## Gerstein lab experience in using LLMs to predict disordered proteins and misfolding

Protein phase transitions (PPTs), especially those observed as a result of disordered states and aggregation, have the potential to interfere with protein-ligand binding by altering target structures, reduce functional availability, and affect binding thermodynamics and kinetics. In our recent work, we leveraged LLMs by fine-tuning the pre-trained ESM-2 model to predict PPTs and aggregation (Frank et al, 2024). Specifically, we demonstrated that the LLM embedding vectors can differentiate between sequences with high aggregation propensity from those with low aggregation propensity.

A graph of a graph showing a number of dots

Description automatically generated with medium confidence

PCA of the embedding vectors, extracted from the LLM, and the biophysical features for both classification tasks.

## Gerstein lab experience in using GNNs to predict Intrinsically Disordered Regions (IDRs)

We also developed a method that is designed to determine whether a specific subregion within a protein adopts disordered states (even if the entire protein itself does not adopt a disordered state) (Wang et al, 2024). Specifically, we introduced the GP-GNN framework, a GNN-based approach designed for predictive tasks where initially unknown subsystems may be crucial for information extraction. Our framework dynamically learns advantageous graph structures, hence, enabling improved graph representation learning. Specifically, we formulated a similarity measure that captures the probability of node pairs jointly exhibiting characteristic patterns within larger systems and used this measure to group nodes into clusters. Connections between clusters are considered unlikely to be relevant to the task and are removed, generating multiple subgraphs before graph representation learning via message passing. This is to confine information sharing within subgraphs during the initial network layers. We leveraged a linear layer to extract meta-features, allowing the similarity measure to be trained with task labels, thereby facilitating task-specific adjustments in the model’s predictions. In the final network layers, the framework captures interactions between the subgraphs. We demonstrated the effectiveness of GP-GNN for predicting protein liquid-liquid phase separation (LLPS), where the input consists of molecular protein graphs, and the output indicates whether the protein undergoes LLPS. Since LLPS is primarily driven by domains such as IDRs, GP-GNN identifies these critical domains as subgraphs. Our model achieved state-of-the-art performance in predicting LLPS proteins, including regulators and scaffolds, using both PDB and AlphaFold-generated graph structures. Additionally, in related work, we showed that a similar approach accurately identified tumor and microenvironment regions that align closely with pathologist annotations. (Song et al, 2024).

## Gerstein lab experience in using protein dynamics to rank protein-ligand pairs

We developed STRESS, a tool to identify allosteric residues from protein structures (Clarke et al, 2016). STRESS employs coarse-grained models of protein dynamics and conformational changes in order to predict surface pockets that may strongly influence dynamics if occluded by ligands (see figure at right). We have used this tool to highlight examples of allosteric residues that help explain disease-associated variants. Using the same dynamic model generation scheme applied by STRESS, we also developed ​​HotCommics, a framework to identify cancer driver genes using a dynamics-based search of mutational hotspot communities within structures (Kumar et al, 2019). Relative to existing cancer hotspot detection methods using structural data, we applied ​​HotCommics to demonstrate how protein dynamics significantly increases the sensitivity of cancer driver detection.

A diagram of different types of clouds

Description automatically generated

STRESS. Identifying key surface pockets using Monte Carlo simulations. This approach is designed to identify candidate sites wherein ligand occlusion may drastically affect conformational changes and dynamics (top-right).

## Gerstein lab experience in using population-level sequence data to quantify evolutionary conservation

We developed multiple computational tools and analysis pipelines to study genomic sequence variation and the associated downstream effects of variants. VAT (Variant annotation tool) annotates protein sequence changes resulting from genetic mutations, categorizing mutations as synonymous, missense, nonsense, or splice-site disrupting (Habegger et al, 2012). We also developed ALoFT (Annotation of Loss-of-Function Transcripts) predicts the disease-causing potential of loss-of-function (LoF) events, distinguishing between deleterious heterozygous and homozygous mutations, and it has been used to analyze somatic variants in cancer exomes (Balasubramanian et al, 2017). We applied the concept of localized protein frustration to evaluate the structural impact of variants by calculating changes in localized physical metrics to assess how genetic variants affect protein structure and stability (Kumar et al, 2016). This has been used to differentiate gain- and loss-of-function variants in oncogenes and tumor suppressors.

A diagram of a model of mutated structure

Description automatically generated

**Fig. 9**. Prioritizing the effect of SNVs based on changes in localized perturbations (as measured by frustration).

We also have extensive experience with conducting analyses related to allele-specific expression (ASE). In particular, we developed AlleleSeq, a computational pipeline that was originally used for identifying and quantifying ASE in GM12878 (Rozowsky et al, 2011). We have applied AlleleSeq to a broad spectrum of personal and functional genomics data. We annotated variants associated with ASE in a large pool of individuals from the 1000 Genomes Project. The results were made available as an online resource, AlleleDB (<http://alleledb.gersteinlab.org>) (Chen et al, 2016).

## Gerstein lab experience in constructing regulatory networks in neurogenomic contexts

In our previous work (Emani et al, 2024), we developed an intercellular communication network by integrating publicly accessible ligand-receptor pair data with snRNA-seq. We identified distinct broad communication patterns among excitatory, inhibitory, and glial cell types, with significant alterations in these patterns observed in patients with neuropsychiatric disorders such as schizophrenia and bipolar disorder. Specifically, we noted that excitatory neurons showed reduced incoming signaling in schizophrenia, whereas inhibitory neurons exhibited increased signaling.

## Bibliography & References Cited

Balasubramanian, S. et al. Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes. Nature communications 8, 382, doi:10.1038/s41467-017-00443-5 (2017).

Chen, J. et al. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. Nature communications 7, 11101, doi:10.1038/ncomms11101 (2016).

Clarke, D. et al. Identifying Allosteric Hotspots with Dynamics: Application to Inter- and Intra-species Conservation. Structure (London, England : 1993) 24, 826-837, doi:10.1016/j.str.2016.03.008 (2016).

Emani, Prashant S., et al. "Single-cell genomics and regulatory networks for 388 human brains." Science 384.6698 (2024): eadi5199.

Frank, M., Ni, P., Jensen, M. & Gerstein, M. B. Leveraging a large language model to predict protein phase transition: A physical, multiscale, and interpretable approach. Proceedings of the National Academy of Sciences of the United States of America 121, e2320510121, doi:10.1073/pnas.2320510121 (2024).

Habegger, L. et al. VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. Bioinformatics 28, 2267-2269, doi:10.1093/bioinformatics/bts368 (2012).

Kumar, S., Clarke, D. & Gerstein, M. Localized structural frustration for evaluating the impact of sequence variants. Nucleic acids research 44, 10062-10073, doi:10.1093/nar/gkw927 (2016).

Kumar, S., Clarke, D. & Gerstein, M. B. Leveraging protein dynamics to identify cancer mutational hotspots using 3D structures. Proceedings of the National Academy of Sciences of the United States of America 116, 18962-18970, doi:10.1073/pnas.1901156116 (2019).

Rozowsky, J. et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. Molecular systems biology 7, 522, doi:10.1038/msb.2011.54 (2011).

Song, T. et al. Predicting spatially resolved gene expression via tissue morphology using adaptive spatial GNNs. Bioinformatics 40, ii111-ii119, doi:10.1093/bioinformatics/btae383 (2024).

Wang, G., Warrell, J., Zheng, S. & Gerstein, M. A variational graph-partitioning approach to modeling protein liquid-liquid phase separation. Cell Reports Physical Science 5, doi:10.1016/j.xcrp.2024.102292 (2024).