3rd semi-annual report

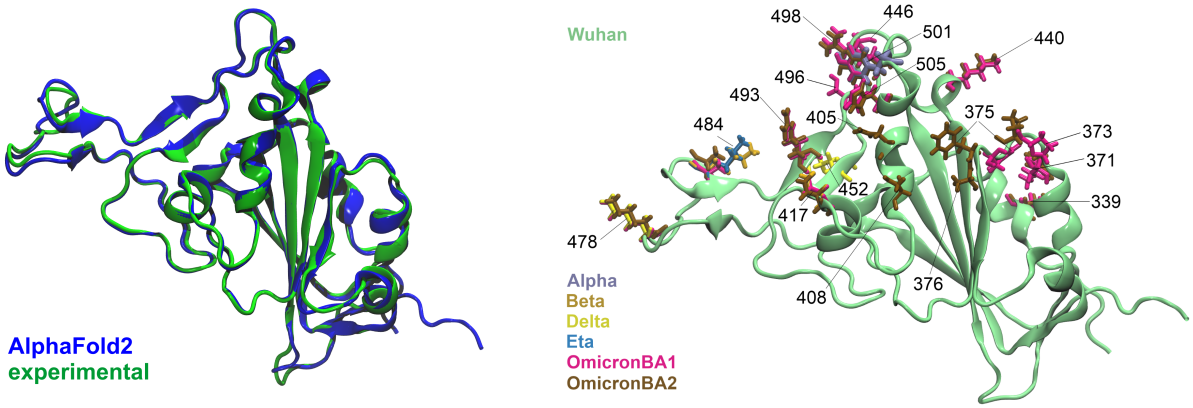
Oz Sam Kilim ozkilim@hotmail.co.uk
Physics PhD.
Supervisors: Prof. Csabai Istvan, Prof. Pollner Peter.

Domain generalization and representation learning for complex systems

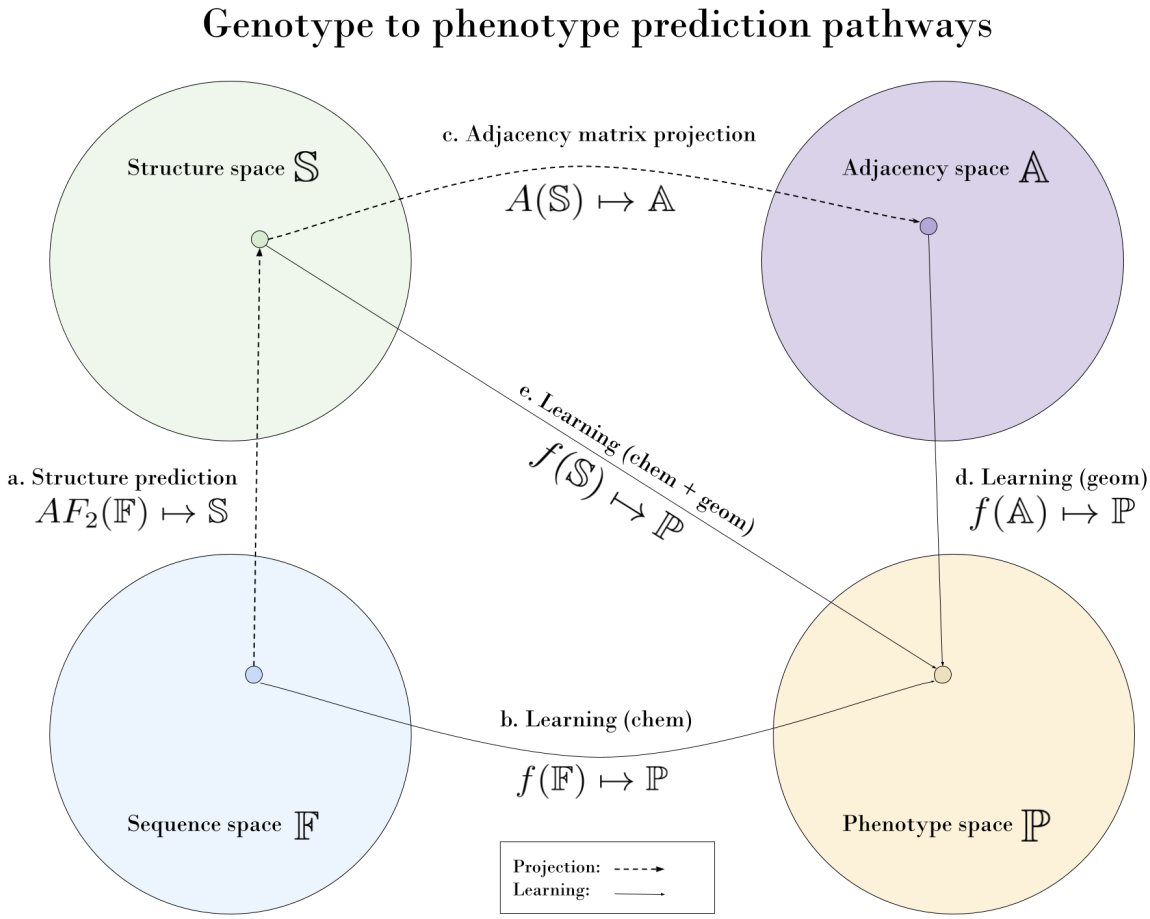
3 main projects undertaken during the 2022 autumn semester.

1. Sars Cov 2 Alphafold2 variants contain phenotypic information

Leveraging recent advancement in computational modeling of proteins with AlphaFold2 (AF2) we provide a complete curated data-set of all single mutations from each of the 7 main SARS-CoV-2 variants spike protein receptor binding domain (RBD) resulting in 3819*7= 26733 PDB structures. We visualize the generated structures and show that AF2 pLDDT values are correlated with state of the art disorder approximations, implying some internal protein dynamics is also captured by the model. Joint increasing mutational coverage of both structural and phenotype data coupled with advances in machine learning can be leveraged to accelerate virology research, specifically future variant prediction. We hope this data release can offer assistance into further understanding of the local and global mutational landscape of SARS-CoV-2 as well as provide insight into the biological prior that 3D structure acts as a bridge between protein genotype and phenotype. During the submission process of this article a large proportion of the technical validation was asked to be removed. This may provide substance for a follow up paper.



Left. AF2 aligned Wuhan WT RBD superimposed onto the experimentally determined 6M0J (RBD-ACE2 complex) clearly shows excellent agreement with respect to local and global structure. Slight deviation between the structures is visible in "loop" areas such as position 371 and 478. **Right**. Variant defining mutations on SARS Cov-19 spike protein RBD. Wuhan RBD is in the cartoon illustration while the variant defining mutations are illustrated with the licorice drawing method. The residue positions and the color codes are indicated.

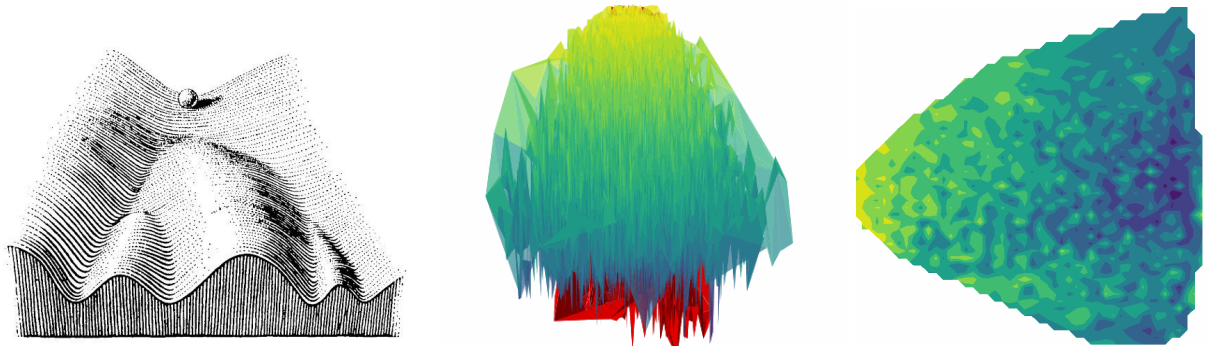


We provided a framework for thinking about how genotype to phenotype tasks can take place. Sketch of protein representations and their projections. $\mathbb{F}, \mathbb{S}, \mathbb{A}, \mathbb{P}$ are the spaces of all possible proteins for a given representation. $\mathbb{S}$ contains $\mathbb{F}$ and $\mathbb{A}$,formally, $\mathbb{F}, \mathbb{A} \subset \mathbb{S}$. Each protein has a FASTA one-hot-encoded representation $F \in \mathbb{F}$, a PDB file $S \in \mathbb{S}$, an adjacency projection of the PDB file $A \in \mathbb{A}$ and some measured phenotypic properties (function) $P \in \mathbb{P}$. We compare the projections $\mathbb{F}$ and $\mathbb{A}$ with respect to how a model f learns from these representations to make predictions about $\mathbb{P}$. **a**. Predict structure with AlphaFold2. **b**. Learning to predict protein-protein binding affinities from FASTA sequences. In the limit of huge amounts of genomic and phenotype data this may even build such a rich internal representation of protein interaction dynamics that explicit structure modeling (the top path of the loop) is not required. **c**. Creation of adjacency matrices from PDB structures. Representations in $\mathbb{A}$ carry no chemical information so can be used to analyze if the AF2 projection to $\mathbb{S}$ actually captured geometric signal that can be leveraged for phenotype prediction tasks, this representation has the added advantage of being rotation agnostic. **d**. Learning to predict protein-protein binding affinities with the adjacency matrices. **e**. $\mathbb{S}$ representations in PDB contain both chemical and geometrical information. An end goal could be to use this representation to build predictive models to predict $\mathbb{P}$ such methods have already been proposed. However this pathway is only worth using if we validate that **d**. is possible to some extent.

2. Visualization of the fitness landscape of Sars Cov 2 evolution

The unprecedented amount of data collected during the SARS-CoV-2 pandemic enables research in viral evolution. We used the antibody escape calculator from the Bloom lab and combinations of real GISAID and generated random combinatorial variant information in the form of sequence embeddings to create a low dimensional visualization of the time dependent evolution of SARS-CoV-2. We saw that the measured sequences of the virus traverse a lower dimensional manifold or “fitness valley” of the entire combinatorial space. This may provide evidence for the deterministic nature of the evolutionary course of the virus as well as show that early insights

into evolutionary landscapes as shown in the figure below may not be as misleading as some academics suggest.

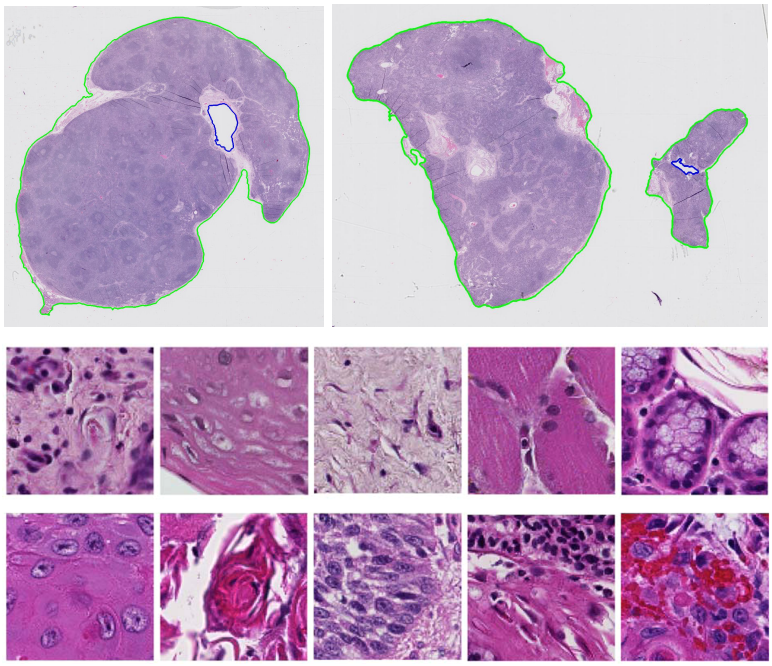


Left. The C. W. Waddington: 1974 Epigenetic landscape (conceptual visualization). **Middle**. View of the evolutionary landscape of sars cov2. The z axis represents the negative viral fitness as calculated by the Bloom lab. The x,y axes are PCA projections of the embeddings of each variant sequence. **Right**. Top down view of the landscape. Red points start on the left hand side where the early Wuhan variant is found and ends up on the right hand side where Omicron is observed.

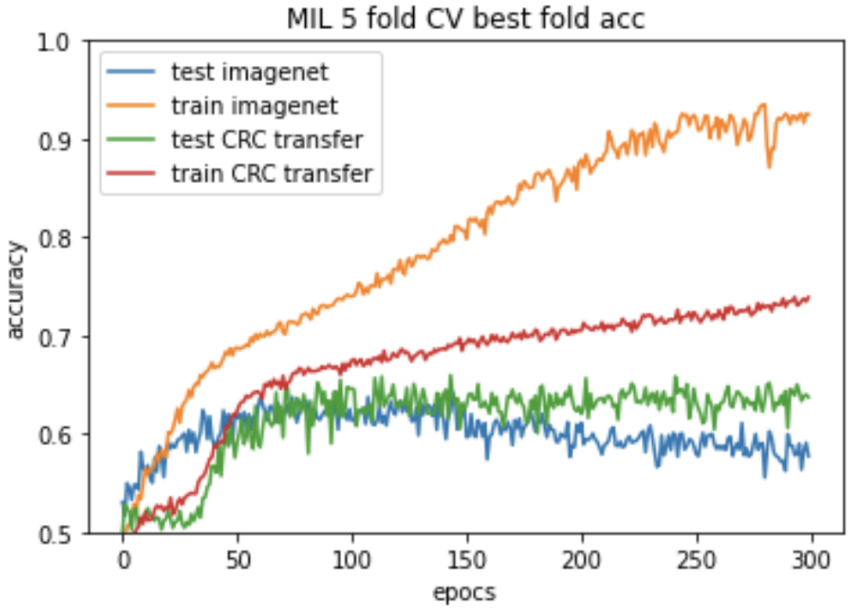
3. WSI phenotype classification

Breast cancer WSIs: Our team won the first part of the [Nightingale high risk breast cancer prediction challenge] where the task was to predict the staging classes of breast cancer whole slide images.

Lymphoma WSIs: Similar techniques were used where a more complex label was required to be predicted for a collaborative project. This project is still in progress and results are promising. Inference on WSIs across hospitals shows a large variance in accuracy. We hypothesize that this domain shift is largely influenced by staining variation. This is to be explored further in the coming semester.



Top. Segmentation results for exemplar WSIs. These gigapixel images contain vast amounts of hierarchical information that can be prognostic. We used a variety of weakly supervised ML techniques to predict either annotation labels or potential phenotypes that are not easily classified by doctors. **Bottom**. Examples of patches extracted from WSIs. These patches can contain rich information in themselves as well as context dependent information.



Binary classification task from WSIs for a phenotype of interest. Exploration of using patch embeddings from a ResNet trained on Imagenet vs the same ResNet trained on colorectal cancer (CRC) images. We see no large transfer of pathology knowledge with the CRC embeddings. This suggests that CRC features are not more appropriate for the Lymphoma modality.

Study activity

1 Module: Collective behavior.
Project work: https://github.com/ozkilim/pidgeon_GNN . This was a project to model the flight of homing pigeons from 3D trajectory data by leveraging a GNN architecture.

Publications

1. https://www.nature.com/articles/s41598-022-23990-4
2. https://www.mdpi.com/2076-2615/13/2/194
3. https://www.biorxiv.org/content/10.1101/2022.10.15.512391.abstract
4. In review: Scientific data: *SARS-CoV-2 RBD deep mutational AlphaFold2 structures*.

Teaching activity

Lectures on neural networks and weekly marking for deep learning course.