Intro to HPC Bootcamp Final Project Presentations

Session 1

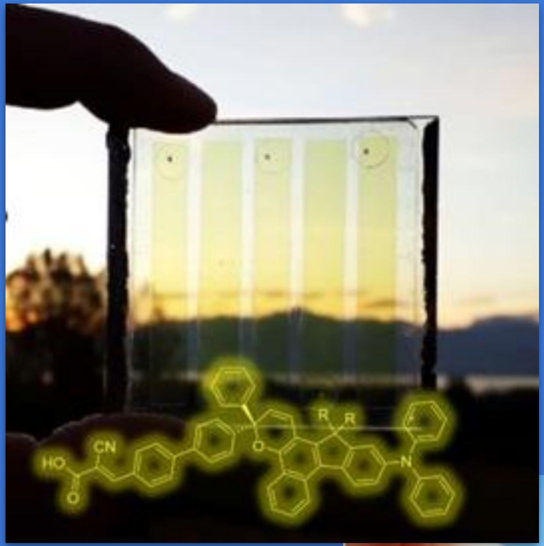
August 15, 2025

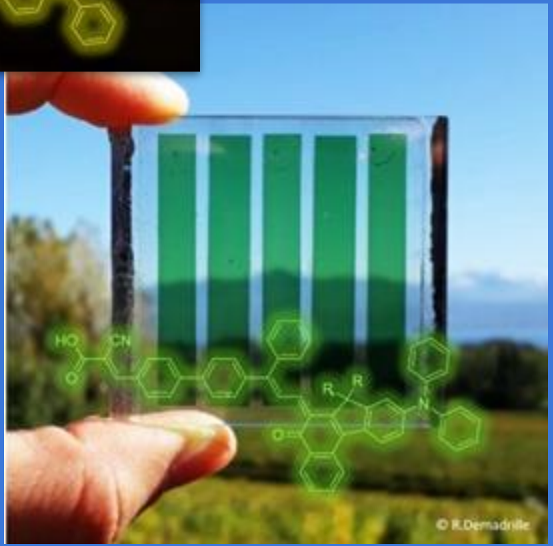
Lightning Talk Session 1	Group #
9:00 - 9:10	Group 7
9:10 - 9:20	Group 2b
9:20 - 9:30	Group 1a
9:30 - 9:40	Group 4b
9:40 - 9:50	Group 6a
9:50 - 10:00	Group 5

Group 7

Project 7: Computational Design of Low-Cost/High-Efficiency Solar Cells

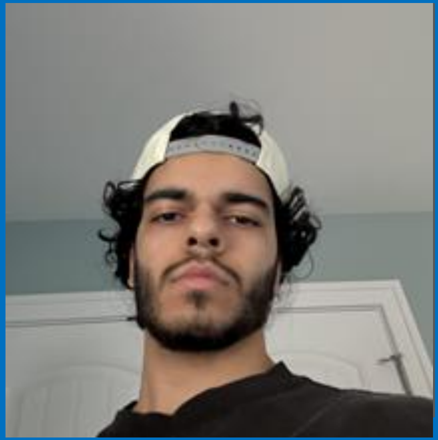
This project explores organic dye-sensitized solar cells (ODSSCs) as a low-cost and flexible clean energy technology.





August 15, 2025

Our Team!



Mohammad Altiwainy

Terry Nguyen

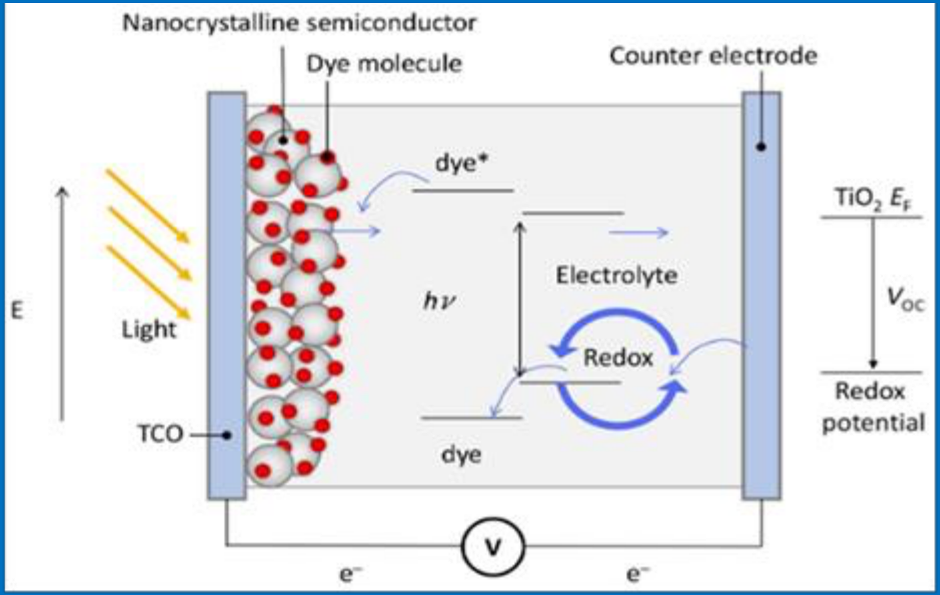
Sophia Johnson

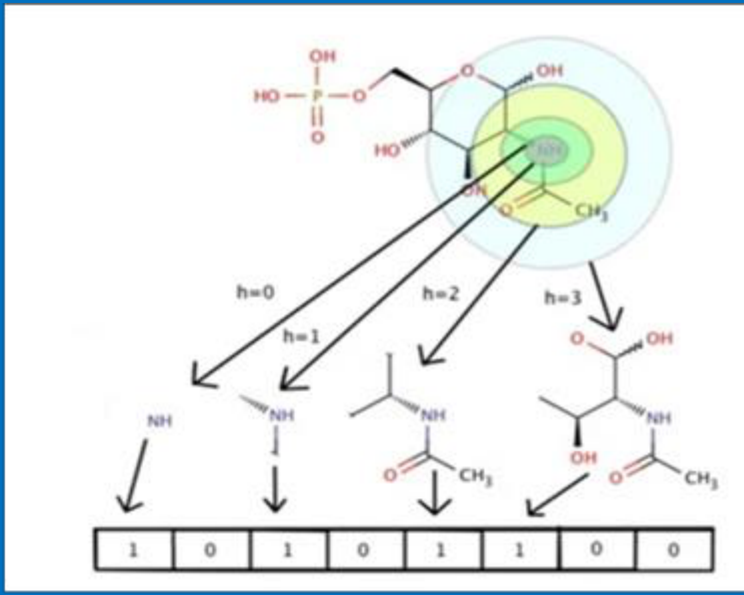
Amelia Mikos

Patricia Ladipo

Project 7

Brief Background / Goal





- •Dye-Sensitized Solar Cells (DSSC)
- •Emerging Technology
- •Cost-Effective

- •DSSC Mechanism
- •Solar cell use organic dye molecules to capture sunlight and convert it into electricity

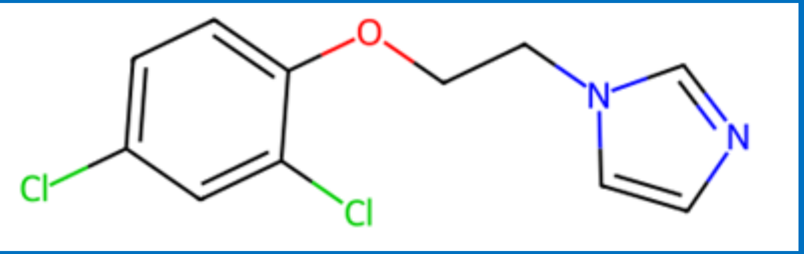
- •DyeDB dataset
- •Data Represented by:
  - •Simplified Molecular Input Line Entry System (SMILES)
  - • Vector fingerprint of a molecule

Our Tools and Methods:

- **Unsupervised learning** - Principal Component Analysis (Scikit library)
- **Supervised Learning:** Gaussian Process Regression and Random Forest Regression (Scikit library)
- **Python libraries** Numpy, Pandas, MatPlotLib to create/process data frames and visuals

Data/Figures

The Molecule

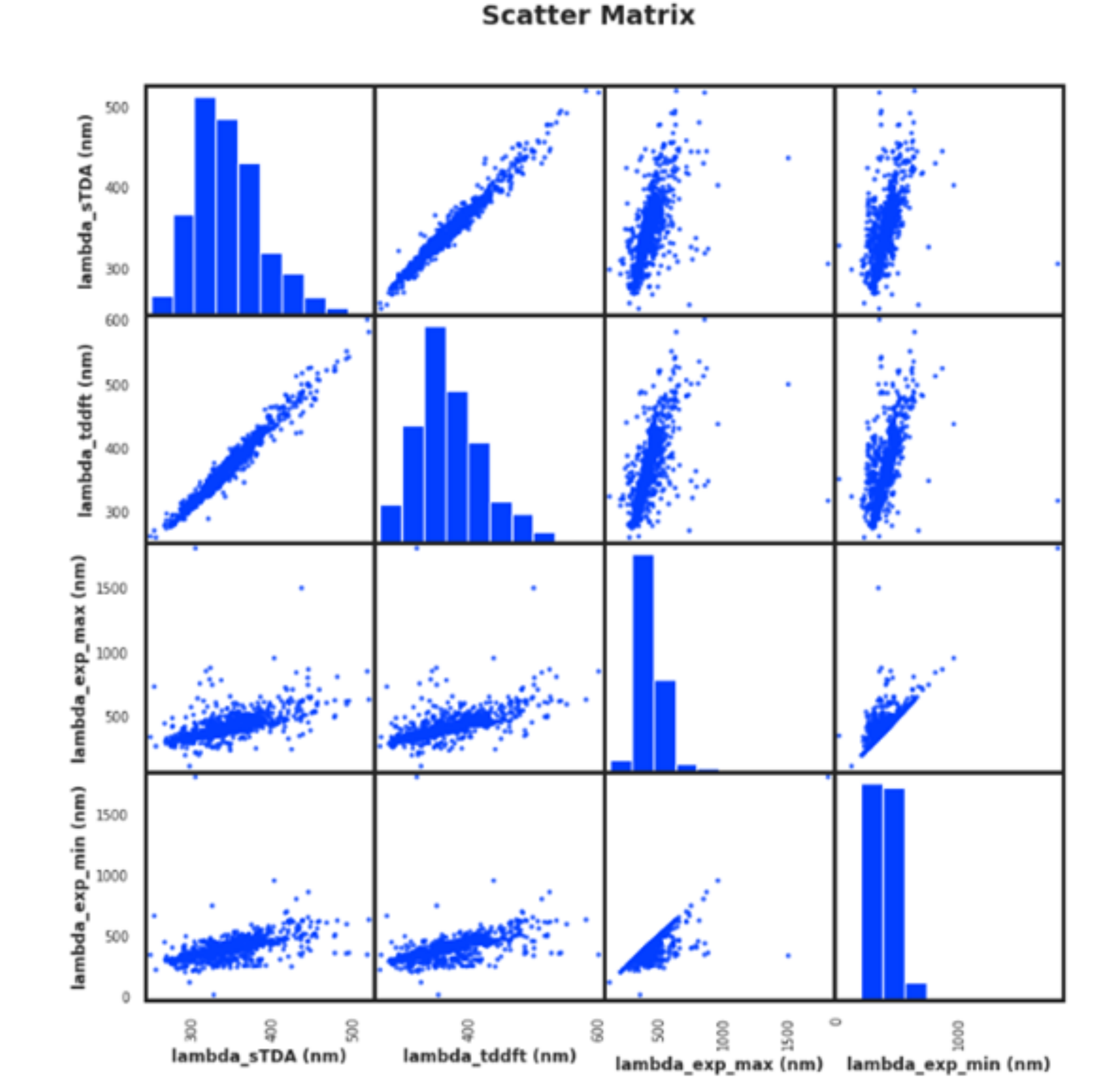


SMILES Generation

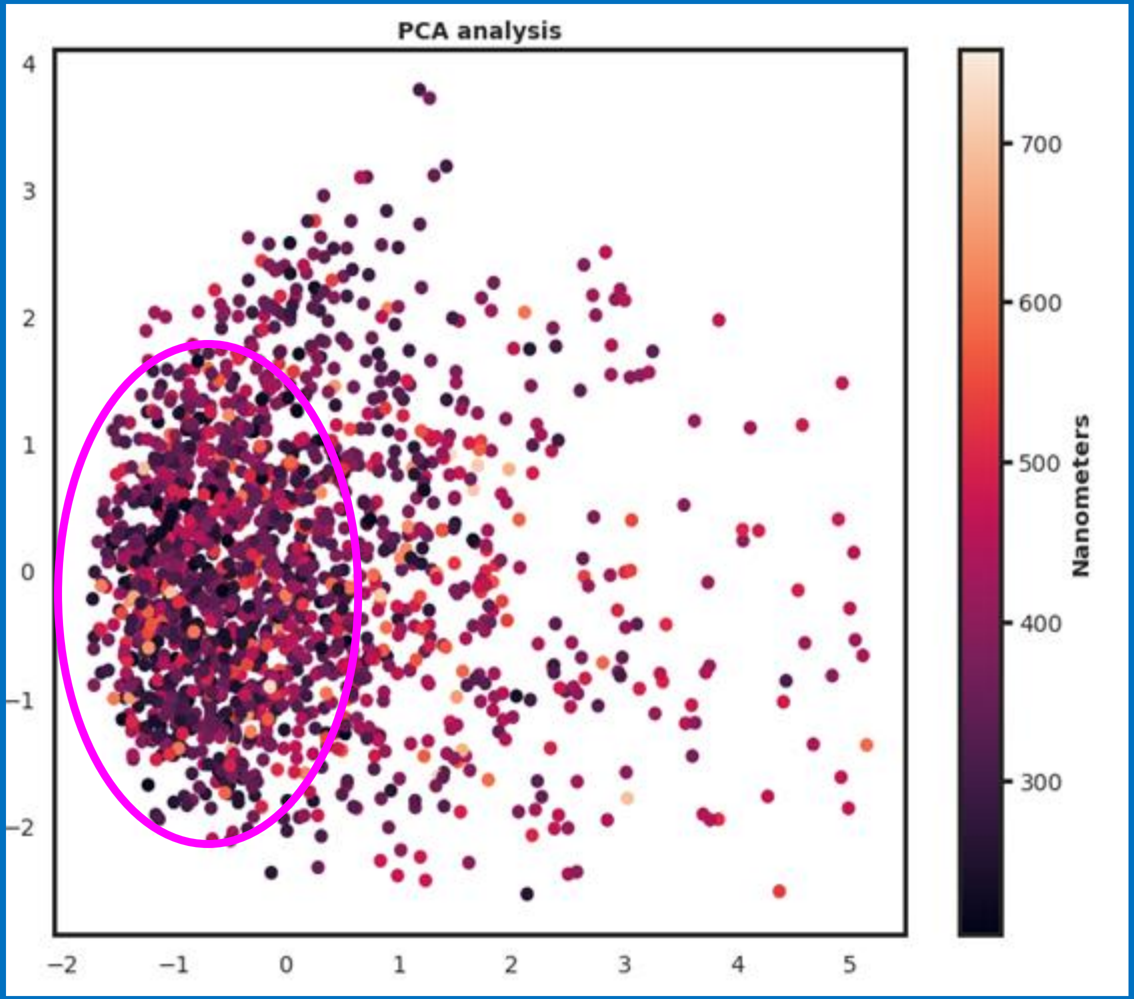
```
The canonical form of Clc1ccc(c(c1)Cl)OCCn1cncc1 -> Clc1ccc(c(c1)Cl)OCCn1cncc1
```

Converting to Morgan Fingerprint

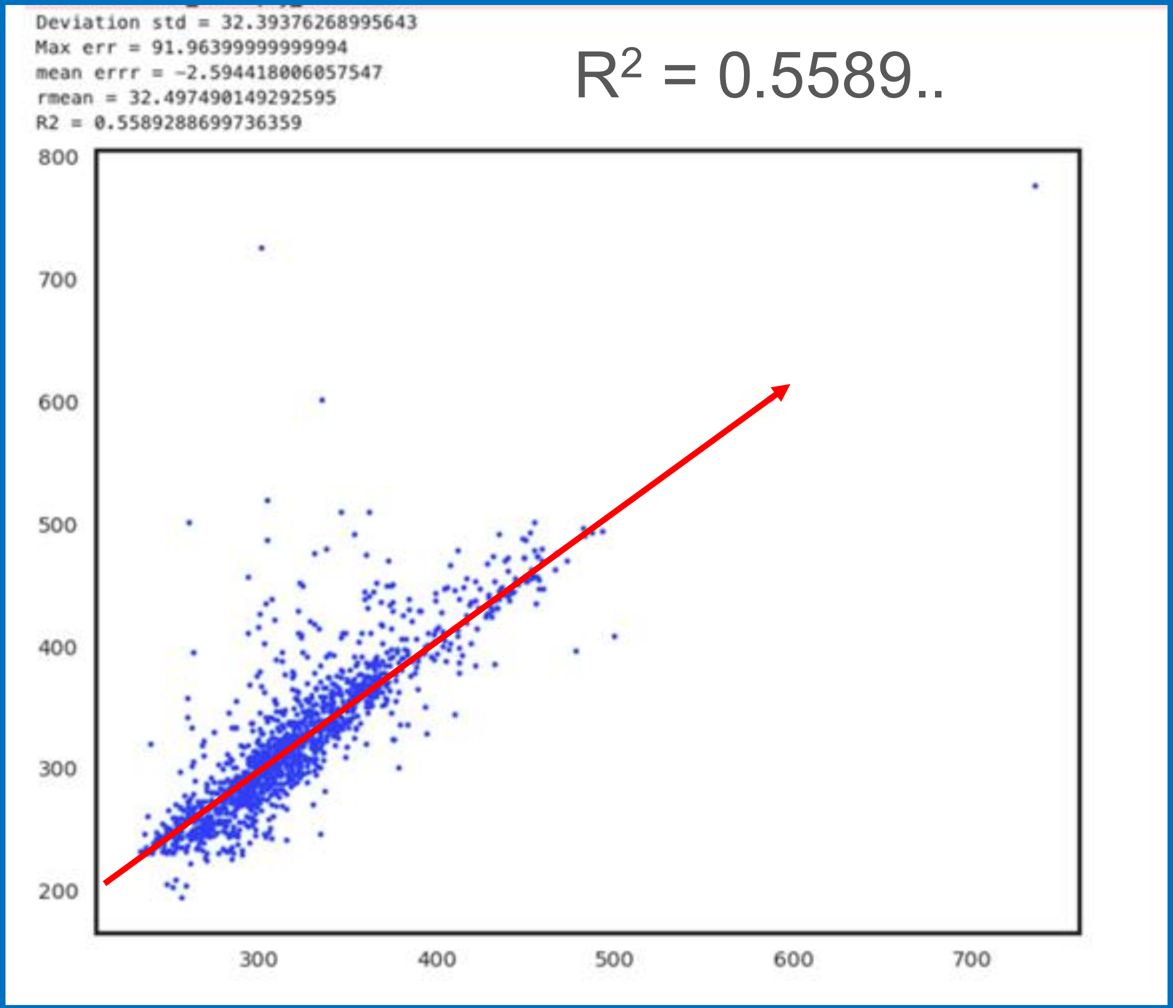
```
Morgan fingerprint of Clc1ccc(c(c1)Cl)OCCn1cncc1 is
[1 0 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1]
```



More Figures!



Random Forest Regression



Challenges

Argonne NATIONAL LABORATORY

What we encountered

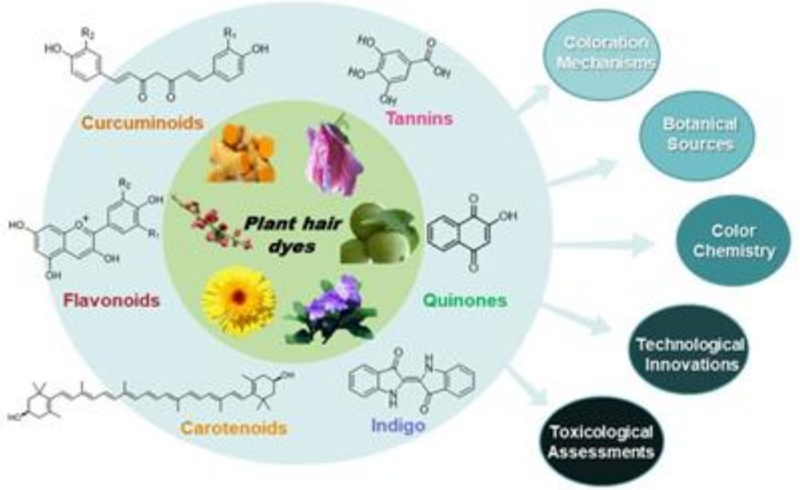
- Mixed experiences with Python and libraries within group
- Small bugs in the base code
- Chemistry heavy content in our project

How we resolved

- Utilized our different STEM backgrounds and skills
- Targeted each problem step by step
- Collaborated with mentors to resolved issues with the code

Conclusions and Next Steps

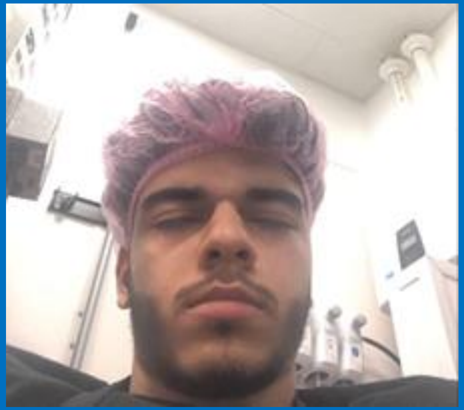
- **Applied** machine learning concepts to utilize dye molecules to increase efficiency for solar cells.
- **Tested** functionality, further research required for implementation(cost, efficiency)
- With more time, we'd learn more **context** of our chemistry-heavy project before diving into our data analysis.



Acknowledgments:

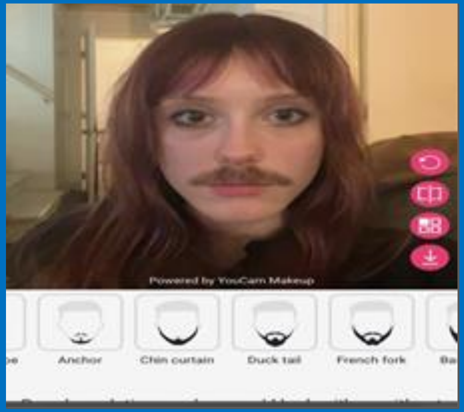
Thank you DOE and ALCF for this opportunity! Thank you Alvaro and Ashton for your help and guidance!

Sincerely:



Mohammad Altiwainy

Terry Nguyen



Sophia Johnson

Amelia Mikos

Patricia Ladipo

Questions?

Group 2b

Argonne Intro to HPC Bootcamp

Understanding and Optimizing Energy Usage from HPC Centers

Isis Reverón
Natalia Mena-Santiago
Gabriela Rodríguez
Jorge López
Daniyal Khokhar

Presenting group 2B

Daniyal Khokhar
University of Illinois
Chicago
CS Undergrad

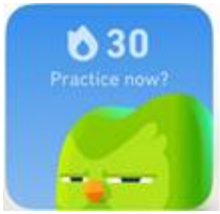
Jorge Jamal López León
University of Puerto Rico
Mayagüez
Physics Undergrad

Gabriela Rodríguez
Harold Washington
College
CS Grad

Natalia C. Mena Santiago
University of Puerto Rico
Cayey
Chem Undergrad

Isis A. Reveron Ortiz
University of Puerto Rico
Mayagüez
ChemE Undergrad

fun fact: 80% of this group speaks/understands Spanish, 20% tried to (using Duolingo)



Project Background/Goal

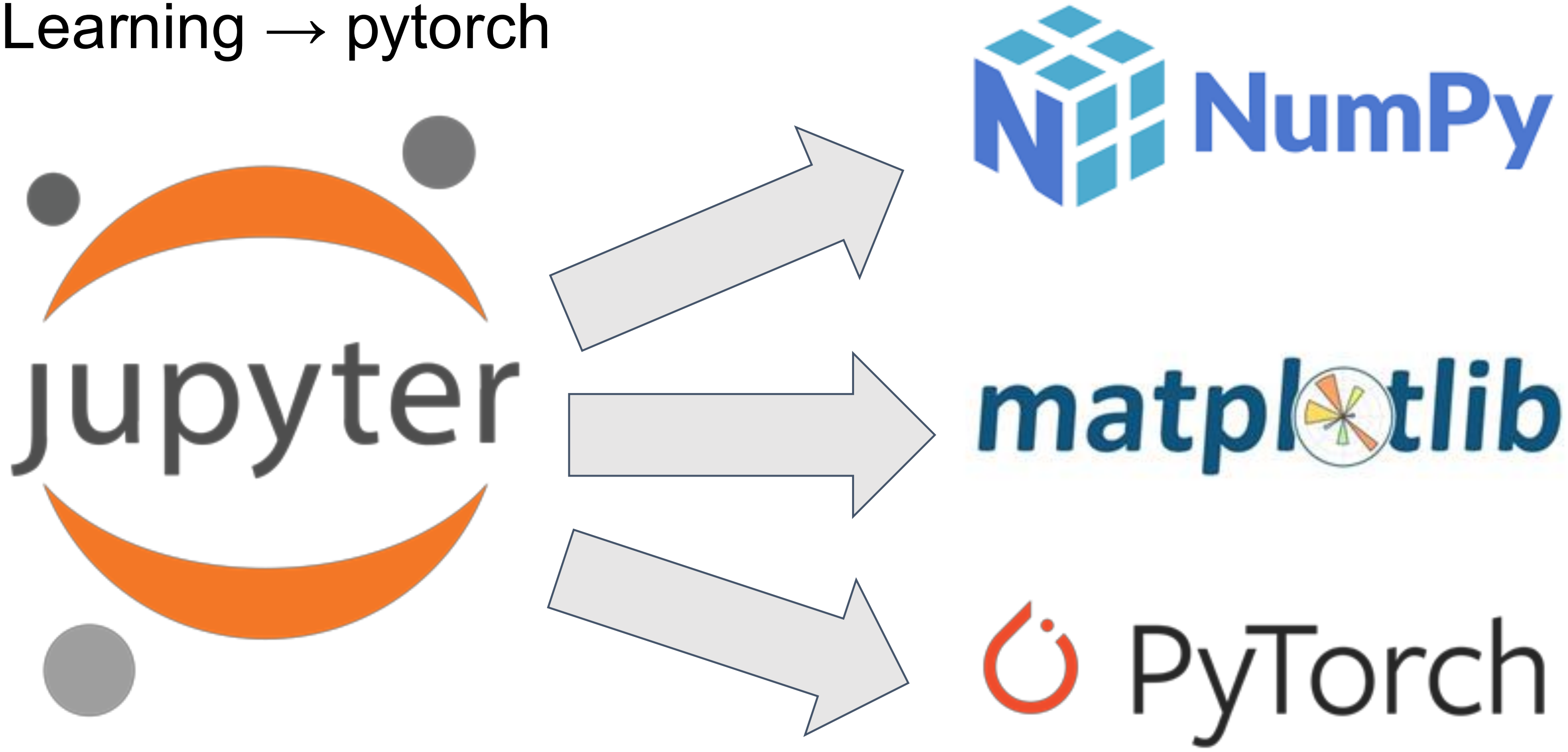
"As there is growing demand for computing resources, HPC systems draw enormous amounts of energy and need to be monitored and optimized"
-HPC Bootcamp organizers

The question we want to answer is: How can we optimize the energy use to balance scientific advancement with global sustainability and social equity?

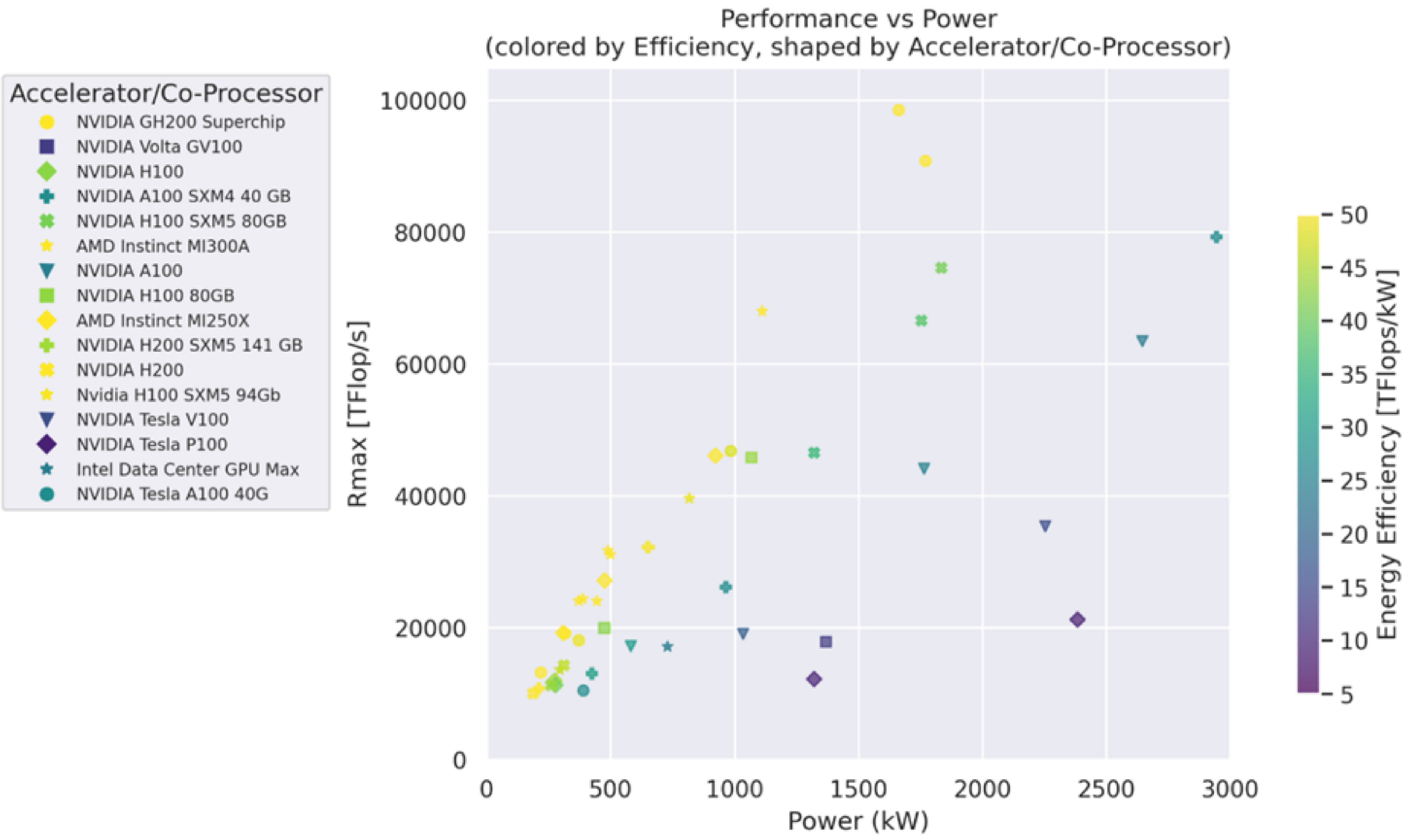
Methods, tools and more!

- Jupyter Notebook → open source interactive computing environment

  In this environment we had:
- Computation tools → numpy/pandas
- Visualization tools → matplotlib
- Machine Learning → pytorch

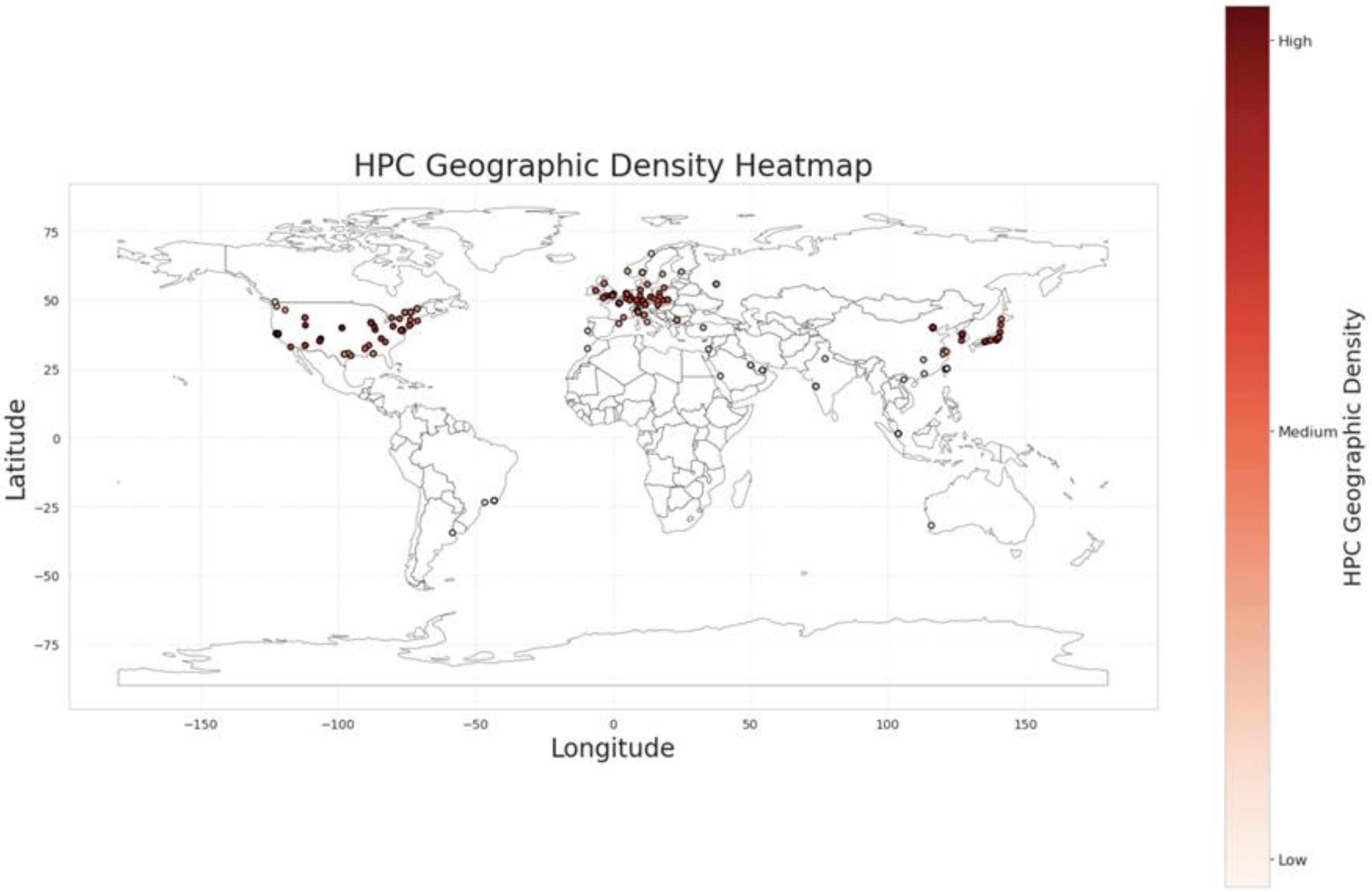


GPU Efficiency: Which has the Best Architecture?



- **Rmax [TFlop/s]** – M*ax measured performance* of the system in TFlops at max power.
- **Energy Efficiency** – How many operations per second for each kiloWatt of power consumed.
- **Power (kW)** – Total power draw of the system during **peak** performance.

Geographic Density of Computing Systems



Challenges… and more challenges

Our favorite ctrl+c, ctrl+v moment:

```
git pull
https://github.com/NERSC/DOE_HPC_Bootcamp_2025
```

Issues with installing packages, libraries etc.

Errors while running

Biggest Takeaways

physical components

location & climate

cooling system

voltage & consumption

This wouldn't be possible without…

Argonne NATIONAL LABORATORY

Project Organizers
Paige Kinsley
Ayesha Shafiuddin

Project Leads
Charles Lively
Helen He
Rebecca Hartman-Baker
Kelly L. Rowland
Lipi Gupta

Peer Mentor
Maricarmen Gonzalez Torres

Funding and Organizing Partners
Funded by ALCF/ANL | Organized by ALCF, NERSC, and OLCF

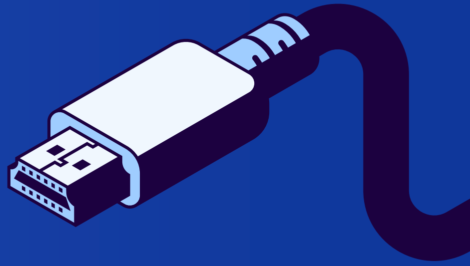
Argonne Intro to HPC Bootcamp

ANY QUESTIONS?

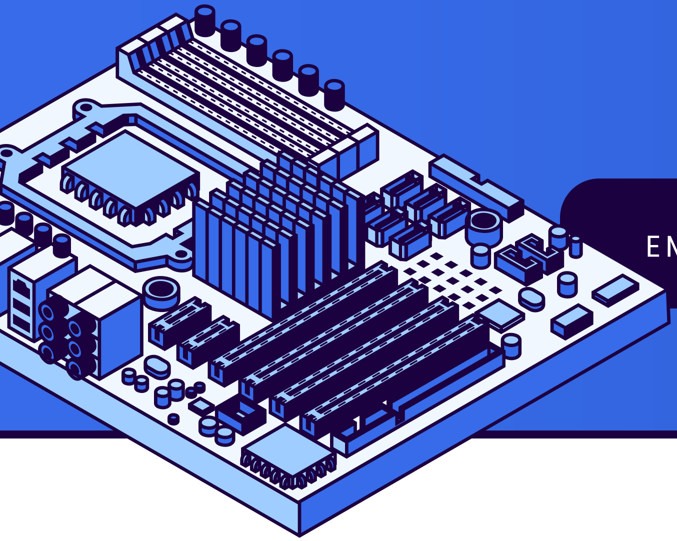
Group 1a

2025

DESIGNING A SUSTAINABLE & EQUITABLE HPC FUTURE



SHAPING THE PRESENT AND THE FUTURE OF SUPERCOMPUTERS

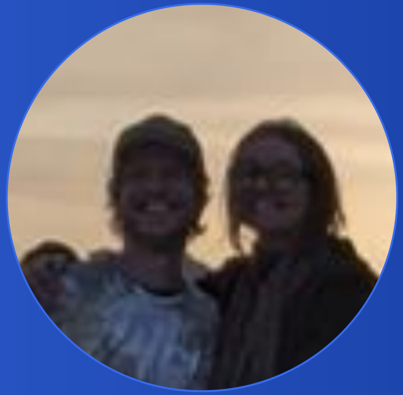


EMMA RONAI, CALEB MORRIS, TRISTIAN DIAZ-GARCIA,JOY WANG, Aly Ayyo

2025

Group Members

ALY AYYOB



CALEB MORRIS

EMMA RONAI

PEER MENTOR JOSE RODRIGUEZ RIOS

JOY WANG

TRISTIAN DIAZ-GARCIA

Project Purpose

AIM

1. DESIGN A NEW SUPERCOMPUTER THAT CAN:
   a. Maximize (fusion) science output
   b. Minimize overall power consumption
2. Create a schedule mimicking the day-to-day usage of *UNION*, our new HPC facility

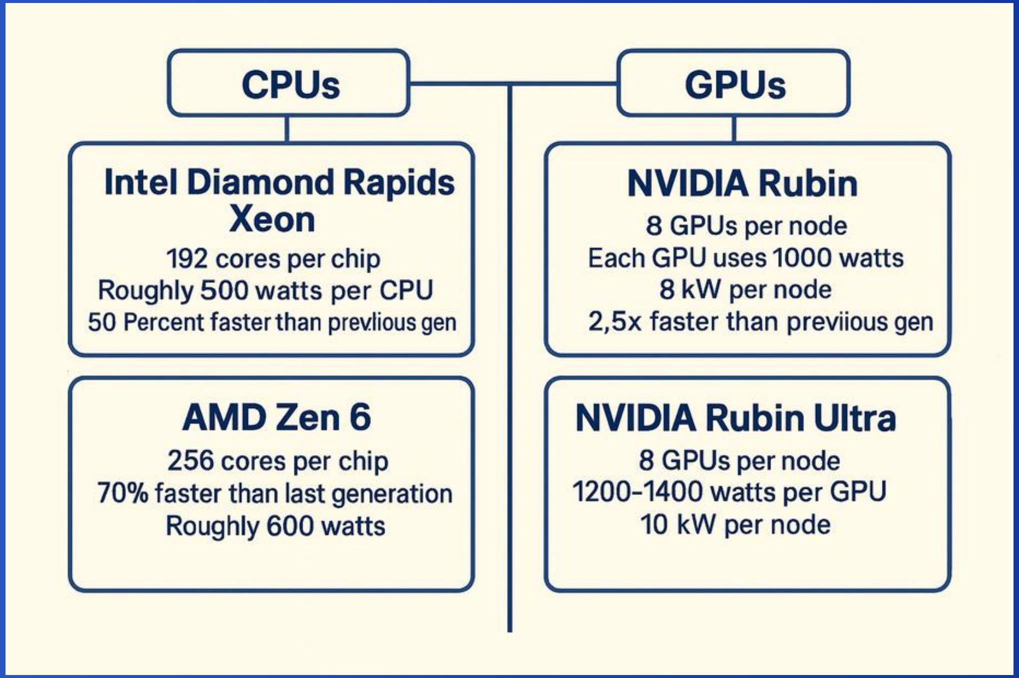
RATIONALE

- HPC powers discovery but consumes huge energy
- Smarter policies = more science per kWh + greater equity.

Hardware

Goal: Identify what hardware should meet our fusion tasks

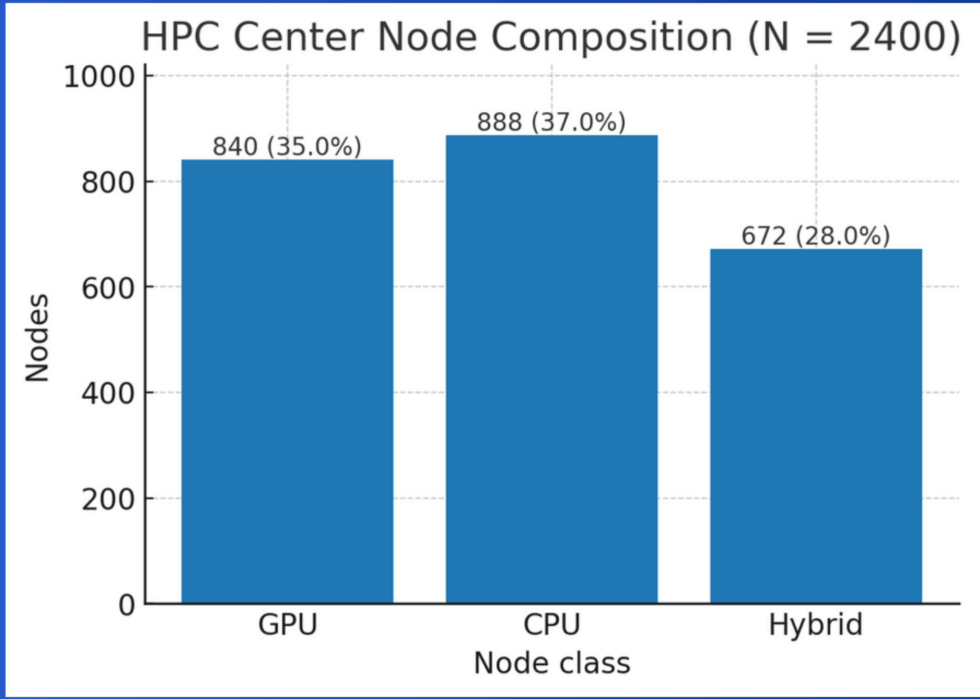
- Cost/benefit analysis of CPU/GPU nodes
- Utilized future Nvidia and AMD chip



IDENTIFY FUSION TASK REQUIREMENTS→ CONSULT EXPERTS → RESEARCH HARDWARE→ SOLUTIONS

Facility Architecture Development

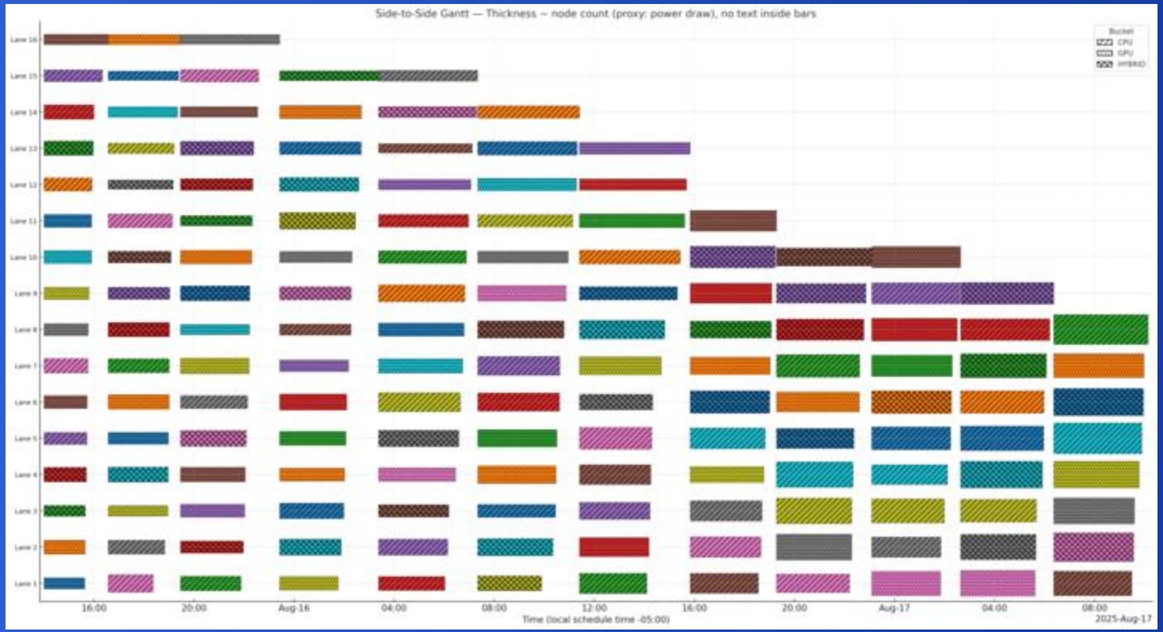
- Average power per node informs how many nodes we can have
- We use as many nodes as we can fit under 12MW
- Then we split the nodes between the 3 node types - GPU, CPU, and HYBRID nodes



Scheduling Tasks

Goal: To use all of the fusion capacity all of the time

- Two “lanes”, fusion and non-fusion
- Fusion balanced to run at 40% back-to-back
- Scheduler fills bulky jobs first, then small jobs fit opertunistically



MAD WHITEBOARDING → NAPKIN MATH → TASK MANAGER SCRIPT → VISUALIZATION SCRIPT

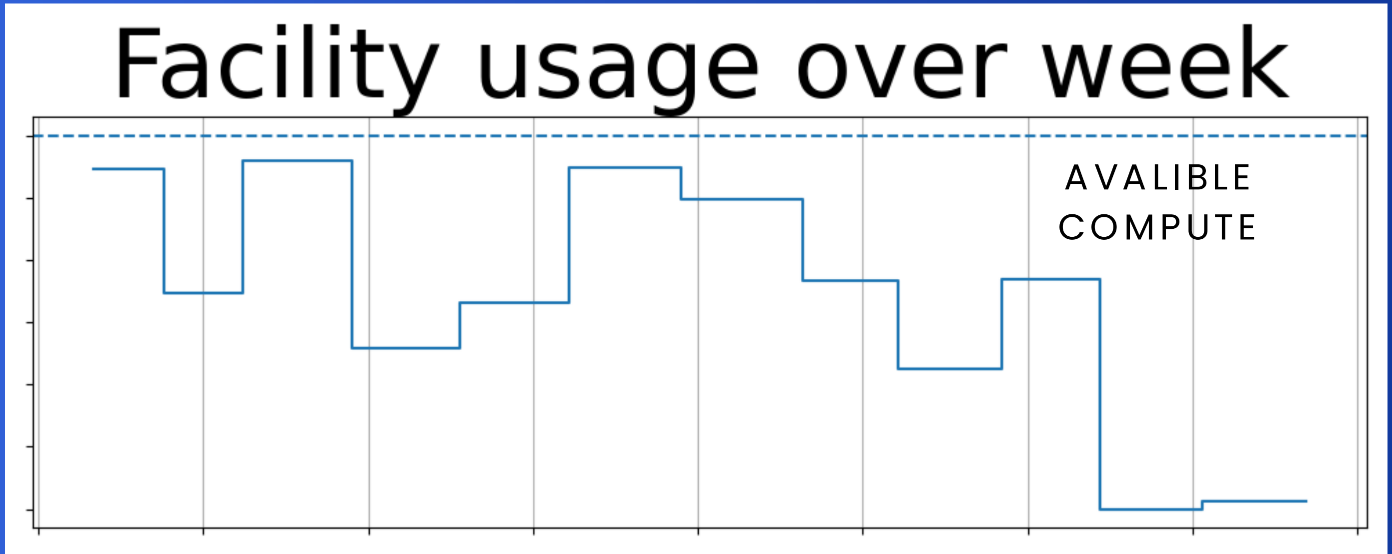
Conclusions and Next Steps

A FIRST STEP:

- HPC ARCHITECURE INFORMED BY FUSION RESEARCH NEEDS
- ESTIMATE FOR NODES REQUIRED BY 12 MW SUPERCOMPUTER
- MOCK 3-DAY SCHEDULE

NEXT STEPS AND IMPROVEMENTS

- HIGHER RESOLUTION APPROXIMATION
- OPTIMIZE SCHEDULER



Acknowledgments

- Peer mentor Jose Rodriguez Rios
- Charles Lively
- The entire HPC bootcamp staff!

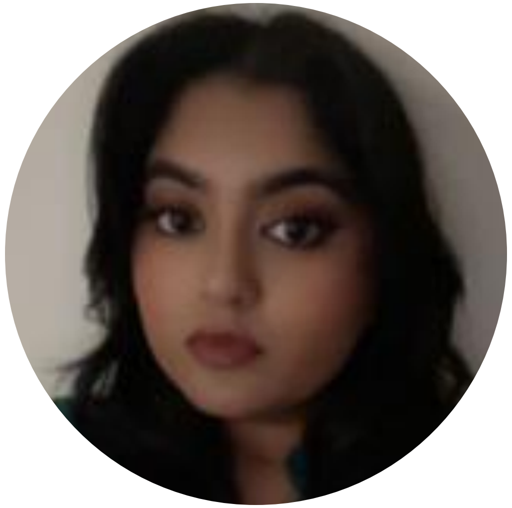
Group 4b

Short-Term Load Forecasting Using Machine Learning (ML)

Group: 4B

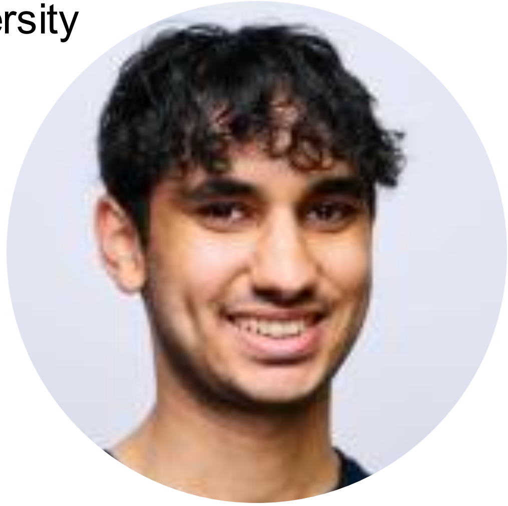
Yasmin, Marques, Kareem, Monica, Pragna

Marques Lewis
Mechanical Engineering (Minor: CS)
Tennessee State University



Pragna Amilineni
Data Science + Business Analytics
University of Illinois Chicago

Yasmin Sawaf
Computer Science
University of Illinois Chicago



Kareem Amin
San Francisco State University

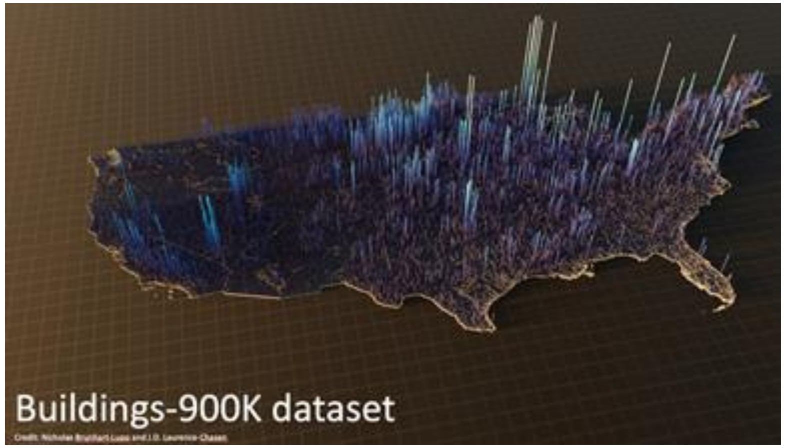
Monica Zapata Villegas
Oregon Sea Grant Fellow at
Tillamook Estuaries Partnership

Project Overview

Short Term Load Forecasting (STLF)
- Uses machine learning to predict the load or amount of energy a building or a unit will use in a certain period of time

Data Source
- BuildingsBench platform

Why do we need STLF?
- Grid reliability
- Cost savings
- Efficient operation
- User insights



NREL

National Renewable Energy Laboratory

Methods Used

Hyperparameter Combinations

- Models
- Activation Functions
- Optimizers
- Epochs

Model Training Metrics

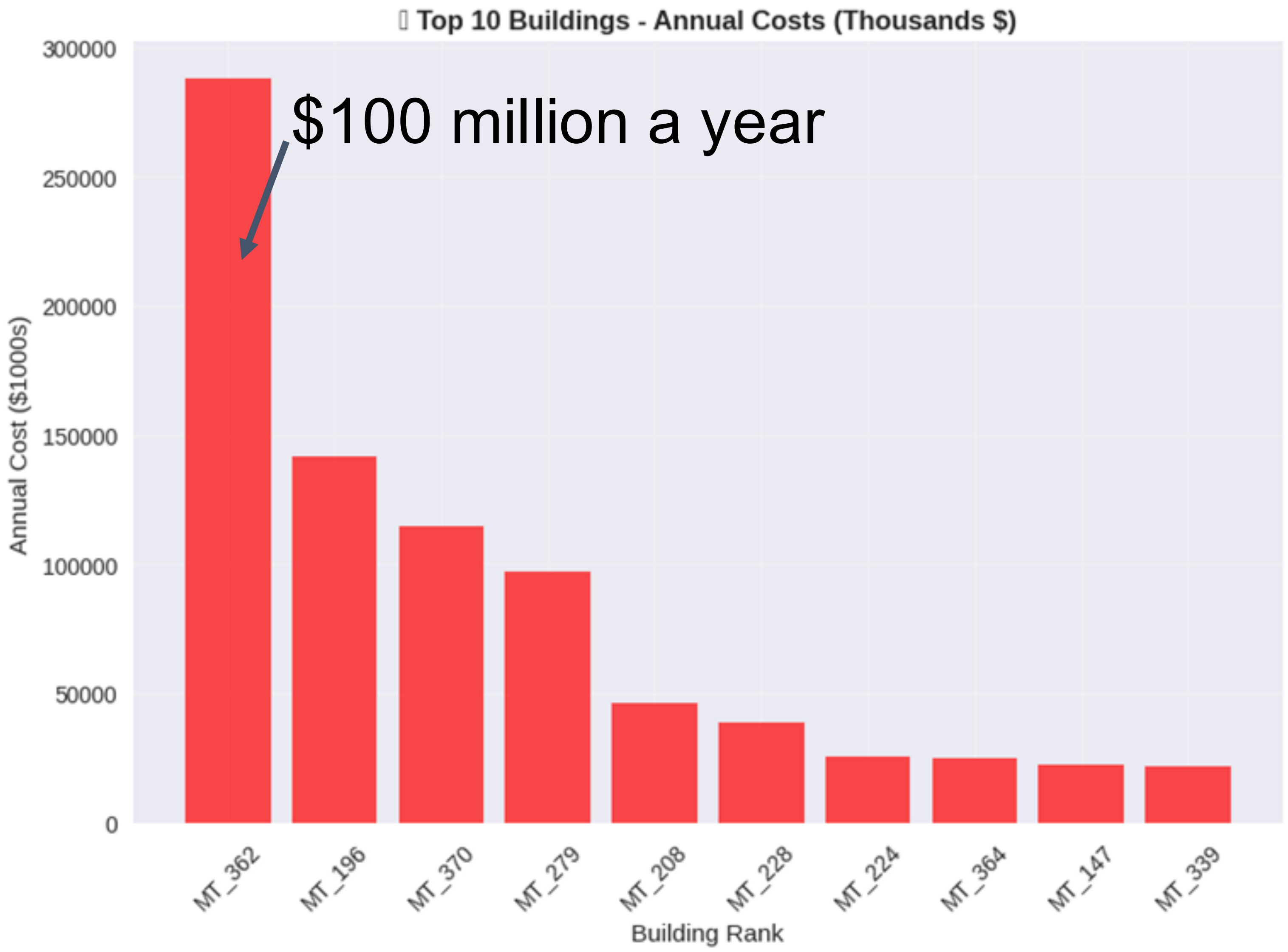
- MAE
- RMSE
- R^2

NERSC Perlmutter

- parallel processing for faster experiments
- handled large datasets & multiple runs efficiently

Electricity Consumption in Commercial Buildings in Portugal (2011–2014)

- Data: **359 buildings**, 3M+ electricity readings (2011–2014)
- Total annual spend: **$1.6 billion** on electricity
- That's **$4.43 million every single day**



All data measured at $0.10/kWh

If we know the spikes, we can cut the cost.

- Avoid peak-time surcharges
- Shift heavy use to cheaper hours
- Reduce grid strain

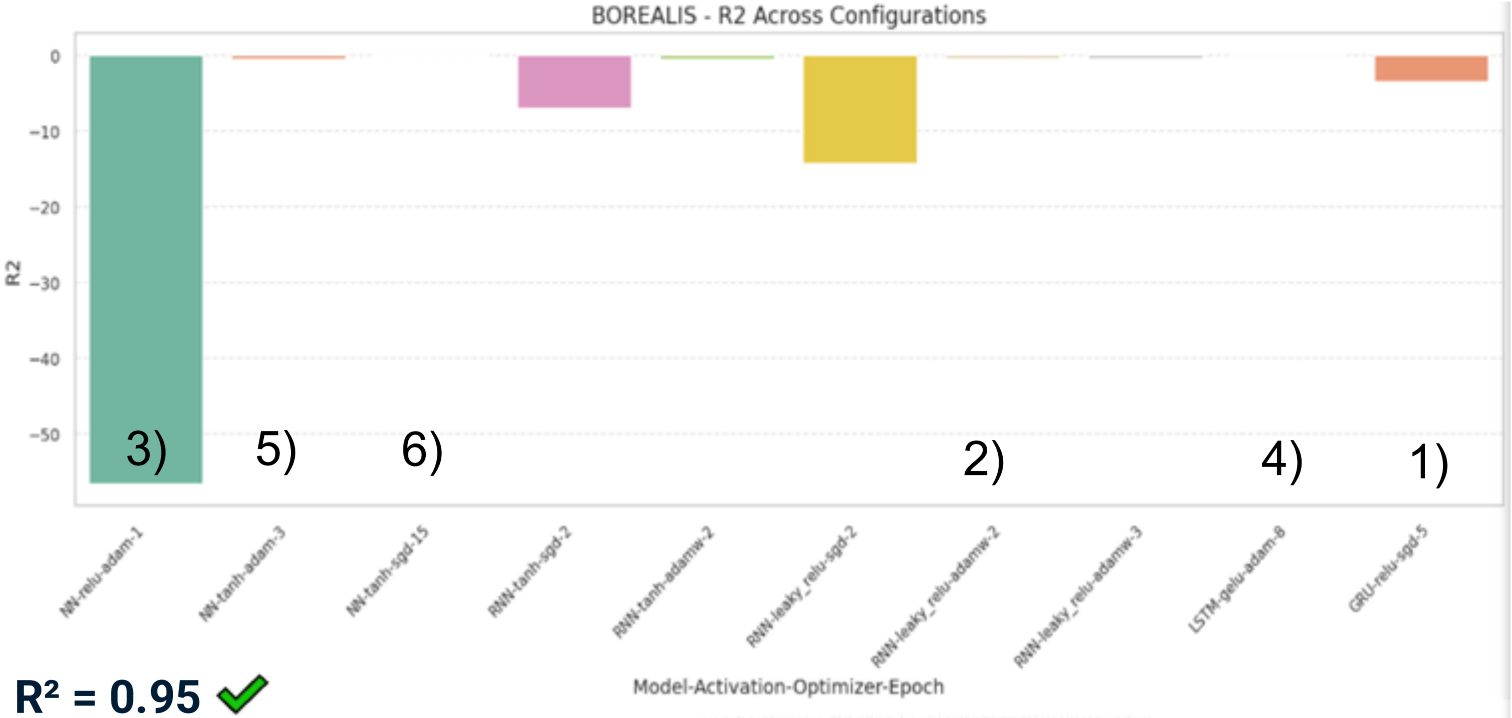
```
--- Training RNN | Activation: leaky_relu | Optimizer: sgd |
[RNN] Epoch 1: Loss = 466.6250
[RNN] Epoch 2: Loss = 303.0813
[RNN] Epoch 3: Loss = 303.5879
[RNN] Epoch 4: Loss = 264.1685
[RNN] Epoch 5: Loss = 241.9533
[RNN] Epoch 6: Loss = 222.1453
[RNN] Epoch 7: Loss = 205.0357
[RNN] Epoch 8: Loss = 189.7810
[RNN] Epoch 9: Loss = 176.9988
[RNN] Epoch 10: Loss = 173.4432
[RNN] MAE: 716.9688, RMSE: 955.4561, R²: 0.4512
Training Time: 4653.00 seconds
```

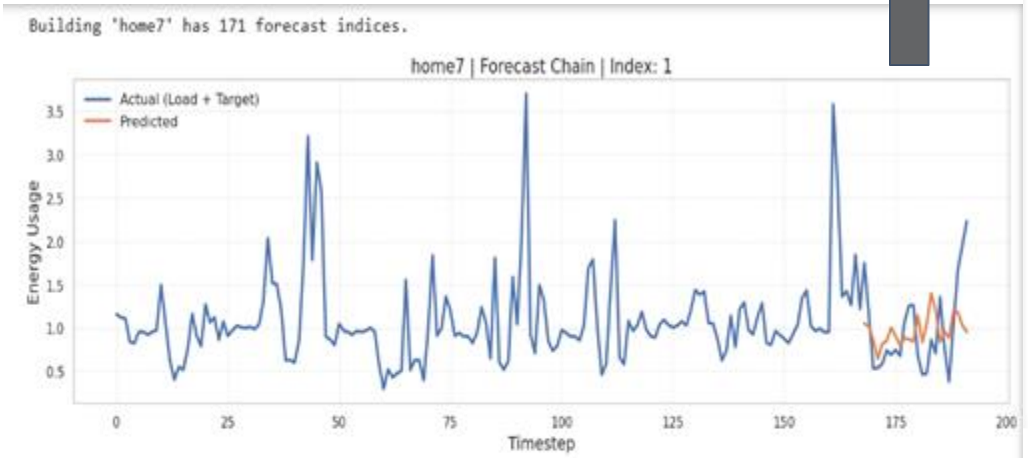
Model Settings	R^2	Train Time
leaky_relu, SGD, 10 epochs	**0.45**	**1h17m**
leaky_relu, SGD, 5 epochs	0.39	41m
tanh, SGD, 10 epochs	0.28	1h19m
tanh, SGD, 5 epochs	0.10	39m

Borealis

Best Model: *NN-tanh-sgd-15*

Combinations	Epochs Number	R^2 Score
1)	5	-3.61631
2)	2	-0.45123
3)	1	-56.5512
4)	8	-0.01833
5)	3	-0.54082
6)	15	0.08816

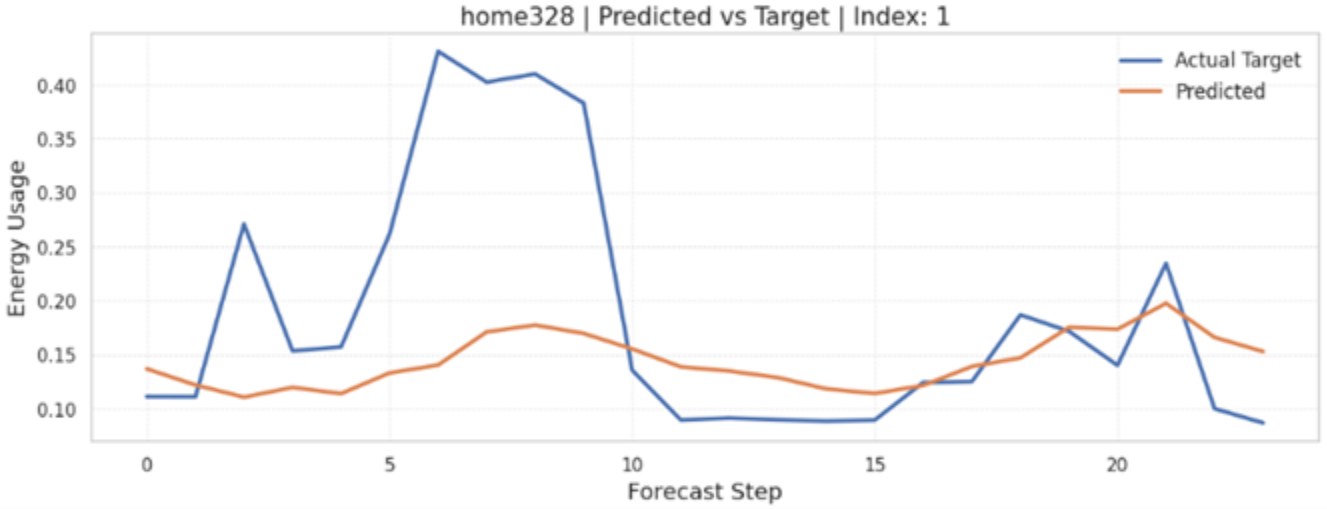


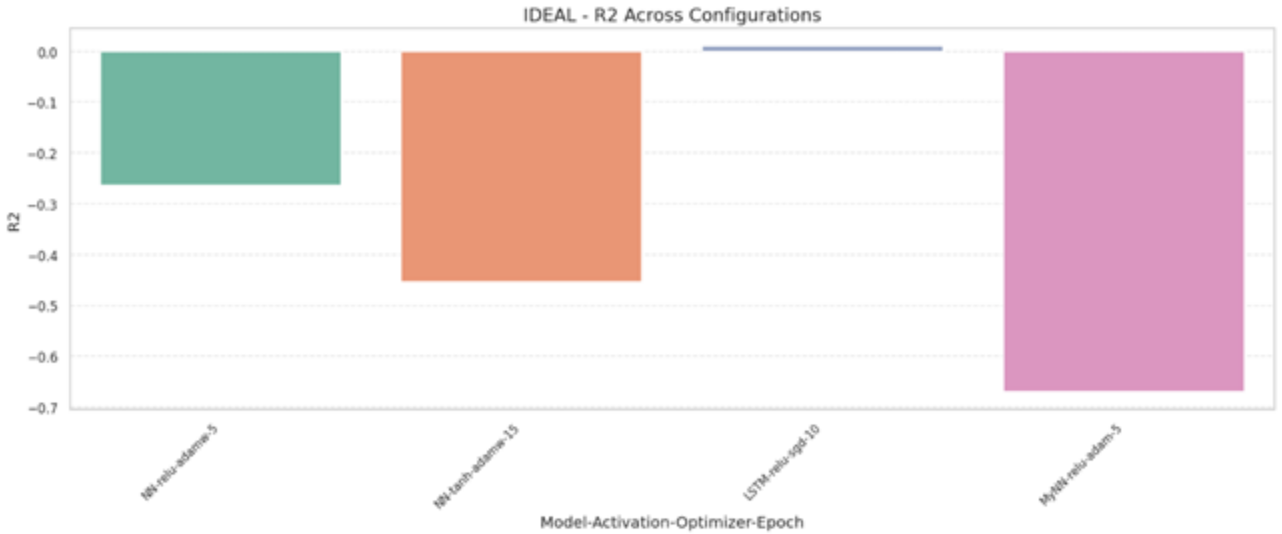


$R^2 = 0.95$ ✓

Plot: Predicted vs Target

IDEAL: electricity meter data from 255 homes in Edinburgh, UK





Predicted vs Actual:

- trend captured; underestimation
- R^2 mostly negative; best ≈ 0.01
  - (LSTM | relu | sgd | 10)

Challenges:

- negative R^2 → poor fit
- underfitting → missed spikes
- long runtimes → fewer tests

Discoveries:

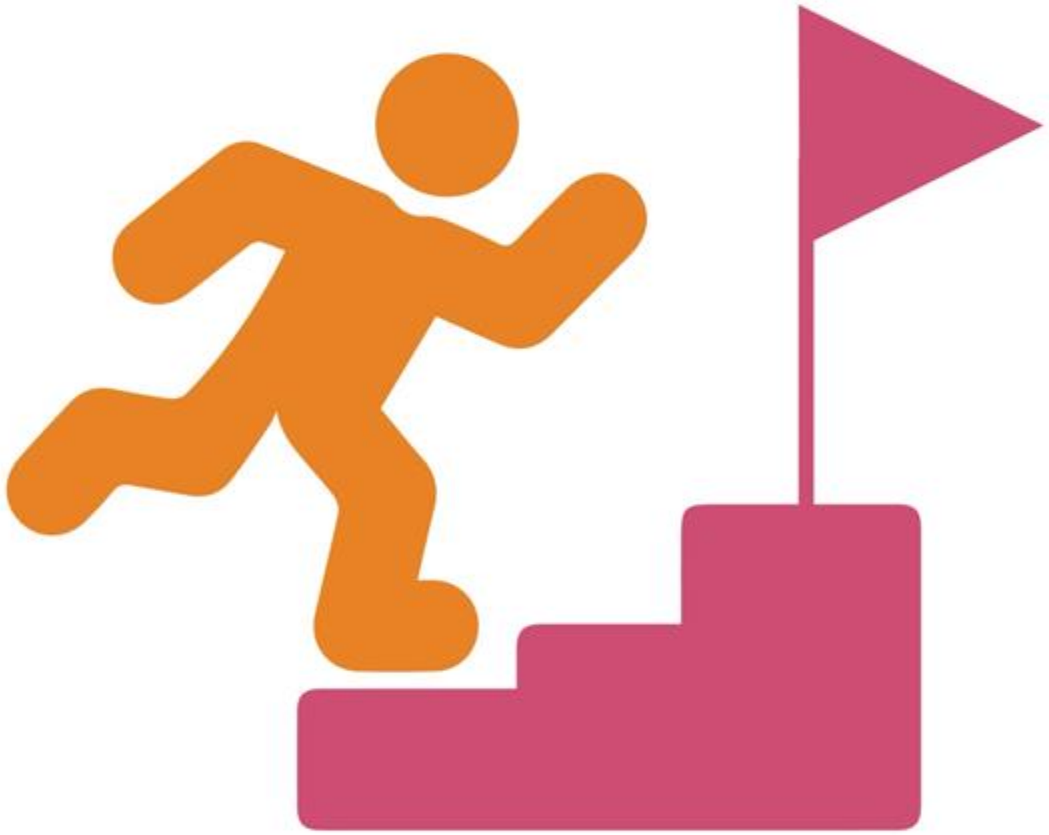
- RNN/LSTM/GRU > NN
- activation (tanh best)
- more epochs → higher R^2 (0.95)

CHALLENGES

- Using supercomputers for the first time & learning about machine learning concepts

- Data Size

- Computational constraints & Model Selection



Acknowledgments








Argonne NATIONAL LABORATORY

NERSC

OAK RIDGE National Laboratory

U.S. DEPARTMENT of ENERGY

Perlmutter

Argonne Intro to HPC Bootcamp

Group 6a

Evaluating Large Language Models for HPC Education

08/15/2025

Michael Manriquez, Leena Nawaz, Maria Jaiyeola,
Justin Tchoffa, Alexis Anderson, Keegan Krawczyk,

Murat Keceli
Project Leader, Argonne National Lab

Oluwaseun Ajayi
Project Leader

Leena Nawaz
Elmhurst University

Michael Manriquez
Harold Washington College

Kevin Tchoffa
University of Illinois Chicago

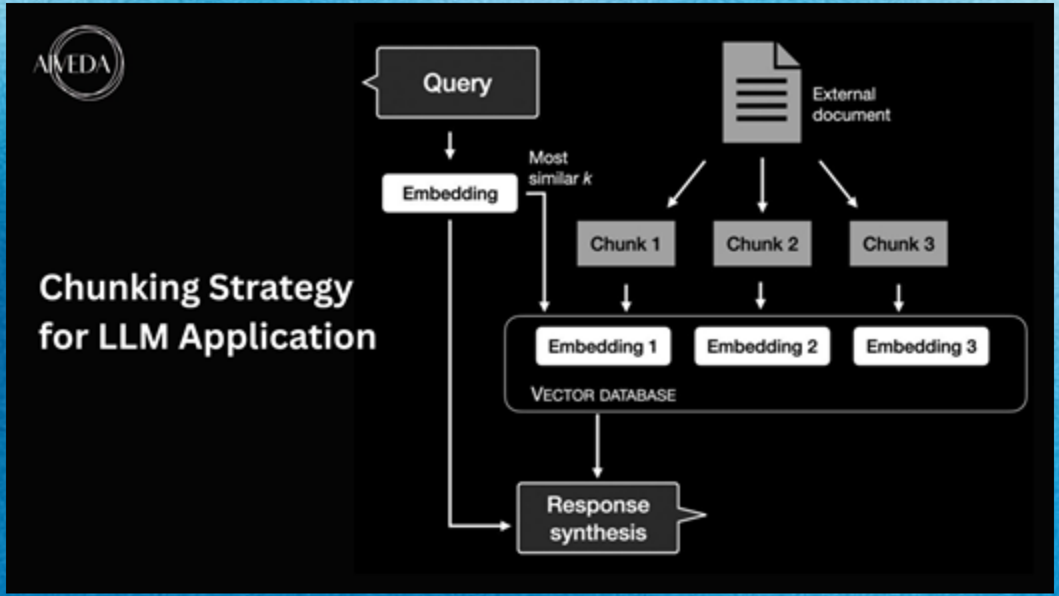
Maria Jaiyeola
Portland State University

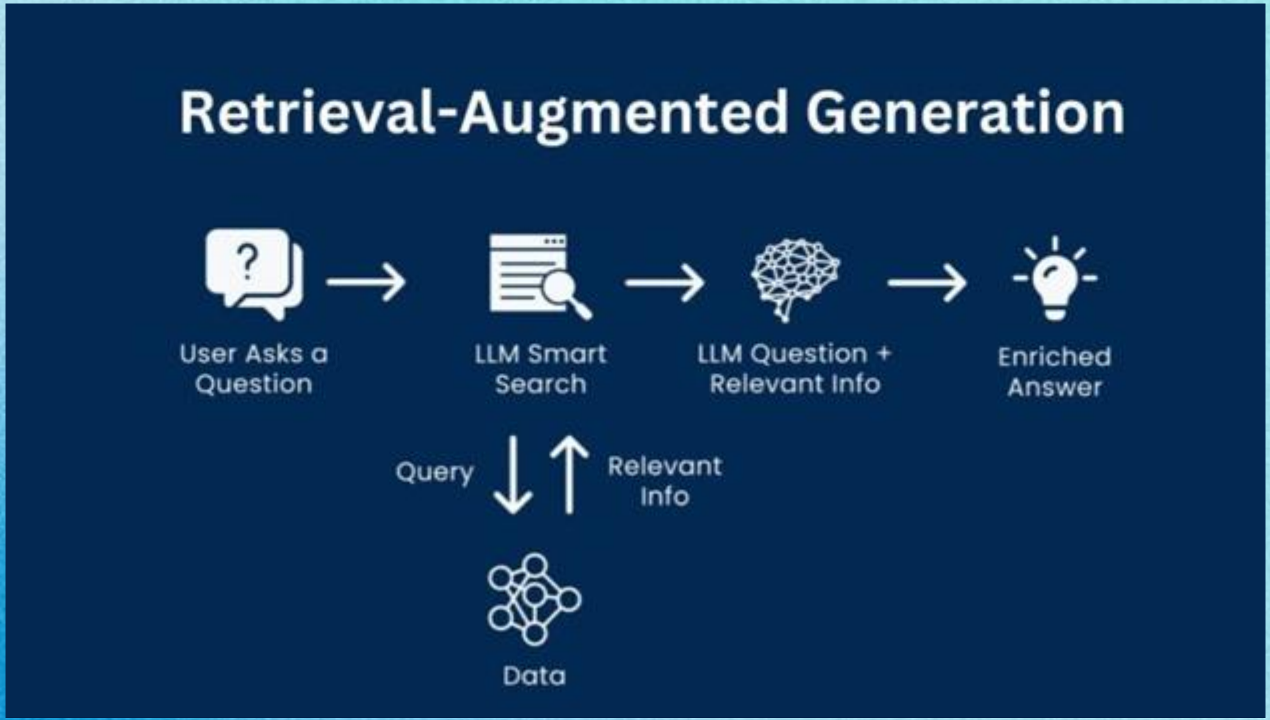
Keegan Krawczyk
Moriane Valley Community College

Alexis Anderson
Peer Mentor

Project Overview

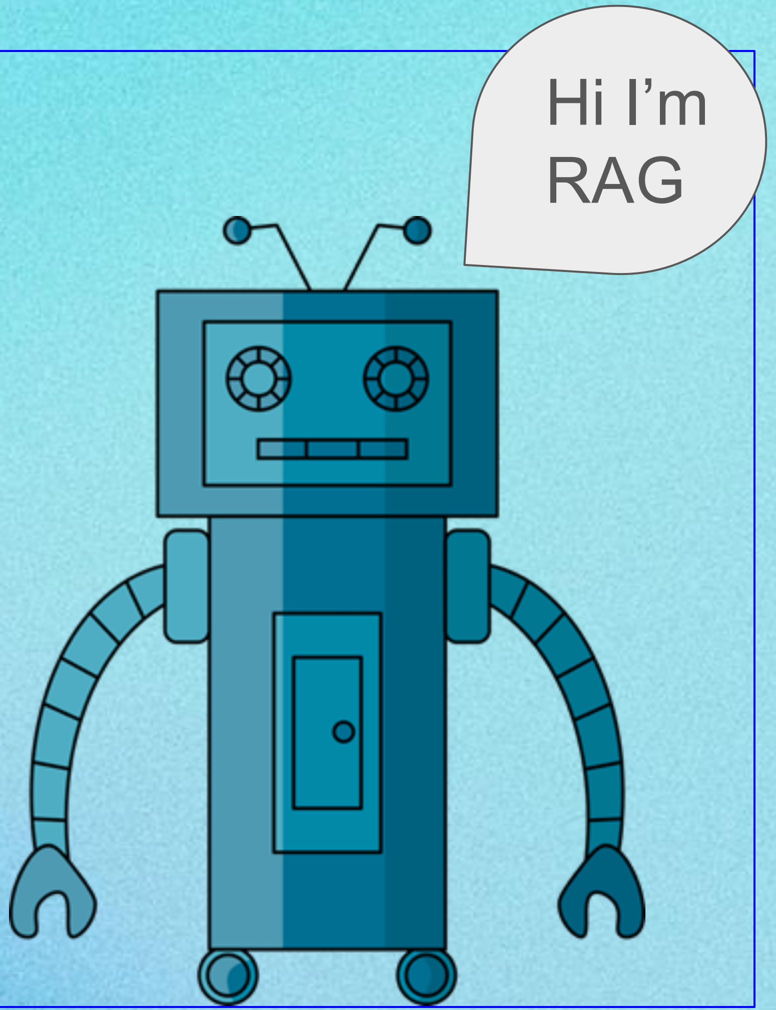
Train a LLM with RAG to assist in HPC education.





The Process

1. Import the documents to feed the RAG
2. The RAG model…
   - Reads the documents
   - Splits it up into chunks
   - Embeds the chunks into tokens
   - Stores chunks into a database
3. RAG + Question = LLM response



Results

Challenges

- Running the code and multiple bugs
- Communication
  - Different groups at the start and then coming together
- Large documents: hard to get accurate answers

Next Step

- Make the interface look better
- Train it for longer
- Use different LLM

Thank You

Project leads: Murat Keceli, Oluwaseun Ajayi
Peer Mentor: Alexis Anderson
HPC Bootcamp Organizers

Group 5

Argonne Intro to HPC Bootcamp

Project 5: Large Language Models for Science

Adam Trojak, Bhumi Choudhary, Fadzai Zivanai,
Fatima Mora Garcia, Jason Griffith, Yumna Hussain

Our Group Members

Project Lead: Sam Foreman
Computational Scientist
Argonne National Laboratory

Peer Mentor: Rene Montelongo
California State University: Northridge

Fadzai
Portland State University

Jason Griffith
NC A&T State Uni.

Adam Trojak
University of Illinois Chicago

Yumna Hussain
University of Illinois Chicago

Bhumi Choudhary
University of Illinois Chicago

Fatima Mora Garcia
University of Illinois Chicago

Background

- What is LLM?

Objectives

- Exploring how LLM's are trained on Scientific Datasets
- Learn about Tokenization and Linear Regression
- Understand challenges in training Scientific Models
- Train a Model from scratch

Tools Learned & Used

Platforms & Libraries

EzPz (HPC setup boilerplate)

Wordplay (LLM training)

Hugging Face

Weights & Biases

Environments

- Jupyter Notebook
- Google Colab
- Vim (for editing scripts)

HPC Tools

- Perlmutter !!
- SLURM for job scheduling
- Bash scripts for requesting and managing resources

Shakespeare Model Training

```
What is an LLM?ZIoZo-om';-'MAhB,RcOVP!JJhhkkJnnUzI''&D&jH!ddWJJhfUUVkRhZoZ:MoJRtDjkkhhdMM'Sdd-
'DqBJtHH;!ozZIZokzoooYlMKLJm.DDmkkXRX'NnhMSccJsH;Ude.tRzDoUtm'JmCd;Jd&j'Qo&'$$DAJTPPVv&j'jjtmn
```

```
What is an LLM?

DURENCK:
Me so my nou, hou ward thes ler noms he he,
Oxt my the my de is by beperd.
```

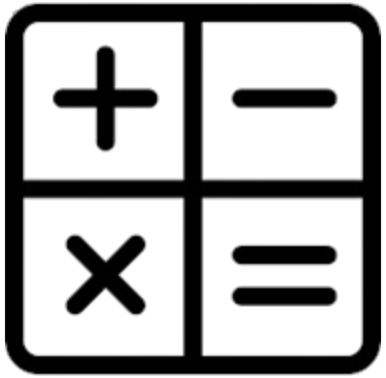
```
What is an LLM? What, that the wild my lord,
And the shal to may so shal that the shall thee.

RICHARD:
What that there thee shal the const the shall so thine.
```

Shakespeare Model Training

Challenges

Debugging
Script Implementation
Distributed Training

Jupyter Notebooks
& Google Colab
Implementation



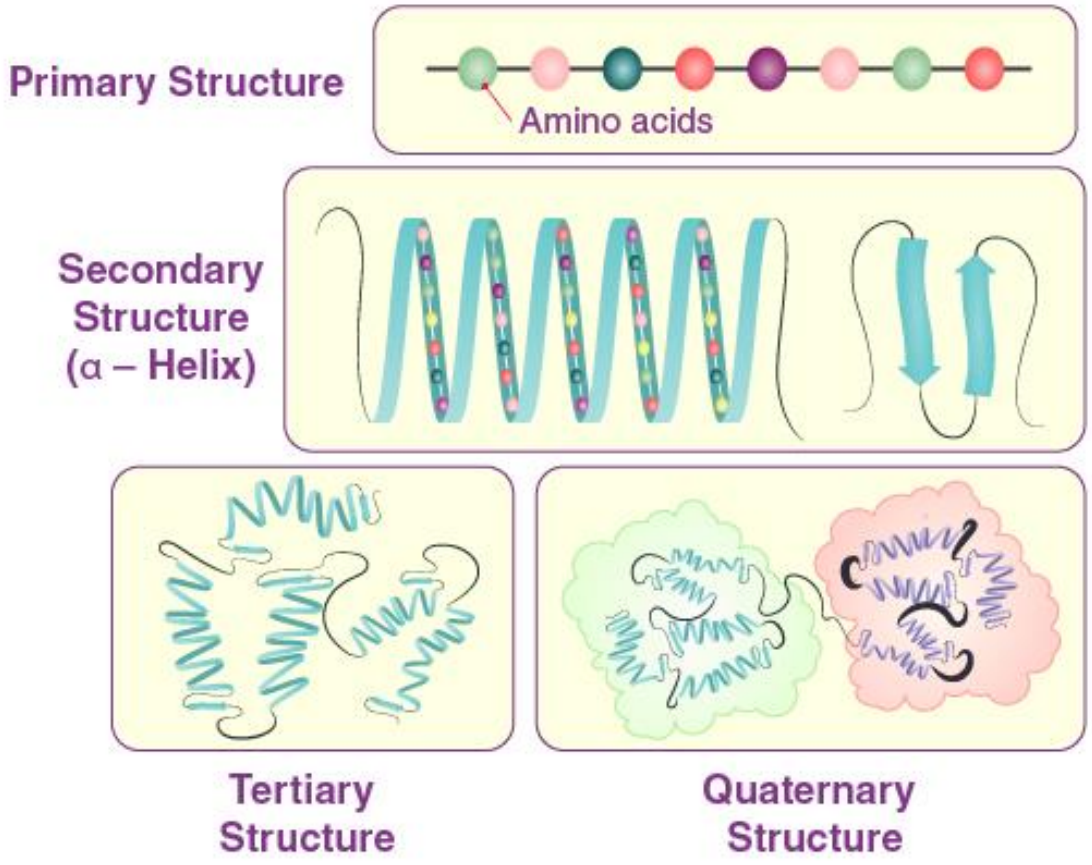
Algorithms
Syntax

Conclusion & Next Steps

- Learned about Python package configurations in different virtual environments
- Fundamentals of LLMs
- Distributed training to maximize performance and minimize cost of LLM training

Next Steps

- Try to train larger models on different datasets (Ex. Genome Model)
- Experiment with different HPC configurations



THANK YOU

Sam and Rene

Thank you to everyone who helped make this bootcamp possible!