Announcing the Open Datasets Initiative
Living datasets backed by automated, open source methods

Machine learning and AI methods show great potential for making new discoveries in the life sciences. For example, the Protein Data Bank (PDB) provided the first open access digital biology dataset, allowing AI-driven tools like AlphaFold and RosettaFold to make huge strides in predicting protein structure. However, this dataset was built up organically over nearly 50 years, and cost an estimated $12 billion to create. Our progress using ML approaches is still currently limited by the lack of appropriate datasets for training algorithms — most existing biology datasets are small, inefficiently curated, and/or privately held. Life science experiments are also notoriously challenging to reproduce: redoing experiments takes human resources and time; is often precluded by funding restrictions; and involves qualitatively dissimilar workflows in different labs. The lack of data availability and the inability to run standardized experiments make it difficult to combine information into large datasets with fidelity suitable for machine learning.

Align to Innovate is looking to change this with the Open Dataset Initiative, a new paradigm for funding, collecting, and sharing large, high-fidelity datasets in biology. We aim to bring together scientists, machine learning specialists, and automation experts in the life sciences to surface and generate important, shared, reproducible datasets using automated laboratories. This strategy has the potential to create datasets faster and more cost effectively than the status quo. Open source automation methods offer a new standard for scientific data collection and sharing because they let us write experimental protocols that are reproducible, scalable, and accessible to scientists worldwide.

Through this initiative, we will connect experimental and computational scientists to identify datasets that are of interest to the machine learning community, that will advance biology research, and that are possible to create with lab automation technology. We will then develop programmable experiments for automation-enabled execution, create the designated standardized datasets, and share them and the methods for creating them with the greater scientific community.
Initiative Overview

Surfacing proposals for target datasets

*What is the next Protein Data Bank?* We will engage with experimental biologists, machine learning researchers, and laboratory automation specialists to identify the next datasets that would be most valuable for life science, as well as the most interesting for training machine learning models. We aim to surface as many ideas as possible and foster discussion between communities during this process through workshops and collaborative conversations.

Creation of automated, open source measurement assays

After deciding upon the datasets and measurement techniques to pursue, we will collaborate with our network of commercial labs and academic partners to develop methods for collecting these datasets in a robust, scalable, and repeatable manner. To ensure open sharing of the scientific process, we will make public the source code for the automated methods.

Open data collection paradigm

With these validated methods, we will ask scientists from around the world to send in relevant samples for standardized analysis. We will return measurements to individual researchers immediately; the same data will also populate a public, shared database after a 1-year default embargo period, which can be extended to 2 years by request. After the embargo period, we will release a public dataset under a permissive license, which will allow it to be reused for commercial purposes.
Initiative Timeline

We will soon (Q2, 2023) open up our initial call for dataset ideas. We encourage people to think broadly, be creative, and not hold back - the more ideas the better. We will monitor entries as they come in, aiming to “matchmake” between groups and disciplines that are surfacing similar intentions. Anyone may provide ideas during this stage, and participants will be able to edit and add to their entries throughout this initial phase. Remember - the goal for this initiative is collaboration, not competition!

We will work with a small group of advisors throughout this phase to surface the most promising datasets to move forward. These advisors will include experimental scientists, machine learning specialists, and automation experts. This will be an iterative process including conversations and workshops with members of the Open Datasets Initiative, Align to Innovate, scientific advisors, and entry participants.

Once datasets and protocols are established, we will start the experiments! At this time (early 2024), we will also call for others to send in samples to be analyzed using these methods.

Questions?

Visit us at: alignbio.org/the-datasets
Contact us at: datasets@alignbio.org