scFlowVis: Streamlining scRNA-seq Analysis through Visual Design

Yiwen Xing [1] Stanley Owomero [1] Sophia Tsoka [1] Rita Borgo [1] Alfie Abdul-Rahman [1]

[1]Department of Informatics, King's College London

Introduction

As the demand for single-cell RNA sequencing (scRNA-seq) analysis grows, a standardized workflow has been established, encompassing preprocessing, dimensionality reduction, clustering, and advanced analysis methods. Numerous data processing and visualization toolkits have been developed to support this workflow, primarily focusing on specific pipeline steps. Integrated platforms such as Galaxy [4] offer tools for various analysis purposes within a single platform. However, they lack seamless integration, making it difficult for users to experiment iteratively between different analysis stages. Python and R notebooks, while powerful and capable of supporting the entire workflow, rely heavily on the user's programming skills. Despite these advancements, there is still a notable lack of visualization tools that span the entire analysis pipeline without requiring extensive programming knowledge. Additionally, tools are significantly needed to support workflow documentation and analysis operation recording [2]. These gaps in the domain motivated us to collaborate closely with domain experts to develop *scFlowVis*, a web-based visualization tool designed to streamline the scRNA-seq analysis pipeline.

Background

scRNA-seq Analysis Pipeline: The field of computational biology has developed sophisticated methods for single-cell RNA sequencing (scRNA-seq) analysis, including quality control, normalization, dimensionality reduction, clustering, trajectory inference, and differential expression analysis, forming the core of a typical scRNA-seq data analysis pipeline [6].

Visualization Techniques: Provenance tracking is crucial across domains for documenting and sharing analytical workflows [3], yet it is underutilized in scRNA-seq analysis tools. Visual programming can simplify the coding process through graphical elements, making it accessible for beginners [5]. It encompasses various graphical presentation formats such as flowcharts and node-based interfaces [1]. However, it is challenging scalability.

Research Goals: In collaboration with domain experts, we established three goals: 1) ***support seamless conduction of the entire data analysis process***, 2) ***support decision-making alongside the entire analysis process***, and 3) ***support workflow and result presenting.***

Requirements Analysis

Derived from the goals, specific requirements were summarized:

R1 *Integrating sufficient data processing and visualization methods to accommodate the entire analysis process.*

R2 *Implementing data structure tracking mechanisms and enabling checking at every analytical phase.*

R3 *Documenting the workflow & supporting visual comparisons.*

R4 *Recommending suitable methods at different analytical phases.*

R5 *Enabling workflow presenting & sharing.*

R6 *Reducing coding workload.*

scFlowVis Design Strategy

scFlowVis is the output of a sustained iterative design study conducted in collaboration with researchers in the scRNA-seq analysis domain. It is a comprehensive visualization tool that assists researchers throughout the entire scRNA-seq analysis pipeline. Integrating a wide range of mainstream data processing and visualization methods, it categorizes them based on the structured workflow informed by domain experts' insights. *scFlowVis* allows for real-time querying and exploration of data and metadata and features a visualization recommendation engine that suggests suitable methods based on data attributes. The tool offers four distinct visualization display layouts, providing versatile data presentation options. All functionalities are seamlessly integrated, enabling users to import data, conduct analyses, and obtain ready-to-use visual results without leaving the interface.

To develop a visual analytical tool like this, where panels are closely integrated and essential to the entire system, we adopted a design strategy (Fig. 1) that simultaneously considers the functionality and positioning of each module.

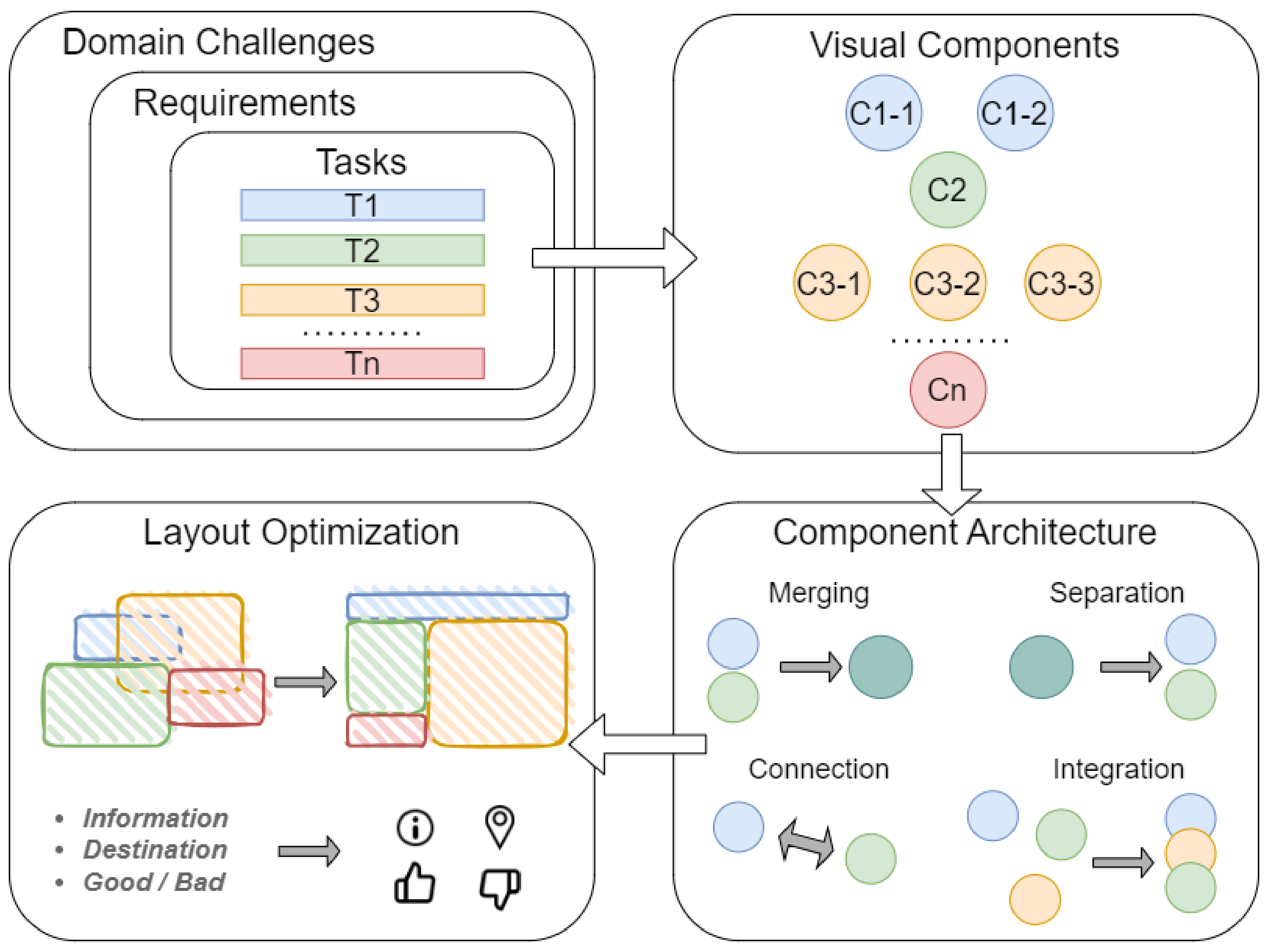
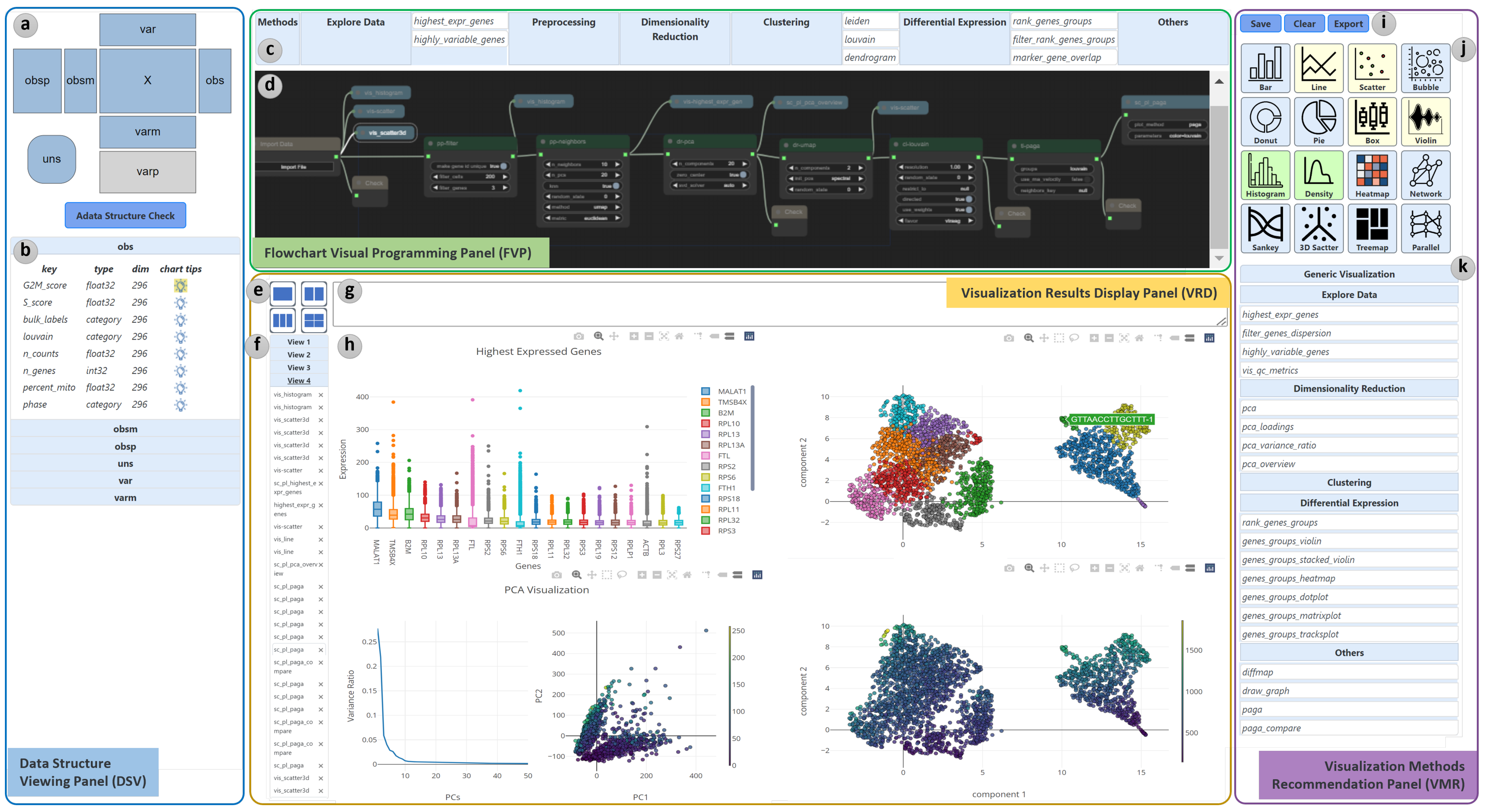


Figure 1. The interface design strategy.

Interface and Components



obs

key	type	dim	chart tips
G2M_score	float32	296	💡
S_score	float32	296	💡
bulk_labels	category	296	💡
louvain	category	296	💡
n_counts	float32	296	💡
n_genes	int32	296	💡
percent_mito	float32	296	💡
phase	category	296	💡

obsm
obsp
uns
var
varm

Figure 2. The interface of *scFlowVis*, highlighting its key components.

Data Structure Viewing Panel: is designed with requirement R2 and the complexities of scRNA-seq data in mind. Using the AnnData [7] structure as a prototype, it provides a thumbnail overview (Fig. 2 a) that highlights key data components and relationships. A detailed data information list (Fig. 2 b) below displays variable information. The DSV panel integrates with other panels, allowing users to dynamically inspect data structure changes and access visualization recommendations via trigger buttons.

Flowchart Visual Programming Panel: serves as the central operational interface in *scFlowVis*, combining flowchart design with visual programming. Users can construct and document their analysis pipeline by connecting blocks for different operations, reducing the programming workload and enhancing workflow presentation (R1, R3, and R6). The panel features a library of scRNA-seq data processing methods (Fig. 2 c) and a user operation board (Fig. 2 d) for adding and linking blocks. It integrates with other panels to reflect updates and maintain consistency across the tool.

Visualization Methods Recommendation Panel: supports data visualization needs with a button group (Fig. 2 i) for managing the workflow and an export function for sharing workflows (R5). The core feature is the visualization recommendation engine (Fig. 2 j), which suggests appropriate methods based on data attributes using a rule-based algorithm (R4). Users can incorporate recommended visualization blocks into their flowcharts, ensuring a seamless transition from data exploration to visualization. An organized list of visualization methods (Fig. 2 k) groups techniques by analysis stage for easy access (R1).

Visualization Results Display Panel: presents all visualization outcomes, designed with four main functional areas. Users can customize the display area (Fig. 2 h) with various layout options to compare multiple visualizations simultaneously (R3). The visualization list (Fig. 2 f) stores results chronologically, allowing users to activate and display them as needed. Interactive charts enable data engagement, with selected elements displaying information in a textbox (Fig. 2 g), which also serves as a note-taking area to integrate data interaction with user insights.

Discussions

Discussions with domain experts highlighted some limitations of *scFlowVis*: 1) its methodological scope, while comprehensive, lacks the latest and deep learning-related methods, and 2) its local-only installation limits sharing and exporting workflows due to the absence of public server deployment. To address data privacy concerns and backend API accessibility, we maintained local installation while releasing the code to an open-source repository. This approach allows users to deploy the tool locally or on their servers, balancing accessibility and privacy and fostering continuous enhancement and community contribution.

Future Work

For future work, we aim to expand *scFlowVis* in two key areas: 1) updating and enriching data processing methods, particularly integrating the latest techniques, and 2) extending the tool's applicability across various bioinformatics domains beyond scRNA-seq analysis, based on user feedback. By making *scFlowVis* open-source, we encourage community contributions to incorporate more methods and enhance functionalities.

References

[1] Margaret M. Burnett and David W. McIntyre. "Visual Programming". In: *Computer* 28.3 (Mar. 1995), pp. 14–16. ISSN: 0018-9162. DOI: `10.5555/618995.620167`.

[2] Batuhan Cakir et al. "Comparison of visualization tools for single-cell RNAseq data". In: *NAR Genomics and Bioinformatics* 2.3 (2020), lqaa052. DOI: `10.1093/nargab/lqaa052`.

[3] Mehdi Chakhchoukh, Nadia Boukhelifa, and Anastasia Bezerianos. "Understanding How In-Visualization Provenance Can Support Trade-Off Analysis". In: *IEEE Transactions on Visualization and Computer Graphics* 29.9 (2023), pp. 3758–3774. DOI: `10.1109/TVCG.2022.3171074`.

[4] Vahid Jalili et al. "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update". In: *Nucleic Acids Research* 48.W1 (June 2020), W395–W402. DOI: `10.1093/nar/gkaa434`.

[5] B. A. Myers. "Visual programming, programming by example, and program visualization: a taxonomy". In: *SIGCHI Bull.* 17.4 (Apr. 1986), pp. 59–66. ISSN: 0736-6906. DOI: `10.1145/22339.22349`.

[6] Tim Stuart et al. "Comprehensive integration of single-cell data". en. In: *Cell* 177.7 (June 2019), 1888–1902.e21. DOI: `10.1016/j.cell.2019.05.031`.

[7] Isaac Virshup et al. "anndata: Annotated data". In: *bioRxiv* (Dec. 2021). DOI: `10.1101/2021.12.16.473007`.

IEEE VIS 2024 Poster ***scFlowVis*** **has been released on** `https://github.com/EavanXing0416/scFlowVis` **Email: yiwen.xing@kcl.ac.uk**