ZEBLOK COMPUTATIONAL

Life Sciences Use Cases

Ai-MicroCloud™ for Life Sciences

Zeblok's Ai-MicroCloud™ is particularly well suited to the needs of life sciences researchers, pharmaceutical clinical and commercial data scientists, biotech startups focused on new drug discovery, data scientists and/or data analysts in CROs focused on improving outcome from study design on behalf of sponsors and technology solutions providers for pharmaceutical and other life sciences companies.

Zeblok deploys its turnkey, cloud native, quantum-safe secured Ai-MicroCloud™ to wherever your data is secured – on-premises data centers, public clouds and edge locations.

Data scientists and data engineers can get started in minutes, with a simple UI, all familiar open-source frameworks, seamless high-performance computing (HPC) orchestration, a growing library of proven, curated AI algorithms, accelerated data lake and a broad network of AI solutions consulting firms. Zeblok's Portable Ai-MicroCloud™ is the most straightforward way to efficiently pipeline data and then quickly & affordably to develop, train and deploy AI/ML models.

> ***"Zeblok's Portable Ai-MicroCloud™ is extremely easy to use. We realized dramatic improvement in our simulations' performance."***

> ***"Ai-Rover™ is powerful and easy to use. We used it to identify novel endpoints and improve study design."***

Why Zeblok?

Ai-MicroCloud™: Proprietary multi-class, multi-cloud orchestration engine for AI workloads, including one-click HPC scalability

Enterprise Grade: Quantum-safe security

Ease of use: Simple UI, leveraging open-source data science tools, enables data scientists to be productive in minutes

Intelligence Marketplace: Proven, original algorithms, ready for model integration

Accelerated Data Lake: 10-15x faster pipelines and faster integration of disparate datasets

Ai-Rover™: Domain-agnostic data discovery tool for large, multi-variate, high dimensional data analysis and data visualization, with patent-pending explainable AI, exclusive to Zeblok

Trusted By

LAUFER CENTER FOR PHYSICAL AND QUANTITATIVE BIOLOGY

milestone group

hosta labs

NVIDIA

IBM

PTC THERAPEUTICS

Athena

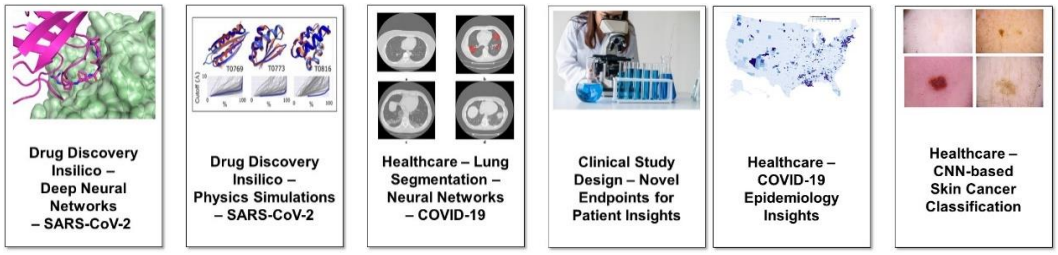
Q→NU

NODEWEAVER Reliable continuity for Your IT

RURAL CLOUD INITIATIVE DIGITAL TRANSFORMATION OF RURAL AMERICA

alef

Use Cases
Ai-MicroCloud™ for Life Sciences



Use Case:
In Silico Drug Discovery: SARS-CoV-2

Project Statement

Understanding the biology of a virus is impossible without uncovering the structure of the viral proteins. Unfortunately, experimental methods for resolving protein structure are complex and time-consuming. Thus, predicting the properties of proteins based on their sequences, which are much easier to determine experimentally, plays an important role in structural biology. Structural biology has greatly benefitted from recent advances in deep learning, including its application to structure prediction.

Simplistically, protein can be thought of as a long string — a linear chain of amino acid residues, the particular sequence of which is determined by organism's DNA (or, for some viruses, including SARS-CoV-2, the organism's RNA). While the sequence, in the vast majority of the cases, fully determines the protein, it is next to impossible to study the protein's behavior without knowing how this chain arranges itself in 3-D space. Though methods for predicting local features of a protein ("secondary structure") have existed for many years, it remains challenging to reliably infer how these local elements are positioned relative to each other. As the local structure is rigid and relatively easy to predict, being able to determine even a few long-range contacts — pairs of residues that are far from each other in the sequence, but close by in the 3-D structure — can help tremendously in determining the overall structure of the protein, or even the relative orientation of two proteins when they form a stable complex.

One way of determining such contacts is by looking at mutations in similar proteins in other organisms and changes in types of residues. When two residues are close in 3-D space, they interact. And if one of them changes, often its neighbor must also change, if protein stability is to be preserved. Such co-evolving residue pairs have been shown to be a great predictor of both intra-protein and inter-protein contacts.

Deep learning techniques are extremely helpful for predicting such long-range contacts, as evidenced by the recent rounds of CASP (Critical Assessment of Structure Prediction — a community-wide competition to compare the performance of different folding methods on previously unpublished proteins).

Building upon these results, running our simulations on the Zeblok platform, our group is working on using co-evolutionary data and deep learning methodology for the prediction of structures of individual proteins and their complexes.

In Silico Drug Discovery: SARS-CoV-2 (continued)

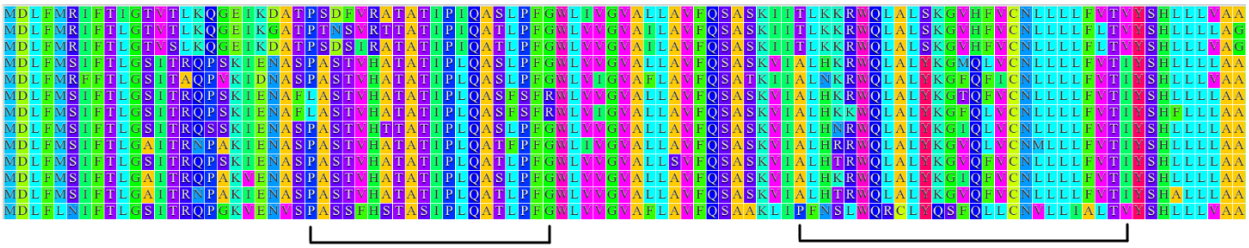


Figure 1. A part of multiple sequence alignment of the orf3a protein from SARS-CoV-2 and homologous proteins from other viruses. Analyzing which residues tend to change together can shed light on their interactions. Black lines at the bottom demonstrate two such pairs (for illustrative purposes only).

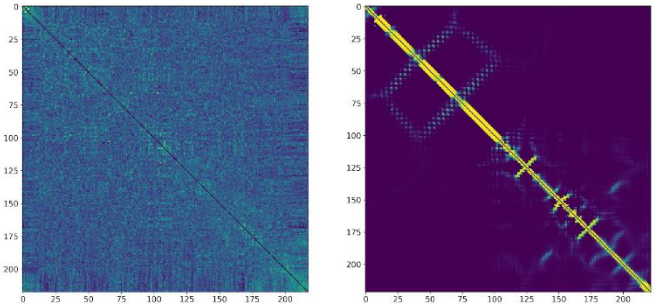


Figure 2. The strength of the co-evolutionary signal (left) and predicted probability of contact (right) for the membrane (M) protein of SARS-CoV-2.

Zeblok Ai-MicroCloud™ Resources Used

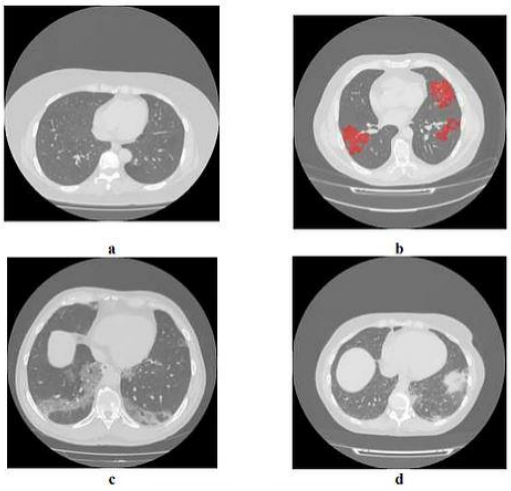
- Zeblok Ai-Rover™ WorkStation
- Multiple containers to support multi-GPU, multi-CPU compute engines
- 128 RTX6000s GPUs
- 640 vCPU
- 3,028 GB RAM
- 800 GB Block Store
- 500 GB of Parallel File System

Use Case:
CNN-based Lung Segmentation Algorithm – COVID-19

Project Statement

The COVID-19 virus, also known as SARS-CoV-2, has severely impacted people's lives all over the world and continuous research effort is being carried out in improving methodologies related to the detection of COVID-19. While treatment of patients, COVID-19 infected regions need to be identified and segmented. An example of COVID-19 infected lung CT slice is shown in Figure 1.

To automate such a task an AI-powered segmentation model is trained using a large amount of unlabeled data available. In this work, we utilized 800 CT volumes of COVID-19 infected patients which results in around 25k CT slices. Training of AI models on such large amounts of data requires parallel computing hardware such as GPUs. Zeblok's platform provides a seamless way to train such models.



The result of this work improves upon the segmentation performance of previous models by around 2-4 dice percentage points and achieves state-of-the-art COVID-19 segmentation results. The automation of such time-consuming segmentation tasks can reduce the workload of clinicians, who are already working under tremendous pressure.

Data Used

For this work, we used the publicly available largest dataset of infected lung CT scans called MOSMEDDATA dataset. It comprises around 1,100 CT volumes with 800 infected and 300 non-infected CT volumes. Using CNN based lung segmentation algorithm we first filtered out the scans containing only lung region. This filtering process results in around 25,000 CT slices from COVID infected lung volumes. Using these volumes, we trained a COVID lesion segmentation model.

Approach

We proposed a novel segmentation approach: using a convolutional neural network (CNN) derived from U-Net for this purpose and we followed a semi-supervised training strategy to train the model. We first trained the CNN using a small amount of labeled data. We then used the trained CNN to get the pseudo masks for the 25,000 CT slices. In the second step, we again trained the proposed CNN from scratch using the pseudo masks of the 25,000 CT slices. Training with such a large amount of data improves the segmentation performance of the model.

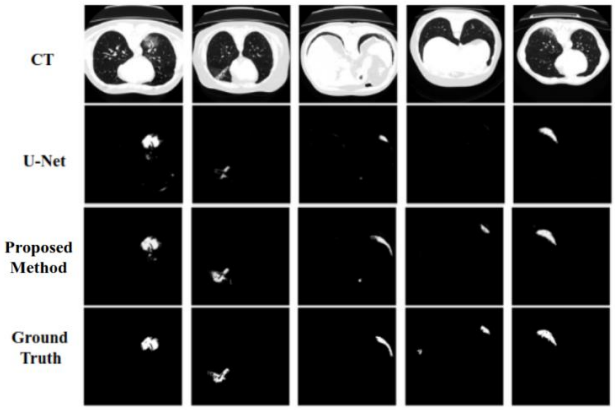
CNN-based Lung Segmentation Algorithm – COVID-19 (continued)

Result

We observed that by using the large amount of unlabeled data the segmentation performance improved by about 2 to 4 dice points.

By following the proposed training approach, our model could reach a segmentation dice score of 0.66 while the current state-of-the-art COVID lesion segmentation model could only obtain a dice score of 0.61.

This is a significant improvement and shows the utility of our semi-supervised training strategy. A qualitative comparison of the output of the proposed method with U-Net is shown in Figure 2.



Zeblok Ai-MicroCloud™ Resources Used

- Zeblok Ai-Rover™ WorkStation
- Multiple containers to support multi-GPU, multi-CPU compute engines
- 128 RTX6000s GPUs
- 640 vCPU
- 3,028 GB RAM
- 800 GB Block Store
- 500 GB of Parallel File System

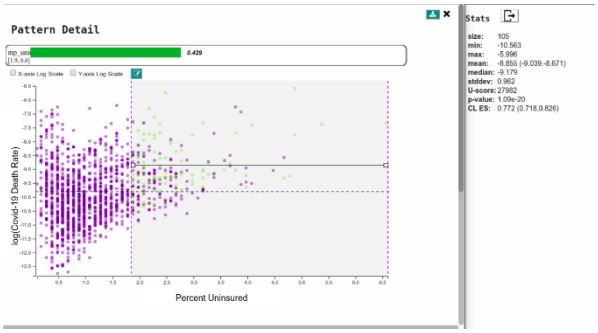
Use Case:
Turnkey AI-WorkStation for COVID-19 Epidemiology

Project Overview

COVID-19 is a disease that has affected everyone globally, with the largest number of cases in the United States. The mortality rate is high in African American and socioeconomically weak communities. We identified key datasets that contain county-level attributes related to economic status, ethnicity, COVID-19 cases, hospitalization rates, mortality rates, etc. and performed visual analytics to obtain more insights, using an explainable AI-analysis tool, developed with our partner, Akai Kaeru, which Zeblok Computational provides, bundled with dataset as a pre-configured AI-WorkStation. We feel that further analysis with other relevant datasets could identify hotspots and clusters, and also assist in emergency preparedness for future disease outbreaks.

Data Used

Fig1. Death rate vs percentage uninsured.



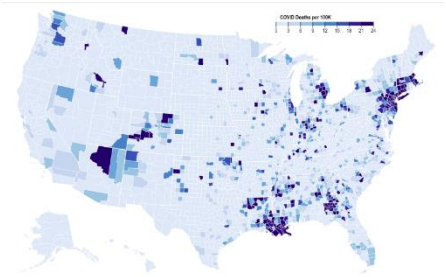
We analyzed several datasets from various publicly available sources, and we identified that the Kaggle repository titled UNCOVER COVID-19 Challenge provided by Roche Data Science Coalition had the relevant data. We used most of the attributes and used various machine learning techniques to obtain some interesting and valuable understanding of the data.

Approach/Tools Used

The goal of this analysis was to identify criteria that put U.S. counties at risk for a higher death rate from COVID-19. We measured the death rate as the number of confirmed COVID-19 deaths per 100,000 people. We used Zeblok's COVID-19 Epidemiology Notebook to identify and explain why some counties have a higher death rate than others. This technology uses a combination of statistical analyses and visual analytics to allow users to identify subgroups of counties with statistically higher or lower COVID-19 death rates.

Results

Figure 2: Multiple maps with insight to data



The analysis resulted in over 100 patterns that succinctly explain differences in county level death rates due to COVID-19. Many of the patterns that describe unusually high death rates are based on socioeconomic factors that also correlate with minority status (i.e. counties with higher poverty rates and higher concentrations of minorities tend to have higher death rates).

Turnkey AI-WorkStation for COVID-19 Epidemiology (continued)

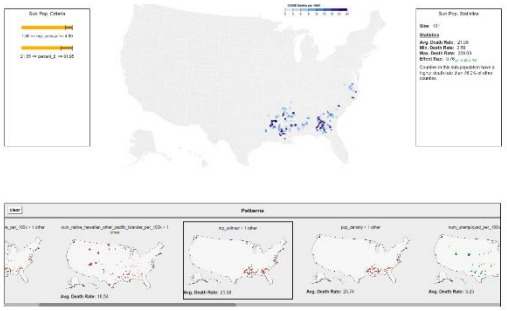


Figure 5. Counties with High Percent Black and Uninsured

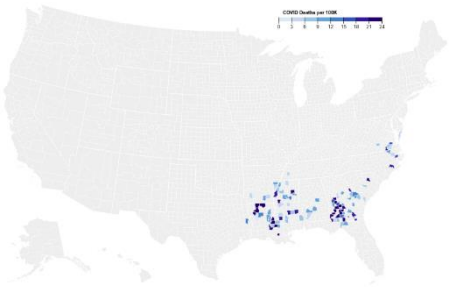


Figure 4. Covid-19 Death rate per US county

For example, counties that have a high percentage of African-American residents and a high percentage of uninsured have about a death rate that is greater than twice the national average i.e. 21 deaths per 100,000 people). These counties are primarily concentrated in the south.

Zeblok Ai-MicroCloud™ Resources Used

- Zeblok Ai-Rover™ WorkStation
- Multiple containers to support multi-GPU, multi-CPU compute engines
- 1 RTX6000 GPU
- 7 vCPU
- 16 GB RAM
- 50 GB Block Store
- 100 GB of Object Store

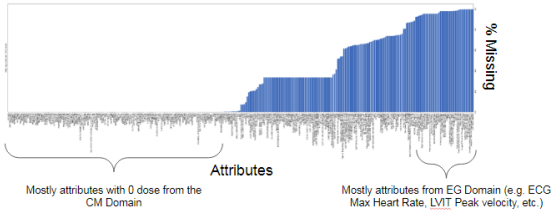
Use Case:
Study Design for Developing Therapies for Rare Disease

Project Overview

Zeblok conducted an analysis, to aid in study design with the goal of efficiently developing safe and effective therapies to treat Friedreich's Ataxia or spinocerebellar degeneration, a rare genetic disease that causes difficulty walking, a loss of sensation in the arms and legs, and impaired speech. The goal was to use data from Friedreich's Ataxia Clinical Outcome Measure Study (FA-COMS) natural history data to identify novel endpoints, biomarkers, and baseline characteristics that may optimize clinical study design. By utilizing machine learning and artificial intelligence-based algorithms, the Zeblok team was more efficiently able to identify correlations/inter dependencies between a large number of variables in the dataset. This approach helped uncover interesting correlations to the onset and progression of the disease. Target attributes such as genetic data, baseline characteristics, and the relationship of cardiac decline to neurological symptoms were of prime importance. The study focused on natural history data, rather than placebo and treatment arm data.

Challenges

Image: Data cleaning



Drug trials are a very time consuming and expensive process. For a disease that progresses over many years or decades, like Friedreich's Ataxia, the success of a drug trial depends in large part on how quickly a subject progresses. If the subjects in both the treatment and control group do not progress very much over a 5 year drug trial, then the efficacy of the drug cannot be determined and another 5 year study has to be performed.

Conversely, any effect a drug has will be much more substantial in subjects whose disease has a much faster progression rate. Therefore, there is a tremendous benefit in identifying subjects that are likely to progress more quickly. The challenge is to use explainable AI to identify subjects that will have a faster disease progression rate based on their natural history data.

About the PoC

Image: Univariate Analysis - Heatmap

We obtained anonymized patient data from the Center for Policy Analysis on Trade and Health (CPATH). The analysis was focused on the FA-COMS dataset. This study contains 1,050 patients with yearly follow-ups over a thirteen-year period. The target variable of interest is the 1-year and 2-year change in MFARS (modified Friedreich's Ataxia Scale, which aggregates a number of tests and ranges from 0, i.e. no FA, to 93, i.e. advanced stage of FA).

Study Design for Developing Therapies for Rare Disease (continued)

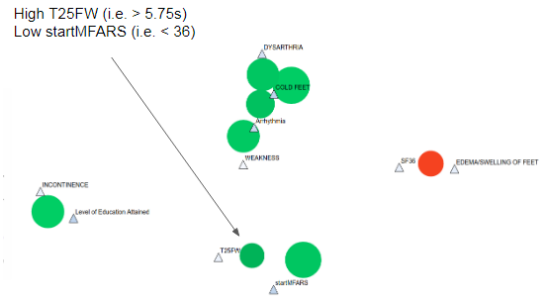
On average, someone with FA will typically progress at a rate of 2 points per year. Our goal was to identify indicators that identify subjects that progress at a significantly faster rate than 2 points per year.

Our approach was to:

1. Use the pattern mining module to identify subpopulations with an unusually high change in MFARS.
2. Use the causal analysis module to remove spurious correlations.
3. Identify subpopulations in which there is a high correlation between the change in MFARS and some independent variable.

Zeblok Advantage

Image: Correlation Mining



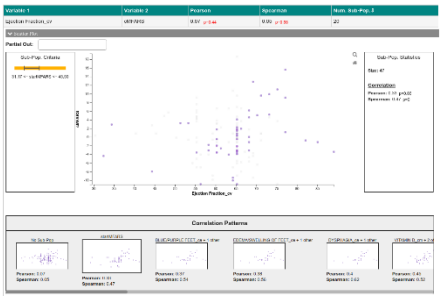
Zeblok has developed a bioinformatics cloud platform for the development of new digital biomarkers, AI-based predictive capabilities for identification of novel endpoints, and real-world data collection for creating new digital health insights. Zeblok works with business partners and academic researchers to develop and deliver best in class tools using Artificial Intelligence. Zeblok intends to provide such software to data scientists to improve efficiency of the drug discovery process. While there is a wide range of machine learning methods, such as decision trees, random forests, neural networks, and deep learning, that can help with prediction tasks, they are primarily black boxes, lacking accountability and trustworthiness. Our approach explains conditions that yield lower or higher values in any target variables and defines causal relationships among different salient data regions. This approach is an emerging field in machine learning, called Explainable AI or XAI. We use techniques of dimensionality reduction through subspace clustering of relevant attribute sets and causality prediction using XAI.

The study design team used Zeblok's AI-Rover™ WorkStation (a pre-configured Jupyter notebook, pre-loaded with the Explainable AI algorithm, developed in collaboration with Akai Kaeru, a leader in visual analytics for complex data, to interact with the FA-COMs data.

Study Design for Developing Therapies for Rare Disease (continued)

Results

Image: Sample Correlation Table



The analysis resulted in the identification of several subpopulations that have a higher rate of progression. One such subpopulation was defined by subjects who have not progressed very far but whose time for the 25 foot walk test was high. These subjects were more likely to progress at a higher rate. This is a criterion that can now be used in the selection of subjects for an upcoming drug trial.

Zeblok Ai-MicroCloud™ Resources Used

- Zeblok Ai-Rover™ WorkStation
- Multiple containers to support multi-GPU, multi-CPU compute engines
- 1 RTX6000 GPU
- 7 vCPU
- 16 GB RAM
- 50 GB Block Store
- 100 GB of Object Store

Platform Features Overview

Ai-MicroCloud™, including **Turnkey HPC Orchestration** and an **Intelligence Marketplace** for curated algorithms

- **Ai-WorkStation:** Customized and virtualized Jupyter Notebook, with access to all familiar open-source frameworks, accelerated data lake and AI algorithms via a simple web interface
- **Ai-HPC-WorkStation:** simple workload distribution to hundreds of GPUs for AI/ML model development, training and simulations
- **Accelerated Data Lake:** Enables a 10-15x reduction in search time
- **Intelligence Marketplace:** Growing library of carefully curated original AI algorithms, including exclusively in-licensed patent-pending software

  Easy to read, easy to use and easy to share

  We fast-track adoption of the best AI algorithms from academia and AI startups
- **Cloud Native:** Scalable architecture running in modern, dynamic environments using containers and declarative APIs
- **Ai-Rover™:** Analytics and data visualization notebook – domain-agnostic data discovery tool for large multi-variate high dimensional data analysis, using patent-pending explainable AI algorithm, exclusive to Zeblok

  Provides crucial data comprehension step as starting point for AI model development – patterns, correlations and causation
- **Quantum-Safe Entropy-as-a-Service:** Truly random numbers, generated by single photon detection (SPD) technology, delivered via container for integration within existing encryption key management
- **Runtime Environment:** Finished model pipeline is easily promoted to a runtime API, including inferences running at the Edge
- **Multi-Cloud from Core to Edge:** Zeblok deploys its Ai-MicroCloud™ anywhere, including enterprise data centers, public clouds and Edge locations

Partner Programs

- **Frontier:** CSPs and MSPs upsell Ai-MicroCloud™ to remain competitive; Specialized Hardware Manufacturers use Zeblok's orchestration to enable AI workloads on their hardware
- **Ingenuity:** Algorithm originators are able to develop their software more easily on our Ai-MicroCloud™ and we facilitate commercialization by including their algorithms in our Intelligence Marketplace
- **Insight:** Data providers benefit from our accelerated search capabilities
- **Build Intelligence Services:** Broad network of AI solutions firms help integrate AI into enterprises' mission-critical process

For more information: email Mouli Narayanan

Zeblok Computational Inc.
1500 Stony Brook Road
Stony Brook, NY 11794
www.computational.zeblok.com
mouli.narayanan@zeblok.com
Phone: +1 (631) 223-8233