

# Visualizing Neural Network Loss Landscapes During Training

Arjun Prakash

PI

arjun\_prakash@brown.edu

Kevin Wang

Co-PI

kevin\_a\_wang@brown.edu

Randal Balestrieri

Collaborator

randal\_balestrieri@brown.edu

October 14, 2024

## Abstract

We aim to develop a novel tool for visualizing the loss landscapes of neural networks. This tool will use adaptive sampling and just-in-time compilation to meaningfully increase the speed of generating loss landscapes such that it can be a useful artifact to use during training, rather than post-training which is the current state of the art.

# 1 Reviews & Responses

## 1.0.1 Review 1

**Reviewer:** asathe1

**Overall Score:** 4

**Interdisciplinary:** 4-5

The proposal deals with artificial neural network models and their training dynamics. This is a field in and of itself, making this a highly interdisciplinary work.

**Scientific:** 3-4

The contributions include potential for new insights into the training dynamics of artificial neural networks (ANNs).

**Visualization:** 4

A new method to sample loss values towards visualization is proposed. Rather than present a novel way to visualize loss landscapes (which are essentially 2-parameter functions), the method improves an existing approach by sampling loss values adaptively and improving efficiency.

**Significant:** 3

This work is significant in advancing our understanding of machine learning theory.

**Novel:** 4

The proposal suggests an improvement on an existing visualization method.

**Goals Clearly Stated:** 3

The goals are clearly stated.

**Likelihood of Success:** 3

The likelihood of success is high.

### **Strengths:**

- Clearly identified niche/target area.
- Growing need for a solution in this domain.
- Clearly identified background work to compare to.
- Clearly stated improvement.

### **Weaknesses:**

- The main contribution is an algorithmic improvement in a visualization technique; the main novel contribution may not be a visualization contribution.

## Other Comments for Discussion:

- What does visualizing the loss landscape contribute to theory-building for language models?
- 

### 1.0.2 Response

We agree that the visualization proposed might not be novel. To improve this, we aim to include volumetric visualization to look at a 3d slice. In addition we will identify critical points and construct a contour tree. This will provide a completely novel way to visualize the complexity of the loss landscape.

---

## 1.1 Review 2

**Reviewer:** bbutaney

**Overall Score:** 3.5

### **Interdisciplinary:** 3

While the proposal does not directly tie itself to a science outside the discipline of computer science, neural network visualization can aid researchers in almost any academic discipline, from biology to linguistics. Some examples of this could be elaborated upon to better get this point across in the proposal.

### **Scientific:** 2

From a machine learning perspective, these visualizations could substantially improve the intuitiveness of training neural networks.

### **Visualization:** 5

Since other methods exist to perform this task, it is hard to rate the visualization contribution too highly.

### **Significant:** 3

The project is overall very significant, since any improvement in our ability to train neural networks has wide-reaching benefits. After all, neural networks are used in almost every major scientific discipline to some extent nowadays.

### **Novel:** 3

The finer details of the proposed project are novel, but the overarching goals are less so.

### **Goals Clearly Stated:** 1

The goals are clearly stated.

### **Likelihood of Success:** 4

It is difficult to tell whether the result will be faster than existing methods, but I am confident that the team can produce a working prototype nonetheless.

**Strengths:**

- Wide-reaching benefits for many fields.
- Baselines and goals are very clear.
- 6-week plan is clear.

**Weaknesses:**

- Not as novel as some other proposals might be.
- There is not enough evidence to make me fully confident that they can develop a method that is actually faster than the baseline.
- Less directly stated significance outside of computer science.

**Other Comments for Discussion:**

---

**1.2 Response**

To address the review's second concern first. We are confident that we can gain significant speed by just-in-time (JIT) compilation and adaptive sampling. [1] demonstrates that just using JIT on a simple function results in a 2.5 times speed-up. These savings compound over a complex program.

As stated in our response to review 1, we will include volumetric visualisations and contour trees.

For the final concern, while we believe that neural networks are inherently interdisciplinary, we will aim to apply our method to physics informed networks as a stretch goal.

---

**1.3 Review 3**

**Reviewer:** bleahey

**Overall Score:** 2.5

**Interdisciplinary:** 3

The proposal relates to deep learning interpretability and visualization, two somewhat related fields. Potential for impact in other fields is present but unstated.

**Scientific:** 1

Mid-training insights and more efficient loss landscape visualization would be a very important and useful development for a variety of deep learning applications. This has not currently been implemented due to computational constraints. Elaboration on the method of acceleration would increase this score.

**Visualization: 4**

The visualization techniques themselves are not necessarily novel, but the improvement in speed and efficiency of visualization is.

**Significant: 3**

The proposal clearly relates to other works and demonstrates how it will build upon them. Targeting more specific insights during training may increase its potential significance.

**Novel: 2**

Loss landscapes have been visualized before but are relatively unexplored and have yet to be implemented at the speed of real-time training.

**Goals Clearly Stated: 1**

The goals are clearly stated.

**Likelihood of Success: 3**

This is a fairly difficult computational task (training speeds are a high bar), but the team is experienced within this field and has a clear plan for the project.

**Strengths:**

- Potential for significant impact in a variety of deep learning applications.
- Team is experienced and has a clear plan for the project.

**Weaknesses:**

- Visualization techniques are not as novel as some other proposals.
- The acceleration method is not elaborated on.

**Other Comments for Discussion:**

Was already very familiar with this proposal and excited to see how it turns out!

---

**1.3.1 response**

We agree that we can improve our visualizations by including a volumetric slice and countour graph. \_\_\_\_\_

**1.4 Review 4**

**Reviewer:** Eric Xia

**Overall Score:** 3.5

**Interdisciplinary: 5**

The interdisciplinary goals of this project relate to the application of neural networks, which are ubiquitous in many different fields, such as biology, physics, mathematics, and language. However, the visualization component itself does not support research outside of the domain of computer science.

**Scientific: 2**

A faster loss landscape visualization could save ML researchers time in developing improved gradient descent techniques, and lead to faster convergence times in general. This would be highly invaluable in effectively utilizing computing resources, given the \$1 trillion investment being made on generative AI across industry. The time saved may be marginal compared to existing methods, but the possibility of enabling online visualization is a valuable direction to explore.

**Visualization: 4**

The contribution given in the "Aims" section, a dynamic view of post-training loss, does not seem to be a visualization contribution. However, the justification in the "Significance" section provides a good visualization contribution: namely, by comparing loss landscape visualizations across pairs of hyperparameters it is possible to understand the optimal parameter space more easily. I am somewhat skeptical of the notion that a spatial visualization (or 2D with isolines) is going to lead to insights in training. One important reason to create these plots is to identify cases of local minima traps; however, my impression is that this is not a huge problem with adaptive moment estimation. If there was a way to identify overfitting, it would be a very helpful tool for ML practitioners.

**Significant: 4**

There already exist tools which produce loss landscape visualizations post-training. However, if the speedups produced by using adaptive sampling are significant enough, this could enable a new way of working with ML models. I have some hesitation in endorsing the effectiveness of using a tool like this during training, as in order to visualize data and derive insights from it, the data needs to be fairly complete. A partial loss landscape (e.g., the top slice of Figure 1) would be quite uninformative, as there would be no way of inferring the final optimal coordinates. I am also not convinced that the original problem of the post-training visualization taking a long time—15 seconds—is completely worthwhile.

**Novel: 5**

To date, no loss-landscape visualization tool has used adaptive sampling to generate visualizations, so this would be a novel approach to the problem. However, the overarching goal of loss-landscape visualization is not a new one. I am somewhat confused as to whether JIT compilation would be effective given the size and density of the data. Perhaps there exist optimization approaches which make sense in the context of 3D post-training visualization.

**Goals Clearly Stated: 3**

The goals are enumerated well. It is not exactly clear how online visual feedback will lead to better or easier characterization of the behavior of a neural network.

**Likelihood of Success: 3**

The goal of making loss-landscape visualization orders of magnitude faster seems to be doable, as the existing methods simply loop over all data points.

**Strengths:**

- Clear pathway to basic implementation.
- Improvements to basic implementation could lead to significant changes in ML methodology, with widespread consequences for the field.

**Weaknesses:**

- Basic implementation lacks scientific significance.
- Existing well-maintained tools in the field.

**Other Comments for Discussion:**

Was already very familiar with this proposal and excited to see how it turns out!

---

**1.4.1 Response**

We thank the reviewer for the very thorough feedback. We push back on the idea that the visualization does not support research outside of computer science. Indeed, training neural networks has proved to be successful in many domains such as robotics, biology and physics. However, we do understand that it is not clear from our proposal how visualizing the loss is helpful. To address this concern, we aim to use physics informed neural networks as one application of our tool. As far as visualisation and novelty are concerned, we agree with the reviewer. To strengthen our contribution we will add volumetric visualisations and contour trees as two new methods to understand the complexity of the loss landscape. These methods have not been applied to neural networks as far as we know.

---

**1.5 Review 5**

**Reviewer:** Thais Del Rosario Hernandez

**Overall Score:** 3

**Interdisciplinary: 3**

The proposal is highly interdisciplinary. Neural networks (NNs) are useful tools in many disciplines, and informing training through visualization of loss landscapes is relevant to any application of NNs. The score could be further improved by citing some reviews that inform the reader about the broad use of NNs.

**Scientific: 2**

The proposed work would provide significant improvements to the NN training workflow, as well as help users understand and optimize the parameters that affect training outcomes.

**Visualization: 4**

The main goal of the proposed work is to increase the generation speed of the visualization (i.e., neural

loss landscape). However, the visualization itself will not be significantly different from those produced by existing tools.

**Significant: 2**

The proposed increase in speed would be a greatly significant and broadly applicable contribution to NN workflows. The proposer also has a plan to confirm that their tool is not only faster but remains accurate.

**Novel: 3**

While the algorithm that will be used as the basis for the speed improvements (adaptive sampling) is not novel, its implementation in this context will be novel.

**Goals Clearly Stated: 1**

The goals are clearly stated throughout the proposal.

**Likelihood of Success: 3**

The implementation goals are definitely achievable; whether it will be significantly faster than the baseline is promising but not guaranteed.

**Strengths:**

- Clear goal explicitly stated throughout the proposal.
- Strong scientific significance of the proposed goals.
- Background research and baseline tools were clearly established.

**Weaknesses:**

- Not obvious what the novel visualization contribution will be.
- Somewhat high risk, as there is a chance that the implementation will not be significantly faster than the baseline.

**Other Comments for Discussion:**

None.

---

### **1.5.1 Response**

We agree with the review that the main weakness is in the visualisation novelty. To improve this we aim to incorporate volumetric visualisation, so that it is possible a 3d slice of parameter space. In addition we will add contour trees which will help give a different sense of the smoothness of the landscape.

---

## 2 Aims

The aim of this project is to develop a visualisation tool for neural network (NN) loss landscapes. The **scientific contribution** is to develop methods to significantly speed up the rate at which these visualisations can be generated such that they can be useful during training and to find the most informative slices. The **visual contribution** is to develop a novel method to dynamically change the vectors along which the visualisation is generated for a more dynamic view of the loss post-training and to visualize critical points in higher dimensions through Reeb graphs [16]. Our stretch goal is to apply our tool to physics informed networks to understand when they are ill-conditioned.

## 3 Significance

Despite the high-dimensionality and non-convexity of neural network loss landscapes, a problem difficult in theory, neural networks are still somehow able to find global minima in practice. Furthermore, some neural network architectures are easier to train than others. Why? Visualisations can help us understand how different architectures and features of the neural net affect the loss landscape. Furthermore, the ability to visualise the loss landscape during training can give the practitioner insights into the training procedure. At the moment, the only feedback available to the practitioner is often the loss value and train/test accuracy. A visual tool would be able to provide much detailed and intuitive feedback that can be used for debugging, early stopping and model tuning. Our hope is that by the end of the semester, this project will be a useful diagnostic tool for anyone training a neural network. In the future, understanding the loss landscape could lead to geometric insights that could help bridge theory and practice. In particular, we aim to bridge the gap between deep learning and geometry. However, given the ubiquity of neural networks in many different scientific fields, we believe this work extends into other domains like graphics, language, and neuroscience.

## 4 Background

Visualising loss landscapes can lead to genuine insight into neural network training as demonstrated by the following works: [4] use affine splines to analytically characterise the local properties of the deep network loss landscape and quantitatively compare different deep network architectures. Back in the 1990's [11] conjectured that flatness of the minima of a loss function found by a NN results in good generalisation. This claim was investigated by [9] who conclude that more geometric properties of the loss landscape should be taken into account to characterise flatness. Crucially, these analyses are all done post-training, our hope is that by providing online visual feedback, we can make these types of mathematical insights into neural networks easier to discover.

## 5 Visualizing Loss Landscapes

This work is primarily inspired by [13] who proposed visualising neural network with linear interpolation: Given two scalars  $\alpha, \beta$ , neural network with parameters  $\theta$  and loss function  $L(\theta)$  choose two direction vectors  $\delta, \eta$  and plot function

$$f(\alpha, \eta) = L(\theta + \alpha\delta + \beta\eta) \tag{1}$$

Which creates a 2D surface of parameter space. In particular, [13] chose  $\delta$  and  $\eta$  which are at the appropriate scale by applying filter normalisation. [2] is a blog by Javier Ideami who renders loss landscapes in extremely high fidelity. These renderings often take days and blend art and science and represent very much an upper bound of what is possible.

## 5.1 Is a 2D Slice enough?

Many convergence proofs often rely on the *local convergence assumption*. For example, adam [12], the most popular optimization algorithm in deep learning is known to locally converge under certain conditions [6]. However, if these conditions are not met, or a different optimizer is used, we may be making a false assumption. By visualizing the loss landscape, if the practitioner is able to find any slice that where local convergence does not happen, they are able to disprove the assumption. Similarly, many optimization regimes often assume well conditioned Hessians (second order derivative of the parameters). If the loss a slice of the loss landscape reveals a valley shape or saddle point instead of a bowl shape [10], the practitioner has visual evidence of ill-conditioning. This appears to be a particular problem with physics informed neural networks (PINNS) [7].

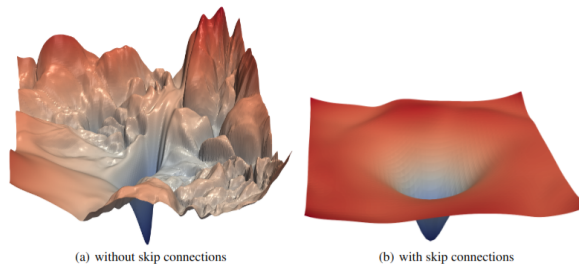


Figure 1: Loss landscape visualisation from [13].

## 6 Going beyond 2D

Nonetheless, going beyond 2D slices does not appear to be done in other works. We propose a novel visual contribution by incorporating Reeb graphs [16]. It should be possible to take a slice of arbitrary dimension and find critical points by sampling the gradient at those points. If gradient is 0, we can add it to the tree. This would be a genuinely novel method for going beyond a 2D slice for neural networks.

## 7 Contributions

### 7.1 Scientific

The main scientific contributions will be the following

- Apply adaptive sampling algorithms to find the most interesting parts of the loss landscape.
- Investigate other dimension reduction algorithms beyond PCA [8] to find the most interesting 2D slices.

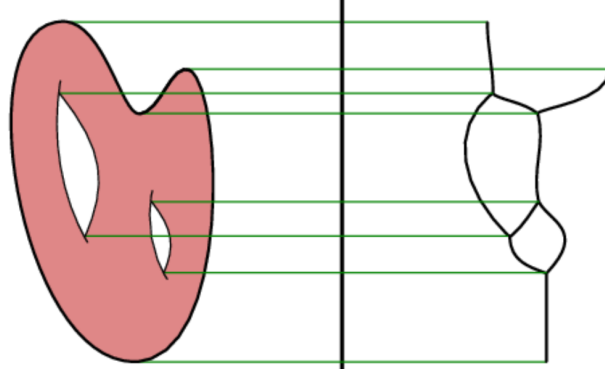


Figure 2: Example of a Reeb graph [5].

- Apply our tool to physics informed networks (stretch goal) to understand conditioning.

Where we define *interesting* as areas that have unstable gradients.

## 7.2 Visual

- Interpolate between slices to gain a more complete picture of the landscape
- Use Reeb graphs to visualize critical points beyond 2D.

## 7.3 Evaluation

### 7.3.1 Quantitative

Gradviz [8] is the most similar tool to ours currently available. Our aim to build a tool which also takes as input a Pytorch model and produces loss landscape visualisations. In this case, GradViz is used as a post-training tool. Our hope to to use adaptive sampling and engineering speedups to allow us to generate similar landscapes during training. Gradviz would be our main benchmark of comparison. [15] solve the opposite problem, — rather than trying to visualise a landscape, they try to shape the landscape to some image. If time permits, we hope to apply this technique as a sense-check to confirm that our tool is working as intended. Finally [14] present Adaptive, a method of sweeping through parameter space and sampling the most interesting parts. We aim to use this algorithm along with just-in-time compilation to make our tool fast enough to be used during training.

### 7.3.2 Qualitative

Our qualitative evaluation will involve a pilot user study of a up to five learning experts. The goal will be to use the tool to design a better neural network architecture to solve a toy problem like MNIST [3]. The aim will be to use the loss. The pilot will see if the users can improve their score on the test set only using the visual tool to inform their neural network architecture.

## 8 Research Plan

Week	Task Description	Deliverables/Outcomes
Week 1	Re-implement [13] as a baseline	Working prototype
Week 2	Integrate Adaptive [14]	Speed benchmarks against [8, 13]
Week 3	Speed improvements	Improved benchmark results
Week 4	Implement change of basis vectors feature	Improved Benchmarks and new feature
Week 5	User study and PINNS	-
Week 6	Final review and project closure	Final report

Table 1: 6-Week Project Plan

## References

- [1] Just-in-time compilation — JAX documentation. <https://jax.readthedocs.io/en/latest/jit-compilation.html>.
- [2] Loss Landscape | A.I deep learning explorations of morphology & dynamics. <https://losslandscape.com/>.
- [3] MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges. <https://yann.lecun.com/exdb/mnist/>.
- [4] R. Balestriero, A. I. Humayun, and R. Baraniuk. On the Geometry of Deep Learning, Aug. 2024.
- [5] U. Bauer, E. Munch, and Y. Wang. Strong Equivalence of the Interleaving and Functional Distortion Metrics for Reeb Graphs. June 2015.
- [6] S. Bock and M. Weiß. A proof of local convergence for the adam optimizer. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019.
- [7] W. Cao and W. Zhang. An analysis and solution of ill-conditioning in physics-informed neural networks, May 2024.
- [8] A. Chatzimichailidis, J. Keuper, F.-J. Pfreundt, and N. R. Gauger. GradVis: Visualization and Second Order Analysis of Optimization Surfaces during the Training of Deep Neural Networks. In *2019 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC)*, pages 66–74, Nov. 2019.
- [9] L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp Minima Can Generalize For Deep Nets, May 2017.
- [10] J. Gill and G. King. What to do when your hessian is not invertible: Alternatives to model respecification in nonlinear estimation. *Sociological methods & research*, 33(1):54–87, 2004.
- [11] S. Hochreiter and J. Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.
- [12] D. P. Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

- [13] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein. Visualizing the Loss Landscape of Neural Nets, Nov. 2018.
- [14] B. Nijholt, J. Weston, J. Hoofwijk, and A. Akhmerov. *Adaptive*: parallel active learning of mathematical functions, 2019.
- [15] I. Skorokhodov and M. Burtsev. Loss Landscape Sightseeing with Multi-Point Optimization, Oct. 2019.
- [16] J. Tierny. *Topological data analysis for scientific visualization*, volume 3. Springer, 2017.

## 9 Collaborator agreement

visualizing loss landscapes of neural nets inbox x

✕ 🖨 🔗

**Prakash, Arjun** <arjun\_prakash@brown.edu>  
 to Randall ▾

Thu, Sep 12, 1:27 PM ☆ ↶ ⋮

Hi Randall,

We met briefly just before the semester started, in office 549.

My name is Arjun, I am a 3rd year PhD supervised by Amy Greenwald and Nora Ayainian.

I am taking a class on interdisciplinary scientific visualization and I would like to visualize the loss landscapes of neural nets using techniques similar to this paper  
 paper: <https://arxiv.org/pdf/1712.09913>

I need an outside collaborator for the class and I was wondering if you would be interested.

It should be a fairly low touch on your end. It mainly satisfies the interdisciplinary collaboration aspect of the class. I remember from your job talk that you had many cool visualizations, so I thought I would ask.

the class: <https://cs.brown.edu/courses/csci2370/2024/ideas.html>

Cheers,  
 Arjun

**Randall Balestriero**  
 to me ▾

Thu, Sep 12, 2:38 PM ☆ ↶ ⋮

Hi Arjun

Thank you for ping me back! Yes I do remember. I would be very happy to be the outside collaborator and help. I liked a lot that paper from Tom and in fact we did a more theoretical follow up: <https://link.springer.com/article/10.1007/s00365-022-09601-5>




Would be very curious to know which specific application you would have in mind!

Best,  
Randall

\*\*\*

# Arjun Prakash



## Computer Science PhD Student (AI/ML)

 arjun-prakash.github.io  a-prakash  arjunp@brown.edu

### Education

Aug 2022 – present	<b>Brown University, PhD, Computer Science (AI/ML)</b> <ul style="list-style-type: none"><li>Researching multi-agent reinforcement learning and its applications to collaborative robotics.</li><li>Supervised by Amy Greenwald and Nora Ayanian.</li></ul>	Providence, USA
2015 – 2019	<b>University of Sydney,</b> <i>Double Bachelor Degree, Computer Science &amp; Business Analytics</i> <ul style="list-style-type: none"><li>First Class Honours</li></ul>	Sydney, Australia

### Research

2024	<b>STA-RLHF: Stackelberg Aligned Reinforcement Learning with Human Feedback,</b> Jacob Makar-Limanov*, <b>Arjun Prakash*</b> , Denizalp Goktas, Nora Ayanian, Amy Greenwald, <i>Coordination &amp; Cooperation for Multi-Agent Reinforcement Learning Methods Workshop (spotlight)</i>
2023	<b>Convex-Concave Zero-Sum Stochastic Stackelberg Games,</b> D Goktas, <b>A Prakash</b> , A Greenwald, <i>Neural Information Processing Systems (NeurIPS)</i> 
2022	<b>Structural Clustering of Volatility Regimes for Dynamic Trading Strategies,</b> <b>A Prakash</b> , N James, M Menzies, G Francis, <i>Journal of Applied Mathematical Finance</i> 

### Awards

2023 – 2024	<b>Quad Fellowship, Schmidt Futures</b> <ul style="list-style-type: none"><li>One of 100 candidates chosen from Australia, India, Japan, USA for STEM research.</li></ul>
2022 – 2023	<b>Paris Kanellakis Graduate Fellowship, Brown University</b>
2020	<b>Dean's List of Excellence in Academic Performance, University of Sydney</b> <ul style="list-style-type: none"><li>Awarded to only two students in the program for a high distinction average across honours research and master's level classes.</li></ul>
2017 – 2019	<b>Taylor Scholarship, St Andrew's College</b> <ul style="list-style-type: none"><li>Awarded twice for contributing to College life with Stan Droid the chatbot.</li></ul>
2019	<b>Microsoft Asia Senior Research Prize, University of Sydney</b> <ul style="list-style-type: none"><li>Awarded for the top senior capstone project, Scopus Miner.</li></ul>

### Professional Experience

Jan 2021 – Jun 2022	<b>Faethm AI, AI Engineer</b> <ul style="list-style-type: none"><li>Created key IP for the successful acquisition by Pearson PLC.</li><li>Led the R&amp;D of NLP models to extract skills from job ads using named entity recognition, Siamese networks and transformers.</li><li>Created node embeddings using graph neural networks to examine the relationships of jobs, skills and career transitions.</li><li>Improved the efficiency of the skill extraction pipeline by 15x in one quarter through vectorization, caching, database redesign and distributing tasks with AWS SQS.</li></ul>	Sydney, Australia
Mar 2020 – Dec 2020	<b>Deloitte Consulting (Analytics and Cognitive), Graduate Consultant</b> <ul style="list-style-type: none"><li>Analysed the use of restrictive language in NSW legislation and generated insights featured in the RegExpore whitepaper.</li><li>Prototyped and prioritised data strategies for NSW Government's COVID-19 response.</li></ul>	Sydney, Australia

Dec 2019 – Feb 2020

**Mitsubishi Electric, Advanced Gakushū (Learning) Group,**

Fujisawa, Japan

*Research Data Science Intern* [↗](#)

- Reviewed literature on geo-trajectory data and built a two-stage classification/regression model to predict when a user returns home for IoT devices.
- Features, insights and recommendations were used in production by the Group.
- Sponsored by the Asia New Zealand Foundation.

Dec 2017 – Dec 2018

**Fund3, Data Science Intern**

Santa Monica, USA

- Implemented complex kernels and support vector machines to forecast volatility of the top cryptocurrencies.
- Managed a team of students to deliver an automatic messaging system that notified users of unusual activity on the blockchain.

## Projects

---

Apr 2021 – Apr 2022

**PZ-Dilemma, Game Theoretic MARL Environments** [↗](#)

- Implemented collective risk dilemmas, centipede game and classic prisoners dilemma for Multi-Agent Reinforcement Learning (MARL).
- Official 3rd party contributor to the open-source Petting Zoo MARL project.

May 2017 – Jan 2020

**Stan Droid, St Andrew's College Chatbot**

- Developed a customer service chatbot for students and admin staff of St Andrew's College, University of Sydney, with continuous integration and unit testing.
- 90% uptake and over 100k messages per year.

Jul 2018 – Nov 2018

**Scopus Miner (Capstone Project), Australian Defense Force Academy** [↗](#)

- Engineered an NLP web app that extracts rich data from academic literature and presents insights in a visual format.
- Connected papers by linking citations and references in a graph database.

## Teaching

---

Jul 2023

**Teaching Associate, Brown University**

- Ethics of Artificial Intelligence and AI Research for the Brown Pre-College Program

## Organizations

---

2018 – present

**Asia New Zealand Foundation, Leadership Network Member**

- Helping strengthen ties between Aotearoa and Asia.

# Visualizing the unfolding high-dimensional residual stream of a transformer neural network model in response to linguistic inputs

Aalok Sathe

PI

aalok\_sathe@brown.edu

Sam Musker

Co-PI

samuel\_musker@brown.edu

October 14, 2024

## Abstract

We describe a novel approach to visualizing the time-varying *unfolding* of representations in the *residual stream* of a transformer. The method allows a joint viewing of the layer-by-layer dynamics of the *residual stream* rather than attention heads, and plots a *trajectory* in an independently-defined reduced dimension space enabling cross-stimulus comparisons.

October 10, 2024: response to reviewers

We thank the reviewers for their time spent reviewing this proposal.

=====

Reviewer: apraka15

=====

Overall: 6.  
ack

Interdisciplinary: 6. The interdisciplinary aspect was not really discussed, apart from a brief mention of cognitive science.  
we will discuss the emerging sub-field of AI, and discuss why it merits studying on its own

Scientific: 2. Adding to the interpretability of transformers would be very helpful to the deep learning community.  
ack

Visualization: 7. The proposal didn't mention any concrete plans for visualizing the residual stream. It seems the first 2 weeks of the project are dedicated to figuring out which tools and methods should be used which doesn't leave much time for the rest of the project. Some pictures of the related work would help.  
Agreed; making the plan more concrete would be a help. The final proposal will include more concrete methods.

Significant: 6. It's unclear what the specific contribution is which makes it hard to evaluate.  
We will clarify the specific contribution in terms of what will be visualized and how.

Novel: 5. It's unclear what the specific contribution is which makes it hard to evaluate.  
See above.

Goals clearly stated:  
I think the goals are quite vague beyond visualizing the residual stream. There was good discussion on the mechanistic interpretability literature, but less on specific visualizations.  
We will include more specific details on the nature of the visualization.

Likelihood of Success: 4/10

Strengths:  
Interesting domain  
Good understanding of existing techniques

Weaknesses:  
unclear what the visualization might look like  
timeline seems ambitious.

Other comments for discussion:  
Some reference visualizations would be helpful.  
How will this be evaluated?  
reference visualizations not available at the time of proposal.  
evaluation will be based on

=====

Reviewer: bbutaney

=====

Score the proposal on the following criteria. For each criteria, provide an NIH-style score from 1 to 9. Refer to the descriptions of scores in the shared google doc "board." Briefly justify your scores under each.

Overall: 3

Interdisciplinary: 2; proposal is substantially interdisciplinary, with relations to cognitive science and neuroscience  
ack

Scientific: 3; While the questions about transformer interpretability are grounded and informed  
(based on my knowledge of the literature surrounding transformers, at least), this could be explained better in the proposal  
we will provide more background on the motivation behind transformer interpretability

Visualization: 3; Strong idea, but vague explanation of how they will actually create interpretable visualizations  
agreed; in the final proposal we will be more concrete about how the visualization will be created  
and why it would be interpretable

Significant: 2; transformers have been highly relevant in NLP and this proposal demonstrates high potential impact on the field of cognitive science.  
ack (see above)

Novel: 3; Idea is innovative with a focus on residual stream and alternative ways to improve interpretability, but this could be elaborated on more  
ack (see above)

Goals clearly stated: 2. Goals clear, but the final outcome desired could be clarified a bit more  
ack (see above)

Likelihood of Success: 3; while this seems promising, dealing with high dimensional data can be  
tricky and interpretability is always hard to evaluate.  
this comment is generally of the nature 'this may be hard'. we take this to interpret it is

hard but  
within the realm of being possible.

Strengths: Very focused goals for residual stream which makes scope feel less daunting. Merging transformers with cognitive science is interesting and could yield very impactful results

Weaknesses: Vague in aspects related to implementation, which decreases my confidence in the proposal as a whole. Also vague in regards to evaluation metrics. agreed; we will aim to be more concrete in terms of specific goals in the final proposal.

Other comments for discussion: What are some specific methods that could be used for visualization? How would they differ from standard methods like PCA?  
we will likely use PCA.

---

Reviewer: bleahey

---

Score the proposal on the following criteria. For each criteria, provide an NIH-style score from 1 to 9. Refer to the descriptions of scores in the shared google doc "board." Briefly justify your scores under each.

Overall: 5;

Interdisciplinary: 5; Relates to large language models. Implies significance in cognitive science but does not elaborate.  
we will connect to the background as it relates to mechanistic interpretability and the implications of studying representations of language in cognitive science.

Scientific: 2; Transformers are extremely novel and significant—improving visualization represents a strong contribution.  
ack

Visualization: 7; presentation demonstrates traditional visualization techniques and a strong understanding of the literature,  
but there is no suggested methodology for the proposed visualization of the residual stream.  
agreed; in the final proposal we will be more concrete about how the visualization will be created  
and why it would be interpretable

Significant: 6; Significance is implied and not elaborated upon.  
we make some connections but will make more connection with literature explaining the necessity of interpreting  
the workings of artificial neural network models.

Novel: 5; It's unclear what the specific contribution is which makes it hard to evaluate.  
see the two responses above

Goals clearly stated:  
I think the goals are quite vague beyond visualizing the residual stream.  
There was good discussion on the mechanistic interpretability literature, but less on specific visualizations.  
we will aim to be more specific in terms of contribution in the final proposal

Likelihood of Success: 7; Proposal is somewhat vague and only allots two weeks for implementing computational methods which strikes me as ambitious.  
we will readjust the proposed time allocation to spend 3-4 weeks on the computational aspects.

Strengths:

Strong understanding of previous work on transformers  
Great potential for novel machine learning contribution

Weaknesses:

Lack of specificity in goals  
Lack of clarity on the specific contribution  
ack; addressed above

Other comments for discussion:

=====

Reviewer: Yang Xiang

=====

Score the proposal on the following criteria. For each criteria, provide an NIH-style score from 1 to 9. Refer to the descriptions of scores in the shared google doc "board." Briefly justify your scores under each.

Overall: 3

Interdisciplinary: 3 Generating visualization for certain transformer block interpretability  
is a typical combination of visualization and ai research.  
(by the way, 3 of the 4 proposals I review have similar topics like this...)  
ack

Scientific: 3 The project seems to try getting new insight for transformer model with visualization, which may help ai researchers build better transformer structure.  
ack

Visualization: 4 The author did not provide too much detail about visualization, yet did a great survey with many related papers.  
agreed; we will provide specifics of how we plan to visualize the residual stream of transformers

Significant: 3 Generating new insight from visualization of a single block is hard.  
However the tool may have great discoveries if the author finds a novel way for visualization.  
ack

Novel: 3 While there are tons of tools doing the similar topic, this project focuses on a small and new direction. Transformer is relatively new and studying a certain block inside it is a quite novel direction.  
ack

Goals clearly stated: 3 There is a brief plan. Maybe more detail is needed for each 2 weeks.  
agreed. more detail will be provided in the final proposal.

Likelihood of Success: 3 The goals need to be clearer. However, it is a really good practice to do enough survey before actually starting coding.

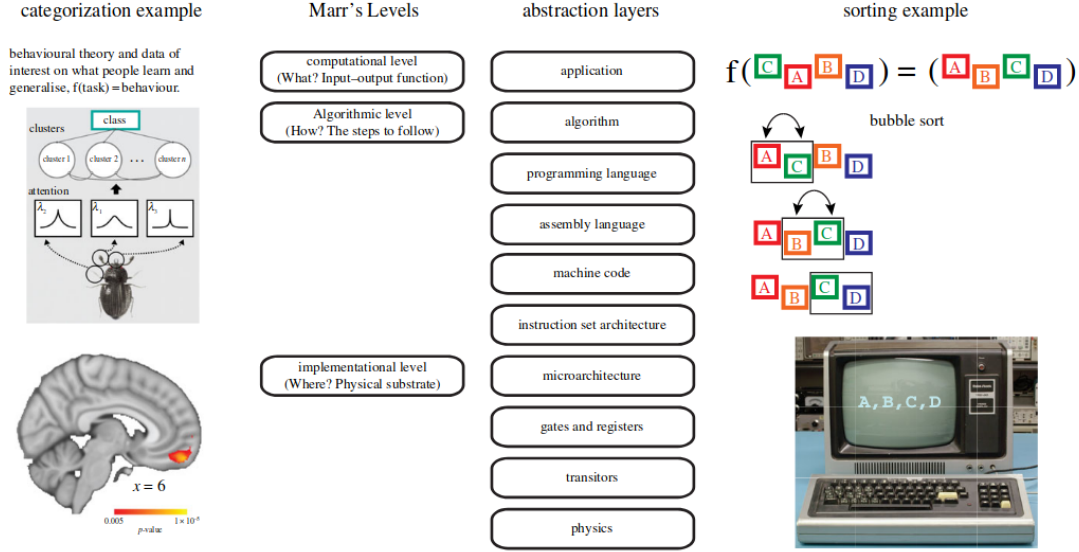
Strengths:

1. do a lot of survey
2. choose a good sub direction

Weaknesses:

1. goals a bit unclear
  2. hard to produce significant insight with the tool
1. goals will be made clearer
  2. a project so small in scope is unlikely to provide significant insight. we will make the limitations clear in our proposal.

Other comments for discussion:



**Figure 1.** Marr’s levels compared to abstraction layers in computing with examples of each. Marr’s levels are clearly influenced by abstraction layers in computer science, though Marr’s levels are less fine grain, particularly for levels of interest to many neuroscientists. On the left, an example from category learning is shown in which an algorithmic model [5] was fit to behaviour and its internal representations are used to interpret BOLD response [6]. On the right, a sorting algorithm addressed the computational level problem of sorting and was implemented by a digital computer. The abstraction layers in computing make clear that moving to a lower layer introduces additional detail (more information) about the computation whereas higher layers introduce abstract constructs that can be realized in multiple ways. (Online version in colour.)

Figure 1: Levels of abstraction relate to levels of understanding. Whereas the implementational details of artificial neural network models are known, and some the computational goals pre-defined in terms of task constraints, their algorithmic workings are unclear. The goal of the current work would be to enable visualization that may provide insights into the algorithms supporting an artificial neural network’s computational goals. (Figure taken from ‘Levels of biological plausibility’, Love (2021)).

## 1 Aims

Transformers (Vaswani & et al, 2017) form the basis of modern language models, including large language models (LLMs). The parameters of a transformer can be fully inspected, and the dynamics of any given stimulus fully and deterministically computed: something that would be a dream come true for neuroscience if this were true of the human brain. Despite their fully accessible structure, their complexity is unlike any neural network model before them, which eludes their understanding at a level of abstraction above their raw weights and mathematical operations. The lowest-level explanation of a system of interest is a full description of the system itself, and provides little to no insight into the workings of it (Figure 1).

In this work, we propose to visualize one component of the transformer model, namely, its *residual stream*. The residual stream forms the basis of transformations to an input *token*, and tracks how it evolves over time up until its output, in the form of next-token prediction.

Traditional visualizations of transformers focus on attention as the primary object of interest (Figure 2). However, despite its allure towards interpretability, the role of attention in shaping representations in an interpretable way is questioned. For instance, Jain & Wallace (2019) note that there is no necessary connection between attention patterns and posited mechanisms that might be the cause of a behavior, and in fact, attention does not correlated with gradient-based feature-importance maps derived directly from an explicit

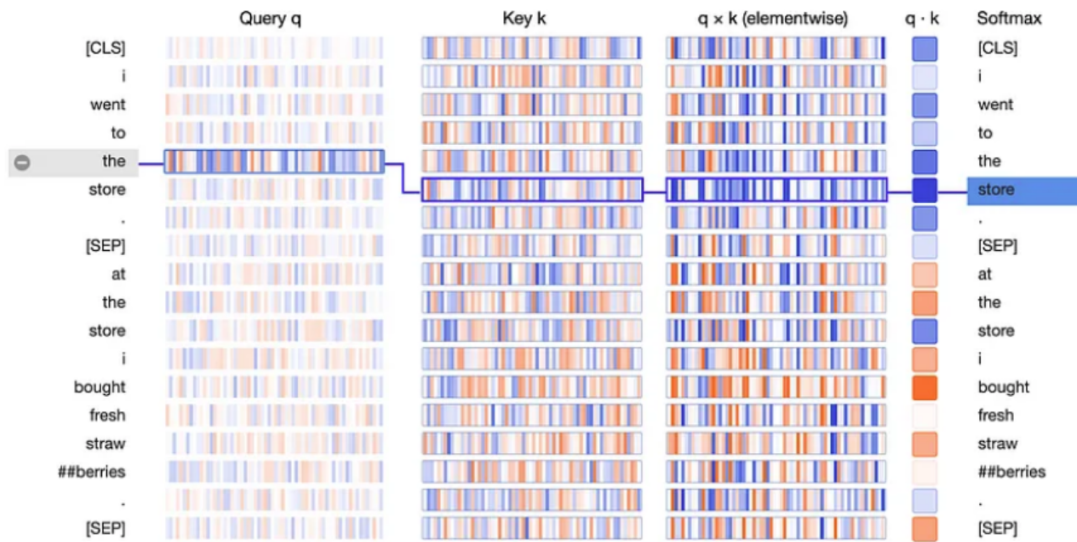


Figure 2: Traditional visualizations of transformer attention (Vig, 2019).

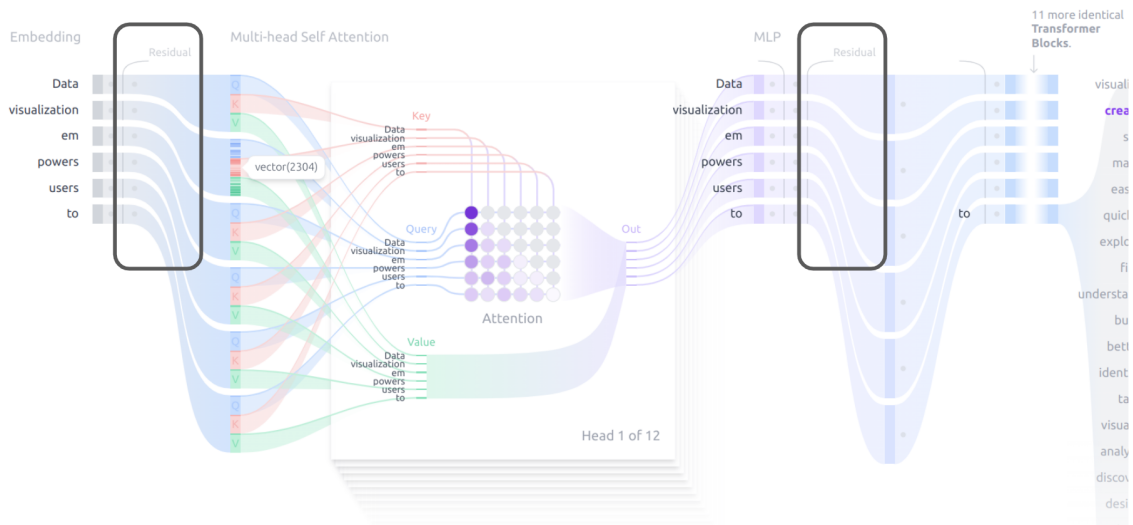


Figure 3: Architecture of a transformer model  
(<https://poloclub.github.io/transformer-explainer>).

error signal.

Instead, we look to the residual stream as an object of interest, and study how the representations of inputs may be evolving over time. It is widely known that neurons in biological and artificial systems engage in multiplexing and construct distributed representations that are more expressive, efficient, and robust to noise (Fusi et al., 2016). Despite this distributedness, there is nevertheless rich structure in how they represent their inputs. Neuroscientists have a long history of studying neural populations of interest and attempting to find structured representations in their activity Khona & Fiete (2022); Chung & Abbott (2021). Such models of neural activity allow an interpretable explanation at an algorithmic level to be described: for instance, keeping track of the position of one’s head is ecologically relevant, as it can help make sense of visual sensory inputs and build a model of the surrounding world. What might be the best implementation for such a circuit? Neuroscientists recording neural activity from the fly brain found evidence of a ring-attractor network where a point in the ring-like activity-space represents a physical instantiation of the head position Vafidis et al. (2022).

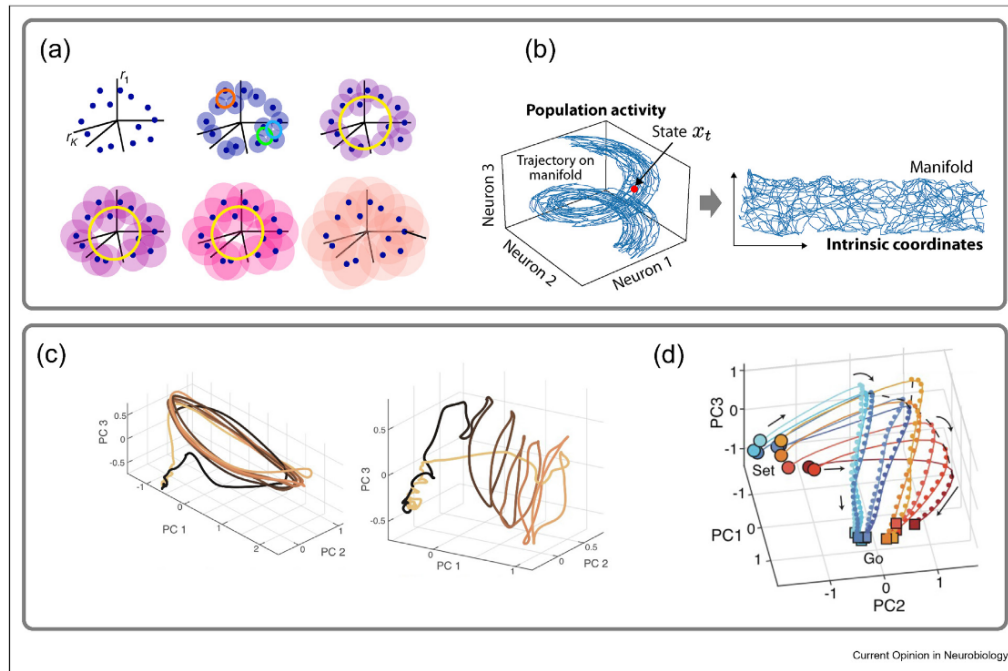
Similarly, despite the vastly high-dimensional nature of representations in an artificial neural network, there may nevertheless exist structure that points to algorithmic-level explanations of its behavior (Chung & Abbott, 2021; Langdon et al., 2023). We propose to look at the *residual stream* (Figure 3) in response to linguistic inputs in the form of sentences or phrases. Since the residual stream is high-dimensional, we plan to use a dimensionality reduction method, such as PCA, to reduce it to 3-dimensional space. Compared to traditional, static, means of visualizing representations at a particular layer, we will use the same PC-space *across* layers, allowing for a longitudinal geometric view into the evolving activations of the residual stream. In order to independently construct such a space and not be biased by the data at hand (to avoid double-dipping), we will come up with two *different* datasets,  $D_1$  and  $D_2$ , one of which will be used strictly towards learning the general properties of variance in this space, and to construct a transformation into a low-dimensional space. The second dataset,  $D_2$ , will be used towards actually constructing visualizations in this space, assuming sufficient variance overlap with  $D_1$  so that the PC-transform is still faithful to the representational properties of the transformer with respect to  $D_2$ .

To evaluate the success of our visualization methods, we will compare them with literature in computational and systems neuroscience on gleaned traces of behavior and mechanisms from neural activity (Figure 4) and qualitatively evaluate whether our method is able to produce any such early signatures of explainability. Furthermore, we will evaluate the usefulness of the current method in relation to traditional visualizations of attention in transformer models (Figure 2).

## 2 Significance

The success of transformers in modeling real-world language use, which had so far eluded artificial models, makes them a compelling target of interpretability work, with potential upshots to both, the field of machine learning as well as cognitive neuroscience. For the joint purposes of understanding these models better, and consequently, understanding the implications for cognitive neuroscience, it is necessary to develop new methods to aid their understanding (Cao & Yamins, 2021a).

Artificial neural network models arose from neuroscience-inspired model-building. Nevertheless, their implementational details have vastly diverged from our understanding of biological neural networks. However, ANNs have also demonstrated, for the first time in history, proficiency in processing language in a human-like way.



**(a–b)** Manifold discovery methods. **(a)** Spine Parameterization for Unsupervised Decoding (SPUD). **(b)** Manifold Inference for Neural Dynamics (MIND) **(c–d)** Population dynamics as cognition. **(c)** (Left) Temporal trajectories during macaque cycling task in M1 and (Right) SMA. **(d)** Dorsomedial Frontal Cortex (DMFC) response profiles during Bayesian computation. Part (a) adapted from Ref. [46]. Part (b) adapted from Ref. [10]. Part (c) adapted from Ref. [51]. Part (d) adapted from Ref. [8].

Figure 4: Chung & Abbott (2021) provide examples of manifolds as an explanatory tool in behavior and brain sciences.

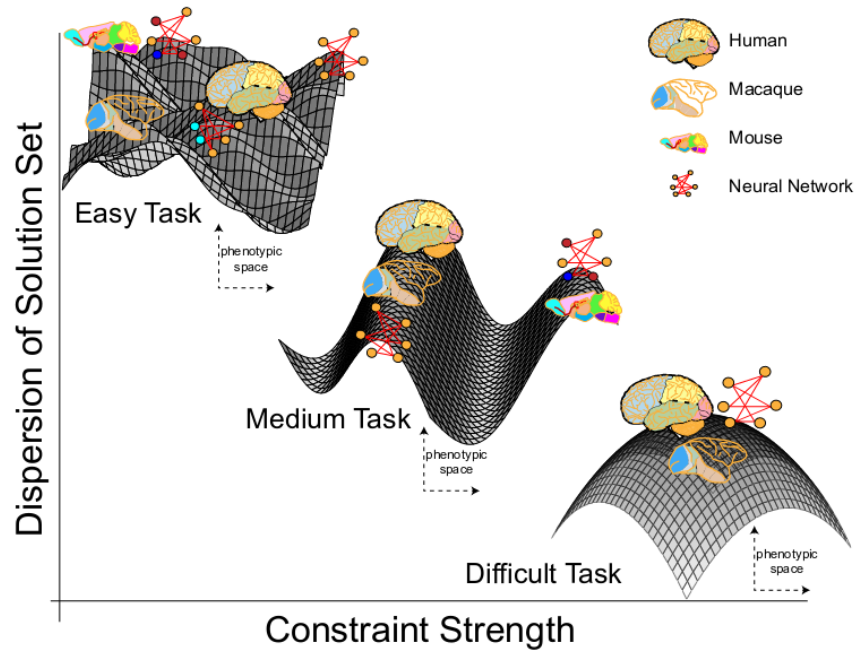


Figure 5: Why is studying the workings of a neural network model important? In Cao & Yamins (2021b), the authors propose the effect of increasing constraint on the solutions intelligent systems come up with to solve the constrained task. Given a complex, high-level task such as language processing, it is possible that two systems of vastly different implementations nevertheless converge on their algorithmic solutions owing to similarity in computational goals: prediction. For this reason, understanding neural network language models has the potential to inform theory in cognitive neuroscience. (Figure taken from Cao & Yamins 2021b).

### 3 Related Work

Molina (2023), use a geometric approach to understanding a specific operation within the transformer architecture, namely, normalizing. Our approach would use similar methods to look at the residual stream rather than the effect of a particular operation like normalization.

Chung & Abbott (2021) Neural population geometry: An approach for understanding biological and artificial neural networks. Authors provide a compelling call to action for using geometric methods to understand the complexities of biological and artificial neural systems. Relationship to your project: Provides a comprehensive review of methods so far and a philosophical perspective motivating such work.

In another work (Sanford et al. (2024)), authors review key properties of representations of inputs in transformers. In our project, this will be relevant in deciding what properties would be most informative to visualize, and to know any limitations of the attempted visualization beforehand.

The authors of Valeriani et al. (2024) study the geometry of hidden representations of large transformer models. Authors consider factors such as 'intrinsic dimensionality', a property of the representation space, and study its unfolding over time (over layers). This is a useful example of studying a particular property rather than all of the representation space in raw form, and then visualizing its unfolding. Furthermore, authors highlight similarities across two very different tasks: protein modeling and image modeling. Similarly, authors Hosseini & Fedorenko (2024); Hénaff et al. (2019) study another key property, 'curvature', of the representation space. In our work, we would be able to use their method to consider displaying some form of aggregate or epiphenomenal property to visualize along with the visualization of the representations of data.

A method we won't be able to get to in our work, but that provides inspiration for an alternative visualization framework, is a developmental framing of the same question: rather than ask how representations evolve over the time-course of processing, the authors Saxe et al. (2019) ask the same question over *development*, or training.

Merullo et al. (2024) Uses singular value decomposition (SVD) on neural network weights as a way of predicting how representations might evolve over time. Relationship to your project: There could be alternate ways of visualizing the dynamics, without even using any input data, but by doing network analysis on its own. Our work may use some methods from this work.

Piantadosi et al. (2024) Posits mechanisms behind a harmony in observed properties of representations so far and our understanding of concepts in the human mind. Posits that despite the mechanistic nature of artificial neural networks, they may have rich conceptual spaces. Provides possible ways this might be realized. Relationship to your project: Useful inspiration for what ways of looking at high-dimensional representation spaces can help inform theory in allied fields such as cognitive science.

### 4 Research Plan

**Weeks 1-2** Narrowing down of what set of stimuli should be used as an example use case of the proposed tool. Constructing a PC-space. Constructing two datasets,  $D_1$  and  $D_2$  as materials supporting visualization.

**Weeks 3-4** Development of tool. Fine-tuning of parameters. Construction of visual imagery.

**Weeks 5-6** Testing, evaluation, comparison with prior methods, and writing of report.

## References

- Cao, R., & Yamins, D. (2021a, April). Explanatory models in neuroscience: Part 1 – taking mechanistic abstraction seriously. (arXiv:2104.01490). Retrieved from <http://arxiv.org/abs/2104.01490>
- Cao, R., & Yamins, D. (2021b, April). *Explanatory models in neuroscience: Part 2 – constraint-based intelligibility* (No. arXiv:2104.01489). Retrieved 2023-04-10, from <http://arxiv.org/abs/2104.01489>
- Chung, S., & Abbott, L. F. (2021). Neural population geometry: An approach for understanding biological and artificial neural networks. *Current opinion in neurobiology*, 70, 137–144.
- Fusi, S., Miller, E. K., & Rigotti, M. (2016, April). Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, 37, 66–74. Retrieved 2023-04-11, from <https://linkinghub.elsevier.com/retrieve/pii/S0959438816000118> doi: 10.1016/j.conb.2016.01.010
- Hénaff, O. J., Goris, R. L., & Simoncelli, E. P. (2019). Perceptual straightening of natural videos. *Nature neuroscience*, 22(6), 984–991.
- Hosseini, E., & Fedorenko, E. (2024). Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language. *Advances in Neural Information Processing Systems*, 36.
- Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 3543–3556).
- Khona, M., & Fiete, I. R. (2022). Attractor and integrator networks in the brain. *Nature Reviews Neuroscience*, 23(12), 744–766.
- Langdon, C., Genkin, M., & Engel, T. A. (2023, April). A unifying perspective on neural manifolds and circuits for cognition. *Nature Reviews Neuroscience*, 1–15. Retrieved 2023-04-14, from <https://www.nature.com/articles/s41583-023-00693-x> doi: 10.1038/s41583-023-00693-x
- Love, B. C. (2021, January). Levels of biological plausibility. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1815), 20190632. doi: 10.1098/rstb.2019.0632
- Merullo, J., Eickhoff, C., & Pavlick, E. (2024). Talking heads: Understanding inter-layer communication in transformer language models. *arXiv preprint arXiv:2406.09519*.
- Molina, R. (2023). Traveling words: A geometric interpretation of transformers. *arXiv preprint arXiv:2309.07315*.
- Piantadosi, S. T., Muller, D. C., Rule, J. S., Kaushik, K., Gorenstein, M., Leib, E. R., & Sanford, E. (2024). Why concepts are (probably) vectors. *Trends in Cognitive Sciences*.
- Sanford, C., Hsu, D. J., & Telgarsky, M. (2024). Representational strengths and limitations of transformers. *Advances in Neural Information Processing Systems*, 36.

- Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23), 11537–11546.
- Vafidis, P., Oswald, D., D’Albis, T., & Kempter, R. (2022). Learning accurate path integration in ring attractor models of the head direction system. *Elife*, 11, e69841.
- Valeriani, L., Doimo, D., Cuturello, F., Laio, A., Ansuini, A., & Cazzaniga, A. (2024). The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36.
- Vaswani, A., & et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.
- Vig, J. (2019). Visualizing attention in transformer-based language representation models. *arXiv preprint arXiv:1904.02679*.

# Aalok Sathe

address: CIT, Brown University, Providence, RI 02912

email: [asathe\[at\]mit.edu](mailto:asathe[at]mit.edu), [aalok\[at\]brown.edu](mailto:aalok[at]brown.edu)web: [aalok-sathe.gitlab.io](https://aalok-sathe.gitlab.io)

## Education

2029 (exp.) (Aug '24–)	<b>Ph.D.</b>	<b>Brown University</b> (Providence, RI, USA) Computer Science	
May 2022 (Jan '22–)		<b>Massachusetts Institute of Technology</b> (Cambridge, MA, USA) Advanced Studies Program (ASP): Introduction to Neural Computation	
May 2021 (Aug '17–)	<b>B.S.</b>	<b>University of Richmond</b> (Richmond, VA, USA) <i>Majors:</i> Computer Science (hons.), Cognitive Science <i>Minors:</i> Linguistics, Math; Church-Kent prize: outstanding grad in CS	GPA: 4/4, rank 1/775
May 2020 (Jan '20–)		<b>University of Edinburgh</b> (Edinburgh, Scotland, UK) Informatics; Philosophy, Psychology & Language Science (PPLS)	

## Research Experience

- Graduate Researcher, **Brown University**: *Computer Science; Carney Institute for Brain Science* Providence, RI  
Mentors: Ellie Pavlick, Michael J. Frank Sep 2024 – present
- Research Associate, **Massachusetts Institute of Technology**: *Brain & Cognitive Sciences* Cambridge, MA  
Mentors: Evelina Fedorenko, Noga Zaslavsky, Greta Tuckute, Anna Ivanova, Cory Shain Jul 2021 – Jun 2024
- Research Intern, **Microsoft Research**: *NLP group* Bengaluru, India  
Mentors: Monojit Choudhury, Somak Aditya May 2020 – Aug 2020
- Undergrad Research Assistant, **University of Richmond**: *Math & Computer Science; Psychology* Richmond, VA  
Mentors (Math+CS): Taylor Arnold, Joonsuk Park, Prateek Bhakta, Heather Russell Aug 2018 – May 2021  
Mentors (Psychology): Cindy Bukach, Matthew Lowder Sep 2017 – Dec 2020

## Peer-Reviewed Publications

&co-first author, #alphabetically listed ([why?](#)), [Google Scholar profile](#)

- Language use is only sparsely compositional: The case of English adjective-noun phrases in humans and LLMs**  
**Sathe, A.**, Fedorenko, E., Zaslavsky, N.  
*Annual Meeting of the Cognitive Science Society*, 46 (CogSci 2024). [\[paper\]](#)
- Conventional and Frugal Methods of Estimating COVID-19-Related Excess Mortality and Undercount Factors**  
Dedhe, A., Chowkase, A., Gogate, N., Kshirsagar, M., Naphade, R., Naphade, A., Kulkarni, P., Naik, M., Dharm, A., Raste, S., **Sathe, A.**, Kulkarni, S., Bapat, V., Joshi, R., Deshmukh, K., Lele, S., Manke, K., Cantlon, J., Pandit, P.  
*Scientific Reports*, 14.1: 10378 (2024). [\[paper\]](#)
- Driving and suppressing the human language network using large language models**  
Tuckute, G., **Sathe, A.**, Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., Kay, K., Fedorenko, E.  
*Nature Human Behavior* (2024). [\[paper\]](#)
- Block symmetries in graph coloring reconfiguration systems**  
Bhakta, P.#, Krehbiel, S.#, Morris, R.#, Russell, H. M.#, **Sathe, A.#**, Su, W.#, Xin, M.#  
*Advances in Applied Mathematics*, 149: 102556 (2023). [\[paper\]](#)

## Samuel Musker

195 Waterman Street | Providence, RI | 267.303.9670 | [musker.sam@gmail.com](mailto:musker.sam@gmail.com) / [samuel\\_musker@brown.edu](mailto:samuel_musker@brown.edu)

### EDUCATION

---

#### Brown University, Providence, RI

Ph.D. student, August 2020 - Present

- Studying towards a Ph.D. in Computer Science and A.M. in Philosophy, GPA: **4.00** out of **4.00**

#### University of Pennsylvania, Philadelphia, PA

B.A., August 2017 - May 2019

- Major in Philosophy (Philosophy and Science concentration), minors in both Mathematics and Logic, Information, and Computation, overall GPA: **3.93** out of **4.00** / in-major GPA: **4.00** out of **4.00**
- Academic achievements: Phi Beta Kappa; Honors and Distinction in Philosophy major; Flower Prize for the best essay in Philosophy; Dean's List 2017-2018 and 2018-2019; Allan Gray Orbis Foundation Fellowship 2017-2018 and 2018-2019

#### University of Cape Town, Cape Town, South Africa

February 2015 – June 2017 (transferred, no degree)

- Registered for B.Sc. Eng. Electromechanical Engineering (2015-2016) / B.Sc. Philosophy and Applied Mathematics (2017), Weighted Average: **81.77%** out of **100%** (First-Class pass is 75%)
- Academic achievements: Dean's List 2015 and 2016; Isaac Ochberg and VC's scholarships 2015, 2016, and 2017; Allan Gray Orbis Foundation Entrepreneurial Innovation Fellowship 2015, 2016, and 2017; Top student Electromechanical Engineering program as at leaving

### EMPLOYMENT

---

#### Boston Consulting Group, Boston, MA | *Associate 1, promoted to Associate 2*

July 2019 – July 2020

- Consulted to a major US pharmacy chain, producing work for executive-level management at the fortune-20 company. Built sophisticated models and workflows in Excel and Alteryx with visualizations in Tableau
- Individually developed data processing and modelling tool to assist BCG North America in reaching reduction targets of GHG emissions from consultant flights. Collaborated with data analytics expert to develop a webapp frontend to interface with the tool. The tool assisted BCG C-suite in decision making, and having exceeded expectations was successfully exported to other global regions including Europe, Asia Pacific, and Australia/New Zealand as the gold-standard flight emissions modelling tool

#### Boston Consulting Group, Johannesburg, SA | *Visiting Associate*

June 2017 - July 2017

- Responsible for drafting infrastructure portion of the company's public report on economic development in East Africa

#### UCT, Cape Town, SA | *Course TA (Mechanics of Solids I - MEC2025F)*

February 2017 - June 2017

- The only undergraduate as one of five TAs for second-year level course; responsible for grading and leading tutorial sessions

#### Debating work (freelance and volunteer), SA | *Adjudicator, Coach, and Selector*

February 2015 - June 2017

- National Deputy Chief Adjudicator; National Team Coach; Western Cape Provincial Selector; Coach at two high schools

### ENTREPRENEURSHIP and INNOVATION

---

- Co-founded the UCT branch of Effective Altruism (February 2017 – June 2017), a global movement founded by philosophers which aims to promote rationality in decision-making about how to do the most possible good
- Secured and completed prestigious 4-year Allan Gray Orbis Foundation entrepreneurship program in parallel with undergraduate degree, inducted as fellow; fellowship provided ~\$24,000 funding towards undergraduate degree
- Finalist pitch at Allan Gray Orbis Foundation national competition (July 2016), for a concept designed to extend top-quality tertiary education opportunities to the developing world using innovative technology and new campus models
- Best pitch at Startup Safari entrepreneurial immersion tour to India (January 2016), for a technical and business concept to support mobile adoption in rural developing contexts, which is a significant driver of socio-economic outcomes
- Category winner (housing and settlement studies) at Eskom International Expo for Young Scientists (October 2012), for an engineering design of a sustainable and packable tent alternative for refugee camps, which often become de facto settlements

### SKILLS, LANGUAGES, and INTERESTS

---

- Programming languages: Python, Java, C, MATLAB, SQL, VBA
- Other tools: Advanced Excel modelling, data processing with Alteryx, data visualization with Tableau
- Languages: English (first language), French (proficiency commensurate with four college courses with an A-grade in each)
- Other interests: soccer, long-distance running, hiking, and abstract photography; UCT and UPenn Outdoors Clubs 2015-2019

### COURSEWORK SUMMARY

---

- Philosophy, excluding Logic: 15
- Mathematics and Formal Logic: 13
- Mechanical Engineering and Applied Physics: 10
- Electrical Engineering and Computer Science: 6
- Basic science (Physics, Chemistry, Neuroscience): 4
- French Language: 4
- Humanities: 2

Note from ext collab

2 messages

Sathe, Aalok <aalok@brown.edu>

Tue, Oct 15, 2024 at 2:08 PM

To: David Laidlaw <david\_laidlaw@brown.edu>

Was awaiting this at the time of submission, but available now:

simple email confirmation for visualization class project

A

Sathe, Aalok

Mon, Oct 14, 10:05 AM (1 day ago)

☆

hi! David would just like to see a quick note from 'external collaborators' about our class project proposals---just confirming you're still interested in visua

P

Pavlick, Ellie

Mon, Oct 14, 10:44 PM (15 hours ago)

☆

Yes confirmed, and still interested! Does this response suffice?

A

Sathe, Aalok <aalok@brown.edu>

Mon, Oct 14, 10:44 PM (15 hours ago)

☆ ↶ ⋮

to Ellie

Yes! Thank you!

\*\*\*

David Laidlaw <laidlaw.david@gmail.com>

Tue, Oct 15, 2024 at 5:49 PM

To: "Sathe, Aalok" <aalok@brown.edu>

Got it. Can you put it in an updated PDF. Leave the original one there and make an asatheXX\_updated.pdf

[Quoted text hidden]

--

David Laidlaw, Professor, Brown Computer Science

Box 1910, Providence, RI 02912, +1-401-354-2819

<http://www.cs.brown.edu/~dhl>

# An Interactive Method For Contextualizing UMAP Visualizations of Population Genetics Data Using Admixture Bar Plots

Byron Butaney

PI

`byron_butaney@brown.edu`

Alexander Diaz-Papkovich

Collaborator

`alex_diaz-papkovich@brown.edu`

Sohini Ramachandran

Collaborator

`sramachandran@brown.edu`

October 14, 2024

## Abstract

We present a novel method for visualizing genotype matrices that uses admixture data and measures of uncertainty to improve standard UMAP visualizations. The method overcomes the tendency of traditional visualizations to exaggerate genetic differences among populations. By incorporating visualized measures of individual ancestral history, our method better prevents misinterpretation and misuse of the data, as will be evaluated through a qualitative user study.

Dear Editor and Reviewers,

We appreciate the constructive feedback on the manuscript, and we have prepared a revision accordingly. Please see below for an itemized list of your feedback and a brief description of how each will be addressed in the revised proposal.

**Reviewer 1:** asathe1

**Reviewer 2:** bleahey

**Reviewer 3:** apraka15

**Reviewer 1**

[R1.1] Overall: 4

[R1.2] Interdisciplinary: 2. target domain is genomics visualization

[R1.3] Scientific: 5. aids the scientific interpretation of genomics data by adding uncertainty estimates to the visualized UMAP points. adds admixture data, increasing our understanding of the data

[R1.4] Visualization: 3. figuring out visualization of admixture on top of points in UMAP space as well as annotating uncertainty is a complex visualization problem, and a solution would be considered a visualization contribution.

[R1.5] Significant: 3. has significance in terms of aiding the understanding of the visualized data to avoid controversy

[R1.6] Novel: 3. novel solution to an existing and important problem

[R1.7] Goals clearly stated: 4-5. yes, at a high level. clear description of the methods needed.

[R1.8] Likelihood of Success: 4. likely

[R1.9] Strengths: good problem identification, good identification of what needs to happen as a solution to said problem.

[R1.10] Weaknesses: Needs more discussion of specifics of how the problem will be solved

**Response:** *We have clarified our approach to merging admixture visualizations with UMAP displays in our Aims sections.*

[R1.11] other comments: - what other areas could your solution be applied to, where you need to visualize uncertainty?

**Response:** *Although we no longer intend to visualize uncertainty (due to time constraints), we have added an explanation of how our visualization solution could be applied to external applications. In particular, we have described how UMAPs have been used in galaxy cluster analysis our Significance section, therefore demonstrating that an improvement in UMAP technology can inform a variety of external disciplines. Generally speaking, researchers in any discipline that use UMAPs could benefit from knowledge on whether informational overlays provide context to their data or if this extra information confuses the viewer more.*

[R1.12] Other comments for discussion: n/a

**Reviewer 2**

[R2.1] Overall: 3

[R2.2] Interdisciplinary: 1; strong relation to genomics, sociology, and visualization with potential applications to other visualization domains

[R2.3] Scientific: 5; Mentions scientific goals as more of an aside on insights into genetic diversity. Emphasis on how this improves education and understanding of genetic diversity is understated and could improve this score.

**Response:** *You are correct. We have improved the Significance section to highlight how the project will educate viewers and aid them in understanding genetic diversity holistically. We have also included an example graphic that I presented in class directly from Sohini's paper (the graphic about the spike in misuse of genetic data on social media following controversial socio-political events).*

[R2.4] Visualization: 2; uncertainty visualization is novel and overall task is well framed in how it ties into visualization and understanding of genetic data.

[R2.5] Significant: 3; has significance in terms of aiding the understanding of the visualized data to avoid controversy

[R2.6] Novel: 4; Uses somewhat dated and known methods. Incorporating more generalized frameworks or approaching this as a proof of concept for other visualization methods (besides umap) and tasks (besides genomics) could make this extremely novel, but also is quite an ambitious goal.

**Response:** *You are correct that an application of our methods in other contexts is ambitious given the 6 week timeframe. We have incorporated an optional milestone towards the end of our 6-week plan, so that (should time permit) we can try to create a more generalized framework that also works for other dimensionality reduction methods (like t-SNE or PCA).*

[R2.7] Goals clearly stated: 2; well defined methods of umap and admixture

[R2.8] Likelihood of Success: 2; umap and admixture are well established methods, and the project lays out a good timeline for incremental goals with these tools.

[R2.9] Strengths: strong interdisciplinary connections, well defined methods, achievable with clear timeline.

[R2.10] Weaknesses: without a strong discussion section or attempt at a more generalized framework, this may fall short on novelty

**Response:** *To improve the strength of our Significance section, we have included a more thorough dialogue regarding the ways in which our framework and results from the user study can be useful to researchers in fields outside of biology. For example, data from the user study on whether interactively merging data with a UMAP display is beneficial or obscuring can be useful for any discipline that deals with dimensionality reduction (ie, physics). We have also given a concrete example of how dimensionality reduction is used in a physics context (galaxy cluster analysis) in order to better highlight this use case. Regarding the more generalized framework, see our response to [R2.6] above.*

[R2.11] Other comments for discussion: n/a

### **Reviewer 3**

[R3.1] Overall: 2

[R3.2] Interdisciplinary: 1. The work is clearly interdisciplinary bringing together genomics, statistics and computer science

[R3.3] Scientific: 2. The proposal aims to visualize an interesting scientific question

[R3.4] Visualization: 2. The proposal aims to combine umap and admixture and measures of uncertainty.

[R3.5] Significant: 2. The significance is clearly motivated.

[R3.6] Novel: 2. Combining UMAP and admixture is interesting, but there is quite a lot of potential in visualizing the uncertainty.

**Response:** *We are glad that the uncertainty visualization idea was of particular interest. Unfortunately, due to the nature of a tight 6-week schedule, we determined that we will not be able to visualize measures of uncertainty as well as admixture data. Further, we decided that doing both might clutter the display. Future work might address the potential with visualizing uncertainty.*

[R3.7] Goals clearly stated: Yes

[R3.8] Likelihood of Success: 8/10

[R3.9] Strengths: clear aims, clear plan, clear visualizations and contributions

[R3.10] Weaknesses: The document could be tidied up - it still has some of the default fields visible from the template. Reference images would be helpful.

**Response:** *The mentioned default fields have been removed. A reference image demonstrating trends in extremist social media posts that use genetic data has been incorporated into the final proposal to aid readers in understanding the significance of misuse by extremist agents.*

[R3.11] Other comments for discussion: - How will this be evaluated?

**Response:** *We have included more details regarding our user study in our Aims section. Specifically, we will be using mechanical turk to gain insights into the interpretability of our improved visualizations. Users will be asked to misinterpret a traditional UMAP baseline as well as our improved UMAP with interactive admixture displays. Users will then rank how difficult it was to obscure the findings of each visualization, and these ratings will be tallied.*

# 1 Aims

The overarching goal of this project is to design a nuanced approach to UMAP-based visualizations of genotype matrices by visually incorporating individual ancestry data, referred to as admixture throughout this proposal. Specifically, we aim to create an interactive visualization allowing users to hover over different segments of an admixture bar plot, which will then highlight the distribution of individuals from the corresponding region within the UMAP display. More broadly, we hope for this novel approach to help prevent the misinterpretation and misuse of population genetics data. Though a UMAP visualization alone may highlight that individuals may appear in distinct clusters based on one metric, our aim is to emphasize that this does not imply they belong to entirely separate populations. In reality, all individuals have complex ancestral backgrounds involving multiple populations, and these ancestral distributions cannot be fully captured by the discrete clusters shown in a UMAP visualization.

Our technical aims are therefore three-fold. First, we aim to develop a streamlined code base that simultaneously performs UMAP dimensionality reduction and computes ancestral admixture proportions for each individual, enabling researchers to see the ancestral origins of individuals within each UMAP cluster. The data of these individuals will be taken from the 1000 Genomes Project’s public dataset [7]. Second, we plan to integrate these visualizations—UMAP clusters and admixture bar plots—into an interactive display that demonstrates how individuals cannot be genetically categorized as belonging to distinct, isolated groups. Finally, we will evaluate the effectiveness of our new visualizations by conducting a user study and getting the opinions of experts. Our plan is to use Amazon’s Mechanical Turk platform to conduct a study where users are asked to create malicious tweets using both traditional UMAP visualizations and our new interactive UMAP/admixture visualizations. Users will then give a rating on how difficult they found it to manipulate each visualization. We will generate “misuse scores” from the resulting data, which will be compared to develop an evaluation of our results. The opinions of Dr. Diaz-Papkovich and Dr. Ramachandran will also be used to qualitatively evaluate the success of our new method.

# 2 Significance

Given the ever-growing polarization that has defined the socio-political landscape of the 21st century, it is crucial to be able to properly and responsibly visualize the genetic relationships between human individuals, with extra emphasis on providing adequate context for the relationships demonstrated. Without sufficient techniques to do so, misinterpretation and misuse of population genetics data has shown to fuel the discriminatory arguments of extremist groups [4]. Carlson et al. showed how rises in social media references to population genetics visualizations increase following divisive social events (see Fig. 1). Current methods for dimensionality reduction of large genotype matrices frequently exaggerate the genetic differences among populations, producing visualizations that are ripe for misinterpretation and misuse. Improving upon these methods provides strong educational merit, with the potential to improve our understanding of complicated population structures and our communication of these structures to a broader audience. By educating a broader audience on the nuance behind human genetic diversity, we can help combat the work of extremists to mislead their viewers with out-of-context population genetics research.

Our project satisfies the need for nuanced and context-driven visualizations of genotype matrices by allowing individuals to explore the data underlying a UMAP display on their own. Having a method that provides this visual context to the complicated genetic relationships represented by genotype matrices will

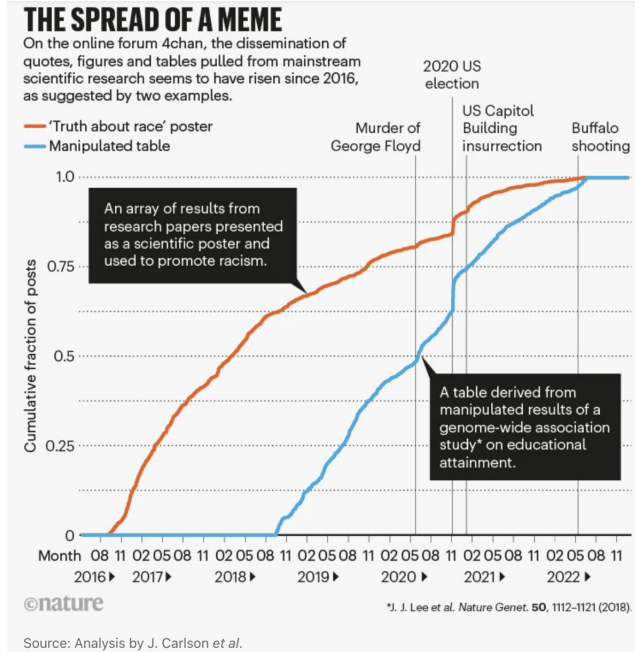


Figure 1: The Spread of a Meme [4]

allow for more accurate and responsible visualizations of genetic diversity within and between human populations. From a visualization standpoint, our project introduces a new approach for presenting complex, high-dimensional data in a way that is accessible to the general public. These enhanced visualizations have applications beyond population genetics, as high-dimensional modeling of complex relationships is also essential in fields such as anthropology, physics, and more. For instance, UMAP visualizations have been used to explore datasets as different from ours as data on the properties of galaxy clusters [8], underscoring the broad need to improve UMAP interpretability. From a science perspective, our project demonstrates unique insights into the role of individual ancestry on genetic diversity. Our results will also provide insight into the complex genetic relationships underlying the genetic data from the 1000 Genomes Project [7]. Finally, results from our user study evaluation will provide insights for researchers across many disciplines who use UMAP. Particularly, we hope to learn more about whether providing extra context to UMAP displays in the form of interactive visualizations is helpful or hurtful to overall interpretability. Though we predict the former, if our results show the latter, this might be evidence that researchers should look for other methods of improving interpretability outside of interactive visualizations.

### 3 Related Work

First proposed in 2018, UMAP has served as the standard for nonlinear dimensionality reduction and has been applied to a plethora of fields [9]. In particular, UMAP allows for researchers to maintain the majority of the structures present in the original data while significantly reducing its dimensionality. Yet, UMAP visualizations do not display any external context, which are a necessity when analyzing the results of any study. In comparison to other dimensionality reduction methods, like t-SNE and PCA, UMAP has been demonstrated to be better in regards to run time and the organization of its generated clusters [2].

Since the introduction of UMAP, the technology has been shown to be able to generalize to different types of biological data. For example, UMAP has been demonstrated to be able to distinguish between biological groups and to differentiate batch effects in single-cell RNA sequencing (scRNA-seq) data [10], although it was not able to demonstrate the significance of these relationships or the context behind them.

More recently, UMAP has been shown to provide unique insights into population genetics data as well [6]. In particular, it was in 2021 shown to be able to visualize complicated population structures and relationships while also finding lower level demographic insights. Similar to the authors of the original UMAP study, however, Diaz-Papkovich et al. found that UMAP visualizations lacked outside context. Further, their methods lacked any notion of ancestral genetic history.

In 2009, ADMIXTURE was presented as a method to evaluate ancestry proportions of individuals in order to aid researchers in understanding the different source populations of individuals [1]. When used alone, however, ADMIXTURE does not provide clustering or dimensionality reduction to high-dimensional datasets. Though other methods, such as PONG [3] admixture bar plots, have since come out to improve admixture visualizations, they have not been applied or integrated with other forms of population genetics visualizations. Our project addresses this limitation by incorporating UMAP visualizations along with these PONG bar plot visualizations in order to simplify the data in a visual context. In particular, users will be able to interact with bar plots that show the proportions of each population from which individuals come from. When hovering over a bar from a specific ancestral population, individuals who come from those populations will be highlighted within our UMAP display. This will show that individuals who discretely fall into one ancestral population are actually distributed across multiple UMAP clusters, minimizing the notion of discrete genetic differences between human beings.

Since the introduction of ADMIXTURE in 2009, efforts have been made to use admixture data to contextualize UMAP displays. In a separate paper to their work in 2021, Diaz-Papkovich et al. worked to color-code UMAP clusters by ancestral proportions in a separate paper [5], the authors mention that these newer visualizations still suffer from the limitation that they have more adjustable parameters than other traditional dimensionality-reduction methods. This means that the data can be presented in many different ways depending on how these parameters are shifted, making it easy to exaggerate the genetic variation of populations [5]. As such, there is still a demonstrated need for a more approachable and interpretable method of combining admixture insights with UMAP displays.

## **4 Research Plan**

### **1. Week 1: Data Preparation**

- (a) Genotype data will be downloaded from the 1000 Genomes Project, a public database for genotype matrices.
- (b) The data will be then converted into a format suitable for use with ADMIXTURE, a common software tool that allows for admixture analysis.

### **2. Week 2: Initial UMAP baseline and ADMIXTURE familiarity**

- (a) A UMAP will be generated to reduce the dimensionality of and visualize the data from the 1000 Genomes Project to show population clustering. This is the standard method currently used, and will be utilized as a baseline for interpretability in the later user study.
  - (b) Familiarity with ADMIXTURE's PLINK format will be developed to support Week 3's goals.
- 3. Week 3: Admixture Analysis and Bar Plot Generation Via PONG**
- (a) The Admixture tool will be run on the 1000 Genomes Project data to acquire ancestry proportions for all individuals that were used to make the UMAP from the prior week.
  - (b) Data will be entered into PONG [3] to generate admixture bar plots
- 4. Week 4: Merging Admixture Display With UMAP.**
- (a) Admixture/PONG bar plots will be merged with the UMAP visualizations made in Week 2. Interactive feature will be implemented, allowing for individuals to be highlighted in UMAP if they have ancestry from the bar plot that the user's mouse is hovering over.
- 5. Week 5: Finalizing Visualizations and Running User Study**
- (a) Interactive display will be finalized if it is not completed by the end of Week 4.
  - (b) If time permits, the code will be organized into a general framework compatible with other dimensionality reduction methods (t-SNE, PCA). This milestone will be ignored if the interactive display from Week 4 is not finished by the start of this week.
  - (c) Mechanical Turk user study will be conducted to generate the "misuse scores" of both our new visualization as well as the basic UMAP.
- 6. Week 6: Presentation and Abstract**
- (a) Develop abstract and final presentation.
  - (b) Practice presenting and ensure all materials are submitted.

## References

- [1] D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, 19(9):1655–1664, July 2009.
- [2] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, 37:38–44, 2018.
- [3] A. A. Behr, K. Z. Liu, G. Liu-Fang, P. Nakka, and S. Ramachandran. pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics*, 32(18):2817–2823, 06 2016.
- [4] J. Carlson, B. M. Henn, D. R. Al-Hindi, and S. Ramachandran. Counter the weaponization of genetics research by extremists. *Nature*, 610(7932):444–447, Oct. 2022.
- [5] A. Diaz-Papkovich, L. Anderson-Trocmé, C. Ben-Eghan, and S. Gravel. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genet*, 15(11):e1008432, Nov. 2019.
- [6] A. Diaz-Papkovich, L. Anderson-Trocmé, and S. Gravel. A review of umap in population genetics. *Journal of Human Genetics*, 66(1):85–91, Jan 2021.
- [7] S. Fairley, E. Lowy-Gallego, E. Perry, and P. Flicek. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research*, 48(D1):D941–D947, 10 2019.
- [8] R. Haggar, F. De Luca, M. De Petris, E. Sazonova, J. E. Taylor, A. Knebe, M. E. Gray, F. R. Pearce, A. Contreras-Santos, W. Cui, U. Kuchner, R. A. Mostoghiu Paun, and C. Power. Reconsidering the dynamical states of galaxy clusters using PCA and UMAP. *Monthly Notices of the Royal Astronomical Society*, 532(1):1031–1048, 06 2024.
- [9] L. McInnes and J. Healy. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, Feb. 2018.
- [10] Y. Yang, H. Sun, Y. Zhang, T. Zhang, J. Gong, Y. Wei, Y.-G. Duan, M. Shu, Y. Yang, D. Wu, and D. Yu. Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. *Cell Rep*, 36(4):109442, July 2021.

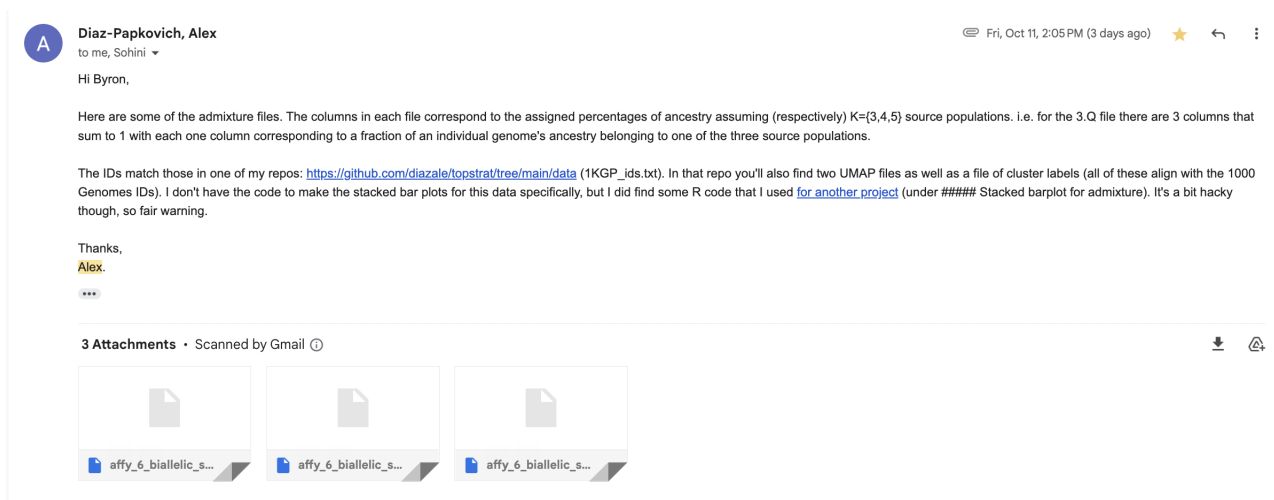


Figure 2: Proof of Collaborator Communication and Support

# Byron Butaney

(480)-862-9109 | 69 Brown St. #3040 Providence, RI 02906 | [byron\\_butaney@brown.edu](mailto:byron_butaney@brown.edu)

## EDUCATION

### Brown University

B.A. Computer Science

Providence, RI

Expected Graduation: May 2025

## PROFESSIONAL EXPERIENCE

### Brown University Department of Chemistry

Tutor for CHEM 0350 (Organic Chemistry I), CHEM 0360 (Organic Chemistry II), and CHEM 0100 (General Chemistry)

Providence, RI  
September 2024 – Present

- Met weekly with groups of organic chemistry students to answer conceptual questions and improve understanding
- Fostered critical thinking, problem solving, and learning strategies by promoting peer-to-peer dialogue

### Hasbro Children's Hospital

Pediatric Emergency Department Volunteer

Providence, RI

September 2023 – Present

- Provided weekly pre-op and post-op logistic and emotional support to children and families in Hasbro's Pediatric Emergency Department

### Singh Lab, Brown University Center for Computational Molecular Biology

Undergraduate Research Assistant

Providence, RI

May 2023 – Present

- Worked under Ghulam Murtaza and Dr. Ritambhara Singh to develop a graph-generative neural network to synthesize single-cell Hi-C data from single-cell RNA and DNA input
- Second-authored a full-length publication in *Bioinformatics*: "scGrapHiC: deep learning-based graph deconvolution for Hi-C using single cell gene expression."
- Work on scGrapHiC was also presented at RECOMB 2024 and ISMB 2024, where it was nominated for the Ian Lawson Van Toch Best Paper Award

### Brown University Department of Computer Science

Undergraduate Teaching Assistant (CSCI 1430: Computer Vision)

Providence, RI

December 2022 – May 2023

- Designed and assessed bi-weekly assignments for a cohort of more than 100 Computer Vision students
- Conducted weekly office hours to provide dedicated support, guidance, and clarification on course materials and debugging strategies

### Brown Infectious Disease Society

Research Lead (September 2023-September 2024), Member (February 2022-Present)

Providence, RI

February 2022 – Present

- Spearheaded a literature review project on therapeutic approaches for microbial infections that address the escalating challenge of antibiotic drug resistance
- Facilitated inclusive weekly discussions for students to engage with the field of infectious diseases under a multidisciplinary lens

### Full Stack @ Brown

Backend Engineer

Providence, RI

January 2022 – Present

- Collaborated with a team on Project Mootual, a mobile app that allows users to solicit friend requests from mutual connections
- Implemented user authentication for the login screen with Firebase Auth

### Brown Journal of Medical Humanities

Editor-in-Chief (September 2023-Present), Editor (September 2021-September 2023)

Providence, RI

September 2021 – Present

- Led club meetings and conducted editor workshops to instruct members in the intricacies of editing submissions for the journal
- Handpicked prose, poetry, and visual works submitted to the journal, emphasizing the intersection of medicine and art in each piece
- Orchestrated the annual editing timeline, overseeing the creation of the Call for Submissions and effectively managing the selection and publication process for the chosen works in each issue.

### Mayo Clinic Department of Biomedical Sciences

Deep Learning Research Intern

Scottsdale, AZ

June 2022 – September 2022

- Wrote a Python program to detect and annotate 51 key body parts in chest X-rays (including the lungs, diaphragm, and ribs)
- Achieved an accuracy of 88% by training a region-based convolutional neural network on 5000 annotated x-rays in TensorFlow
- Wrote a Python program to detect whether individuals had any of 14 lung diseases by analyzing a patient's chest x-rays
- Trained a Swin Transformer in PyTorch on 100k+ images NIH chest x-rays and achieved an accuracy of 76%

### Bluebonnet Data

Data Science Fellow

Bellaire, TX

May 2022 – September 2022

- Led team of 4 to develop data-driven goals for Patrick Belmont's Utah House of Representatives Campaign
- Created a web scraping program in Python to collect and plot precinct-level data for votes on ballot measures from 2018 and 2020
- Remodeled open-source code from the Bluebonnet code repository to find the most effective yard sign locations for the campaign

### Stem Cell Laboratory at MD Anderson at Banner Health

Research Intern, Clinical Shadow

Gilbert, AZ

Jun 2020 – Aug 2021

- Collaborated on research with Dr. Sergio Torloni, to validate a better method of stem cell collection for multiple myeloma patients.

- Contributed to a research abstract using statistical analyses to improve stem cell collection efficiency for multiple myeloma patients.
- Shadow Dr. Torloni as he diagnoses patients, performs apheresis and blood transplants and manages stem cell collections.

## PROJECTS

---

### Trust Your Tesla? Collision Course Prediction for Simulated Autonomous Vehicles

- Implemented next-frame prediction to determine whether a car in Carla simulator is at risk for collision with 91.7% accuracy
- Designed a convolutional LSTM-based neural network that allows for the classification of time-series image data
- Organized and collected 14000 sequences of 8 image frames from the simulator to train our model

### Interactive Redlining and Weather Data Visualizer

- Wrote a full-stack, interactive web program that allows users to view redlining data in the US
- Created a web API with Spark Java that parses redlining JSON data and returns data within requested bounds of latitude and longitude for frontend use
- Backend also allows user to check the weather of a given location by accessing the NWS API
- Wrote frontend with TypeScript (React) and Mapbox GL that displays a map of the US with redlining data overlayed on user-specified regions and supports dragging and zooming

### Travel Assistant

- Created a Python program that stores public Airbnb listing data in a PostgreSQL database and can run SQL commands to help the user plan a trip to any of 155 cities around the world
- Designed program to be capable of answering complex questions based on 74 different data points such as number of rooms, minimum nights, etc.

## SKILLS & INTERESTS

---

**Programming Languages:** Fluent in Python, Java, Typescript, Javascript, C++, OCAML, Racket/Lisp; Familiar with SQL, HTML, CSS

**Other Skills:** Git, React, Web APIs, PyTorch, Tensorflow, Keras, Sklearn, NumPy, Firebase, Statistics

# Calendar Based Cancer Treatment Regimen Browser

Brendan Leahey

PI

brendan\_leahey@brown.edu

Jasmine Liu

Co-PI

jasmine\_c.liu@brown.edu

Sanjay Mishra

Collaborator

Sanjay\_Mishra@brown.edu

Jeremy Warner

Collaborator

Jeremy\_Warner@brown.edu

Sandeep Jain

Collaborator

Sandeep\_Jain@brown.edu

October 15, 2024

## Abstract

We aim to enhance cancer treatment understanding for patients by converting regimens on the HemOnc database into a user-friendly calendar format using the Google Calendar API. After implementing visual encodings such as color, saturation, opacity, and density and fuzzy scheduling techniques, we evaluate visual and oncology contributions through user studies and surveys.

# 1 Response to Reviews:

We appreciate the time and effort put into reviews and have included responses to reviewers' feedback on the attached proposal.

## 1.1 Aalok Sathe

**Overall:** 3-4

**Interdisciplinary:** 2; target domain is cancer therapeutics

**Scientific:** 6; this is not intended to be a scientific contribution; though it is definitely aligned with the longer-term translational goal of science: treatment. Additionally, enabling better patient tools for treatment will someday enable better administration of clinical trials, etc.

**Response:** *We hope that patient interaction with the browser will be seen as a novel contribution towards oncology (and other medical fields with similar regimens). We have received feedback from health professionals we have previously worked with that a personal calendar approach to treatment regimens is appealing. As such, we have detailed more concrete methods to evaluate this interaction in our proposal and timeline and will continue to develop these methods with our collaborators and their patients.*

**Visualization:** 3; the primary contribution of this project. The authors propose to intuitively visualize a calendar of cancer treatment timelines.

**Significant:** 3; seems to be very important work.

**Novel:** 4-5; it will involve the exploration of some visualization techniques well suited to a tabular and temporal format, including exploration of color schemes to display data.

**Response:** *To emphasize the novelty of these techniques, we have added a discrete set of novel visualization methods and their combinations as part of our aims section. We have also detailed a comparison trial with a baseline visualization of a Google Calendar and HemOnc to evaluate the effectiveness of our visualization methods.*

**Goals clearly stated:** 3; yes.

**Likelihood of Success:** 3; highly likely.

**Strengths:**

- Concrete goals
- Clear significance
- Good identification of prior work to compare to

**Weaknesses:**

- Authors may want to hypothesize or propose some minimal extensions for future work that may be scientific contributions and simply mention them in the proposal to woo funding agencies.

**Response:** *We have elaborated on the aims section, adding a future work/discussion section to the proposal. This details the collaborator's future goals of integrating an even more generalized pipeline for the treatment browser, encompassing practitioner views, hematology treatment regimens, and comparative visualization methods.*

**Other comments:**

- Could you do a human study contrasting a few approaches to find out what is more intuitive?

**Response:** *In response to this concern, we have added greater detail to our mention of "user study."*

## 1.2 Byron Butaney

**Overall:** 3

**Interdisciplinary:** 1; mix of visualization with healthcare and patient care is highly interdisciplinary.

**Scientific:** 4; while meaningful, further clarification is needed for how the visualization can be used in current clinical settings.

***Response:** We have added greater detail to our background section to clarify the current clinical settings. We have also added greater detail in our aims section to how both ourselves and our collaborators see this being used by patients and how it can be properly evaluated for its effectiveness in a clinical setting.*

**Visualization:** 3; novel methods, but visual elements of the visualizations are not as clear as they could be.

***Response:** See the novelty response to reviewer 1—we have detailed which specific visualization methods we hope to explore after further consulting the literature.*

**Significant:** 2; enhancing patient care is highly significant, and integration could impact many people from many backgrounds.

**Novel:** 3; visualization method and overall idea of treatment planning is unique, but the inclusion of life milestones and experimental visualizations could be better fleshed out.

***Response:** We have discussed with our collaborators the level of detail that should be included on life milestones. One potential solution we have discussed, using the Google Calendar API, will allow us to integrate a patient's personal calendar into the treatment browser. We will proceed for now using this method as our primary option and will build out on wrapping around their API as we explore further methods.*

*See previous response on experimental visualizations.*

**Goals clearly stated:** 1

**Likelihood of Success:** 3; project seems mostly feasible, but the 3D visualizations and integration of user feedback seem a bit challenging... maybe focus on more manageable methods first?

***Response:** In concurrence with this feedback, we have pushed off heightened 3D visualization methods and some of the more experimental visualizations to future work. Getting detailed feedback from users on more basic visualizations while experimenting with factors like color and layout will be our priority to establish a concrete baseline for future work.*

**Strengths:** Patient-centered scope yields a very meaningful proposal. Highly interdisciplinary and novel visualization methods.

**Weaknesses:** Some details are lacking for the calendar layout and for evaluation. Feasibility of 3D visualizations is questionable.

***Response:** Addressed in previous responses.*

## 1.3 Key points from Prof. Laidlaw

**Significant:** 8; the “Time-permitting” phrase on the evaluation greatly reduced the potential significance. Even if patients cannot be surveyed, some kind of evaluation is essential. Non-patients could provide feedback. So could the collaborators. This score would be much higher with more evaluation specifics and commitment. The proposal claims the approach will be “approachable” and “relatable.” Also, that a calendar framework is “novel” and “approachable.” Those are further claimed to constitute a novel scientific contribution. Can you measure those properties? Do they really constitute a novel scientific contribution? There were other claims that might suggest evaluations: accessible, effective, understanding.

***Response:** This was a key point when looking at feedback overall. This contributed greatly to our decision to cut back on some of the more experimental visualizations and focus on a more concrete evaluation*

*plan. As stated, we have detailed more concrete visualization methods we will use and have explored literature in order to best evaluate significance through questions about whether the visualization is more “approachable” and “relatable” than the wiki format of HemOnc or a plain Google Calendar.*

**Novel:** 6; is “Calendar-based” novel? I thought that was one of the ways that the collaborators already use. Other approaches mentioned appear to be combinations of earlier research. Combinations can be novel, even of known approaches.

**Response:** *As far as we know, collaborators have not previously explored a calendar-based approach to this specific regimen dataset, just the browser and network visualizations shown previously. However, we have more concretely outlined which methods we plan to combine in hopes of building on specific points discussed in previous user studies.*

## 2 Aims

We will transfer key treatment plans from the extensive cancer treatment database, HemOnc, into an easy to navigate calendar view, a step towards a patient-focused, generalized framework for displaying treatment plans. Utilizing the Google Calendar API, we incorporate patient’s life milestones towards a more relatable understanding of regimens. Finally, we implement and combine visualization methods overlaying the calendar for displaying longitudinal clinical data. This include proven methods such as color saturation, density, and opacity of datapoints. Further, we employ novel calendar visualization methods such as “fuzzy scheduling” (see Related Work) with respect to factors like occlusion, priority, and workload limits which have not been implemented in an oncology context.

To evaluate these methods, we will carry out a user study of patients beginning treatment and a survey of patients who have gone through treatment already. This evaluation is carried out in terms of the scientific and visualization impact. A comparison study with HemOnc and a baseline calendar view will assess our contribution to the task of patient experience in cancer treatment regimens. We evaluate all permutations of our visualization methods in terms of its accessibility and practicality. Through this user study, we hope to gain key insights on improving patient’s understanding of their cancer treatments and visualizing longitudinal data in a calendar format.

## 3 Significance

Cancer treatment regimens are structured plans of cancer treatment, including factors like chemotherapy drugs, surgeries, radiation, and various other treatments. The selection of regimen has been shown to have significant affects on the quality of life and cognitive function of breast cancer patients, often requiring multidisciplinary support [1]. Further, a survey of patients showed their strong interest in knowing available treatment options, timing and risks, as well as having multiple modes in which this information is presented [2]. Typical information presentation varies, but can include discussion with doctors, pamphlets, and video/audio sources. This also includes treatment calendars, which are generally created by nurses or advanced practitioners [3]. A survey of these patients found approximately **95% indicated the creation of digital calendar generation software to be “moderately to extremely helpful”** [3]. We hypothesize that a calendar based approach will be particularly effective in personalizing a patient’s understanding of these regimens. Automatically synthesizing these personalized calendar visualizations from the HemOnc database represents a novel contribution (see Related Work). Further, by providing a generalized framework that may be updated with future clinical data, we can allow for continued accessible communication in the future.

Calendar or scheduling visualizations have, in general, been used for a variety of applications such as personal health, job planning, and more. Exploring and combining previous visualization methods then collecting feedback from a user study can allow for continued progression of digital calendar and scheduling applications (see Related Work for specific methods and applications). In particular, the visualization of “fuzzy scheduling” in the context of healthcare is a novel presentation of a growing logical approach to regimens and treatments.



## 4.2 Healthcare Calendars

Calendar or longitudinal approaches exist to map factors like genomic features alongside cancer treatment regimens. For example, the cBioPortal visualizes changes in protein and other longitudinal data over years of treatment using color coded data [6]. Manual entry calendar approaches specific to patient regimens also exist [7, 8], but remain underexplored with strong patient desire for more accessible software [3]. We may use existing methods cBioPortal as a loose framework for effective existing methods specific to cancer regimens. However, applying more sophisticated visualizations, as well as contexts that steer away from the single cell and protein applications, we provide an extremely desired patient contribution and a basis for future automated calendar softwares.

## 4.3 Visualization Methods

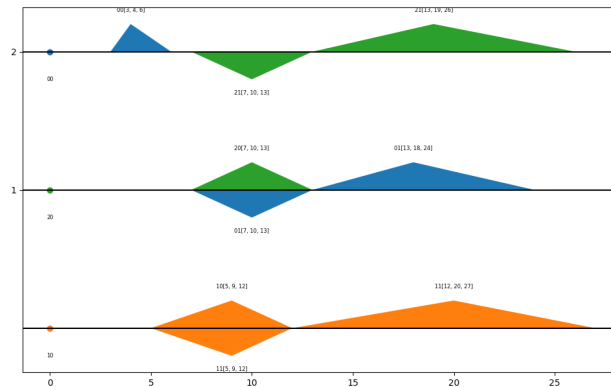


Figure 2: Gantt Chart for Fuzzy Scheduling Visualization

Other existing calendar visualization methods have been explored within the context of personal health calendars, which have been evaluated more formally. User tests from a dissertation provide a framework for effective use of visual encodings such as color, saturation, hue, slope, and density [9]. Effectively applying these results will be necessary to build upon previous developments in calendar-based visualization. Further, visualization methods for fuzzy scheduling (detailed below) remain untested within a personal calendar or healthcare context. Incorporating some of these methods and receiving user feedback from patients will be both a visualization and oncology contribution.

“Fuzzy Scheduling” relates to the broader concept of fuzzy logic, which deals with imprecise information using ranges. Fuzzy scheduling in particular refers to scheduling without a fixed start or endpoint. Recent novel applications in chemotherapy have used fuzzy scheduling alongside controllers, which are mathematical models that help guide treatment [11]. In general, it presents an excellent way for treatments to handle uncertainty in patient responses and model adaptive treatment. Other applications of fuzzy scheduling, such as for shipyard management, have displayed it within a Gantt chart to display uncertainty in ship arrival as shown in Figure 2 [15]. Other applications of fuzzy scheduling for time series data include shop management, algorithm timing, and construction [12, 13, 14]. To our knowledge, there is no existing visualization of fuzzy scheduling for cancer treatment regimens, especially not in calendar format. Applying fuzzy scheduling and receiving user feedback may represent a significant improvement over other calendar based methods. Additionally, better understanding how users interact with fuzzy scheduling visualizations can have significant reach into the applications described above and more.

## 5 Research Plan

### 5.1 Timeline

Task	Time Estimate
HemOnc/Raw Regimen data to generalized data type pipeline	1-2 weeks
Base Calendar View	0.5-1 week
Implement Novel Visual Elements	1-2 weeks
Write up method results and novel technological contributions	2 weeks
Concurrent small scale user study comparing visualization methods	

Table 1: Timeline for Project Tasks

### 5.2 Visualization Methods

Key visual encoding methods we will explore include color, saturation, density, and opacity, motivated by previous works [5, 9]. Color will be used to represent categories of dates, the saturation will represent its relative impact, density will represent relative importance, and opacity will model certainty of information based on previous clinical trials. We will also include fuzzy scheduling, which we have motivated above. These methods will be implemented incrementally on a plain Google Calendar representation of our data.

Google Calendar’s API allows us to interface with a patient’s existing calendars, given permissions, as well as create predefined event and calendar visual objects [16]. These objects allow us to specify color saturation, but we will have to implement opacity and size changes wrapping around these objects. Further, the Google Calendar API does not provide “fuzzy” start and end times, so we will implement a novel class wrapping around their API. We will experiment with exploiting the “all-day” event or the “parent event” features built into their library to create these fuzzy events.

### 5.3 Evaluation

With our collaborators, we will obtain as many current cancer patients and previous patients as possible to survey during our 2 week window. This may include those with personal relationships to the PI, for which we will document conflicts of interest. Motivated by previous literature on calendar assessment surveys, we will include a pointed 1-5 survey on the practicality and clarity of whether the calendar improved their understanding of desired information such as [2, 9]:

1. treatment options (chemotherapy, radiation, surgery, hormones etc.)
2. tests and results
3. impact of treatments (timing, aftereffects)
4. clinical trials
5. treatment duration

Further, we will include a general interview of current and previous patients to establish visualization significance. These will be more open ended questions such as:

1. Which visualization method do you prefer? why?
2. Would you see yourself using the calendar during your treatment (or would you have for previous patients)?
3. What visual elements stood out to you the most? how did they make you feel?

Between these two sub-studies, we hope to get a better sense of the calendar’s merit and how it could be improved in the future (if found to be as desirable as in previous surveys).

## **5.4 Facilities**

For our calendar implementation, we use the Google Calendar API, a free service. Some API usage quotas and operational limits are imposed, but we do not anticipate reaching these limits or requiring a paid subscription through the user testing stage [16].

The HemOnc Knowledge base has been provided by our collaborators, which has been detailed below in the Note From Collaborator (6).

No additional computational resources should be required for this project.

## References

- [1] Grusdat, N. P., Stäuber, A., Tolkmitt, M., Schnabel, J., Schubotz, B., Wright, P. R., Heydenreich, M., Zermann, D. H., & Schulz, H. (2022). Cancer treatment regimens and their impact on the patient-reported outcome measures health-related quality of life and perceived cognitive function. *Journal of Patient-Reported Outcomes*, 6(1), 16. <https://doi.org/10.1186/s41687-022-00422-5>.
- [2] Chua, G. P., Tan, H. K., & Gandhi, M. (2018). What information do cancer patients want and how well are their needs being met?. *Ecancermedicalscience*, 12, 873. <https://doi.org/10.3332/ecancer.2018.873>.
- [3] Mueller, E. L., Cochrane, A. R., & Carroll, A. E. (2023). Perceptions of chemotherapy calendar creation among US pediatric oncologists. *Pediatric Blood & Cancer*, 70(12), e30688. <https://doi.org/10.1002/pbc.30688>.
- [4] Warner, J. L., Dymshyts, D., Reich, C. G., Gurley, M. J., Hochheiser, H., Moldwin, Z. H., Belenkaya, R., Williams, A. E., & Yang, P. C. (2019). HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. *Journal of Biomedical Informatics*, 96, 103239. <https://doi.org/10.1016/j.jbi.2019.103239>.
- [5] Warner, J., Yang, P., & Alterovitz, G. (2013). Automated synthesis and visualization of a chemotherapy treatment regimen network. *Studies in Health Technology and Informatics*, 192, 62–66.
- [6] de Bruijn, I., Kundra, R., Mastrogiacomo, B., Tran, T. N., Sikina, L., Mazor, T., Li, X., Ochoa, A., Zhao, G., Lai, B., Abeshouse, A., Baiceanu, D., Ciftci, E., Dogrusoz, U., Dufilie, A., Erkoc, Z., Garcia Lara, E., Fu, Z., Gross, B., Haynes, C., et al. (2023). Analysis and Visualization of Longitudinal Genomic and Clinical Data from the AACR Project GENIE Biopharma Collaborative in cBioPortal. *Cancer Research*, 83(23), 3861–3867. <https://doi.org/10.1158/0008-5472.CAN-23-0816>.
- [7] ChemoExperts Foundation, Inc. (2024). ChemoExperts Treatment Tracker®. Retrieved October 14, 2024, from <https://www.chemoexperts.com/calendar.html>.
- [8] OncoLink. (2024). Regimen Calendars. Retrieved October 14, 2024, from <https://www.oncolink.org/cancer-treatment/cancer-medications/support/regimen-calendars>.
- [9] Tavakkol, S. (2014). *Personal Analytical Calendar* (Doctoral dissertation).
- [10] Hartl, P. R. (2008). *Visualization of Calendar Data* (Doctoral dissertation).
- [11] Ghasemabad, E. S., et al. (2022). Design and implementation of an adaptive fuzzy sliding mode controller for drug delivery in treatment of vascular cancer tumours and its optimisation using genetic algorithm tool. *IET Systems Biology*, 16(6), 201–219. <https://doi.org/10.1049/syb2.12051>.
- [12] Behnamian, J. (2016). Survey on fuzzy shop scheduling. *Fuzzy Optimization and Decision Making*, 15, 331–366. <https://doi.org/10.1007/s10700-015-9225-5>.
- [13] Jalali Khalil Abadi, Z., & Mansouri, N. (2024). A comprehensive survey on scheduling algorithms using fuzzy systems in distributed environments. *Artificial Intelligence Review*, 57(4). <https://doi.org/10.1007/s10462-023-10632-y>.

- [14] Ghaffari, A., & Khatami, H. (2024). Development of a decision-making framework for the selection of sustainable construction materials: An integrated fuzzy-AHP and fuzzy-TOPSIS approach. *Journal of Construction Engineering and Management*, 150(4). [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001996](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001996).
- [15] Meng, C., Feng, Z., Zhao, D., Li, X., Yu, J., & Yang, L. (2024). A Fuzzy Scheduling Method for Pipeline Processing in Shipyards Incorporating the Black Widow Optimization Algorithm. *Applied Sciences*, 14(13), 5639. <https://doi.org/10.3390/app14135639>.
- [16] Google Developers. (2024). Overview of the Google Calendar API. Retrieved October 14, 2024, from <https://developers.google.com/calendar/api/guides/overview>.

## Brendan Leahey

E-Mail: [brendan\\_leahey@brown.edu](mailto:brendan_leahey@brown.edu) — LinkedIn: [linkedin.com/in/brendan-leahey/](https://www.linkedin.com/in/brendan-leahey/)

### Education

**Brown University**, Providence, RI — Sc.B., Math-Computer Science — Expected Graduation: 05/25

#### Relevant Coursework:

- Foundations of Living Systems, Deep Learning in Genomics
- Software Engineering, Data Structures and Algorithms
- Interdisciplinary Scientific Visualization, Computer Vision

### Work Experience

**Flox AB**, Stockholm, SE Detection (Artificial Intelligence) Intern, June 2023 - August 2023

Drone-engineering startup focusing on easing the impact of wildlife on agriculture

#### Responsibilities:

- Leveraged and modified long-ranged and multimodal deep learning techniques (YOLO) to perform real-time recognition of geese from UAVs
- Cleaned and augmented an extensive paired, multimodal dataset of geese (~700mb) for current and future training of models using data processing libraries in Python
- Created scripts for image homography between LFIR and RGB camera modes for paired detection using OpenCV

**Lalmba Association**, Software Engineering Intern, June 2024 - Present

Healthcare nonprofit based in Kenya and Ethiopia

#### Responsibilities:

- Full-stack Android development of malnutrition tracking app
- Consulted with healthcare professionals to center around simple user experience for practitioners in disadvantaged areas

**University Of Pennsylvania Department of Oncology**, Visiting Scholar, 2023

#### Responsibilities:

- Researched the role of computer vision in classifying post-lumpectomy cancer patients by risk of recurrence.

### Research Projects

**Honors Thesis**, 2024 - Present

- Title: Computationally Mitigating Environmental Noise in Multispectral UAV-Based Object Detection
- Abstract: Multispectral object detection has fruitful applications in defense, agriculture, industry, and more. We explore potential improvements of object detection capabilities in unmanned aerial vehicles (UAVs) through computational methods of incorporating environmental variance into multispectral models. Combining forward-looking infrared (FLIR) and visible light (RGB) modalities, we address challenges in object detection accuracy caused by thermal fluctuations and emissivity variations in infrared sensing. We propose a novel adaptation of multispectral YOLO object detection frameworks

that integrate RGB information to circumvent this noise, aiming to outperform traditional RGB- or FLIR-only networks.

**Celltype Heterogeneity Extraction and Encoding in Single-cell Integrated Multimodalities (CHEESGrate), 2023**

- Developed a multimodal variational autoencoder (VAE) to align different omics datasets, such as genomics, transcriptomics, proteomics, and epigenomics in a latent space structured by inputted cell-type information
- Discovered new insights and interpretations of cell-type specific factors in single cell omics data

**CaDance — A fine tuned Running and Listening Experience, 2023**

- Developed a full-stack web application in Typescript utilizing Spotify's API to generate short queues of songs for runners
- Programmed a full-stack Garmin companion application in Monkey C (native Garmin language) to provide certain fields to users
- Collaborated with 3 other developers to integrate components and test the application

**Diagnosing the problem with breast cancer: Unbiasing predictive, patch-based models, 2022**

- Extracted, scored, stain-normalized, filtered, and clustered patches from whole slide images using Python libraries (histolab, etc.)
- Synthesized patches using StyleGan2-ADA to augment data from females of color and attempt to reduce racial bias during predictive model training
- Tuned the existing VGG16 architecture for binary classification of malignant versus benign patches

## **Skills**

**Programming Languages:** JavaScript, TypeScript, HTML, CSS, React, MATLAB, Python, Java

**Web Development Tools:** VSCode, HTTP, OAuth, SQL and other database management

**Version Control:** Git, GitHub

**Cloud Platforms:** AWS, GCP, Oscar

# Jasmine C. Liu

jasmineliu0114@gmail.com — linkedin.com/in/jasmine-liu/

## Education

- **Brown University, Providence, RI** *Sept. 2024 – Present*  
Candidate for Doctor of Philosophy in Computer Science, 2029
- **Northeastern University, Boston, MA** *Sept. 2020 – May 2024*  
Khoury College of Computer Sciences  
Bachelor of Science in Data Science and Mathematics, 2024

## Technical Knowledge

- **Languages:** Python — Java — SQL — R — C++ — ACL2s
- **Systems:** MacOS — Linux — Windows
- **Skills/Tools:** AWS — Git — Excel — Tableau — Redis — MongoDB — Neo4j — PySpark

## Publications

- Yi, Y., Li, Y.Q., Wang, R., Yu, X.F., Liu, Q., Yum, C.H., Szczepanski, A., Li, Q.Q., Fazli, L., Shen, J.C., Wang, X., Liu, J.C., Schaeffer, E.M., Hundley, H.A., Niu, H.Y., Wang, L., Jin, J., Dong, X., Zhao, W., Chen, K.F., Cao, Q. A dual role of EZH2 in regulating A-to-I RNA editing and mRNA stability through ADAR. *Submitted to Nature, under revision.*

## Work and Research Experience

- **Orna Therapeutics, Watertown, MA** *Jul. 2023 – Mar. 2024*  
Data Science Co-op
  - Automate and scale workflows using infrastructure as code methodology (CloudFormation templates) with serverless architectural patterns in AWS to increase efficiency and productivity for research and development.
  - Build a queryable database to organize the large amounts of data being processed throughout the company and make the data more accessible to scientists.
  - Assist the IT team with handling tickets and providing technical support for Orna employees.
- **Joslin Diabetes Center, Harvard Medical School, Boston, MA** *Aug. 2023 – Sept. 2024*  
Intern, Principal Investigator: Yu-Hua Tseng, PhD
  - Analyze single-nucleus RNA sequencing data of human adipose tissue using cell communication packages in R and Python (CellChat, MEBOCOST).
- **Boston Children’s Hospital, Harvard Medical School, Boston, MA** *Jul. – Aug. 2022, May – June 2023*  
Intern, Principal Investigator: Kaifu Chen, PhD

- Performed single-cell RNA-sequencing on breast cancer patient data via Scanpy in Python.
  - Completed accuracy tests to finalize the development of MEBOCOST, a Python package for cell communication.
  - Trained using the Seurat and MAGeCKFlute R packages and converting between Scanpy and Seurat objects.
  - Detected Adenosine-to-Inosine (A-to-I) editing sites from direct-RNA sequencing data using DIno-PORE.
- **Mass General Brigham, Enterprise Research IS, Somerville, MA** *Jan. – Aug. 2022, Dec. 2022 – Jul. 2023*  
Data Analyst Intern
    - Automated the data transfer of inactive users to make available space on the organization's High-Performance Computing (HPC) Linux clusters.
    - Maintained R Shiny apps to keep users up to date on available virtual desktop servers and storage space.
    - Facilitated office hours and online support via Zoom to assist users with HPC Linux cluster issues.
    - Taught introduction training sessions to over 50 physicians and researchers to promote Python and R within the Mass General Brigham community.

**Jeremy L. Warner, MD, MS, FAMIA, FASCO**

Professor of Medicine, Brown University

Professor of Biostatistics, Brown University

Director, Center for Clinical Cancer Informatics and Data Science (CCIDS)

Associate Director of Data Science, Legorreta Cancer Center

Editor-in-Chief, JCO Clinical Cancer Informatics

he/him/his

October 13, 2024

In my role as Chief Software Architect of the HemOnc knowledge base (KB) and Professor of Medicine and Biostatistics at Brown University, I am pleased to collaborate with Brendan Leahey and Yang Xiang on their proposed projects. I confirm that the full HemOnc KB is freely available for academic use under the Creative Commons BY-NC-SA license. I will work with Brendan and Yang to best utilize the KB in their projects.

Additionally, I am required to provide the following disclosure to any student working with HemOnc content:

*I am a founder of HemOnc.org LLC and serve as Chief Technology Officer. I also receive research grant funding from the National Institutes of Health towards enhancing the HemOnc.org website and the HemOnc ontology. This relationship has been identified as having the potential to create a conflict of interest with my responsibilities as a faculty member. I have fully disclosed these interests to Lifespan and have in place an approved plan for managing any potential conflicts arising from this involvement.*

I understand that your work on the project should be for academic reasons to further your studies and your professional career endeavors. If at any time you have concerns about whether your work is inappropriately focused towards my outside relationships, or that your ability to publish has been impeded in any way, I encourage you to contact the Lifespan Office of Research, attention Jacqueline Poore, Manager, Research Compliance Program at 401-444-5843 or [jpoore@lifespan.org](mailto:jpoore@lifespan.org).

Best,

Jeremy

# Syntax-based Contextual Visualizations for Improving SAE & LLM Interpretability

Eric Xia  
PI

`eric_xia@brown.edu`

Brendan Leahey  
Co-PI

`brendan_leahey@brown.edu`

Gonalo Paulo  
Collaborator  
`goncalo@eleuther.ai`

October 14, 2024

## Abstract

Through automated high-quality feature detection, SAEs promise a more complete understanding of how language models function. This paper proposes a novel method for visualizing SAE feature contexts using probabilistic syntactic trees. Through the replacement of linear text-based activation contexts with probabilistic syntax trees, the method simplifies feature comparison and grouping. Among other applications, this visualization enables higher-level investigations of universality, and facilitates the identification of occlusion and oversplitting within SAE training. The resulting visualization of “language-conditional functions” will also be compatible with any set of activating contexts, making it a long-lasting contribution to interpretability beyond specific techniques.

# 1 Response to Reviewers

We appreciate the constructive feedback on the manuscript, and we have prepared a revision accordingly. Please see below for an itemized list of your feedback and a brief description of how each is addressed in the revised manuscript.

**[ChatGPT.1] The experimental nature of the proposed visualization method presents some risk in terms of feasibility and success.**

Response: We acknowledge that there are several failure cases in which visualizations fail to have a measurable impact on interpretability research. However, we believe that there are measurable improvements, such as context scope tagging, re-organizing of contextual text, and activation frequency in relation to linguistic properties which can be developed which do not depend on feasibility of contextual visualization.

**[Laidlaw.1] How will tree knowledge be incorporated into the visualization? Will you augment fig 1 with fig 2? The opposite? Something else? It is also unclear how the claims will be evaluated. What will be done and how it will be evaluated is insufficient to understand if the proposed work will realize any potential significance.**

Response: We aim for the probabilistic syntax tree to be a direct replacement for text contexts. As a fundamental improvement to Anthropic/Neuronpedia’s exploratory feature browser, we envision this enabling new kinds of theories about features. For instance, it should be possible to identify input sequences that don’t seem to activate any features identified with syntactic processing. At minimum, this should serve as an effective diagnosis of sparse coverage in SAEs, but ideally it should also enable insights about the LM’s processes themselves.

**[Laidlaw.2] Related work does not state the relationship of the literature to the proposed work, it seems to only summarize the related work and the conclusions drawn. That is motivational but does not anchor or define the proposed work.**

Response: The revised related work section will make connections between the literature and proposed work clearer. However, the related works will still primarily be defining problems which the proposed work may help with. There are no taxonomically close works except the parent work, Monosemantic Features.

**[Byron.1] mentioning how the results could improve interpretability (how this improved interpretability could help researchers) might improve the scientific contributions of the work.**

Response: The reviewer was correct in pointing out that the work assumed mechanistic interpretability was an inherent contribution. The revised work point to the many practical domains in which interpretability can be useful, including controlling model outputs, ablating bias and discrimination in models, correcting responses by LLMs, making models smaller, faster, and more efficient, and robust guardrails for model behavior.

**[Byron.2] Needs more details for evaluation and for the specific visual implementation details (ie, how is the graph layed out?). Task could also be too computationally intensive for 6 weeks.**

Response: The revised work will contain more comprehensive breakdowns of a range of features, as seen in the slideshow. It will also include mentions of milestones that can be implemented in the event that there is insufficient time to create useful graph visualizations.

**[Jasmine.1] The proposal would benefit from defining evaluation methods of both the probabilistic**

## trees and effect of different design principles.

Response: As stated in response to ChatGPT.2, user studies comparing comprehension of activation contexts as opposed to syntactic visualizations will almost certainly lead to better scores for the latter. The revised work will likely not define evaluation methods for the visualization, but it will focus on user feedback on whether the visualization enables two goals: 1. effective comparison between features in order to identify larger patterns, and 2. grouping of features at a larger scale.

**[Kevin.1] I'm somewhat skeptical that trees will be a big help in visualizing these features. (Are trees really that understandable?) Do we anticipate this to be helpful for all features or a subset of them? What about features which activate on tokens which are not entire words?**

Response: The concerns of the reviewer are valid, and its possible that syntax visualizations will not be useful with the current set of SAE activations. Indeed, Anthropic in their original paper chooses to focus exclusively on token-level features, which identify foreign languages, dates, and other special characters [2]. However, there also exist features extracted which exhibit syntactic patterns in their contexts, and these will likely be able to be captured by the proposed visualization. Because our implementation will be suitable for any set of contextual inputs, we expect syntactic visualizations to be useful in conjunction with other methods of feature extraction.

## 2 Aims



Figure 1: Examples of SAE contextual features and syntax trees at the word, phrase, and sentence level. We propose replacing activating contexts with syntax tree representations of contributions, reducing information load while enabling the critical tasks of feature *comparison* and *grouping*.

This project proposes visualizing feature contexts for language models by syntactic representation, with the goal of facilitating comparisons, making searches more accessible, and introducing formal structure. Features are learned from intermediate layer activations through training sparse autoencoders (also known as SAEs). Due to the large quantity of monosemantic features extracted through an automated process, SAEs are considered the state-of-the-art in understanding how large language models function [15]. Some

features are core notions which exclusively co-occur with single concepts or ideas, but can also be far more complex, e.g. exclusively modifying final tokens of relative clauses across sentences. The varying scope of features extracted – from token to paragraph level – point to the possibility of obtaining a more comprehensive *formal* understanding of the factors leading to a language model’s output.

Consequently, the goal of this project is to investigate the applicability of syntax structures in abstracting over activating contexts for individual SAE features. The critical issue with the current feature visualizations is that it displays list of linear text contexts. While this can work for giving a general notion of a feature, it is not simple, elegant, or readily accessible. Text is not inherently linear, and readers do not treat it as such [9]. Consequently, a list of contexts is far from the ideal format for highlighting the abstract qualities of a feature.

We draw interdisciplinary inspiration from linguistics, where syntax trees are commonly used to reveal

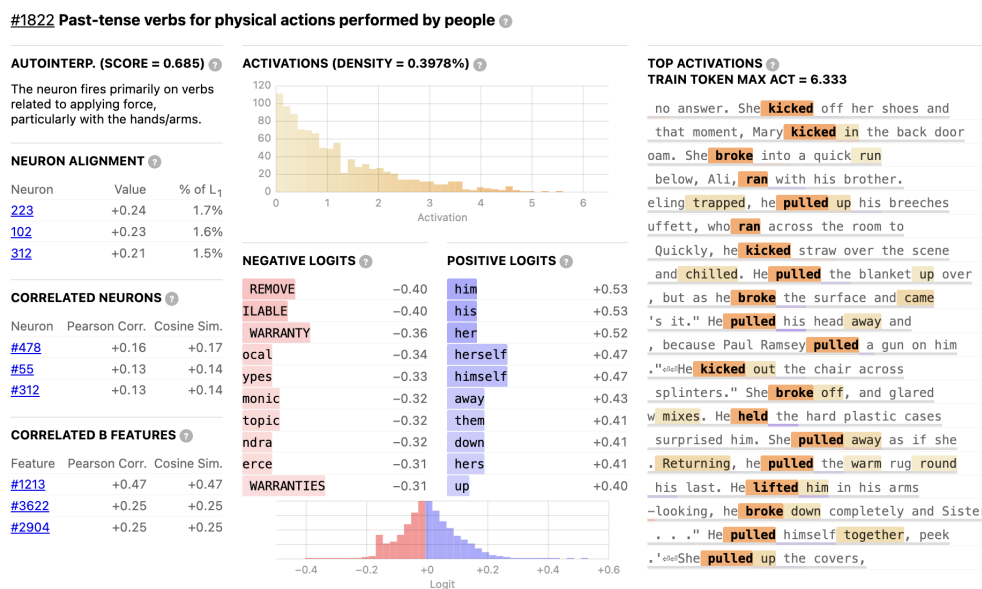


Figure 2: The current activation dashboard to display activations does not effectively display the contexts in which this feature activates. This feature activates only on *past tense physical actions*, and might have an even more restrictive patterns over contexts which are hard to see.

deeper internal structure. We plan to explore this paradigm applied to activation contexts. In order to do this quickly and effectively, we will employ traditional natural language processing on contexts to generate parse trees, and combine them into a probabilistic model. These trees, as aggregate representations of feature contexts, lend themselves to comparison and grouping methods.

### 3 Significance

In recent years, large language-based models such as ChatGPT have undergone an explosion in interest and applications[5]. A large part of the success of these models have been due to their emergent ability to utilize contextual information, and output coherent and structured language. Yet due to their complex nonlinear architecture, it is still unclear how they achieve either of these capabilities. The field of mechanistic inter-

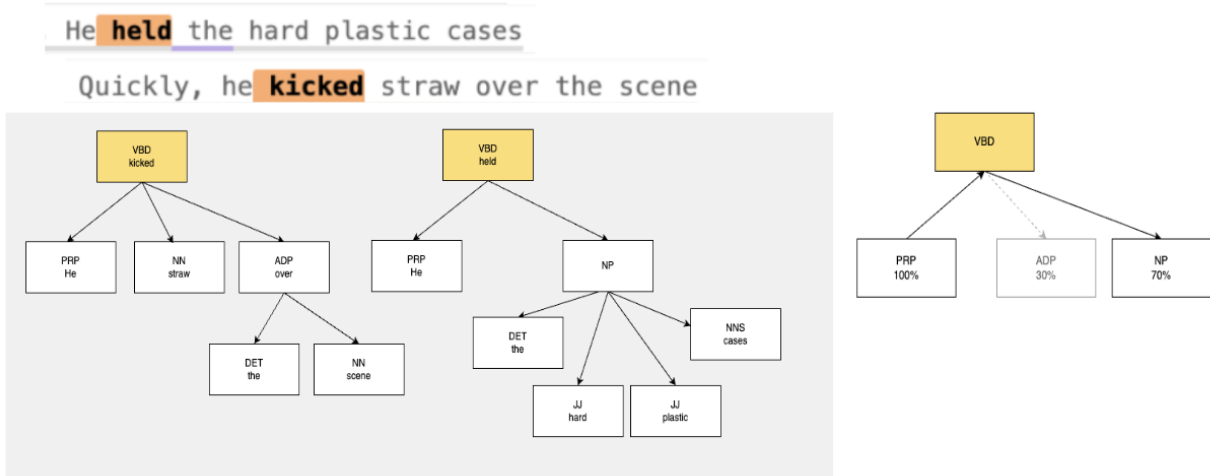


Figure 3: As pictured above, a single probabilistic tree of common activating contexts could greatly simplify the representation of any particular feature, e.g. for *past tense physical actions*. We envision this unified representation replacing the textual activating contexts on the right of Figure 1.

pretability strives for a precise understanding of model behavior through identifying mechanisms carried out by neural networks. This is crucial to developing models which are trustworthy, beneficial and effective [1]. To date, the field has led to the development of a variety of demonstrably useful techniques such as correcting factual recall through activation patching [14], steering model outputs through task vectors [7], and mitigating bias from training data [3].

One recent approach to mechanistic interpretability has been training sparse autoencoders on transformer weights in order to learn features [2]. This automated feature discovery uncovers an exceptionally large number of features with context-specific characteristics. However, SAE feature dashboards lack the capabilities to identify syntactic patterning, while LLMs clearly learn syntax to a degree [10]. Additionally, the sheer number of features learned by SAEs has consistently been a critical issue with their interpretability [15]. An appropriate visualization with two critical properties, *comparison* and *grouping*, would enable a taxonomic model of transformer features.

Through identifying consistently activating contexts, visualization can lead to clearer evaluation of the degree of monosemanticity for particular features. As stated in Anthropic’s SAE decomposition of Claude 3 Sonnet, the current optimization for reconstruction accuracy and sparsity is only a proxy for what we really are interested in, interpretable features [15].

As a new kind of data, feature-specific activating contexts lend themselves to novel visualization research. Probabilistic, usage-based models of language such as exemplar theory have accumulated a large body of supporting evidence in recent years [4]. This raises the question of how probabilistic notions of linguistic structure be most effectively conveyed. Specifically, in the context of large language models, an interesting problem is how the functional contribution of many features could be aggregated visually over a sentence. We plan to explore this question by experimenting with formal design principles, including continuity, proximity, color, and contrast.

## 4 Related Work

Our work relates to current research being published in the mechanistic field. Many of these are announcements of SAE feature and contexts being trained on larger and more capable models, such as Gemma Scope [12] and Claude 3.5 Sonnet [15]. However, establishing monosemantic features only lay the foundations for interpretability research [2]. Crucial to utilizing features in interpretability is a way to understand the roles features serve within models. Our visualization method provides two distinct advantages over existing approaches: the ability to *compare* and *group* features.

One key concept within interpretability is the universality of features. More precisely, this is the notion that features identified for one model can be *compared* across others, and some can be generalized. Work has been done on universality of neurons across GPT-2 models [6]. This work was able to identify consistently activating neurons on punctuation, date, and medical terms. Our approach would enable researchers visually *compare* decompositions coming from several SAEs for higher-level features. For example, it might lead to the identification of a phenomenon where all language models implement a feature for keeping track of a temporal deictic center, as utilized in expressions such as *yesterday*, *last week*, and *tomorrow*, but only some have this feature within relative clauses.

Other papers in interpretability which utilize SAE features do so through identifying groups of features which work together. Over the Indirect Object Identification task for GPT-2, both supervised and unsupervised dictionary learning were employed, where unsupervised learning is equivalent to SAE reconstruction. This comparison led to the identification of issues with feature occlusion and over-splitting [13]. Through syntax visualization, features could be *grouped* by shared modifying contexts. This would lead to increased understanding of the specific kinds of features which exhibit occlusion and over-splitting in SAE training.

Finally, the probabilistic analysis of feature contexts draws on successful approaches in generative linguistics to syntax. Hidden Markov Models were used effectively in early approaches to Part-of-Speech tagging [11]. Probabilistic context-free grammars designate a probability distribution over possible derivations[8]. SAE features differ from probabilistic grammars in providing an *inverse problem*: given empirical data of feature activations, the goal is to elucidate an underlying model for the feature.

## 5 Research Plan

Week 1: Data analysis and intake. Activation, context, and explanatory data have been sourced and time spent looking over it. The collaborator will work with us to identify the plausibility of the approach and related avenues of research.

Week 2: Prototype. A basic visualization has been created for the data at hand, with some novel features added. These will include tagging features by context scope (e.g. token/clause/sentence level), the reorganization of activation contexts, and activation frequency in relation to part-of-speech. This serves as an achievable and measurable initial contribution, while enabling the construction of a tree view as below.

Week 3: We rapidly iterate on the project. At this stage, a tree view for custom user inputs should be fully developed, and some associated features tagged and linked in the overall UMAP visualization.

Week 4: We continue to develop, with the goal of getting feedback from external users, such as researchers using Neuronpedia.

Week 5: Finalize iteration and prepare the writeup. At this stage, several novel visualization approaches have been attempted and their successes and failures characterized. Promising directions for further work have been identified.

Week 6: At this stage, we will prepare a final abstract and presentation.

## References

- [1] L. Bereska and E. Gavves. Mechanistic interpretability for ai safety—a review. *arXiv preprint arXiv:2404.14082*, 2024.
- [2] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, and C. Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- [3] S. Dai, C. Xu, S. Xu, L. Pang, Z. Dong, and J. Xu. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6437–6447, New York, NY, USA, 2024. Association for Computing Machinery.
- [4] N. C. Ellis, M. B. O'Donnell, and U. Römer. Does language zipf right along? *Georgetown University Round Table on Languages and Linguistics*, pages 33–50, 2014.
- [5] Goldman Sachs Insights. Gen ai: Too much spend, too little benefit? <https://www.goldmansachs.com/insights/top-of-mind/gen-ai-too-much-spend-too-little-benefit>, 2024. Accessed: 2024-10-13.
- [6] W. Gurnee, T. Horsley, Z. C. Guo, T. R. Kheirkhah, Q. Sun, W. Hathaway, N. Nanda, and D. Bertsimas. Universal neurons in gpt2 language models. *arXiv preprint arXiv:2401.12181*, 2024.
- [7] G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi. Editing models with task arithmetic, 2023.
- [8] F. Jelinek, J. D. Lafferty, and R. L. Mercer. *Basic methods of probabilistic context free grammars*. Springer, 1992.
- [9] E. Kaiser. Experimental paradigms in psycholinguistics. *Research methods in linguistics*, pages 135–168, 2013.
- [10] A. Kulmizev and J. Nivre. Schrodinger’s tree — on syntax and neural language models. *Frontiers in Artificial Intelligence*, 5:796788, 2022.
- [11] J. Kupiec. Robust part-of-speech tagging using a hidden markov model. *Computer speech & language*, 6(3):225–242, 1992.
- [12] T. Lieberum, S. Rajamanoharan, A. Conmy, L. Smith, N. Sonnerat, V. Varma, J. Kramár, A. Dragan, R. Shah, and N. Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.
- [13] A. Makelov, G. Lange, and N. Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. *arXiv preprint arXiv:2405.08366*, 2024.
- [14] K. Meng, A. S. Sharma, A. Andonian, Y. Belinkov, and D. Bau. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022.

- [15] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, and T. Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

## EDUCATION

**Brown University**, *Math-Computer Science, Linguistics*

**Providence, RI** 2022 - 2026

*Selected coursework: Deep Learning, Computational Linguistics, Abstract Algebra.* Brown Linguistics department undergraduate leader, symposium organizer. Language Understanding and Representation Lab guest presenter. CSA Family Head, Outdoor Leadership Training, Brown Opinion Project Research.

## WORK :)

**UCCS NLP REU**, Researcher

**Colorado Springs, CO** 2024

One of ten undergraduates nationwide selected for a ten-week NSF-funded research program. Full-time independent research, including literature review, experimentation, analysis, and authoring publications. Guest presentation on sinusoidal encoding of relative attention in transformers.

**Brown University Library**, GIS & Data Assistant

2023-

Created series of scripting and geoprocessing tutorials, including compositing satellite imagery with Google Earth Engine's Python API & Colab. Facilitator of introductory GIS workshops for professional researchers. Performed high volume queries on commercial geospatial data.

**Northshore Utility District**, Seasonal Utility Worker

2023

**NORC at the University of Chicago**, Field Interviewer

2022

Conducted comprehensive social science surveys for the General Social Survey

## PROJECTS

**Linear Decoding of Morphology Relations in Language Models** (ICLR submission)

2024

Over morphological relations in large language models, we find that multiplying a base form hidden state by model derivatives faithfully approximates final hidden states of derived forms, providing the first evidence of a truly linear relational embedding being implemented by models.

**Understanding Arctic Sea Ice Melt with Airborne Observations**

2024

Undergraduate research with the School of Engineering and Thermal Sciences, rewrote algorithm for matching LiDAR to ICESAT2 altimetry. Through an automated window matching computer vision pipeline, improved manual accuracy by 15% while reducing processing time by 90%.

**Terraformation with Pix2Pix and Satellite-Elevation Image Pairs**

2023

Implemented Pix2Pix and U-Net architecture for terraformation of extraterrestrial imagery.

**word.golf**

2021

Created **word.golf**, an online sport played with word vectors. Using semantic connections between words, players shift through the English language through nearest neighbors. I received a grant from **Emergent Ventures**, a fellowship program that supports entrepreneurs with highly scalable ideas for meaningfully improving society.

**Recurse Center** (Hacker School)

**Brooklyn, NY** S'1 2022

**1517 Fund Medici Grantee**

2020-

**Grand Prize, DefHacks**

2020

**Best Use of SnapKit, DefHacks**

2019

## +SKILLS

Python, Pytorch, Pandas, Flask, NumPy, Huggingface, Baukit, Docker, R, QGIS, Technical Writing, JavaScript, Google Earth Engine, Illustrator, Photoshop

# Gonçalo Santos Paulo



## About me

I've always wanted to be a scientist. I love learning about a problem, brainstorm about solutions and discuss with other people my ideas. I've always enjoyed learning, and I think that helps me be a good discussion partner.

## personal

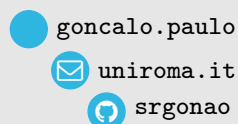
Gonçalo Santos Paulo  
Nationality: Portuguese  
Date of birth: 25/08/1997

## Areas of specialization

Condensed matter Physics  
• Computational Physics

## Interests

I'm very interested in AI safety.  
I'm also very sociable and active around people.



## SHORT RESUMÉ

### 2024 Interpretability research Intern

• EleutherAI

Currently researching interpretability methods using Sparse Autoencoders, mainly using automated interpretability tools.



### 2023–2024 PostDoc

• Sapienza university of Rome

Research focused on memristive behaviour due to hydrophobic gating, nanofluidics, and intrusion in hydrophobic materials.



## DEGREES

### 2023 Theoretical and Applied Mechanics

SAPIENZA UNIVERSITY OF ROME



### 2020 Physics

MASTER DEGREE  
Faculty of Science  
University of Lisbon



### 2018 Physics

BACHELORS DEGREE  
Faculty of Science  
University of Lisbon



## PROGRAMMING

L<sup>A</sup>T<sub>E</sub>X

python

Processing

## CERTIFICATES & GRANTS

2022 Grant of computer resources on the Italian SuperComputing Resource Allocation (IRCA-C project)

## LANGUAGES

Portuguese  
English  
Italian

C2 mother tongue  
C2  
B1

## TALKS

March 2022 "Building an artificial neuron with simple hydrophobic nanopores: a one component memristor", at: American Physical Society online.

## PUBLICATIONS

- 2024 Does Transformer Interpretability Transfer to RNNs?, ArXiv.
- 2023 Hydrophobically gated memristive nanopores for neuromorphic applications, Nature Communications.
- 2023 The impact of secondary channels on the wetting properties of interconnected hydrophobic nanopores, Communications Physics.
- 2023 Optimization of the Wetting-Drying Characteristics of Hydrophobic Metal Organic Frameworks via Crystallite Size: The Role of Hydrogen Bonding between Intruded and Bulk Liquid, Journal of Colloid and Interface Science
- 2023 An atomistically informed multi-scale approach to the intrusion and extrusion in hydrophobic nanopores, Journal of Chemical Physics

Gonçalo Paulo ✉ goncalo.paulo@uniroma1.it 📍 Rome

# Visualizing Measures of Uncertainty in Population Genomics

Jasmine Liu

PI

`jasmine_c_liu@brown.edu`

Richard Huang

Co-PI

`richard_huang2@brown.edu`

Alex Diaz-Papkovich

Collaborator

`alex_diaz-papkovich@brown.edu`

Sohini Ramachandram

Collaborator

`sramachandran@brown.edu`

October 13, 2024

## Abstract

We present a proposal to develop visualization tools for analyzing population genomic data, with a focus on highlighting shared genetic variation while minimizing the risk of misinterpreting such genomic visualizations. We will integrate measures of uncertainty into uniform manifold approximation and projection (UMAP) plots, allowing researchers to explore complex patterns of genetic variation, helping prevent the misuse of genetic data and promoting a more nuanced understanding of population genomics. Through a small-scale user study, we will evaluate various methods of visually encoding measures of uncertainty.

# 1 Reviews and Responses

Dear Editor and Reviewers,

We appreciate the constructive feedback on the manuscript, and we have prepared a revision accordingly. Please see below for an itemized list of your feedback and a brief description of how each is addressed in the revised manuscript.

## [R1] Brendan Leahey

### Overall: 2

[R1.1] Interdisciplinary: 1; strong relation to genomics, sociology, and visualization with potential applications to other visualization domains.

[R1.2] Scientific: 2; Strong explanation of scientific background and how visualization techniques improve understanding of genetic data.

[R1.3] Visualization: 2; uncertainty visualization is novel. clearly explains the impact of visualization techniques and how they tie into the scientific goals. Elaboration on methods like blurring and examples could be helpful (presentation provides greater detail), but understandably difficult to define this early on. Also unclear how users may evaluate these techniques.

*Response: We have elaborated on different methods of blur and further explained what sorts of qualitative and quantitative measures we will obtain from the user study in the Aims section of the proposal.*

[R1.4] Significant: 2; clearly defines the significance of the work in population genomics and visualization techniques. High potential for contributions towards other visualization tasks that can apply uncertainty for clarity is relatively underexplored, but also fall out of the scope of a 6 week project. elaborating on something like this in a discussion section of the final paper could be beneficial.

*Response: In the Significance section, we elaborated more on the broader impacts of the project, including the possible application of these methods to other fields and data types.*

[R1.5] Novel: 3; Strong overall case study in multiple visualization techniques and how they can be applied to population genomics. Extending this to other domains can improve novelty. Further, clearly defining the uncertainty visualization techniques will give a better sense of how this work is novel.

[R1.6] Goals clearly stated: 3; Tools to display uncertainty visualization can use a bit of elaboration. Overall goals and impact of outcomes are clearly defined.

*Response: See response to R1.3.*

[R1.7] Likelihood of Success: 2; clearly defined methods and timeline. user study may be unrealistic depending on the number of participants and if previous work stays true to the timeline.

[R1.8] Strengths:

- strong interdisciplinary connections
- demonstrates strong understanding of scientific context
- achievable with clear timeline

[R1.9] Weaknesses:

- without a strong discussion section or attempt at a more generalized framework, this may fall short on novelty
- user study may be unrealistic given the timeline

- could use more detail on the uncertainty visualization techniques and how they may be explored and evaluated

**Response:** *To adjust to the 6-week timeframe, we will conduct a small-scale user study.*

**[R1.10]** Other comments for discussion:

- would be cool to see demonstrations of blurring etc if you have any!

**Response:** *We have added relevant figures from other papers, as well as our own preliminary sketches.*

## **[R2] Kevin Wang**

**Overall: 2**

**[R2.1]** Interdisciplinary: 1, The project combines genomics and visualization.

**[R2.2]** Scientific: 3, This project will produce better genomic visualizations, while reducing potential for misinformation, which will help with scientific presentation and communication of genomic research.

**[R2.3]** Visualization: 3, Visualizing uncertainty and effective visualization without miscommunication are visualization challenges.

**[R2.4]** Significant: 2, As above, better genomic visualization is significant, as is visualization of uncertainty.

**[R2.5]** Novel: 2, Uncertainty for genomic data seems novel.

**[R2.6]** Goals clearly stated: 1, The goals are clearly stated in the Aims section: creation of blurred plots to represent uncertainty.

**[R2.7]** Likelihood of Success: 2, The 6-week plan looks feasible.

**[R2.8]** Strengths:

- Good research plan, concise and limited in scope.
- Proposal is easy to read and clear.

**[R2.9]** Weaknesses:

- It's not immediately clear if and how much uncertainty visualization will accomplish the goal of reducing misinformation. Visuals would have been helpful in the proposal.

**Response:** *See response to R1.10. We have also included more about previous research on uncertainty visualizations in the Related Works section that describe the possible benefits of using such methods.*

## **[R3] Kei Yoshida**

**Overall: 3**

**[R3.1]** Interdisciplinary: 1. The proposal is interdisciplinary as it incorporates computer science (visualization) and genomics, while borrowing ideas from other fields (dimensionality reduction techniques).

**[R3.2]** Scientific: 3. With the tool, researchers would be able to explore complex patterns of genetic variation, and this has a potential to advance the field. The score would improve if it was how this exactly would be achieved – while there are some descriptions of the visualization, the explanation could be clearer.

**Response:** *We have elaborated more on the scientific contributions of the project in the Significance section.*

**[R3.3]** Visualization: 1. The visualization tool has many implications; (1) researchers would be able to explore complex patterns of genetic variation, (2) it would prevent the misuse of genetic data, and (3) it would promote a more nuanced understanding of population genomics.

**[R3.4]** Significant: 3. It is clear that many people would benefit from it. But as explained in "Novel" section, it is not very clear and convincing what this tool adds to the existing frameworks.

**Response:** *See the response to R3.5.*

**[R3.5]** Novel: 7. It is not very clear what exactly is novel. The related work section somewhat explains it, but Sun et al. already have a framework that incorporates uncertainty. The novel aspect that was not included in their work (blurred plots?) should be more explicitly explained and emphasized.

**Response:** *To address the novelty of the proposal, we have made more of a distinction of what the limitations of DynamicViz are and how we are building on the methods of DynamicViz using blurred plots.*

**[R3.6]** Goals clearly stated: 1. The goal is to integrate measures of uncertainty into visualizations to provide clarity on the biological significance of clusters, which could help researchers to understand the data and the public to avoid misinterpreting.

**[R3.7]** Likelihood of Success: 5. This score would significantly improve if the novelty aspect was made more explicit, and the research plan was more detailed.

**Response:** *We have added more detail to each week of the Research Plan, including goals/deliverables for each step.*

**[R3.8]** Strengths:

- Very clear benefits to the research community the society, which creates a high motivation for the project.
- Good explanations of the techniques that you plan to use (dimensionality reduction).
- Risks of visualization misinterpretation is well described.
- Good description of planned evaluation.

**[R3.9]** Weaknesses:

- Novelty (the gap in the existing work) is not well explained.
- More related work that leads to this work should be cited to motivate the proposed project.

**Response:** *See responses to R2.9 and R3.5.*

## 2 Aims

We propose to develop population genomic visualizations that focus on highlighting shared genetic variation while avoiding misinterpretation of population differences. To do so, we aim to achieve the following goals:

1. Integrating measures of uncertainty into visualizations to provide clarity on the biological context of clusters.
2. Evaluating the effectiveness and robustness of the developed visualizations.

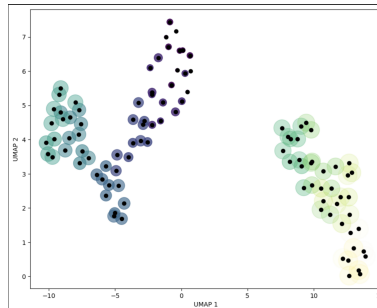


Figure 1: Preliminary sketch of visualizing uncertainty measures on randomly generated data. The color indicates the label of each data point, while the size and opacity reflect the variation of each point after bootstrapping.

We aim to integrate measures of uncertainty into genomic UMAP plots to provide clarity on the biological context of clusters. To do so, we will bootstrap over genetic markers to generate multiple resamples of the data and apply UMAP dimensionality reduction to each resample. This will create several different embeddings for the same dataset. For each data point, its location in the reduced space will vary slightly between embeddings, depending on the variability in the original data. Our methods will use the multiple embeddings generated through bootstrapping to calculate the variation of each data point’s position. This allows for the visual encoding of these uncertainty measures as blur on the dimensionality reduction plots. We aim to use multiple visual encoding methods, including varying degrees of transparency, fuzziness, size, density contours, or a combination of techniques using the matplotlib and seaborn packages in Python (Figure 1). We will then compare the efficacy of each method with each other, as well as with UMAP plots with no measures of uncertainty. In the context of genomic studies, we believe bootstrapped blurred plots will help to emphasize both the overall genetic similarity of populations and the underlying genetic uncertainties. They offer a powerful visual tool for interpreting complex patterns of genetic variation, showing not only the genetic similarities among specific populations but also the fuzziness that arises from migration, admixture, and other evolutionary processes.

To evaluate the effectiveness of our methods, we will conduct a small-scale, task-oriented user study to obtain quantitative and qualitative measurements of how effective the measures of uncertainty in our visualizations are for reducing misinterpretation. Our user group will include experts in the field, including our collaborators Dr. Alex Diaz-Papkovich and Dr. Sohini Ramachandran, as well as some non-experts, such as our fellow classmates. As a baseline, we will use UMAP plots of genomic data without any measures of uncertainty. As a secondary baseline, we will compare our methods to the visualizations produced by DynamicViz [6], which creates animated and stacked plots from bootstrapping. This user study will include

tasks such as asking participants to purposefully misinterpret the plots. We will then take measurements such as task completion time and accuracy, as well as user-provided scores and feedback.

### 3 Significance

Population genetics plays a crucial role in understanding human ancestry, disease risk, and evolutionary patterns. However, current visualization techniques, such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and UMAP, often overstate population differences. These visualizations are susceptible to misinterpretation and have been misused to promote harmful narratives about biological racial differences. This project addresses these challenges by developing visualizations that make population genetic data more interpretable and inclusive.

By integrating uncertainty into dimensionality reduction visualizations, this project will offer a more accurate and reliable representation of genetic variation. Researchers will be able to differentiate between stable, biologically meaningful clusters and those influenced by data variability. This will not only improve the robustness of genomic analyses but also mitigate the risk of oversimplifying complex genetic relationships, therefore, counteracting harmful narratives that falsely emphasize biological racial differences.

Additionally, by comparing different methods of visualizing these measures of uncertainty, we will further visualization research by identifying effective methods of visually encoding measures of uncertainty in dimensionality reduction plots. Furthermore, these methods can extend to other fields and data types, broadening the impact across multiple domains of research.

### 4 Related Work

Recent advancements in dimensionality reduction techniques have significantly impacted population genetics research, allowing for the visualization of high-dimensional genomic data. UMAP [5], in particular, has been adopted due to its ability to preserve local relationships between data points, making it well-suited for uncovering fine-scale genetic structures. Diaz-Papkovich et al.’s research applied UMAP to explore cryptic population structures and phenotype heterogeneity in datasets such as the 1000 Genomes Project, UK Biobank, and Health and Retirement Study [2]. They found that UMAP can reveal previously undetected subpopulations and fine-scale relationships between genetic variation, geography, and phenotypes. This approach provides an important framework for identifying subtle demographic and genetic patterns that might otherwise be overlooked. Building upon this, Diaz-Papkovich et al. further reviewed the use of UMAP in population genetics, emphasizing its effectiveness in visualizing ancestral composition and subtle genetic structures within large datasets [3].

However, while methods like UMAP effectively cluster high-dimensional data, they often distort global genetic structures, leading to potential misinterpretations of population differences. Previously, Sun et al. introduced DynamicViz, a tool used to assess the robustness of dimensionality reduction visualizations. DynamicViz uses similar bootstrapping methods to visualize dimensionality reduction plots in either an interactive, animated, or stacked visualization [6]. Their approach emphasizes the need for dynamic, interactive tools to explore the variability and uncertainty inherent in dimensionality reduction techniques. Although these methods highlight uncertainty in dimensionality reduction techniques, they do not visually represent uncertainty directly and become more challenging to interpret as datasets increase in size.

Building upon these methods, the need to effectively communicate uncertainty in scientific visualizations has been widely acknowledged [4]. A comprehensive review by Bonneau et al. highlights methods of

visualizing uncertainty, like attribute modification, where uncertainty is conveyed through adjustments in visual properties such as color, opacity, and size [1].

Our proposal builds on these concepts by not only integrating uncertainty specifically into genomic visualizations but also by visually encoding these measures of uncertainty through the concept of blurred plots. Additionally, we will compare different visual encoding methods, allowing us to determine what visualization techniques are most effective in this regard. This work directly addresses the challenges posed by dimensionality reduction distortions, reducing the misinterpretation of population differences while offering a more comprehensive understanding of the groupings that manifest from these techniques.

## **5 Research Plan**

### **Week 1**

Get familiarized with the biobank datasets and begin performing dimensionality reduction using methods such as PCA and UMAP. Create plots using DynamicViz. (Deliverables: Baseline UMAP plots without uncertainty and animated/stacked DynamicViz plots.)

### **Week 2**

Perform bootstrapping and dimensionality reduction and calculate variance score for each point. Begin trying out visually encoding uncertainty with opacity and size. (Deliverables: Preliminary plots with uncertainty represented through opacity and/or size.)

### **Week 3**

Visual encoding uncertainty through clouds and density contours while fine-tuning previous week's methods. (Deliverables: More uncertainty plots.)

### **Week 4**

Continue fine-tuning methods. Define tasks and questions for the small-scale user study. (Deliverable: Questionnaire for user study.)

### **Week 5**

Conduct task-oriented user study to evaluate the tool's effectiveness in conveying uncertainty. Gather feedback and refine the tools based on results. (Deliverables: User study results.)

### **Week 6**

Finalize the visualizations based on feedback from testing and complete the final report. (Deliverables: Final report and presentation.)

## References

- [1] G.-P. Bonneau, H.-C. Hege, C. R. Johnson, M. M. Oliveira, K. Potter, P. Rheingans, and T. Schultz. Overview and state-of-the-art of uncertainty visualization. *Scientific visualization: Uncertainty, multi-field, biomedical, and scalable visualization*, pages 3–27, 2014.
- [2] A. Diaz-Papkovich, L. Anderson-Trocmé, C. Ben-Eghan, and S. Gravel. Umap reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS genetics*, 15(11):e1008432, 2019.
- [3] A. Diaz-Papkovich, L. Anderson-Trocmé, and S. Gravel. A review of umap in population genetics. *Journal of Human Genetics*, 66(1):85–91, 2021.
- [4] C. Johnson. Top scientific visualization research problems. *IEEE Computer Graphics and Applications*, 24(4):13–17, 2004.
- [5] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- [6] E. D. Sun, R. Ma, and J. Zou. Dynamic visualization of high-dimensional data. *Nature Computational Science*, 3(1):86–100, 2023.

# Jasmine C. Liu

(781)-492-7502 | [jasmineliu0114@gmail.com](mailto:jasmineliu0114@gmail.com) | [linkedin.com/in/jasmine-liu/](https://www.linkedin.com/in/jasmine-liu/)

## Education

---

**Brown University**, Providence, RI Sept. 2024 – Present  
*Candidate for Doctor of Philosophy in Computer Science, 2029*

**Northeastern University**, Boston, MA Sept. 2020 – May 2024  
**Khoury College of Computer Sciences**  
*Bachelor of Science in Data Science and Mathematics, 2024*

## Technical Knowledge

---

Languages: Python | Java | SQL | R | C++ | ACL2s  
Systems: MacOS | Linux | Windows  
Skills/Tools: AWS | Git | Excel | Tableau | Redis | MongoDB | Neo4j | PySpark

## Publications

---

Yi, Y., Li, Y.Q., Wang, R., Yu, X.F., Liu, Q., Yum, C.H., Szczepanski, A., Li, Q.Q., Fazli, L., Shen, J.C., Wang, X., **Liu, J.C.**, Schaeffer, E.M., Hundley, H.A., Niu, H.Y., Wang, L., Jin, J., Dong, X., Zhao, W., Chen, K.F., Cao, Q. A dual role of EZH2 in regulating A-to-I RNA editing and mRNA stability through ADAR. *Submitted to Nature, under revision.*

## Work and Research Experience

---

**Orna Therapeutics**, Watertown, MA Jul. 2023 – Mar. 2024  
Data Science Co-op

- Automate and scale workflows using infrastructure as code methodology (CloudFormation templates) with serverless architectural patterns in AWS to increase efficiency and productivity for research and development.
- Build a queryable database to organize the large amounts of data being processed throughout the company and make the data more accessible to scientists.
- Assist the IT team with handling tickets and providing technical support for Orna employees.

**Joslin Diabetes Center, Harvard Medical School**, Boston, MA Aug. 2023 – Sept. 2024  
Intern, *Principal Investigator: Yu-Hua Tseng, PhD*

- Analyze single-nucleus RNA sequencing data of human adipose tissue using cell communication packages in R and Python (CellChat, MEBOCOST).

**Boston Children's Hospital, Harvard Medical School**, Boston, MA Jul. – Aug. 2022, May – June 2023

Intern, *Principal Investigator: Kaifu Chen, PhD*

- Performed single-cell RNA-sequencing on breast cancer patient data via Scanpy in Python.
- Completed accuracy tests to finalize the development of MEBOCOST, a Python package for cell communication.

- Trained using the Seurat and MAGeCKFlute R packages and converting between Scanpy and Seurat objects.
- Detected Adenosine-to-Inosine (A-to-I) editing sites from direct-RNA sequencing data using DInoPORE.

**Mass General Brigham, Enterprise Research IS**, Somerville, MA      Jan. – Aug. 2022, Dec. 2022 – Jul. 2023

Data Analyst Intern

- Automated the data transfer of inactive users to make available space on the organization's High-Performance Computing (HPC) Linux clusters.
- Maintained R Shiny apps to keep users up to date on available virtual desktop servers and storage space.
- Facilitated office hours and online support via Zoom to assist users with HPC Linux cluster issues.
- Taught introduction training sessions to over 50 physicians and researchers to promote Python and R within the Mass General Brigham community.



**Diaz-Papkovich, Alex** <alex\_diaz-papkovich@brown.edu>

to me ▾

Wed, Sep 11, 3:10 PM

Hi Jasmine,

Here's a link to my full thesis: <https://escholarship.mcgill.ca/concern/theses/qn59q986v>

It's made up of three papers---if you'd like to see the lions of the individually, see:

\* <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008432>

\* <https://www.nature.com/articles/s10038-020-00851-4>

\* <https://www.biorxiv.org/content/10.1101/2023.07.06.548007.abstract>

If you're interested in Lior Pachter's paper on single-cell genomics:

\* <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011288>

(For what it's worth it has generated a response paper here: <https://www.biorxiv.org/content/10.1101/2024.03.26.586728v2.abstract>)

Thanks,  
Alex.

# Response to Reviewers – Loss Landscape Visualization

Kevin A. Wang

PI

kawang@brown.edu

Arjun Prakash

Co-PI

arjun\_prakash@brown.edu

Randall Balestrieri

Collaborator

randall\_balestrieri@brown.edu

October 14, 2024

Dear Reviewers,

We thank you for your time, and we appreciate all of your thoughtful feedback. We agree with many of your suggestions on how to improve our preliminary proposal. We incorporated these improvements and the final proposal is much stronger as a result.

In the following pages, we responded to each piece of feedback in turn.

The final proposal follows.

At the end is a note from our collaborator (Professor Randall Balestrieri) stating his participation, and a copy of my C.V.

-Kevin Wang

Summary of major changes:

Added a new section (Intro and Background) to give context and an example of loss landscape visualization.

Added a new section (Evaluation).

Updated research plan to be scaled down.

# 1 Secondary Reviewer (Jasmine Liu)

Overall: 2.7

Interdisciplinary: 3, The project combines deep neural networks with visualizations.

**Response:** Indeed. To emphasize the interdisciplinary content of the proposal, I will add a section to the proposal that explicitly states the two fields being bridged (visualization and geometric deep learning).

Scientific: 3, The proposal has clear scientific contributions toward analyzing model behavior and optimizing neural networks.

**Response:** Indeed. And I will add even more specific examples to the Deep Learning Significance section about how visualizations help deep learning researchers (including concrete examples from the collaborator).

Visualization: 2, Producing faster visualizations through adaptive sampling and visualizing high-dimensional data through 2-D slices are significant visualization contributions.

**Response:** Thank you for the positive feedback.

Significant: 3, The work is significant, as it has the potential to improve the efficiency and effectiveness of deep learning research.

Novel: 4, This project improves upon previous work on analyzing loss landscapes.

**Response:** I will make the Related Works section (and the beginning of the paper) more explicit about how our work differs from existing visualization tools.

Goals clearly stated: 1, Each of the three goals are clearly stated and well-defined.

**Response:** Thank you for the positive feedback.

Likelihood of Success: 3, The research plan and goals are well-defined but may be a little ambitious for the timeframe of the project.

**Response:** Good point. Nailing all three contributions may be ambitious. I will edit the research plan to include a “safety” plan wherein we focus on a single contribution, if progress is slow.

Strengths:

- Goals are clear.
- Methods for carrying out each goal are well-defined.

Weaknesses:

- Lacks evaluation methods.

**Response:** Good point. I will add a section on evaluation methods. We will design an evaluation using our collaborator to measure his performance on a task such as picking good neural network architectures. We can measure accuracy, speed, and subjective preference on the task.

- Would be helpful to have baseline or preliminary figures, especially for those who are unfamiliar with loss landscapes.

**Response:** Great point. I will add figures throughout the proposal. I think they will help a lot.

## 2 Tertiary Reviewer 1 (Kei Yoshida)

Overall: 3

Interdisciplinary: 8. Although it can be inferred, it is not explicitly explained how interdisciplinary this work is outside of computer science. This could be improved by adding the implications (e.g., how this novel tool can be used by researchers in other fields).

**Response:** In my view, the project is not interdisciplinary because it connects CS with a non-CS discipline, but because it bridges the two fields of visualization and geometric deep learning. I will add a short section to the proposal to make this point clear.

Scientific: 1. The scientific contribution is very clear and well-explained (what researchers can do with this visualization, and how their workflow would be better).

**Response:** Thank you for the positive feedback.

Visualization: 4. The visualization contribution could be clearer; this would be improved by describing the existing work and what the proposal adds to it.

**Response:** Good point. I will investigate what previous work has been done on visualizing very high dimensional scalar fields. I'll add concrete examples to the Visualization Significance section (Section 4.3), to connect our proposal with other visualization research.

Significant: 1. The contributions are very clearly stated: (1) easier integration into existing workflows, (2) faster visualizations, and (3) the ability to visualize custom 2D slices.

**Response:** Thank you for the positive feedback.

Novel: 3. A NOVEL tool for visualizing loss landscapes. The score would be improved by explaining the novelty in more detail in the related work section – what currently exists and doesn't exist.

**Response:** Great point. I will flesh out the Related Works section with a paragraph about the seminal Li, et al. paper [4], including at least one figure. I will then add a sentence or two to clarify that all three of our contributions are novel compared to existing work.

Goals clearly stated: 1. The goals are explicitly stated as contributions.

**Response:** Thank you for the positive feedback.

Likelihood of Success: 3. Clear contributions make this project likely to succeed. This can be improved by making the research plan more detailed with clear milestones.

**Response:** Good point – my research plan wasn't very precise. I will add explicit milestones.

Strengths:

- Very clear contributions with explicit novelty – this also provides clear goals of the proposal and motivates the work.
- Clear significance in Deep Learning research

**Response:** Thank you for the positive feedback.

Weaknesses:

- It would be better to explain a little bit about the existing open-source tool (citation [3]), rather than only referring to the citation so that the readers don't have to actually check out the reference. You do this in Related Work section but not in the earlier sections.

**Response:** Great feedback. I'll add a short sentence to the abstract introducing the existing tool. I'll also add a sentence or two about the existing tool in the beginning of the proposal.

- Related Work could have more details, explaining what is currently there or if there is anything you could build upon, even if you are not directly following up on them. Are there any similar visualization in other topics/fields?

**Response:** I will flesh out the Related Works section with the seminal paper, as well as two lesser-known, more recent tools I found.

Lacks description of planned evaluation.

**Response:** Good point. I'll add a section on planned evaluations.

Other comments for discussion:

- Organization: I like how clear the scientific and visualization significances are. Some parts in significance could be moved to Related Work with added citations.

**Response:** I'll flesh out the Related Works section.

### 3 Tertiary Reviewer 2 (Musa Tahir)

Overall: 3/4. Strong and well put together proposal that proposes a novel tool to enhance deep learning workflows and visualization

Interdisciplinary: 6. Maybe consider emphasizing the specific ways deep learning visualization can be used in an interdisciplinary manner, since the algorithm is inherently interdisciplinary.

**Response:** As stated in Review 2, I will clarify the interdisciplinarity in an additional section of the proposal. Additionally: although it's not my primary argument for interdisciplinarity of the proposal, I can also add a sentence about the inherent interdisciplinarity of deep learning research, as evidenced by two recent Nobel Prizes.

Scientific: 4. There's a significant scientific contribution this proposal makes (generating novel insights for deep learning loss landscapes). Be more specific in how this will concretely facilitate deep learning research.

**Response:** Good suggestion. I will include specific examples of how this helps deep learning research, including examples from the collaborator.

Visualization: 4. Although there are certainly visual contributions, compare them more to baseline methods and focus on what practical improvements researchers can expect in terms of any quantifiable metrics if possible

**Response:** I'm not sure what quantifiable metrics I can reference here, but as in Review 2, I will reference the literature for visualizing high-dimensional scalar fields.

Significant: 2. This project has the potential to push deep learning research forward by improving visualization workflows and methods.

Novel: 2. The novelty primarily lies in the custom 2D slices and adaptive sampling, but scientific novelty would depend on research results

**Response:** Good feedback. We can improve the claims of novelty even more by giving concrete examples of the kinds of research results we want to achieve, e.g. by the outside collaborator.

Goals clearly stated: 4. The goals are clearly stated, but the proposal could benefit from more detail explaining the process of achieving them.

**Response:** I will flesh out the proposal by adding a section between Related Work and Research Plan that has more detail about the aims.

Likelihood of Success: 4. I think visualization will be quite doable in a six-week timeframe, but both the adaptive sampling and custom metrics may or may not be feasible depending on potential implementation issues

**Response:** Good point. This is probably right. As in Review 1, I will rewrite the 6-week plan so that the "safety plan" is to focus on just custom metrics, with a "reach plan" to accomplish both custom metrics and adaptive sampling.

Strengths:

Clearly defined problem with practical applications

While the proposal is built in prior work, the contributions proposed are meaningful and advance the field

forward

Researcher focused by placing emphasis on workflows and usability

Weaknesses:

Could benefit from more detail on the empirical evaluation of the tool's effectiveness

**Response:** As in Review 1, we will add description of an evaluation to the proposal.

Could emphasize the interdisciplinary nature of deep learning itself

**Response:** Addressed above in this review (Interdisciplinary).

More emphasis needed on the difference between 2D custom slices and existing tools

**Response:** As in Review 2 (Novel), I will make the proposal more explicit in how our work differs from existing tools.

Other comments for discussion:

Will the success of the adaptive sampling approach be measured or quantified?

**Response:** We will add a section to the proposal about evaluations. For adaptive sampling, we can measure this using (A) the “end-to-end” evaluation proposed in Review 1 (Weaknesses) where we measure performance on some task, and (B) a specific evaluation of adaptive sampling using quantitative metrics (how close an adaptively sampled landscape is to the actual, high-fidelity landscape, as measured by a mathematical metric that captures an aspect we care about, such as smoothness.)

What will the graphical interface look like, or will it interact with the user through code?

**Response:** We primarily care about interacting with the user through code. We will amend the beginning of the proposal, or the new section with details on the aims, to specify this.

## 4 Tertiary Reviewer 3: Brendan Leahy

Overall: 3

Interdisciplinary: 5; Relates to deep learning and visualization. Could outline more specific interdisciplinary connections to other fields that may be implicated by deep learning contributions.

**Response:** See Review 3 (Interdisciplinary).

Scientific: 2; Clearly demonstrates how visualization insights are important and how this may be integrated into future workflows.

Visualization: 3; Framing this as a high-dimensional visualization problem without specifying what challenges are associated with this seemed poorly framed. However, the potential for improving efficiency of visualization and the custom 2D slices are an interesting visualization problem.

**Response:** Indeed. See Review 2 (Visualization) for the changes I'll make.

Significant: 2;

Novel: 2-3. Adaptive sampling and improved visualization speed of loss landscapes are novel contributions. Similar approaches have been taken but this builds on them nicely, amount achieved will determine novelty.

**Response:** Thanks for the positive feedback! See Review 3 (Novel) for how I will make the proposal even stronger in this regard.

Goals clearly stated: 3;

Likelihood of Success: 3. This is a challenging problem, and the three listed contributions are not super easy to achieve. However, listed PI and CO-PI have a strong background and have worked together on projects of similar scope.

**Response:** Indeed. See Review 1: I will rewrite the 6-week plan so that the “safety plan” is to focus on just custom metrics, with a “reach plan” to accomplish both custom metrics and adaptive sampling.

Extremely novel practical applications for deep learning

Well laid out conceptual goals

**Response:** Thanks for the positive feedback!

Weaknesses:

Goals can be more specific in what methods will be used to achieve them (aside from pytorch integration, which is a strong starting point)

**Response:** Good point. See Review 3 (Goals clearly stated) for improvements on this front.

Other comments for discussion:

waited on slides for visual insights (which were helpful), and computational methods (which are still not detailed)

again, really like this project idea overall, and cool to see how Arjun and this proposal took different angles on things!

**Response:** Thanks, I appreciate the reviewer's positive feedback! I'll add visual insights to the proposal as well. I will add a little bit of detail to the proposal about how we will achieve the aims.

# Visualization of Loss Landscapes of Deep Neural Networks

Kevin A. Wang

PI

kawang@brown.edu

Arjun Prakash

Co-PI

arjun\_prakash@brown.edu

Randall Balestrieri

Collaborator

randall\_balestrieri@brown.edu

October 14, 2024

## Abstract

We will present a novel tool for visualizing loss landscapes of deep neural networks. The tool will be based on the current state-of-the-art [4]. Our contributions will include easier integration into existing workflows, faster visualizations, and the ability to specify custom metrics to pick which 2D “slice” of the high-dimensional space to visualize.

# 1 Introduction and Background

Deep neural networks have millions or billions of real-valued parameters called *weights*. The “goodness” of any neural network for a given task is usually measured by a function called the *loss*. The goal of training a neural network is to find out how to set all the parameters, in order to achieve the lowest loss.

Although the standard technique for training deep neural networks (gradient descent on training data) has achieved great results empirically, its success has historically been a mystery to deep learning researchers [5].

Researchers are interested in understanding why neural networks work well, and what makes certain neural network design choices (such as choice of architecture, or the use of some regularization) easier to train than others. This understanding can then lead to improved neural networks.

To gain this understanding, researchers such as our collaborator, Randall, often use visualizations of the neural network’s loss landscape. The loss landscape is the scalar field that maps each possible parameterization of the neural network to its corresponding loss. This scalar field is a map from  $\mathbb{R}^N$  to  $\mathbb{R}$ , where  $N$  is the number of weights of the neural network (e.g. in the hundreds of millions). Since such a high-dimensional thing is impossible to fully visualize, researchers instead visualize 2D “slices” of such a field. Despite the infinitesimal size of such a slice compared to the full, high-dimensional field, such visualizations are widespread and useful in practice.

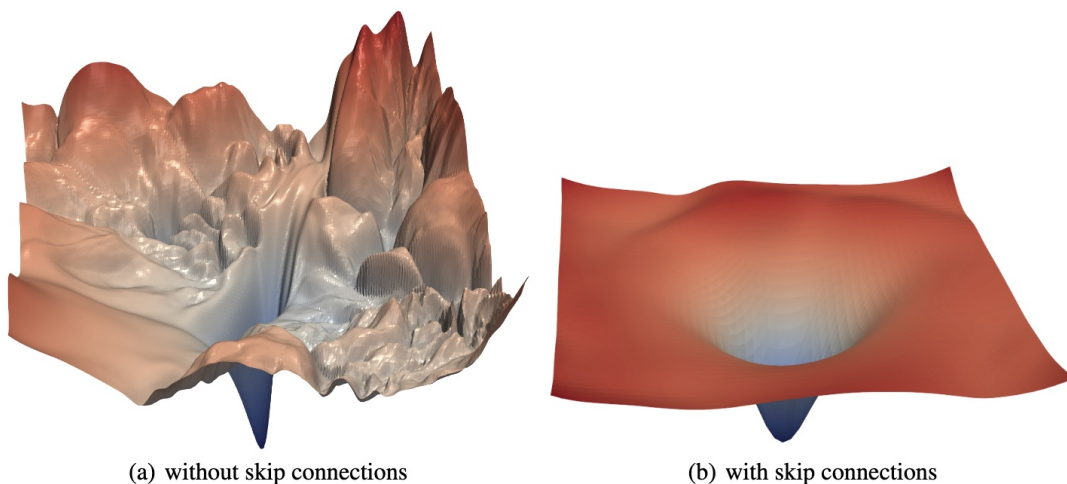


Figure 1: Figure from the Li, et al. [4] showing that the choice of neural network architecture dramatically affects the loss landscape

Such visualizations began in the seminal 2018 paper “Visualizing the Loss Landscape of Neural Nets” [4]. Since then, visualizations have been oft-used, and the paper has been cited 2131 times. According to our collaborator, researchers often use visualizations for understanding during their research process, even when they may not include visualizations to present their findings to others.

Although such visualizations are widespread and useful, there has been very little follow-up on improving the visualization tools themselves. Until now, that is!

## 2 Aims

We will develop a new tool for visualizing loss landscapes of deep neural networks. We will do this by extending the existing open-source tool from the seminal work “Visualizing the Loss Landscape of Neural Nets” [4].

Our tool will allow deep learning researchers to easily visualize the loss landscapes of neural networks during and after training. Compared to existing tools, it will have the following contributions:

1. Easy integration into existing deep learning workflows: Our tool will be a Python module with a simple API for popular deep learning frameworks such as Pytorch.
2. Visualizing custom 2D slices: The parameter space of modern neural networks is very high-dimensional (there may be millions of parameters, known as “weights”). However, we can only effectively visualize over a 2D domain, so visualization tools pick a 2D slice of the high-dimensional space to visualize. Our tool will allow researchers to pick from a variety of strategies to pick such a 2D slice, such as “the slice with the smoothest landscape”, or “the slice with the least smooth landscape”.
3. (If time permits) Faster visualizations through adaptive sampling: Sampling the loss at a given point in the domain can be expensive, and a “2D  $\rightarrow$  1D” visualization requires sampling the loss at many  $(x, y)$  points. By adaptively sampling points which are more likely to be interesting, we will speed up the visualization process.

## 3 Evaluation

As part of our project, we will evaluate our visualization tool. For each of our contributions, we will compare it against the baseline (the existing open-source tool [4]).

To evaluate the speedup of visualization through adaptive sampling, we can perform a fully quantitative evaluation. We will first pick a metric or set of metrics which summarize a loss landscape (e.g. the condition number of the loss landscape). We will calculate the true value of the metric, or a very close approximation by sampling a very large number of points. We will then pick a threshold value. We will then measure how many points each approach needs to sample in order to get within the threshold value of the true value. Our method will be successful if it can reach the threshold with fewer samples than the baseline, for a variety of threshold values.

To evaluate the custom 2D slices and integration, we must perform an experiment with a deep learning researcher. We will perform experiments with our outside collaborator. We will pick a task to be performed using visualization. The task will consist of a set of different neural network architectures. The goal of the researcher will be to pick the neural network architecture which will train the best. We will measure the ground truth by training the neural networks. The researcher will perform the task multiple times for our method and for the baseline, with different architectures each time. The result of the experiment will be the accuracy of the researcher’s pick, the speed that they perform the task, and their subjective enjoyment using the tool.

## 4 Significance

### 4.1 Interdisciplinarity

This project is interdisciplinary, bridging the fields of *visualization* and *deep learning*. Our proposal (and previous work) is based on the idea that visualization techniques help deep learning researchers develop insights and generate novel scientific knowledge about deep learning.

Further, deep learning research is inherently interdisciplinary – it is a field of computer science, but is frequently applied to other scientific domains, sometimes with great effect [1].

### 4.2 Significance to Deep Learning Research

Researchers visualize loss landscapes to gain insights into deep learning. For example, by inspecting a loss landscape, a researcher can gain intuition into why or why not a neural network will reach a global optimum through training (e.g. if the loss landscape is smooth and has no spurious local minimum, it seems probable that it will easily reach the global optimum).

Since the popularization of loss landscape visualization in 2018 [4], researchers have used visualizations to understand and explain a variety of phenomena with neural networks, such as the advantages of a neural network architecture over another.

Our tool will be easier to integrate with existing deep learning workflows, and will use faster visualizations. This reduced friction will enable deep learning researchers to perform more visualizations *during* training itself, and to perform more visualizations while iterating and exploring deep neural networks, which should lead to more insights with less work.

By allowing researchers to choose custom 2D slices such as “smoothest” and “least smooth”, researchers can gain a richer understanding of the loss landscape than by simply looking at one (random) 2D slice of a millions-dimensional space. With existing tools, any understanding may be incomplete or misleading due to this simplification, and our contribution will attempt to ameliorate this.

### 4.3 Visualization Significance

The neural network parameter space that we truly care about is extremely high-dimensional (millions or even billions of dimensions). The visualization challenge inherent is thus visualizing a million-D to 1-D dataset. To simplify the problem, we are reduced to visualizing a 2-D to 1-D slice of the dataset.

Our project will attempt to better visualize this high-dimensional dataset. Primarily, we will visualize multiple 2D slices, for example by showing the maximum and minimum 2D slices for some aspect that the user cares about. This should be of broad application to the challenge of visualizing high-dimensional spaces.

## 5 Related Work

Visualization research was largely pioneered in Li, et al.[4]. Precursors include work visualizing 1D slices and theoretic work analyzing loss landscapes [3]. Their work picks a random 2D slice, or picks a 2D slice using PCA, but does not allow the user to define a custom function and request the slice that maximizes and/or minimizes the function.

There has been very little follow-up work on improved visualization methods. A recent work [2] speeds up visualization for validation loss by simply noting that very few validation samples are needed to get an accurate visualization, but they do not use adaptive sampling.

## 6 Research Plan

- Week 1: We will sit down and outline the project in greater detail: What will the API of the software be? Which contributions do we need to prioritize? We will play around with existing software and understand how it is used.

Milestone: detailed outline of the project, images generated with existing software

- Week 2: We will create or obtain example experiment training code and completed neural networks to use for our testing. We will outline our code. We will implement code to show a plot during neural network training.

Milestone: end-to-end demonstration of code to train a neural network while showing loss landscapes

- Week 3: We will begin work on selecting custom 2D slices, using “smoothest” and “least smooth” metrics. We will get ideas for other metrics. We will produce a first prototype of it.

Milestone: Code that accepts a custom function and produces a 2D slice that maximizes the function.

- Week 4: We will continue work on selecting custom 2D slices.  
If selecting 2D slices is complete, we will integrate adaptive sampling.

Milestone: End-to-end demo – input a custom function and get multiple 2D slices while training.

- Week 5: We will finalize work on selecting custom 2D slices  
If selecting 2D slices is complete, we will work on adaptive sampling. We will draft the project report.

Milestone: Project report draft

- Week 6: We will finalize the project report.

## References

- [1] The Nobel Prize in Chemistry 2024.
- [2] R. Bain. Visualizing the loss landscape of winning lottery tickets. *arXiv preprint arXiv:2112.08538*, 2021.
- [3] I. J. Goodfellow, O. Vinyals, and A. M. Saxe. Qualitatively characterizing neural network optimization problems, 2015.
- [4] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein. Visualizing the loss landscape of neural nets. In *Neural Information Processing Systems*, 2018.
- [5] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, Feb. 2021.

## **A Note from Collaborator**

Hi Kevin,

I confirm that I am participating in the loss landscape visualization project and that we already meet two times to discuss about it. As mentioned in the last meeting, I am looking forward to working on this together during the semester as I believe your project answers some important needs for the AI community.

Best,

Randall

**Skills and capabilities:** programming, algorithms, machine learning

---

## Education

**PhD Student at Brown University**

*Fall 2023 - ???*

PhD student with Professor Amy Greenwald

**University of California, Irvine - BS in Computer Science**

*2014 - June 2018*

---

## Research Interests

Broadly, I'm interested in machine learning and game theory, with a focus on search / planning / "system 2 thinking".

Currently, I'm interested in computational game theory, multi-agent reinforcement learning, and the pursuit of solving large, partially-observable games, both theoretically and practically.

---

## Publications

Publications can be viewed at <https://scholar.google.com/citations?user=Q06Rh6oAAAAJ>

---

## Research and Industry Experience

**Poker Consulting**

*August 2023*

Consulted for a Poker AI startup. Reviewed code and advised on state-of-the-art methods.

**Meta AI (FAIR) - AI Resident**

*July 2022 - July 2023*

Worked with Noam Brown on research in computational game theory.

**UC Irvine - [Visiting Researcher](#)**

*Mar 2021 - 2022*

Worked with Prof. Roy Fox and Stephen McAleer to research methods to approximate Nash equilibria in zero-sum, 2-player, imperfect-information games, including depth-limited search.

**Shift - Software Engineer**

*Aug 2018 - Dec 2020*

Shift was an online car marketplace startup. I built and maintained the test-drive scheduling system using a constraint programming solver. I also did full-stack web development with Golang, Typescript, and React. This work contributed to the company going public in 2020 at a \$700+ million valuation. It is unfortunately now defunct.

---

## Fellowships

**CSGrad4US - NSF Fellowship**

*2023-???*

[NSF fellowship](#) for CS graduate school applicants with \$34,000 annual stipend and \$12,000 annual cost-of-education allowance for 3 years. One of 35 students selected nationally in 2021.

# Synchronized 2D and 3D Visualization of Disease Spread and Control Measures

Kei Yoshida (PI)  
Computer Scientist, Psychologist  
kei\_yoshida@brown.edu

Richard Huang (Co-PI)  
Computer Scientist  
richard\_huang2@brown.edu

Simon Su  
Computer Scientist  
simon.su@nist.gov

Dachollom Sambo  
Applied Mathematician  
dachollom.sambo@nist.gov

October 14, 2024

## Abstract

We propose to develop synchronized 2D and 3D visualizations of Lassa fever transmission and the effects of non-pharmacological control measures. This will improve understanding of spatial and temporal aspects of disease data, making complex information accessible to researchers, public health professionals, and the public, aiding infectious disease management.

Dear Editor and Reviewers,

We appreciate the constructive feedback on the manuscript, and we have prepared a revision accordingly. Please see below for an itemized list of your feedback and a brief description of how each will be addressed in the revised proposal.

Reviewer 1: jliu239

Reviewer 2: kawang

Reviewer 3: mtahir1

Reviewer 4: myoon15

*Reviewer 1*

[R1.1] Scientific: This project aims to make Lassa fever transmission data more interpretable, allowing users to better understand the spread of the disease and the effectiveness interventions. The proposal could elaborate a bit more on how this could further research in this area.

**Response:** The final proposal addressed the scientific contributions in more detail by emphasizing how this project can further research in the field of epidemiology. The ability to visualize complex transmission dynamics in both humans and rats over time and space allows them to intuitively compare the patterns and trends in both real and simulated data.

[R1.2] Novel: The methods of visualization themselves may not be novel, but the application to this particular data and context is novel.

**Response:** I made it more clear in the final proposal that the novelty is not with the idea of "synchronized 2D & 3D visualization", but the aspects that our visualization can display. The ability to see the dynamic interaction between multiple variables in one platform provides new insights, making the tool particularly novel and useful in this context.

[R1.3] Likelihood of Success: Research plan is well-defined, accompanied by key milestones for each week. However, the number of aims might be a bit ambitious for the given timeframe of the project.

**Response:** I reduced the scope of the project in the final proposal to ensure feasibility within the timeline. Specifically, I removed the aspects related to interactive web-based functionality (keeping as optional), focusing instead on delivering pre-rendered animations (e.g., GIFs or short videos).

[R1.4] Weaknesses: The proposal would benefit from defining evaluation methods. For example, how does one measure if a 3D visualization is more interpretable than a corresponding 2D plot? In other words, how will the new visualizations be evaluating in relation to the baseline?

**Response:** I included the planned evaluation methods in the final proposal. It includes feedback from one of the collaborators who is a domain expert in mathematics and modeling of epidemiological data. Additionally, we will conduct a small-scale user study where participants are asked to interpret 2D and 3D visualizations of the same data and report on their understanding and insights.

*Reviewer 2*

[R2.1] Scientific: Scientific understanding of disease spread is significant. However, it's not intuitively clear why 3D visualizations are necessary for this topic.

**Response:** I included how engaging 3D figures are compared to 2D.

[R2.2] Visualization: The project will combine 3D and 2D visualizations, synchronized. Its specific novelty is not clear.

**Response:** I clarified that the idea of "synchronized 2D & 3D visualization" itself may not be novel, but the specific application of synchronized visualizations for disease transmission dynamics and such dataset is. Our visualization will show a combination of geographical data, time-series transmission data, and the impact of control measures in an integrated format. This approach allows users to understand the data in a way that highlights the spatial spread of the disease alongside the temporal effects of interventions, which is a novel application in this field.

[R2.3] Significant: The project seems motivated by scientists who think this proposal would be useful, but as stated above, it's not clear to the reader in what way a 3D visualization is more useful than a 2D one in this setting.

**Response:** I expanded on why the multi-dimensional visualization is more useful than traditional 2D only representations. 2D visualizations often fail to capture the complexity of interactions between different variables such as time, geography, and intervention effects. By incorporating all these elements into a single platform, we can provide a more comprehensive view of the data, allowing researchers to observe patterns and trends that may not be visible with 2D methods.

[R2.4] Likelihood of Success: The project seems ambitious and it's not clear if it will all be accomplishable in 6 weeks. The proposal mentions a web-based tool, but it's not clear if this is part of the 6-week plan.

**Response:** To address this concern, I scaled down the project to focus on deliverables that can be realistically achieved within the 6-week time frame. Please see response to [R1.3].

[R2.5] Weaknesses: It's not immediately easy to understand the Aims of the proposal. How are the aims different from each other, and are all of them novel compared to existing work?

**Response:** I clarified and distinguished between the aims. Each aim will be clearly defined in terms of its specific goals and contributions to the project. I also highlighted the novelty of each aim, particularly how they go beyond existing work. For example, I highlighted that the emphasis of the first aim is on the ability to display over a geo-graphical region, and the second aim on the time-series data.

### *Reviewer 3*

[R3.1] Overall: This project's proposed tool for interactive and dynamic visualization of infectious disease spread overall seems quite promising, some more clarity on some techniques would strengthen the proposal.

**Response:** The final proposal provides more detail on the technical implementation of the visualization tool, particularly how ParaView will be utilized.

[R3.2] Interdisciplinary: The project combines CS, Public Health, and Applied Math. Project may benefit from more direct interdisciplinary collaboration with public health experts.

**Response:** While we do not have existing collaborations with public health experts, we will try to find to receive insights from experts who are in relevant fields (e.g., public health or medical students) as part of the user study.

[R3.3] Scientific: The visualization is scientifically valuable by enhancing the understanding of fever spread, but the proposal would benefit from more discussion on how the tool will specifically help researchers develop novel insights.

**Response:** The final proposal provides specific examples of how researchers can use this tool to generate

novel insights. For example, the ability to visually compare the effects of different interventions over time will enable researchers to identify intervention strategies that may not be apparent from raw data or traditional 2D plots. Additionally, the exploration of multiple dimensions (e.g., geography, time, intervention) will enable researchers to generate new hypotheses about the factors driving transmission, improving the design of new models and response strategies.

[R3.4] Visualization: The interactive visualization is novel, but I am curious how you will ensure the tool is UI friendly without overwhelming the user. More detail on this point would strengthen the proposal.

**Response:** The revised proposal includes a mention of how we plan to balance simplicity and usability. This may be partially solved by removing the interactive feature in the project (please see response to [R1.3]).

[R3.5] Significant: Contributions would improve public health knowledge and better inform experts. Perhaps more emphasis on validating the tool beyond Lassa fever, potentially referencing other case studies to broaden the impact.

**Response:** I mentioned the tool's potential application to other infectious diseases. I changed the narrative of the proposal so that Lassa fever is mentioned as one use case, rather than the main application.

[R3.6] Novel: The tool is novel by integrating synchronized 2D and 3D visualizations with live interaction for infectious disease data. There's a clear baseline here that I think is identified and incrementally improved upon.

**Response:** I reiterated the novelty of integrating synchronized 2D and 3D visualizations, particularly in the context of infectious disease transmission. I explicitly stated the baseline (existing visualization with 2D or 3D only and no synchronization component).

[R3.7] Goals clearly stated: The goals are all clearly defined, but the proposal would benefit from more clarity on user testing techniques.

**Response:** The final proposal includes details on user testing techniques. Please see response to [R1.4].

[R3.8] Likelihood of Success: I think success is probable, but it depends on the technical implementation issues posed by synchronizing 2D and 3D visualizations.

**Response:** To ensure technical feasibility, we detail the synchronization method between 2D and 3D visualizations using ParaView.

[R3.9] Weaknesses: Visualizations may be complex for non-expert users, so this must be accounted for in the design.

**Response:** I addressed this concern by designing visualizations that balance complexity with clarity. Please see response to [R3.4].

[R3.10] Weaknesses: Synchronization of 2D and 3D data may be difficult to implement, but certainly possible depending on progress and technical limitations of ParaView/Trame.

**Response:** The revised proposal includes more technical details of Paraview related to synchronizing 2D and 3D data that would make the proposal achievable.

[R3.11] Comment: Will you collaborate with any public health policymakers in the user testing phase?

**Response:** Please see response to [R3.2].

[R3.12] Comment: How will you accommodate larger datasets that might cause performance issues?

**Response:** We will leverage the parallel rendering capabilities of ParaView and use optimized data pipelines to ensure smooth and responsive visualizations, even with larger datasets. Additionally, we have access to and plan to utilize high-performance computing resources to support the rendering of large-scale simulations without compromising performance.

*Reviewer 4*

[R4.1] Novel: As previously mentioned, this project addresses a gap in visualization methods for Lassa fever data. As such, I believe it's novel. However, 3D visualizations of disease transmissions have been done before.

**Response:** We agree that 3D visualizations of disease transmission have been done before, but the novelty of our approach lies in the integration of 2D and 3D synchronized visualizations. Please see responses to [R1.2] and [R2.2].

[R4.2] Goals clearly stated: Goals are clearly stated, but I would like a little more detail on the implementation.

**Response:** In the revised proposal, we provide more detail on the technical implementation of the visualization system, particularly for the use of Paraview.

[R4.3] Likelihood of Success: I believe a potential challenge could be linking the 2D and 3D data to effectively convey data in a way that's not redundant, but only adds information.

**Response:** I addressed this challenge by carefully designing visualizations the 2D and 3D views complement each other rather than duplicate information. Please see response to [R3.4]. For example, 2D views may focus on time-series data of a particular variable, while 3D views provide a spatial representation of the disease spread. The synchronization will be done in a way that enhances the user's understanding of both aspects simultaneously.

[R4.4] Weakness: It's not clear how researchers will evaluate the success/efficacy of these new visualizations and compare them to traditional 2D representations.

**Response:** Please see responses to [R1.4] and [R3.6].

[R4.5] Comment: I'm just curious how this data is formatted and whether it's a plug-and-play into ParaView, or if it needs preprocessing.

**Response:** We have yet to acquire the full dataset (only a sample dataset), but the data will most likely require some preprocessing before being imported into Paraveiw. In the final proposal, we incorporate the preprocessing steps in the research plan, which may include cleaning the data, and converting it into formats compatible with ParaView for rendering.

# 1 Aims

The overall goal of this research is to develop animated visualizations that synchronize a 3D representation of disease transmission over a geographical region with a 2D time-series view of related variables, demonstrating the effects of non-pharmacological control measures on disease spread. This novel approach transitions from traditional 2D visualization methods to a more engaging and intuitive 3D format, which better conveys complex information. By combining 2D and 3D visualization, we can represent spatial disease spread, control measures, and related variables without over-complicating individual figures. The animations will further enhance the understanding of both temporal and spatial dynamics in disease transmission. Synchronized views across two panels will allow researchers to explore complicated datasets without missing key insights, and make the visualizations accessible to public health professionals and the public, offering significant societal benefits. While this project focuses on Lassa fever as a case study, the approach is generalizable to other infectious diseases.

To address the gaps in understanding disease dynamics, we propose the following specific aims, each contributing uniquely to the visualization framework:

1. **Develop 3D visualizations of infectious disease data over a geographical region.** The 3D visualization will integrate data from both human and rodent populations and map disease transmission across a geographical region, offering a comprehensive and engaging view of the spatial dynamics that drive the epidemic.
2. **Develop 2D visualizations showing the effects of control measures and other related variables.** We will also employ a traditional 2D visualization, displaying control measures (e.g., quarantine, isolation, hospitalization, and public awareness campaigns) and other related variables as a function of time, demonstrating the impact of these variables.
3. **Develop animated visualizations of time-series data of infectious disease.** With animated visualizations, we can display the temporal evolution of Lassa fever, allowing users to observe how the disease progresses over time. Additionally, animation allows us to show the effectiveness of interventions as a function of time observe how the disease progresses over time.
4. **Synchronize a traditional 2D visualization of control measures with a 3D visualization of disease spread.** Animations of synchronized 2D and 3D visualizations can display complex information without making them too inaccessible to non-experts. It can retain the advantage of 2D visualizations to display detailed and quantitative information, as well as the engaging experience of 3D visualization, enhancing both spatial understanding and data analysis.

# 2 Significance

The proposed work addresses significant gaps in current visualization methods for infectious diseases research. Traditional 2D visualizations are suited for presenting detailed quantitative variables, but they can be difficult for non-experts to interpret. 3D visualizations are often more engaging and better at conveying complex, multidimensional information, but they are heavily underutilized in infectious disease studies.

Using these traditional 2D visualizations as a baseline, our approach seeks to bridge this gap by offering accessible, intuitive visualizations that make complex disease data easier to interpret. The key contribution of this project is the development of animated visualizations that integrate the display of **synchronized 2D and 3D views**, integrating multivariate data into a single platform.

Achieving previously mentioned aims and transitioning from traditional 2D methods to a combined 3D approach will offer a novel technique for visualizing epidemiological data, offering benefits from multiple perspectives:

1. **For researchers:** The visualizations will facilitate deeper exploration of data. Currently, researchers rely on deductive reasoning from raw data to identify trends, such as seasonal outbreaks, intervention impacts, or non-compliance consequences. This process can lead to overlooked details. By visualizing transmission dynamics over time and across space for both humans and rats, researchers can more easily spot patterns and compare different data types (e.g., human vs. rodent infections, real-life vs. simulated data, and the effects of various control measures).
2. **For policymakers, healthcare professionals, and the public:** The accessible, easy-to-understand visualizations will enable informed decision-making in disease management, allowing policymakers and healthcare professionals to base strategies on a clearer understanding of disease spread and intervention effects. Additionally, the public can improve their understanding of infectious diseases and the effectiveness of control measures.

Ultimately, this project fills a critical gap in infectious disease research by developing a versatile, accessible tool that enhances both scientific exploration and public communication.

## 3 Related Work

### 3.1 Background

Lassa fever is an infectious disease transmitted to humans via the rodent multimammate rat. It is endemic to West African countries, and it has spread to other parts of the world. Rodents spread the virus for their lifetime, and the virus can cause severe disease in affected humans, leaving significant public health implications in the impacted regions. Therefore, effective disease management is crucial. Non-pharmacological approaches, such as quarantine, hospitalization, isolation, and mass awareness, have been used to mitigate the disease spread.

Our collaborators have been working on developing a mathematical compartmental model that tracks disease transmission in two hosts (i.e. human and rat populations), incorporating these non-pharmacological control measures [3]. Effective data visualizations of complex multivariate data are crucial for these researchers to visually identify patterns in the data and better understand different factors that can affect disease dynamics.

### 3.2 Existing visualizations

**2D visualizations** are most commonly used to represent infectious disease data and the effects of control measures. While, they can be effective in certain contexts [2], they are limited by the lack of spatial depth and clarity, limiting their ability to fully convey the complexities of disease transmission.

In contrast, **3D data visualization** offers several advantages, such as being more engaging, easier to understand [8], and better at conveying complex information. Some 3D visualization methods are particularly effective in communicating virus outbreaks over geographical regions [10] [5]. It enhances the understanding of disease spread in geographical space, improves data interaction, and supports the integration of multivariate data, visualization of temporal changes, and the use of virtual reality and educational

tools. Despite these benefits, 3D visualizations are underutilized in infectious disease modeling, with only a few notable exceptions [6].

However, multi-dimensional visualizations come with their own challenges. As more layers of information are added, the visualization can become too complex and hard to interpret. To address this, visualization researchers have developed **hybrid 2D and 3D visualizations** where two views display related but different information [7, 4, 9]. By integrating these views and keeping simplicity and clarity in each element, researchers can represent complex data without overcrowding each element of the visualization and overwhelming the viewers.

Additionally, **animated visualizations** have proven particularly useful for exploring large spatio-temporal datasets [1]. These animations enable users to follow the progression of disease transmission over time, enhancing both spatial and temporal understanding.

## 4 Research Plan

### 4.1 Schedule

Over the course of the 6 weeks, we will continue to work on preparing a final presentation and writing detailed weekly reports. Additionally, we will frequently get feedback from collaborators as topic experts.

Week	Goal	Milestones
<b>Week 1</b>	Basic 2D & 3D mapping in Paraview	(1) Data preparation: gather data & pre-process if needed. (2) ParaView setup: create two panels & import data. (3) Spatial representation: import geographical data and set up the projection. (4) Discussion with collaborators.
<b>Week 2</b>	Static 3D visualization of disease spread in space	(1) Implement 3D visualization of a geographical map. (2) Use color mapping to represent disease intensity in different areas for humans and rodents.
<b>Week 3</b>	Dynamic 3D visualization of disease spread in time	(1) Implement an animation using time-series data.
<b>Week 4</b>	Synchronized 2D & 3D visualization, showing the effects of control measures	(1) Incorporate non-pharmacological control measures and other demographic factors in the 2D view. (2) Synchronize between 2D and 3D views.
<b>Week 5</b>	Refinement & user study	(1) Refinement of the visualization. (2) Prepare & start the user study (over the Thanksgiving break). (3) Discussion with collaborators (expert evaluation). (4) Prepare presentation & report materials.
<b>Week 6</b>	Processing user study results & final refinement	(1) Process user study results. (2) Make any adjustments necessary based on the results. (3) Finalize presentation & report materials.
If time permits:	Web-based and/or interactive application	(1) Utilize Trame to create a web-based application. (2) Use Paraview to implement interactivity.

## 4.2 Evaluation

Our evaluation of the visualizations will rely on both expert insights and user studies to assess the effectiveness and usability of the visualizations.

First, we will leverage feedback from our collaborators who have been developing a Lassa fever mathematical model. Their expertise will provide valuable insights from a researcher's perspective, particularly in terms of which aspects of the visualization enhance or hinder his ability to interpret real-life and simulated data. Their feedback will help us identify which features are necessary or unnecessary to improve workflow when exploring complex epidemiological data.

Additionally, we will conduct a user study to evaluate how effectively the visualizations communicate patterns and trends. This study will likely be conducted through Amazon Mechanical Turk or within the context of the CS237 class. Participants will be asked to interact with our visualizations, as well as baseline visualizations, and identify and list the patterns they can observe. Since the goal of our visualizations is to intuitively convey complex information, the number and diversity of patterns identified by users will be a key measure of effectiveness.

The baseline visualizations will include traditional 2D visualizations from our collaborators' previously published articles, as well as existing 3D visualizations from a select number of studies that utilize 3D methods for disease modeling.

Furthermore, if possible, we hope to gather additional insights from individuals in public health, medicine, or policy-making fields. While this may require more assistance from our collaborators, incorporating feedback from professionals in these areas will help ensure that the visualization tool can be directly applied to decision-making processes at higher levels, making it more impactful in real-world scenarios.

## 4.3 Facilities

For the development of our 3D visualization, we will primarily use the scientific visualization tool ParaView, and potentially Trame, both of which are open-source. ParaView allows us to set up multiple view panels within a single interface, making it ideal for synchronizing different types of visualizations, such as 2D and 3D views.

We may utilize high-performance computing resources provided by the Center for Computation and Visualization at Brown University (Oscar), the Brown University Department of Computer Science (Hydra), and the Texas Advanced Computing Center (through allocations from the National Institute of Standards and Technology).

## References

- [1] T. Becker. Visualizing time series data using web map service time dimension and svg interactive animation. Master's thesis, University of Twente, 2009.
- [2] L. N. Carroll, A. P. Au, L. T. Detwiler, T. chieh Fu, I. S. Painter, and N. F. Abernethy. Visualization and analytics tools for infectious disease epidemiology: A systematic review. *Journal of Biomedical Informatics*, 51:287–298, 2014.
- [3] S. Dachollom and C. E. Madubueze. Mathematical model of the transmission dynamics of lassa fever infection with controls. *Mathematical Modelling and Applications*, 5(2):65–86, 2020.
- [4] E. Le Malécot, M. Kohara, Y. Hori, and K. Sakurai. Interactively combining 2d and 3d visualization for network traffic monitoring. In *Proceedings of the 3rd International Workshop on Visualization for Computer Security*, VizSEC '06, page 123–127, New York, NY, USA, 2006. Association for Computing Machinery.
- [5] C. K. Leung, Y. Chen, C. S. Hoi, S. Shang, Y. Wen, and A. Cuzzocrea. Big data visualization and visual analytics of covid-19 data. In *2020 24th International Conference Information Visualisation (IV)*, pages 415–420, 2020.
- [6] A.-S. Stensgaard, C. F. Saarnak, J. Utzinger, P. Vounatsou, C. Simoonga, G. Mushinge, C. Rahbek, T. K. Kristensen, et al. Virtual globes and geospatial health: the potential of new tools in the management and control of vector-borne diseases. *Geospatial health*, 3(2):127–141, 2009.
- [7] S. Su, V. Perry, L. Bravo, S. Kase, H. Roy, K. Cox, and V. R. Dasari. Virtual and augmented reality applications to support data analysis and assessment of science and engineering. *Computing in Science Engineering*, 22(3):27–39, 2020.
- [8] M. Teplá, P. Teplý, and P. Šmejkal. Influence of 3d models and animations on students in natural subjects. *International Journal of STEM Education*, 9(1), Oct 2022.
- [9] M. Vuckovic, J. Schmidt, T. Ortner, and D. Cornel. Combining 2d and 3d visualization with visual analytics in the environmental domain. *Information*, 13(1), 2022.
- [10] J. Zhang, Y. Wang, W. Wanta, Q. Zheng, and X. Wang. Reactions to geographic data visualization of infectious disease outbreaks: an experiment on the effectiveness of data presentation format and past occurrence information. *Public Health*, 202:106–112, 2022.



Yoshida, Kei <kei\_yoshida@brown.edu>

---

## Request for Collaboration and Support Note

---

**Su, Simon M. Dr. (Fed)** <simon.su@nist.gov>

Wed, Oct 2, 2024 at 1:36 PM

To: "Yoshida, Kei" <kei\_yoshida@brown.edu>

Cc: Dachollom Sambo <dasam7@morgan.edu>, "Sherman, William R. (Fed)" <william.sherman@nist.gov>

Hi Kei,

Thank you for your interest in the project and we are thrilled to have you working on the visualization part of the project.

Please consider this as confirmation of our involvement and support for the project. We are also working on different aspects of visualization using ParaView and Trame and would be happy to share what we have learned. Although not necessary, if you needed any hardware (HPC) access, we can also look into trying to get you access on TACC HPC systems. NIST has TACC HPC computing resource allocations that can be leveraged given we have enough time to process the lengthy governmental paperwork.

Thank you again and we are looking forward to our collaboration

Best

-simon

# Kei Yoshida

Brown University (Providence, RI 02912 USA) | [kei\\_yoshida@brown.edu](mailto:kei_yoshida@brown.edu)

## EDUCATION

---

- 2020 – current    **Ph.D. in Cognitive Science**, Department of Cognitive and Psychological Sciences (CoPsy), Brown University, Providence, RI  
Dissertation: *Pedestrian interactions at local and global levels in human crowds* (Advisor: Dr. William H. Warren)
- 2023 – current    **M.S. in Computer Science**, Department of Computer Science, Brown University, Providence, RI  
[Open Graduate Education Program Scholar](#)
- 2016 – 2020       **B.A. in Computer Science & Psychology**, Coe College, Cedar Rapids, IA  
Honors: *Magna Cum Laude*, Phi Beta Kappa  
Senior Honors Thesis: *Perceptual-motor recalibration in naturalistic and virtual environments* (Advisor: Dr. Benjamin Chihak)

## RELEVANT RESEARCH EXPERIENCE

---

- 2019 – 2020       **Undergraduate Student Researcher**, Dept. of Psychology, Coe College
- Senior Honors Thesis: *Perceptual-motor recalibration in naturalistic and virtual environments* (Advisor: Dr. Benjamin Chihak). Designed series of experiments systematically investigating recalibration effects in rotational locomotion in naturalistic and virtual environments. Developed a virtual environment in Unity using C#.
- 2019                **Undergraduate Student Researcher**, Dept. of Computer Science, Coe College
- Research Project: *Technology assisted review with iterative classification* (Advisor: Dr. Stephen Hughes). Developed a software tool to explore data mining using Python and Java, and implemented a Naive Bayes classifier for text classification.
- 2018                **Programming Technician**, Dept. of Biology, Coe College
- Research Project: *GIS-based study on topographical preference of common tree species in Palisades-Kepler State Park, IA* (Research lead: Abhinav Shrestha). Created Python scripts used to analyze geographical data in ArcGIS software.

## RELEVANT COURSES & SKILLS

---

### *Courses*

#### **Brown University** (2020 – current)

Data Science, Computer Vision, Deep Learning, Statistical Inference, Human-Computer Interaction, Perceiving and Acting in 3D, Perception and Action, Applied Regression Analysis

#### **Coe College** (2016 – 2020)

Principles of Computer Graphics, Data Structures & Algorithms, Programming Languages, Interactive System Design, Object Oriented Programming, Software Engineering, Foundations of Computer Science, Foundations of Advanced Mathematics, Research Methods, Statistical Methods and Data Analysis, Sensation and Perception

### *Programming Languages*

Python, MATLAB, C++, Julia, C#, Bash, C, SQL, R, HTML/CSS, Javascript, Java

### *Software & Instruments*

VS Code, Visual Studio, Git/GitHub, Jupyter Notebooks, SPSS Statistics Software, Unity, DJI Mavic 3 Pro ([Part 107 certified](#)) UAS pilot), HTC Vive, SSH, RStudio, Atom, IntelliJ IDEA, Xcode, Adobe (Photoshop, Premiere, Illustrator)

## PUBLICATIONS & PRESENTATIONS

---

### *Publications*

Warren, W. H., Falandays, J. B., **Yoshida, K.**, Wirth, T. D., & Free, B. A. (2024). Human crowds as social networks: Collective dynamics of consensus and polarization. *Perspectives on Psychological Science*, 19(2), 522–537.  
<https://doi.org/10.1177/17456916231186406>

**Yoshida, K.**, di Bernardo, M., & Warren, W. H. (2024). *Reconstruction of visual influence networks in walking crowds*. Under review.

# Richard Huang

[rzhang@brown.edu](mailto:rzhang@brown.edu)

[richardhuangz.github.io](https://richardhuangz.github.io)

## Research

---

I'm broadly interested in theoretical computer science and combinatorics, particularly in approximation and graph algorithms. I'm also interested in algorithm design in the context of incentives and uncertainty.

## Education

---

- Brown University** 2023 -  
PhD in Computer Science  
Advised by Ellis Hershkowitz
- Princeton University** 2019 - 2023  
AB in Computer Science, Minor in Cognitive Science

## Papers

---

- Simple Length-Constrained Minimum Spanning Trees** SOSA 2025  
With Ellis Hershkowitz
- Prophet Inequality of Partition Matroid Intersection** Undergraduate Senior Thesis, 2023  
Advised by Matt Weinberg
- On Multidimensional Stable Matching** Undergraduate Junior Paper, 2022  
Advised by Mark Braverman

## Teaching

---

- At Brown University**  
An Algorithmist's Toolkit (CSCI 2952T) 2024
- At Princeton University**  
Computational Geometry (COS 451) 2022  
Economics of Computation (COS 445) 2022 - 2023  
Reasoning about Computation (COS 340) 2021  
Computer Science: An Interdisciplinary Approach (COS 126) 2020 - 2023

## Service

---

- Conference Reviewing**  
SOSA
- University Service**  
Student Organizer for Computer Science PhD Orientation at Brown University 2024

## Activities

---

- Brown University Orchestra 2023 -  
Princeton Pianists Ensemble 2021 - 2023  
Princeton University Orchestra 2019 - 2023  
Greater Dallas Youth Orchestra 2015 - 2019

## (Old) Employment

---

### **Amazon Web Services**

Software Dev Engineer Intern 2023

### **Engineering Library at Princeton University**

L<sup>A</sup>T<sub>E</sub>X Instructor 2022

### **Amazon**

Software Dev Engineer Intern 2022

### **Office of the Dean of the College at Princeton University**

Academic Advising Assistant 2021 - 2022

### **Department of Computer Science at Princeton University**

Course Development Intern 2021

Research Assistant in Computing Education 2020 - 2021

Teaching Assistant Admin 2020 - 2023

# Enhancing Genotype Visualizations with Contextual Data Insights

Musa Tahir

PI

musa\_tahir@brown.edu

Kevin Durant [TBD, in process of choosing classmate]

Co-PI

easymoneysniper@goat.com

Alex Diaz-Papkovich

Collaborator

alex.diaz-papkovich@brown.edu

Sohini Ramachandram

Collaborator

sramachandran@brown.edu

October 14, 2024

## Abstract

This project presents a method for visualizing genotype matrices using PCA/UMAP to better contextualize genetic differentiation. Incorporating uncertainty metrics (e.g confidence ellipses), site contribution overlays, and additional genomic data (e.g RNA, epigenetic), this approach reduces the risk of overemphasizing genetic differences, offering more accurate and interpretable visualizations.

# 1 Reviews and Responses

Dear Editor and Reviewers,

We appreciate your feedback and have revised our proposal accordingly. Because of your feedback, our proposal addressed many of its gaps and is much stronger as a result. Please see below for an itemized list of your feedback and a brief description of how each was addressed in the final proposal.

Reviewer 1: kyoshid1

Reviewer 2: jliu239

Reviewer 3: myoon15

Reviewer 4: rhuang79

## Review 1:

[R1.1] Overall: 2.14

[R1.2] Interdisciplinary: 1. The proposal is interdisciplinary as it incorporates computer science (visualization) and genomics (biology), while borrowing ideas from other fields (e.g., dimensionality reduction techniques, uncertainty metrics).

**Response:** In the final proposal, I am more specific and explained how genetic researchers can incorporate these methods to safeguard against misinterpretation while advancing the field of genetics.

[R1.3] Scientific: 4. It is stated that "Other biological researchers will benefit from this analysis and could implement it in their research to guard against misinterpretation", but the "how" part is less clear here.

**Response:** In the final proposal, I clarified how the methods can be directly applied by other researchers.

[R1.4] Visualization: 1. An urgency for the proposed visualization is well motivated by explaining the risks of misinterpretation and misuse.

[R1.5] Significant: 4. It is very clear that many people (researchers and the general public) would benefit from it. It could be improved by different benefits that each objective brings.

**Response:** In the final proposal, I provided more detailed information on how each target group is benefited in the significance section (i.e general public/society and scientists).

[R1.6] Novel: 1. A NOVEL method for visualizing genotype matrices. It is clear that the existing work does not have all the components proposed in this project.

[R1.7] Goals clearly stated: 1. Four objectives are clearly described in the Aims section, and all are based on existing work and achievable.

[R1.8] Likelihood of Success: 3. The Research Plan is very detailed and achievable, which makes it very likely to succeed. This could be improved by more clearly describing the planned techniques and evaluation methods of misinterpretation.

**Response:** I have revised the proposal to more clearly explain the planned techniques for visualizing uncertainty and adding site contribution overlays, and explained the evaluation methods, which involve crafting a user study to measure misinterpretation tendencies.

[R1.9] Strengths: - It is clear what the proposed activity adds to the existing work (using dimensionality reduction, uncertainty metrics, site contribution overlays, and additional genomic data). - It explains the issue with the existing visualization. - The Aims section is well-organized and easy to follow.

[R1.10] Weaknesses: - The first sentence in the Related Work section could be explained in more detail so that readers/reviewers who may not be experts understand the techniques' importance. - It's unclear how current visualizations may be misinterpreted (e.g., Fig 1 in Related Work). - Less description of techniques

(e.g., uncertainty metrics).

**Response:** I explained the background of these techniques to help non-experts understand their importance. I also included specific examples of misinterpretations of existing visualizations.

[R1.11] Other comments: - In Related Work, "Figure 1." instead of "figure one". - Significance section could be divided into paragraphs for easier understanding. - Who are the "users"? (in Significance) - Related Work could focus more on motivating the proposed work rather than describing it explicitly.

**Response:** I clarified the meaning of "users". The Related Work now focuses on studies that motivate the proposed approach.

## Review 2:

[R2.1] Overall: 3

[R2.2] Interdisciplinary: 2. The project integrates concepts from genomics with dimensionality reduction and visualization techniques.

[R2.3] Scientific: 4. There are two clear directions for scientific contributions: reducing misinterpretation of genomic visualization and furthering genomics research. However, it is unclear how these contributions will be evaluated.

**Response:** Please refer to [R1.8], where I mention the integration of visualization evaluations.

[R2.4] Visualization: 4. The idea of incorporating site contribution overlays is well-defined, but it is less clear how the uncertainty aspect will be executed.

**Response:** In my final proposal, I better define how uncertainty metrics will be executed by using bootstrapping to output shaded uncertainty regions and density heatmaps.

[R2.5] Significant: 2. If successful, the potential to mitigate risks associated with misinterpreted genomic data is significant.

[R2.6] Novel: 3. The focus on uncertainty metrics and specific genomic overlays is novel.

[R2.7] Goals clearly stated: 3. There is a well-defined problem, and the goals are clearly stated, but methods of implementation could be more in-depth.

**Response:** Please refer to [R2.4].

[R2.8] Likelihood of Success: 3. The research plan is clear, but without defining the uncertainty measures, it's hard to gauge feasibility.

**Response:** Please refer to [R2.4].

## Review 3:

[R3.1] Overall: 4.6

[R3.2] Interdisciplinary: 3. This proposal aims to create better visualizations of genotype matrices.

[R3.3] Scientific: 5. This proposal has strong societal contributions, as it prevents data from being misused by extremists. If it was societal rather than scientific, I would rate this 2.

**Response:** The revised proposal goes into more depth on the scientific contributions, including better methods to contextualize genotype visualizations and advancing knowledge through genomic overlay data.

[R3.4] Visualization: 7. Dimensionality reduction of core genomic data has been done, but adding site contribution overlays and uncertainty metrics is new. The visualization could be stronger if data was displayed in a new way rather than just adding overlays.

**Response:** I agree that a more transformative visualization approach could be valuable. I discussed exploring ensemble methods and plan to mention heat maps and point clouds for uncertainty visualization in the final proposal.

[R3.5] Significant: 3. The project addresses misinterpretation risks, which could save lives and protect people from biases.

[R3.6] Novel: 6.5. Having only 3 sources raises concerns about novelty, though the techniques for genotype matrices are new.

**Response:** I included more research in the final proposal and clarified the novelty of applying these methods specifically to genotype matrices.

[R3.7] Goals clearly stated: 2. Larger goals are clear.

[R3.8] Likelihood of Success: 3.5. High likelihood of success, but unclear implementation details.

**Response:** Please refer to [R2.4].

[R3.9] Strengths: The broader application is useful and will save lives.

[R3.10] Weaknesses: - Only 3 sources. - Lacks specific methodology. - Ambiguity around development environment and data preprocessing.

**Response:** I clarified the use of Python, VSCode, and libraries like Pandas, NumPy, and Scikit-learn, with preprocessing steps for normalization and data preparation.

#### **Review 4:**

[R4.1] Overall: 2

[R4.2] Interdisciplinary: 2. Combines scientific visualization with genetic data and motivations from social sciences.

[R4.3] Scientific: 4. The main impact can't be directly measured, since you can't measure how many shootings are avoided.

**Response:** While we can't measure this directly, we can measure how well the method guards against misinterpretation through a user study.

[R4.4] Visualization: 2. The visualization has a clear goal in what users should take away.

[R4.5] Significant: 2. This project could save lives, though it's hard to prove.

**Response:** Please refer to [R4.3].

[R4.6] Novel: 2. The visualization is novel for this data type with a unique motivation.

[R4.7] Goals clearly stated: 1. Goals are clear but could be more specific.

[R4.8] Likelihood of Success: 3. Visualization execution is likely, but evaluation method remains unclear.

**Response:** Please refer to [R1.8].

[R4.9] Strengths: Clear aims and vision, well-motivated.

[R4.10] Weaknesses: - Lacks discussion of evaluation method. - Difficulty in evaluating social impacts, as you can't survey extremists.

**Response:** Please refer to [R1.8].

## 2 Aims

Although current PCA/UMAP visualizations are useful for dimensionality reduction, they often lack key contextual information, which can lead to misinterpretation—particularly the overemphasis on genetic clusters, potentially reinforcing harmful narratives and biologically deterministic conclusions [1]. This project aims to improve traditional PCA/UMAP visualizations by making them more transparent, nuanced, and resistant to misinterpretation by users (i.e., genetic scientists and the general public). To achieve this, we will add genomic context (e.g., site contribution and environmental data overlays) and uncertainty measures (e.g., confidence ellipses, shaded regions, point clouds). We will validate these techniques through a user study and expert feedback. Specifically, this research focuses on three key objectives.

1. **Integrate Uncertainty Metrics:** We will add uncertainty regions (e.g confidence ellipses, point clouds) to indicate areas of the data that have less reliable differentiation. To determine such regions, we will use bootstrapping techniques, which will entail generating multiple subsamples of the data and performing PCA/UMAP on each one. By observing how the clusters shift across samples and projections, we will be able to quantify variability. Using bootstrapping, we can also generate ensemble visualizations to further inform the stability and consistency of our PCA/UMAP models.
2. **Incorporate Data Overlays::**
  - **Environmental Data Overlays:** We will add environmentally influenced data (e.g RNA, epigenetic data) overlays on top of our existing PCA/UMAP visualizations. Using color gradients or heatmaps, we will visualize how environmental data varies across clusters, demonstrating that different clusters are not biologically predetermined but dynamic and shaped by more than just inherited DNA.
  - **Site Contribution Overlays:** We will add site contribution overlays, which indicate specific genomic regions that contribute to clustering. In order to represent each site contribution, we will use color coded markers to signify how much specific genomic sites contribute to each cluster in a PCA/UMAP plot.
3. **Optimize based on User Study and Expert Feedback:** In order to ensure that these improvements effectively reduce misinterpretation, we will conduct a user study as well as gather expert feedback. Based on the results from both the user study and expert evaluations, we will refine and enhance the visualizations accordingly.
  - **User Study — Misinterpretation Detection Testing:** Using crowdsourcing platforms like Amazon Mechanical Turk or surveying our classmates, we will ask participants to interpret both traditional baseline visualizations and enhanced visualizations to identify potential misinterpretations and insights. User feedback will help us assess whether our enhanced visualizations reduce misinterpretation. We will iterate and revise designs based on this feedback and the patterns of misinterpretation we notice from the study, using metrics like task completion time and user scores to quantify results [2].
  - **Expert Feedback:** Our collaborators, Dr. Diaz-Papkovich and Dr. Ramachandram, will be consulted to evaluate the enhanced visualizations as well. Their expertise in genetic visualization will help ensure that the visualizations prevent misinterpretation while maintaining scientific and research value.

### 3 Significance

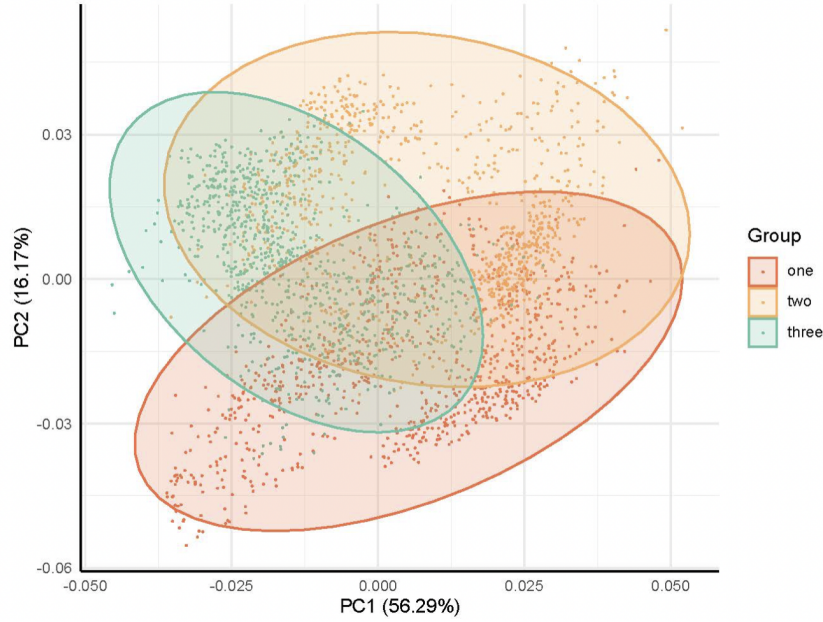


Figure 1: Example of proposed PCA visualization plotted with confidence ellipses, demonstrating variation and population overlap between groups

Visualizations in any field can be misinterpreted, but the repercussions for genotype visualizations are especially severe. Misinterpreted genetic data, as seen in the 2022 Buffalo shooting, can fuel extremism and domestic terrorism [3]. This is not an isolated issue, as this type of misuse of genomic data fuels extremist ideology [3]. To counter this, incorporating context-driven visualizations with overlays and uncertainty metrics is critical. Uncertainty regions will highlight overlap and variability, making it harder to draw deterministic conclusions. By explicitly showing uncertainty with confidence ellipses and point clouds, we enhance clarity and reduce the risk of extremist distortion, ensuring responsible scientific communication.

From a scientific perspective, this project will address the gap in existing visualizations in literature by proposing a novel visualization method that better contextualizes genetic diversity. Other genetic researchers will incorporate these methods proposed in this paper. Specifically, by integrating uncertainty metrics and incorporating data overlays, genetic researchers can better safeguard against misinterpretation while advancing their field forward. Furthermore, by incorporating site contribution overlays, users will gain a clearer understanding of the genetic drivers of clustering, enhancing scientific nuance and preventing oversimplification. More specifically, researchers will be able to pinpoint specific genomic regions that are driving genetic differentiation. Understanding which genetic markers drive clustering adds significant value to medical and genetic research, as different diseases and conditions may correlate with geographical groups [4]. This overlay also serves to combat misinterpretation by highlighting how localized genomic differences rather than broad genetic differences are what is driving clustering, contradicting extremist narratives of biological distinction. Lastly, environmental data overlays will help highlight how clusters are dynamic and reduce biologically deterministic overinterpretations. This will give users an enhanced understanding of how gene expression and environmental factors influence genotypic diversity, offering more nuanced, less rigid

insights. Ultimately, from a scientific, visualization, and broader perspective, this project will significantly advance the way genetic diversity is represented and understood.

## 4 Related Work

Genotype matrices are data structures that represent genetic variation across individuals in a population. Due to its high dimensional nature, Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) are both widely used dimensionality reduction techniques to visualize genotype data from massive genetic datasets [5] [6]. While PCA is a linear transformation that preserves global structure, UMAP is a non-linear technique that preserves primarily local structure [5]. By drawing on both PCA and UMAP techniques, our analysis will be more comprehensive and robust than existing methods which only use one visualization per study. We will gain better insights into broad trends (PCA) and detailed groupings (UMAP) depending on which one we use, leveraging their different strengths to strengthen our analysis.

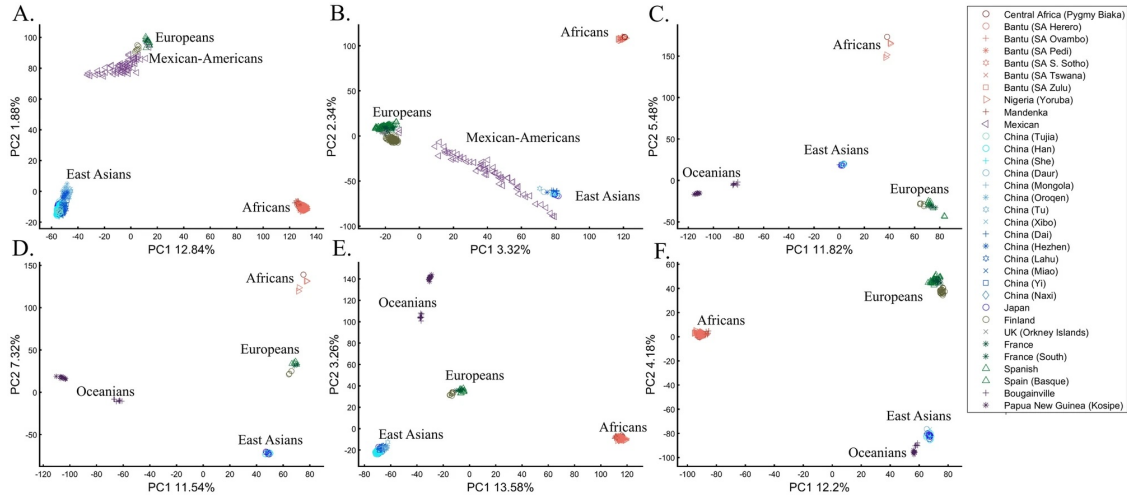


Figure 2: PCA plots of genotype data for different geographical groups under different sampling conditions [1]

However, existing approaches using these algorithms have significant limitations in terms of interpretability and uncertainty representation. For example, as seen in figure two, the PCA visualizations lack any uncertainty representations or additional genomic context. These current approaches tend to overemphasize clusters without any consideration for other types of contextual genetic data. In fact, none of these approaches directly address how environmental data or specific genomic regions influence clustering outcomes. Therefore, our visualizations are both more transparent by showing confidence in clustering and more scientifically robust by highlighting specific genomic contributions as opposed to vague, generalized groupings.

Based on these visualizations, clusters can be over-interpreted as biologically determined when, in reality, the boundaries between geographical groups are often fluid and more influenced by sampling, environmental factors, and statistical noise than rigid biological data [1]. In figure two, PCA results show that arbitrary sampling parameters can lead to African data plotted separately in its own section, leading to misin-

interpretations that reinforce racial hierarchical narratives. While mathematically valid, this can lead to cherry-picked interpretations. By integrating uncertainty metrics and using bootstrapping to generate PCA/UMAP ensemble visualizations, we aim to provide a clearer picture and avoid these misinterpretations. Previous studies have used bootstrapping for confidence intervals, supporting our bootstrapping methodology [7].

Many visualizations fail to follow standards that prevent misinterpretation. Some suggest that journals should require annotations indicating the proportion of genetic variation explained by the analysis, but this is rarely adopted [3]. Our method will include not only these annotations but also visual techniques like overlays and bootstrapping to better contextualize the data.

To implement PCA/UMAP, we will use tools like scikit-learn for machine learning and Plotly for interactive visualizations, which can handle overlays such as environmental or genomic data, making it ideal for our project [8] [9].

## 5 Research Plan

### Week 1: Data Preparation

Clean and integrate genotype and environmental data (e.g., SNPs, RNA) from 1000 Genomes Project.

*Tools:* Python, Pandas, NumPy.

*Deliverable:* Ready-to-use datasets for dimensionality reduction.

### Week 2: PCA/UMAP and Bootstrapping

Apply PCA/UMAP to data and begin generating bootstrapped samples for variability.

*Tools:* scikit-learn, umap-learn.

*Deliverable:* Initial projections and bootstrapped data.

### Week 3: Uncertainty Visualization

Add uncertainty metrics (confidence ellipses, point clouds) to visualizations.

*Tools:* matplotlib, seaborn.

*Deliverable:* Uncertainty-enhanced plots.

### Week 4: Environmental Data Overlays

Incorporate environmental data overlays into PCA/UMAP visualizations.

*Tools:* Plotly, matplotlib.

*Deliverable:* Visualizations with environmental layers.

### Week 5: User Study & Expert Feedback

Conduct user study (e.g. MTurk/Class Participants) and gather expert feedback.

*Tools:* Google Forms.

*Deliverable:* Quantitative/qualitative feedback on visualizations.

### Week 6: Final Presentation & Report

Finalize visualizations, complete the report, and prepare the presentation.

*Deliverable:* Final report and presentation incorporating feedback.

## References

- [1] E. Elhaik, "Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated," *Sci. Rep.*, vol. 12, article 14683, 2022, doi: 10.1038/s41598-022-14395-4.
- [2] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale, "Empirical Studies in Information Visualization: Seven Scenarios," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 9, pp. 1520–1536, 2012, doi: 10.1109/TVCG.2011.279.
- [3] J. Carlson et al., "Counter the weaponization of genetics research by extremists," *Nature*, vol. 610, no. 7932, pp. 444–447, 2022, doi: 10.1038/d41586-022-03252-z.
- [4] N. Brattig, R. Bergquist, D. Vienneau et al., "Geography and health: role of human translocation and access to care," *Infect. Dis. Poverty*, vol. 13, article 37, 2024, doi: 10.1186/s40249-024-01205-4.
- [5] A. Diaz-Papkovich, L. Anderson-Trocmé, and S. Gravel, "A review of UMAP in population genetics," *J. Hum. Genet.*, vol. 66, no. 1, pp. 85–91, Jan. 2021, doi: 10.1038/s10038-020-00851-4.
- [6] The 1000 Genomes Project Consortium, "A global reference for human genetic variation," *Nature*, vol. 526, pp. 68–74, 2015, doi: 10.1038/nature15393.
- [7] P. Moorjani and G. Hellenthal, "Methods for Assessing Population Relationships and History Using Genomic Data," *Annu. Rev. Genomics Hum. Genet.*, vol. 24, pp. 305–332, Aug. 2023, doi: 10.1146/annurev-genom-111422-025117.
- [8] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011. [Online]. Available: <http://jmlr.org/papers/v12/pedregosa11a.html>
- [9] Plotly Technologies Inc., "Plotly: The interactive graphing library for Python," Plotly, [Online]. Available: <https://plotly.com/python/>. [Accessed: Oct. 13, 2024].

# Musa Tahir

Providence, RI | 503-726-9202 | [musa\\_tahir@brown.edu](mailto:musa_tahir@brown.edu) | [linkedin.com/in/musa-tahir-78aa011ba](https://www.linkedin.com/in/musa-tahir-78aa011ba)

## SUMMARY

---

Brown Computer Science master's student with a background in software engineering, artificial intelligence, and business development. I'm eager to apply my software and data science skills to improve PCA/UMAP visualizations by incorporating uncertainty metrics and contextual overlays to prevent data misinterpretation.

## SKILLS

---

Languages & Frameworks: Python, Java, C, HTML/CSS, JavaScript, React, SQL, Tensorflow, PyTorch, Pandas

Development Tools: Git, Docker

## EDUCATION

---

**Sc.M. in Computer Science** – Expected May 2025

Brown University, Providence, RI

**Sc.B. in Computer-Science-Economics** (GPA: 3.95/4.00)

Brown University, Providence, RI

- Relevant Coursework: Software Engineering, Object Oriented Programming, Data Science, Deep Learning, Computer Vision, Machine Learning, Artificial Intelligence, User Interface & Web Development Design

## EXPERIENCE

---

**CANYAS – Portland, OR**

**Jul 2024 – Sep 2024**

*AI Software Engineering Intern*

- Designed and implemented a robust audio processing pipeline using OpenAI's Whisper and PyAnnote for transcribing audio with accurate speaker labels, improving accuracy by 9.5% and reducing Diarization Error Rate by 6.3%.
- Developed an intuitive Streamlit interface that enabled users to upload audio files and seamlessly view and download labeled transcripts

**INTEL CORPORATION, INTEL LABS – Portland, OR**

**Jun 2022 – Sep 2022**

*Engineering Intern*

- Developed and optimized RISC processors using ASIP Designer, customizing pipelines and reducing execution times by 70% with SHA-256 extensions and post-increment addressing
- Enhanced pipeline functionality by adding single-precision, floating-point support using the nML architecture description language

**SAFEMODE – Remote**

**Jun 2021 – Sep 2021**

*Business Development Intern*

- Developed Cost Savings Calculator using Google Sheets for fleet clients, demonstrating up to 30% in savings per fleet
- Enabled a 20% increase in client acquisition by producing weekly fleet KPI analyses

**INTEL CORPORATION, INTEL LABS – Portland, OR**

**Jun 2020 – Sep 2020**

*AI Software Engineering Intern*

- Built deep learning algorithms to predict yield of integrated circuits based on standard cell layout images using AutoKeras and Scikit-learn, improving model's performance over 5x and achieving R-Value over 0.7
- Resolved GPU memory constraints during pre-processing by deploying down sampling and pooling algorithms with Python and NumPy, effectively reducing dataset size by 95%

## ADDITIONAL EXPERIENCES

---

**Capstone – Optimized Multi-Style Neural Style Transfer, Computer Vision**

**Apr 2023 – May 2023**

*Team Lead*

- Implemented three different integration techniques of multi-style neural style transfer – a deep learning method that combines the content of one image with the style of others – for nuanced style representation using Python & TensorFlow
- Enabled user-driven prioritization of style images for enhanced artistic control using weighted losses in third method

**Put It On: Full Stack Outfit Recommendation Web App, Software Engineering**

**Nov 2022 – Dec 2022**

*Team Lead*

- Developed backend Java server with Spark, featuring API endpoints for storing, retrieving, and recommending personalized outfits based on weather and closet data
- Integrated OpenWeatherAPI to access live weather data, validating through comprehensive unit and integration testing



**Diaz-Papkovich, Alex**

to me ▼

Hi Tahir,

Hope this doesn't arrive too late--here are my slides.

Thanks,  
Alex.

# Three-dimensional Visualizations of Ecology LIDAR Waveforms

Matthew Yoon

PI

matthew\_yoon@brown.edu

Kei Yoshida

Co-PI

kei\_yoshida@brown.edu

Ziang Liu

Collaborator

ziang\_liu@brown.edu

October 13, 2024

## **Abstract**

We propose developing a Unity-based 3D visualization tool for NASA GEDI Lidar waveforms. By transforming them into spatial rotational meshes, this tool will allow interactive analysis of Relative Height metrics and enhance ecology research.

## Response to Reviewers

Thank you everyone for the reviews. This is very valuable insight on how I can make my final proposal stronger.

Reviewer 1: David Laidlaw

Reviewer 2: Sam Musker

Reviewer 3: Kei Yoshida

Reviewer 4: Richard Huang

Reviewer 5: Musa Tahir

**[R1.1] The broad scientific goals are compelling, although there is not much meaningful evaluation of them.**

*Response: I agree that the goals are lofty and they lack a proper evaluation framework. I will significantly narrow down the proposal focus and include clear methods of evaluating the effectiveness of my system (e.g., user studies, comparisons with 2D systems). I have many ideas and will speak to David for further discussion.*

**[R1.2] The broad visualization goals are similarly impressive, although vastly too large and insufficiently clearly evaluated.**

*Response: See R.1.1*

**[R1.3] There are no clear research contributions that seem likely to emerge as proposed**

*Response: Our current approach is too broad and unspecialized to yield meaningful research results for experts in the field. We'll focus on specific visual elements and their evaluation and downplay the haptic/audio components.*

**[R1.4] Haptics and sound are novel partly because they haven't been very useful compared to visuals. So, while they help with a novelty score, they lower several other scores.**

*Response: See R1.3*

**[R1.5] A feature list for software is articulated reasonably clearly. It is probably many months of work, at the very least. It also does not appear to have sufficient evaluation in the context of the value to scientists of the proposed features. A revised proposal with just the visual elements and their evaluation would likely score much better. The value of haptics that are possible today seems quite low, although it is certainly an intriguing concept. Sound could augment visuals, but the literature suggests that visuals are much higher bandwidth and probably the main channel to use.**

*Response: See R1.1 and R1.3. The proposal includes many inspirational papers, but lacks a baseline comparison (why my idea is better). On top of adjusting my research goals, the revised proposal will include more papers that create a baseline and will explain why it's better than that baseline.*

**[R1.6] This proposal is way too large to complete even a small portion. Focusing it significantly and pivoting it toward specific research evaluations clearly described would greatly increase enthusiasm.**

*Response: See R1.3*

=====

**[R2.1] Interesting and ambitious but utility could be low and risks to successful implementation.**

*Response: I agree that utility could be low, considering this tool is meant for experts in the field. Research contributions will be adjusted (one current idea is shifting focus to evaluation methods)*

**[R2.2] Does auditory and haptic data representation support scientific inferences effectively or is it just overwhelming?**

*Response: To answer your question, literature suggests that sound operates on top of space and can be effective in identifying variance or hidden patterns in the data. However, the raw utility of this is dubious and will be subject to further review before the final proposal.*

**[R2.3] Auditory and haptic data representations may be neglected for good reasons.**

*Response: See [R1.3]*

**[R2.4] How much is being built from scratch?**

*Response: Agreed that the scope of the project is too large. See [R1.1] and [R1.5]*

**[R2.5] It is unclear that haptic and auditory representations of data are useful. Is it easy to compare levels of haptic and auditory representations across different locations in the data? Are these representations stable across users? Is the representation overwhelming? Is it difficult to communicate about haptic and auditory representations? I suspect that there are good reasons why auditory and haptic representations of data are not commonplace, in humans these senses are serial and low precision.**

*Response: I agree. One of the papers admitted that a large part of their research was in evaluating users to figure out their parameters. This proposal lacks that evaluation framework that's essential to creating a useful tool. This will be addressed.*

**[R2.6] I also have concerns about the scope of the project and whether it is feasible in the short period, especially since the proposal states that the visual representation will be built in Unity and the first item on the work plan is to get familiar with Unity. It could easily take six weeks just to learn Unity.**

*Response: See [R1.1]*

=====

**[R3.1] The scientific contributions can be made more clear. How exactly ecological researchers would benefit from this new visualization that the existing methods do not offer? One paragraph in the Introduction ("Traditional 2D...") explains this, but this could be explained more in detail with more emphasis.**

*Response: The final proposal will address this issue and pivot focus towards creating an evaluation framework. See [R1.1]*

**[R3.2] Three goals stated make this work significant. The score could be improved by more clearly explaining the novelty and how exactly that helps researchers.**

*Response: Please see [R1.3].*

**[R3.3] The novelty can be inferred, but it is not made explicit as there are no references for existing work visualizing such representation.**

*Response. I agree, there lacks explicit baseline comparisons for this specific implementation. This will be addressed. See [R1.5]*

**[R3.4] Some of the Week 1 activities could be done before, as Week 2-6 activities may be somewhat ambitious. It is hard to gauge how achievable it is, and this could be improved by describing the work that has already been established previously.**

*Response: Scope will be narrowed down to make it more achievable in 6 weeks. Please see [R1.5]*

**[R3.5] This is good, so it could be explained more to motivate the study: "Even advanced 3D visualizations may not properly represent the multi-dimensional relationships between data points, often limited by user interface and screen immersion."**

*Response: Thanks for pointing this out! I think this could be a foundational idea in my evaluation framework (optimizing UI and screen immersion in VR).*

=====

**[R4.1] It wasn't clear to me what the scientific contribution or evaluation methods were besides this being a cool visualization.**

*Response: Please see [R1.3]*

**[R4.2] Similarly to the scientific criteria it wasn't clear to me why this should be made and who would benefit most from using it.**

*Response: Please see [R1.3] and [R1.5]*

**[R4.3] The methods don't seem novel but they are applied to new data.**

*Response: Please see [R2.1]*

**[R4.4] Not sure how this is being evaluated or what the author hopes users can measure or learn/take away from the visualization.**

*Response: Please see [R1.3] and [R2.1]*

=====

**[R5.1] The paper proposes a promising approach to visualizing complex ecological data with scientific and practical applications. The project is ambitious, so a methodical plan will be important to ensure its completion within six weeks.**

*Response: I agree that the current plan is very ambitious for 6 weeks. The focus will be narrowed down significantly, likely towards user evaluation methods.*

**[R5.2] The scientific contribution is strong, but the proposal would benefit from a more in depth explanation of how this tool would improve pattern recognition beyond traditional methods. The proposal could also more directly tie the tool to impacting ecological research outcomes.**

*Response: This proposal can definitely benefit from head-to-head comparisons to existing work to explain how this tool would improve recognition. This will be included for the final submission.*

**[R5.3] The visualization techniques are innovative, combining 3D visualizations with auditory and haptic feedback. I'm curious, however, what specific steps will be taken to avoid overwhelming the user. Furthermore, the proposal could benefit from more details on how the auditory and haptic feedback would function in practice.**

*Response: This will be addressed. One paper that implemented a similar system included a large part for user studies to ensure there's no overstimulation, and I believe shifting focus to evaluation could provide a large contribution.*

**[R5.4] The proposal would have a significant impact on ecological research. A more thorough analysis on how the system's effectiveness would be measured would also strengthen its significance. Furthermore, more emphasis on how this tool would be utilized by the broader scientific community would improve the proposal.**

*Response: Please see [R5.3]*

**[R5.5] The novelty lies in the way the tool combines auditory, visual, and haptic feedback to represent complex data. The proposal could provide more details on the technological challenges of integrating all these components together and potentially shed light on why this combination hasn't been attempted or overlooked.**

*Response: Please see [R2.1]*

**[R5.6] Limited discussion on evaluation/validation method for how system will improve data comprehension. Proposal would benefit from more detail on how**

**feedback will be incorporated in the iteration process to design and refine the system**

*Response: Please see [R5.2] and [R1.1]*

# 1 Aims

The Global Ecosystem Dynamics Investigation (GEDI) mission by NASA has revolutionized how we analyze Earth's ecosystems. By scanning the Earth's surface using Lidar, GEDI data captures detailed vertical profiles of forest canopies and other information about vegetation structure, biomass distribution, and ecological habitats.

Currently, ecology researchers at Brown University use 2D maps to display scalar values such as canopy height and Relative Height metrics like RH98 (the height where 98% of the waveform energy is returned). They also compare individual Lidar waveforms in isolation. While these visualizations provide valuable information, they have notable limitations:

1. Lack of spatial context: analyzing individual waveforms and 2D maps separately disconnects the data from its spatial environment, which makes it hard to perceive patterns across a geographical area.
2. Dimensional constraints: 2D representations of high-dimensional data struggle to effectively display critical ecological features, like under-canopy habitats or canopy layering.
3. Combining attributes: 2D tools make it difficult to effectively integrate multiple data attributes (different RH levels, terrain features, scalar values like canopy) into a cohesive visualization. This places a ceiling on how much information a 2D visualization can convey.

The lack of a 3D environment that spatially places NASA GEDI Lidar waveforms limits researchers' ability to analyze important spatial patterns within the data, which could hinder advancements in ecological research and applications like wildlife conservation and fire spread analysis.

The aim of this proposal is to develop a Unity-based platform that visualizes GEDI waveforms inside an immersive virtual reality environment. Specifically, we have the following goals:

1. Create 3D representations of Lidar waveforms. We will transform individual Lidar waveforms into 3D generalized cylinders by revolving the waveform data around a vertical axis. This produces detailed meshes that accurately represent wave amplitude at various RH levels.
2. Visualize RH metrics. We plan to compute and interpolate multiple RH metrics to create wireframe-esque continuous surfaces for each height. For example, RH50 would have a surface that reflects mid-canopy structures relevant to animal habitats.
3. Incorporate detailed terrain models using scalar value texturing. Our model will apply textures to the ground surface to display additional scalar data such as entropy or soil moisture.
4. Evaluate the efficacy of this tool. Because of its novelty, we will collect input from ecology researchers to make sure this platform is useful.

We hope that by creating a new 3D visualization tool that spatially places GEDI Lidar waveforms, researchers are able to better understand forest structures and improve data interpretation.

# 2 Significance

## Scientific Contributions

The development of such a platform holds substantial significance in ecological research and environmental studies. We address critical limitations in current data analysis by opening new avenues for scientific

research. Specifically, spatial representation of Lidar waveforms allows for a more comprehensive understanding of forest structures. This includes detailed information about canopy height, density, and layering, which are essential for analyzing climate modeling, biodiversity conservation, and sustainable resource management [1].

Additionally, by spatially visualizing different RH metrics, researchers can identify and analyze crucial habitats for various species. For example, RH50 can show mid-canopy spaces that act as shelters or migration paths for animals, which can aid wildlife conservation efforts. Similarly, analyzing vegetation density through RH metrics can improve models predicting fire spread to mitigate wildfire risk.

### **Visualization Contributions**

From a visualization perspective, this proposal introduces novel techniques to represent high-dimensional waveform data. The use of generalized cylinders to represent Lidar waveforms is a novel approach that transforms abstract data into tangible spatial structures. Additionally, implementing filters for different RH levels and opacity will allow researchers to interact with the data dynamically, increasing engagement and deeper analytical insights.

Most importantly, the methodologies developed can be adapted for other types of spatial or environmental data, making this proposal significant beyond the direct scope of this project.

## **3 Related Work**

Rendering Lidar point cloud datasets in virtual reality is a well-established practice in the field of 3D visualization and spatial data analysis. Many studies have focused on optimizing data pipelines to efficiently process and render Lidar data that allows display of both minimally and additionally processed point clouds, as well as developing mechanisms to render more accurate physical aspects of spaces [5]. These efforts have significantly improved the rendering of physical spaces by improving performance, handling large datasets, and developing specialized algorithms to increase realism of 3D reconstruction from point clouds.

Current literature emphasizes techniques for managing large amounts of data generated by Lidar scans. For example, methods have been developed such as data decimation, level-of-detail rendering, and spatial indexing to optimize rendering performance while minimizing compromise [4].

However, the visualization of raw Lidar waveforms, representing the returned energy as a function of distance, is pretty unexplored. Most existing tools and techniques focus on returning discrete Lidar data by simplifying waveforms into individual points. Although this simplification is helpful to many researchers, it results in loss of detailed structural information contained in the full waveform [3][2]. Specifically, the raw waveform data provides a continuous signal return, which captures subtle variations in vegetation structure like tree density, branching patterns, and undergrowth characteristics [2].

Currently, researchers resort to analyzing individual waveforms in isolation without spatial context. The gap in literature shows a clear need for visualization tools that can display raw Lidar waveforms with 3D spatial information. By working with raw data, researchers can directly interact with the complete dataset and potentially find hidden relationships that may be otherwise hidden.

## **4 Research Plan**

I will need access to the visualization lab, VR headsets, and VR-ready computers.

**Week 1:** Acquire and preprocess NASA GEDI Lidar waveform data. This includes studying the structure and format of GEDI data, and developing scripts to parse and preprocess the waveform data. We will

also convert data into a format for Unity such as CSV or JSON.

**Week 2:** Unity environment setup and basic 3D visualization. We will create basic waveform representation by using simple 3D objects such as vertical bars to represent individual waveforms at their geographic locations, and map amplitude to color or height as a preliminary proof-of-concept. Also implement camera and navigation controls.

**Weeks 3-4:** Implement advanced 3D waveform visualizations. Here, we'll use mesh generation for waveforms, interpolate RH metrics, create top surfaces, and adjust textures for data attributes such as reflectivity or entropy.

**Week 4:** Collect feedback from ecology researchers and implement relevant changes.

**Week 5:** Add interactivity, additional layers, and incorporate feedback. We will develop UI controls to filter and display various RH levels (RH98, RH50, etc) and opacity controls. Also importing terrain elevation data and generate a terrain mesh using tessellated triangles.

**Week 6:** Documentation and presentation. This includes final testing and optimization.

## References

- [1] R. Dubayah, J. B. Blair, S. Goetz, L. Fatoyinbo, M. Hansen, S. Healey, M. Hofton, G. Hurtt, J. Kellner, S. Luthcke, et al. The global ecosystem dynamics investigation: High-resolution laser ranging of the earth's forests and topography. *Science of remote sensing*, 1:100002, 2020.
- [2] A. Hovi, D. Schraik, N. Kuusinen, T. Fabiánek, J. Hanuš, L. Homolová, J. Juola, P. Lukeš, and M. Rautiainen. Synergistic use of multi-and hyperspectral remote sensing data and airborne lidar to retrieve forest floor reflectance. *Remote Sensing of Environment*, 293:113610, 2023.
- [3] C. Mallet and F. Bretar. Full-waveform topographic lidar: State-of-the-art. *ISPRS Journal of photogrammetry and remote sensing*, 64(1):1–16, 2009.
- [4] R. Schnabel and R. Klein. Octree-based point-cloud compression. *PBG@ SIGGRAPH*, 3:111–121, 2006.
- [5] R. Tredinnick, M. Broecker, and K. Ponto. Experiencing interior environments: New approaches for the immersive display of large-scale point cloud data. In *2015 IEEE Virtual Reality (VR)*, pages 297–298. IEEE, 2015.

[1] [5] [4] [3] [2]

# Matthew Yoon

(571) 598-2123 | matthew\_yoon@brown.edu

## EDUCATION

---

### Brown University

*B.S. Computer Science and Applied Mathematics*

Expected graduation: May 2026

*Providence, RI*

- **GPA:** 4.0/4.0
- **Relevant Coursework:** 2D Game Engines, Advanced Visualization, Computer Graphics, Software Engineering, Machine Learning, Data Structures and Algorithms, Computational Linear Algebra, Discrete Math

## EXPERIENCE

---

### Software Development Intern

*NASA | National Aeronautics and Space Administration*

June 2024 – August 2024

*Kennedy Space Center, FL*

- Intern on Simulations & Modeling team with Jacobs under NASA COMET contract
- Developed an algorithm to decode embedded data streams containing CH10 telemetry for the ICPS rocket utilizing SQL for database management and Bash for script automation
- Built a real-time simulation sniffer using Python, dpkt, and tcpdump to extract packets, and implemented multiprocessing with stdin/stdout interface handling

### Computer Graphics Researcher

*CRaGL | George Mason University*

May 2023 – Aug. 2023

*Fairfax, VA*

- Research intern for Computational Reality Creativity & Graphics Lab under Yotam Gingold
- Developed programming proficiency in differentiable rendering and optimization in the context of geometric texture analysis, leveraging Mitsuba3 renderer and Dr.Jit compiler
- Conducted literature review on 3D mesh generation and hand-object interaction to guide research on inverse grasp synthesis for extended reality applications

### Undergraduate Researcher

*Virginia Tech*

Oct. 2021 – June 2022

*Blacksburg, VA*

- Built a lightweight data processor using Matplotlib, pandas, and NumPy to efficiently analyze thousands of empirical data points, resulting in substantial time savings for 20+ researchers
- Investigated electrochemical effects of niobium pentoxide cathode coating on reaction kinetics, rate capability, and coulombic efficiency of Li-NMC-532 cells

## PROJECTS

---

### Flight Simulator | *React, TypeScript, Three.js, CSS*

December 2023

Built a React and Three.js-based flight simulator with infinite procedurally-generated terrain, accurate physics, and raycaster interactions

Implemented Perlin noise landscape and chunk generation/deletion for an efficient and dynamic environment

Designed flight mechanics and physics for realistic plane movement, including forward vector interpolation based on z-axis rotation

### Ray Tracing Engine | *C++, OpenGL*

October 2023

- Engineered a GPU-accelerated ray tracer featuring implicit shape definitions, UVs, transformations, and texturing
- Shot per-pixel rays for intersection and reflection calculations, employing Phong lighting model for realistic illumination
- Calculated surface normals and implemented UV coordinate mapping for photorealistic lighting and texturing

## SKILLS

---

**Languages:** Python, Java, C++, C#, Typescript, MATLAB

**Frameworks/Libraries:** React, OpenGL, Three.js, PyTorch, pandas, NumPy

**Tools/Skills:** Unity, CI/CD, Git, Linux, Bash, PowerShell, real-time rendering, full-stack, back-end

**Soft:** Exceptional communicator, meticulous attention for detail, execution-focused, and robust organization skills

---

**Extracurriculars:** Brown Cycling Team, Hack@Brown, Brown Outing Club

# Subcircuit visualization for enhancing neural network interpretability

Sam Musker  
PI

`samuel_musker@brown.edu`

Aalok Sathe  
Co-PI

`aalok_sathe@brown.edu`

October 12, 2024

## **Abstract**

We propose to develop visualization methods for neural network subnetwork identification techniques, aiding experts in scientific discovery. By revealing and visualizing the internal subcircuits responsible for specific tasks, our work will enhance understanding of neural networks and support advances in neuroscience and cognitive science.

# 1 Response to reviewers

## Response to reviewers

Reviewer 1, Matthew Yoon

R 1,1: I'm not too familiar with neural networks, so I'm unsure about this claim: "research is interdisciplinary due to the clear relevance to neuroscience and cognitive science." However, from the proposal, this seems like a cross-disciplinary project between visualization and machine learning, which are both CS. I could be wrong, but the fact I'm uncertain indicates that this could be an area of clarity in future proposal drafts.

Response: The proposal has been updated to make clearer the connections between the project and disciplines such as neuroscience and cognitive science, and to make clear what research questions in cognitive science can be answered using the proposed tool.

R 1,2: It's not clear exactly how these subnetworks will be visualized.

Response: The proposal has been updated to clarify what visualizations will be produced.

R 1,3: It's not exactly clear what the new "interpretability tools" will be for subnetwork identification

Response: The proposal has been updated to clarify that new tools for subnetwork identification won't be developed. Rather, existing nascent tools for subnetwork identification will be used as inputs to a new visualization tool.

R 1,4: I'm unsure if the scope of this proposal can be completed in 6 weeks.

Response: The proposal has been amended to include a research plan that clarifies how the project goals can be reached within the given timeframe.

R 1,5: Considering NSF will have experts reviewing this, this is not necessarily a weakness, but you assume the reader already knows a lot about NNs, which makes this difficult to read.

Response: The proposal now includes two survey articles on neural network interpretability and visualization respectively to provide background to a broader audience.

Reviewer 2: Richard Huang

R 2,1: This project combines visualization with AI research. It felt more like CS vis + CS ai; I'm missing the deeper connection with cog sci/neuro besides neural nets representing real brain processes.

Response: Please see the response to R 1,1 above.

R 2,2: I feel like lots of AI research already involves plotting and visualizing data, which is what this proposal proposes to do.

Response: The proposal has been updated to emphasize what features of the visualization will be novel and what insights can be generated from such visualizations that cannot be derived from existing visualization tools.

R 2,3: I think the author should argue more about what needs to be improved by Lepori et al. and why.

Response: The proposal has been updated to more clearly characterize the shortcomings of the (very preliminary) visualization work included in Lepori et al.

R 2,4: The research plan was missing, hard to judge

Response: The proposal has been updated to include a research plan.

R 2,5: The usability of the tool is fairly limited to neural net researchers, possibly not accessible to others.

Response: The proposal has been updated to explain how the tool could be useful for non-experts (for example, in pedagogical applications).

Reviewer 3: Thais Del Rosario Hernandez

R 3,1: It's a bit unclear if the proposed visualization tool aims to aid in identifying and extracting subnetworks, improve the interpretability of already extracted subnetworks, or both.

Response: See the response to R 1,3 above.

R 3,2: The related work section does a great job at outlining the current state of visualization for neural networks. However, the proposal seems to draw inspiration from each citation without going in depth about how the proposed tool will be built as a whole to solve a specific problem. This score could be improved by compiling the differing aims from each citation into a concise set of aims (2-3) in the aims section.

Response: The proposal has been updated to more clearly synthesize project aims and relate these to prior literature.

R 3, 3: the approach to tackle the problem is not outlined either in the aims nor research plan.

Response: The proposal has been updated to more clearly describe the visualization approach and methodology.

R 3,4: Likelihood of Success: Difficult to estimate without a research plan.

Response: See the response to R 2,4 above.

Reviewer 4: Yang Xiang

R 4,1: If achieved success, the tool will push the ai theory a large step. May be really hard though. [...] The final goal seems too hard

Response: The proposal has been updated to clarify that existing tools for subnetwork identification will be used. This significantly increases the likelihood of success of the project.

R 4,2: didn't see the tool and steps detail [...] Maybe more specific weekly plans and tool feature design should be added.

Response: The proposal has been amended to clarify the steps involved in tackling the problem and to more clearly characterize the envisaged output.

## 2 Aims

New neural network interpretability techniques have emerged in recent years, providing the opportunity to build associated visualization tools that help researchers to derive insights into neural networks.

One such development involves techniques to identify the part of a neural network that is active in solving a particular task or subtask. These techniques involve training a mask over neural network parameters to ablate a certain functionality. For example, if a neural network is trained to identify object shape and color, then a mask can be trained such that with the mask applied the neural network is still capable of identifying object color but is no longer able to identify object shape. The masked parameters then can be reconstructed into a subnetwork that is responsible for shape identification alone. One such technique is introduced by Lepori et al. [2023]. A need exists for better tools to visualize the subcircuits that are identified by nascent tools as being responsible for particular task components.

The proposal is therefore to develop visualization methods to operate with new interpretability tools, in particular subnetwork identification techniques. No new subnetwork identification techniques will be developed. Rather, existing nascent subnetwork tools will be utilized, on top of which new visualization tools will be developed.

The approach is to visualize the neural network at the parameter level as a graph structure with nodes and edges representing neurons and their connections. All connections will be shown in the visualization, with the subnetworks corresponding to different subtasks shown in different colors. The shading of the subnetwork colors will increase through training as the subnetworks become more entrenched, and this will be shown diachronically as a GIF or animation through training. Overlap between subnetworks will be shown as the color produced by the mixing of the different subnetwork colors. This methodology will ensure that users can view the formation of subnetworks through time during training, including the overlap between subnetworks.

The specific visualization aims are as follows: 1. Visualize subnetwork structure at the parameter level (this is not present in Lepori’s visualization). 2. Visualize multiple subnetworks simultaneously. 3. Visualize the overlap between subnetworks. 4. Visualize the change in subnetworks over time during training (this is not present in Lepori’s visualization).

The visualization tool will be compared to the rudimentary visualization present in Lepori et al. [2023]. Due to the novelty of subnetwork analysis in general, we are not aware of other subnetwork visualization tools. Those authors present a visualization that shows what proportion of parameters in each layer of a network is dedicated to one subnetwork, another subnetwork, their overlap, and neither subnetwork. However, this visualization does not show subnetwork strength and structure and does not show any changes over time during training. Such information is necessary for drawing substantive conclusions about the nature of the subnetworks. For example a highly disparate subnetwork may indicate a learned solution that patches together various heuristics that may not generalize, whereas a compact subnetwork structure may indicate a more integrated solution mechanism. Such structure is not visible in Lepori’s existing visualization. We aim to present a visualization that is better than the existing benchmark due to providing the above information that is not visible in existing visualizations. A sketch of the proposed visualization is shown in Figure 1 and an output of the visualization tool due to Lepori is shown in Figure 2.

The evaluation methodology will utilize an insights-based approach, using the framework and definition outlined by Saraiya et al. [2005]. In particular, we will seek to learn whether expert users are able to derive insights into neural subnetworks that are qualitatively different or more informative than the insights that can be derived from using the visualizations presented by Lepori et al. [2023]. This will be done through a three-part process, in which users are taught how to use the tool, encouraged to use the tool on a learning task, and

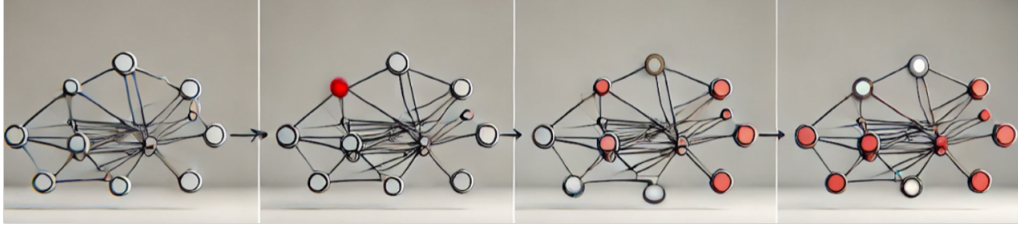


Figure 1: A sketch of the proposed visualization, representing frames of a GIF or animation. Created using Dall-E and Adobe Firefly.

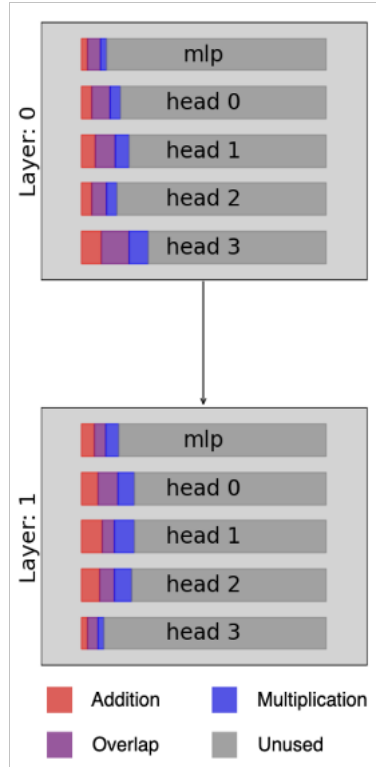


Figure 2: The rudimentary visualization due to Lepori et al. [2023].

then interviewed by the PI and co-PI. The interview process will seek to establish what insights the user has gained into neural subnetworks, and these insights will be coded according to various metrics (e.g. novelty and scientific value). The coded insights will be compared to the insights that can be derived from using the existing subnetwork visualization tool. Expert users to be assessed will include the sponsoring faculty, being Roman Feiman (Psychology) and Ellie Pavlick (Computer Science / Linguistics). Expert users to be assessed will also include graduate students from Brown’s Language Understanding and Representation (LUNAR) lab.

The research is interdisciplinary due to the clear relevance to neuroscience and cognitive science. A better understanding of neural subnetworks and their changes over time during training can inform a better understanding both of how neural mechanisms emerge to solve different tasks, and how different tasks are related to each other by virtue of being solvable with shared underlying mechanisms.

### 3 Significance

The development of advanced visualization methods for subnetwork identification techniques holds substantial significance for both the field of artificial intelligence and its interdisciplinary applications. As neural networks become increasingly complex and are deployed in critical sectors, the need to understand their internal workings becomes increasingly important.

Firstly, enhancing interpretability directly addresses one of the paramount challenges in deep learning: the “black box” nature of neural networks. By visualizing the specific subcircuits responsible for distinct tasks or functionalities within a model, researchers can gain insight into how information is processed and represented. This transparency is crucial for diagnosing and mitigating biases, understanding decision-making processes, and ensuring that models behave as intended.

Secondly, the proposed work has the potential to accelerate scientific discovery by facilitating cross-disciplinary collaboration. The ability to map and visualize subnetworks aligns closely with neuroscientific methods of studying brain functionality. By drawing parallels between artificial neural networks and biological neural systems, this research can contribute to understanding in neuroscience and cognitive science. For example, a cognitive scientist could use subnetwork visualization methods to identify that two diverse tasks can be solved by a network using a common subnetwork or two largely overlapping subnetworks, thus supporting hypotheses regarding the relatedness of those tasks.

Moreover, from an engineering perspective, these visualization tools can aid in optimizing neural network architectures. Identifying and extracting subnetworks dedicated to specific tasks can lead to more efficient models that require fewer resources, which is particularly valuable for deployment in resource-constrained environments. It can also inform the development of modular AI systems, where components can be independently analyzed, tested, and improved.

Last, the subnetwork visualization tool can help to build confidence in the reliability of the underlying subnetwork tool - or to aid in discovering ways in which these methods are not reliable. If the subnetwork identification tool “invents” a subnetwork that does solve the task but is not used by the original network to do so, then evidence of this should be available through our visualization tool. In particular, we hypothesize that if the subnetwork identification tool is unreliable in this way then the subnetwork visualization will show large frame-by-frame differences between the identified subnetwork, corresponding to a different non-utilized subtask “solution” being discovered each time.

In summary, the proposed visualization methods are significant because they enhance the interpretability of neural networks to promote knowledge and trust, they may yield insights for neuroscientists and cognitive scientists, and they may potentially contribute to optimizing AI systems and stress-testing the relied upon subnetwork identification tools. While the tool is primarily intended for expert users in a scientific discovery context, the tool might be appropriate for a secondary use case as a pedagogical tool for demonstrating how neural networks learn subtask solutions.

### 4 Related Work

Enhancing neural network interpretability is an important aim that may support scientific, engineering, and safety goals [Zhang et al., 2021]. Numerous visualization techniques have arisen to aid researchers in understanding the internal processing mechanisms in neural networks [Matveev et al., 2021]. Such techniques include, for example, the early decoding of intermediate layers of a convolutional neural network (CNN) to allow a user to view a read-out of a processed image at a certain stage of the network [Wang et al., 2021]. New interpretability tools have arisen which provide the opportunity to develop corresponding visualization

tools.

Interactive visualization tools have been developed to help users understand neural networks at different levels of expertise. For instance, Wang et al. [2021] introduced *CNN Explainer*, an interactive tool designed for non-experts to better understand the inner workings of CNN models. Their work focuses on pedagogical purposes and is tailored to CNN architectures. In contrast, our work is intended for an expert audience and is not limited to specific neural network architectures.

Recent advances in subnetwork identification have provided new avenues for neural network interpretability. Lepori et al. [2023] introduced *NeuroSurgeon*, a toolkit for identifying the subnetworks involved in a neural network solving a particular part of a composite task. This technique enables the extraction of subnetworks responsible for specific functionalities, offering a granular view of neural network operations. Our proposed visualization tools aim to complement such analysis tools, facilitating the exploration and understanding of these identified subnetworks.

Visualization of parameter changes during training has also been explored to enhance interpretability. Schneider et al. [1997] presented methods for real-time visualization of interactive parameter changes in image processing systems. Their approach allows users to observe how parameter adjustments affect processing outcomes dynamically. Similarly, although applied to image processing systems, their methodology informs our intention to include methods for visualizing diachronic parameter changes within subnetworks of neural networks, thereby aiding in understanding how subnetworks evolve during training.

Attention mechanisms in neural networks, particularly in transformer models, have been a focus of visualization efforts to understand model activations in response to specific inputs. Vig [2019] developed a multiscale visualization of attention in transformer models, providing insights into how models process input data. However, their work centers on activations, which are input-specific. Our work differs by concentrating on visualizing the learned weights of neural networks, which are input-agnostic parameters that define the network’s behavior irrespective of specific inputs.

Feature visualization is another approach that has been employed to interpret neural networks. Olah et al. [2017] provided techniques for visualizing the features learned by a neural network as intermediate representations. This line of work sheds light on what features a network learns at different layers, enhancing understanding of the network’s hierarchical feature extraction. Our focus, however, is on identifying and visualizing the circuits or subnetworks that extract particular features, thus moving from feature representation to the structural basis of feature extraction within the network.

Dimensionality reduction techniques have been utilized as visualization tools for data analysis. Nasser et al. [2006] explored the use of Kernel Principal Component Analysis (KPCA) for visualizing data clusters, aiding in the identification of the number of clusters within datasets. Their work emphasizes visualization of features of the data rather than aspects of the model itself. In contrast, our project aims to visualize features of the trained network, specifically the subnetworks responsible for particular tasks, thereby contributing to model interpretability rather than data analysis.

Beyond neural networks, subnetwork identification has been examined in the context of complex networks. Gao et al. [2023] proposed a method for identifying key nodes in complex networks based on subnetwork feature extraction, utilizing graph convolutional networks. While their work addresses general complex networks and focuses on key node identification, it highlights the broader applicability and importance of subnetwork analysis. Our project is aligned with this perspective but is specifically tailored to neural networks, aiming to visualize and understand the subnetworks within these models.

Subnetwork analysis has also found significant applications in biological networks. Su et al. [2010] identified diagnostic subnetwork markers for cancer in human protein-protein interaction networks, demonstrating how subnetwork identification can aid in disease classification and understanding of biological pro-

cesses. Although our work is focused on neural networks, the methodologies and motivations share common ground with such biological network studies, emphasizing the utility of subnetwork visualization in complex systems analysis.

In summary, there is a rich body of work on neural network interpretability and visualization, ranging from educational tools for non-experts to advanced techniques for visualizing features, activations, and parameter changes. The exploration of subnetwork identification in both artificial and biological networks underscores the importance of understanding the internal structures that underpin complex functionalities. Our proposed work seeks to build upon these efforts by developing visualization methods tailored to emerging subnetwork identification techniques in neural networks, thereby providing experts with powerful tools to facilitate scientific discovery and deepen our understanding of neural network architectures and their functionalities.

## 5 Research Plan

Week 1:

Create and train a neural network to solve a task that can be broken down into two sub tasks Retrieve parameter values and create a static visualization of the final parameter values for the full network

Week 2:

Modify final parameter visualization into a dynamic visualization of parameter values as they change through training

Week 3:

Apply existing subcircuit identification methodology to identify subnetworks responsible for each task component

Week 4:

Layer subcircuit identification onto the parameter visualizations to introduce different shadings for different subcircuits, changing through the course of training

Week 5:

Conduct pilot user study with expert users (Ellie and Roman) and assess the insights derived

Week 6:

Write final report Create final presentation Submit final report and give final presentation

## References

- Luyuan Gao, Xiaoyang Liu, Chao Liu, Yihao Zhang, Giacomo Fiumara, and Pasquale De Meo. Key nodes identification in complex networks based on subnetwork feature extraction. *Journal of King Saud University - Computer and Information Sciences*, 35(7):101631, 2023. ISSN 1319-1578. doi: <https://doi.org/10.1016/j.jksuci.2023.101631>. URL <https://www.sciencedirect.com/science/article/pii/S1319157823001854>.
- Michael A. Lepori, Ellie Pavlick, and Thomas Serre. Neurosurgeon: A toolkit for subnetwork analysis, 2023. URL <https://arxiv.org/abs/2309.00244>.
- S. A. Matveev, I. V. Oseledets, E. S. Ponomarev, and A. V. Chertkov. Overview of visualization methods for artificial neural networks. *Computational Mathematics and Mathematical Physics*, 61(5):887–899, May 2021. ISSN 1555-6662. doi: 10.1134/S0965542521050134. URL <https://doi.org/10.1134/S0965542521050134>.
- Alissar Nasser, Denis Hamad, and Chaiban Nasr. Kernel pca as a visualization tools for clusters identifications. In Stefanos Kollias, Andreas Stafylopatis, Włodzisław Duch, and Erkki Oja, editors, *Artificial Neural Networks – ICANN 2006*, pages 321–329, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-38873-9.
- Christopher Olah, Ludwig Schubert, and Alexander Mordvintsev. Feature visualization. *Distill*, 2017. URL <https://distill.pub/2017/feature-visualization/>.
- Purvi Saraiya, Chris North, and Karen Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE transactions on visualization and computer graphics*, 11:443–56, 07 2005. doi: 10.1109/TVCG.2005.53.
- Wolfgang Schneider, Wolfgang Eckstein, and Carsten T. Steger. Real-time visualization of interactive parameter changes in image processing systems. In Georges G. Grinstein and Robert F. Erbacher, editors, *Visual Data Exploration and Analysis IV*, volume 3017, pages 286 – 295. International Society for Optics and Photonics, SPIE, 1997. doi: 10.1117/12.270324. URL <https://doi.org/10.1117/12.270324>.
- Junjie Su, Byung-Jun Yoon, and Edward R. Dougherty. Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network. *BMC Bioinformatics*, 11(6):S8, Oct 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-S6-S8. URL <https://doi.org/10.1186/1471-2105-11-S6-S8>.
- Jesse Vig. A multiscale visualization of attention in the transformer model. *CoRR*, abs/1906.05714, 2019. URL <http://arxiv.org/abs/1906.05714>.
- Zijie J. Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng Polo Chau. Cnn explainer: Learning convolutional neural networks with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1396–1406, 2021. doi: 10.1109/TVCG.2020.3030418.
- Yu Zhang, Peter Tiño, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021. doi: 10.1109/TETCI.2021.3100641.

## Samuel Musker

195 Waterman Street | Providence, RI | 267.303.9670 | [musker.sam@gmail.com](mailto:musker.sam@gmail.com) / [samuel\\_musker@brown.edu](mailto:samuel_musker@brown.edu)

### EDUCATION

---

#### Brown University, Providence, RI

Ph.D. student, August 2020 - Present

- Studying towards a Ph.D. in Computer Science and A.M. in Philosophy, GPA: **4.00** out of **4.00**

#### University of Pennsylvania, Philadelphia, PA

B.A., August 2017 - May 2019

- Major in Philosophy (Philosophy and Science concentration), minors in both Mathematics and Logic, Information, and Computation, overall GPA: **3.93** out of **4.00** / in-major GPA: **4.00** out of **4.00**
- Academic achievements: Phi Beta Kappa; Honors and Distinction in Philosophy major; Flower Prize for the best essay in Philosophy; Dean's List 2017-2018 and 2018-2019; Allan Gray Orbis Foundation Fellowship 2017-2018 and 2018-2019

#### University of Cape Town, Cape Town, South Africa

February 2015 – June 2017 (transferred, no degree)

- Registered for B.Sc. Eng. Electromechanical Engineering (2015-2016) / B.Sc. Philosophy and Applied Mathematics (2017), Weighted Average: **81.77%** out of **100%** (First-Class pass is 75%)
- Academic achievements: Dean's List 2015 and 2016; Isaac Ochberg and VC's scholarships 2015, 2016, and 2017; Allan Gray Orbis Foundation Entrepreneurial Innovation Fellowship 2015, 2016, and 2017; Top student Electromechanical Engineering program as at leaving

### EMPLOYMENT

---

#### Boston Consulting Group, Boston, MA | *Associate 1, promoted to Associate 2*

July 2019 – July 2020

- Consulted to a major US pharmacy chain, producing work for executive-level management at the fortune-20 company. Built sophisticated models and workflows in Excel and Alteryx with visualizations in Tableau
- Individually developed data processing and modelling tool to assist BCG North America in reaching reduction targets of GHG emissions from consultant flights. Collaborated with data analytics expert to develop a webapp frontend to interface with the tool. The tool assisted BCG C-suite in decision making, and having exceeded expectations was successfully exported to other global regions including Europe, Asia Pacific, and Australia/New Zealand as the gold-standard flight emissions modelling tool

#### Boston Consulting Group, Johannesburg, SA | *Visiting Associate*

June 2017 - July 2017

- Responsible for drafting infrastructure portion of the company's public report on economic development in East Africa

#### UCT, Cape Town, SA | *Course TA (Mechanics of Solids I - MEC2025F)*

February 2017 - June 2017

- The only undergraduate as one of five TAs for second-year level course; responsible for grading and leading tutorial sessions

#### Debating work (freelance and volunteer), SA | *Adjudicator, Coach, and Selector*

February 2015 - June 2017

- National Deputy Chief Adjudicator; National Team Coach; Western Cape Provincial Selector; Coach at two high schools

### ENTREPRENEURSHIP and INNOVATION

---

- Co-founded the UCT branch of Effective Altruism (February 2017 – June 2017), a global movement founded by philosophers which aims to promote rationality in decision-making about how to do the most possible good
- Secured and completed prestigious 4-year Allan Gray Orbis Foundation entrepreneurship program in parallel with undergraduate degree, inducted as fellow; fellowship provided ~\$24,000 funding towards undergraduate degree
- Finalist pitch at Allan Gray Orbis Foundation national competition (July 2016), for a concept designed to extend top-quality tertiary education opportunities to the developing world using innovative technology and new campus models
- Best pitch at Startup Safari entrepreneurial immersion tour to India (January 2016), for a technical and business concept to support mobile adoption in rural developing contexts, which is a significant driver of socio-economic outcomes
- Category winner (housing and settlement studies) at Eskom International Expo for Young Scientists (October 2012), for an engineering design of a sustainable and packable tent alternative for refugee camps, which often become de facto settlements

### SKILLS, LANGUAGES, and INTERESTS

---

- Programming languages: Python, Java, C, MATLAB, SQL, VBA
- Other tools: Advanced Excel modelling, data processing with Alteryx, data visualization with Tableau
- Languages: English (first language), French (proficiency commensurate with four college courses with an A-grade in each)
- Other interests: soccer, long-distance running, hiking, and abstract photography; UCT and UPenn Outdoors Clubs 2015-2019

### COURSEWORK SUMMARY

---

- Philosophy, excluding Logic: 15
- Mathematics and Formal Logic: 13
- Mechanical Engineering and Applied Physics: 10
- Electrical Engineering and Computer Science: 6
- Basic science (Physics, Chemistry, Neuroscience): 4
- French Language: 4
- Humanities: 2

# Aalok Sathe

address: CIT, Brown University, Providence, RI 02912

email: [asathe\[at\]mit.edu](mailto:asathe[at]mit.edu), [aalok\[at\]brown.edu](mailto:aalok[at]brown.edu)web: [aalok-sathe.gitlab.io](https://aalok-sathe.gitlab.io)

## Education

2029 (exp.) (Aug '24–)	<b>Ph.D.</b>	<b>Brown University</b> (Providence, RI, USA) Computer Science	
May 2022 (Jan '22–)		<b>Massachusetts Institute of Technology</b> (Cambridge, MA, USA) Advanced Studies Program (ASP): Introduction to Neural Computation	
May 2021 (Aug '17–)	<b>B.S.</b>	<b>University of Richmond</b> (Richmond, VA, USA) <i>Majors:</i> Computer Science (hons.), Cognitive Science <i>Minors:</i> Linguistics, Math; Church-Kent prize: outstanding grad in CS	GPA: 4/4, rank 1/775
May 2020 (Jan '20–)		<b>University of Edinburgh</b> (Edinburgh, Scotland, UK) Informatics; Philosophy, Psychology & Language Science (PPLS)	

## Research Experience

- Graduate Researcher, **Brown University**: *Computer Science; Carney Institute for Brain Science* Providence, RI  
Mentors: Ellie Pavlick, Michael J. Frank Sep 2024 – present
- Research Associate, **Massachusetts Institute of Technology**: *Brain & Cognitive Sciences* Cambridge, MA  
Mentors: Evelina Fedorenko, Noga Zaslavsky, Greta Tuckute, Anna Ivanova, Cory Shain Jul 2021 – Jun 2024
- Research Intern, **Microsoft Research**: *NLP group* Bengaluru, India  
Mentors: Monojit Choudhury, Somak Aditya May 2020 – Aug 2020
- Undergrad Research Assistant, **University of Richmond**: *Math & Computer Science; Psychology* Richmond, VA  
Mentors (Math+CS): Taylor Arnold, Joonsuk Park, Prateek Bhakta, Heather Russell Aug 2018 – May 2021  
Mentors (Psychology): Cindy Bukach, Matthew Lowder Sep 2017 – Dec 2020

## Peer-Reviewed Publications

&co-first author, #alphabetically listed ([why?](#)), [Google Scholar profile](#)

- Language use is only sparsely compositional: The case of English adjective-noun phrases in humans and LLMs  
**Sathe, A.**, Fedorenko, E., Zaslavsky, N.  
Annual Meeting of the Cognitive Science Society, 46 (CogSci 2024). [\[paper\]](#)
- Conventional and Frugal Methods of Estimating COVID-19-Related Excess Mortality and Undercount Factors  
Dedhe, A., Chowkase, A., Gogate, N., Kshirsagar, M., Naphade, R., Naphade, A., Kulkarni, P., Naik, M., Dharm, A., Raste, S., **Sathe, A.**, Kulkarni, S., Bapat, V., Joshi, R., Deshmukh, K., Lele, S., Manke, K., Cantlon, J., Pandit, P.  
Scientific Reports, 14.1: 10378 (2024). [\[paper\]](#)
- Driving and suppressing the human language network using large language models  
Tuckute, G., **Sathe, A.**, Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., Kay, K., Fedorenko, E.  
Nature Human Behavior (2024). [\[paper\]](#)
- Block symmetries in graph coloring reconfiguration systems  
Bhakta, P.#, Krehbiel, S.#, Morris, R.#, Russell, H. M.#, **Sathe, A.**#, Su, W.#, Xin, M.#  
Advances in Applied Mathematics, 149: 102556 (2023). [\[paper\]](#)



🔍 laidlauw



99+

## Compose

Mail

## Inbox

11,522

Starred

Snoozed

Sent

## Drafts

39

More

## Labels

Fwd: Course project concept for CSCI 2370, Visualization

External

Inbox x



David Laidlaw

to me

----- Forwarded message -----

From: **Roman Feiman** <[roman\\_feiman@brown.edu](mailto:roman_feiman@brown.edu)>

Date: Wed, Sep 18, 2024 at 11:44AM

Subject: Re: Course project concept for CSCI 2370, Visualization

To: David Laidlaw <[laidlaw.david@gmail.com](mailto:laidlaw.david@gmail.com)>

Cc: Ellie Pavlick <[ellie\\_pavlick@brown.edu](mailto:ellie_pavlick@brown.edu)>

Hi David,

Sam has mentioned this to us, and absolutely -- I think this would be a valuable visualization tool that I'd be interested in using.

Best,  
Roman

On Wed, Sep 18, 2024 at 10:02AM David Laidlaw <[laidlaw.david@gmail.com](mailto:laidlaw.david@gmail.com)> wrote:

Hi Ellie and Roman

Sam Musker, a student in my visualization class, has proposed a class project to use visualization to aid neural network int  
interested in the project's results -- a user. I've attached his more detailed summary of the project idea. Would the two of y  
Please let me know if any more info would help in deciding.

Thanks for considering!

Cheers,

-David

—

David Laidlaw, Professor, Brown Computer Science  
Box 1910, Providence, RI 02912, +1-401-354-2819  
<http://www.cs.brown.edu/~dhl>



🔍 laidlaw



99+



--  
Roman Feiman

Thomas J. and Alice M. Tisch Assistant Professor  
Department of Cognitive and Psychological Sciences  
Program in Linguistics  
Brown University  
Room 241, 190 Thayer St., Providence, RI  
p: 617-834-7008

Room 241, 190 Thayer St., Providence, RI  
p: 617-834-7008

## Labels

David Laidlaw, Professor, Brown Computer Science  
Box 1910, Providence, RI 02912, +1-401-354-2819  
<http://www.cs.brown.edu/~dhl>



to me

Cc: Roman Feiman <[roman\\_feiman@brown.edu](mailto:roman_feiman@brown.edu)>

Absolutely! Our lab would definitely use something like this. (I told Sam to do it...he is my student...not sure if that is against the

Reply

Forward

# 1 Response to Reviewers

Dear Editor and Reviewers,

We are grateful for the time and effort that you have dedicated to providing feedback on our proposal. We have prepared a point-by-point response to your comments and suggestions, which can be found below.

## 1.1 Reviewer 1: Richard Huang

### Overall: 3

Interdisciplinary: 2

The project combines sci vis with cardiac anatomy.

Scientific: 3

The project provides the first non-invasive visualization method for certain areas of the heart. I am curious what makes this non-invasive.

**Response:** *We have elaborated on the protocol used to obtain this data in the Biological Reference section.*

Visualization: 2

They apply a technique (tractography) mainly used in brain anatomy visualization, but has also been used in the heart anatomy. This project's visualization will be the first of QSI atrial images.

Significant: 2

It is significant in that it not only visualizes an important organ of the body but does so in a non-invasive way. This makes the techniques transferrable beyond mice.

Novel: 3

(As stated previously) The project provides the first non-invasive visualization method for certain areas of the heart. The techniques seem to not be novel.

Goals clearly stated: 3

The 3 aims are clearly stated at the beginning with the necessary tools. I think it would help to write more on the motivation for these goals at the beginning before stating them.

**Response:** *We have expanded on our motivations and hopes for this project in the Aims section.*

Likelihood of Success: 3

I am wondering if the author needs a specific facility to perform the visualization? I.e. do they need to be at a lab where the mouse hearts are in person? This wasn't clear to me. The weekly plan is well-defined and has a baseline to reference.

**Response:** *We have provided more details about the generation of the data (see previous response addressing Scientific Contributions)*

Additional Comments:

"Maybe a brief discussion on animal experimentation ethics would be helpful."

**Response:** *We have added more references regarding the use of animal models in heart disease research.*

”More can be said about how this impacts human health research.”

**Response:** *We expand on the future implications of our proposal in the Significance section*

## 1.2 Reviewer 2: Samuel Musker

**Overall: 3**

Interdisciplinary: 1

Clear medical application

Scientific: 3

Diagnosing heart conditions is important but is the macrostructure of the heart understudied?

**Response:** *We have added more references outlining how both the macro- and micro-structures within the atria (and atrioventricular connectivity) have yet to be properly defined to our Related Works section*

Visualization: 6

“Our project will produce the first visualization of atrial images acquired using QSI. We will use tractography software DSI Studio [3] and TrackVis [4] to determine and optimize tractography parameters that capture macrostructures and substructures in different regions of the atria and their surrounding vessels.”

Is the project relying too heavily on DSI Studio and TrackVis to create visualizations of the data?

**Response:** *We thoroughly agree with the reviewer on this comment, and have adjusted our methodology to 1. evaluate other available tools and 2. more thoroughly describe a novel visualization interface that combines the most appropriate features from these tools. We have also outlined the current limitations of existing tools, especially regarding cardiac fiber visualization and using small animal models.*

Significant: 3

If the tool can support insights into heart disease diagnosis then it would be very significant but I don’t know how much this will improve relative to current understanding.

**Response:** *We have expanded on future implications of our work.*

Novel: 4

New data used for a heart visualization but there are several existing heart visualizations with other data

**Response:** *We have emphasized how our data differs from previously published data, especially in the context of atrial visualization.*

Goals clearly stated: 1

Good clarity breaking it down into three

Likelihood of Success: 1

Well scoped project

### 1.3 Reviewer 3: Yang Xiang

#### Overall: 2

Interdisciplinary: 2

The project is combining visualization and biology in a great way, utilizing the research results of both sides to create a new interdisciplinary finding.

Scientific: 3

Assessing the myoarchitectural patterns of mouse atrium is scientific and interesting, may be used to extract some insights from the complex raw data.

Visualization: 3

The first visualization of atrial images acquired using QSI. Some softwares are also used during the visualization, which may enhance the visual quality while keeping good efficiency. A new interface will be created too.

Significant: 4

The application may not have so many real-world applications, as it is restricted in a small field.

**Response:** *Refer to previous response to Reviewers 1 (Additional comments) and 2 (Significant contributions).*

Novel: 3

The method is novel, combining QSI with advanced visualization techniques. While MRI has been studied a lot, the project chooses a subdirection and aims to create a new tool for it.

Goals clearly stated: 1

Plans are detailed and clear.

Likelihood of Success: 2

With careful planning and clear knowledge of the project, it is highly likely to succeed.

### 1.4 Reviewer 4: Arjun Prakash

#### Overall: 2

Interdisciplinary:

Combines biology and imaging

Scientific: 2.

The scientific contribution doesn't seem as clear since it is applying this existing technique on a new subject? although maybe the hope is to find something significant with this technique. I think this is aim 3.

**Response:** *The reviewer's assessment is correct - we have made the goals and outcomes of aim 3 more clear throughout the proposal.*

Visualization: 2

the visualization objectives are very clear in the Aims.

Significant: 2

Well motivated and clear articulation of the gap

Novel: 1

Seems like this is a new use of this diffusion tractography which could lead to new insights

Goals clearly stated: Yes

Likelihood of Success: 8/10

Additional Comments:

"Not clear what the benchmark or evaluation would be"

**Response:** *Refer to previous response to Reviewer 2 (Visualization Contribution).*

## 1.5 Reviewer 5: Eric Xia

**Overall: 4**

Interdisciplinary: 4.

The research is interdisciplinary, utilizing high-resolution imaging techniques in a new setting. The computational contribution is very minimal. The proposal elects to focus primarily on the novel application of existing software to data. The multimodal interface is only briefly mentioned in Aim 2, and not elaborated upon.

**Response:** *We have expanded the description of our computational contributions - namely, optimization of tractography parameters and building a visualization interface capable of interpreting and combining data from multiple sources (i.e. exported data from current visualization tools), incorporated as part of aims 1 and 2, respectively.*

Scientific: 3.

The identification of architectural features in the atrial region of the heart could have widespread applications in diagnosing disease-related structural changes. The combination of the tractogram model with anatomical renderings could allow for identification of architectural patterns, furthering our understanding of mammal anatomy. This section of the heart is known to be highly complex, which motivates the development of novel visualization methods to aid understanding its structure. However, there is no justification provided for the proposed interface as an improvement over previous methods.

**Response:** *Refer to previous response to Reviewer 2 (Visualization contribution).*

Visualization: 6.

The proposal will primarily use existing software, DSI Studio and TrackVis to visualize the atrial chamber of the mouse heart, specifically with a focus on atrial fibrillation. It proposes the development of a multimodal

interface with overlaid anatomical renderings in order to study the atrial region ex vivo. This would be a novel visualization application, as QSI tractography can achieve higher resolution imagery than previous techniques. The overlay interface could be described in more detail. In particular, the extensibility of the interface to other problem domains is unclear.

**Response:** *We have provided a more detailed description of the interface, as well as some comments on possible future use cases. We thank the reviewer for this insightful suggestion.*

Significant: 3.

The primary visualization contribution would be in the interface. It would potentially allow for the identification of patterns beyond the scope of the scientific contribution outlined. However, it is not especially clear what anatomical renderings would add to the visualization, as the goal seems to characterize novel architecture in the atria region. For the scientific contribution here, the use of the mouse as a model organism may complicate generalizations to the human body.

**Response:** *We appreciate the reviewer's concern - we believe that anatomical renderings contextualize and provide an important "ground truth" to visualizations. Additionally, most of the cited papers in this proposal make use of anatomical models to complement other visualization methods.*

Novel: 5.

The analysis would be a novel application of QSI to an understudied region of the heart. However, the proposed interface overlay does not noticeably benefit from the technology being employed: the anatomical models, assuming they are from existing sources, will not provide additional information. The focus of the proposal is clearly on the application of existing technology.

**Response:** *We further emphasize the benefits of our proposed interface by identifying the shortcomings of current tools in processing and visualizing our type of data. See also previous response to Reviewer 2 (Visualization contribution).*

Goals clearly stated: 1.

The goals are clearly stated and well defined: the first visualization of atrial images acquired using Q-space diffusion MRI.

Likelihood of Success: 2.

Assuming that the software supports QSI imagery, the visualization of the data does not seem to be problematic. However, the development of a multimodal interface with anatomical overlays seems to be tangential to the focus of the proposal.

# Architectural Analysis of the Mouse Atrium using Q-Space MRI

Thais Del Rosario Hernandez

PI

thais\_del\_rosario\_hernandez@brown.edu

Richard Gilbert, MD

Collaborator

rjgilbert12@gmail.com

October 14, 2024

## Abstract

We propose to utilize imaging data obtained with high geometric resolution Q-space MRI (QSI) to assess the intricate myoarchitectural patterns of the mouse atrium. Our approach seeks to address the limitations of lower-resolution imaging frameworks, offering a novel perspective on the spatial organization and connectivity of cardiac tissue.

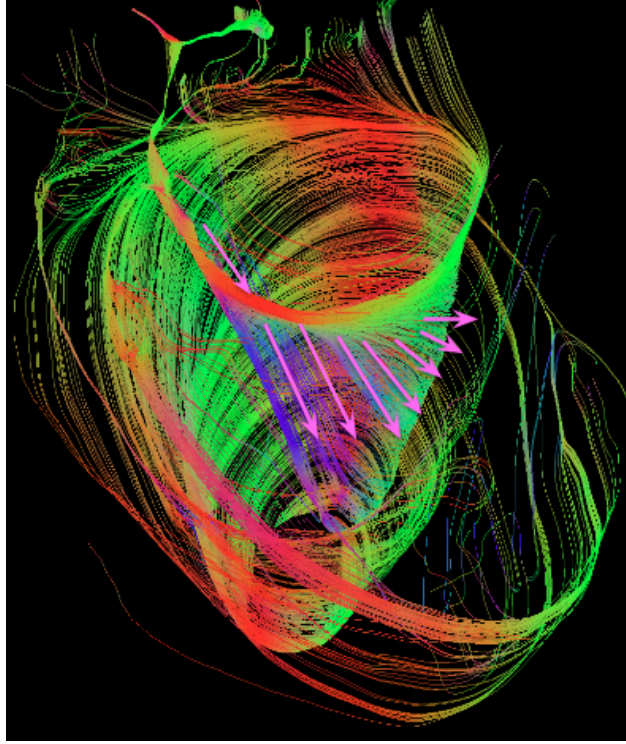


Figure 1: Fiber architecture of the ventricular wall in a mouse heart [1].

## 2 Aims

The atria are the upper chambers of the heart responsible for receiving blood from the veins and distributing electrical signals throughout the heart. Our goal is to derive a preliminary atlas that optimally maps fiber tracts to disease-relevant regions of the atria. We will use mouse hearts imaged ex vivo using QSI to accomplish three main goals:

**Aim 1:** Generate diffusion tractography images of the mouse atrium employing raw diffusion DICOM files obtained from previously imaged normal mouse hearts (n=10 hearts).

**Aim 2:** Develop a multimodal interface to visualize tractogram models overlaid with anatomical renderings of the heart.

**Aim 3:** Assess dominant architectural patterns in the right/left atria of the mouse hearts in the vicinity of the major vessels, and in relation to the anatomical locations of the sinoatrial & atrioventricular nodes.

This data has been previously analyzed in the context of the ventricular regions of the heart (Figure 1), which have a well defined helix structure. The atrium displays a higher degree of architectural complexity, making it more difficult to characterize as a single-patterned structure. Our hope is to define a set of relevant architectural features that can be later used to detect disease-related structural changes.

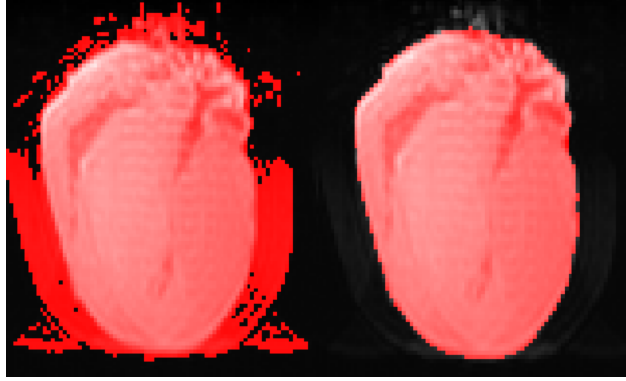


Figure 2: Examples of different masking and thresholding parameters for processing data in DSI Studio.

### 3 Significance

#### 3.1 Biological Reference

Atrial fibrillation (AF) is the most common clinical arrhythmia, affecting more than 33 million people around the world. AF is characterized by aberrant electrical signaling and disorganized waves. Fiber orientation is an important factor in electrical propagation in the heart, which is divided into four chambers and further sectioned into the upper (atria) and lower (ventricles) chambers. While the ventricles have been extensively studied with a wide range of imaging techniques - including QSI [1] -, the atria remain mainly studied by destructive methods such as tissue sectioning. Non-invasive imaging methods prove difficult due to the thinness of the atrial wall and the high complexity of fiber organization. This project seeks to fill this gap by first investigating atrial architecture in the mouse - a well established model organism used to study AF [2]. Moreover, mouse hearts are easier to acquire and image than human hearts while still providing relevant findings for clinical applications [3, 4]. The main data for this project was obtained from excised mouse hearts imaged using high resolution QSI.

#### 3.2 Tractography Optimization

Our project will produce the first visualization of atrial images acquired using QSI. The raw data (DICOM files) will be processed using three software widely used for diffusion MRI data: Diffusion Toolkit [5], DSI Studio [6], and Quantitative Imaging Toolkit (QIT) [7]. The processing parameters have been well defined for white matter (i.e. brain) studies, but not yet explored in the context of cardiac imaging. Specifically, we will investigate reconstruction parameters such as manual and precomputed masking, image thresholding, reconstruction methods, and diffusion sampling length ratio (Figure 2). We believe it is paramount to first evaluate and establish a method for processing our raw files, particularly when brain and heart structure are notably different in terms of tissue composition and functional dynamics.

We will then use tractography software TrackVis [5], and the visualization tools in DSI Studio and QIT to determine and optimize tractography parameters that capture macrostructures and substructures in different regions of the atria and their surrounding vessels (Figure 3). We will also highlight tract clusters of interest and investigate their anatomical connection(s) by overlying tracks with histology slices and 3D morphology models. Selected models and data files will be exported from each corresponding software and compiled in a comprehensive interface using Paraview and/or Dipy. Altogether, this novel visualization will allow us to

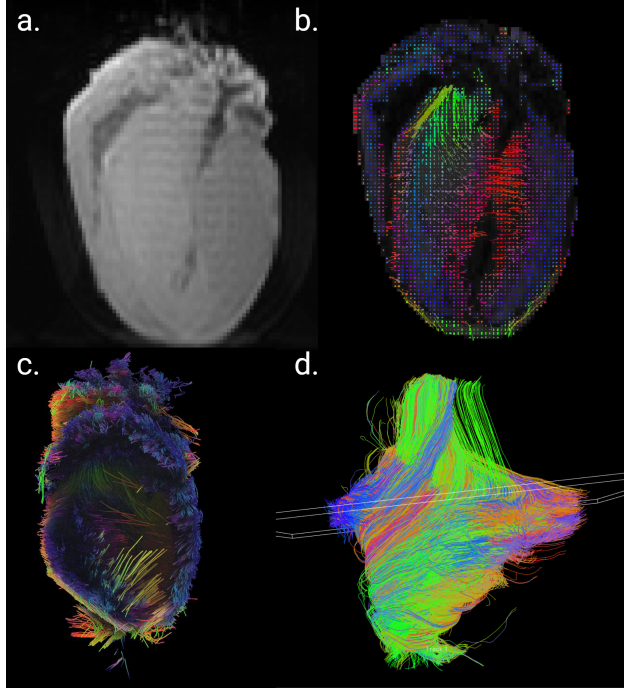


Figure 3: Preliminary visualizations of raw data (a), 2D reconstruction results with fiber orientation overlay (b), fiber microstructure (c), and macrostructure (d) using our mouse heart data.

extract meaningful insights from the intricacy of high resolution tract data.

## 4 Related Work

Tractography is a technique mainly used to explore brain structure - specifically, white matter tracts. However, it has been previously used in the context of the heart, though mostly to investigate ventricular architecture. There is still a prominent need for further research in the upper region of the heart and its mechanistic effects, especially in the context of AF [8]. Previous work exploring the atrium suffers from limitations regarding the resolution of the imaging methodologies [9, 10]. In contrast, QSI tractography is sensitive to microstructural changes occurring at the subvoxel scale, allowing us to demonstrate fiber tracts at the scale of single myocytes in rodent models, and therefore should possess the geometric resolution needed to resolve atrial myocardial architecture.

Furthermore, tractography analysis has been used to quantify differences in cardiac fiber alignment in a mouse model of congenital hypertrophic cardiomyopathy [11]. We strive to institute a useful reference for comparison between healthy and pathological models of cardiac disease.

## 5 Research Plan

- **Week 1:** Process DICOM files and generate tractography data; document optimized parameters.
- **Week 2:** Compare and validate the tract generation parameters with the previous analysis in the ventricular wall [1].

- **Weeks 3-4:** Implement overlay interface showing tracts over anatomical 3D renderings of the heart. Highlight anatomical regions of interest based on prior literature.
- **Week 5:** Identify and annotate regions of inter-sample variation. Begin documenting findings.
- **Week 6:** Finalize presentation and write-up.

## References

- [1] E. N. Taylor et al. “Alterations in Multi-Scale Cardiac Architecture in Association With Phosphorylation of Myosin Binding Protein-C”. In: *Journal of the American Heart Association* 5.3 (2016). DOI: 10.1161/JAHA.115.002836.
- [2] D. Schüttler et al. “Animal Models of Atrial Fibrillation”. In: *Circulation Research* 127.1 (2020). DOI: 10.1161/CIRCRESAHA.120.316366.
- [3] Breckenridge R. “Heart failure and mouse models”. In: *Disease models mechanisms* 3.138–143 (2010). DOI: 10.1242/dmm.005017.
- [4] Hasenfuss G. “Animal models of human cardiovascular disease, heart failure and hypertrophy”. In: *Cardiovascular research* 39.60-76 (1998). DOI: 10.1016/s0008-6363(98)00110-2.
- [5] R. Wang and V.J. Wedeen. *TrackVis*. URL: <https://trackvis.org/>.
- [6] F. Yeh. *DSI Studio*. URL: <http://dsi-studio.labsolver.org>.
- [7] Ryan P Cabeen, David H Laidlaw, and Arthur W Toga. “Quantitative imaging toolkit: software for interactive 3D visualization, data exploration, and computational analysis of neuroimaging datasets”. In: *ISMRM-ESMRMB Abstracts* (2018), pp. 12–14.
- [8] O. Berenfeld et al. “Animal models of human cardiovascular disease, heart failure and hypertrophy”. In: *Cardiovascular research* 39.60-76 (1998). DOI: 10.1016/s0008-6363(98)00110-2.
- [9] F. Pashakhanloo et al. “Myofiber Architecture of the Human Atria as Revealed by Submillimeter Diffusion Tensor Imaging”. In: *Circulation. Arrhythmia and electrophysiology* 9.4 (2016). DOI: 10.1161/CIRCEP.116.004133.
- [10] R. Kamali et al. “Contribution of atrial myofiber architecture to atrial fibrillation”. In: *PloS one* 18.1 (2023). DOI: 10.1371/journal.pone.0279974.
- [11] T. T. Wang et al. “Resolving myoarchitectural disarray in the mouse ventricular wall with diffusion spectrum magnetic resonance imaging”. In: *Annals of biomedical engineering* 38.9 (2010). DOI: 10.1007/s10439-010-0031-5.

# Thaís Del Rosario Hernández

123 Packard St. Cranston, RI 02910

(401) 837-9992 • [t.delrosarioh@gmail.com](mailto:t.delrosarioh@gmail.com) • [www.linkedin.com/in/thaís-del-rosario-59b13114a/](http://www.linkedin.com/in/thaís-del-rosario-59b13114a/)

## EXPERIENCE

**Senior Research Assistant,** *Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence, RI.* July 2021 - Present

- Acquired and analyzed high-throughput behavioral data for a research program on the discovery of novel treatments for neurodegenerative disorders. Developed and revised new methods for behavioral tracking and classification. Performed drug exposure experiments in zebrafish larvae using commercially available, large scale drug libraries.
- Prepared manuscripts and generated figures for multiple scientific publications.
- Mentored undergraduate and graduate students. Assisted them with the development, data acquisition and analysis, and materials for their specific research projects. Instructed them about laboratory safety, experimental guidelines, and kept track of their progress with regular meetings and opportunities to present and discuss their data.
- Maintained zebrafish and mice colonies, keeping track of experimental groups and transgenic lines. Managed project supply inventories and communicated with vendors about the best solutions for our laboratory needs.
- Constantly ensured bio-chemical and radiation safety compliance, as well as IACUC-compliant animal use for research.

**Research Laboratory Technician,** *Department of Neurology, University of Massachusetts Medical School, Worcester, MA.* August 2020 – June 2021

- Worked on the development of AAV based gene therapies for neurodegenerative diseases using small animal models. Performed tasks at each step of this cycle, from experimental design to production and administration of viral vectors, molecular and behavioral data collection, analysis, interpretation, and presentation of results.
- Performed downstream procedures for the acquisition of outcome measures – in experiments both in vivo and in vitro –, such as collection of tissues, RNA/DNA isolation, RT-qPCR/qPCR, and imaging assays.
- Maintained mouse colony and managed corresponding colony records, encompassing various transgenic mouse strains. Trained staff on colony management software usage.
- Created and maintained a cross-platform, relational database software for the synchronous collection of behavioral data.

**Undergraduate Research Assistant,** *Carleton College Biology Department, Northfield, MN.* January - August 2019

- Assembled, annotated, and queried metagenomic data from deep sea hydrothermal vents using various bioinformatic tools.
- Analyzed processed data to investigate the abundance of mobile genetic elements in taxonomically classified deep sea samples.

**Undergraduate Research Assistant,** *Carleton College Computer Science Department, Northfield, MN.*

June - August 2018

- Implemented a pipeline for the generation of simulated next-generation DNA sequencing data for heterogeneous cancer tumors using online databases of somatic and germline mutations as templates.

---

## SKILLS

**Laboratory Techniques:** Behavioral testing, Mammalian cell culture, Competent cell preparation, Plasmid engineering, DNA assembly and cloning, AAV vector production, Fluorescence microscopy.

**Programming Languages:** Proficient in Python, Java, C#, C.

**Data Analysis and Visualization:** RStudio, GraphPad Prism • EthoVision XT • Fiji, ImageJ.

**Technological Experience:** Microsoft Office (Word, Excel, PowerPoint, and Outlook) • SQL, JUnit, HTML, CSS, JavaScript • Bash • DeepLabCut • FileMaker Pro.

**Computer Graphics:** Blender • Adobe Photoshop, Adobe Illustrator.

**Computational Biology:** BLAST, Bowtie2, IGV, IDBA-UD, Prokka, Samtools, Virsorter.

---

## PUBLICATIONS

- **Del Rosario Hernandez T**, Joshi N, Gore SV, Kreiling JA, Creton R. Combining supervised and unsupervised analyses to quantify behavioral phenotypes and validate therapeutic efficacy in a triple transgenic mouse model of Alzheimer's disease [preprint]. 2024 June. Available from: <https://doi.org/10.1101/2024.06.07.597924>
- Abrams KB, Wilson A, **Del Rosario Hernandez T**, Choate A. Virtual Reality-Based Simulated Hallucinations to Enhance Empathy Toward Individuals With Schizophrenia. *J Nerv Ment Dis*. 2024 Jun 1; 212(6):312-316. doi: 10.1097/NMD.0000000000001772
- Gore SV, **Del Rosario Hernandez T**, Creton R. Behavioral effects of visual stimuli in adult zebrafish using a novel eight-tank imaging system. *Front. Behav. Neurosci.*, 2024 Mar 10; 18:1320126. doi: 10.3389/fnbeh.2024.1320126.
- Zhong J, Osborn T, **Del Rosario Hernandez T**, Kyrasyuk O, Tully BJ, Anderson RE. Increasing transposase abundance with ocean depth correlates with a particle-associated lifestyle. *mSystems*. 2024 Feb 21; 9(3):e0006724. doi: 10.1128/msystems.00067-24
- **Del Rosario Hernandez T**, Gore SV, Kreiling JA, Creton R. Drug repurposing for neurodegenerative diseases using Zebrafish behavioral profiles. *Biomed Pharmacother*. 2024 Feb 17;171:116096. doi: 10.1016/j.biopha.2023.116096
- **Del Rosario Hernández T**, Joshi NR, Gore SV, Kreiling JA, Creton R. An 8-cage imaging system for automated analyses of mouse behavior. *Sci Rep*. 2023 May 19;13(1):8113. doi: 10.1038/s41598-023-35322-1. PubMed PMID: 37208415; PubMed Central PMCID: PMC10199054.
- Gore SV, Kakodkar R, **Del Rosario Hernández T**, Edmister ST, Creton R. Zebrafish Larvae Position Tracker (Z-LaP Tracker): a high-throughput deep-learning behavioral approach for the identification of calcineurin pathway-modulating drugs using zebrafish larvae. *Sci Rep*. 2023 Feb 23;13(1):3174. doi: 10.1038/s41598-023-30303-w. PubMed PMID: 36823315; PubMed Central PMCID: PMC9950053.
- Tucker Edmister S, **Del Rosario Hernández T**, Ibrahim R, Brown CA, Gore SV, Kakodkar R, Kreiling JA, Creton R. Novel use of FDA-approved drugs identified by cluster analysis of behavioral profiles. *Sci Rep*. 2022 Apr 21;12(1):6120. doi: 10.1038/s41598-022-10133-y. PubMed PMID: 35449173; PubMed Central PMCID: PMC9023506.

---

## EDUCATION

**Carleton College** - *Northfield, MN.* September 2016 - June 2020  
Bachelor of Arts in Computer Science, Bachelor of Arts in Biology

## QSI and atrial myoarchitecture

External

Inbox x



**richard gilbert**

to me ▾

Sun, Sep 22, 5:13 PM



Dear Thais:

I am looking forward to getting started on our project. I have attached:

- 1) Two seminars describing QSI & assessment of architectural array/disarray.
- 2) Several papers defining QSI tractography for architectural analysis and architecture-driven FEM (tongue, heart). Included also is the paper defining the spiralis muscle that we hope to submit very soon to Science.
- 3) Several papers of ours looking at ventricular fiber architectural organization and disorganization in pathological settings.
- 4) Several papers from others depicting atrial myoarchitecture in excised tissue.

My recommendation for you is to get familiar with the method of QSI tractography (tongue, heart) via my presentations/papers and the tutorial on the DSI Studio website (program can be downloaded for free), and to review for the big picture the papers done by others regarding atrial myoarchitecture.

The first goal will be a short paragraph for the course providing the rationale for the work (clinical, scientific), the core methodology, and the projected results and significance. My hope is to get the mouse and rat diffusion data from Dana any day. Let's plan to meet next Saturday at SS, same time.

Welcome to the world of diffusion. Best regards

Richard

# Interactive Cancer Regimen Visualization and Analysis

Yang Xiang

PI

yang\_xiang@brown.edu

Brendan Leahey

Co-PI

brendan\_leahey@brown.edu

Warner Jeremy

Collaborator

jeremy\_warner@brown.edu

October 14, 2024

## Abstract

Cancer treatment regimens are highly complex, involving a combination of anti-cancer drugs given in numerous routes (eg, oral, IV, intramuscularly) and supportive therapies administered over specific periods. An interactive regimen visualization tool would give healthcare providers, patients, and researchers a way to visually explore the details of specific regimens, making complex treatment plans more accessible and easier to communicate. Thus, we present a novel tool to automatically visualize cancer regimen from databases. User only need to select an regimen or input some query text, the tool will deal with the input, find the related regimen and visualize it for the user. The user can then interact with the regimen, and see its details and ask questions about it. The system supports automatic visualization for the graph structure of the regimen and question-answering for regimen details and regimen selection, which is through analysis of the regimen data with finetuned LLM models.

# 1 Response to reviewers

## 1.1 Reviewer 1: Arjun Prakash

1. Overall: 4
2. Interdisciplinary: 1. Clearly combines medicine with NLP and visualization
3. Scientific: 5. Not clear what the scientific contribution is other than the use of the dataset.

**Response:** the original dataset are unorganized and do not have a uniform format; so using llm for uniform format processing helps visualize them uniformly

4. Visualization: 4. Seems ambitious to figure out the best visualizations during the first week
5. Significant: 4. Having an LLM being able to wrangle a complex dataset is interesting.
6. Novel: 4. Unclear what the novelty is if the authors are not planning on pretraining.

**Response:** Pretraining is hard and cost a lot, so currently the strategy is to use fine-tune

7. Goals clearly stated: Yes, but they might be ambitious
8. Likelihood of Success: 5/10

9. Strengths:

Good presentation. Well motivated.

Clear timeline.

Weaknesses:

Probably too ambitious.

Not clear what the evaluation will be, or how to prevent hallucinations.

**Response:** This is very correct, and I will add the detailed evaluation and hallucination-prevention methods in the final proposal.

## 1.2 Reviewer 2: Sathe, Aalok

1. Overall: 4
2. Interdisciplinary: 2  
Target domain is cancer therapeutics.
3. Scientific: 6

This is not intended to be a scientific contribution; though it is definitely aligned with the longer-term translational goal of science: treatment. Additionally, enabling better patient tools for treatment will someday enable better administration of clinical trials, etc.

**Response:** Yes, this is more like a tool rather than something that directly generates insight for a domain. Maybe insight generation feature will be added, but it is hard and currently not planned.

4. Visualization: 5  
The primary contribution of this project. The authors propose to use LLMs to parse textual data into structured formats and visualize these using intuitive graphs. The nature of these contributions is a bit vague.

**Response:** The preliminary proposal indeed is not clear about this point. I think the contribution will include generating uniform visualization for some messy data with several given inputs, so that it can be also used in other similar scenes.

5. Significant: 3

Seems to be very important work.

6. Novel: 3-4

Using LLMs to parse medical data would reduce burden from patients and avoid an overload of information. However, how this information will be visualized is not clear.

**Response:** I will add images in the final proposal I think

7. Goals clearly stated: 5

Some goals are clearly stated (target domain, the nature of the tool, its purpose, some components of it such as LLMs for parsing and graphs for viewing), but others are not (methods for visualization; what data will be visualized, etc.).

**Response:** The detail of methods will be added in final proposal.

8. Likelihood of Success: 4

Likely to succeed, but unclear in the absence of information about methods of visualizing, and what data will be visualized and why.

9. Strengths:

- Well motivated domain and purpose.
- Novel additional use of LLMs for an appropriate purpose.

Weaknesses:

- Could be more specific in its methods.

Other comments:

- LLMs might omit information from the data and cause fatal problems with patients' treatment.

**Response:** This is a very good review, pointing out the problems correctly: lack of detail and inaccuracy of LLMs. I will add them in final proposal, like the images of the visualization method, the ways of avoiding hallucination, etc.

### 1.3 Reviewer 3: Eric Xia

1. Overall: 3

2. Interdisciplinary: 2. This work is highly interdisciplinary, as it involves utilizing computer science and human-centered design techniques to visualize an accumulation of scientific knowledge relating to cancer treatment. I think this could be improved with more integrated biochemical visualizations, e.g. of the specific effects and interactions of treatments, but this may not be as practical for an end-user application.

**Response:** I think it will be combined with the patient's conditions to customize a calendar for them, so yes!

3. Scientific: 3. This approach would clearly benefit cancer patients in allowing them to compare potential treatment plans and make better decisions about their healthcare. There does not seem to be a substantial scientific component in the health sciences. As a design intervention, there is the possibility of performing studies comparing the visualization design or LLM to existing plans.

4. Visualization: 4. Using LLMs or retrieval techniques could lead to better parsing of cancer treatment regimens. I'm not convinced that it will allow for consistent visualization or more flexible accommodation. The goal of developing a universal approach for cancer regimen visualization should ideally specify more on the visual techniques being used. Are there novel ways of visualizing time intervals for medication, comparing different medications, appointments, checkups? The LLM contribution seems somewhat distinct from this: the approach appears to be using LLMs to generate visualizations e.g. svg flowcharts, while simultaneously manually implementing a framework. In the research plan, they outline the usage of visual techniques e.g. node-link diagrams, gantt charts, and process flow diagrams. Both approaches seem like they would require user studies early on to be effective.

**Response:** This reminds me of some vague ideas that I had when designing the proposal. I had wanted to do visualization for almost structures automatically, like most structures can be represented in some kind of graph, and most of them can be represented in tree without losing too much info. I think I will consider more about this and try to find something useful in a more general way.

5. Significant: 2. This would be a very significant contribution in improving wellbeing and making healthcare more accessible. Care should be taken to ensure that the visualization is actually providing a benefit, and not rehashing text information in an unclear or misleading manner. Visual interfaces lend themselves to comparisons and exploration in particular. I do not think this interface would be particularly generalizable.

**Response:** Comparison is important and also a great idea. It can be integrated by comparing the original paper infos of 2 different regimen. A great idea I did not come up with...

6. Novel: 4. If the visualization component succeeds in being usable, it would be a novel contribution and has a very clear benefit for patients. I have some concerns on the practicality of the visualization; some demos sketching out visualizations would be helpful.

**Response:** Sure, talk is cheap and demo is important.

7. Goals clearly stated: 1. The goals are very clearly stated, and the research plan is extremely comprehensive.

8. Likelihood of Success: 3. I think there is a reasonable chance of success. The two different directions outlined of using LLMs to automate regimen visualization and designing a visual language for modular construction may be somewhat contradictory.

9. Strengths:

- Very comprehensive research plan
- Solves a real and important problem
- Widely available base of data
- The problem seems like it would benefit from visualization

Weaknesses:

- Not clear what the novel visualization contribution is; dynamic visualization as opposed to static?
- The 0.5-1 week allocated to visualization design is frighteningly short

- Somewhat overlapping goals with LLM and manual visualization design.

**Response:** Thanks to the help of gpt, the framework can be built quickly.

#### 1.4 Reviewer 4: Thais Del Rosario Hernandez

1. Overall: 3
2. Interdisciplinary: 1. The proposal is highly interdisciplinary. It proposes to tackle complex (and often variable) oncology regimen data and provide the user with an accessible interface to visualize their treatment plan(s).
3. Scientific: 3. This visualization tool's user base is both patients and care providers (i.e. doctors and other medical staff). Reducing the complexity of the tasks required to browse through the regimen database might facilitate the discovery of patterns in patient response to certain treatments. The proposer draws parallels between other fields where the same time-dependent and divergent nature of the data (e.g. epidemiology) reduces interpretability, suggesting that this tool could also be applicable in that context.
4. Visualization: 3. The database that this proposal will use as a data source is entirely text-based and hard to navigate, especially in the case of oncology non-experts (i.e. recently diagnosed patients and their families). The score could be improved with a more detailed description of how the visualization will improve on current visual paradigms.
5. Significant: 2. A visualization tool for this data will increase accessibility and ease communication in patient care. It will also aid clinicians and researchers in discovering patterns in the data. Since the proposed database mostly contains clinical trial data, recognizing what works is important in order to establish more concrete treatment protocols.
6. Novel: 4. The proposal mentions previous visualizations of cancer regimen data, but does not go in depth about what will make the proposed visualization novel, nor how it will be evaluated against the previous work (i.e. baseline).

**Response:** There are not many works doing this, so finding a baseline is hard...

7. Goals clearly stated: 3. Goals are mostly clear, but are wide-reaching and evaluations are not included at each step.

**Response:** True. I will add them separately for each feature.

8. Likelihood of Success: 4. Research plan and number of proposed methods seems ambitious.
9. Strengths:
  - Proposal is thorough and aims are well connected to the overall problem.
  - Highly relevant to a wide user base.Weaknesses:
  - The proposal to use LLM as a way to filter data seems overly ambitious and prone to more problems than its worth. It could be replaced with a simpler but more stable filtering algorithm for the data

(perhaps using metadata from the existing regimen database, if available).

- Data collection involves reducing the amount of regimens, but then mentions "exploring additional databases". Given the timeline of the project, it would be more reasonable to focus on a small subset within one database, and optimizing the visualization for that type of data.
- The research plan timeline is very tight, and does not include time for preparing the final write-up/presentation for the course.

**Response:** The data are somewhat messy and do not have a uniform format, that's one of the reasons I use an LLM. It indeed is reasonable to just explore some certain database, and now I plan to use just hemonc.org. The timeline may be altered — I did not save time for presentation as I thought it is something to be done quickly after finishing the project, but saving some time for it surely would be better.

## 1.5 Reviewer 5: David Laidlaw

1. Overall: 7

2. Interdisciplinary: 5  
Good

3. Scientific: 7

The proposal mentions some positive properties of the proposed system and its use. Will your system be better in those ways? If so, how will you measure that and what will the system be better than?

**Response:** I will measure them with accuracy + user study, and for user study I plan to design a multi-choice survey form to collect data. Maybe this is a bit simple and I will add more user study methods...

4. Visualization: 7

How will you establish the value and contributions of the proposed system? You list a number of positive properties, like user-friendly, consistent, and accurate. As above, will your system be better in those ways? If so, how will you measure that and what will the system be better than?

**Response:** This can be organized by comparing the acc/time results between experts, non-experts and llms.

5. Significant: 7

See the previous two criteria

6. Novel: 7

Not clearly articulated. The proposal lists some other approaches, but it isn't clear how the proposed work will differ.

7. Goals clearly stated: 7

At a high level, the goals are impressive and valuable. At a more detailed level of methodology and evaluation, they are insufficient to evaluate.

8. Likelihood of Success: 8

The LLM element seems like it is unlikely to fit in the time we will have. It seems like it is very important to all of the aims, however, so having it optional seems like it is a problem. It also seems like it could take a very long time.

**Response:** I have some previous experience related, so I can save some time for it, and can at least give a version with not-bad performance.(compared to perfect)

A revision focusing on the evaluation and user study would score much better.

**Response:** Yes, this is the most important part.

9. Other:

- Don't know what RAG is in the abstract.
- The proposal needs a title.
- Collaborator(s) not listed.

**Response:** Will surely add them in final proposal! And delete RAG because I think it may not be used.

## 1.6 Reviewer 6: ChatGPT

1. Overall: 2

This proposal presents a novel approach to visualizing complex cancer treatment regimens using LLMs and data retrieval techniques. The project addresses a critical need in healthcare for better communication and understanding of treatment plans.

2. Interdisciplinary: 1

The project bridges healthcare, data science, and AI, making it highly interdisciplinary.

3. Scientific: 2

The scientific merit is strong, particularly in the application of AI to healthcare and the potential to improve patient outcomes.

4. Visualization: 1

Visualization is central to the project, and the proposed use of interactive graphs and diagrams is both novel and impactful.

5. Significant: 2

The potential impact on healthcare, particularly in improving communication between healthcare providers and patients, is significant.

6. Novel: 1

The integration of LLMs with data visualization for cancer treatment regimens is a novel and timely approach.

7. Goals clearly stated: 1

The goals are clearly articulated and aligned with the project's objectives.

8. Likelihood of Success: 2

The project is likely to succeed, though there may be challenges in ensuring the accuracy and real-time updating of cancer regimen data.

9. Strengths:

The proposal addresses a critical need in healthcare with a novel and impactful approach.

The research plan is well-structured and achievable within the given timeframe.

Weaknesses:

The project may face challenges in integrating real-time updates and ensuring the accuracy of the LLM-generated visualizations.

10. Other comments for discussion:

Collaboration with healthcare professionals and oncologists could further enhance the accuracy and relevance of the tool.

**Response:** The suggestion for collaboration indeed is great... For other parts I think gpt is just giving praise. Maybe the prompt should be changed to let gpt give more criticism.

## 2 Aims

1. **Develop a novel approach for visualizing cancer regimens:** The tool will establish a flexible framework that can accommodate the diverse structures of cancer treatment regimens, including various drug combinations, delivery methods, and time frames. By leveraging LLMs and retrieval methods, the system will dynamically analyze regimen data, allowing for consistent visualization of even the most complex regimens.
2. **Achieve accurate and visually compelling regimen representations:** The tool will utilize LLMs to parse text input related to cancer regimens and automatically convert this information into an intuitive visual format. The aim is to ensure that healthcare providers, researchers, and patients can easily comprehend complex treatment protocols through user-friendly, interactive graphs and diagrams, reducing the potential for misinterpretation.
3. **Integrate a dynamic question-answer system:** The tool will incorporate a QA system that allows users to ask specific questions about their disease or a selected node within the regimen. Leveraging LLMs, the system will provide contextually relevant treatment information based on the latest medical guidelines and research data. This feature aims to enhance user engagement by offering personalized, real-time answers, empowering patients and healthcare providers to make informed decisions while navigating complex treatment options interactively.
4. **Enable real-time updates for cancer regimen knowledge:** The system will support real-time integration of new and updated cancer regimen data. By fine-tuning the model with the latest developments in cancer treatments and utilizing retrieval-based augmentation, the tool will maintain accuracy and ensure that users always have access to the most current information.

## 3 Significance

Cancer treatment regimens are inherently complex, often involving intricate combinations of anti-cancer drugs, supportive therapies, and individualized treatment plans based on patient-specific factors. These

regimens are challenging to interpret due to the diversity in drug combinations, delivery methods (oral, intravenous, intramuscular), and timing schedules. As a result, healthcare providers, researchers, and patients face difficulties in understanding and communicating these plans effectively. Visualizing cancer regimens in a structured, clear, and interactive way will significantly improve the accessibility of this information.

The proposed tool is not only a valuable asset for healthcare professionals who need to quickly grasp and communicate treatment plans, but it is also essential for researchers looking to analyze treatment patterns and their outcomes. Patients, often overwhelmed by complex medical information, can benefit from a more intuitive understanding of their treatments, leading to better engagement and adherence to prescribed regimens. By simplifying the interpretation of these regimens, the tool can improve communication between doctors, patients, and caregivers, ultimately contributing to better patient outcomes.

Furthermore, the framework developed for this tool has the potential to be adapted beyond cancer regimen visualization. Many fields in biology, social sciences, and beyond encounter similarly complex datasets that are difficult to analyze and visualize in an accessible manner. For example, in genomics or epidemiology, researchers deal with large datasets that require detailed exploration and representation. The same visualization approach can be used to streamline the analysis of these datasets, saving researchers time in developing manual visualization codes and allowing them to focus on the scientific questions at hand. This framework can therefore serve as a generalizable solution for various disciplines, reducing the need for field-specific visualizations and promoting cross-disciplinary innovations.

## 4 Related Work

Some previous works have made attempts for cancer regimen visualization [3] [9]. Some database collect and validate existing cancer regimen, including a large amount of valuable data [10]. However, the database do not give a intuitive visualization of the regimen, leaving the complicated and unorganized long text to users. This is because cancer regimen data is complex and diverse, and merely organizing and validating the regimen in text form is already a huge amount of work, let alone analyzing and organizing them into a unified structure. Due to this, some current visualization methods for cancer regimen [3] [9] utilize of a processed dataset with fixed structure.

In a broader view, there exist systems that automatically generate visualization [4] which needs only a input dataset and will do automatic analysis and visualization generation. However, this tool is used for generating overall information of the whole dataset(e.g. histograms) instead of generating high-quality visualization for unprocessed sample points inside the dataset. It also requires a input dataset and does analysis for the whole dataset, which is of high cost and low efficiency when the dataset is updated frequently.

For feeding specific knowledge to a pretrained LLM, the most widely used methods include fine-tune(with models like lora [7] ) and rag[8, 6]. They are cost-effective ways to customize the LLM for specific domain usage. However, their outputs are sometimes unstable and faces problems of hallucination and low accuracy. The quality of input dataset matters when customizing LLMs.

## 5 Research Plan

1. **Data Collection:** The first step involves gathering relevant cancer regimen data from various sources, such as hemonc.org, which contains a wide range of chemotherapy regimens and clinical oncology data. However, since the dataset is too broad, we will focus on selecting regimen properties that are most commonly used or of high clinical significance. This phase will include selecting the properties

to remove noise, standardize the format, and ensure that the regimens are represented in a way that facilitates both LLM processing and visualization. (1 week)

2. **Visualization Design:** In this phase, we will determine the most effective way to visually represent cancer regimens. This involves working closely with oncologists and healthcare professionals to understand their requirements for regimen visualization. We will explore different visualization techniques, such as node-link diagrams, Gantt charts, or process flow diagrams, to identify the format that best captures the complexity of drug combinations, treatment schedules, and delivery methods. The goal is to design a visualization that is both clinically useful and easy to interpret for non-experts. (0.5-1 week)
3. **Web Framework Construction:** The visualization tool will be built using modern web technologies such as React for the front-end and D3.js [1], ECharts [5], or similar libraries for creating dynamic and interactive visualizations. There is a library, react-d3-graph, [2] which can be really useful. This stage will involve designing the front-end user interface (UI) and ensuring that the visualizations are rendered efficiently and responsively. We will prioritize a modular design to allow future extensions and integration with other medical data systems. This phase will also involve basic backend setup for data management and interaction with the LLM-based components. (0.5 week)
4. **Automatic Visualization with LLMs:** This step will involve integrating LLMs to automatically parse and analyze the regimen data. The LLMs will process input queries (e.g., regimen names, drug combinations) and provide some potential regimens, and user will select the one they want. Then the system will generate visualizations that accurately reflect the regimen's structure and timeline. Optimization methods such as cycle self-improving techniques will be employed to iteratively improve the accuracy of the LLM-generated visualizations. Additionally, special attention will be given to handle variations in regimen format and structure in a generalized way. (1 week)
5. **Fine-tuning LLMs for Advanced Features:** We will fine-tune the LLMs using the collected cancer regimen dataset to enhance domain-specific performance. Fine-tuning will allow the LLM to generate more accurate and relevant visualizations for specific regimens, capturing subtle details like drug interactions, dosage variations, and treatment schedules. In this step, we will experiment with different fine-tuning techniques including LoRA and retrieval-augmented generation (RAG), as well as regular evaluation to reduce issues like hallucination and increase accuracy. If time permits, this step will also include auto-update integration with the database. (1-2 weeks)
6. **Evaluation and Writing Report:** The evaluation will focus on two aspects: LLM output accuracy and visualization effectiveness. For LLM accuracy, we will compare outputs to ground-truth data from clinical oncology sources using precision, recall, and F1 score. For visualization effectiveness, a user study will involve healthcare practitioners (doctors, researchers, students) and patients, assessing usability, task completion time, error rates, and satisfaction. Upon completion, a detailed report will summarize the evaluation results, highlight system strengths and limitations, and suggest future improvements. Comprehensive documentation will also be provided in the report to ensure ease of future development and use. (2 weeks)

## References

- [1] M. Bostock. D3.js - data-driven documents, 2024. <https://d3js.org/>.
- [2] D. Caldas. React d3 graph documentation, 2024. <https://danielcaldas.github.io/react-d3-graph/docs/>.
- [3] I. de Bruijn, R. Kundra, B. Mastrogioacomo, T. N. Tran, L. Sikina, T. Mazor, X. Li, A. Ochoa, G. Zhao, B. Lai, et al. Analysis and visualization of longitudinal genomic and clinical data from the aacr project genie biopharma collaborative in cbiportal. *Cancer research*, 83(23):3861–3867, 2023.
- [4] V. Dibia. Lida: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. *arXiv preprint arXiv:2303.02927*, 2023.
- [5] A. S. Foundation. Echarts: A powerful, interactive charting and visualization library for browser, 2024. <https://echarts.apache.org/en/index.html>.
- [6] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.
- [7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- [9] J. Warner, P. Yang, and G. Alterovitz. Automated synthesis and visualization of a chemotherapy treatment regimen network. *MEDINFO 2013*, pages 62–66, 2013.
- [10] J. L. Warner, A. J. Cowan, A. C. Hall, and P. C. Yang. Hemonc. org: A collaborative online knowledge platform for oncology professionals. *Journal of Oncology Practice*, 11(3):e336–e350, 2015.

EDUCATION

---

- **Brown University** Providence, RI  
*Master of Science in Computer Science* Aug. 2024 – May. 2026
- **Tsinghua University** Beijing, China  
*Bachelor of Science in Applied Mathematics* Aug. 2020 – Jun. 2024

EXPERIENCE

---

- **Microsoft** Suzhou, China  
*SDE Intern* Apr 2024 - Jul 2024
  - **Metaversity**  
Developed a multiple-player online virtual school **Metaverse** from scratch. Implemented various features like object manipulation, user movement with **C#**. Designed customized ethical **AI chatbots**, utilized **C++** for mesh subdivision to refine models, wrote **HLSL** shaders for rendering and implemented server requests with **PUN**.
  - **Copilot LLM Router**  
Built a **Copilot** multi-LLM router which reduces GPU costs by **47%** while preserving accuracy and performance. Implemented **C#** functions to retrieve and process data from **CosmosDB**, trained a model with customized **MLP**, **embedding** and **Attention** layers and used **Azure Kubernetes Service** for deployment.
  - **Multimodal AI Psychotherapist**  
Implemented **TTS** and painting generation with customized multimodal input in an AI Psychotherapist Miniapp that provides users with mental help. Utilized **Prompt Engineering** to better simulate personalized AI character.
- **Kuaishou** Beijing, China  
*SDE Intern* Jul 2023 - Dec 2023
  - **Internal Websites Construction**  
Built internal dashboard for efficiency improvement, serving over **500** company staff. Wrote **SQL** commands to retrieve data, used **React** to build websites and utilized **Antd** and **Echarts** for UI and data visualization.
  - **Predictive Analysis Platform**  
Re-implemented predictive **data analysis** feature of Salesforce's Einstein Discovery platform. Used **Python Flask** for backend, utilized **Ridge regression**, encoding algorithm and served over **200** businesses and vloggers.
  - **Live Scripts Automatic Generation**  
Automatically generated live scripts for influencer marketing. Trained a **Transformer** model for image caption and wrote **KwaiYii LLM API** in **async Javascript** to generate expressive live scripts that serves over **800** businesses.
- **Tsinghua University** Beijing, China  
*Lab Member* Mar 2023 - Jun 2023
  - **Furniture Layout Showcase**  
Designed data strcture and recursive algorithm to provide flexible furniture layouts. Implemented the algorithm and processed data in **Python**, utilized **D3.js** and **Javascript** for website visualization.

PROJECTS

---

- **PitchPerfecter** <https://singleview11.github.io/>  
*Leader, Developer* Feb 2024 - Present
  - **Music Training Website**  
Designed and developed PitchPerfecter, a website for pitch perfect training which has gained over **20k** PV.
  - **Frontend and backend Development**  
Used **React** framework for frontend development and utilized **Tone.js** for music playing. Used **SpringBoot** framework for backend development, built an MVC structure with **Mysql** for Database and **Mybatis** for DAO. Designed **RESTful** APIs and customized **Java** classes. Utilized **OAuth 2** for convenient user authentication.

PROGRAMMING SKILLS & INTERESTS

---

- **Languages:** Python, C++, Java, C#, Javascript      **Technologies:** React, SpringBoot, Git, Linux, SQL
- **Interests:** Electronic music production / 3D Animation production, with over 1000 fans on Bilibili