we changed this physical modal to a more efficient algorithmic solution.

2.2.3 Implementation Description (Backend, Frontend, Development Environment)

All implementation is done in a standalone Python script¹⁰, which heavily relies on the *NumPy*¹¹ package for linear algebra. The paths for all markers are computed and saved to an Scalable Vector Graphics (SVG) output.

2.2.4 Demonstration of Preliminary Use Case(s)

The design was evaluated on the contest dataset and presented as a poster [Lengauer et al., 2024] at IEEE VIS 2024 (Figure 7). Even though this use case is very specific, we are convinced that the concept itself can be generalized for a variety of other domains, such as sociodemographic relations, election statistics, vector embeddings. etc. Any data consisting of both proportional relations and locations could be candidate.

Recent developments also showed that the layout could be computed efficiently (a few seconds on commodity hardware), rendering it a viable concept for interactive and dynamic visualizations.

2.2.5 Development Roadmap: What to Develop Next?

Based on the published results [Lengauer et al., 2024], we pursue several tracks for extending and improving the technique. Our first preliminary goal is the translate the existing Python prototype to JavaScript, to clear the way for interactive web-based extensions. We consider several updates and alternative to the base design, combining it with other designs, adding detail-on-demand functionalities, etc.

We, furthermore, plan to demonstrate the capabilities of the droplets visualization for other domains with spatially distributed, but proportional data. These domains include

- landuse data over Europe, with a proportional aspect for each grid cell [Schultz et al., 2024],
- Copernicus vegetation data, which is a high-dimensional spatial dataset with time series and semantically interesting dimensions for each spatial point [Copernicus Land Monitoring Service, 2024], and
- European Parliament plenary voting data [European Parliament, 2024], where we would preprocess the dataset using embedding [Reimers and Gurevych, 2020] to arrive at a spatially distributed proportional dataset. This dataset, however, would require an extension of the droplet algorithm to generalize to non-uniform grids.

¹⁰<https://pluto.cgv.tugraz.at/slengaue/droplets>

¹¹<https://numpy.org/>

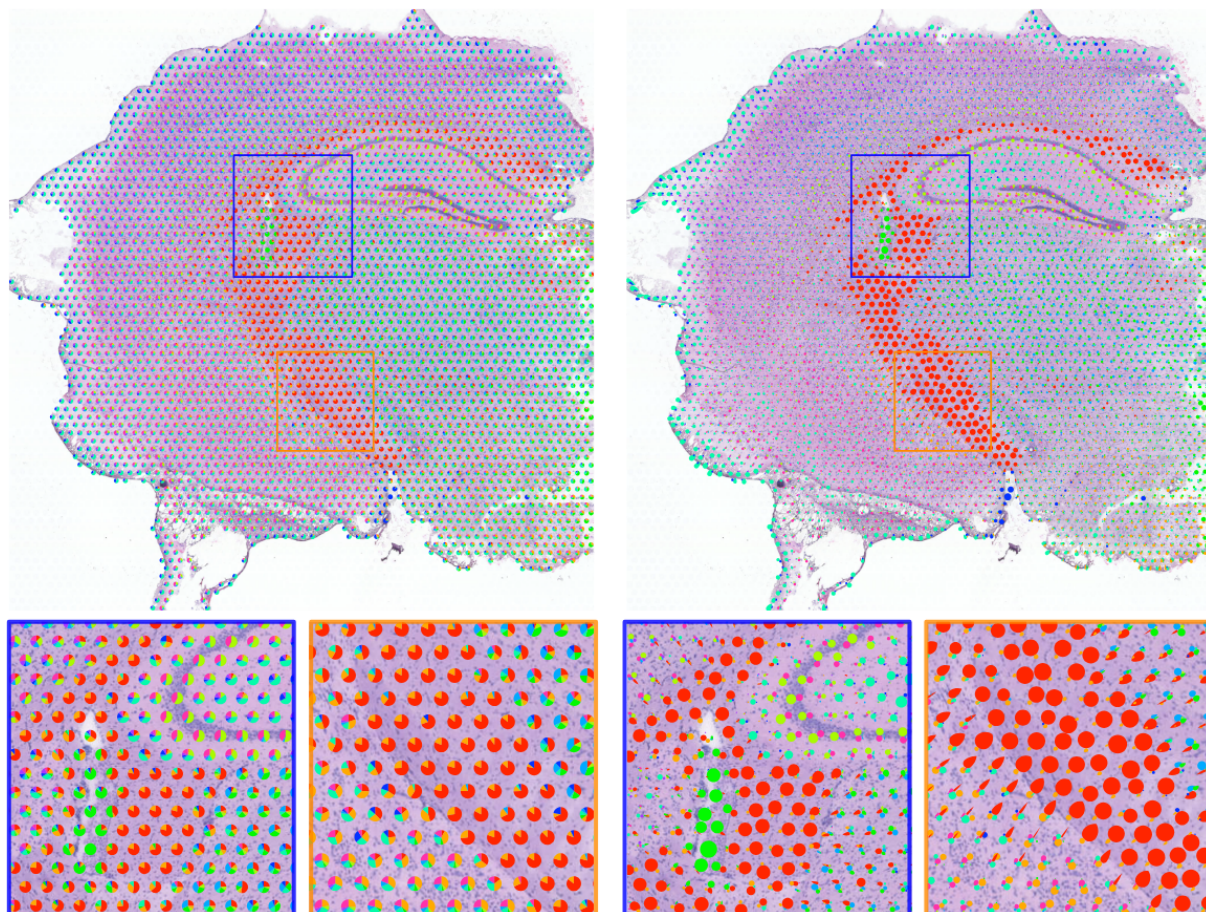


Figure 7: Side by side comparison of the given baseline visualization (left) with our Droplets design, applied to the tissue sample provided by the redesign challenge.

3 Conclusion

During the course of the year that spanned deliverable 5.3, we collaborated with project partners and initiatives that reflected several of the use-cases and which were considered as a good starting point for our initial effort in developing the visualization components. We achieved to conceptualize and develop two visualization components that help to visualize and analyze different types of spatial, imaging and simulated data-sets.

With these achieved visualizations we can initiate more collaboration with project partners to increase the adaptability and responsiveness of our visualizations in the future of WP5. This combined effort can lead to a shared visualization platform which can be utilized to visualize the fast array of different analytical results of the work packages WP3 and WP4.

References

- Copernicus Land Monitoring Service (2024). Vegetation — copernicus land monitoring service. <https://land.copernicus.eu/en/products/vegetation>. (Accessed on 11/13/2024).
- European Parliament (2024). Results of votes | votes | plenary | european parliament. <https://www.europarl.europa.eu/plenary/en/votes.html?tab=votes>. (Accessed on 11/13/2024).
- Groves, A. R., Beckmann, C. F., Smith, S. M., and Woolrich, M. W. (2011). Linked independent component analysis for multimodal data fusion. *Neuroimage*, 54(3):2198–2217.
- Kohn, N., Szopinska-Tokov, J., Llera Arenas, A., Beckmann, C., Arias-Vasquez, A., and Aarts, E. (2021). Multivariate associative patterns between the gut microbiota and large-scale brain network connectivity. *Gut Microbes*.
- Lengauer, S., Waldert, P., and Schreck, T. (2024). DROPLETS: A Marker Design for visually enhancing Local Cluster Association.
- Lipman, Y., Cohen-Or, D., Levin, D., and Tal-Ezer, H. (2007). Parameterization-free projection for geometry reconstruction. *ACM Trans. Graph.*, 26(3):22–es.
- Lu, M., Wang, S., Lanir, J., Fish, N., Yue, Y., Cohen-Or, D., and Huang, H. (2020). Winglets: Visualizing association with uncertainty in multi-class scatterplots. *IEEE TVCG*, 26(1):770–779.
- Miller, B. F., Huang, F., Atta, L., Sahoo, A., and Fan, J. (2022). Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. *Nature communications*, 13(1):2339.
- Mörth, E., Haldorsen, I. S., Bruckner, S., and Smit, N. N. (2020). Paraglyder: Probe-driven interactive visual analysis for multiparametric medical imaging data. In *Advances in Computer Graphics*, pages 351–363. Springer.
- Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Schultz, M., Li, H., Wu, Z., Wiell, D., Auer, M., and Zipf, A. (2024). OpenStreetMap land use for Europe “Research Data”.
- Stoppel, S., Hodneland, E., Hauser, H., and Bruckner, S. (2016). Graxels: information rich primitives for the visualization of time-dependent spatial data. In *VCBM/MedViz*, pages 183–192. Eurographics Association.