4th 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 2023 spring semester.

1. Broad Institute Immunotherapy single-cell RNAseq Gene regulatory network inference challenge 2023.

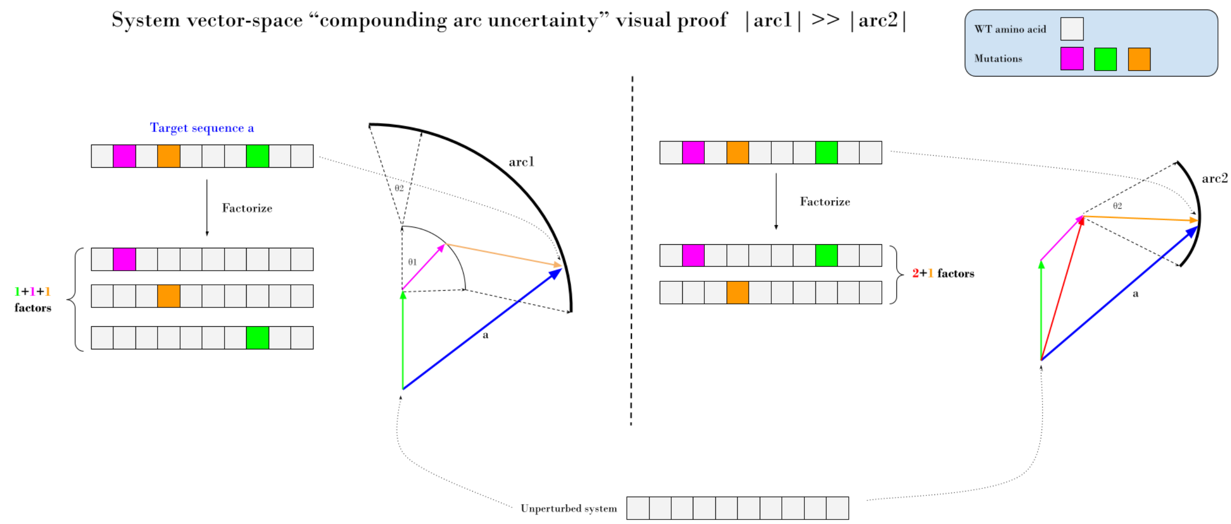
We took part in an international challenge for prediction of the effect of single gene knockouts on B cell cycle state relating to potency of immunotherapy in mice models. We took the GEARS [1] model based on a graph neural network (GNN) trained to learn knockout "perturbation" embeddings leveraging co-expression and gene ontology graphs. The original model aimed to learn the change in gene expression from an unperturbed cell to a perturbed one. We removed the MLP cell expression head after the GNN layers and used this embedding as an input for our model. The motivation to use this model was its natural representation of the inductive biases in the data. Due to such a small number of examples we needed to inject domain knowledge as a constraint in order to have a model that could generalize to unseen knockouts.

A gene's mechanism to impact the expression of other genes is partially through the proteins it expresses. In order to leverage this knowledge we included protein information as an input to enrich the models internal representation of the system. Each knockout can be represented by a bag of expressed protein embeddings where some proteins may have a larger impact on the downstream process we are trying to model (for example transcription factors) and some less. This problem is aligned naturally with the weakly-supervised method of multi-instance-learning (MIL), where we used an attention layer to learn weights for the importance of each "item" (protein embedding in our case) in providing predictive power for the final state distribution of a given knockout experiment. Our MIL MLP model outputted an embedding that was fed into the final model. Auxiliary Chromatin information was used as it should contain information about the accessibility of genes. This data was embedded with an MLP and these embeddings were concatenated with the outputs of a multi-instance learning MLP and GEARS embeddings. The 3 branches were concatenated and fed into an MLP. During public validations steps we achieved scores in the top 25 out of ~200 participating groups with 900 registered. The standard deviation of scores was very small with only a few groups breaking out ahead for Challenge 1 and 2. https://www.broadinstitute.org/news/machine-learning-experts-around-world-compete-improve-cancer-immunotherapy

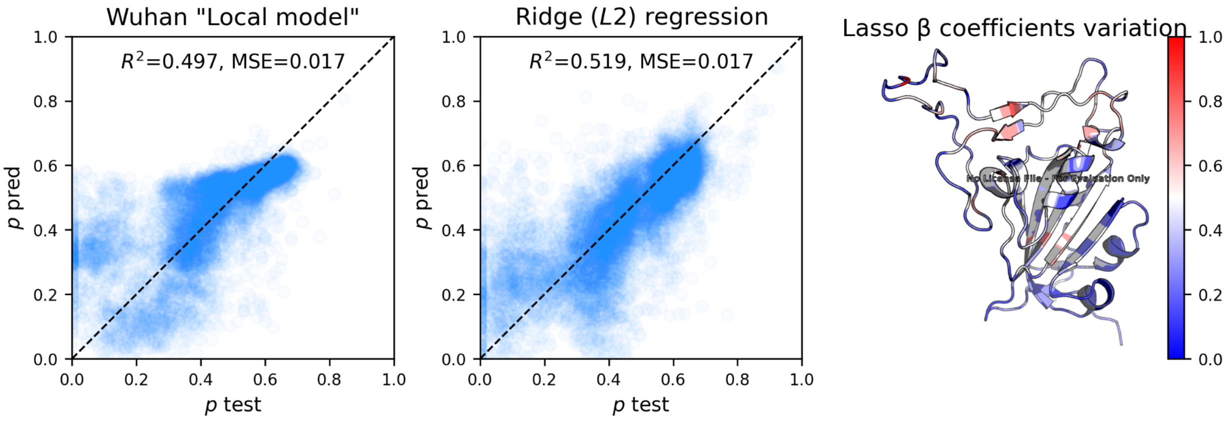
2. Modeling epistasis in fitness landscapes.

We explored sequence-to-function mapping with further modeling techniques and various protein protein interaction datasets of systems including the receptor binding domain (RBD) of Sars-Cov-2. We explored predicting Omicron ACE2 and Antibody binding from early Wuhan sequences and lab measurements to see if we could generalize from early in the phylogenetic tree to later stages with Machine Learning [2].

For ACE2 binding we found that the RBD of Omicron was under purifying selection. Omicron has increased antibody escape in comparison to Wuhan and its only ACE2 binding selective pressure was to bind sufficiently to this receptor to be contagious. We see that the local ACE fitness landscape topology of Omicron and Wuhan are similar and a "local model" can perform as well as highly hyperparameter optimized regularized regression models when predicting Omicron phenotypes. We explored factorization as a concept to enable distant landscape predictions however we were limited by a lack of relevant deep mutational scanning datasets in our ability to validate our hypotheses fully. We have built a modular framework for extraditing this stream of research https://github.com/ozkilim/combinatorial_learning .Users can change sequences, embedding functions and models in order to further explore their systems of interest. We hope to release this as an open source codebase with benchmarks on various deep mutational scanning datasets. This project has involved in-depth study into theoretical epistasis literature including NK-landscapes.



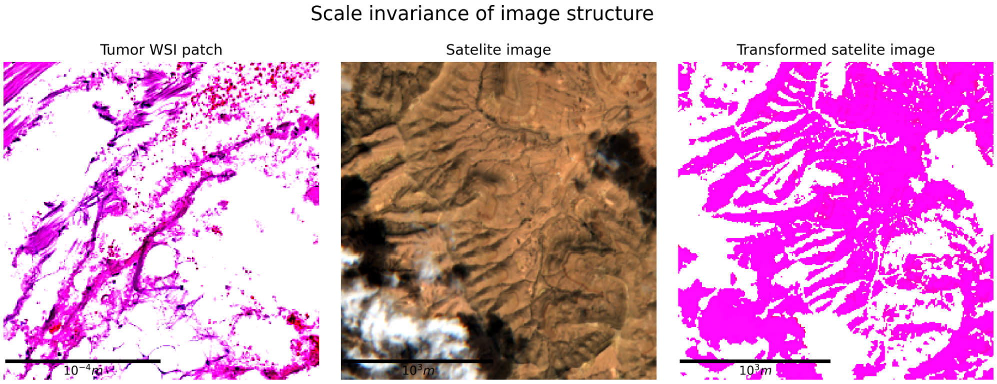
Visual proof that larger factors of targets should provide a less uncertain estimate of the target when combined. We can think of representing a sequence as a colored vector where for example its angle is related to its amino acid and its color is its position in the protein. For each combinatorial mutant, the factors add in a specific way. We assume that vector addition is a good model however there is some noise that models the epistatic non-linear effects. Due to uncertainty apriori in how mutants will combine we can think of each consequential mutant existing on some arc in the vector space. When we have more factors we have more positions where there is uncertainty in how the vectors add so the final arc is larger than the arc after the addition of larger factors. We are yet to validate or invalidate this hypothesis successfully due to limited data availability. We have set up a framework for "neural cross attention factorization".

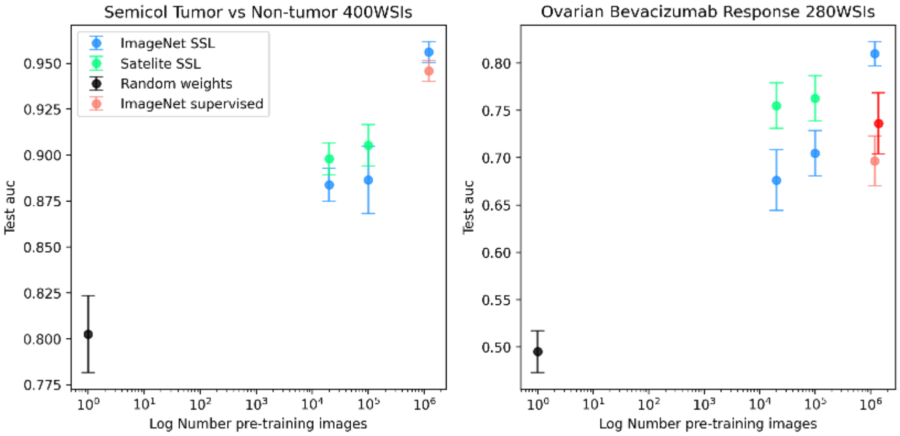


Analysis of generalization to the rugged landscape with regularization. a. Non-epistatic model. Here we assume that all mutations are additive. (Here the local landscape is similar to the omicron local landscape.) b. Ridge L2 regularized linear regression results. Slightly out perform the summation model. The regularizing factor relieves over-fitting. c. Visualization of regression coefficients on the RBD structure reveal the (top) binding interface. These results give evidence for a purifying selection mechanism in the Omicron system.

3. Transfer learning across physical scales

We were part of the Nightingale challenge for the task of cancer staging from WSIs and it was observed that embedding networks seem to be a bottleneck in model predictive accuracy. Inspired by the similarity between whole slide images (WSI) and satellite imagery we explored the extent to which satellite images can be used within the self supervised learning framework DINO [3] to create better embeddings for downstream WSI tasks utilizing the weakly supervised learning framework CLAM [4].


Scale invariance of image structure
Tumor WSI patch
Satelite image
Transformed satelite image
$10^{-4}m$
$10^{3}m$
$10^{3}m$



Initial results of SSL image scales vs downstream accuracy for WSI tasks. For most of the tasks satellite imagery provided better embeddings than ImageNet. CUrrently we are in the process of scaling up our experiments to larger dataset sizes.

We aim to show that self-supervised pre-training of neural networks with satellite data set can create superior embeddings to ImageNet for various data scales as tested on WSI digital pathology downstream tasks across 5 data-sets representing diverse cancer types. We give an example where non-domain but similarly structured data-sets can be leveraged for important medical tasks. We aim to release our Dino-Trained ResNet50 model on ~10 Million earth images that can be used for downstream ML-pathology tasks. We hypothesize that the shared constraints and complexity of systems of different physical scales allows for this efficient transfer learning as well as the ability to extract a very large "almost unlimited" scale pre-training dataset. This concept could be generalized and we hope that ML practitioners may be inspired to search for external modality data-sets that may aid their task at hand. This has been validated at low dataset sizes and with promising results.

Study activity

1 Module: Networks.
Complex exam: Networks, ML methods, Bioinformatics.

Teaching

Supervision of Tamás Zsiga for thesis project on modeling epistasis in the Sars-Cov-2 RBD system for multiple antibody escape measurements.

Publications

MTMT profile: https://m2.mtmt.hu/gui2/?type=authors&mode=browse&sel=10087749
Google scholar: https://scholar.google.com/citations?user=DtBgwP8AAAAJ&hl=en&oi=ao

1. Kilim, Oz, et al. "Physical imaging parameter variation drives domain shift." *Scientific Reports* 12.1 (2022): 21302.
2. Kilim, Oz, et al. "SARS-CoV-2 receptor-binding domain deep mutational AlphaFold2 structures." *Scientific Data* 10.1 (2023): 134.
3. Nagy, Sára Ágnes, et al. "Impact evaluation of score classes and annotation regions in deep learning-based dairy cow body condition prediction." *Animals* 13.2 (2023): 194. (played a minor role in finalizing the machine learning method and its explanation in the paper)
4. Gréta Tóth, Adrienn, et al. "Ixodes ricinus tick bacteriome alterations based on a climatically representative survey in Hungary." *bioRxiv* (2022): 2022-10. (played a minor role in reviewing the paper internally)

References

1. Roohani, Yusuf, Kexin Huang, and Jure Leskovec. "GEARS: Predicting transcriptional outcomes of novel multi-gene perturbations." *BioRxiv* (2022): 2022-07
2. Starr, Tyler N., et al. "Deep mutational scans for ACE2 binding, RBD expression, and antibody escape in the SARS-CoV-2 Omicron BA. 1 and BA. 2 receptor-binding domains." *PLoS pathogens* 18.11 (2022): e1010951.
3. Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.
4. Lu, Ming Y., et al. "Data-efficient and weakly supervised computational pathology on whole-slide images." *Nature biomedical engineering* 5.6 (2021): 555-570.