Emory CXR An Embeddings Study

Thoracic Rangers

What?

- An interactive visualization tool for guiding subgroup analysis.
  - Comparing how different models interpret CXR embeddings.
  - Understand outlier regions

Methods

- Dataset: Emory CXR
- The models:
  - RaD-DINO
  - MedGemma
  - CheXagent
  - MedImageInsight
  - BiomedCLIP
- Visualization Tool: Embedding Atlas

Process

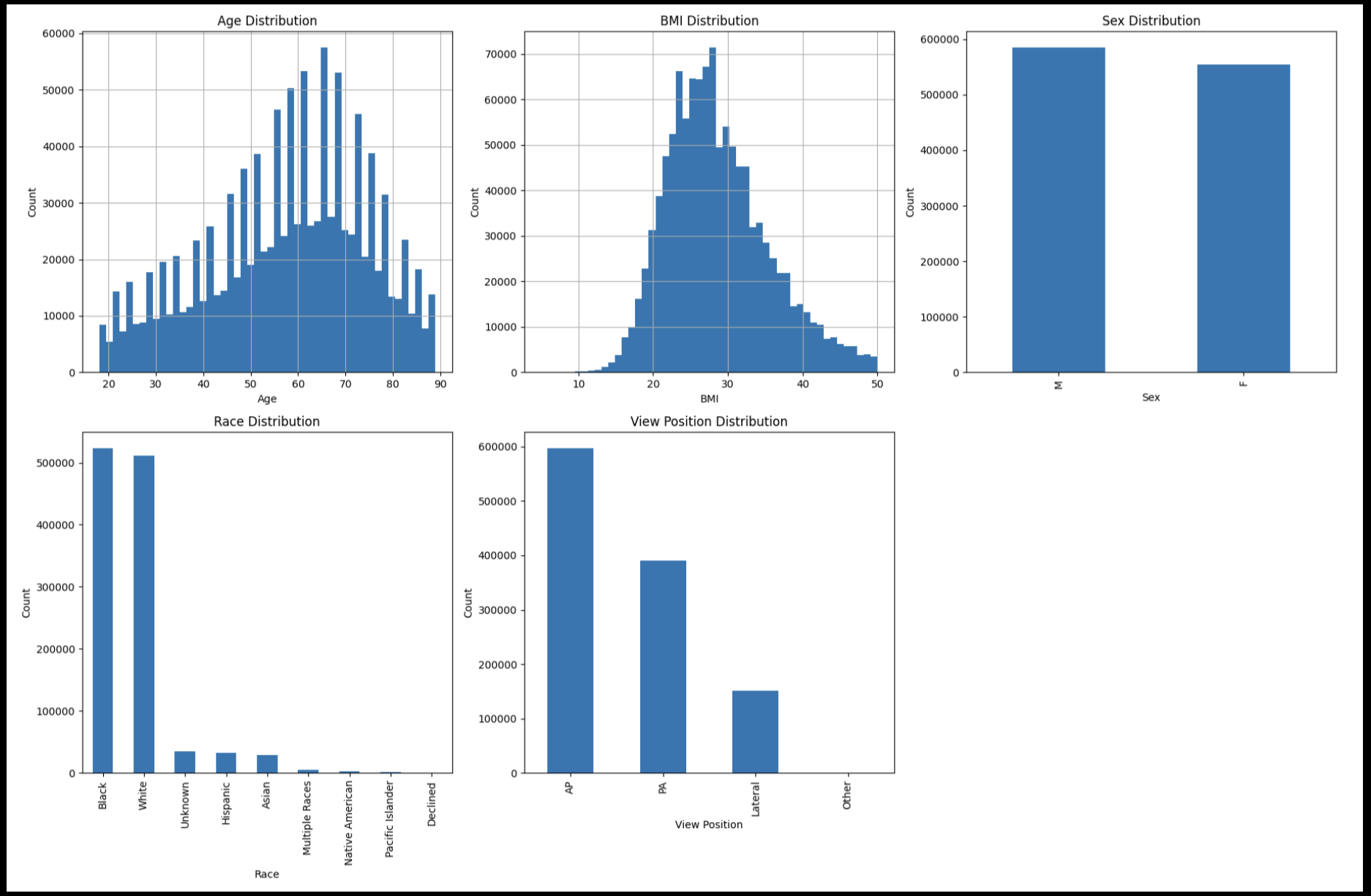
+ Step 1 – Exploratory Data Analysis on CXR all metadata and labels
+ Step 2 – Select 100k samples for each embedding model, same SOP for direct comparison
+ Step 3 - Dimensionality reduction of the embeddings.
+ Step 4 – Generate 2D Visualization of Embeddings using UMAP with Cosine Similarity
+ Step 5 – Interactive Real-time Clustering (Embedding Atlas)
+ Step 6- Example usage: Sample images from each cluster to provide insights and guide subgroup Analysis.

EDA - EMORY CXR

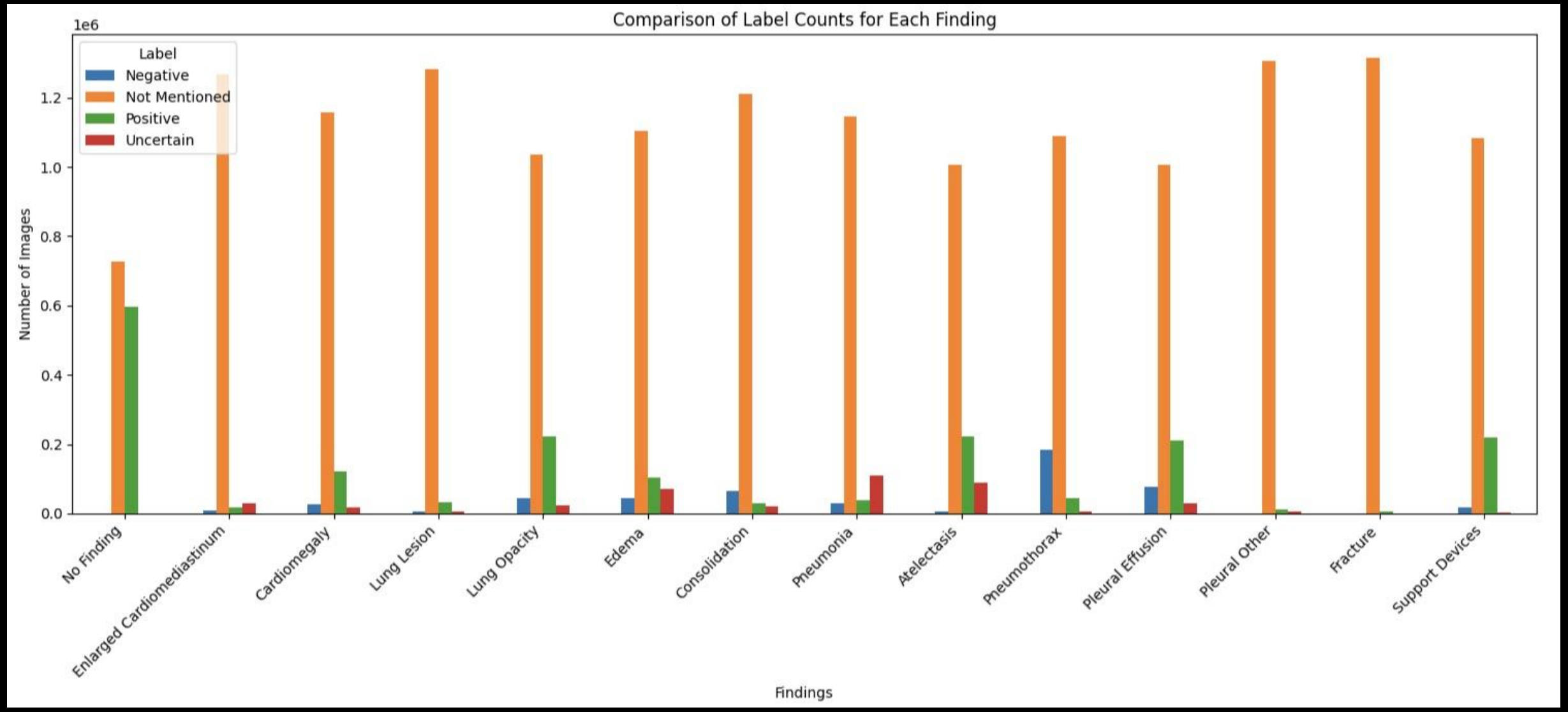
- The Emory CXR dataset contains Chest X-Ray (CXR) images embeddings

```
Metadata: (2430209, 19)
Findings: (1323308, 15)
Cardiomegaly: (2486502, 4)
Manufacturer: (1853067, 4)
SDOH: (159946, 388)
```

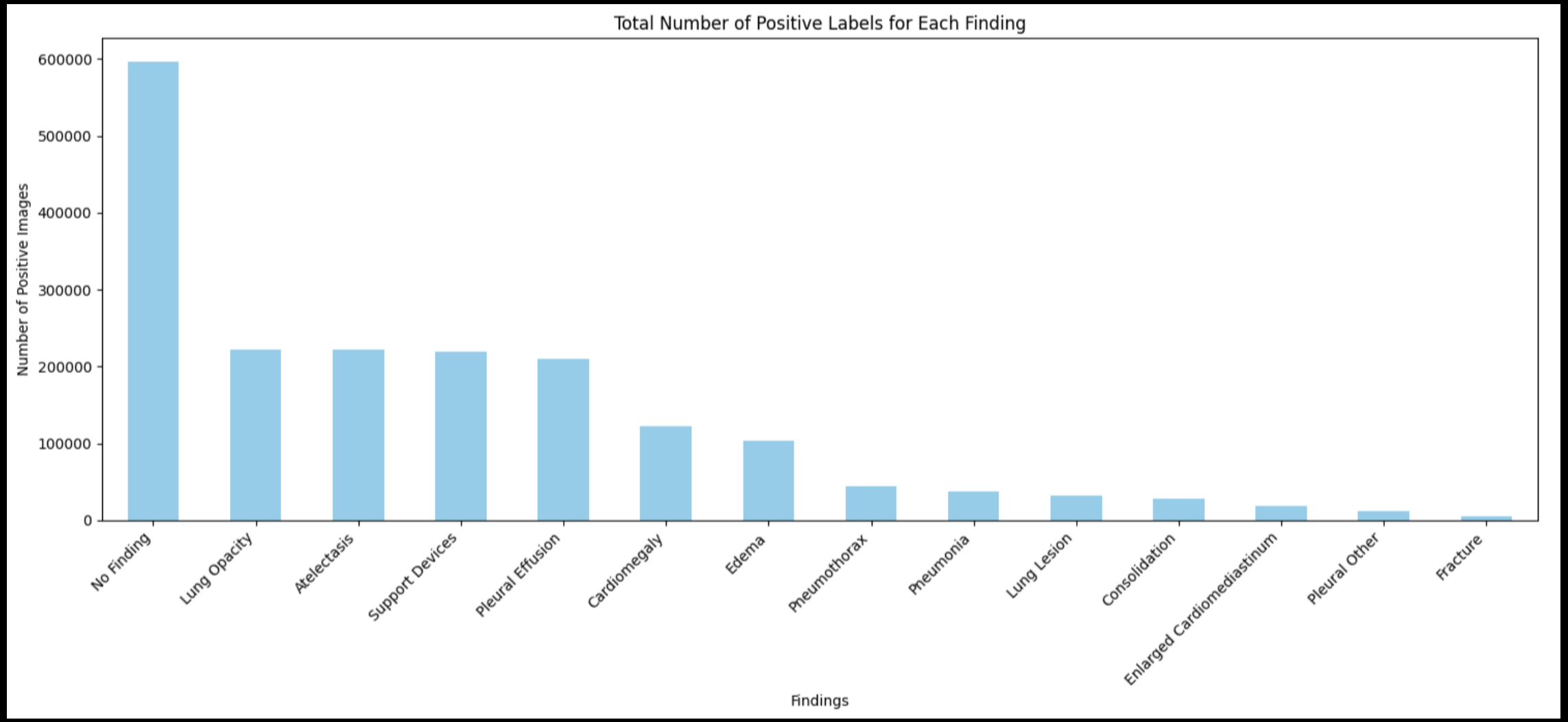
EDA - EMORY CXR



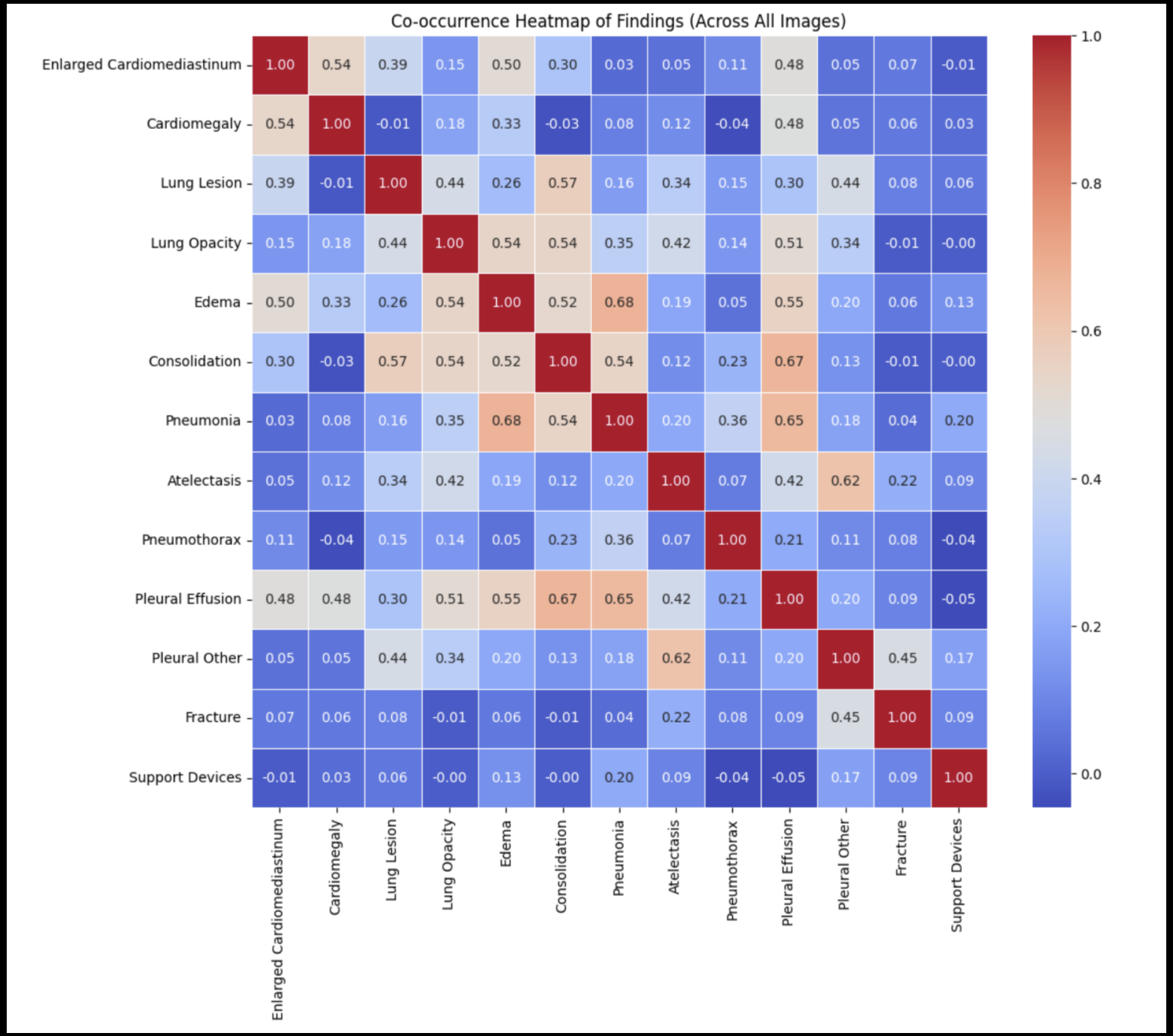
EDA - EMORY CXR



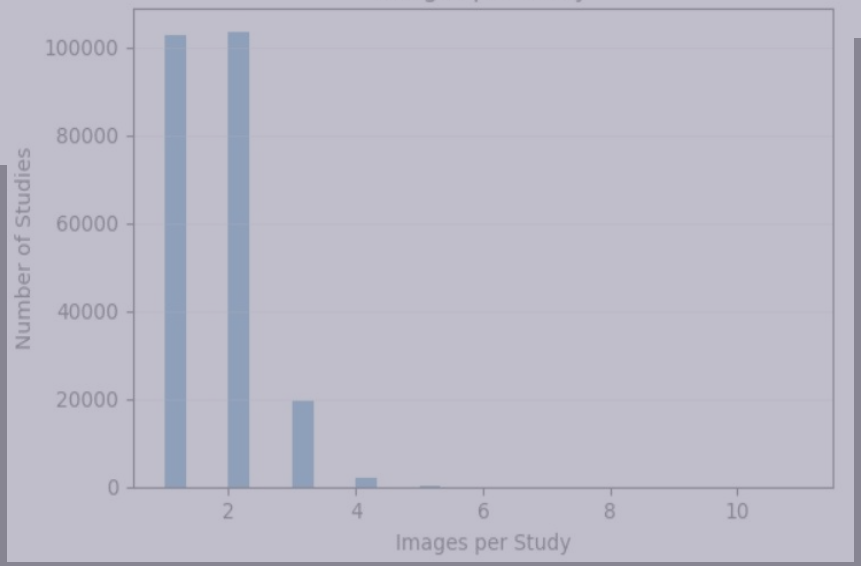
EDA - EMORY CXR



EDA - EMORY CXR

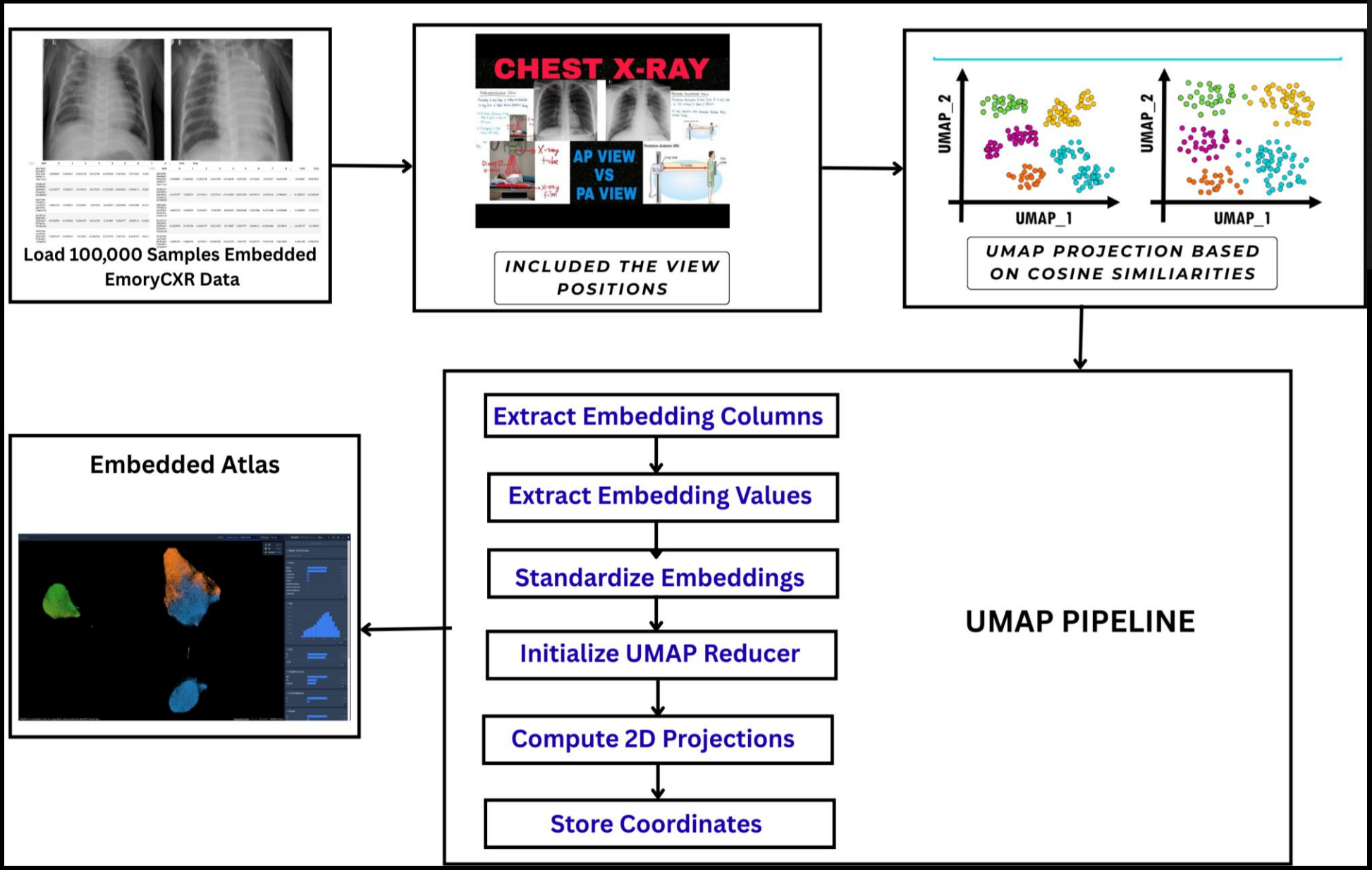


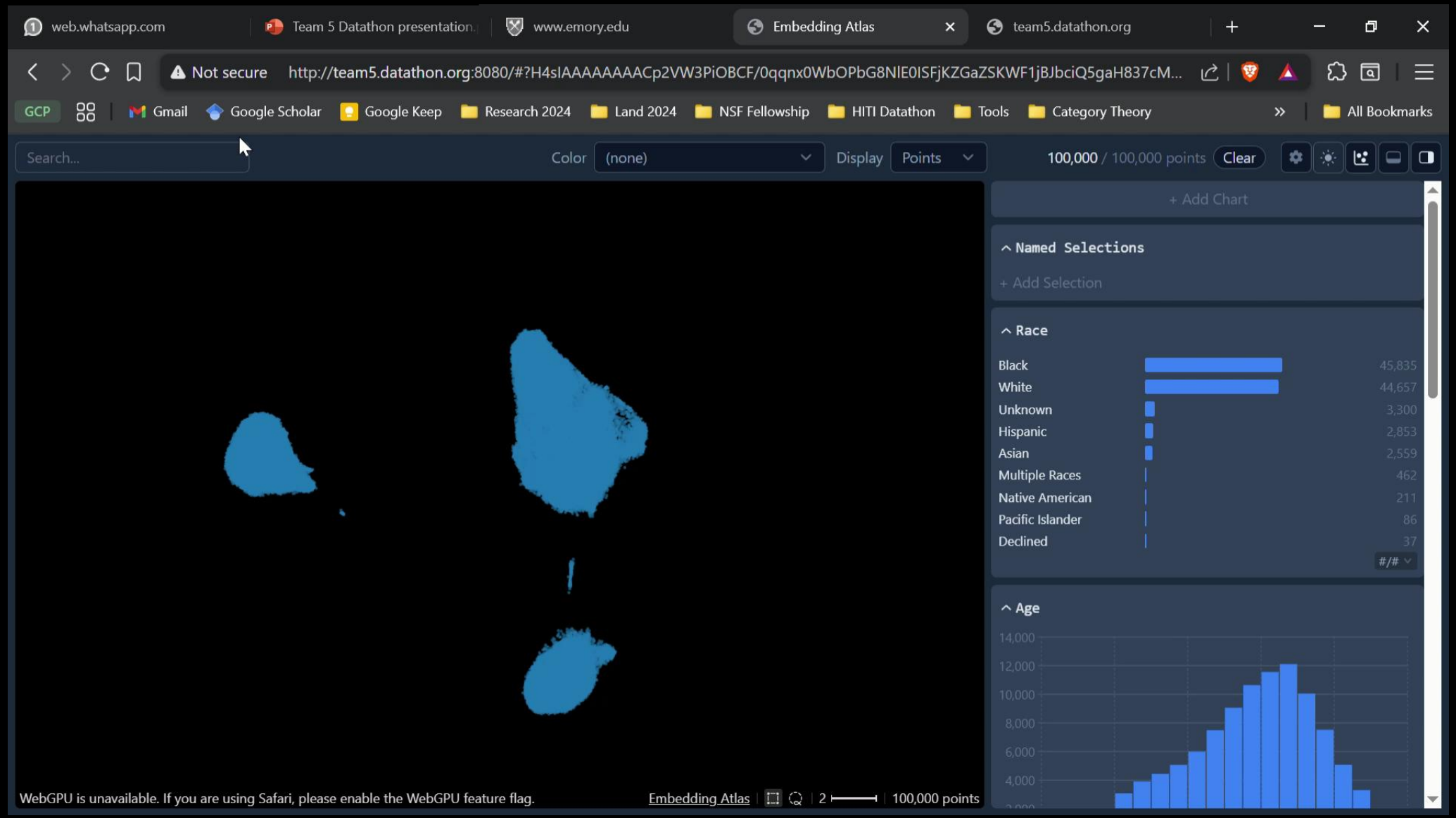
EDA - MIMIC CXR



ADDENDUM: REDACTED

DEMO PIPELINE




web.whatsapp.com
Team 5 Datathon presentation
www.emory.edu
Embedding Atlas
team5.datathon.org
Not secure
http://team5.datathon.org:8080/#?H4sIAAAAAAAACp2VW3PiOBCF/0qqnx0WbOPbG8NIE0ISFjKZGaZSKWF1jBJbciQ5gaH837cM...
GCP
Gmail
Google Scholar
Google Keep
Research 2024
Land 2024
NSF Fellowship
HITI Datathon
Tools
Category Theory
All Bookmarks
Search...
Color
(none)
Display
Points
100,000 / 100,000 points
Clear
+ Add Chart
Named Selections
+ Add Selection
Race
Black 45,835
White 44,657
Unknown 3,300
Hispanic 2,853
Asian 2,559
Multiple Races 462
Native American 211
Pacific Islander 86
Declined 37
#/#
Age
14,000
12,000
10,000
8,000
6,000
4,000
WebGPU is unavailable. If you are using Safari, please enable the WebGPU feature flag.
Embedding Atlas
2
100,000 points

Our Limitations - Design

- Sampling method may not be accurate representative of the full dataset.
- Quantification and comparison of embeddings is a difficult unsolved problem.
- The tool still requires a lot of expansion to interpret/analyze the embedding atlas.

Our Limitations - Practical

- Time consumed identifying team dynamics and fully understanding our data and tools.

- Resource limitations for fine-tuning the embeddings on foundation models.

What is the way forward?

- Compare clustering when a single feature is mis-labelled.
- Comparing different CXR datasets on a same embedding model.
- How well models trained on each set of embeddings classify a single outcome (e.g. pneumonia)


THORACIC RANGERS

Appendix

Model Classification

Model	Input/Output	Use Case	Clinical Relevance
MedGemma	Image + Text	Radiology Reports	Improve Workflow Increase Efficiency
BioMedCLIP	Image + Text	Radiology Reports	Improve Workflow Increase Efficiency Knowledge Transfer
RaD-DINO	Image	Classification	Encoding
MedImageInsight	Image	Classification	Encoding
CheXpert	Image + Text	Labels – Embeddings	Training Models