SMART INDIA HACKATHON 2025

- **Problem Statement ID – 25042**
- **Problem Statement Title- Identifying Taxonomy and Assessing Biodiversity from eDNA**
- **Theme - Miscellaneous**
- **PS Category- Software**
- **Team ID-**
- **Team Name (Registered on portal) : STRIVE**

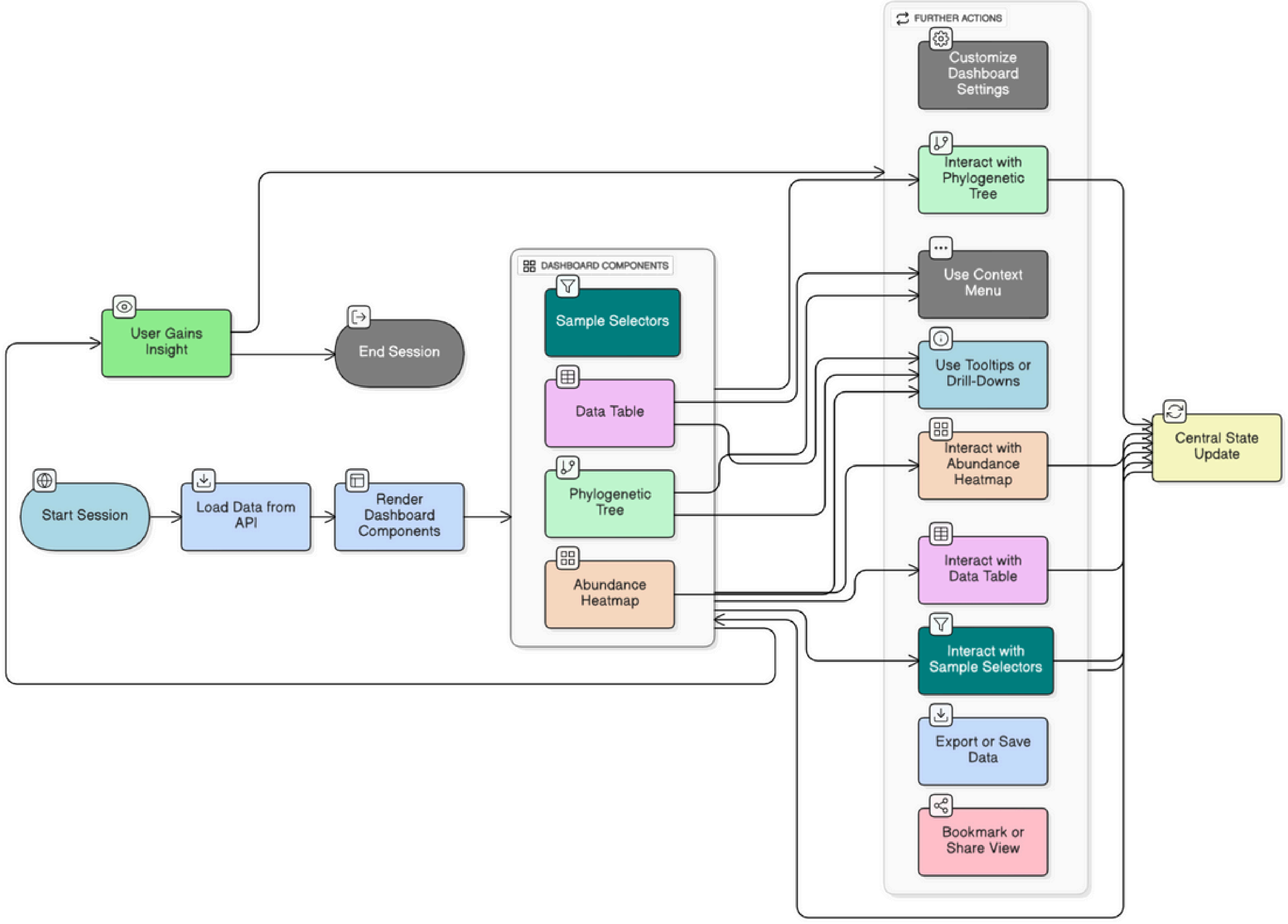
Strive

AQUANOVA

SMART INDIA HACKATHON 2025

❖ Proposed Solution

- **AI-driven pipeline** using deep & unsupervised learning on raw eDNA reads.
- **Minimizes** reliance on reference databases.
- Performs **taxon identification**, annotation, and abundance estimation in a **single workflow**.
- Made **FCGR images** from **ASVs**, which we need to train our **CNN model** and did the clustering using **HDBSCAN** for phylogenetic placement of novel taxa.
- **Scale of the problem:** Oceans cover 71% of Earth, yet 80% of marine species remain undiscovered. Existing eDNA pipelines rely heavily on reference databases that poorly represent deep- sea taxa.
- **Impact on science & society:** Incomplete biodiversity data → weak conservation policies. Slower analysis → delays in detecting ecological threats.
- **Why it matters?** Protecting marine ecosystems = safeguarding climate balance, fisheries, and human livelihoods



Strive

TECHNICAL APPROACH

SMART INDIA HACKATHON 2025

❖ Technical Stack

Frontend: React, TypeScript, Redux, Bootstrap

Backend: Fast API, Java SpringBoot

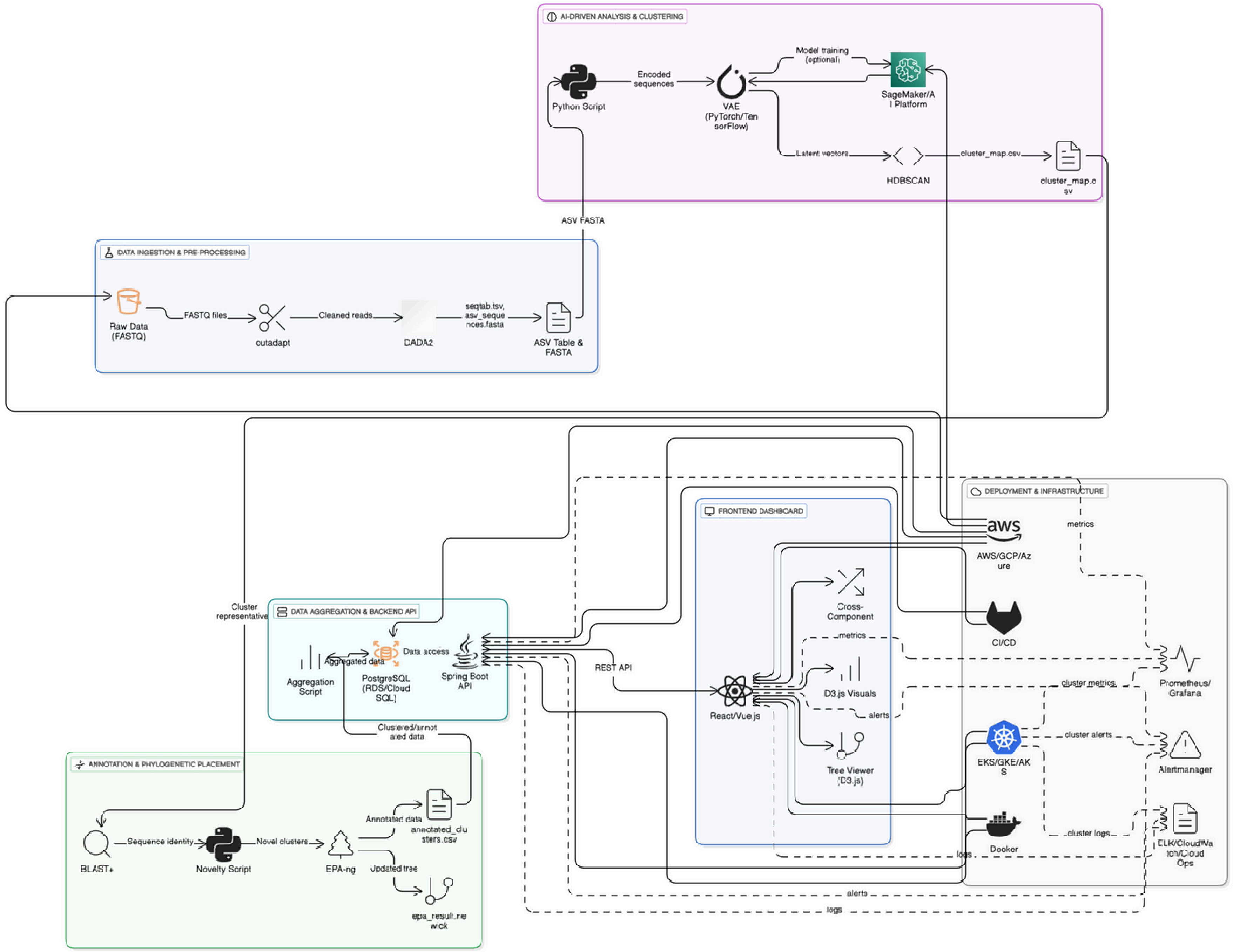
AI/ML: PyTorch (model training), HDBSCAN (novel lineage detection), BLAST+, EPA-ng (taxonomic assignment) Core classification and prediction
Tasks: Classifying waste / eDNA reads / pattern recognition

Database: PostgreSQL, MongoDB

Infrastructure: AWS, GitHub Actions, Jenkins

Other Services: Docker, Kubernetes, Monitoring Tools

System Architecture



Strive

FEASIBILITY AND VIABILITY

SMART INDIA HACKATHON 2025

Feasibility:

- Efficiency: It reduces the costs and logistical challenges of traditional deep-sea exploration.
- Scientific: It accelerates the discovery of deep-sea biodiversity.
- Modularization: The pipeline is broken into smaller, reusable components.

Viability:

- Cost-Effective: It uses open-source tools to minimize initial investment.
- Usability: The project has a clear roadmap for development.
- High Demand: There's a high demand for such solutions in deep-sea biodiversity research because the current version is really time consuming and not effective

Challenges:

1. Data Quality & Heterogeneity
2. Computational Resources
3. Reference Database Gaps

Solutions:

1. Implement rigorous filtering and normalization; robust error correction.
2. Optimize algorithms for cloud scalability; leverage GPU acceleration.
3. Utilize unsupervised learning (VAE, HDBSCAN) for novel lineage detection.

Buissness Potential:

Go-to-Market Strategy:

- Industry: Direct sales to biotech, energy, and environmental consulting firms.
- Academia/Gov't: Engage via scientific publications, conferences, and research partnerships.

Market Size:

- $1.2B in Marine eDNA Biomonitoring.

Use Cases:

- Environmental Agencies: The solution can be used for large-scale biodiversity and pollution monitoring to ensure compliance.
- Marine & River Conservation Projects: It facilitates eDNA-based species detection to protect ecosystems.

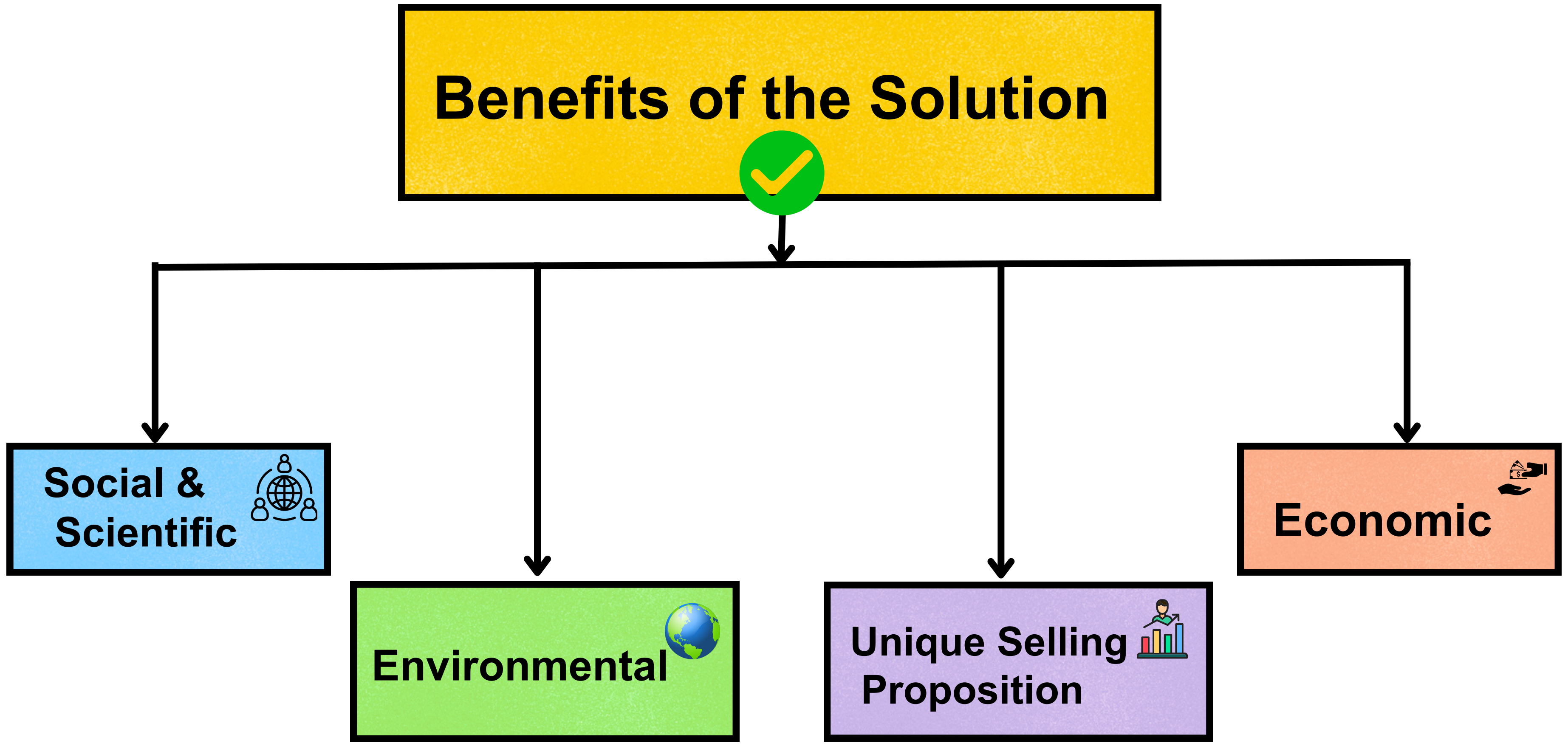
Supporting facts for Feasibility and Viability

- Open-source use: 70% of marine genomics projects rely on tools like QIIME2 & Kraken2 (Nature Biotechnology, 2023).
- Cost savings: eDNA reduces survey costs by up to 60% vs. trawling/submersibles (Marine Technology Society, 2022).
- Faster discovery: Species identification sped up from months to weeks (Frontiers in Marine Science, 2021).
- Global need: IUCN (2023) stresses urgent demand for user-friendly eDNA tools for deep-sea monitoring.

Strive

IMPACT AND BENEFITS

SMART INDIA HACKATHON 2025



Benefits of the Solution

Social & Scientific

Environmental

Unique Selling Proposition

Economic

❖ Potential Impact on the Target Audience

- Cost reduction: The solution reduces the high costs and logistical challenges of traditional deep-sea exploration, thanks to its use of open-source tools.
- Biotechnological potential: It can help identify novel species with potential for biotechnological applications.

❖ Benefits of the solution

Unique Selling Proposition (USP):

- An AI-driven pipeline for novel species discovery.
- Operates offline.
- It's low-cost and scalable.

Social & Scientific

- Accelerates the pace of deep-sea biodiversity discovery.
- Contributes to a comprehensive understanding of global biodiversity, supporting long-term research.

Economic:

- Identifies novel eukaryotic species with potential for biotechnological applications.
- Reduces high costs and logistical challenges of traditional deep-sea exploration.

Environmental:

- Informs critical conservation strategies for vulnerable marine habitats.
- Facilitates ecosystem monitoring for environmental changes.

Strive

RESEARCH AND REFERENCES

SMART INDIA HACKATHON 2025

❖ References

- environmental DNA (eDNA)
- Environmental DNA analysis
- Comparison of Bioinformatics
- Pipelines and Operating Systems
- Comparing bioinformatic pipelines
- The Deep Search Project
- Deep-sea water amplicon metagenomes

Dataset Type	Source/Generator	Ground Truth	Primary Metric	Formula/Definition	Purpose
In Silico Mock Community	Grinder	Known species composition & abundance	Adjusted Rand Index (ARI)	Measures similarity between true and predicted clusterings, corrected for chance.	To validate the accuracy of the unsupervised AOTU clustering (Module II).
***In Silico* Mock Community**	Grinder	Known species composition & abundance	Root-Mean-Square Error (RMSE)	∑*i*=1*n*(*yi*−*y*^*i*)2	To validate the accuracy of the corrected abundance estimates (Module V).
***In Silico* Mock Community**	Grinder	Known species composition & abundance	F-Measure	∑*i*=1*n*(*yi*−*y*^*i*)2	To validate the accuracy of the taxonomic annotation (Module III).
Real-World Deep-Sea Dataset	Field Samples	Traditional survey data (e.g., trawl, ROV)	Concordance Analysis	Jaccard/Bray-Curtis similarity between eDNA and traditional species lists.	To assess pipeline performance and discovery power in a real-world context.
Replicated Field Samples	Field Samples	N/A(comparison between replicates)	Clustering Stability (e.g., NMI between replicate runs)	Measures consistency of AOTU partitions across technical replicates.	To assess the reproducibility and robustness of the clustering algorithm.