

# Large-scale Biological Data Engineering with HDF5 + the Highly Scalable Data Service

Open Science Research at  
Scale Workshop

April 1st, 2022

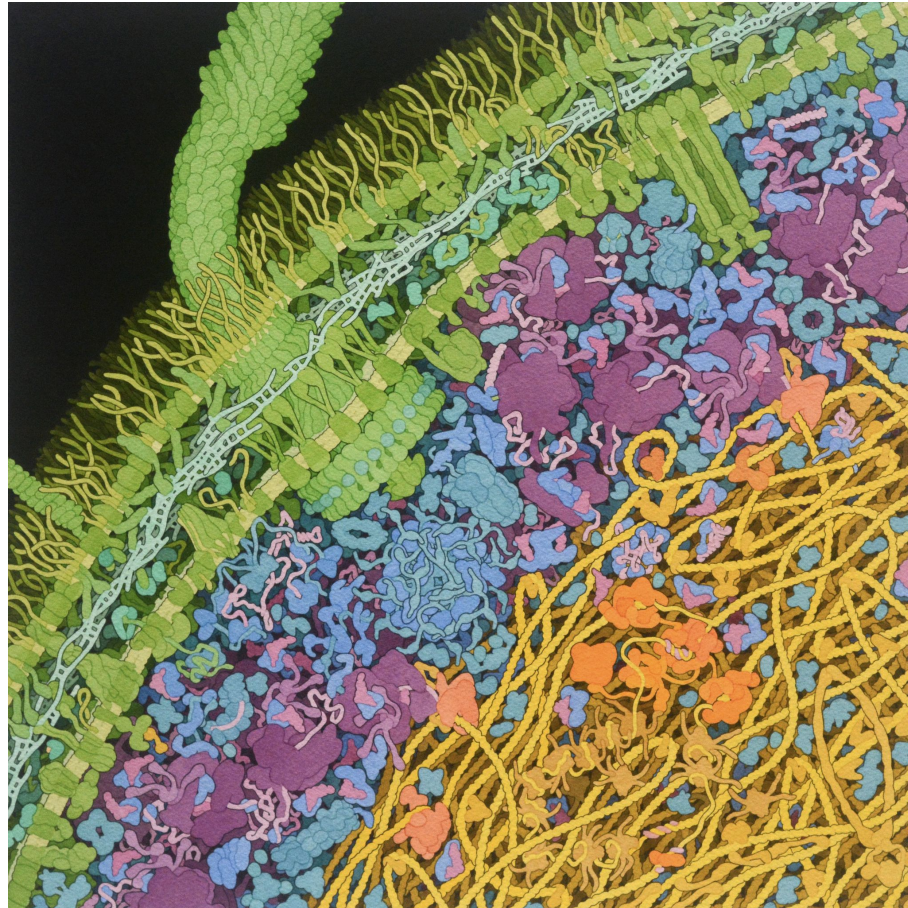
**Eli Draizen**

Phil Bourne's Lab

<http://bournelab.org>

University of Virginia

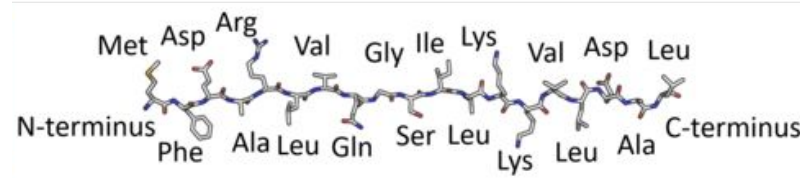
# Motivation



} ~ 200 Å =  
200 x 10<sup>-8</sup> m

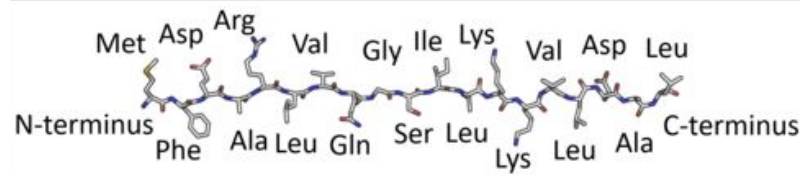
- Proteins mediate many biological functions and are crucial to understanding biological pathways
- Most biological problems concern:
  - Protein interactions
  - Drug targets
  - Interrelationships (evolution)
- Robust data infrastructures are needed to analyze large-scale data about protein structures

# Background: Protein Primary Sequence of Amino Acids



**MLPHFMCPEYCRENLPHFMCPRKV**

# Background: Protein Secondary Structure



**MLPHFMCPEYCRENLPHFMCPRKV**

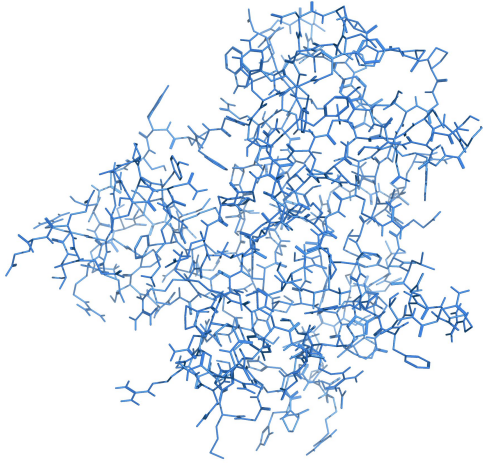


$\beta$ -Sheet (3 strands)

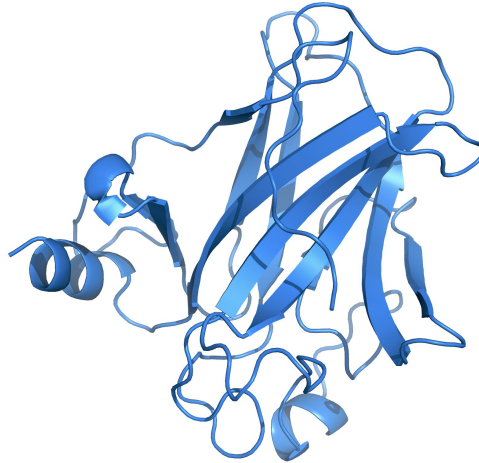


$\alpha$ -helix

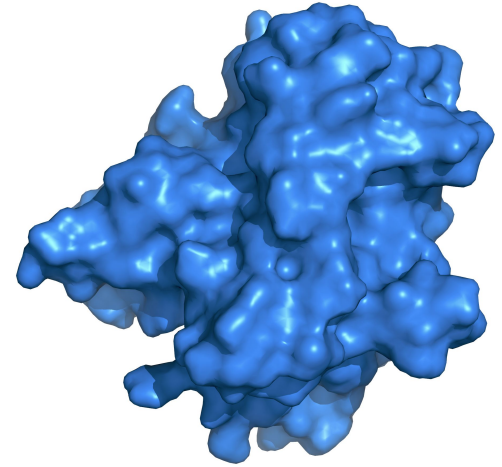
# Background: Protein 3D Structure



**Full Atom  
Structure**

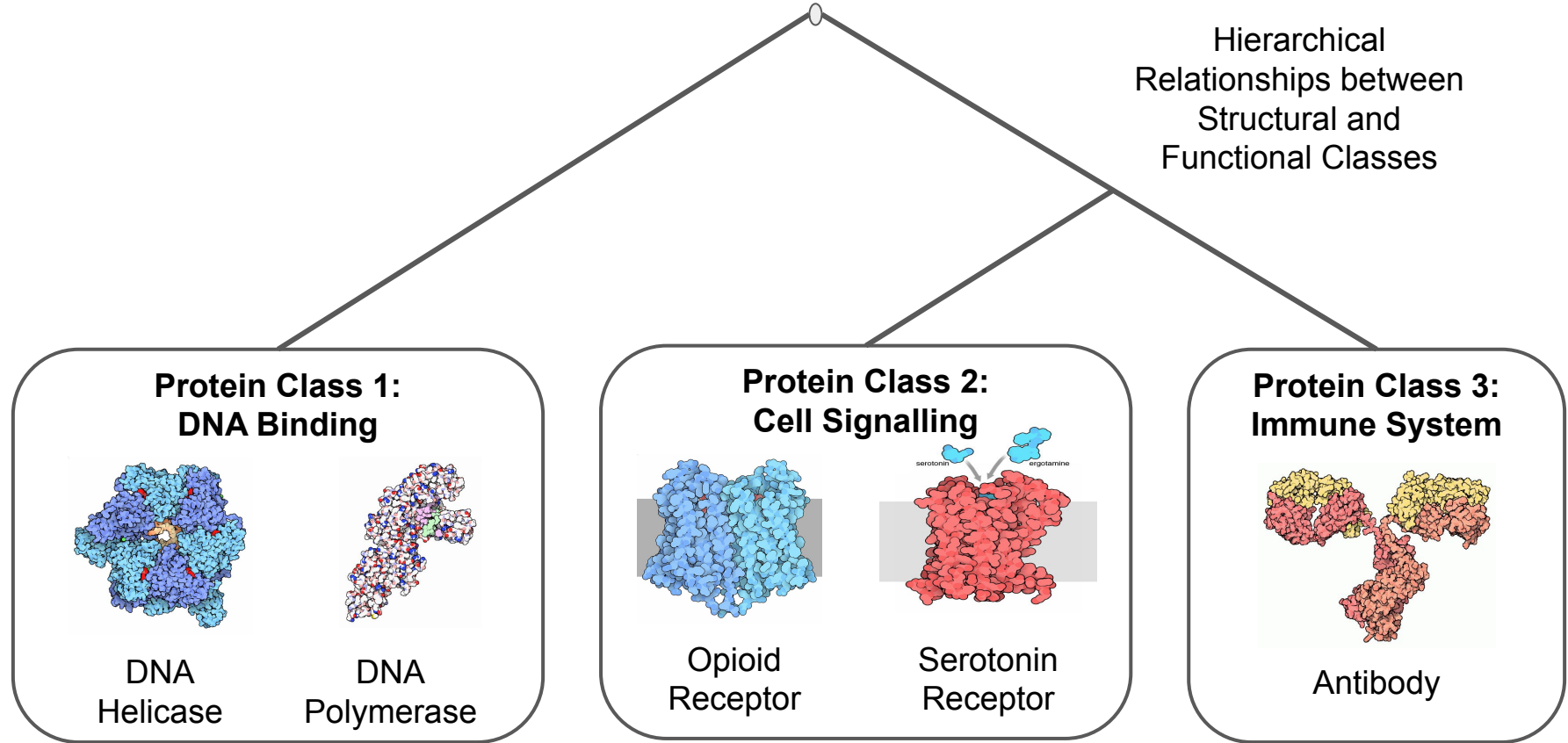


**Cartoon  
Representation**

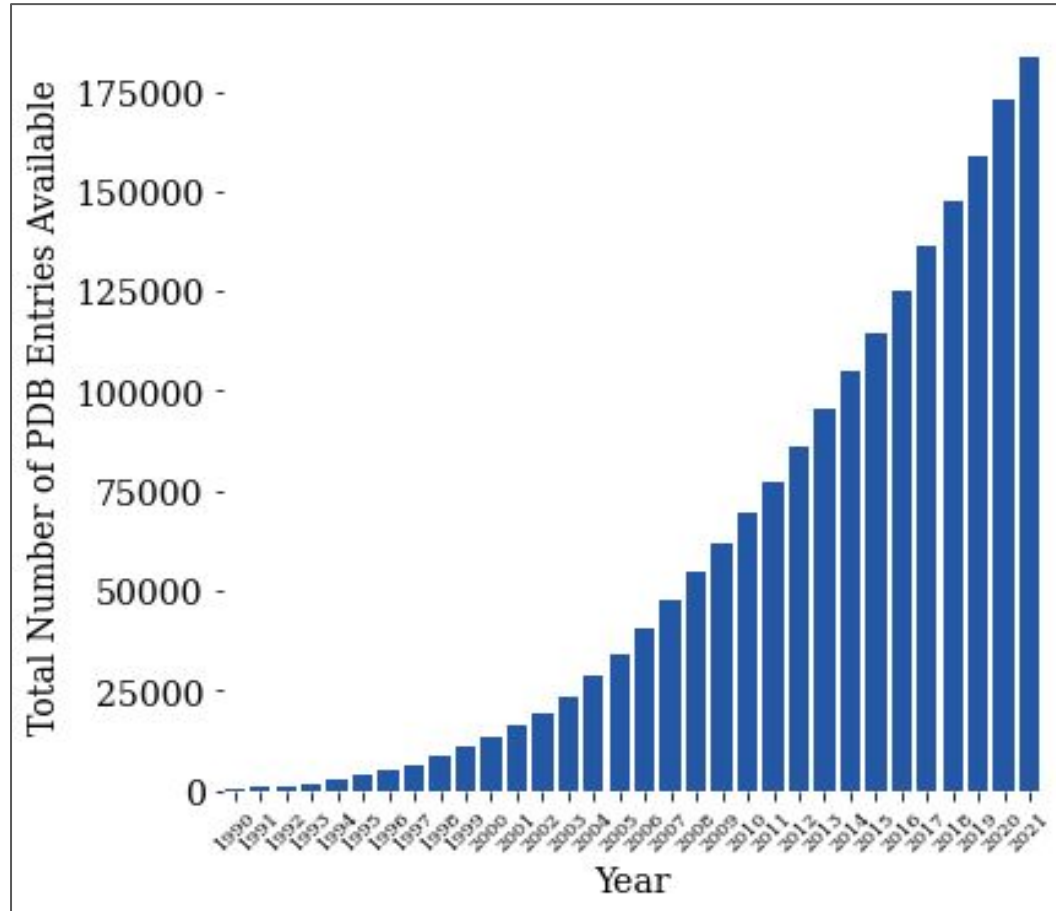


**Surface  
Representation**

# Protein Classification by Structure and Function



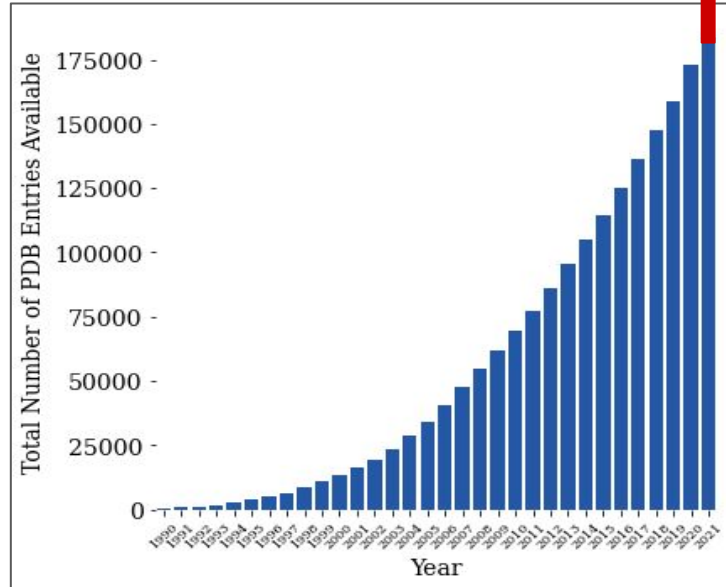
# Robust infrastructure is needed to handle large-scale protein data





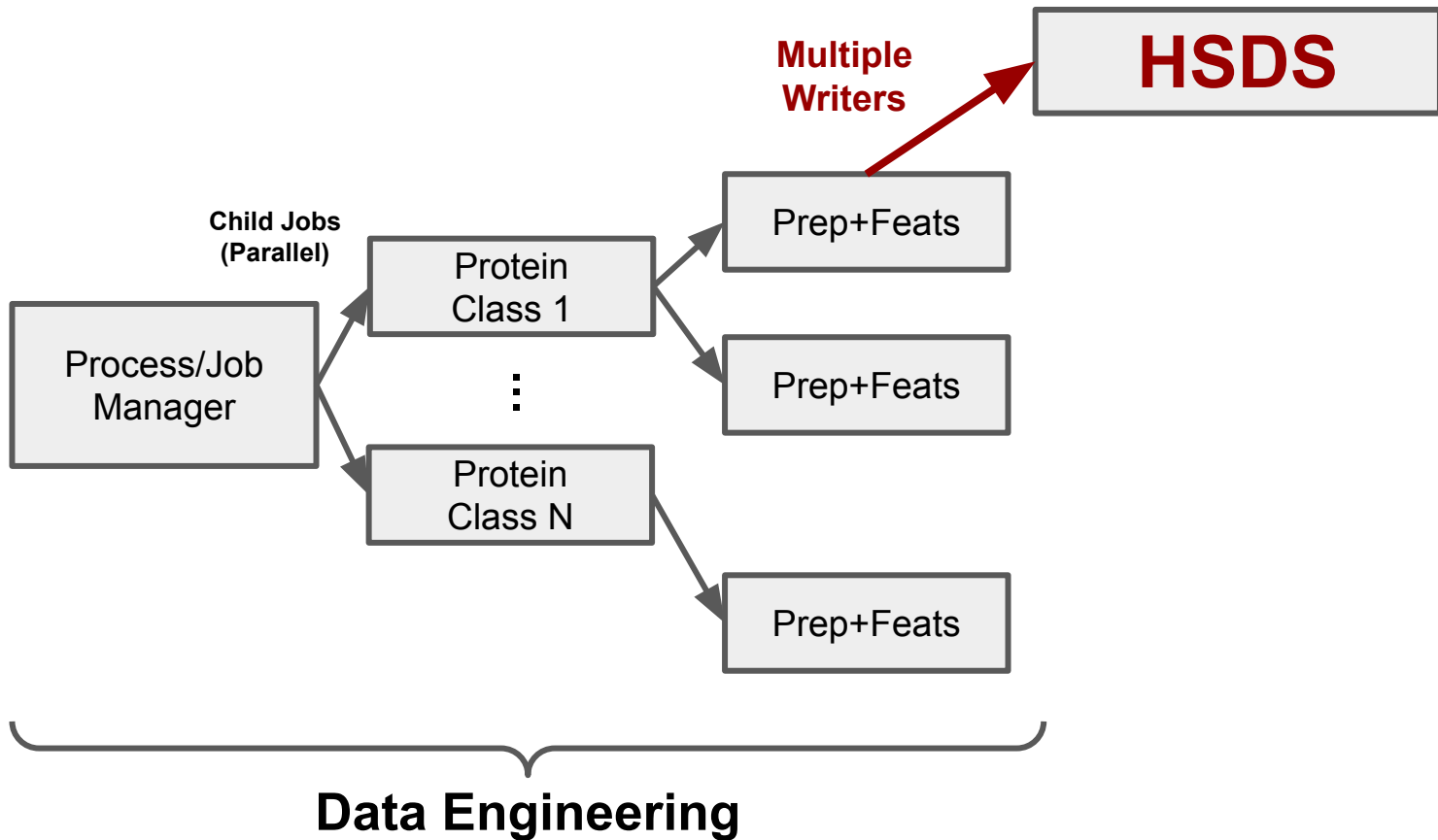
**~50 Million New  
AlphaFold2  
structures!**

Robust infrastructure is needed to handle large-scale protein data

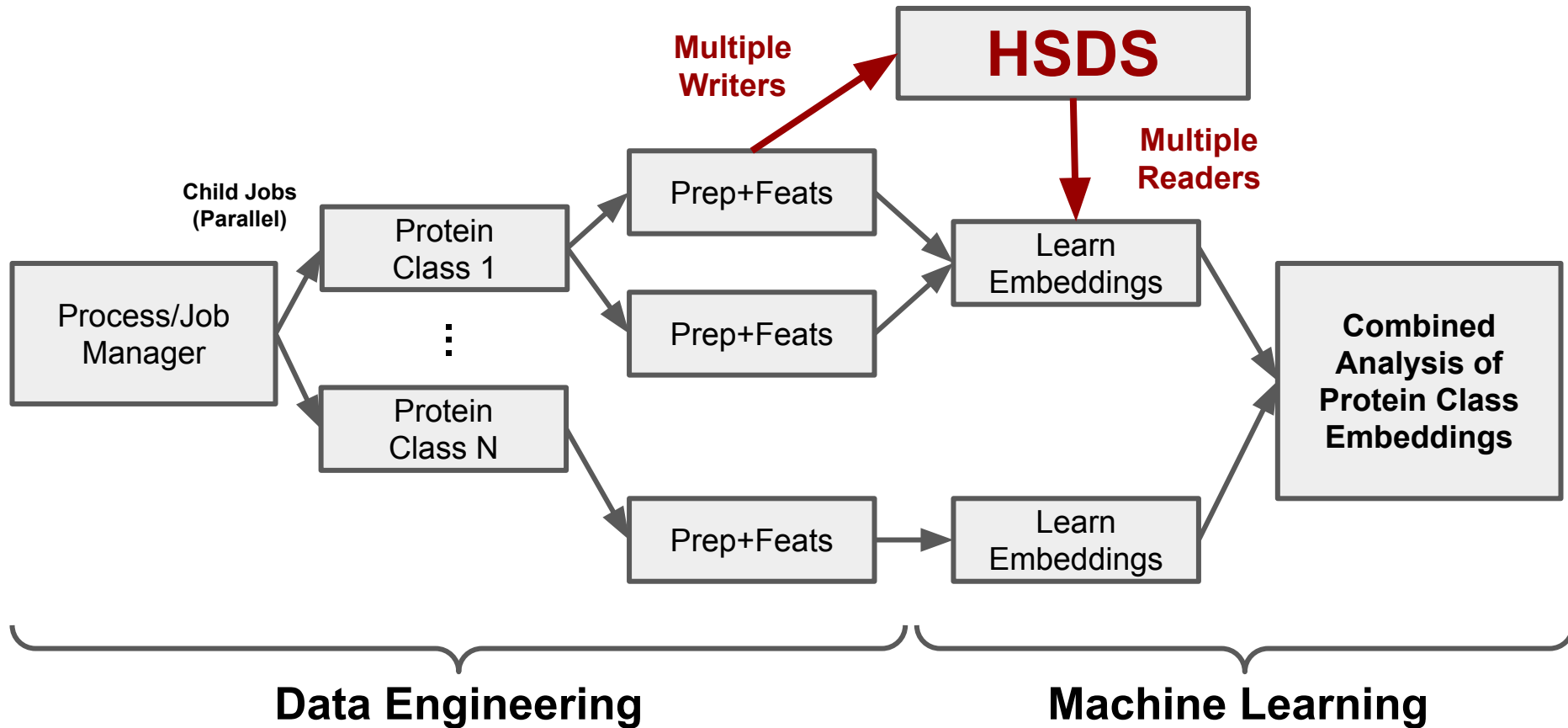




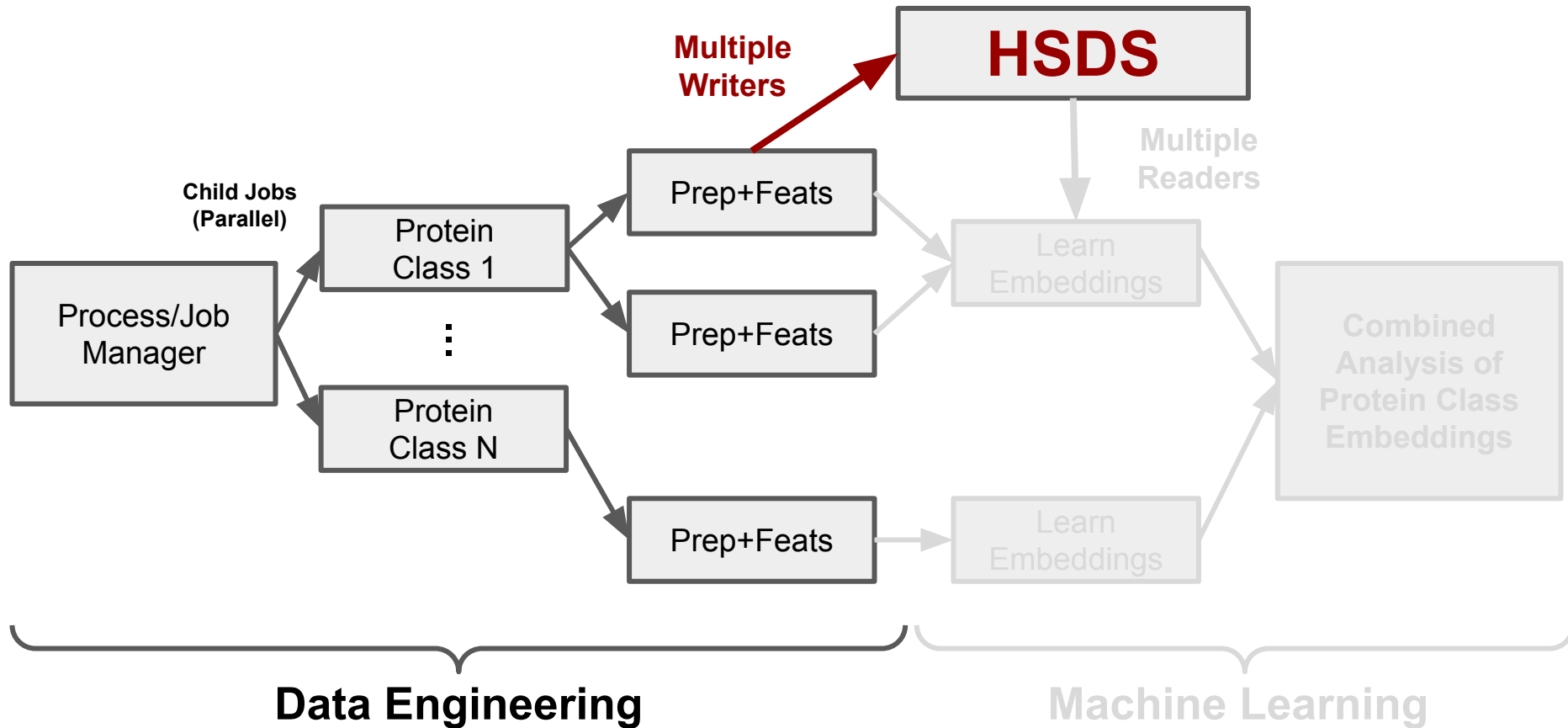
# HSDS Workflow for Biological Data Analysis



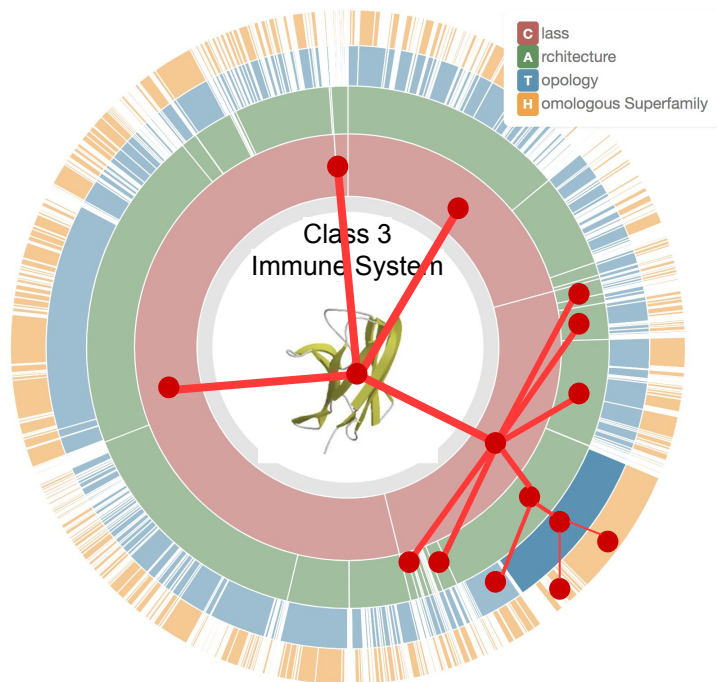
# HSDS Workflow for Biological Data Analysis



# HSDS Workflow for Biological Data Analysis

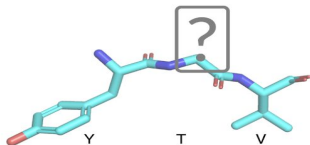


# Massively Parallel Workflows with TOIL are used to process the Protein Class Hierarchy in the Cloud and HPCs

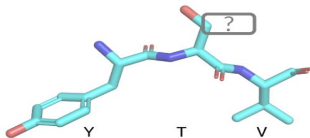


# Data Engineering Stage: Preparation and Feature Calculation

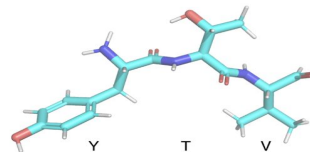
## Step 1: Protein Structure Preparation



1. Add missing amino acids

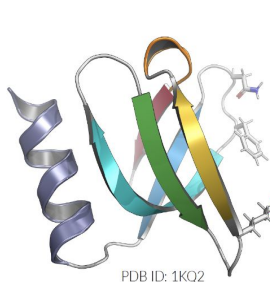


2. Add missing atoms



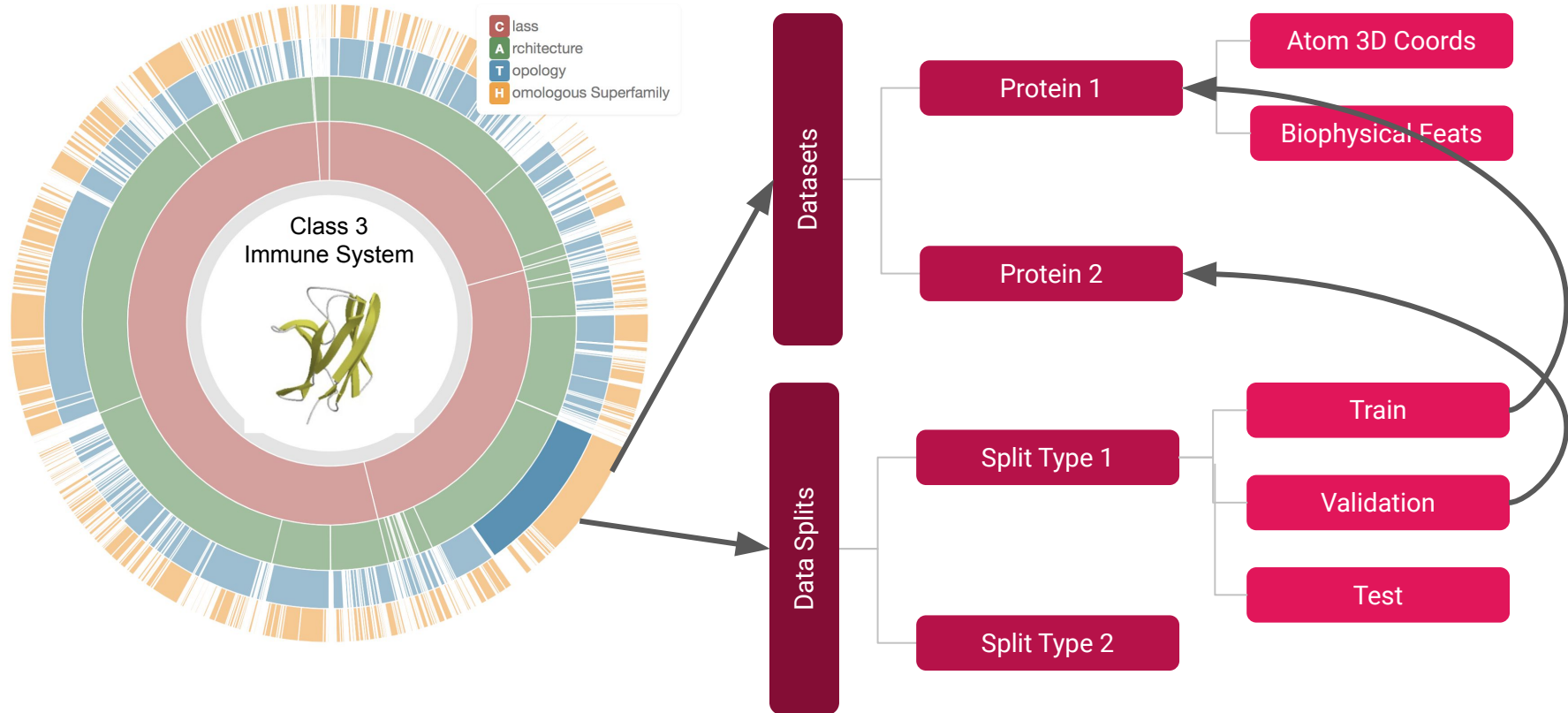
3. Add hydrogens and energy minimize structure

## Step 2: Calculate Biophysical Properties

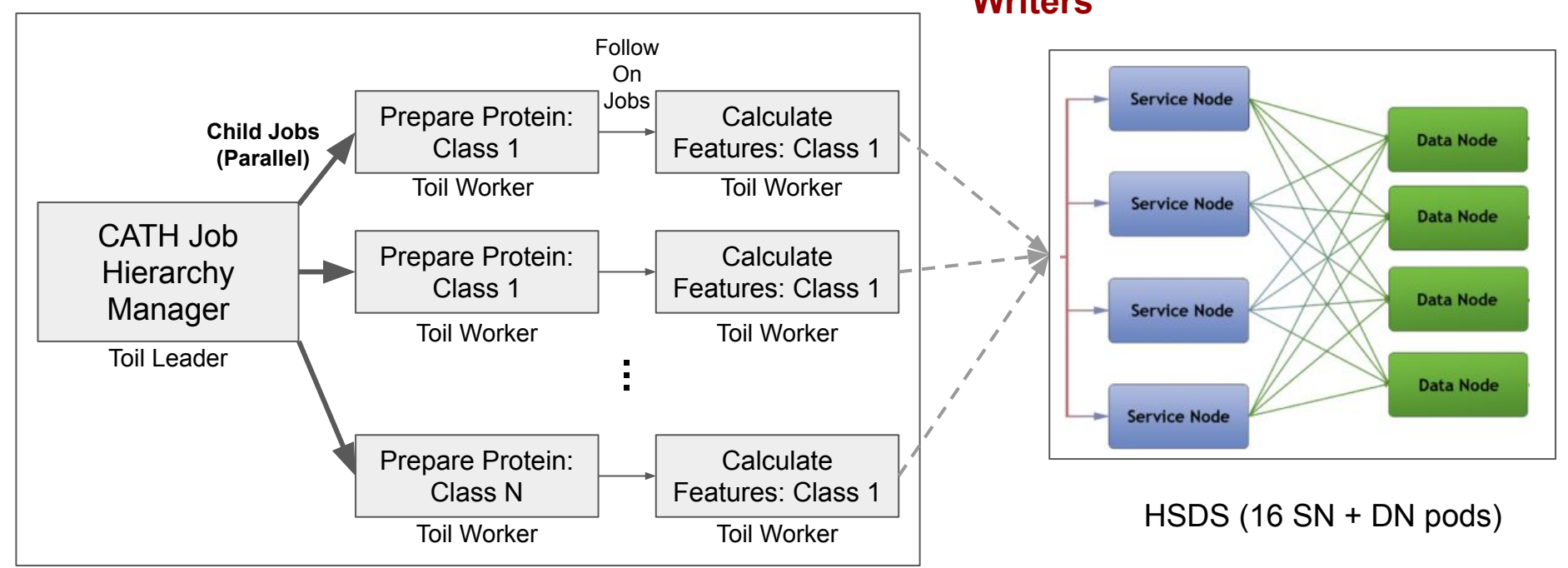


- Atom Type
- Partial Charge + Electrostatics
- Hydrophobicity
- Secondary Structure
- Evolutionary Conservation

# Hierarchical Data Format (HDF) files can chunk and compress the protein class hierarchy in a scalable way



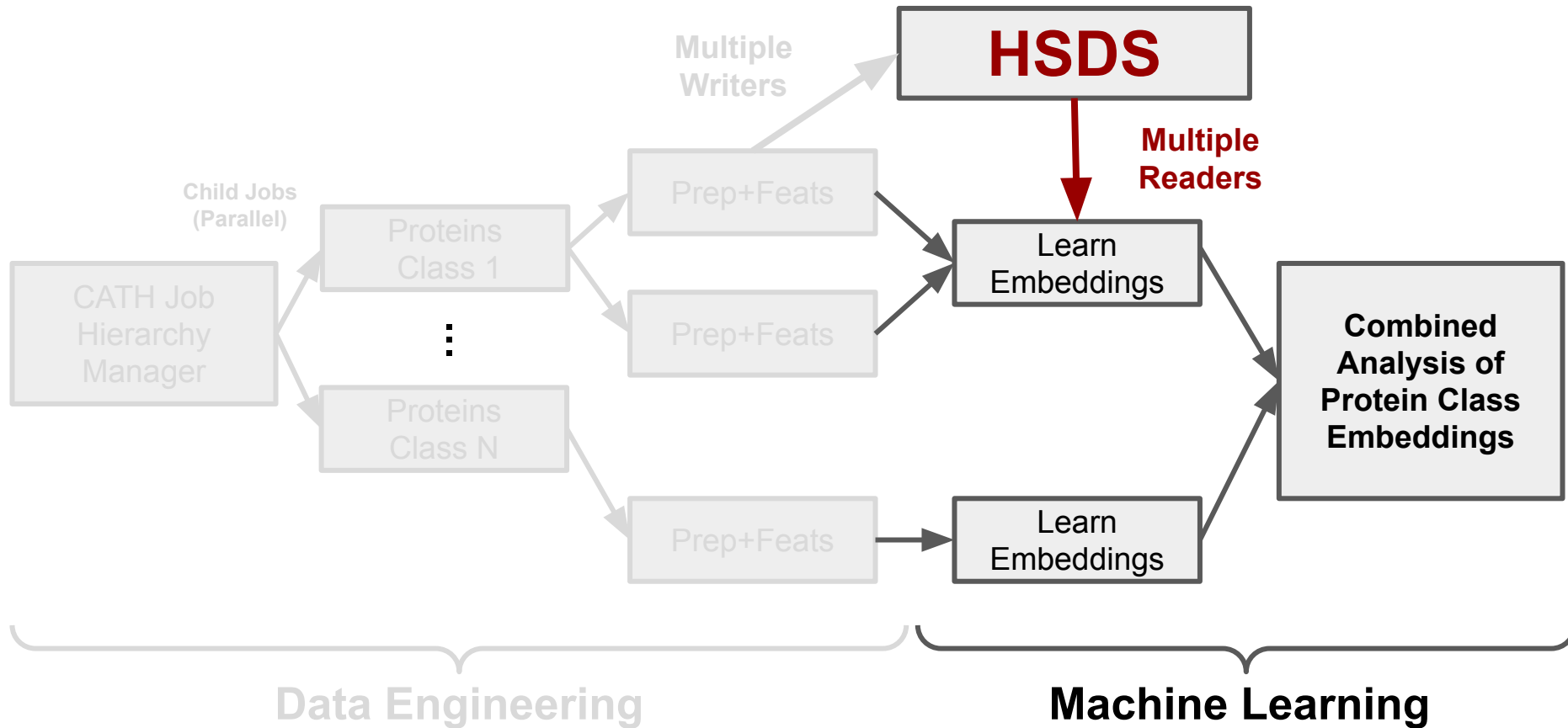
# Data Engineering Stage



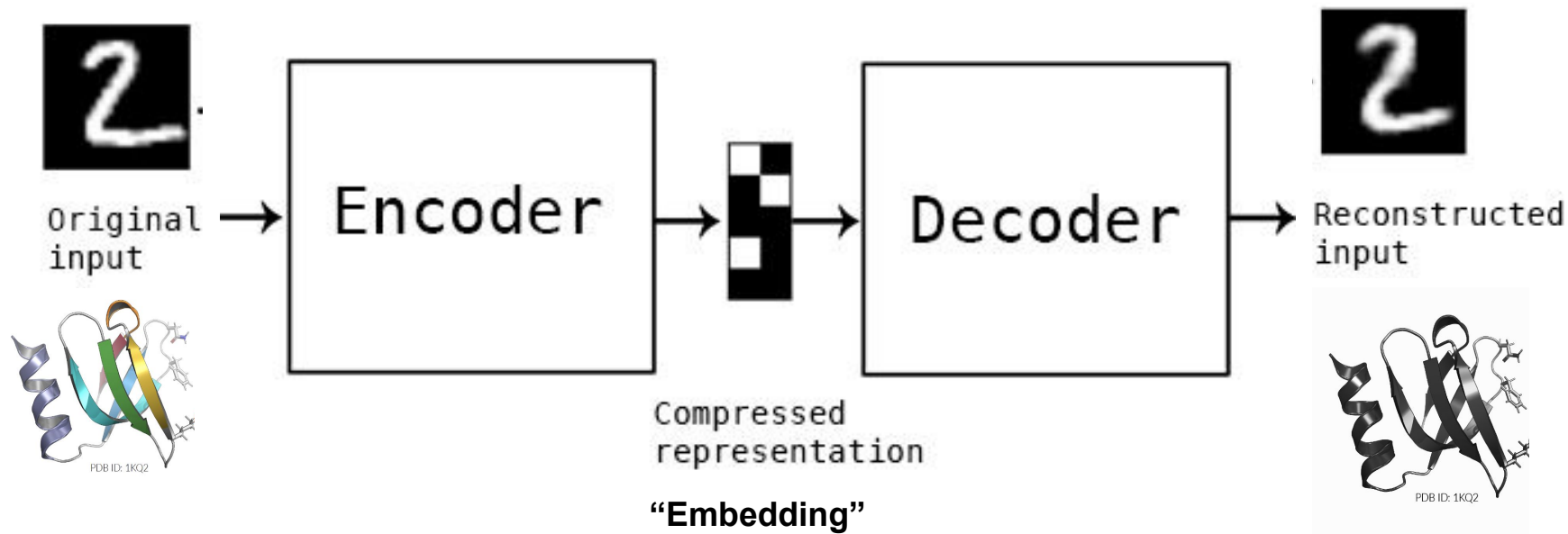
**150,000 domain structures in 20 Classes takes 6 hours, instead of 3+ months**



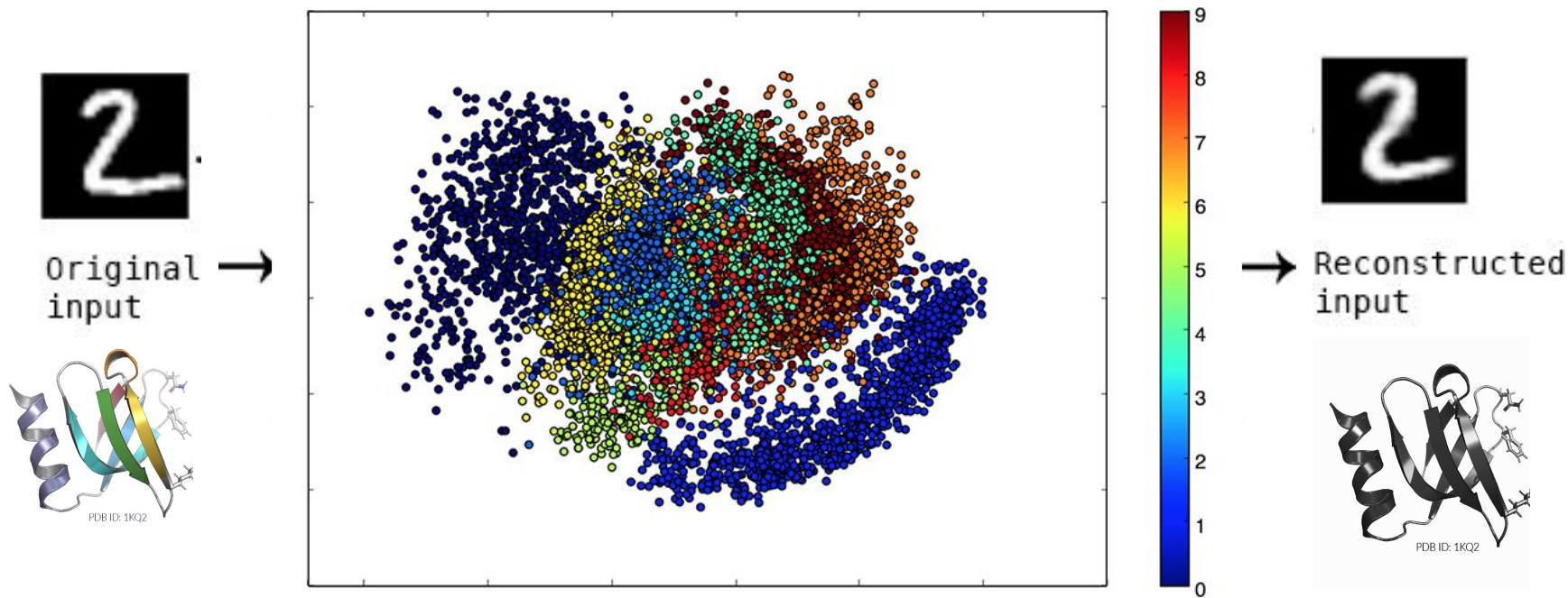
# HSDS Workflow for Biological Data Analysis



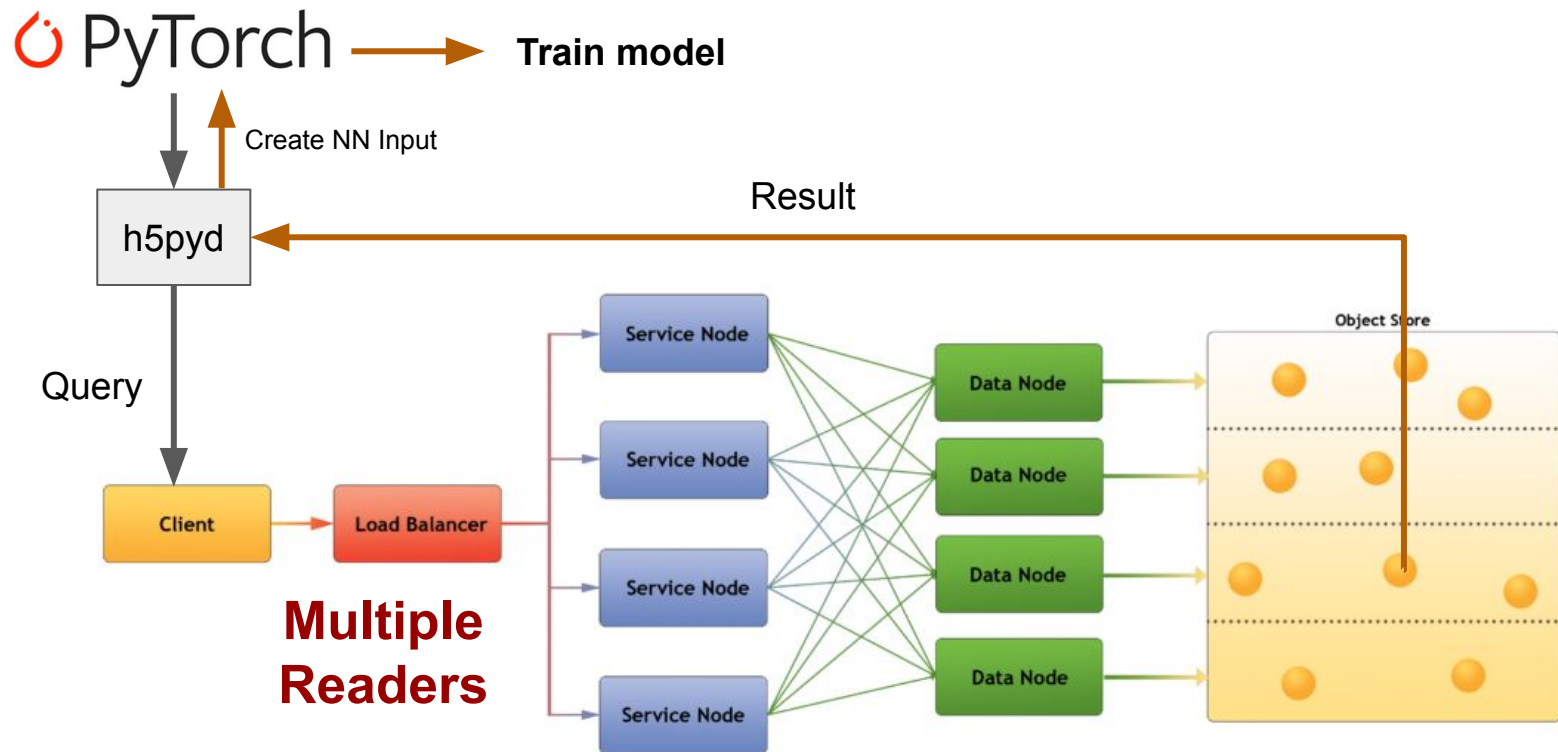
# Machine Learning Model: AutoEncoder / Anomaly Detection



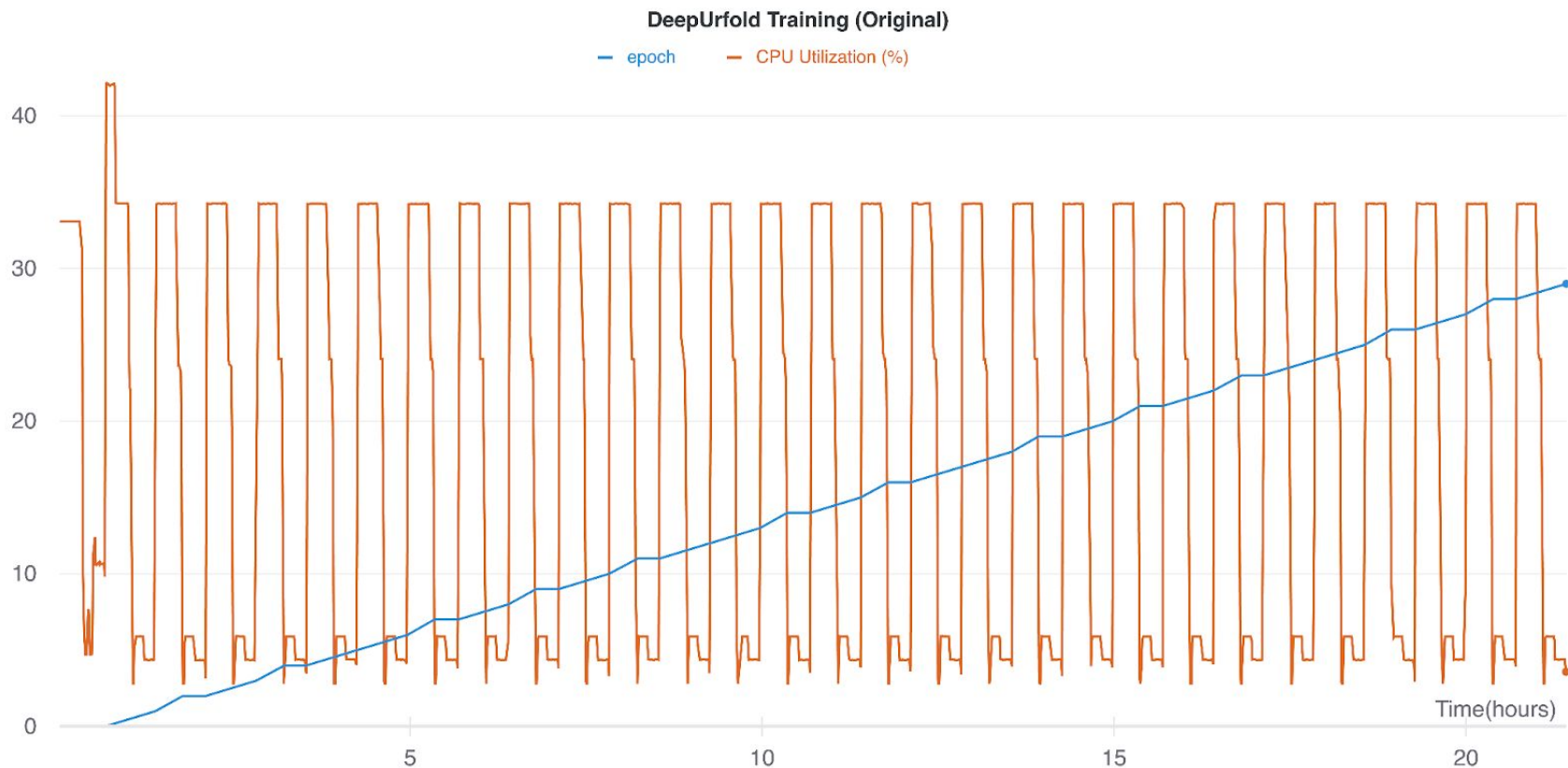
# Machine Learning Model: AutoEncoder / Anomaly Detection



# Create a Highly Scalable Data Service (HSDS) with REST API to access biophysical properties

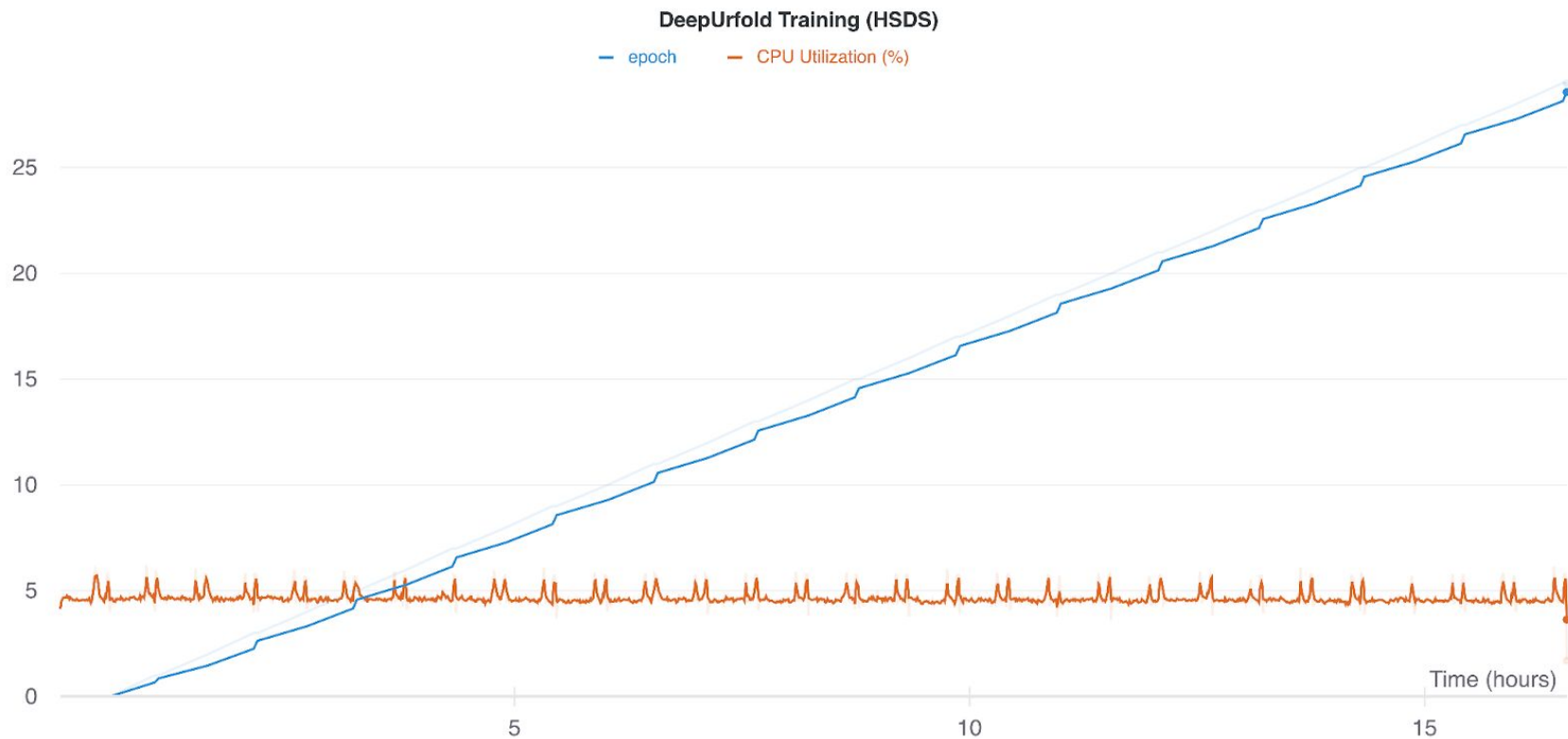


# Before HSDS: **24 Hours** + Constant CPU Use to Read Data

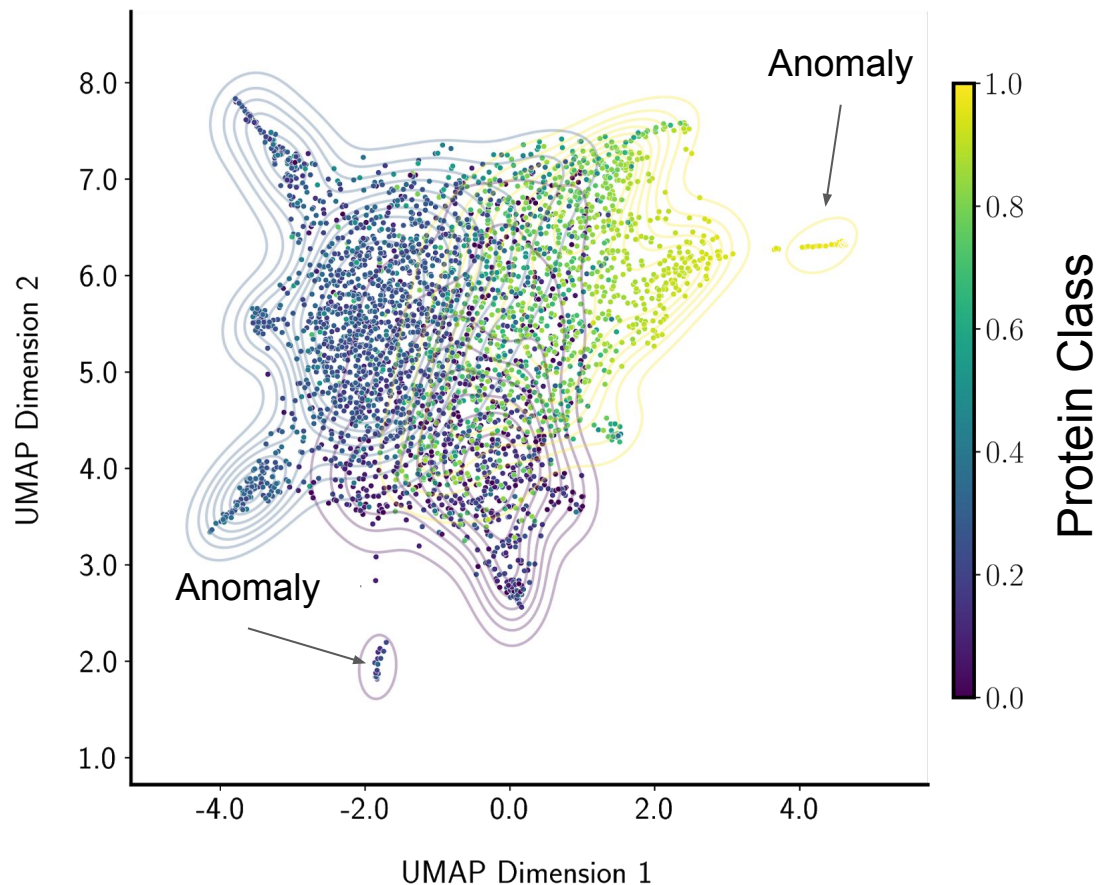


Old Input: Text files of Atomic Coordinates (PDB) + CSV of biophysical properties

# After HSDS: **16 Hours** + Efficient CPU Use to Read Data



# Embeddings Show Similarity b/w Classes + Anomalous Data





# Conclusions

- HSDS handles large-scale biological data **quickly and efficiently**
- Multiple Writers allows for **parallel data generation** to be combined into a single file or dataset
- HSDS + Multiple Readers allows for quick access to the data to speed up training of ML models
  - Data is stored as binary rather than text/CSV files
- **AutoEncoder models can be used to identify anomalies** in any dataset including biological and micro-meteorological data
- **Robust data infrastructures** are needed to accommodate large-scale data analytic workflows

# Acknowledgements

## **Bourne Lab Members**

Phil Bourne

Cam Mura

Lei Xie (Sabbatical Visitor)

Zheng Zhao

Stella Veretnik

Abby Newbury

Skylar Brodowski



Phil Bourne



Cam Mura

## **UVA SDS/Notre Dame**

Filipe Murillo



**UVA DATA SCIENCE**

Funding: Presidential Fellowship in Data Science program

## **HDF Group**

John Readey

<http://bournelab.org>