



VI SEMESTER

**SEE Presentation
on
Crystoper–Prediction of Protein Crystallization
Conditions Using Big Data Techniques**

Domain: Bioinformatics



Team Members

Go, change the world

K preethi	Roopa Iranna Bagalkoti	Gagan Gowda V S
1RV23AI402	1RV23AI404	1RV23AI400

MENTOR

Dr. Vijayalakshmi M N
Associate Professor
Dept. of AIML
RVCE



Contents

1. Introduction
2. Existing system
3. Identification of problem
4. Literature Review
5. Problem definition
6. Objectives
7. Methodology
8. Tools and Techniques used
9. Results and Discussion
10. Conclusion
11. References



Introduction

1. Crystoper is a Big Data-driven system developed to analyze protein crystallization conditions—a crucial step in structural biology and drug discovery.
2. The traditional trial-and-error process of finding suitable crystallization conditions is time-consuming and resource-intensive.
3. With the availability of large-scale experimental datasets, this project leverages MapReduce and Apache Spark to extract insights from over 15,000+ records, providing visual analytics to help researchers determine optimal pH, temperature, and methods for crystallization



Existing System

Traditional Methods:

- Relies on trial-and-error in labs
- Requires testing various combinations of pH, temperature, chemicals
- Low success rate, time-consuming, and costly

Computational Tools

- Some models use sequence-based features for prediction
- Pipelines extract patterns from PDB records
- Mostly use static datasets with limited scope.

Limitations

- No support for large-scale (Big Data) analysis
- Lack of visualization tools for pattern discovery
- No heat/resource monitoring (CPU, memory, time)
- Often complex or costly to set up and use



Identification of problem

Observations from Existing Systems:

1. Manual trial-based crystallization is time-consuming and inefficient.
2. Existing models are not scalable to handle large datasets (10K+ records).
3. Lack of interactive visualizations makes insight extraction difficult.
4. No support for real-time analysis or predictions.
5. Researchers have no visibility into resource usage (CPU, memory, execution time).
6. Many tools are either costly, complex, or require advanced infrastructure.

There is no unified, scalable, and interactive platform that can efficiently analyze, visualize, and predict protein crystallization outcomes using Big Data technologies.



Literature Review

Go, change the world

Title	Author	Summary
Large-scale analysis of protein crystallization conditions	Newman et al	Created a crystallization condition database and identified frequent successful conditions using clustering.
Predicting successful crystallization using machine learning	Overton et al	Used decision trees to predict crystallization outcomes from protein sequence and experimental features.
Towards automated protein crystallization: using AI to select conditions	Cumbaa et al	Developed an AI system to suggest likely crystallization conditions based on historical success data.
Machine learning in protein crystallization	Georgiev	Explored use of SVMs and neural networks to predict crystallizability from protein sequence data.
Big Data in Structural Biology	Marx	Described how large-scale structural databases (like PDB) and data mining help in understanding macromolecular crystallization.



Literature Review

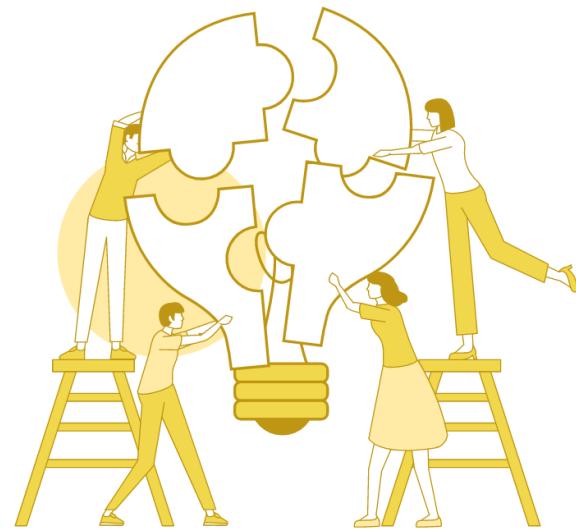
Recent research in protein crystallization since 2020 has increasingly integrated AI, machine learning, and big data approaches to improve prediction and success rates.

- Newman et al. developed a crystallization condition database to identify optimal conditions using clustering techniques.
- Overton et al. applied decision tree models to predict crystallization outcomes based on protein sequences and experimental parameters.
- Cumbaa et al. introduced an AI-based recommendation system for selecting suitable crystallization conditions.
- Georgiev employed SVMs and neural networks to analyze protein sequences for crystallization potential.
- Marx highlighted the role of big data and large-scale biological databases in advancing structural biology through data-driven insights.



Problem Definition

To develop an intelligent Big Data-based system named Crystoper that can efficiently process, analyze, and visualize protein crystallization experiments using MapReduce, Spark simulation, and an interactive dashboard—helping researchers identify optimal crystallization conditions and understand system resource usage at each processing stage.





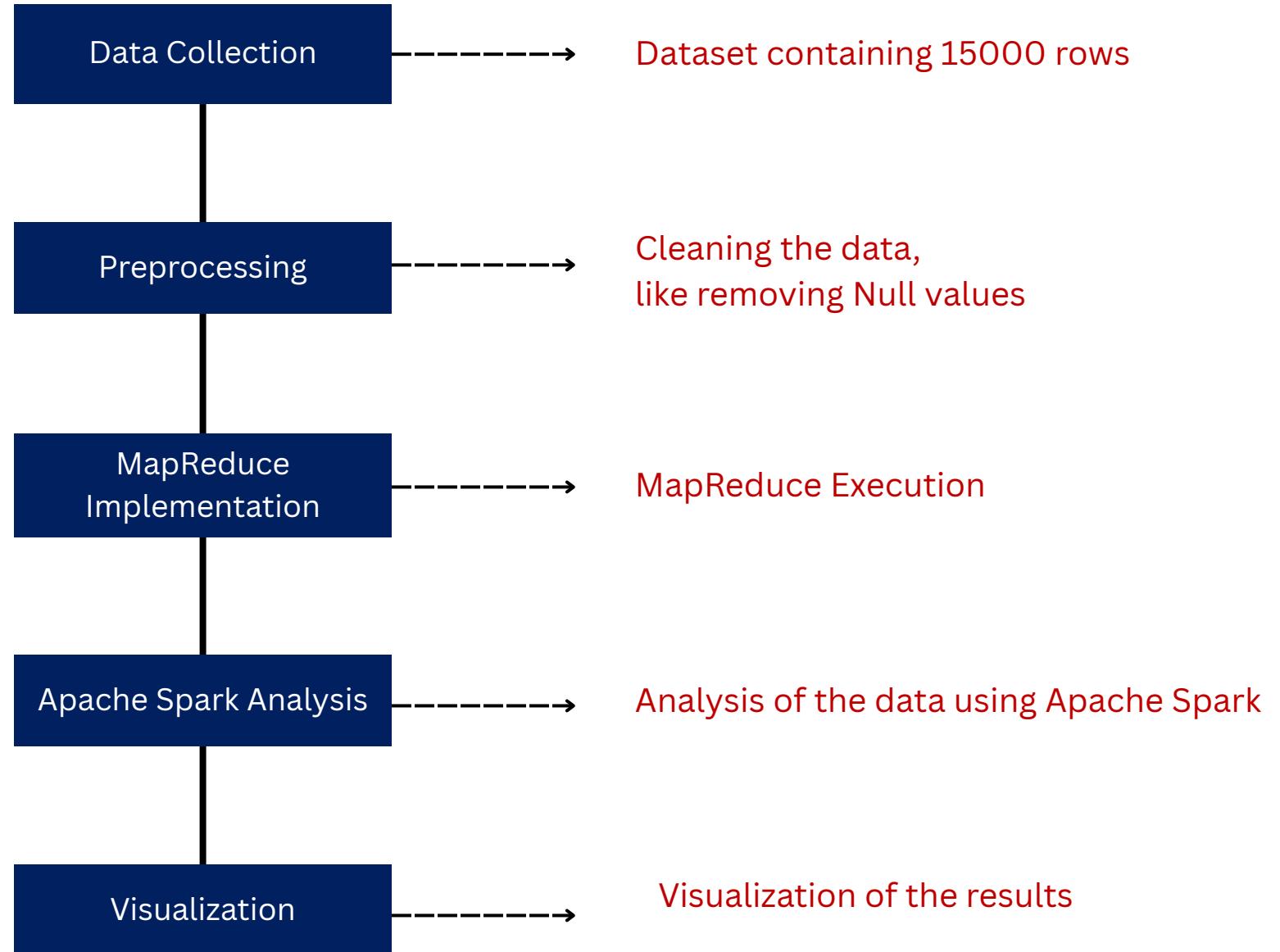
Objectives

1. Analyze a large-scale dataset (15,000+ records) of protein crystallization experiments.
2. Apply MapReduce to calculate Average pH, Average temperature, Average sequence length
3. Number of trials per method
4. Use Apache Spark for scalable distributed aggregation and processing.
5. Visualize trends with bar graphs, heatmaps, pie charts, and line plots.
6. Develop an interactive Streamlit dashboard for real-time user interaction.
7. Build a Random Forest ML model for crystallization success prediction



Methodology

Go, change the world





Methodology

Go, change the world

Data Collection

- Synthetic dataset with 15,000+ crystallization records
- Includes features like: pH, temperature, sequence length, method, chemicals.

Preprocessing

- Handled missing values, normalized categories
- Converted data types (e.g., float for pH, temperature)
- Balanced the dataset ($\approx 50\%$ crystallized vs non-crystallized)

MapReduce Implementation

- Mapper extracts method-wise:
 - Count, pH, temperature, sequence length
- Reducer aggregates to compute averages and totals
- Outputs summarized method-wise statistics



Methodology

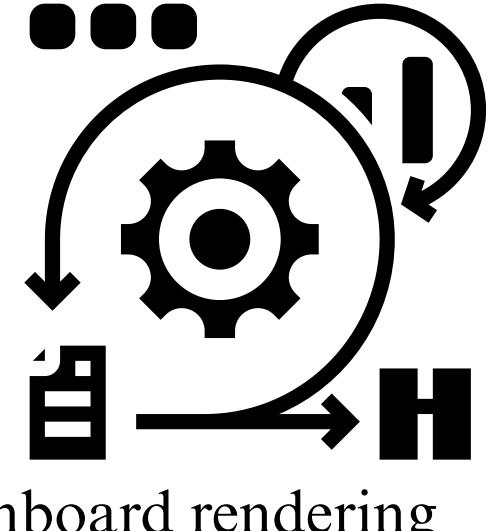
Go, change the world

Apache Spark Analysis

- Spark used for scalable grouped aggregation

Performed:

- Avg pH, temperature, sequence length
- Grouped by crystallization method
- Converted Spark DataFrame → pandas for dashboard rendering



Visualization & Dashboard (Streamlit)

- Plots: bar charts, pie charts, heatmaps
- Interactive UI with upload and result display
- Monitored:
 - Execution time
 - CPU & memory usage (Heat analysis)



Tools and Techniques Used

Big Data Processing

MapReduce

- Used for distributed summarization (method-wise stats)
- Implemented using Python

Apache Spark

- In-memory processing
- Faster aggregations on large datasets

Visualization

- Matplotlib & Seaborn
 - Bar graphs, pie charts, heatmaps, line plots
- Streamlit
 - Interactive web-based dashboard
 - Real-time display of stats and system metrics





Tools and Techniques Used

Performance Monitoring

psutil (Python)

- CPU usage
- Memory consumption
- Execution time

Machine Learning

• Random Forest Classifier

- For predicting crystallization success
- Integrated into Streamlit app.





Results and Discussion

MapReduce Results

1. Analyzed 15,000+ records
2. Identified top methods: Vapor Diffusion, Sitting Drop
3. Computed: Avg pH, Avg Temperature, Avg Sequence Length
4. Execution Time: ~1.23 seconds

Apache Spark Insights

1. Performed grouped aggregation at scale
2. Avg pH, Temp, and Sequence

Length per method

1. Execution Time: ~3.14 seconds
2. Memory Usage: ~250 MB
3. Better scalability for larger datasets

The dashboard displays a preview of the protein crystallization dataset with columns including Protein_ID, Sequence_Length, Molecular_Weight_kDa, Isoelectric_Point, Hydrophobicity, Secondary_Structure, Buffer_Type, Precipitant_Type, Precipitant_Concentration_N, pH, Temperature_C, Crystallization_Method, and Crystallized. It also shows a summary of the dataset with 15000 total records, 4856 crystallized, and 10144 non-crystallized entries.

Protein_ID	Sequence_Length	Molecular_Weight_kDa	Isoelectric_Point	Hydrophobicity	Secondary_Structure	Buffer_Type	Precipitant_Type	Precipitant_Concentration_N	pH	Temperature_C	Crystallization_Method	Crystallized
0	152	45.96	9.22	0.381	Alpha	Tris	Ammonium sulfate	15.41	9.2	20	Free interface diffusion	0
1	485	85.5	7.85	-0.523	Alpha	Tris	PEG 3350	31.72	8.9	25	Vapor diffusion	1
2	918	110.91	9.57	-0.768	Beta	Phosphate	PEG 3350	32.88	5.6	25	Batch under oil	0
3	339	82.27	9.05	0.771	Beta	HEPES	Ethanol	29.5	5.1	37	Microbatch	0
4	154	140	4.05	-0.173	Beta	Tris	PEG 3350	5.58	8.5	25	Dialysis	0

Dataset Summary

Total Records	15000
Crystallized (0)	4856
Not Crystallized (0)	10144

Missing Values:

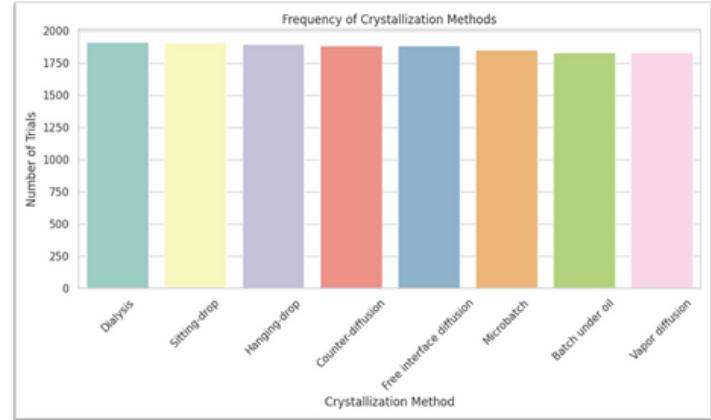
Protein_ID	0
Sequence_Length	0
Molecular_Weight_kDa	0
Isoelectric_Point	0
Hydrophobicity	0
Secondary_Structure	0



Results and Discussion

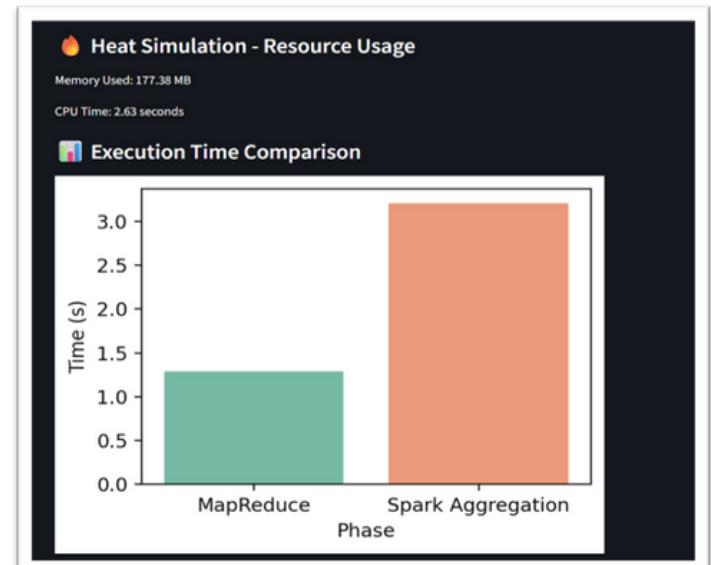
Visual Insights

- Pie Chart: ~50–60% crystallization success rate
- Bar Graph: Vapor Diffusion most commonly used method
- Line Chart: Different methods favor specific pH ranges
- Heatmap: Moderate positive correlation between pH and crystallization success.



System Load Analysis (Heat Monitoring)

- Spark consumed more CPU & memory, ideal for large-scale future use
- MapReduce better suited for medium-sized datasets





Results and Discussion

1. Data-driven approach helps prioritize optimal crystallization conditions
2. Visualization improves understanding for biologists
3. System performance metrics support real-world feasibility
4. Platform is modular and integrated with ML models
5. The model achieved an accuracy of 90.96%, indicating strong predictive capability.

Predict Crystallization Method

Make a Prediction

Precipitant Concentration (%)
25.00 - +

pH
7.00 - +

Buffer Type
Citrate

Secondary Structure
Alpha

Molecular Weight (kDa)
45.00 - +

Sequence string

Predict

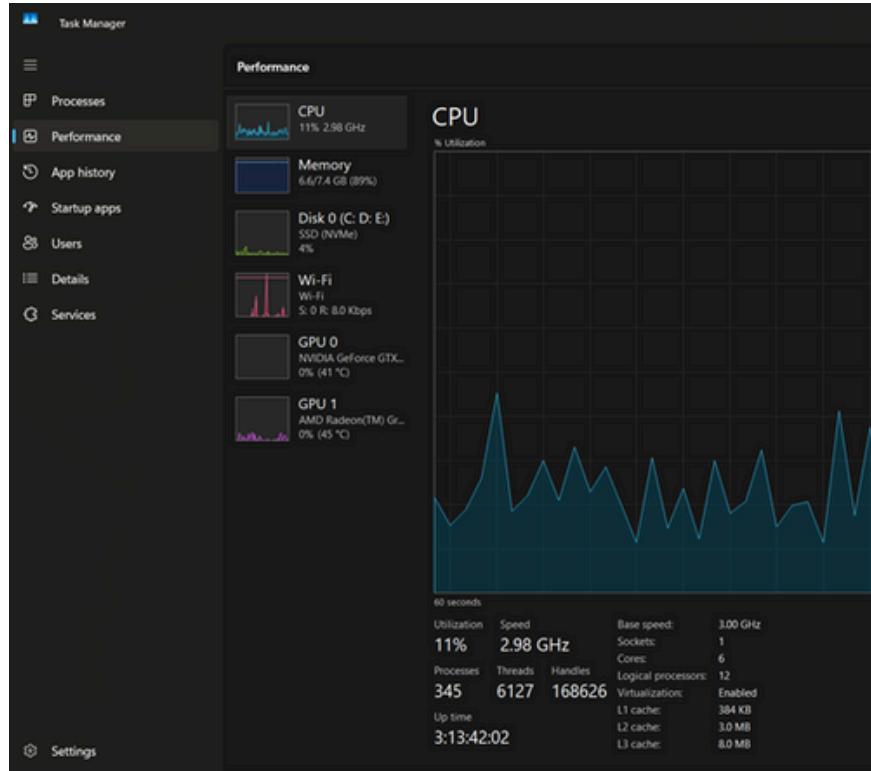
Predicted Crystallization Method: Vapor diffusion



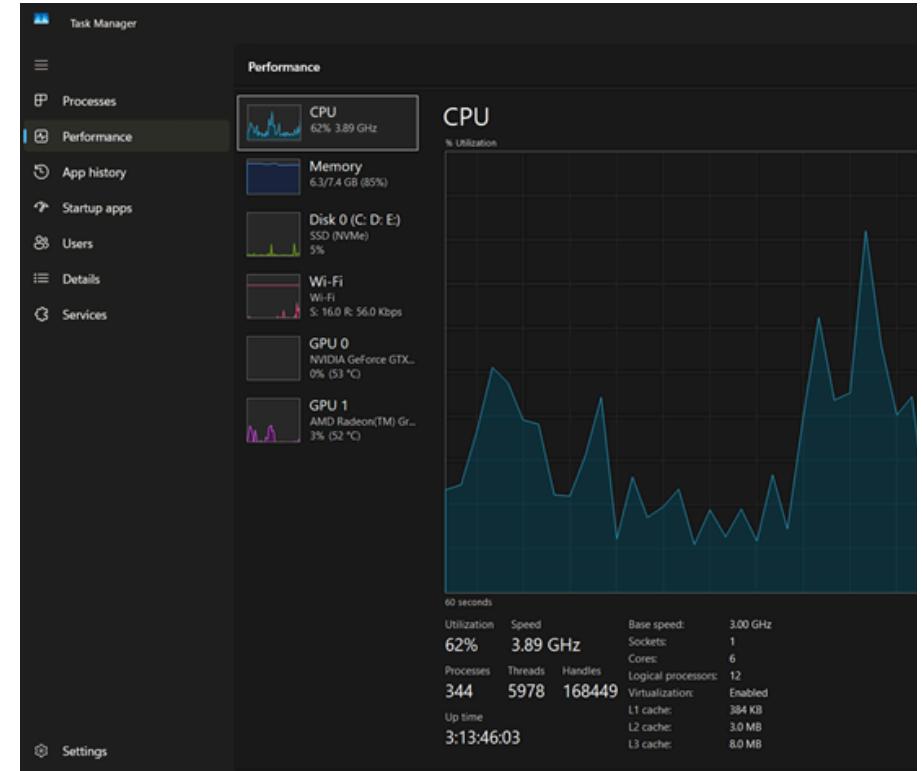
Results and Discussion

Heat Analysis of CPU:

CPU usage Before Processing



CPU usage After Processing





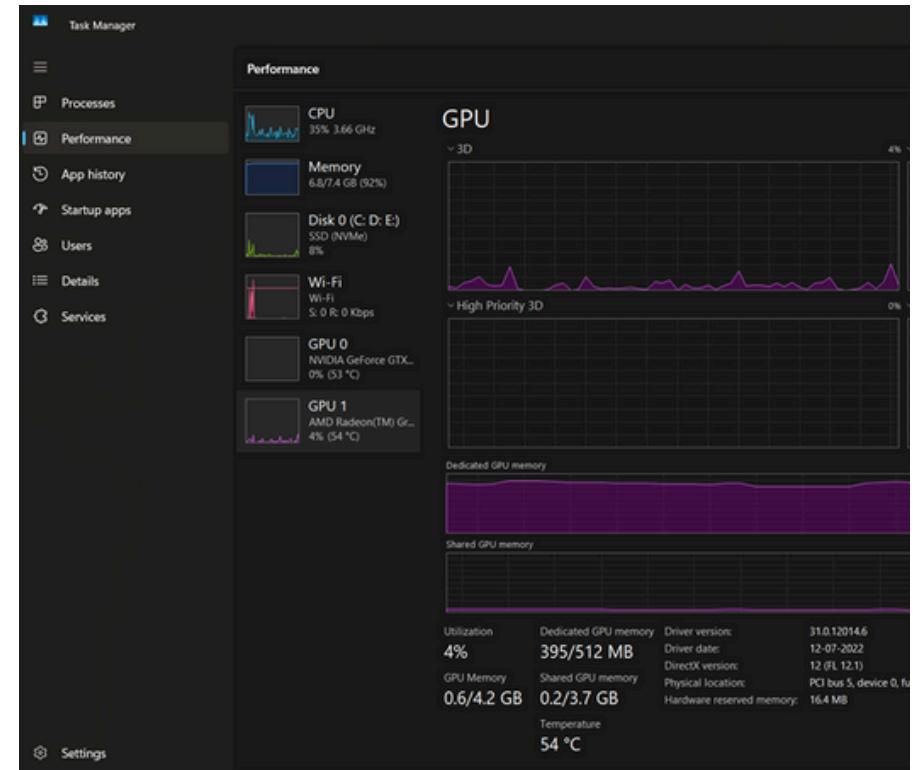
Results and Discussion

Heat Analysis of GPU:

GPU usage Before Processing



GPU usage After Processing





Results and Discussion

GPU usage Using HW-Monitor during Processing

The image shows two side-by-side screenshots of the HWMonitor application interface. Both screenshots display sensor data for two different systems: 'GAGAN' and 'SAMSUNG MZVLQ512HBLU-0'. The left screenshot for 'GAGAN' shows detailed information for an AMD Ryzen 5 Mobile 4600H processor, including Voltages, Temperatures, Powers, Currents, Levels, Utilization, and Clocks. The right screenshot for 'SAMSUNG MZVLQ512HBLU-0' shows information for an NVIDIA GeForce GTX 1650 GPU, including Speed, Temperatures, Powers, Currents, Counters, Clocks, and Utilization.

GAGAN (AMD Ryzen 5 Mobile 4600H)

- Voltages
 - CPU VDD: 0.750 V
 - SoC VDD: 0.875 V
 - VID (Max): 0.725 V
- Temperatures
 - Package: 61.8 °C
 - Cores (Max): 64.0 °C
 - L3 Cache #0: 61.9 °C
 - L3 Cache #1: 62.9 °C
- Powers
 - Package: 8.79 W
 - Cores: 2.48 W
 - PPT Limit: 31.50 W
- Currents
 - CPU VDD: 7.27 A
 - SoC VDD: 24.94 A
- Levels
 - Energy Performance: 69.8 %
- Utilization
 - Processor
 - CCX #0: 31.6 %
 - CCX #1: 32.2 %
- Clocks
 - Core #0: 1395.4 MHz
 - Core #1: 1395.4 MHz
 - Core #2: 2167.9 MHz
 - Core #3: 2167.9 MHz
 - Core #4: 2167.9 MHz
 - Core #5: 2167.9 MHz
 - CPU BCLK: 99.7 MHz
 - Core #0: 1234.3 MHz
 - Core #1: 1236.0 MHz
 - Core #2: 1193.4 MHz
 - Core #3: 1234.3 MHz
 - Core #4: 1195.6 MHz
 - Core #5: 1234.3 MHz
 - CPU BCLK: 99.0 MHz
 - Core #0: 3995.8 MHz
 - Core #1: 3995.8 MHz
 - Core #2: 3995.8 MHz
 - Core #3: 3991.2 MHz
 - Core #4: 3995.8 MHz
 - Core #5: 3991.2 MHz
 - CPU BCLK: 100.3 MHz

SAMSUNG MZVLQ512HBLU-0...

- Temperatures
 - Assembly: 49.0 °C
 - Sensor 1: 49.0 °C
- Counters
 - Power-On-Hours: 2459 hrs
 - Power Cycle Count: 40066
 - Health Status: 93 %

NVIDIA GeForce GTX 1650

- Speed
 - Read Rate: 0.04 MB/s
 - Write Rate: 0.11 MB/s
- Temperatures
 - GPU: 61.2 °C
 - Hot Spot: 70.0 °C
- Powers
 - GPU: 5.76 W
 - Core Power Supply: 2.14 W
 - Frame Buffer Power Su...: 2.62 W
- Currents
 - GPU: 0.33 A
 - Core Power Supply: 0.15 A
 - Frame Buffer Power Su...: 0.18 A
- Counters
 - PCIe Lanes Errors: 0
 - PCIe Replay Counter: 0
 - PCIe Replay Rollover C...: 0
 - PCIe Fatal Errors: 0
 - PCIe Non Fatal Errors: 0
 - PCIe Correctable Errors: 0
 - PCIe CRC Errors: 0
 - PCIe Receiver Errors: 0
 - PCIe Unsupported Req...: 0
 - PCIe Bad DLLP: 0
 - PCIe Bad TLP: 0
 - PCIe NAK Received: 0
 - PCIe NAK Sent: 0
 - PCIe PEX Errors Recov...: 451
 - PCIe PEX Errors Recovered: 11
- Clocks
 - Graphics: 300.0 MHz
 - Memory: 405.0 MHz
 - Video: 540.0 MHz
- Utilization
 - GPU: 0.0 %
 - Memory: 3.9 %
 - Frame Buffer: 0.0 %



Conclusion

Go, change the world

1. Built Crystoper – a Big Data-driven system for protein crystallization analysis
2. Processed 15,000+ experimental records
3. Used MapReduce and Apache Spark for distributed data processing
4. Supports data-driven decision-making in structural biology
5. Replaces expensive, inefficient trial-and-error lab processes
6. Platform is scalable and modular for research and academic use



References

Go, change the world

- [1] A. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, “Spark: Cluster Computing with Working Sets,” Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing (HotCloud '10), Boston, MA, USA, June 2010.
- [2] J. Dean and S. Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters,” Communications of the ACM, vol. 51, no. 1, pp. 107–113, Jan. 2008.
- [3] M. Zaharia et al., “Apache Spark: A Unified Engine for Big Data Processing,” Communications of the ACM, vol. 59, no. 11, pp. 56–65, Nov. 2016.
- [4] RCSB Protein Data Bank, “PDB Statistics: Growth of Released Structures per Method,” [Online]. Available: <https://www.rcsb.org/stats/>. [Accessed: June 2025].
- [5] P. Bhatotia, A. Wieder, R. Rodrigues, U. A. Acar, and R. Pasquin, “Incoop: MapReduce for Incremental Computations,” in Proceedings of the 2nd ACM Symposium on Cloud Computing, 2011, pp. 1–14.
- [6] S. Ghosh, “Big Data Analytics Using Spark for Massive Datasets,” International Journal of Computer Applications, vol. 179, no. 27, pp. 15–19, Feb. 2019.
- [7] D. Loshin, Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph, Morgan Kaufmann, 2013.
- [8] M. Armbrust et al., “Scaling Spark in the Real World: Performance and Usability,” in Proceedings of the VLDB Endowment, vol. 8, no. 12, pp. 1840–1851, Aug. 2015.
- [9] S. Ahmad and J. Lee, “Big Data Analytics in Bioinformatics,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 12, no. 1, pp. 13–24, Jan.–Feb. 2015.
- [10] L. Neumann et al., “Protein Crystallization: From Traditional Methods to Microfluidics,” Angewandte Chemie International Edition, vol. 47, no. 36, pp. 6832–6850, Sept. 2008.



RV College of Engineering®

Mysore Road, RV Vidyanketan Post,
Bengaluru - 560059, Karnataka, India

Go, change the world

Thank You