Position Title: Volunteer Database / Data Engineering Intern – ETL, PostgreSQL, and Scientific Data Pipelines

Organization: PredictaBio Innovations
Location: Remote
Type: Unpaid Internship (Part-Time, Flexible)
Duration: 2–3 months (with potential for extension)

About PredictaBio

PredictaBio Innovations is an AI-first biotechnology startup building data and intelligence systems for life sciences research. Our platform relies on robust data pipelines that ingest, normalize, and organize scientific information from diverse sources including research papers, biological databases, and experimental metadata.

We are looking for volunteer interns who are excited to work on database design, ETL pipelines, and scalable data infrastructure that powers downstream AI and analytics workflows.

Position Overview

We are seeking highly motivated Volunteer Database / Data Engineering Interns to help build reliable, well-structured data systems for scientific and AI-driven applications. In this role, you will work on database schema design, ingestion workflows, data transformation pipelines, and storage layers for both structured and semi-structured biomedical data.

This opportunity is ideal for students or early-career engineers who want practical experience working with PostgreSQL, ETL pipelines, cloud-native data workflows, and data models supporting AI applications.

Key Responsibilities

• Design, build, and maintain ETL pipelines for ingesting data from scientific databases, APIs, and literature-derived outputs.
• Work with PostgreSQL to model and manage structured scientific and application data.
• Support ingestion and normalization of data from sources such as UniProt, AlphaFold, biomedical literature, and internal processing pipelines.
• Create data validation, transformation, and cleaning workflows for downstream analytics and AI systems.
• Design schemas for metadata, entities, experimental attributes, and document-derived structured outputs.
• Assist in integrating relational and non-relational storage systems where appropriate.
• Write efficient SQL queries for analysis, debugging, and backend support.
• Help monitor data quality, consistency, and reproducibility across pipeline stages.
• Collaborate with AI and backend teams to ensure databases support retrieval, annotation, and application requirements.
• Contribute to documentation, pipeline testing, and weekly team check-ins.

Desired Qualifications

• Strong Python skills for scripting, automation, and data processing.
• Good understanding of SQL and relational database concepts.

• Experience with PostgreSQL or similar relational databases.
• Familiarity with ETL pipeline design and data transformation workflows.
• Comfort working with structured and semi-structured formats such as CSV, JSON, XML, and API responses.
• Exposure to cloud platforms, Docker, or workflow automation tools is a plus.
• Interest in scientific data, life sciences, or biomedical informatics is helpful.
• Strong analytical thinking and attention to data quality.
• Ability to work independently in a fast-paced startup environment.
• Prior project work, GitHub repositories, or relevant coursework is a plus.

Application

If you are interested in this position, please send your resume along with a brief motivational paragraph outlining your interest and relevant experience to **predictabiogeneral@gmail.com**.