Q&A Bank for Judges

1. How is your AI model different from just clustering sequences by genetic distance?

Our model is fundamentally different. Instead of just calculating a single distance score, our CNN learns the complex, high-dimensional *features* of the sequences from their image-like fingerprints. This allows it to find more robust, biologically relevant groups that simple distance metrics might miss. It's the difference between matching fingerprints by eye and using a machine learning system to analyze every single whorl and ridge.

2. Isn't a 96% reduction in BLAST searches an exaggeration?

Not at all; it's a direct result of the hybrid workflow. For our COI dataset, a traditional pipeline would BLAST all 17,255 unique sequences. Our AI first grouped these into 692 clusters. By only BLASTing one representative from each cluster, we reduce the number of queries from 17,255 to 692, which is a 96% reduction.

3. What happens if a cluster contains multiple species?

This is a great question about resolution. By default, we propagate the annotation of the most abundant ASV. However, our annotated_clusters.csv output contains the mapping for every ASV. A future version of the pipeline could easily calculate within-cluster diversity or flag clusters where multiple high-confidence taxonomies are present, indicating a need for finer-grained analysis.

4. How do you validate that your "Potentially Novel" clusters are actually new?

The "Potentially Novel" flag is the first step, not the last. The definitive validation requires two more steps, which are part of our future roadmap: first, placing the sequence on a global phylogenetic tree using tools like EPA-ng to see if it forms a new branch; and second, synthesizing the DNA for further lab-based analysis. Our pipeline provides the critical first step: a high-confidence list of candidates to focus on.

5. Your pipeline still uses BLAST, so isn't it still database-dependent?

Yes, but we've changed the *nature* of that dependency. We use the database for what it's good at—labeling known organisms. But we don't depend on it for the initial, crucial step of *discovery*. Our AI finds the clusters first, so even if 100% of our data were novel, our pipeline would still produce a structured output of all the distinct biological groups present. A traditional pipeline would simply return "unassigned."

6. Why did you choose HDBSCAN over a more common algorithm like K-Means?

We chose HDBSCAN for two critical reasons. First, K-Means forces you to specify the number of clusters beforehand, but in biology, we never know the true number of species in a sample. HDBSCAN discovers the natural number of clusters from the data. Second, HDBSCAN can identify "noise," which is perfect for eDNA, as it allows us to isolate the truly unique, rare sequences that don't belong to any major group.

7. How does this scale to hundreds or thousands of samples?

The architecture is designed for scalability. The DADA2 and AI training steps can be parallelized across samples or run on more powerful GPUs. The most significant advantage is that the annotation step (BLAST) scales incredibly well, because the number of clusters grows much more slowly than the number of total sequences. This efficiency advantage becomes even greater with larger datasets.

8. Is the FCGR image conversion scientifically validated?

Yes, Frequency Chaos Game Representation is a well-established technique in bioinformatics for visualizing and analyzing genomic data. It has been shown to preserve the underlying statistical properties of DNA sequences, making it a valid input for deep learning models.

9. What's the business model? How would CMLRE use this?

We envision this as an open-source platform that CMLRE can deploy on its own cloud infrastructure, ensuring data privacy and security. A potential business model would involve providing enterprise support, custom feature development, and managed cloud deployments for other research institutions or commercial entities.

10. You skipped the bias correction step. How accurate are your abundance estimates?

That's correct; the ML-based bias correction is a key feature on our roadmap. The current abundance estimates are based on the direct output from DADA2, which are the standard in the field today. While they are subject to known PCR biases, they are sufficient for the community-level analyses shown here. Implementing the XGBoost correction model using a mock community would be the next step to achieve superior quantitative accuracy.

11. Why did you use both R and Python?

We used the best tool for each job. DADA2 is the undisputed gold standard for ASV generation and is written in R. The deep learning and data manipulation steps are best handled by the extensive libraries available in the Python ecosystem, such as PyTorch and Pandas. Our pipeline demonstrates how to effectively integrate these two powerful languages.

12. What was the most challenging technical problem you solved?

The most challenging part was the series of subtle bugs related to integrating multiple tools, each with its own data formats and environment requirements—especially making Python's subprocess call blastn reliably. We solved it by creating a robust script that explicitly manages environment variables and uses absolute paths, which is a lesson in building production-grade bioinformatics software.