SUPPLEMENTARY FILE - HIERARCHICAL UNIFORM MANIFOLD APPROXIMATION AND PROJECTION

Wilson E. Marcílio-Jr, Danilo M. Eler, Fernando V. Paulovich, Rafael M. Martins

November 6, 2021

1 Introduction

In this **Supplementary File**, we provide all the additional embeddings and analyses of the HUMAP paper. For metrics in which we compare the distributions using boxen-plots, we also use t-test to measure the statistical significance of their differences. Thus, we always show the statistical significance of the difference between HUMAP and the best performing technique other than HUMAP. Finally, we use the following code strategy to accommodate all of the results in tables:

- HUMAP $\rightarrow$ T1;
- UMAP $\rightarrow$ T2;
- HSNE GPU $\rightarrow$ T3;
- HSNE CPU $\rightarrow$ T4;
- MPHATE $\rightarrow$ T5.

2 Hierarchical Dimensionality Reduction comparison

Figures 1 and 2 show the all of the results for embedding hierarchy levels demonstrated in the paper. Tables 1, 2, and 3 show the the p-values after running t-test using HUMAP and the better performing technique other than UMAP.

	Hierarchy fit		Level 2		Level 1		Level 0	
Dataset	Techniques	p-value	Techniques	p-value	Techniques	p-value	Techniques	p-value
mammals	T1, T5	1.71-e5	T1, T5	0.048	T1, T5	2.46e-33	T1, T3	2.45e-39
EB	T1, T3	7.46e-9	T1, T3	0.247	T1, T3	1.49e-12	T1, T3	9.07e-22
MNIST	T1, T3	2.79e-25	T1, T3	0.069	T1, T3	1.39-e56	T1, T3	3.10e-59
FMNIST	T1, T3	9.16e-21	T1, T3	4.94e-28	T1, T3	8.07e-41	T1, T3	3.62e-58

Table 1: p-values for the running time metric.

	Level 2		Level 1		Level 0	
Dataset	Techniques	p-value	Techniques	p-value	Techniques	p-value
mammals	T1, T5	9.73e-19	T1, T5	1.52e-18	T1, T2	0.71
EB	T1, T2	0.09	T1, T2	0.016	T1, T5	3.08e-12
MNIST	T1, T2	6.70e-9	T1, T2	7.24e-14	T1, T2	3.05e-7
FMNIST	T1, T2	0.99	T1, T2	0.15	T1, T4	2.83e-18

Table 2: p-values for the DEMaP metric.

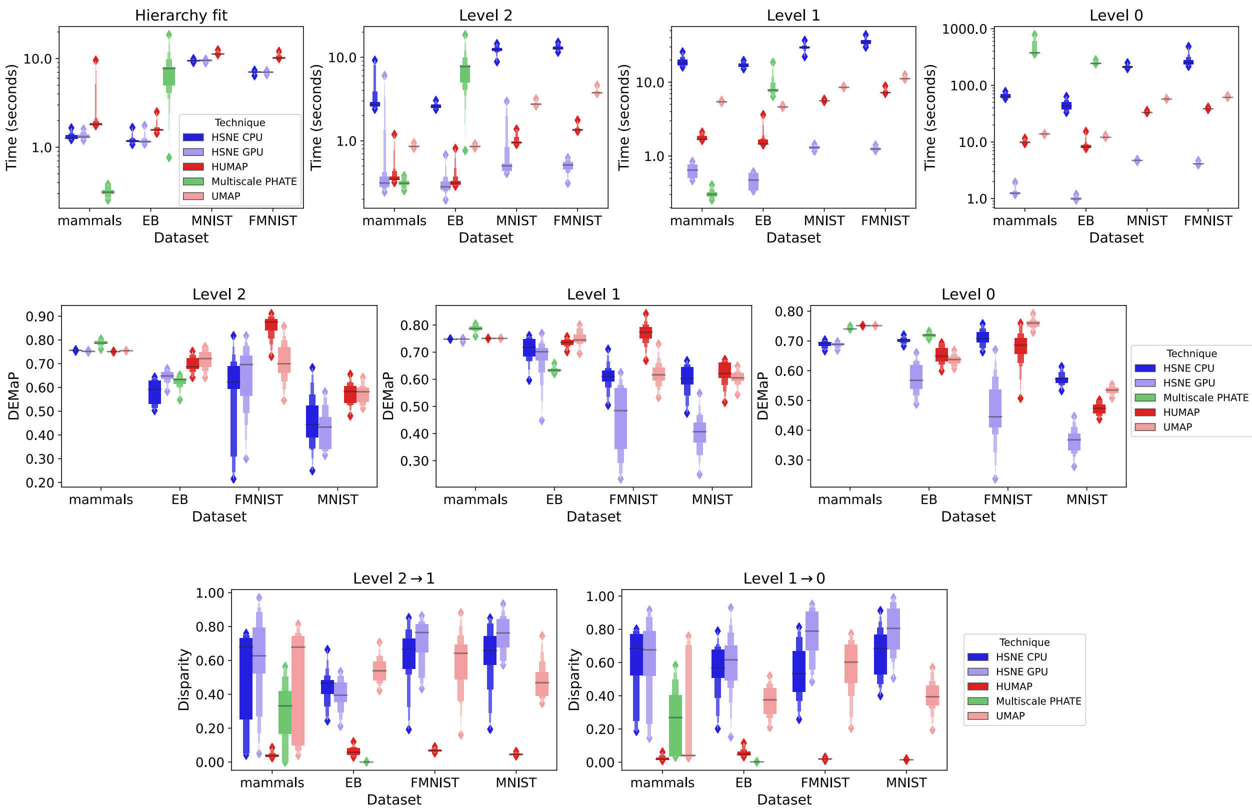

Figure 1: Results for Time, DEMaP, and Disparity for twenty executions of each HDR technique.

	Level 2 → 1		Level 1 → 0	
Dataset	Techniques	p-value	Techniques	p-value
mammals	T1, T5	1.44e-6	T1, T2	0.21
EB	T1, T5	1.02e-13	T1, T5	4.07e-13
MNIST	T1, T2	1.48e-15	T1, T4	8.82e-17
FMNIST	T1, T2	8.56e-21	T1, T2	1.50e-20

Table 3: p-values for the disparity metric.

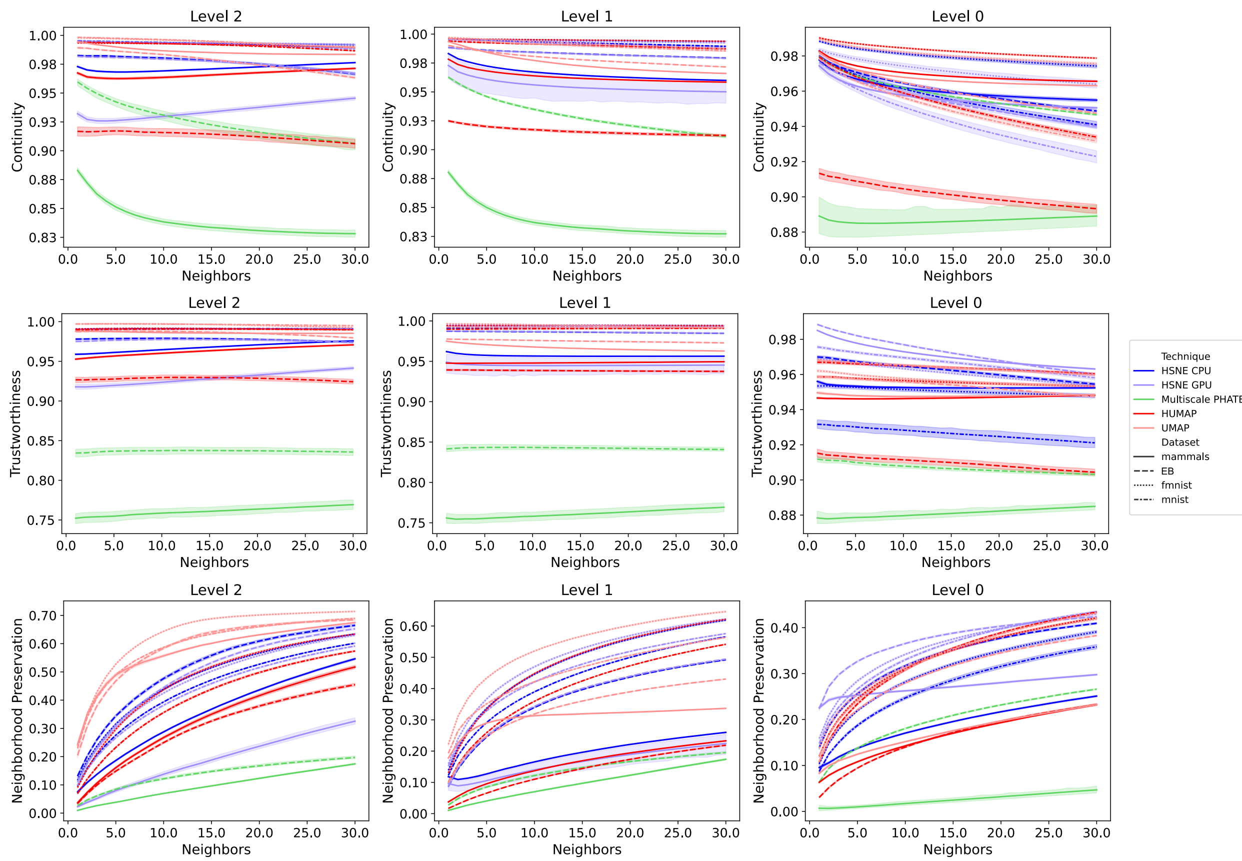

Figure 2: Results for Neighborhood Preservation, Continuity, and Trustworthiness for twenty executions of each HDR technique.

3 Local neighborhood analysis

HUMAP has a parameter (β) that can be used to tune the preservation of local neighborhoods. In the paper, we use $\beta = 0$. However, increasing such a value up to 1 makes HUMAP preserve more of local neighborhoods in higher hierarchy levels.

Figure 3 shows the behavior of HUMAP as we change β parameter according to the metrics used in the paper. We also compare it against HSNE on GPU.

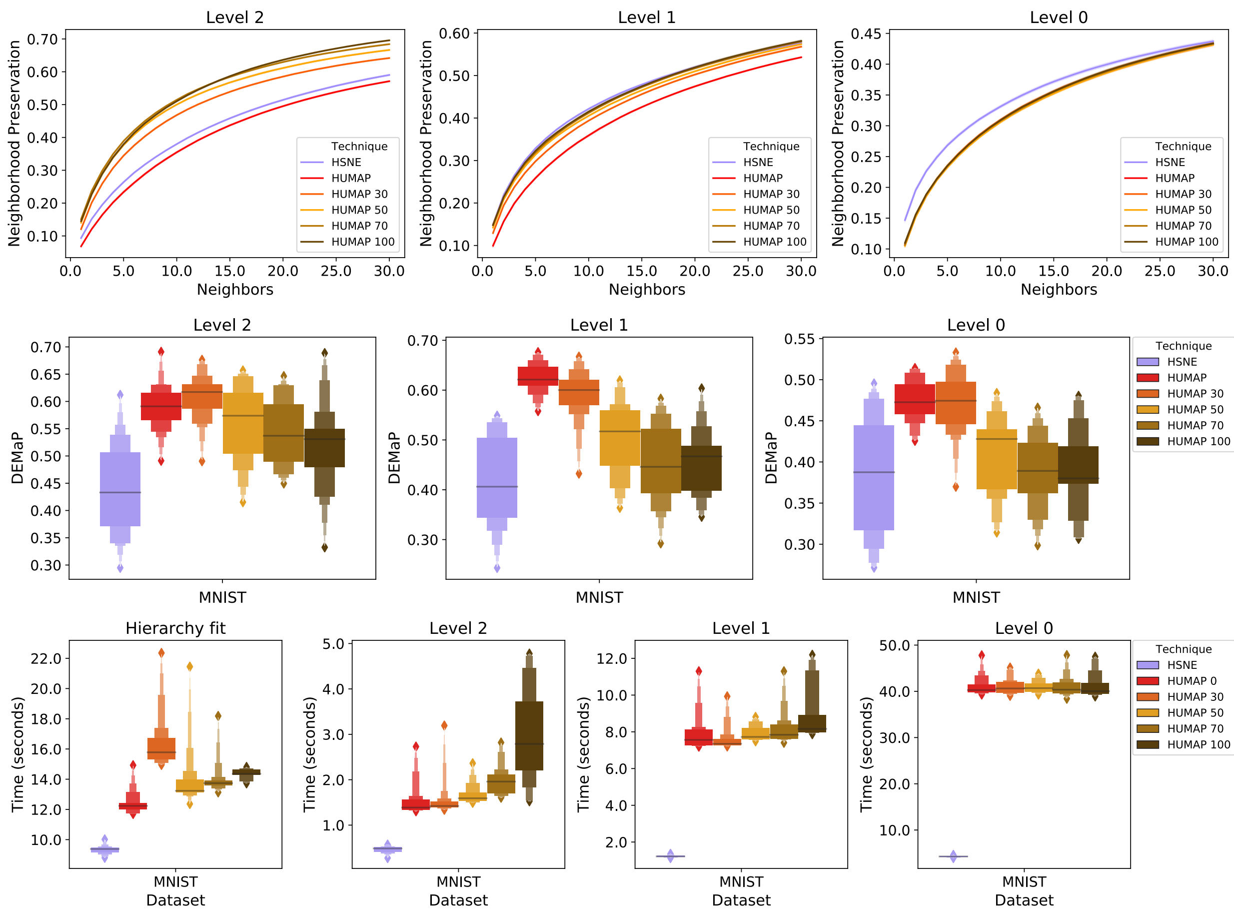

Figure 3: Metrics for HUMAP with different β values.

Figure 4 shows the projections with best neighborhood preservation for different HUMAP configurations and HSNE on GPU.

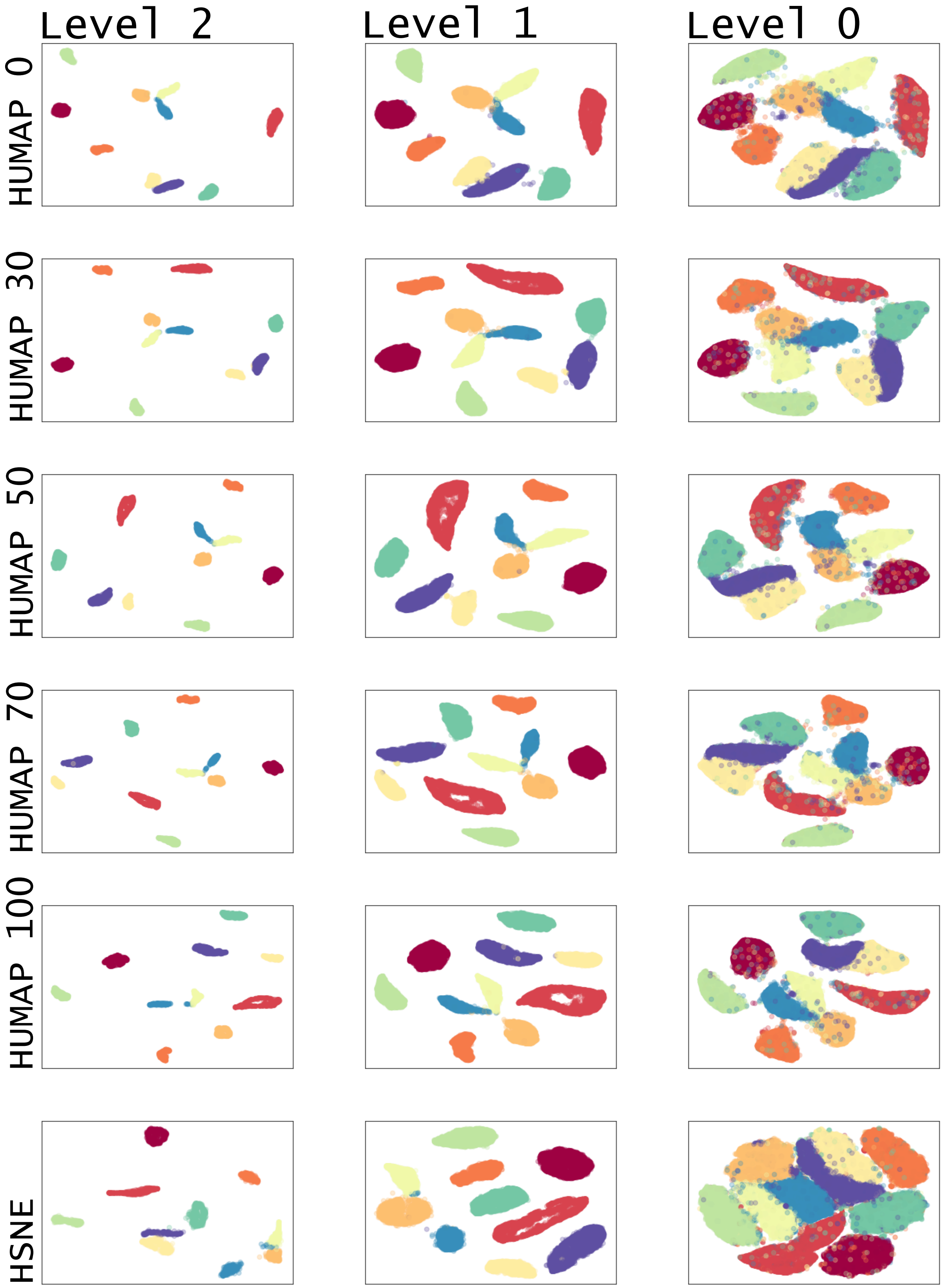

Figure 4: Projections for HUMAP with different β values.

4 Drill-down analysis

Figures 5 and 6 show the all of the results for the drill-down operation demonstrated in the paper. Tables 4, 5, and 6 show the the p-values after running t-test using HUMAP and the better performing technique other than UMAP.

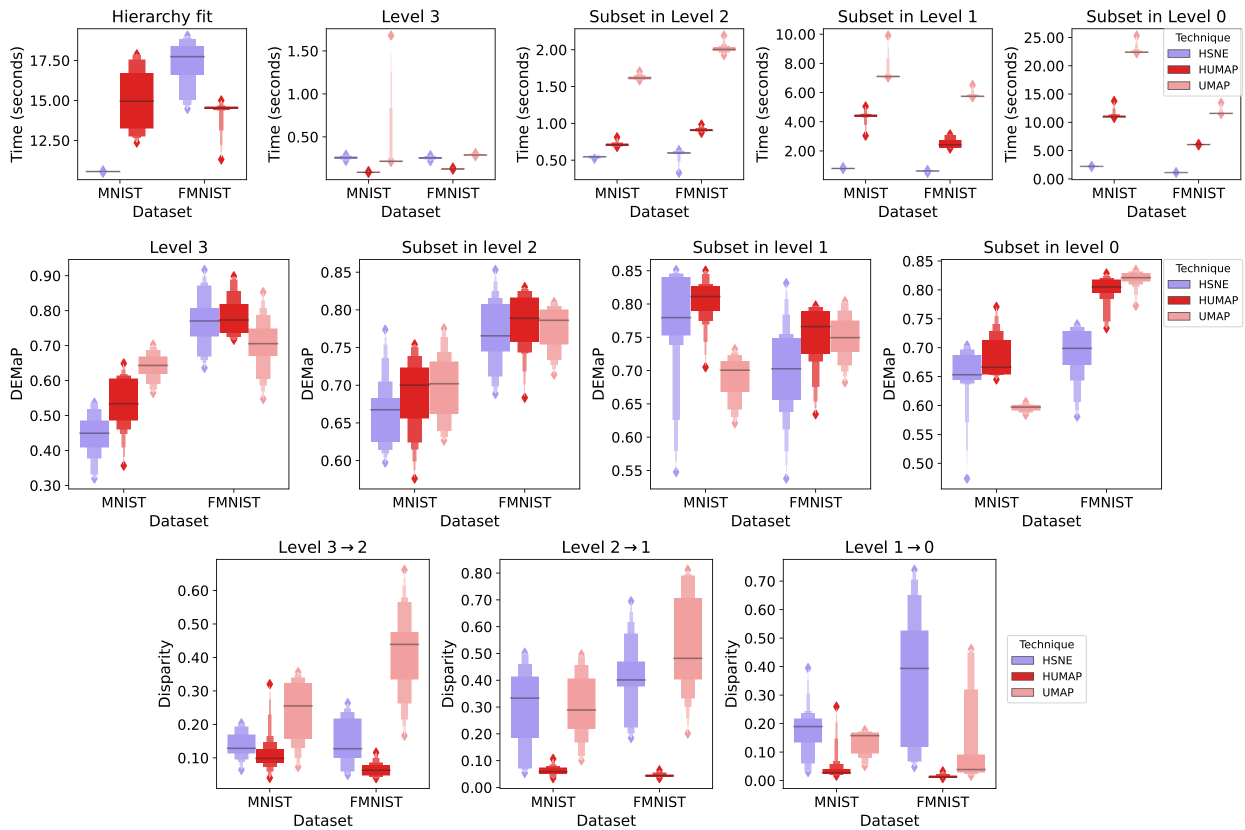

Figure 5: Results for Time, DEMaP, and Disparity for twenty executions of each HDR technique.

	Hierarchy Fit		Level 3		Subset in Level 2		Subset in Level 1		Subset in Level 0	
Dataset	Tech.	p-value	Tech.	p-value	Tech.	p-value	Tech.	p-value	Tech.	p-value
MNIST	T1, T4	3.19e-9	T1, T2	2.33e-46	T1, T4	1.32e-34	T1, T4	0.05	T1, T4	0.05
FMNIST	T1, T4	1.12e-12	T1, T4	1.59e-22	T1, T4	1.46e-24	T1, T4	2.81e-11	T1, T4	2.81e-11

Table 4: p-values for the running time metric.

	Level 3		Subset in Level 2		Subset in Level 1		Subset in Level 0	
Dataset	Techniques	p-value	Techniques	p-value	Techniques	p-value	Techniques	p-value
MNIST	T1, T2	2.17e-6	T1, T2	0.46	T1, T4	0.10	T1, T4	0.05
FMNIST	T1, T4	0.57	T1, T2	0.43	T1, T2	0.96	T1, T4	2.81e-11

Table 5: p-values for DEMaP metric.

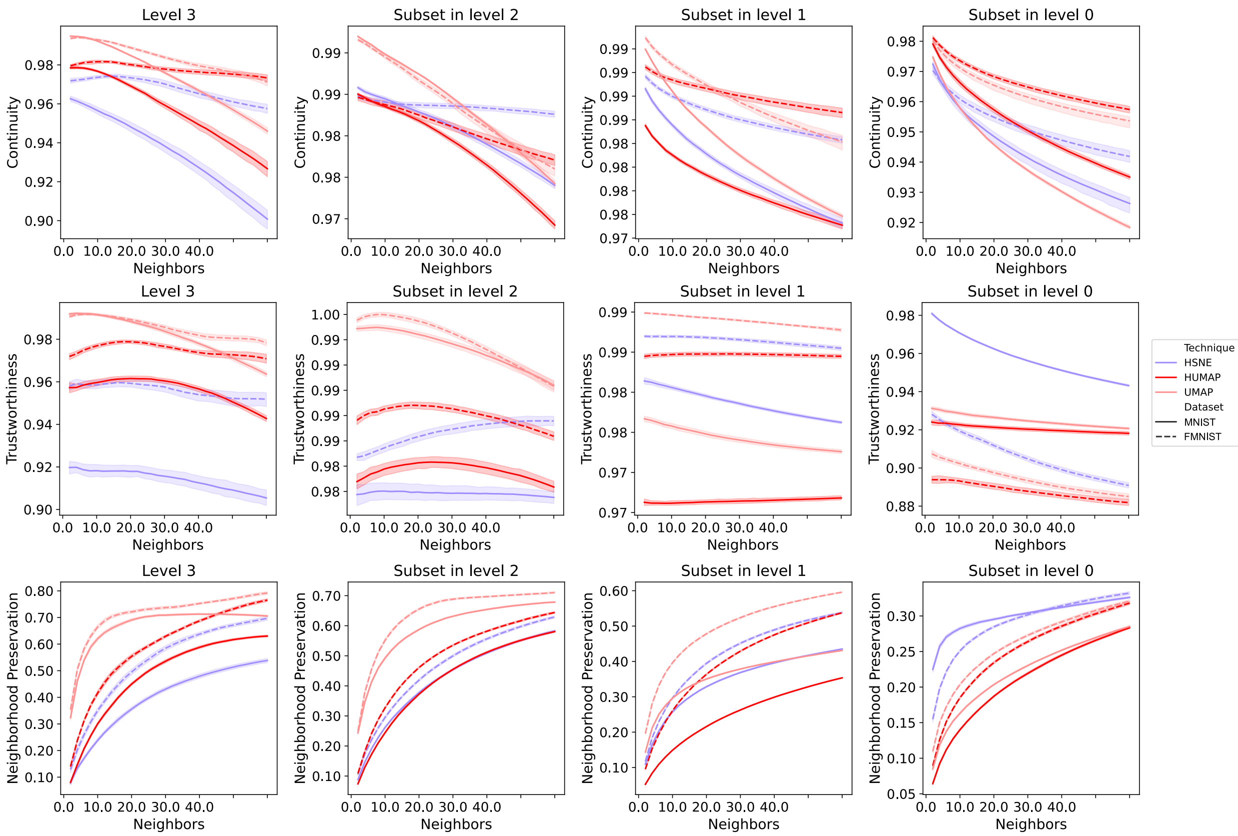

Figure 6: Results for Neighborhood Preservation, Continuity, and Trustworthiness for twenty executions of each HDR technique.

	Level 3 → 2		Level 2 → 1		Level 1 → 0	
Dataset	Techniques	p-value	Techniques	p-value	Techniques	p-value
MNIST	T1, T4	1.28e-5	T1, T2	6.93e-14	T1, T2	0.006
FMNIST	T1, T4	0.12	T1, T4	6.77e-8	T1, T2	4.06e-7

Table 6: p-values for the disparity metric.

5 Hierarchical Dimensionality Reduction without preserving the mental map

In the paper and Section 2, we show that HUMAP loses quality regarding DEMaP metric when projecting whole hierarchy levels. The following results show the same comparison of the paper and Section 2 by setting the fixing term (θ) to 1. As shown in the figures, by setting free the projection we get higher DEMaP values at the cost of losing the mental map.

Figures 7 and 8 show the projections of all hierarchy levels for the datasets used in the study.

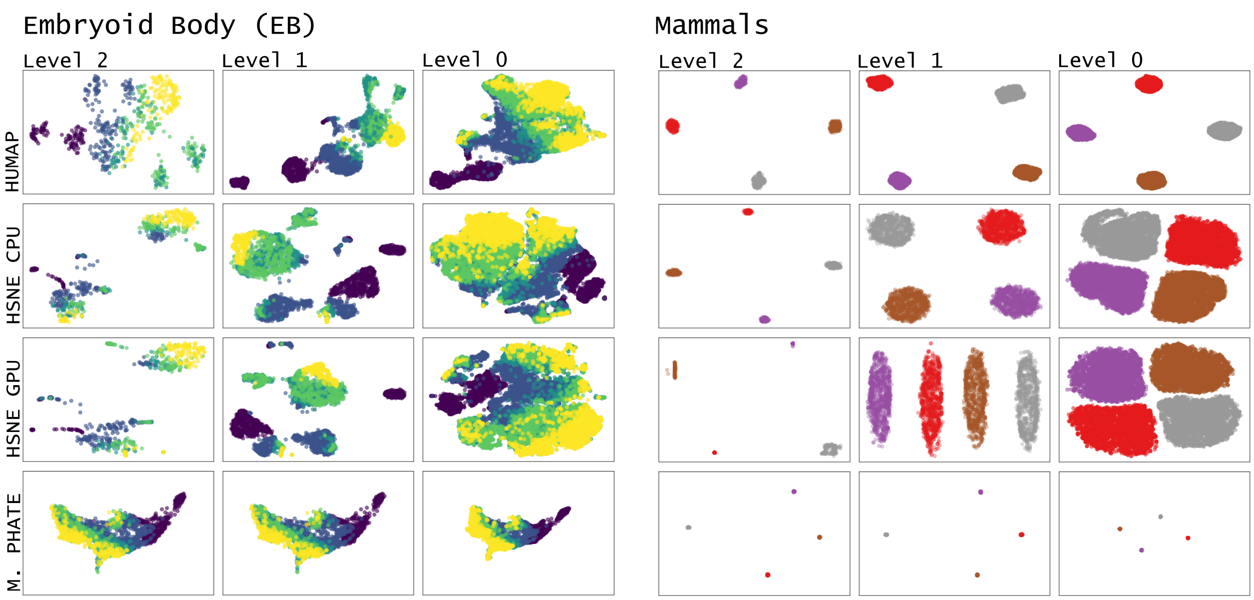

Figure 7: Projections for Embryoid Body and Mammals datasets.

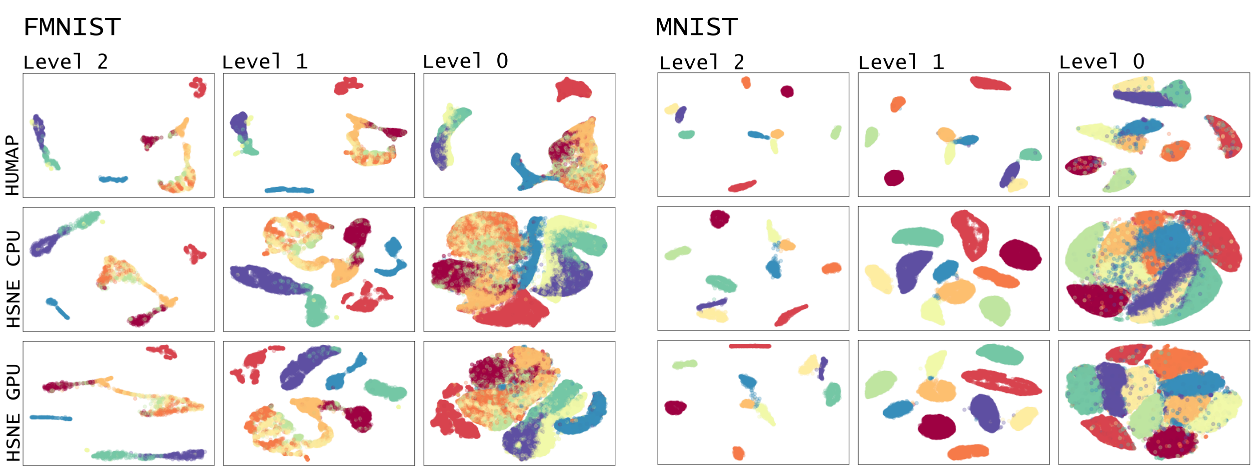

Figure 8: Projections for FMNIST and MNIST datasets.

Figure 9 shows the results after twenty runs for each hierarchical dimensionality reduction technique,

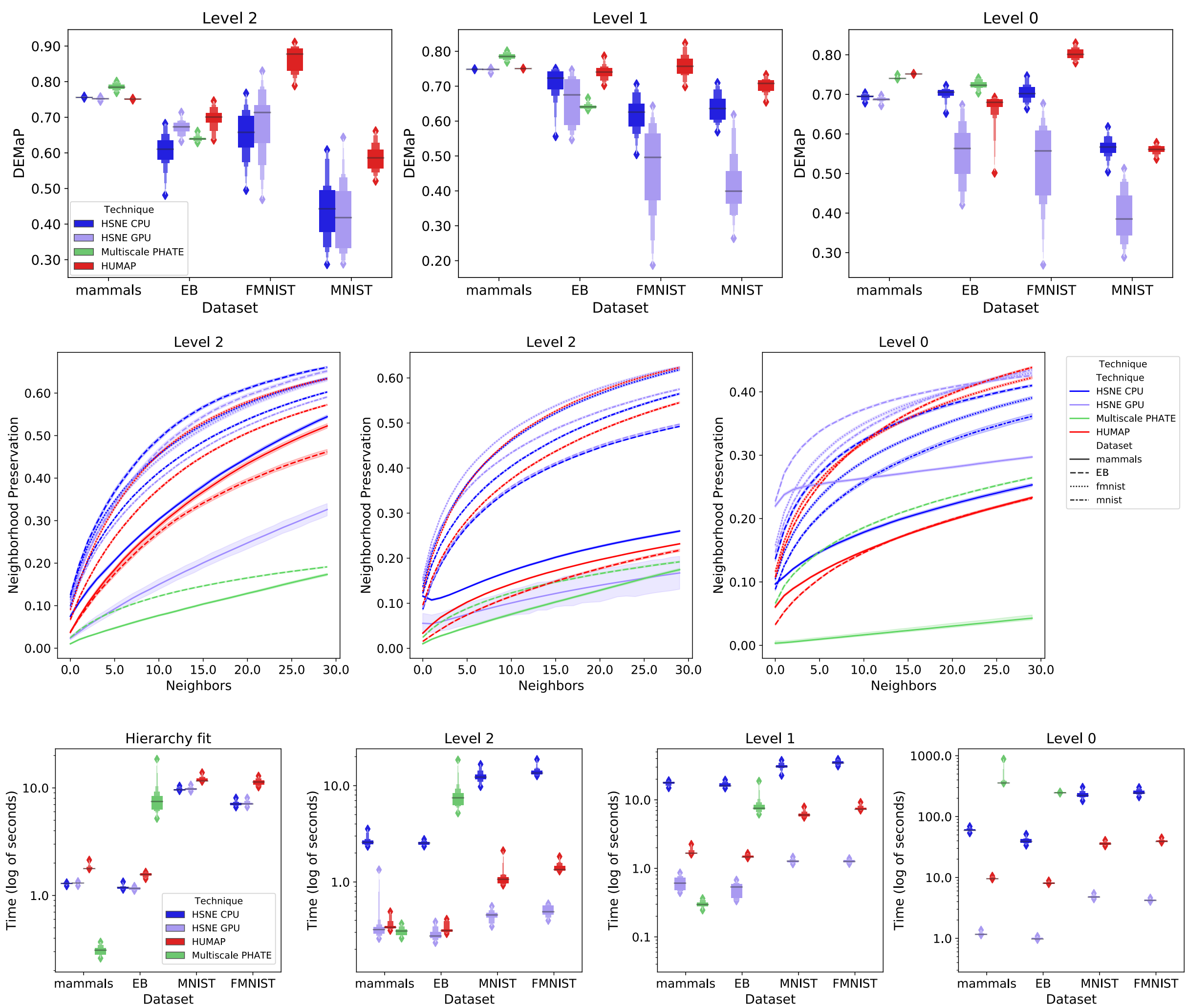

Figure 9: Numerical analysis for twenty executions of each HDR technique.

6 Additional use case - Exploration of COVID-19 tweets

This section aims to investigate the relationship of *tweets* about COVID-19 symptoms in the São Paulo state, Brazil. This dataset first appears in [1], where the authors investigate the whole dataset using contrastive analysis. Here, we use HUMAP to explore it using two hierarchical levels, aiming at discovering dominant structures (using the top-level) and detailed information about these dominant structures through interaction. The dataset was acquired from Twitter by querying COVID-19 symptoms (fever, high fever, cough, dry cough, difficulty breathing, and shortness of breath) retrieved from São Paulo state (Brazil) from March 2020 to August 2020. The authors classified the *tweets* according to their relevancy (relevant or not relevant) to COVID-19 infection. For this case study, we set HUMAP to freely find the embeddings as we drill-down the hierarchy by not projecting the low-levels bases on the higher levels (i.e., setting the fixing term $\theta = 1.0$).

To explore the dataset, we manually defined clusters in the visual space using lasso selection, as shown in Fig. 10(A-B). After associating each data point to a cluster using this procedure, we compute topics (F) and proceed to the second and lowest level of the hierarchy, choosing the desired cluster (e.g., cluster ▭). Then, we also compute the topics for the manually defined cluster of the new hierarchical level.

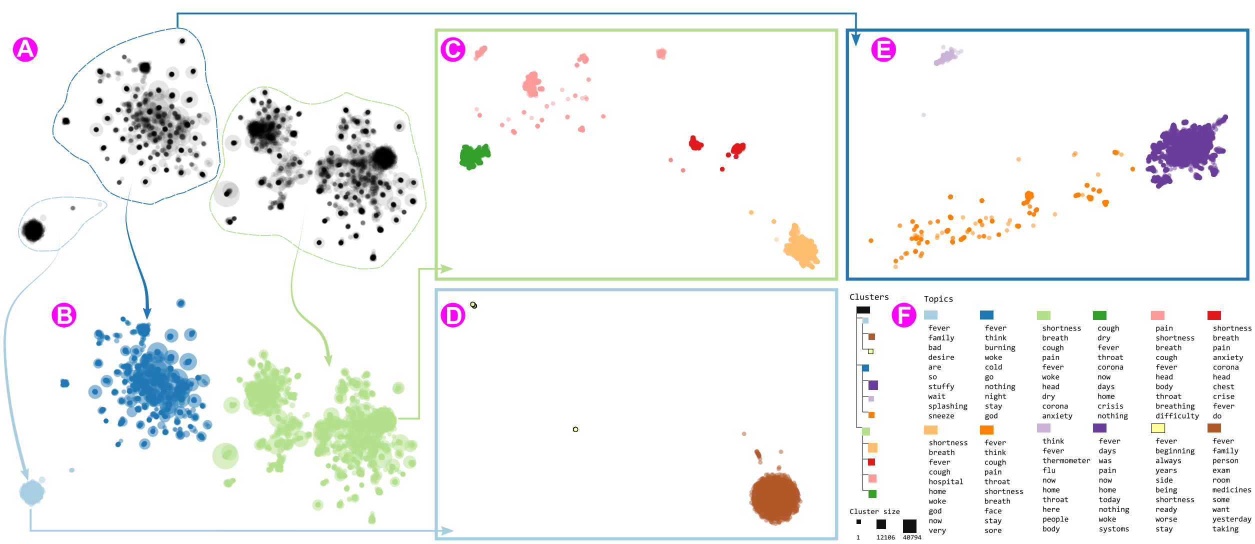

Figure 10: HUMAP exploration and annotation of a document collection of COVID-19 *tweets*. The top-level hierarchy level shows unlabeled data points and three major structures (A). We annotate these three clusters (B) and compute their topics computed (F). For each cluster in (B), we also project their corresponding level (and final) hierarchy to look for other patterns, annotating the dataset (C, D, E) and computing their topics.

Fig. 10(B) shows three clusters with different characteristics. First, cluster ▭ is very cohesive and dissimilar from the other two dominant structures. Second, due to visual proximity, cluster ▬ and ▭ share some information. Finally, these clusters (mainly ▬ and ▭) contain various substructures that need further investigation. The topics for these three major clusters further explain their organization in the visual space (F). That is, cluster ▭ presents important terms related to respiratory problems that COVID-19 infection may cause, such as `stuffy` nose, `sneeze`, or `splashing`. Other important terms in this cluster can indicate *tweets* about individuals waiting (`wait`) for COVID-19 `exams` or are associated with `desire` to `sneeze`, which supports a hypothesis that this cluster corresponds to individuals worrying about symptoms that are not well-associated with the disease. Clusters ▬ and ▭ are associated with well-known COVID-19 symptoms. Cluster ▬ shows three terms in the topic associated with fever: `fever`, `burning`, and `cold`. These words describe the way individuals feel while worrying about COVID-19 infection or even describing the symptoms after a (possibly) positive test. Other terms such as `think`, `woke`, and `night` might be associated to phrases describing experiences with COVID-19 symptoms, such as: "I **think** I have **fever**", "Today I **woke-up** in the middle of the **night burning** in **fever**". Lastly, cluster ▭ is associated with dry cough and shortness of breath. The terms for this cluster in Fig. 10(F) show that the *tweets* talk about these symptoms while the term `anxiety` could be the cause of shortness of breath [2].

We proceed to cluster ▭ to investigate its substructures and retrieve more information about the *tweets* associated with it. Then, we define the major structures and compute the topics for the clusters, as shown Fig. 10 (C) and (F). Here, there are a few interesting patterns. The first one is that the previous hierarchy level sufficiently gives an overview of the data organization since the topic for cluster ■ encodes most of the information expressed in these sub-clusters. The second and most interesting aspect is the global relationship among these structures apparent in the embedding. The topics retrieved from each local structure explains this aspect. The leftmost cluster (■) shows terms related to `dry cough`, `throat` pain, and a few other important terms. Cluster ■ has a relationship with cluster ■, which also adds `difficult breathing`, `body pain`, and headache ("pain in the **head**", using a direct translation from Portuguese). In cluster ■, `shortness of breath`, `anxiety`, `crisis`, and breast become more important. These symptoms might be easily confused with anxiety crisis, a common problem during COVID-19 pandemic [3]. The last cluster (■) seems to be the most related to COVID-19 symptoms, showing the majority of term: `shortness of breath`, `fever`, and `cough`.

Analyzing the substructures of cluster ■, we define three clusters as shown in Fig. 10 (E). As well-defined by the higher-level cluster, all data points refer mainly to the COVID-19 symptom of fever. However, there are a few characteristics that might explain the differentiation of these clusters. For example, cluster ■ present terms related to `cough`, `shortness` of `breath`, which do not appear in the other clusters. Lastly, cluster ■ has an interesting characteristic since its topic suggest that individuals are commenting about the period in which they experience fever: `fever`, `days`, `was`, `today`, and `home`.

The most cohesive cluster of top-level (■) and its two subclusters (■ and ■)revealed by drilling-down the hierarchy show tweets that seem to talk more about daily aspects than truly COVID-19 symptoms.

In this case study, our technique reveals with little effort (looking at top-level structure) aspects of the classified COVID-19 dataset that a considerable number of *tweets* may not indicate relevant information. Further that, we can also augment the trust in this analysis by expanding the cluster of interest.

References

[1] Wilson E. Marcílio-Jr, Danilo M. Eler, and Rogério E. Garcia. Contrastive analysis for scatterplot-based representations of dimensionality reduction. *Computers Graphics*, 2021.

[2] Pietro Smirni, Gioacchino Lavanco, and Daniela Smirni. Anxiety in older adolescents at the time of covid-19. *Journal of Clinical Medicine*, 9(10), 2020.

[3] Raquel Brandini De Boni, Vicent Balanzá-Martínez, Jurema Correa Mota, Taiane De Azevedo Cardoso, Pedro Ballester, Beatriz Atienza-Carbonell, Francisco I Bastos, and Flavio Kapczinski. Depression, anxiety, and lifestyle among essential workers: A web survey from brazil and spain during the covid-19 pandemic. *J Med Internet Res*, 22(10), Oct 2020.