DQVis Dataset: Natural Language to Biomedical Visualization

Devin Lange
Harvard Medical School
devin@hms.harvard.edu

Pengwei Sui
Harvard Medical School
pengwei_sui@hms.harvard.edu

Shanghua Gao
Harvard Medical School
shanghua_gao@hms.harvard.edu

Marinka Zitnik
Harvard Medical School
marinka@hms.harvard.edu

Nils Gehlenborg
Harvard Medical School
nils@hms.harvard.edu

Abstract

Biomedical research data portals are essential resources for scientific inquiry, and interactive exploratory visualizations are an integral component for querying such data repositories. Increasingly, machine learning is being integrated into visualization systems to create natural language interfaces where questions about data can be answered with visualizations, and follow-up questions can build on the previous state. This paper introduces a framework that takes abstract low-level questions about data and a visualization grammar specification that can answer such a question, reifies them with data entities and fields that meet certain constraints, and paraphrases the question language to produce the final collection of realized data-question-visualization triplets. Furthermore, we can link these foundational elements together to construct chains of queries, visualizations, and follow-up queries. We developed an open-source review interface for evaluating the results of these datasets. We applied this framework to five biomedical research data repositories, resulting in DQVis, a dataset of 1.08 million data-question-visualization triplets and 11.4 thousand two-step question samples. Five visualization experts provided feedback on the generated dataset through our review interface. We present a summary of their input and publish the full reviews as an additional resource alongside the dataset. The DQVis dataset and generation code are available at `https://huggingface.co/datasets/HIDIVE/DQVis` and `https://github.com/hms-dbmi/DQVis-Generation`.

1 Introduction

Natural language interfaces show promise for exploratory data analysis [42]. The central engine of such a system converts a natural language query into a visualization. There are different approaches to this engine, but many require training data in the form of natural language queries and visualization responses (NL2VIS). With modern techniques, such datasets carry various requirements. First, they must be large enough to fine-tune an LLM. Second, many applications are tailored towards specific domains, often requiring the questions and visualizations to be rooted in this domain. Finally, the goal of NLI is not to answer a single question with a visualization, but to provide an interface for exploring data, where the past questions and visualizations are known by the NL2VIS generation

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

engine, in other words, multi-step reasoning. All of these requirements for sufficient training data coalesce into a challenging, labor-intensive task.

Such training data is worthwhile because of the potential value of NLI. In biomedical research, large research consortia typically build and curate data repositories to collect and distribute data generated in the course of collaborative multiyear projects. Researchers exploring these repositories can lead to discoveries explaining the mechanisms of human biology and the development of new treatments for diseases like cancer [3, 14, 8]. Many biomedical research data repositories contain sophisticated interfaces on their data portal for interacting with the data[44, 10, 37]. However, such interfaces require time and expertise to develop, and it is difficult to account for every possible way a user may want to query the data and design visualizations and interfaces that can accommodate all of these. Here is the promise of natural language interfaces. An ideal system can bridge the gap between diverse user queries and one-size-fits-all interfaces by responding to each individual with the response to the exact question they have, whether that's a developer trying to understand the size and type of datasets on the repository, or a researcher seeking the next critical piece of information in the path to solve cancer.

Building an ideal NLI interface is challenging with insufficient training data. Unfortunately, the construction of datasets requires time, expertise, and computational resources. LLMs can be fragile when dealing with terms in evolving specialized domains [13], such as new drug names or biological assays. Thus, any curation of a domain-specific training dataset will require expertise in the construction and review of the data. The time required for this construction is exacerbated by the large scale of data required — experts can't construct all the required data manually. Therefore, computational resources are often employed to assist in the creation of these large-scale datasets. Finally, supporting multi-step further complicates the task by introducing branching complexity of possible paths.

NVBench [30] is an existing dataset over 25K data points and has already shown utility in training NL2VIS models [31]. However, as a domain-agnostic dataset, it may not be appropriate for domain-specific applications. Additionally, although the dataset size was sufficient for techniques a few years ago, it may not scale to popular modern-day approaches, such as fine-tuning LLMs. Finally, NVBench does not consider multi-step reasoning.

We work towards addressing these challenges with the **two main contributions of this paper**, a data-question-visualization generation framework, and a new dataset (DQVis) for biomedical research repositories constructed with this framework. DQVis includes **1.08 million data-question-visualization triplets** that address the challenge of domain specificity and data scale for this particular domain. Additionally, we provide **11.4K two-step question samples** that are representative of common visual exploration tasks. The framework we introduce eases the burden for other domains to create datasets by taking as input a list of dataset schemas and a relatively small number of abstract questions, reifying those questions with the provided data, and paraphrasing them to produce a massive number of resulting data-question-visualization data points. We also include an interface for reviewing generated data as part of this framework. Finally, we illustrate how data from DQVis can be linked together, forming multi-step chains of user queries and visualization responses.

2 Related Work

2.1 Natural Language to Data Visualization

Data visualization is a ubiquitous way for people to interact with data. However, selecting a visualization manually for a given dataset is tedious. Thus, the visualization community has long researched visualization recommendation systems that can automatically produce visualizations for a given dataset or integrate AI into visualization systems [34, 54, 12, 36, 55]. Taking it further, natural language interfaces introduce a system that allows users to produce or update visualizations [42]. This technique can be used to generate visualizations given a prompt [6, 35, 31], to update interactive visualization systems [41, 19, 21, 57], and to support visualization authoring systems [52, 43, 32]. In the last few years, the use of LLMs for such systems has exploded [32, 28, 50, 26, 59, 29, 43, 56, 49, 24, 11]. In addition, work has been done to understand what the visualization design preferences of LLMs are [51], their promises and pitfalls [7], and how to evaluate them [5]. Many of these systems are designed for a specific domain, such as health information seeking [56], education [15], and sports [26, 29, 59]. Furthermore, many are tailored towards a specific data type, like graphs [45] or trajectories [21]. Although some systems are able to use standard LLMs, others fine-tune the LLM [16, 15], highlighting the need for domain-specific datasets that can readily be used for this task.

Natural-language-to-visualization (NL2VIS) requires models to analyze and transform user queries expressed in natural language into visualizations. NVBench [30] repurposed Spider [58] to generate Vega-Lite specifications from text, focusing on the general domain. NVBench is an important contributions to the field, however, DQVis covers additional challenges related to scale, domain, and multi-step queries. NVBench includes 25,750 datapoints, whereas DQVis includes 1,075,190 datapoints, a more than 40x increase in scale. DQVis is a domain-specific dataset centered around biomedical research repositories, whereas NVBench is a general-purpose dataset. Finally, DQVis contains multi-step data, which is increasingly vital for conversational LLMs. NVBench does not include multi-step data. Dial-NVBench builds on the NVBench dataset to include dialog-style constructions of data visualizations [46], however, it does not cover address the challenge of scale and domain. VizNet [20] assembled a large repository of real-world examples across general datasets but did not consider natural language input. Srinivasan et al. [47] collected specification language utterances for describing data visualizations. AVA [28] introduced iterative refinement of visualizations using multimodal agents, although this was limited to narrow domains. While these works focus on general-purpose visualization datasets, they do not address biomedical research data, which requires an understanding of domain-specific knowledge. In contrast, our dataset is specifically designed for the biomedical domain. Furthermore, we account for the complexity of biomedical visualization queries, which often demand detailed and multi-step reasoning to produce accurate and meaningful visualizations.

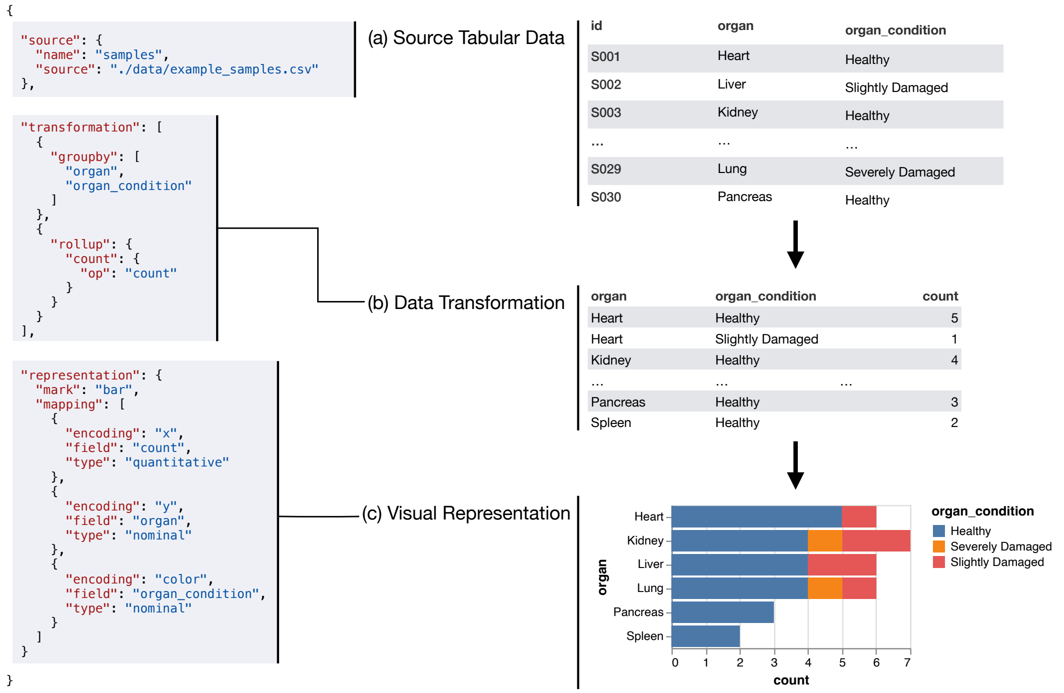
2.2 Data Visualization Question Answering

Data visualization question answering (DVQA), inspired by visual question answering (VQA) aims to train models to answer questions about visualizations [18, 9]. The goal of DVQA is similar to our aim of answering questions about data *with* visualizations, but also has critical distinctions. The high-level approach of creating datasets [23, 22, 2] in order to train models [2] is similar to our goals. Although some data retrieval questions could exist in both DVQA and DQVis datasets, e.g. "What category contains the most records?" could be asked of a dataset, and a bar chart of category record counts, the *best* visualization for answering this retrieval question may not be the same as a visualization that could contain the answer. Additionally, some questions should exist in an NL2VIS dataset, but not in a DVQA dataset. One example is questions holistically characterizing data, e.g. "What is the distribution of values" can be answered with a histogram, but asking "What is the distribution of values" for a histogram is not an appropriate question. Some questions can exist in DVQA, but not in DQVis. In particular, questions about the structure of a chart "Is the y-axis scale linear?" fall into this category [22]. Furthermore, while DVQA datasets can omit the dataset the data visualization is representing, DQVis must include the data since this is the artifact on which the questions are posed. Still, some aspects can be applied to both scenarios. Ko et al. introduce a framework for generating DVQA datasets, and we use a similar technique for paraphrasing in our framework, inspired by their work [25]. In short, DVQA and DQVis are both important areas with overlap, but one cannot simply reverse a dataset in one domain to produce the other.

3 Domain Background

The framework we developed is designed to support creating domain-specific datasets. In this paper, we illustrate its utility for biomedical research data portal metadata. The details of data portals vary, but a common theme is that they include metadata on **donors** (e.g., humans or mice) who provide biological **samples** (e.g., blood or tissue), which are analyzed and result in a **dataset**. We distinguish the metadata from the datasets themselves and focus on metadata in this work. Whereas the datasets contain the results of various biological assays, the metadata records information about the datasets, e.g., which assay was run and how large the dataset is. Datasets are essential for deep analysis of an individual sample, and metadata is needed to identify relevant datasets or understand patterns across multiple samples.

We use an entity relationship model to represent the metadata. **Entities** (E) correspond to a data table, such as donors. The **fields** (F) correspond to columns in the data table that represent attributes of the entity, such as weight or height. Fields can have different data types, such as *quantitative*, *ordinal*,



```
{
  "source": {
    "name": "samples",
    "source": "./data/example_samples.csv"
  },

  "transformation": [
    {
      "groupby": [
        "organ",
        "organ_condition"
      ]
    },
    {
      "rollup": {
        "count": {
          "op": "count"
        }
      }
    }
  ],

  "representation": {
    "mark": "bar",
    "mapping": [
      {
        "encoding": "x",
        "field": "count",
        "type": "quantitative"
      },
      {
        "encoding": "y",
        "field": "organ",
        "type": "nominal"
      },
      {
        "encoding": "color",
        "field": "organ_condition",
        "type": "nominal"
      }
    ]
  }
}
```

(a) Source Tabular Data

id	organ	organ_condition
S001	Heart	Healthy
S002	Liver	Slightly Damaged
S003	Kidney	Healthy
...	...	...
S029	Lung	Severely Damaged
S030	Pancreas	Healthy

(b) Data Transformation

organ	organ_condition	count
Heart	Healthy	5
Heart	Slightly Damaged	1
Kidney	Healthy	4
...	...	...
Pancreas	Healthy	3
Spleen	Healthy	2

(c) Visual Representation

Figure 1: The biomedical visualization grammar we are developing has three main components: (a) specifying one or more sources of tabular data; (b) transforming that data by filtering, grouping, and aggregating; (c) mapping transformed data to visual channels.

and *nominal*. Entities can also be related through identifying columns as one-to-one, one-to-many, or many-to-many. For instance, you might expect a biological sample to have a one-to-one or many-to-one relationship with donors.

There are many tools and libraries for visualizing such data. In our case, we focus on a visualization grammar we are developing for biomedical data portal exploration.[1] This grammar is defined by a JSON Schema, which defines how visualizations can be constructed with a declarative JSON specification (spec). Our visualization toolkit renders these specifications into interactive visualizations (see Figure 1).[2] At a high level, this specification will define one or more tabular sources of data (e.g., CSV files), how the data is transformed (e.g., grouped and aggregated), and how the transformed data is mapped to visual channels (e.g., color and position). This grammar of graphics approach is popular among the visualization community [53, 39, 33]. In particular, our grammar is similar to a popular library, vega-lite [39], and uses it for part of the rendering implementation. The most significant difference is that our grammar includes additional support for tabular representations (see Figure 2e), an essential component of biomedical research data portals.

4 Data Description

The core component of the DQVis dataset is triplets of **data** (D), **query** (Q) that could be posed about such data, and **visualizations** (Vis) that could answer those questions. The **data** column references an entity relationship definition. The **query** is in the form of a natural language query a user may have about a dataset. Finally, the **visualization** is a JSON specification for our grammar. DQVis contains 1.08 million data-query-visualization triplets across various data repositories and chart complexity, sumarized in Table 1. Examples of generated questions and visualizations is shown in Figure 2.

[1] https://hms-dbmi.github.io/udi-grammar
[2] https://github.com/hms-dbmi/udi-grammar

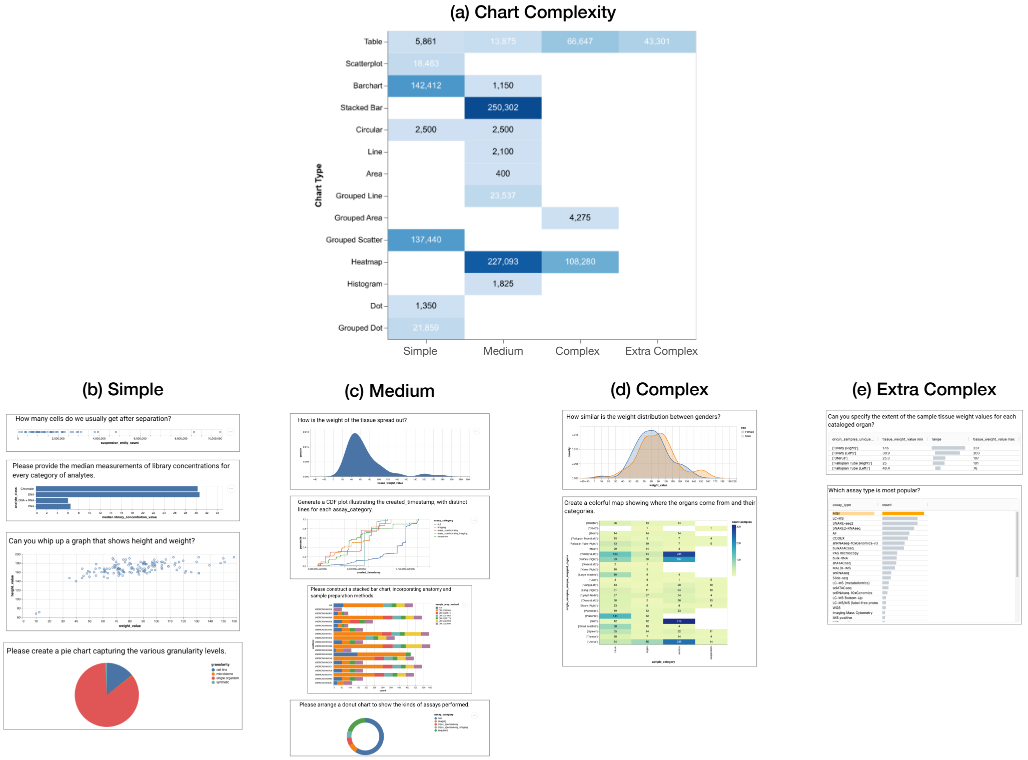


Figure 2: (a) DQVis contains 14 chart types with varying degrees of complexity. Complexity is determined by the number of keys in the visualization specification. (b) Simple visualizations have <= 12 keys; (c) Medium <= 24; (d) Complex <= 36; and (e) Extra Complex is > 36.

	Dataset Dimensions		**Chart Complexity**			
Repository	Entities	Fields	Simple	Medium	Complex	Extra Complex
HuBMAP	3	320	321,069	514,813	147,680	41,695
MW	22	99	3,375	3,994	10,950	681
4DN	20	101	2,661	1,800	10,225	525
MoTrPAC	14	68	2,075	1,525	7,000	275
SenNet	6	35	725	650	3,347	125

Table 1: DQVis includes data from 5 different biomedical research data repositories: HuBMAP [44], MW[48], 4DN[10, 37], MoTrPAC[38], and SenNet [27]. The different datasets contain a varied number of Entities and Fields, visualized with different types of chart complexity.

The entity relationship model for the biomedical research *data* is saved in a Frictionless Data Package.[3] This standard provides a consistent definition for entity relationships and fields within each entity and can be extended with additional information. For our framework, we extend data packages to include other useful information for each field, such as the number of unique values recorded. The Common Fund Data Ecosystem (CFDE [4] publishes metadata packages for multiple data portals in the Crosscut Metadata Model (C2M2) [4] format, which also adheres to the Frictionless Data Package standard. DQVis contains five different data packages from biomedical data portals. SenNet [27], Metabolomics Workbench (MW) [48], MoTrPAC [38], and 4DN [10, 37] are in the C2M2 format. Although HuBMAP [44] also has a C2M2 package, we instead created a data package from donor, sample, and dataset metadata since more information was available on the data portal.

[3] https://datapackage.org/standard/data-package/

[4] https://commonfund.nih.gov/dataecosystem

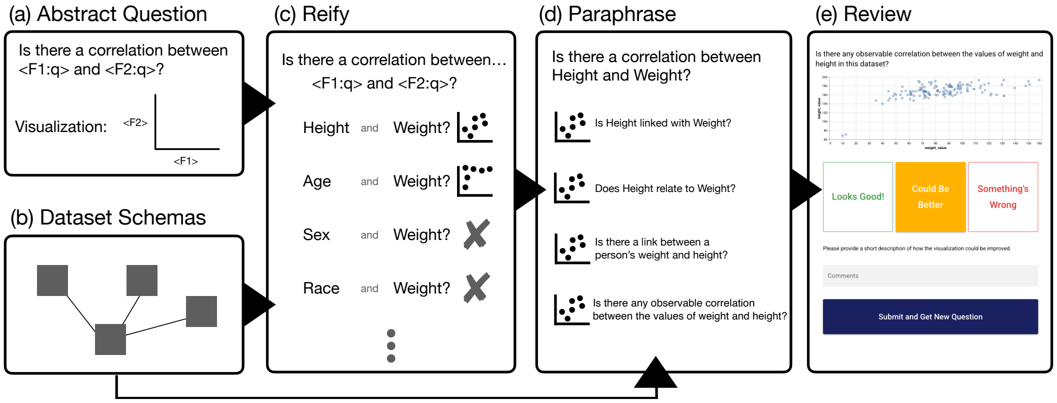

(a) Abstract Question
Is there a correlation between <F1:q> and <F2:q>?
Visualization:
<F2>
<F1>
(b) Dataset Schemas
(c) Reify
Is there a correlation between… <F1:q> and <F2:q>?
Height and Weight?
Age and Weight?
Sex and Weight?
Race and Weight?
(d) Paraphrase
Is there a correlation between Height and Weight?
Is Height linked with Weight?
Does Height relate to Weight?
Is there a link between a person's weight and height?
Is there any observable correlation between the values of weight and height?
(e) Review
Is there any observable correlation between the values of weight and height in this dataset?
height_value
weight_value
Looks Good!
Could Be Better
Something's Wrong
Please provide a short description of how the visualization could be improved.
Comments
Submit and Get New Question

Figure 3: Overview of data synthesis pipeline. (a) Abstract questions are written as templates with placeholders and visualizations that reference those placeholders. (b) Dataset schemas define entity relationship models. (c) Templates are combined with data and placeholders are resolved as long as they satisfy all constraints. (d) The question is paraphrased with an LLM to produce more diverse questions. (e) The final data is reviewed in our review software for issues or potential improvements.

Since HuBMAP includes more data fields than the other data packages, the majority of the queries in DQVis are about HuBMAP (see Table 1).

The natural language $query$ can be in the form of a specific question about the data or an utterance requesting a specific visualization of the data. The query will always include references to entities or fields from the dataset. For instance, "How many donors are there?" would refer to the donor entity, and "What is the average age?" would refer to the age field in the donor entity. A more specified question would be "What is the average age of donors?", but omitting the entity in a query is plausible and can be inferred. It is also possible to ask questions that require multiple entities. For instance, "How many samples are there for each sex?" is asking for the count of records in the sample entity grouped by the sex field in the related donor entity. Queries that are utterances will typically reference a specific chart type, e.g. "Make a scatterplot of height and weight." Such queries can have implied underlying questions, e.g. "Is there a correlation between height and weight?" or can be a step in a more open-ended data exploration process. The answers to these queries for DQVis are visualization specifications ($spec$) in the form of our biomedical visualization grammar.

In addition to the data-question-vis triplets, additional information related to the data creation is recorded in the DQVis dataset. The $query\_template$ and $spec\_template$ values are described in Section 5.1; **query_base** and $constraints$ in Section 5.2; and $expertise$ and $formality$ in Section 5.3.

5 Data Synthesis Pipeline

Synthesizing the DQVis dataset of 1.08M data-question-visualization triplets consists four major steps (see Figure 3). First, templates define abstract queries and visualization answers. Next, these are reified with data entities and fields that meet imposed constraints. Then the base query is paraphrased to produce a diverse phrasing of the same question. Finally, the resulting data is reviewed both iteratively to refine and debug issues and to capture expert visualization and domain knowledge.

5.1 Abstract Queries and Visualizations

The goal of this pipeline is to capture a wide range of queries that could be posed for a dataset. Our work focuses on low-level analysis tasks that comprise a more involved analysis session. Amar et al. identify ten types of abstract low-level analysis tasks that "cover the vast majority of the corpus of analytic questions [they] studied." [1]. DQVis includes 64 abstract queries that span these ten types of tasks. (Computed Derived Value: 39, Determine Range, 14, Correlate: 4, Cluster: 5, Find Anomalies: 3, Find Extremum: 9, Retrieve Value: 8, Sort: 1, Characterize Distribution: 10, Filter: 10). Some abstract queres relate to multiple low level tasks, especially utterance queries.

These abstract queries and visualizations are in the form of a template with entity and field placeholders. So instead of a dataset specific question “How many biological samples are there for each organ?” the template questions would be “How many <E> are there for each <F>?” where the <E> takes the place of the entity and the <F> takes the place of the data field. This template question is stored as the $query\_template$ in DQVis. Next, a visualization specification is produced that includes these placeholders, which is stored in the $spec\_template$ column.

5.2 Reify Queries and Visualizations

To reify questions, we replace the entity and field placeholders with concrete entity and field names. However, abstract questions cannot be posed to every possible entity and field. Some questions cannot be logically resolved, like asking what the maximum value is for a nominal field. To account for this, abstract question-vis pairs also include constraints that limit the applicable entities and fields. Constraints are question-specific, though some constraints are applied to multiple questions. There are between 1 and 11 constraints for each question. Every possible question that satisfies all of the constraints is generated with a constraint satisfaction solver library. Depending on the input data and constraints this could lead to an combinatorial explosion of possible solutions. However, for this is not the case for the DQVis dataset and was run on a single personal laptop. The most common type of constraint includes constraints on the field types. Thus, these have a special shorthand in our framework. e.g. “How many <E> are there for each <F:n>?” will add a type constraint that <F> must be a nominal data type. There are additional implied constraints in this question related to entity field relationships. In this example, <F> must exist within <E>. Or, in a more complex case “How many <E1> are there for each <E2.F:n>”, <F> exists in <E2> and <E1> and <E2> are two different entities.

Additional constraints are defined as a list in the $constraints$ column of DQVis. Each constraint is a boolean expression that allows references to the placeholders (E1, E2, E1.F, E2.F1, E2.F2, etc.) as well as attributes of those entities and fields defined on the data package. For instance, F.c will resolve to the cardinality (unique number of values), in that column. Imposing these constraints helps ensure that the questions are logical and that the visualization responses adhere to visualization design best practices (e.g., limiting the number of distinct categorical values encoded with color). Constraints are also essential for handling different types of data transformations required for the visualization. For instance “How many <E> are there for each <F:n>?’ and “How many <E1> are there for each <E2.F:n>” are similar questions with visually similar representations. However, the latter question requires a more complex specification that combines data from multiple data entities. Constraints can ensure that entities have a defined relationship and the type of relationship (e.g., E1 $\rightarrow$ E2 is one-to-one or many-to-one).

Once constraints are defined, reifying the abstract queries on dataset schemas is formulated as a constraint satisfaction problem. Find all entity-field combinations in the dataset schemas that meet all of the constraints, resulting in a list of solutions that map abstract entities and fields to real entity and field names. Then, for each solution, a data point is created that replaces the template placeholders with these solutions. These are included in the $query\_base$ column for queries and $spec$ for the final specification result.

5.3 Query Paraphrasing

Replacing the placeholder entities and fields with real entity and field names will result in real, but repetitive queries, e.g. “How many donors are there for each sex?”, “How many donors are there for each race” and “How many samples are there for each organ” will result in some grammatically incorrect queries. E.g. “How many donor are there for each sex”. Paraphrasing these queries with an LLM rectifies both of these issues.

We use Ko et al.’s [25] definition for question technicality and formality of language and use a similar technique for paraphrasing across these dimensions We expand the approach to also include the relevant dataset schema in the prompt template for the LLM. Since the dataset schemas include descriptions of entities and fields, this increases the potential of the LLM to generate better paraphrased queries. In addition to varying the style of query, this also replaces exact variable names with synonyms. For instance, if a donor entity has a field named “age_value”, we want the phrases “construct a distribution plot of ’age_value”’, “show me the age distribution of donors”, and “How old are the donors?” to all result in the same plot.

For the DQVis dataset, we used gpt-4o and varied technicality and formality between 1 and 5 exhaustively, resulting in 25 paraphrased $queries$ for each $query\_base$. We also store the $expertise$ and $formality$ scores used to generate these queries.

5.4 Multi-Step Generation

To illustrate the potential for DQVis to be applied to multi-step reasoning applications, we provide 11,447 two-step question samples that mimic the realistic user interaction around a data visualization. To implement multi-step question generation with DQVis, we first extracted all single-step queries and their structured $solution$ metadata from our 1.08 million corpus, and deduplicate on the underlying $query\_base$ to isolate prototypical questions. We then designed 17 templates links. When designing the logic to link follow-up questions, we considered models of the information-seeking process summarized in Chapter 3 of Search User Interfaces [17]. One model proposed by Shneiderman et al. 1997 [40] includes four steps. 1. Query Formulation, 2. Action, 3. Review Results, 4. Refinement. The other models have slightly different formulations, but all agree on an interactive cycle where results are acquired and actions are refined based on those results. So in our case, the multi-step questions include an initial question (steps 1 and 2), a resulting visualization (step 3). Then, based on the information presented, a refined follow-up query that requests additional information (Step 4). For example, if the first query asks to view the distribution of donor weight, the second query could ask for the distribution of donor weight grouped by sex. By enforcing matching constraints on entity names, fields, and underlying solution metadata, we collected up to 50 linked question pairs $(Q1, Q2)$ that satisfy the constrains per dataset schema (e.g, HuBMAP, MW, 4DN, MoTrPAC, SenNet), ensuring balanced coverage and yielding 1,273 unique pairs for coherent two-step dialogues. Finally, we apply an LLM paraphrasing step — varying expertise and formality scores in the scale $\{1, 3, 5\}$ — to produce linguistically diverse question pairs. We reduced the number of paraphrased sentences for each multi-step chain and instead prioritized generating more questions. The paraphrasing also introduces diectic phrases for the follow-up query, e.g. group *this* by donor sex. This pipeline can be readily extended to generate multi-round reasoning dialogues of varying complexity, demonstrating DQVis's ability to link broad analytical tasks to detailed follow-ups.

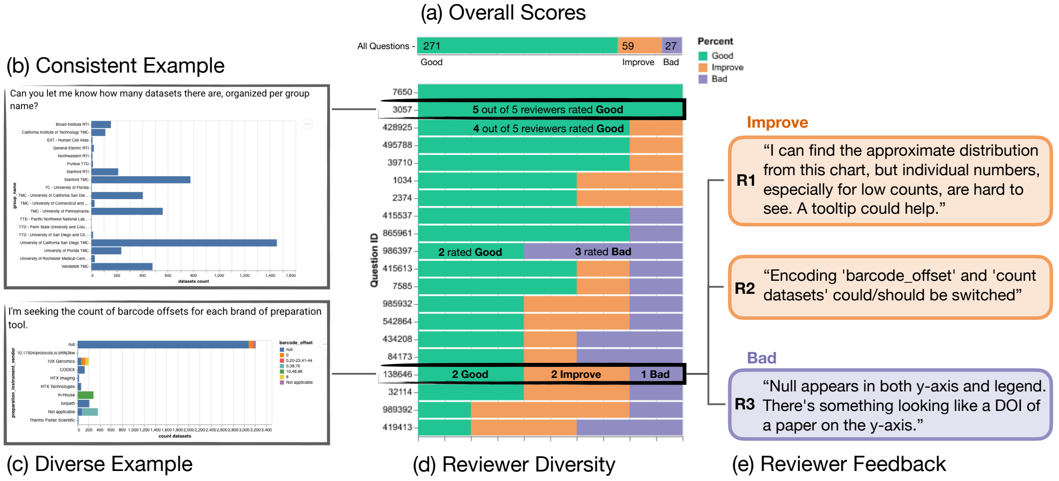
6 Data Evaluation

There are multiple ways to evaluate the quality of data-query-visualization triplets. The most basic checks include validation that the visualization specification adheres to our grammar, which is easy to confirm programmatically. However, valid specifications could still result in a malformed or empty visualization or reveal an error in the underlying visualization software. Furthermore, even for well-formed visualizations, there can be better and worse visualizations for a particular question, as well as different individual preferences. Since visualizations are intended for human interpretation, such an evaluation requires a human evaluation. For this evaluation, we recruited five individuals with advanced degrees in computer science, data visualization, biomedical informatics, and professional experience in these domains. All individuals are familiar with HuBMAP and its metadata model. To facilitate this process, we developed an interactive data review interface.

6.1 Data Review Interface and Methods

The data review interface allows reviewers to review an individual query and visualization (Figure 3e). For each data point, the reviewer can confirm that this is a reasonable question with a visualization response that satisfies the query. Alternatively, if there is a significant issue, the reviewer can select from a list of predefined options and provide free text feedback on the issue. Finally, when the visualization can still answer the question, but could be improved, the reviewer can select the middle option with a free text suggestion for improvement.

We recruited five individuals with advanced degrees in computer science, data visualization, biomedical informatics, and professional experience in these domains. All individuals are familiar with HuBMAP and its metadata model. In order to compare the similarity of reviewer responses, the first 20 data points are randomly selected once and shown to all reviewers. Then, data points are selected randomly with balance across template queries, resulting in every query template being reviewed. In total, 357 reviews were submitted for 274 unique questions.



Can you let me know how many datasets there are, organized per group name?

I'm seeking the count of barcode offsets for each brand of preparation tool.

R1 "I can find the approximate distribution from this chart, but individual numbers, especially for low counts, are hard to see. A tooltip could help."

R2 "Encoding 'barcode_offset' and 'count datasets' could/should be switched"

R3 "Null appears in both y-axis and legend. There's something looking like a DOI of a paper on the y-axis."

Figure 4: Highlights of review data. (a) Overall scores data was marked positively, with room for improvement. (b) An example of 1 of the 2 questions that was given the same score by all reviewers. (c) An example with 2 positive scores, 2 improve scores, and 1 bad score. (d) The reviewer scores for the 20 overlapping questions all reviewers saw. (e) Qualitative user feedback that points to additional features, different encodings, and data issues — all from the same question.

6.2 Results

Overall, 271 out of 357 (75.9%) questions were labeled as good; 59 (16.5%) were marked as needing improvements; and 27 (7.6%) were marked as bad. When viewing the same questions, reviewers responded diversely (See Figure 4d). Of the 20 questions shown to all reviewers, only two had the same score from everyone. The diversity of reviews points to one challenge in designing automated systems for generating visualizations — the diversity of the skills, domain knowledge, visualization knowledge, and preferences of those viewing those visualizations. Reviewers also provided free text descriptions for 73 questions that were marked as 'bad' or 'could be improved'. This feedback provides valuable insight into the process, dataset, and associated tools. The reviewed data is published with the DQVis dataset as an additional resource. Several qualitative themes are present within the reviews. Eleven comments reference the large number of null values, making the visualization difficult to read and implying it would be improved with those values filtered out. One example of this is R3 in Figure 4e, which also calls out a potential issue with the data itself — an unexpected DOI listed as an instrument vendor. Many comments point to issues with the paraphrasing. For instance, the question "Would you please furnish the distribution specifics of the number of input cells or nuclei?", which was paraphrased from "What is the distribution of number_of_input_cells_or_nuclei?" was considered "weird". More importantly, some paraphrased questions changed the meaning of terms. Three comments called out Rh factor as being distinct from blood group or blood type — the variable "rh_blood_group" had been paraphrased to "blood type" or "RH[sic] blood type." In this case, "rh_blood_group" had an empty description in the dataset schema, which may have contributed to this failure. The paraphrasing also changed the requested chart type in some instances, e.g., from "heatmap" to "bar chart" and from "pie chart" to unspecified "chart". Similarly, it altered some aggregation function words, e.g., changing "range of" to "common" values. Other comments point to enhancements for the visualization tool, such as highlighting an entire row in the tabular view or including tooltips (see R1 in Figure 4e). Alternatively, R2 suggests an alternative encodings for the same chart and similar suggestions appear for other charts, such as swapping the x and y axes on a scatterplot. In the scatterplot review, the visualization was still correct, but given the variables, swapping the axes would've conformed more to the convention of placing the independent variable on the x-axis and the dependent variable on the y-axis.

7 Discussion

7.1 Limitations

Some data fields, or combinations of data fields, result in certain chart types. For instance, placing independent variables on the x-axis or displaying population distributions split by gender as a population pyramid are conventions that visualization designers follow. The template-based approach we take loses these variable-specific conventions. Fortunately, breaking conventions does not make the visualizations inaccurate. Still, DQVis could be further enhanced by including non-template-based data.

The results indicate that the paraphraser can introduce some issues by changing the requested chart types, aggregation functions, and variable names. Hence, the review software included in the dataset generation framework is essential. Although the paraphrasing introduces some imprecision, humans are also imprecise or even incorrect when specifying queries with natural language. Still, more work can be done to characterise imprecise prompts and how best to respond to them.

The DQVis dataset is not balanced with respect to question types, visualization type, or dataset schema. This imbalance is a result of our choice to generate as many questions as possible given our set of constraints and the supplied dataset schemas. In particular, the questions generated from different dataset schemas is impacted by the number of entities and fields in the schema. Although this increases the number of data points and increases the potential of DQVis it also may require users to rebalance the dataset depending on their tasks by subsampling data points from overrepresented categories.

7.2 Ethical Considerations

New technology like LLMs introduces the possibility that generated visualizations are incorrect, even ones trained with DQVis. The generation of visualization specifications does provide some inherent guardrails compared to generating images or Python code, since it will always operate within the well-defined bounds of the visualization toolkit. Still, it is possible that data could be transformed incorrectly or presented in a misleading way. Visualization systems that use LLMs to generate visualizations should communicate with users the possibility of such outcomes.

8 Conclusion

We introduce DQVis, a dataset of 1.08 million natural language questions and visualization responses for the domain of biomedical research data repositories. This dataset can be used for fine-tuning an LLM for a biomedical natural language interface, potentially enabling critical scientific discoveries. Additionally, it could serve as a reference dataset to benchmark, compare, augment, and synthesize other work in the domain of NL2VIS. Additional domain-specific datasets could utilize the generation and review framework introduced in this work. Such domain-specific datasets have potential for specialized domains that require domain-specific visualizations like body maps and genomics visualizations. Finally, DQVis lays the foundation for multi-step reasoning datasets. By linking together the elemental data points in DQVis, we illustrate how chains of conversation can be constructed.

Acknowledgments and Disclosure of Funding

This work was supported by ARPA-H AY2AX000028. The authors wish to thank members of the HIDIVE lab for their support and feedback throughout this project, and to the expert reviewers who were essential to this project.

References

[1] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pp. 111–117, Oct. 2005. doi: 10.1109/INFVIS.2005.1532136

[2] V. S. Bursztyn, J. Hoffswell, E. Koh, and S. Guo. Representing Charts as Text for Language Models: An In-Depth Study of Question Answering for Bar Charts. In *2024 IEEE Visualization and Visual Analytics (VIS)*, pp. 266–270, Oct. 2024. doi: 10.1109/VIS55277.2024.00061

[3] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schultz. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404, May 2012. doi: 10.1158/2159-8290.CD-12-0095

[4] A. L. Charbonneau, A. Brady, K. Czajkowski, J. Aluvathingal, S. Canchi, R. Carter, K. Chard, D. J. B. Clarke, J. Crabtree, H. H. Creasy, M. D'Arcy, V. Felix, M. Giglio, A. Gingrich, R. M. Harris, T. K. Hodges, O. Ifeonu, M. Jeon, E. Kropiwnicki, M. C. W. Lim, R. L. Liming, J. Lumian, A. A. Mahurkar, M. Mandal, J. B. Munro, S. Nadendla, R. Richter, C. Romano, P. Rocca-Serra, M. Schor, R. E. Schuler, H. Tangmunarunkit, A. Waldrop, C. Williams, K. Word, S.-A. Sansone, A. Ma'ayan, R. Wagner, I. Foster, C. Kesselman, C. T. Brown, and O. White. Making Common Fund data more findable: Catalyzing a data ecosystem. *GigaScience*, 11:giac105, Nov. 2022. doi: 10.1093/gigascience/giac105

[5] N. Chen, Y. Zhang, J. Xu, K. Ren, and Y. Yang. VisEval: A Benchmark for Data Visualization in the Era of Large Language Models. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):1301–1311, Jan. 2025. doi: 10.1109/TVCG.2024.3456320

[6] W. Cui, X. Zhang, Y. Wang, H. Huang, B. Chen, L. Fang, H. Zhang, J.-G. Lou, and D. Zhang. Text-to-Viz: Automatic Generation of Infographics from Proportion-Related Natural Language Statements. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):906–916, Jan. 2020. doi: 10.1109/TVCG.2019.2934785

[7] Y. Cui, L. W. Ge, Y. Ding, L. Harrison, F. Yang, and M. Kay. Promises and Pitfalls: Using Large Language Models to Generate Visualization Items. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):1094–1104, Jan. 2025. doi: 10.1109/TVCG.2024.3456309

[8] I. de Bruijn, R. Kundra, B. Mastrogiacomo, T. N. Tran, L. Sikina, T. Mazor, X. Li, A. Ochoa, G. Zhao, B. Lai, A. Abeshouse, D. Baiceanu, E. Ciftci, U. Dogrusoz, A. Dufilie, Z. Erkoc, E. Garcia Lara, Z. Fu, B. Gross, C. Haynes, A. Heath, D. Higgins, P. Jagannathan, K. Kalletla, P. Kumari, J. Lindsay, A. Lisman, B. Leenknegt, P. Lukasse, D. Madela, R. Madupuri, P. van Nierop, O. Plantalech, J. Quach, A. C. Resnick, S. Y. A. Rodenburg, B. A. Satravada, F. Schaeffer, R. Sheridan, J. Singh, R. Sirohi, S. O. Sumer, S. van Hagen, A. Wang, M. Wilson, H. Zhang, K. Zhu, N. Rusk, S. Brown, J. A. Lavery, K. S. Panageas, J. E. Rudolph, M. L. LeNoue-Newton, J. L. Warner, X. Guo, H. Hunter-Zinck, T. V. Yu, S. Pilai, C. Nichols, S. M. Gardos, J. Philip, AACR Project GENIE BPC Core Team, AACR Project GENIE Consortium, K. L. Kehl, G. J. Riely, D. Schrag, J. Lee, M. V. Fiandalo, S. M. Sweeney, T. J. Pugh, C. Sander, E. Cerami, J. Gao, and N. Schultz. Analysis and Visualization of Longitudinal Genomic and Clinical Data from the AACR Project GENIE Biopharma Collaborative in cBioPortal. *Cancer Research*, 83(23):3861–3867, Dec. 2023. doi: 10.1158/0008-5472.CAN-23-0816

[9] A. C. A. M. de Faria, F. d. C. Bastos, J. V. N. A. da Silva, V. L. Fabris, V. d. S. Uchoa, D. G. d. A. Neto, and C. F. G. dos Santos. Visual Question Answering: A Survey on Techniques and Common Trends in Recent Literature, June 2023. doi: 10.48550/arXiv.2305.11033

[10] J. Dekker, A. S. Belmont, M. Guttman, V. O. Leshyk, J. T. Lis, S. Lomvardas, L. A. Mirny, C. C. O'Shea, P. J. Park, B. Ren, J. C. R. Politz, J. Shendure, and S. Zhong. The 4D nucleome project. *Nature*, 549(7671):219–226, Sept. 2017. doi: 10.1038/nature23884

[11] V. Dibia. LIDA: A Tool for Automatic Generation of Grammar-Agnostic Visualizations and Infographics using Large Language Models. In D. Bollegala, R. Huang, and A. Ritter, eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 113–126. Association for Computational Linguistics, Toronto, Canada, July 2023. doi: 10.18653/v1/2023.acl-demo.11

[12] V. Dibia and Ç. Demiralp. Data2Vis: Automatic Generation of Data Visualizations Using Sequence to Sequence Recurrent Neural Networks, Nov. 2018. doi: 10.48550/arXiv.1804.03126

[13] J. Gallifant, S. Chen, P. Moreira, N. Munch, M. Gao, J. Pond, L. A. Celi, H. Aerts, T. Hartvigsen, and D. Bitterman. Language Models are Surprisingly Fragile to Drug Names in Biomedical Benchmarks, June 2024. doi: 10.48550/arXiv.2406.12066

[14] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, 6(269):pl1, Apr. 2013. doi: 10.1126/scisignal.2004088

[15] L. Gao, J. Lu, Z. Shao, Z. Lin, S. Yue, C. Leong, Y. Sun, R. J. Zauner, Z. Wei, and S. Chen. Fine-Tuned Large Language Model for Visualization System: A Study on Self-Regulated Learning in Education. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):514–524, Jan. 2025. doi: 10.1109/TVCG.2024.3456145

[16] S. Gao, R. Zhu, Z. Kong, A. Noori, X. Su, C. Ginder, T. Tsiligkaridis, and M. Zitnik. Txagent: An ai agent for therapeutic reasoning across a universe of tools, 2025.

[17] M. Hearst. *Search User Interfaces*. Cambridge University Press, Sept. 2009.

[18] E. Hoque, P. Kavehzadeh, and A. Masry. Chart Question Answering: State of the Art and Future Directions. *Computer Graphics Forum*, 41(3):555–572, June 2022. doi: 10.1111/cgf.14573

[19] E. Hoque, V. Setlur, M. Tory, and I. Dykeman. Applying Pragmatics Principles for Interaction with Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):309–318, Jan. 2018. doi: 10.1109/TVCG.2017.2744684

[20] K. Hu, S. N. S. Gaikwad, M. Hulsebos, M. A. Bakker, E. Zgraggen, C. Hidalgo, T. Kraska, G. Li, A. Satyanarayan, and Ç. Demiralp. VizNet: Towards A Large-Scale Visualization Learning and Benchmarking Repository. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12. ACM, Glasgow Scotland Uk, May 2019. doi: 10.1145/3290605.3300892

[21] J. Huang, Y. Xi, J. Hu, and J. Tao. FlowNL: Asking the Flow Data in Natural Languages. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1200–1210, Jan. 2023. doi: 10.1109/TVCG.2022.3209453

[22] K. Kafle, B. Price, S. Cohen, and C. Kanan. DVQA: Understanding Data Visualizations via Question Answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5648–5656, June 2018. doi: 10.1109/CVPR.2018.00592

[23] S. E. Kahou, V. Michalski, A. Atkinson, A. Kadar, A. Trischler, and Y. Bengio. FigureQA: An Annotated Figure Dataset for Visual Reasoning, Feb. 2018. doi: 10.48550/arXiv.1710.07300

[24] J. Kim, S. Lee, H. Jeon, K.-J. Lee, H.-J. Bae, B. Kim, and J. Seo. PhenoFlow: A Human-LLM Driven Visual Analytics System for Exploring Large and Complex Stroke Datasets. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):470–480, Jan. 2025. doi: 10.1109/TVCG.2024.3456215

[25] H.-K. Ko, H. Jeon, G. Park, D. H. Kim, N. W. Kim, J. Kim, and J. Seo. Natural Language Dataset Generation Framework for Visualizations Powered by Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–22. Association for Computing Machinery, New York, NY, USA, May 2024. doi: 10.1145/3613904.3642943

[26] C. Lee, T. Lin, H. Pfister, and C. Zhu-Tian. Sportify: Question Answering with Embedded Visualizations and Personified Narratives for Sports Video. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):12–22, Jan. 2025. doi: 10.1109/TVCG.2024.3456332

[27] P. J. Lee, C. C. Benz, P. Blood, K. Börner, J. Campisi, F. Chen, H. Daldrup-Link, P. De Jager, L. Ding, F. E. Duncan, O. Eickelberg, R. Fan, T. Finkel, D. Furman, V. Garovic, N. Gehlenborg, C. Glass, I. Heckenbach, Z.-B. Joseph, P. Katiyar, S.-J. Kim, M. Königshoff, G. A. Kuchel, H. Lee, J. H. Lee, J. Ma, Q. Ma, S. Melov, K. Metis, A. L. Mora, N. Musi, N. Neretti, J. F. Passos, I. Rahman, J. C. Rivera-Mulia, P. Robson, M. Rojas, A. L. Roy, M. Scheibye-Knudsen, B. Schilling, P. Shi, J. C. Silverstein, V. Suryadevara, J. Xie, J. Wang, A. I. Wong, L. J. Niedernhofer, S. S. Wang, H. Anvari, J. Balough, C. Benz, J. Bons, B. Brenerman, W. Evans, A. Gerencser, H. Gregory, M. Hansen, J. Justice, P. Kapahi, N. Murad, A. O'Broin, M. E. Pavone, M. Powell, G. Scott, E. Shanes, M. Shankaran, E. Verdin, D. Winer, F. Wu, A. Adams, P. D. Blood, A. Bueckle, I. Cao-Berg, H. Chen, M. Davis, S. Filus, Y. Hao, A. Hartman, E. Hasanaj, J. Helfer, B. Herr, Z. B. Joseph, G. Molla, G. Mou, J. Puerto, E. M. Quardokus, A. J. Ropelewski, M. Ruffalo, R. Satija, M. Schwenk, R. Scibek, W. Shirey, M. Sibilla, J. Welling, Z. Yuan, R. Bonneau, A. Christiano, B. Izar, V. Menon, D. M. Owens, H. Phatnani, C. Smith, Y. Suh, A. F. Teich, V. Bekker, C. Chan, E. Coutavas, M. G. Hartwig, Z. Ji, A. B. Nixon, Z. Dou, J. Rajagopal, N. Slavov, D. Holmes, D. Jurk, J. L. Kirkland, A. Lagnado, T. Tchkonia, K. Abraham, A. Dibattista, Y.-W. Fridell, T. K. Howcroft, C. Jhappan, V. P. Montes, M. Prabhudas, H. Resat, V. Taylor, M. Kumar, V. Suryadevara, F. Cigarroa, R. Cohn, T. M. Cortes, E. Courtois, J. Chuang, M. Davé, S. Domanskyi, E. A. L. Enninga, G. N. Eryilmaz, S. E. Espinoza, J. Gelfond, J. Kirkland, G. A. Kuchel, C.-L. Kuo, J. S. Lehman, C. Aguayo-Mazzucato, A. Meves, M. Rani, S. Sanders, A. Thibodeau, S. G. Tullius, D. Ucar, B. White, Q. Wu, M. Xu, S. Yamaguchi, N. Assarzadegan, C.-S. Cho, I. Hwang, Y. Hwang, J. Xi, O. A. Adeyi, C. F. Aliferis, A. Bartolomucci, X. Dong, M. J. DuFresne-To, S. Ikramuddin, S. G. Johnson, A. C. Nelson, L. J. Niedernhofer, X. S. Revelo, C. Trevilla-Garcia, J. M. Sedivy, E. L. Thompson, P. D. Robbins, J. Wang, K. M. Aird, J. K. Alder, D. Beaulieu, M. Bueno, J. Calyeca, J. A. Chamucero-Millaris, S. Y. Chan, D. Chung, A. Corbett, V. Gorbunova, K. M. Gowdy, A. Gurkar, J. C. Horowitz, Q. Hu, G. Kaur, T. O. Khaliullin, R. Lafyatis, S. Lanna, D. Li, A. Ma, A. Morris, T. M. Muthumalage, V. Peters, G. S. Pryhuber, B. F. Reader, L. Rosas, J. C. Sembrat, S. Shaikh, H. Shi, S. D. Stacey, C. S. Croix, C. Wang, Q. Wang, A. Watts, L. Gu, Y. Lin, P. S. Rabinovitch, M. T. Sweetwyne, M. N. Artyomov, S. J. Ballentine, M. G. Chheda, S. R. Davies, J. F. DiPersio, R. C. Fields, J. A. J. Fitzpatrick, R. S. Fulton, S.-i. Imai, S. Jain, T. Ju, V. M. Kushnir, D. C. Link, M. Ben Major, S. T. Oh, D. Rapp, M. P. Rettig, S. A. Stewart, D. J. Veis, K. R. Vij, M. C. Wendl, M. A. Wyczalkowski, J. E. Craft, A. Enninful, N. Farzad, P. Gershkovich, S. Halene, Y. Kluger, J. VanOudenhove, M. Xu, J. Yang, M. Yang, SenNet Consortium, Writing Group, Brown University TDA, Buck Institute for Research on Aging TMC/TDA, Consortium Organization and Data Coordinating Center (CODCC), Columbia TMC, Duke University TMC, Massachusetts General Hospital TDA, Mayo Clinic TDA, National Institute of Health (NIH), Stanford TDA, University of Connecticut TMC, University of Michigan TDA, University of Minnesota TMC, University of Pittsburgh TMC, University of Washington TDA, Washington University TMC, and Yale TMC. NIH SenNet Consortium to map senescent cells throughout the human lifespan to understand physiological health. *Nature Aging*, 2(12):1090–1100, Dec. 2022. doi: 10.1038/s43587-022-00326-5

[28] S. Liu, H. Miao, Z. Li, M. Olson, V. Pascucci, and P.-T. Bremer. AVA: Towards Autonomous Visualization Agents through Visual Perception-Driven Decision-Making. *Computer Graphics Forum*, 43(3):e15093, 2024. doi: 10.1111/cgf.15093

[29] Z. Liu, X. Xie, M. He, W. Zhao, Y. Wu, L. Cheng, H. Zhang, and Y. Wu. Smartboard: Visual Exploration of Team Tactics with LLM Agent. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):23–33, Jan. 2025. doi: 10.1109/TVCG.2024.3456200

[30] Y. Luo, N. Tang, G. Li, C. Chai, W. Li, and X. Qin. Synthesizing Natural Language to Visualization (NL2VIS) Benchmarks from NL2SQL Benchmarks. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, pp. 1235–1247. Association for Computing Machinery, New York, NY, USA, June 2021. doi: 10.1145/3448016.3457261

[31] Y. Luo, N. Tang, G. Li, J. Tang, C. Chai, and X. Qin. Natural Language to Visualization by Neural Machine Translation. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):217–226, Jan. 2022. doi: 10.1109/TVCG.2021.3114848

[32] S. L'Yi, A. van den Brandt, E. Adams, H. N. Nguyen, and N. Gehlenborg. Learnable and Expressive Visualization Authoring through Blended Interfaces. *IEEE Transactions on Visualization and Computer Graphics*, 2024. doi: 10.1109/TVCG.2024.3456598

[33] S. L'Yi, Q. Wang, F. Lekschas, and N. Gehlenborg. Gosling: A Grammar-based Toolkit for Scalable and Interactive Genomics Data Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):140–150, Jan. 2022. doi: 10.1109/TVCG.2021.3114876

[34] J. Mackinlay, P. Hanrahan, and C. Stolte. Show Me: Automatic Presentation for Visual Analysis. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '07)*, 13(6):1137–1144, 2007. doi: 10.1109/TVCG.2007.70594

[35] A. Narechania, A. Srinivasan, and J. Stasko. NL4DV: A Toolkit for Generating Analytic Specifications for Data Visualization from Natural Language Queries. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):369–379, Feb. 2021. doi: 10.1109/TVCG.2020.3030378

[36] A. Pandey, S. L'Yi, Q. Wang, M. A. Borkin, and N. Gehlenborg. GenoREC: A Recommendation System for Interactive Genomics Data Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):570–580, Jan. 2023. doi: 10.1109/TVCG.2022.3209407

[37] A. L. Roy, R. S. Conroy, V. G. Taylor, J. Mietz, I. M. Fingerman, M. J. Pazin, P. Smith, C. M. Hutter, D. S. Singer, and E. L. Wilder. Elucidating the structure and function of the nucleus—The NIH Common Fund 4D Nucleome program. *Molecular Cell*, 83(3):335–342, Feb. 2023. doi: 10.1016/j.molcel.2022.12.025

[38] J. A. Sanford, C. D. Nogiec, M. E. Lindholm, J. N. Adkins, D. Amar, S. Dasari, J. K. Drugan, F. M. Fernández, S. Radom-Aizik, S. Schenk, M. P. Snyder, R. P. Tracy, P. Vanderboom, S. Trappe, M. J. Walsh, J. N. Adkins, D. Amar, S. Dasari, J. K. Drugan, C. R. Evans, F. M. Fernandez, Y. Li, M. E. Lindholm, C. D. Nogiec, S. Radom-Aizik, J. A. Sanford, S. Schenk, M. P. Snyder, L. Tomlinson, R. P. Tracy, S. Trappe, P. Vanderboom, M. J. Walsh, D. Lee Alekel, I. Bekirov, A. T. Boyce, J. Boyington, J. L. Fleg, L. J. O. Joseph, M. R. Laughlin, P. Maruvada, S. A. Morris, J. A. McGowan, C. Nierras, V. Pai, C. Peterson, E. Ramos, M. C. Roary, J. P. Williams, A. Xia, E. Cornell, J. Rooney, M. E. Miller, W. T. Ambrosius, S. Rushing, C. L. Stowe, W. Jack Rejeski, B. J. Nicklas, M. Pahor, C.-j. Lu, T. Trappe, T. Chambers, U. Raue, B. Lester, B. C. Bergman, D. H. Bessesen, C. M. Jankowski, W. M. Kohrt, E. L. Melanson, K. L. Moreau, I. E. Schauer, R. S. Schwartz, W. E. Kraus, C. A. Slentz, K. M. Huffman, J. L. Johnson, L. H. Willis, L. Kelly, J. A. Houmard, G. Dubis, N. Broskey, B. H. Goodpaster, L. M. Sparks, P. M. Coen, D. M. Cooper, F. Haddad, T. Rankinen, E. Ravussin, N. Johannsen, M. Harris, J. M. Jakicic, A. B. Newman, D. D. Forman, E. Kershaw, R. J. Rogers, B. C. Nindl, L. C. Page, M. Stefanovic-Racic, S. L. Barr, B. B. Rasmussen, T. Moro, D. Paddon-Jones, E. Volpi, H. Spratt, N. Musi, S. Espinoza, D. Patel, M. Serra, J. Gelfond, A. Burns, M. M. Bamman, T. W. Buford, G. R. Cutter, S. C. Bodine, K. Esser, R. P. Farrar, L. J. Goodyear, M. F. Hirshman, B. G. Albertson, W.-J. Qian, P. Piehowski, M. A. Gritsenko, M. E. Monore, V. A. Petyuk, J. E. McDermott, J. N. Hansen, C. Hutchison, S. Moore, D. A. Gaul, C. B. Clish, J. Avila-Pacheco, C. Dennis, M. Kellis, S. Carr, P. M. Jean-Beltran, H. Keshishian, D. R. Mani, K. Clauser, K. Krug, C. Mundorff, C. Pearce, A. A. Ivanova, E. A. Ortlund, K. Maner-Smith, K. Uppal, T. Zhang, S. C. Sealfon, E. Zaslavsky, V. Nair, S. Li, N. Jain, Y. Ge, Y. Sun, G. Nudelman, F. Ruf-zamojski, G. Smith, N. Pincas, A. Rubenstein, M. Anne Amper, N. Seenarine, T. Lappalainen, I. R. Lanza, K. Sreekumaran Nair, K. Klaus, S. B. Montgomery, K. S. Smith, N. R. Gay, B. Zhao, C.-J. Hung, N. Zebarjadi, B. Balliu, L. Fresard, C. F. Burant, J. Z. Li, M. Kachman, T. Soni, A. B. Raskind, R. Gerszten, J. Robbins, O. Ilkayeva, M. J. Muehlbauer, C. B. Newgard, E. A. Ashley, M. T. Wheeler, D. Jimenez-Morales, A. Raja, K. P. Dalton, J. Zhen, Y. Suk Kim, J. W. Christle, S. Marwaha, E. T. Chin, S. G. Hershman, T. Hastie, R. Tibshirani, and M. A. Rivas. Molecular Transducers of Physical Activity Consortium (MoTrPAC): Mapping the Dynamic Responses to Exercise. *Cell*, 181(7):1464–1474, June 2020. doi: 10.1016/j.cell.2020.06.004

[39] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vega-Lite: A Grammar of Interactive Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):341–350, Jan. 2017. doi: 10.1109/TVCG.2016.2599030

[40] B. Schneiderman, D. Byrd, and B. W. Croft. Clarifying Search: A User-Interface Framework for Text Searches. *DL Magazine*, 1997. doi: 10.1045/january97-shneiderman

[41] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang. Eviza: A Natural Language Interface for Visual Analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, pp. 365–377. Association for Computing Machinery, New York, NY, USA, Oct. 2016. doi: 10.1145/2984511.2984588

[42] L. Shen, E. Shen, Y. Luo, X. Yang, X. Hu, X. Zhang, Z. Tai, and J. Wang. Towards Natural Language Interfaces for Data Visualization: A Survey. *IEEE Transactions on Visualization and Computer Graphics*, 29(6):3121–3144, June 2023. doi: 10.1109/TVCG.2022.3148007

[43] C. Shi, W. Cui, C. Liu, C. Zheng, H. Zhang, Q. Luo, and X. Ma. NL2Color: Refining Color Palettes for Charts with Natural Language. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):814–824, Jan. 2024. doi: 10.1109/TVCG.2023.3326522

[44] M. P. Snyder, S. Lin, A. Posgai, M. Atkinson, A. Regev, J. Rood, O. Rozenblatt-Rosen, L. Gaffney, A. Hupalowska, R. Satija, N. Gehlenborg, J. Shendure, J. Laskin, P. Harbury, N. A. Nystrom, J. C. Silverstein, Z. Bar-Joseph, K. Zhang, K. Börner, Y. Lin, R. Conroy, D. Procaccini, A. L. Roy, A. Pillai, M. Brown, Z. S. Galis, L. Cai, J. Shendure, C. Trapnell, S. Lin, D. Jackson, M. P. Snyder, G. Nolan, W. J. Greenleaf, Y. Lin, S. Plevritis, S. Ahadi, S. A. Nevins, H. Lee, C. M. Schuerch, S. Black, V. G. Venkataraaman, E. Esplin, A. Horning, A. Bahmani, K. Zhang, X. Sun, S. Jain, J. Hagood, G. Pryhuber, P. Kharchenko, M. Atkinson, B. Bodenmiller, T. Brusko, M. Clare-Salzler, H. Nick, K. Otto, A. Posgai, C. Wasserfall, M. Jorgensen, M. Brusko, S. Maffioletti, R. M. Caprioli, J. M. Spraggins, D. Gutierrez, N. H. Patterson, E. K. Neumann, R. Harris, M. deCaestecker, A. B. Fogo, R. van de Plas, K. Lau, L. Cai, G.-C. Yuan, Q. Zhu, R. Dries, P. Yin, S. K. Saka, J. Y. Kishi, Y. Wang, I. Goldaracena, J. Laskin, D. Ye, K. E. Burnum-Johnson, P. D. Piehowski, C. Ansong, Y. Zhu, P. Harbury, T. Desai, J. Mulye, P. Chou, M. Nagendran, Z. Bar-Joseph, S. A. Teichmann, B. Paten, R. F. Murphy, J. Ma, V. Y. Kiselev, C. Kingsford, A. Ricarte, M. Keays, S. A. Akoju, M. Ruffalo, N. Gehlenborg, P. Kharchenko, M. Vella, C. McCallum, K. Börner, L. E. Cross, S. H. Friedman, R. Heiland, B. Herr, P. Macklin, E. M. Quardokus, L. Record, J. P. Sluka, G. M. Weber, N. A. Nystrom, J. C. Silverstein, P. D. Blood, A. J. Ropelewski, W. E. Shirey, R. M. Scibek, P. Mabee, W. C. Lenhardt, K. Robasky, S. Michailidis, R. Satija, J. Marioni, A. Regev, A. Butler, T. Stuart, E. Fisher, S. Ghazanfar, J. Rood, L. Gaffney, G. Eraslan, T. Biancalani, E. D. Vaishnav, R. Conroy, D. Procaccini, A. Roy, A. Pillai, M. Brown, Z. Galis, P. Srinivas, A. Pawlyk, S. Sechi, E. Wilder, J. Anderson, HuBMAP Consortium, Writing Group, Caltech-UW TMC, Stanford-WashU TMC, UCSD TMC, University of Florida TMC, Vanderbilt University TMC, California Institute of Technology TTD, Harvard TTD, Purdue TTD, Stanford TTD, T. C. V. HuBMAP Integration, and Engagement (HIVE) Collaboratory: Carnegie Mellon, T. C. Harvard Medical School, M. C. Indiana University Bloomington, I. a. E. C. Pittsburgh Supercomputing Center and University of Pittsburgh, C. C. University of South Dakota, M. C. New York Genome Center, and NIH HuBMAP Working Group. The human body at cellular resolution: The NIH Human Biomolecular Atlas Program. *Nature*, 574(7777):187–192, Oct. 2019. doi: 10.1038/s41586-019-1629-x

[45] S. Song, J. Chen, C. Li, and C. Wang. GVQA: Learning to Answer Questions about Graphs with Visualizations via Knowledge Base. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 1–16. Association for Computing Machinery, New York, NY, USA, Apr. 2023. doi: 10.1145/3544548.3581067

[46] Y. Song, X. Zhao, and R. C.-W. Wong. Marrying Dialogue Systems with Data Visualization: Interactive Data Visualization Generation from Natural Language Conversations. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 2733–2744. Association for Computing Machinery, New York, NY, USA, Aug. 2024. doi: 10.1145/3637528.3671935

[47] A. Srinivasan, N. Nyapathy, B. Lee, S. M. Drucker, and J. Stasko. Collecting and Characterizing Natural Language Utterances for Specifying Data Visualizations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–10. ACM, Yokohama Japan, May 2021. doi: 10.1145/3411764.3445400

[48] M. Sud, E. Fahy, D. Cotter, K. Azam, I. Vadivelu, C. Burant, A. Edison, O. Fiehn, R. Higashi, K. S. Nair, S. Sumner, and S. Subramaniam. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Research*, 44(D1):D463–D470, Jan. 2016. doi: 10.1093/nar/gkv1042

[49] P. Vaithilingam, E. L. Glassman, J. P. Inala, and C. Wang. DynaVis: Dynamically Synthesized UI Widgets for Visualization Editing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–17. Association for Computing Machinery, New York, NY, USA, May 2024. doi: 10.1145/3613904.3642639

[50] C. Wang, J. Thompson, and B. Lee. Data Formulator: AI-Powered Concept-Driven Visualization Authoring. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1128–1138, Jan. 2024. doi: 10.1109/TVCG.2023.3326585

[51] H. W. Wang, M. Gordon, L. Battle, and J. Heer. DracoGPT: Extracting Visualization Design Preferences from Large Language Models. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):710–720, Jan. 2025. doi: 10.1109/TVCG.2024.3456350

[52] Y. Wang, Z. Hou, L. Shen, T. Wu, J. Wang, H. Huang, H. Zhang, and D. Zhang. Towards Natural Language-Based Visualization Authoring. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–11, 2022. doi: 10.1109/TVCG.2022.3209357

[53] L. Wilkinson. *The Grammar of Graphics*. Springer, 2nd ed., 2005.

[54] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):649–658, Jan. 2016. doi: 10.1109/TVCG.2015.2467191

[55] A. Wu, Y. Wang, X. Shu, D. Moritz, W. Cui, H. Zhang, D. Zhang, and H. Qu. AI4VIS: Survey on Artificial Intelligence Approaches for Data Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):5049–5070, Dec. 2022. doi: 10.1109/TVCG.2021.3099002

[56] Y. Yan, Y. Hou, Y. Xiao, R. Zhang, and Q. Wang. KNOWNET: Guided Health Information Seeking from LLMs via Knowledge Graph Integration. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):547–557, Jan. 2025. doi: 10.1109/TVCG.2024.3456364

[57] B. Yu and C. T. Silva. FlowSense: A Natural Language Interface for Visual Data Exploration within a Dataflow System. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1–11, Jan. 2020. doi: 10.1109/TVCG.2019.2934668

[58] T. Yu, R. Zhang, K. Yang, M. Yasunaga, D. Wang, Z. Li, J. Ma, I. Li, Q. Yao, S. Roman, Z. Zhang, and D. Radev. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3911–3921. Association for Computational Linguistics, Brussels, Belgium, Oct. 2018. doi: 10.18653/v1/D18-1425

[59] C. Zhu-Tian, Q. Yang, X. Xie, J. Beyer, H. Xia, Y. Wu, and H. Pfister. Sporthesia: Augmenting Sports Videos Using Natural Language. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):918–928, Jan. 2023. doi: 10.1109/TVCG.2022.3209497

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We published a dataset on Hugging Face and data generation code on GitHub.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 7.1.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: No theoretical results in paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Data is published on Hugging Face, and the code used to generate the data is published on GitHub.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Same as above.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: This paper did not train any models.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: Paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research adheres to the NeurIPS Code of Ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Section 7.2 and 8.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: All data is already published and easily available or synthesized by the authors.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Related data sources are cited in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Data is published on Hugging Face along with documentation in the readme.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No crowdsourcing or human subjects research was performed.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No crowdsourcing or human subjects research was performed.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: See Section 5.3.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Technical Appendices and Supplementary Material

A.1 Code and Resources

The following are the primary DQVis resources.

- Data: `https://huggingface.co/datasets/HIDIVE/DQVis`
- Synthesis Code: `https://github.com/hms-dbmi/DQVis-Generation`
- Review Code: `https://github.com/hms-dbmi/DQVis-review`

A highly-relevant resource that is not contributed by this paper, but used by it is the visualization grammar that DQVis uses. We include the versions used for the creation and review of DQVis.

The **udi-grammar-py** Python package was used for the generation of template specifications.

- Version: 0.2.6
- GitHub: `https://github.com/hms-dbmi/udi-grammar-py`
- PyPI: `https://pypi.org/project/udi-grammar-py/`

The **udi-toolkit** JavaScript package was used in the review software.

- Version: 0.0.24
- GitHub: `https://github.com/hms-dbmi/udi-grammar`
- NPM: `https://www.npmjs.com/package/udi-toolkit`

A.2 Paraphrasing Prompts

The following is the complete prompt template used for paraphrasing questions.

LLM Prompt Template

You are a paraphrasing assistant. Your task is to rewrite a given sentence with various styles of language usage. The sentence will either be a question about data, or request to construct a data visualization.

The input sentence will include entity names and fields names from the data. The dataset schema will also be provided to you to enable better paraphrasing of the field and entity names. More technical language may use the exact field names, while more colloquial language may use more general terms, synonyms, and will likely not use the exact field names. e.g. "What is the value of the age_value field?" vs "How old is the person?".
Dataset schema: `{dataset_schema}`

Score-A of 1 indicates a higher tendency to use Colloquial language and a Score-A of 5 indicates a higher tendency to use Standard language.
Score-B of 1 indicates a higher tendency to use Non-technical language and a Score-B of 5 indicates a higher tendency to use Technical language.
Rewrite the following sentence as if it were spoken by a person with a given score for language usage.

Sentence: `{sentence}`

##
Score-A 1, Score-B 1: Score-A 1, Score-B 2: Score-A 1, Score-B 3: Score-A 1, Score-B 4:
Score-A 1, Score-B 5: Score-A 2, Score-B 1: Score-A 2, Score-B 2: Score-A 2, Score-B 3:
Score-A 2, Score-B 4: Score-A 2, Score-B 5: Score-A 3, Score-B 1: Score-A 3, Score-B 2:
Score-A 3, Score-B 3: Score-A 3, Score-B 4: Score-A 3, Score-B 5: Score-A 4, Score-B 1:
Score-A 4, Score-B 2: Score-A 4, Score-B 3: Score-A 4, Score-B 4: Score-A 4, Score-B 5:

Score-A 5, Score-B 1: Score-A 5, Score-B 2: Score-A 5, Score-B 3: Score-A 5, Score-B 4: Score-A 5, Score-B 5:

`Placeholder`	**Example Value / Description**
`{dataset_schema}`	A JSON schema showing dataset fields and entity types.
`{sentence}`	The input text to paraphrase, e.g., `"Is there a correlation between age_value and weight_value?"`

Table 2: Placeholders and example input values used in the LLM paraphrasing prompt template.

In addition to the prompt we use structured outputs to require a list of paraphrased sentences with a formality and expertise score. The description of the formality and expertise score are included in the structured output definition.

Formality. Colloquial (Score=1) language is informal and used in everyday conversation, while standard language (Score=5) follows established rules and conventions and is used in more formal situations.

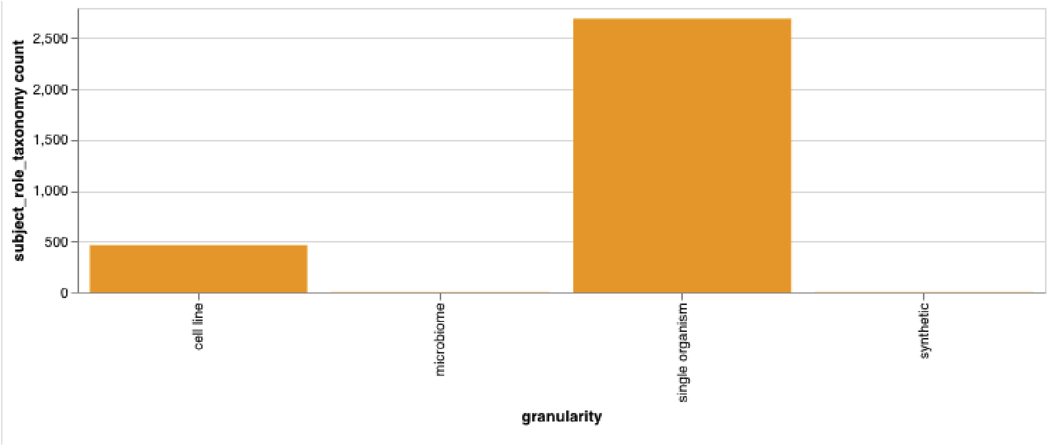
Expertise. Technical language (Score=5) is often used by experts in a particular field and includes specialized terminology and jargon. Non-technical language (Score=1), on the other hand, is more accessible to a general audience and avoids the use of complex terms.

A.3 Reviews

This section shows all of the data that is summarized in Figure 4. These are not all the results collected, but just the 20 data points that every reviewer saw. All review data is available on the Hugging Face repository. Each page in this section includes the question ID, the query, and the visualization shown to the reviewers, along with all five reviewer responses. The review interface allowed users to select a score of good, improve, or bad. If the selection was not good, then the user could select predefined issue categories and leave free-text comments.

Question ID: 7650

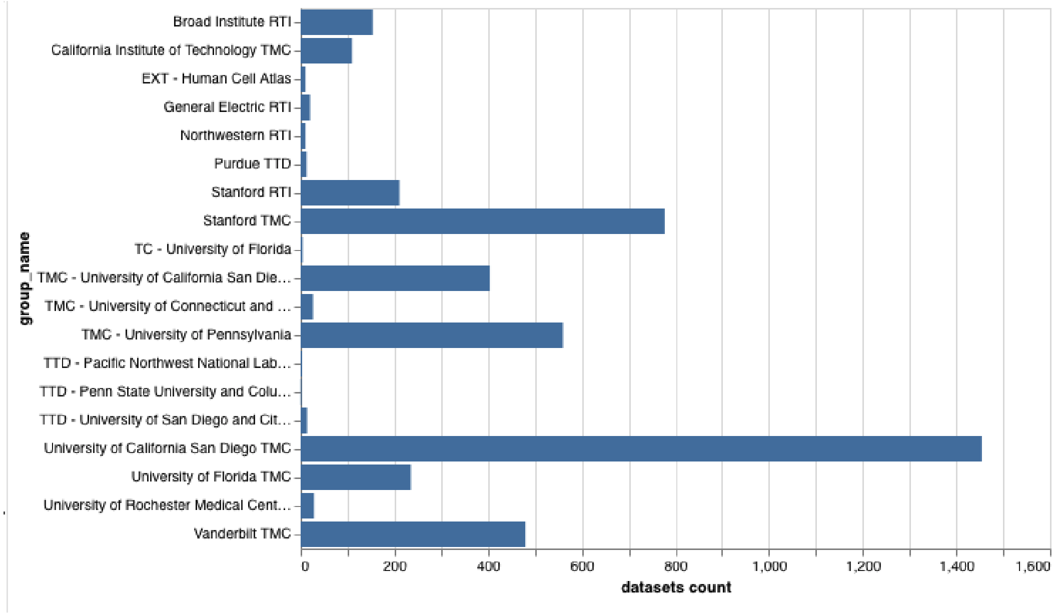
Query: Can you tell me the number of subject categories we've got, broken down by the level of detail?



Reviewer	Score	Issue Categories	Comments
R1	Good		
R2	Good		
R3	Good		
R4	Good		
R5	Good		

Question ID: 3057

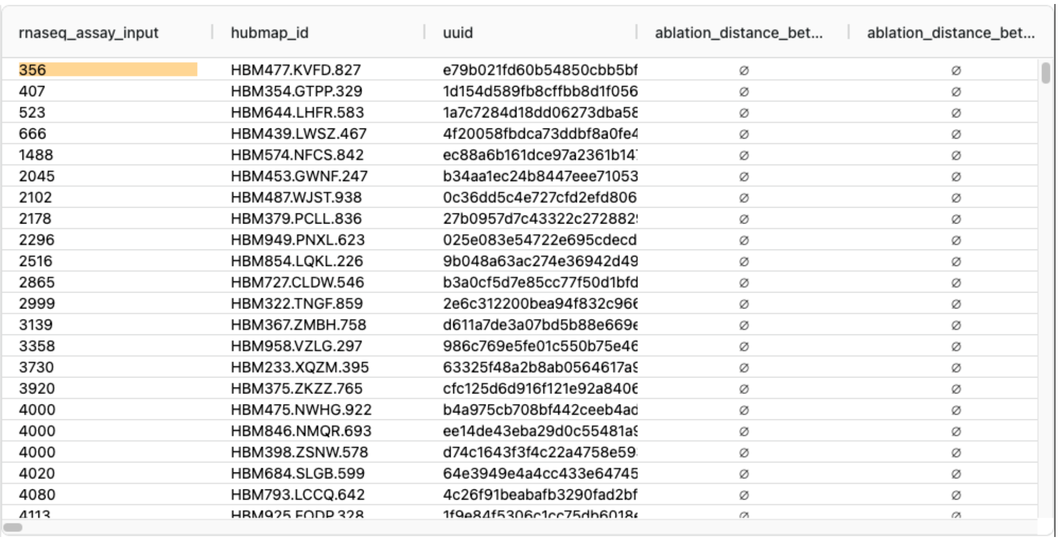
Query: Can you let me know how many datasets there are, organized per group name?



Reviewer	Score	Issue Categories	Comments
R1	Good		
R2	Good		
R3	Good		
R4	Good		
R5	Good		

Question ID: 428925

Query: Could you highlight the record with the lowest RNA-seq input amount listed?

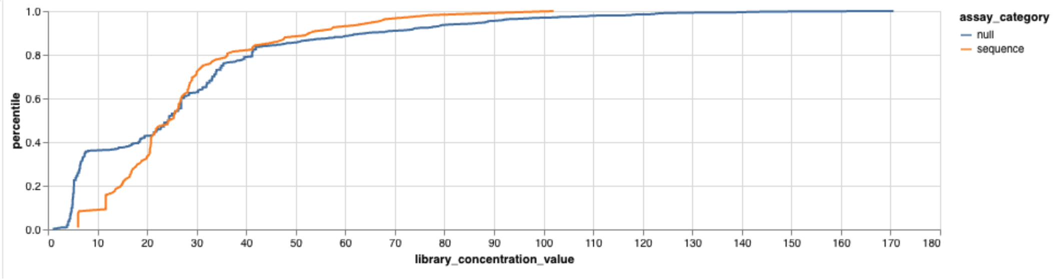


rnaseq_assay_input	hubmap_id	uuid	ablation_distance_bet...	ablation_distance_bet...
356	HBM477.KVFD.827	e79b021fd60b54850cbb5bf	∅	∅
407	HBM354.GTPP.329	1d154d589fb8cffbb8d1f056	∅	∅
523	HBM644.LHFR.583	1a7c7284d18dd06273dba58	∅	∅
666	HBM439.LWSZ.467	4f20058fbdca73ddbf8a0fe4	∅	∅
1488	HBM574.NFCS.842	ec88a6b161dce97a2361b14	∅	∅
2045	HBM453.GWNF.247	b34aa1ec24b8447eee71053	∅	∅
2102	HBM487.WJST.938	0c36dd5c4e727cfd2efd806	∅	∅
2178	HBM379.PCLL.836	27b0957d7c43322c272882	∅	∅
2296	HBM949.PNXL.623	025e083e54722e695cdecd	∅	∅
2516	HBM854.LQKL.226	9b048a63ac274e36942d49	∅	∅
2865	HBM727.CLDW.546	b3a0cf5d7e85cc77f50d1bfd	∅	∅
2999	HBM322.TNGF.859	2e6c312200bea94f832c96	∅	∅
3139	HBM367.ZMBH.758	d611a7de3a07bd5b88e669	∅	∅
3358	HBM958.VZLG.297	986c769e5fe01c550b75e4	∅	∅
3730	HBM233.XQZM.395	63325f48a2b8ab0564617a	∅	∅
3920	HBM375.ZKZZ.765	cfc125d6d916f121e92a840	∅	∅
4000	HBM475.NWHG.922	b4a975cb708bf442ceeb4ad	∅	∅
4000	HBM846.NMQR.693	ee14de43eba29d0c55481a	∅	∅
4000	HBM398.ZSNW.578	d74c1643f3f4c22a4758e59	∅	∅
4020	HBM684.SLGB.599	64e3949e4a4cc433e64745	∅	∅
4080	HBM793.LCCQ.642	4c26f91beabafb3290fad2bf	∅	∅
4113	HBM925.FODP.328	1f9e84f5306c1cc75db6018	∅	∅

Reviewer	Score	Issue Categories	Comments
R1	Good		
R2	Good		
R3	Good		
R4	Good		
R5	Improve		maybe highlight the whole row, but this will do.

Question ID: 495788

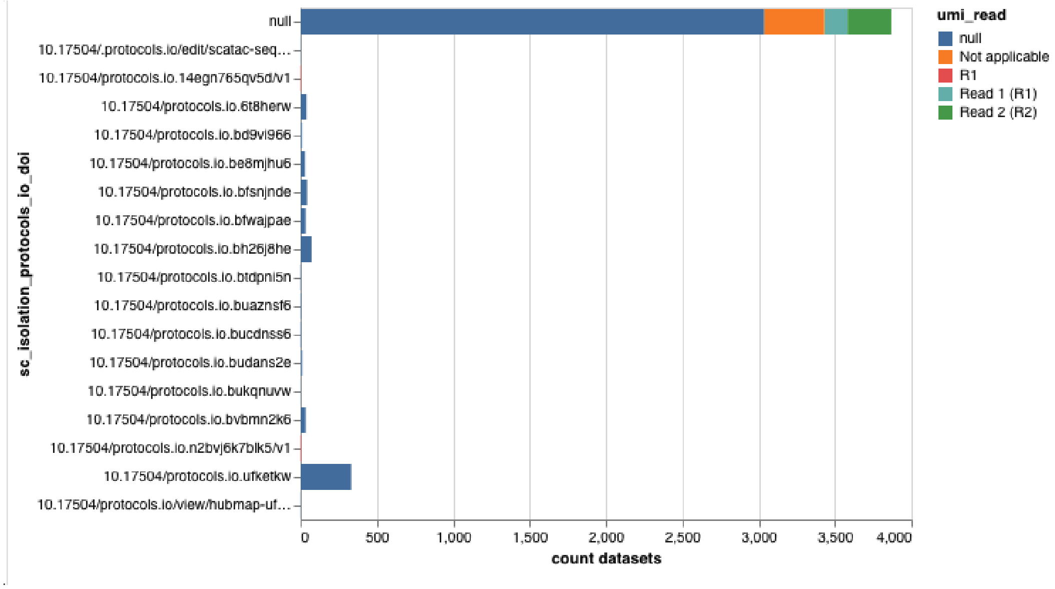
Query: What’s the complete spread of library concentration values for each assay category?



Reviewer	Score	Issue Categories	Comments
R1	Improve		library_concentration_value is missing unit
R2	Good		
R3	Good		
R4	Good		
R5	Good		

Question ID: 39710

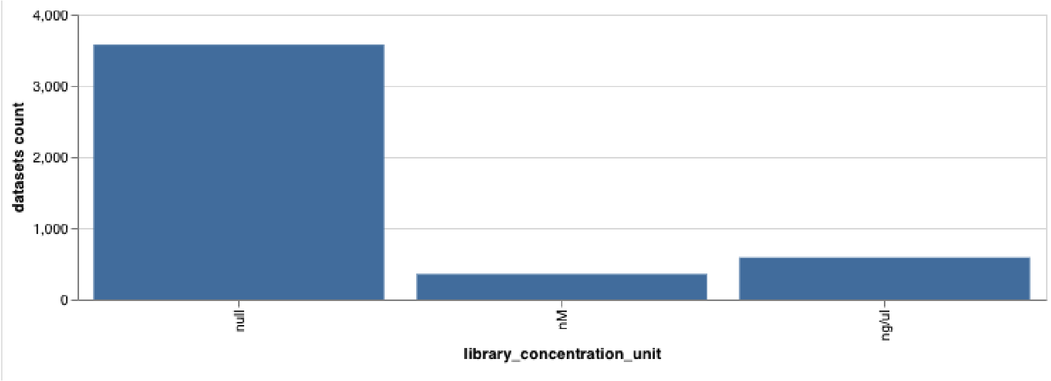
Query: Could you give me the number of datasets grouped by the fields ’umi_read’ and ’sc_isolation_protocols_io_doi’?



Reviewer	Score	Issue Categories	Comments
R1	Good		
R2	Improve		null value again messes readability of the visualization
R3	Good		
R4	Good		
R5	Good		

Question ID: 1034

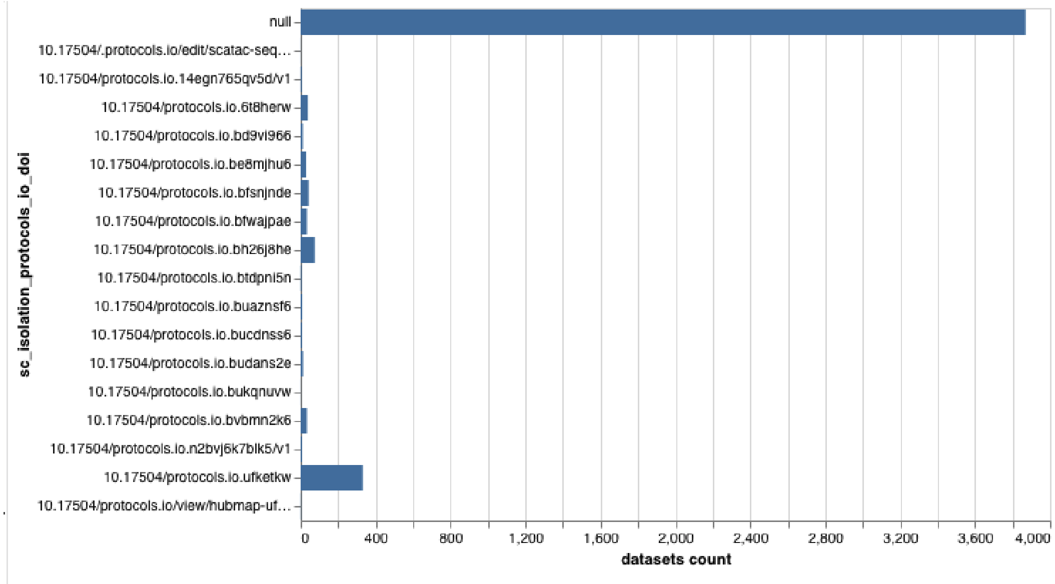
Query: Can you identify how many datasets are grouped by the respective units used for library concentration?



Reviewer	Score	Issue Categories	Comments
R1	Good		
R2	Improve		Approximately: yes, but finer grid would be easier to get more precise values
R3	Good		
R4	Improve		Can estimate but with only bars and grid lines every 1000 units will be rough
R5	Good		

Question ID: 2374

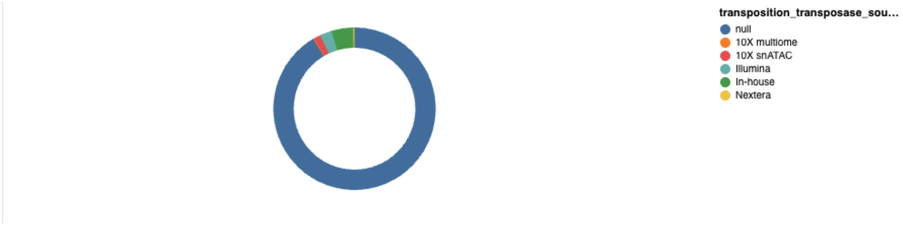
Query: Could you determine the number of datasets, segmented based on sc_isolation_protocols_io_doi identifiers?



Reviewer	Score	Issue Categories	Comments
R1	Good		
R2	Improve		null value leads to very difficult distinction of the other (actual) values, impossible to get a precise number
R3	Improve		Not sure what segmented means in this context
R4	Good		
R5	Good		

Question ID: 415537

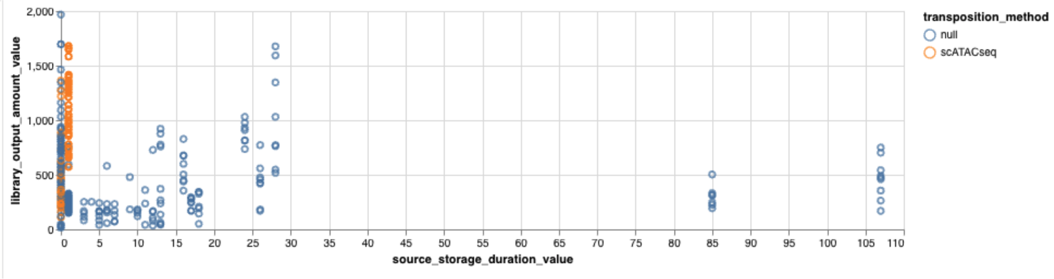
Query: Could you kindly create a circular diagram highlighting the transposition transposase origin?



Reviewer	Score	Issue Categories	Comments
R1	Good		
R2	Bad	Other	null value breaks the visualization + not sure whether "circular diagram" = donut chart
R3	Good		
R4	Good		
R5	Good		

Question ID: 865961

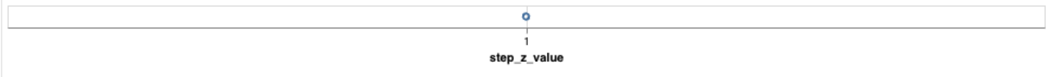
Query: Are there clusters of storage duration and library output values based on different chromatin capture methods?



Reviewer	Score	Issue Categories	Comments
R1	Bad	Other	transposition_method is chromatin capture methods? given that there only is 1 (and null), can't really answer this also again time value is missing unit
R2	Good		
R3	Good		
R4	Good		
R5	Good		

Question ID: 986397

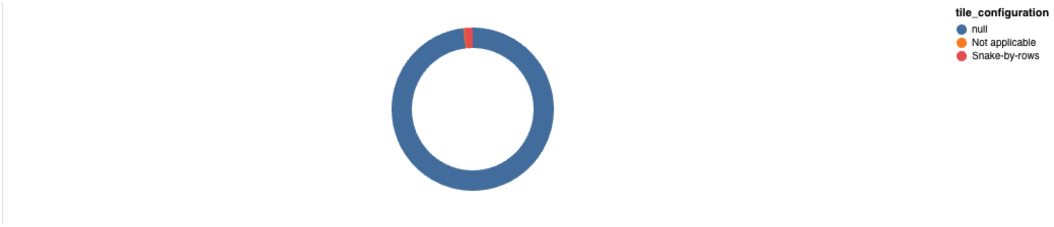
Query: Could you describe the pattern in step_z_value distribution?



Reviewer	Score	Issue Categories	Comments
R1	Bad	Malformed Visualization, Question Not Answered	
R2	Bad	Malformed Visualization	
R3	Bad	Malformed Visualization	Only one data point visible
R4	Good		
R5	Good		

Question ID: 415613

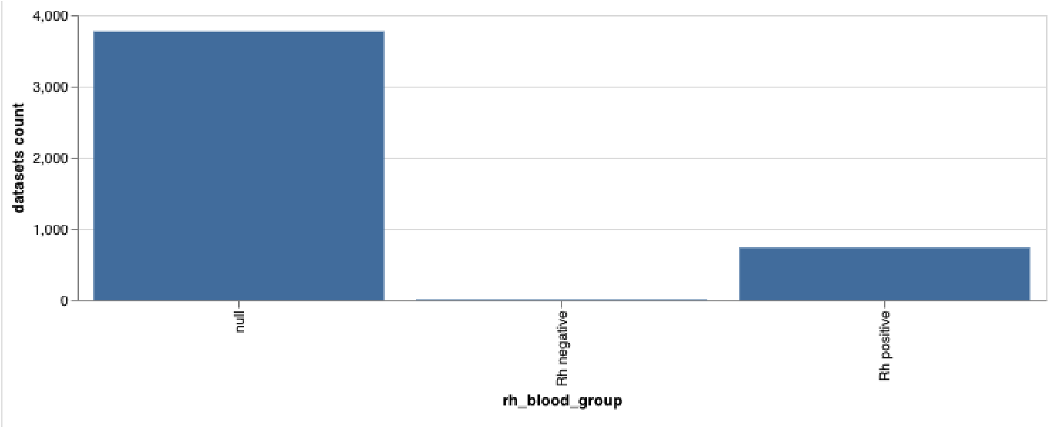
Query: Could you assemble a donut chart detailing the tile arrangement?



Reviewer	Score	Issue Categories	Comments
R1	Good		
R2	Bad	Other	null value overtakes whole visualization
R3	Improve		Need percentage number for detailing
R4	Good		
R5	Good		

Question ID: 7585

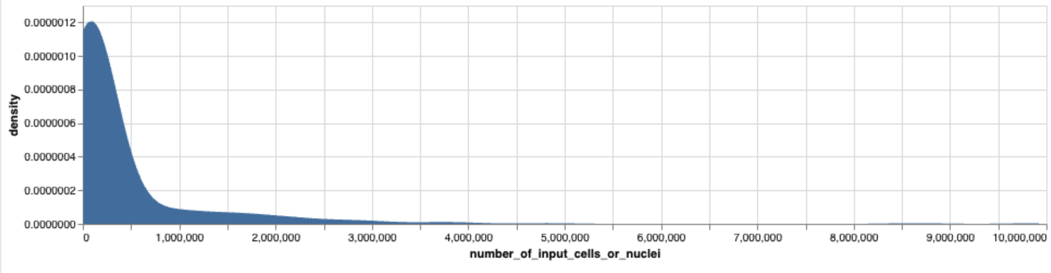
Query: How many dataset bundles are there when arranged by blood type?



Reviewer	Score	Issue Categories	Comments
R1	Bad	Question Not Answered	Visualization just showing rhesus factor, not blood group
R2	Good		
R3	Good		
R4	Improve		Rh factor is not the same as blood type
R5	Good		

Question ID: 985932

Query: Would you please furnish the distribution specifics of the number of input cells or nuclei?

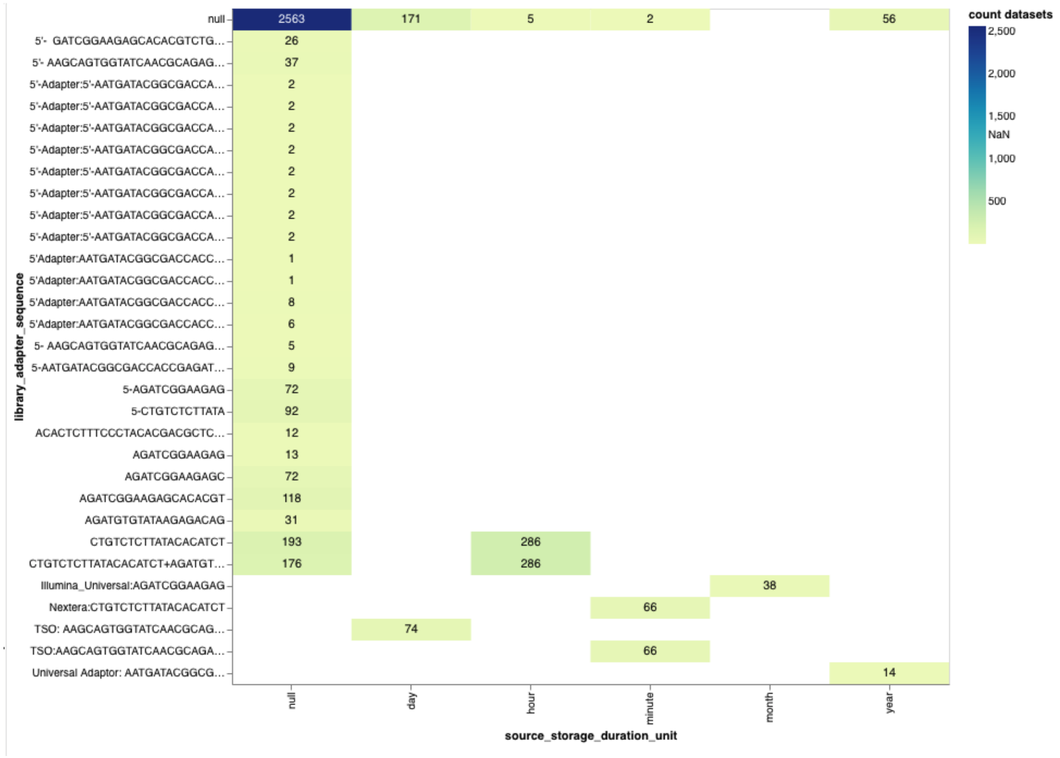


Reviewer	Score	Issue Categories	Comments
R1	Improve		decent visualization, but how can the number of input cells/nuclei be negative?
R2	Bad	Bad Question	
R3	Improve		question wording sounds weird
R4	Good		
R5	Good		

Note: This figure is slightly different from the version shown to the participant. In an earlier version of the visualization toolkit the visualization was not clipped at zero. The visualization itself has not changed.

Question ID: 542864

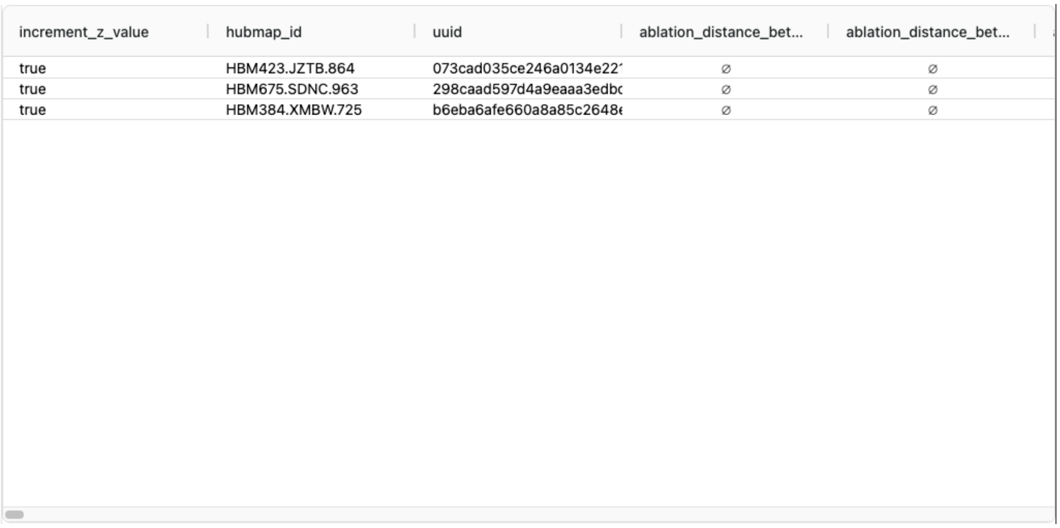
Query: Are there any visible clusters in datasets concerning library_adapter_sequence and time items were preserved?



Reviewer	Score	Issue Categories	Comments
R1	Bad	Malformed Visualization	time is grouped by unit, rather than time amount
R2	Good		
R3	Improve		The value for null badly skewed the color range
R4	Good		
R5	Improve		It was difficult for me to connect the two, i.e., color was showing clusters. Maybe colors could be more differentiable. or use shape as a visual encoding as well.

Question ID: 434208

Query: Can you sort these datasets by the gap in Z coordinates?

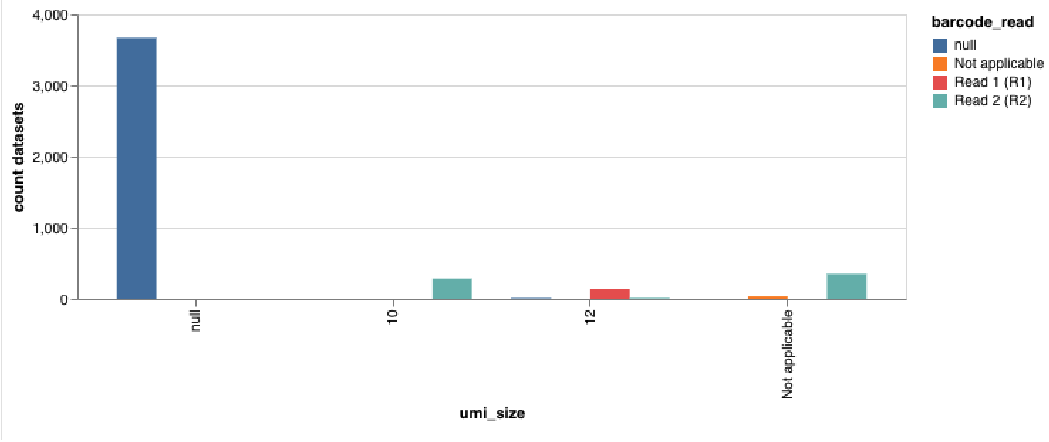


increment_z_value	hubmap_id	uuid	ablation_distance_bet...	ablation_distance_bet...
true	HBM423.JZTB.864	073cad035ce246a0134e22	∅	∅
true	HBM675.SDNC.963	298caad597d4a9eaaa3edb	∅	∅
true	HBM384.XMBW.725	b6eba6afe660a8a85c2648	∅	∅

Reviewer	Score	Issue Categories	Comments
R1	Bad	Question Not Answered	
R2	Bad	Other	the z coordinate of interest is either boolean, or somewhere we I need to scroll too far to the right, which makes it difficult to validate the question at first glance.
R3	Good		
R4	Improve		Columns could be sorted to reduce horizontal scrolling, as there are many columns and I am not sure which corresponds to "gap in Z coordinates"
R5	Good		

Question ID: 84173

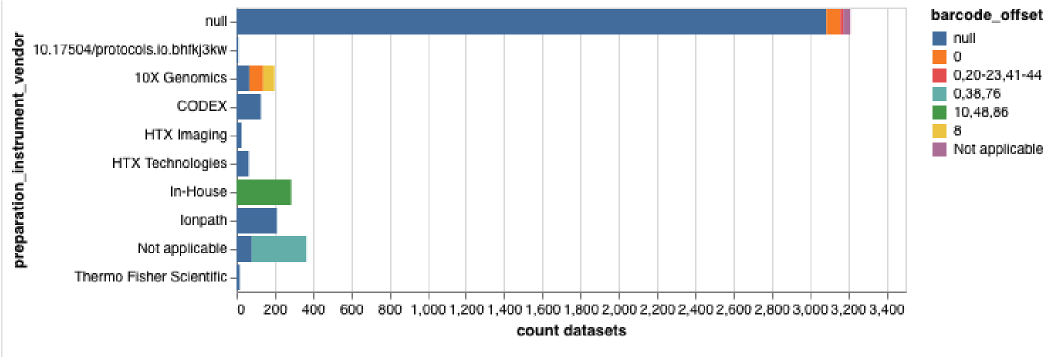
Query: Identify the distinct frequency for barcode reads within varying unique molecular index sizes.



Reviewer	Score	Issue Categories	Comments
R1	Bad		
R2	Improve		Low count items almost indistinguishable, could use different y-axis scale.
R3	Bad	Malformed Visualization, Other	The bars are not aligned with ticks. There are also two bars with the same colors.
R4	Good		
R5	Good		

Question ID: 138646

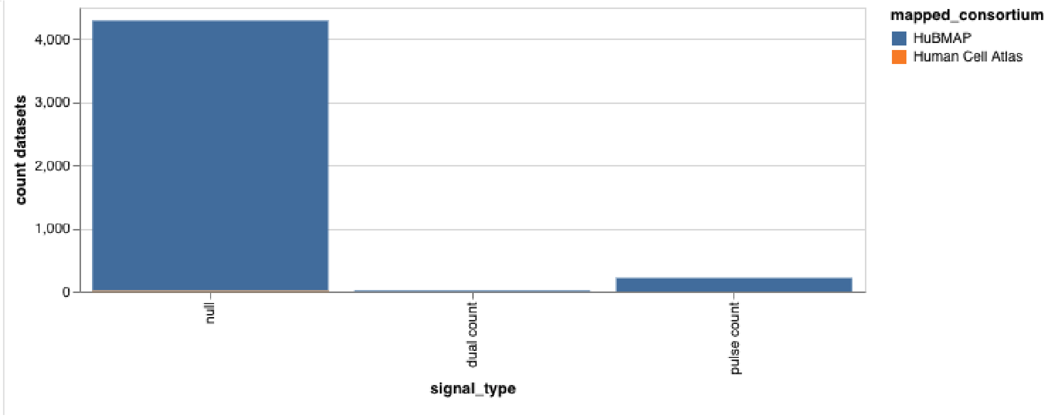
Query: I’m seeking the count of barcode offsets for each brand of preparation tool.



Reviewer	Score	Issue Categories	Comments
R1	Improve		I can find the approximate distribution from this chart, but individual numbers, especially for low counts, are hard to see. A tooltip could help.
R2	Improve		Encoding ’barcode_offset’ and ’count datasets’ could/should be switched
R3	Bad	Other	Null appears in both y-axis and legend. There’s something looking like a DOI of a paper on the y-axis.
R4	Good		
R5	Good		

Question ID: 32114

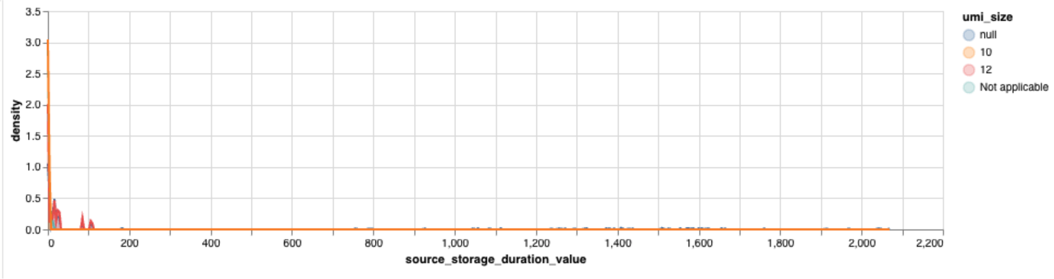
Query: Please provide the dataset count, categorized by consortium and kind of signal.



Reviewer	Score	Issue Categories	Comments
R1	Bad	Question Not Answered	HCA is mentioned in the legend but not shown in the visualization.
R2	Improve		the hubmap cell atlas data are almost not visible
R3	Improve		Need specific number annotated to each bar to be able to provide specific dataset count
R4	Good		
R5	Good		

Question ID: 989392

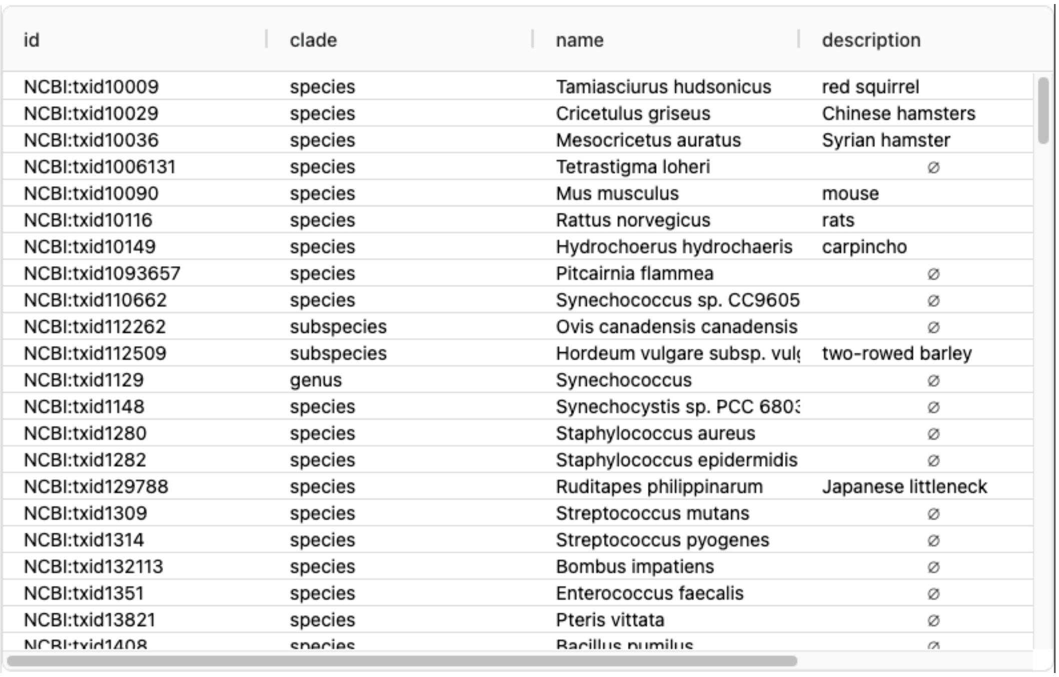
Query: Is the storage duration consistent with each barcode length?



Reviewer	Score	Issue Categories	Comments
R1	Good		
R2	Improve		Difficult to answer
R3	Bad	Question Not Answered	
R4	Improve		What does it mean by "consistent"? Also, I cannot distinguish the different lines from each other visually.
R5	Improve		I see distribution, rather than answer for consistency. If correlation is meant, maybe a different graph would be helpful?

Question ID: 419413

Query: How does the data for taxonomy numbers like those found in NCBI appear?



id	clade	name	description
NCBI:txid10009	species	Tamiasciurus hudsonicus	red squirrel
NCBI:txid10029	species	Cricetulus griseus	Chinese hamsters
NCBI:txid10036	species	Mesocricetus auratus	Syrian hamster
NCBI:txid1006131	species	Tetrastigma loheri	∅
NCBI:txid10090	species	Mus musculus	mouse
NCBI:txid10116	species	Rattus norvegicus	rats
NCBI:txid10149	species	Hydrochoerus hydrochaeris	carpincho
NCBI:txid1093657	species	Pitcairnia flammea	∅
NCBI:txid110662	species	Synechococcus sp. CC9605	∅
NCBI:txid112262	subspecies	Ovis canadensis canadensis	∅
NCBI:txid112509	subspecies	Hordeum vulgare subsp. vul	two-rowed barley
NCBI:txid1129	genus	Synechococcus	∅
NCBI:txid1148	species	Synechocystis sp. PCC 680	∅
NCBI:txid1280	species	Staphylococcus aureus	∅
NCBI:txid1282	species	Staphylococcus epidermidis	∅
NCBI:txid129788	species	Ruditapes philippinarum	Japanese littleneck
NCBI:txid1309	species	Streptococcus mutans	∅
NCBI:txid1314	species	Streptococcus pyogenes	∅
NCBI:txid132113	species	Bombus impatiens	∅
NCBI:txid1351	species	Enterococcus faecalis	∅
NCBI:txid13821	species	Pteris vittata	∅
NCBI:txid1408	species	Bacillus pumilus	∅

Reviewer	Score	Issue Categories	Comments
R1	Improve		Question is vague - "how does the data appear" - so unsure if the table answers this question.
R2	Bad	Bad Question	
R3	Bad	Bad Question	
R4	Good		
R5	Improve		Maybe this should be phrased like, give me a sample or summary of the data. When I see the question, I'd expect this to give me some analysis rather than a dataset?