Knowledge-Guided Statistical Learning Methods for Analysis of High-Dimensional Omics Data in Precision Medicine

Qi Long, Ph.D.

University of Pennsylvania

September 24, 2020

Background

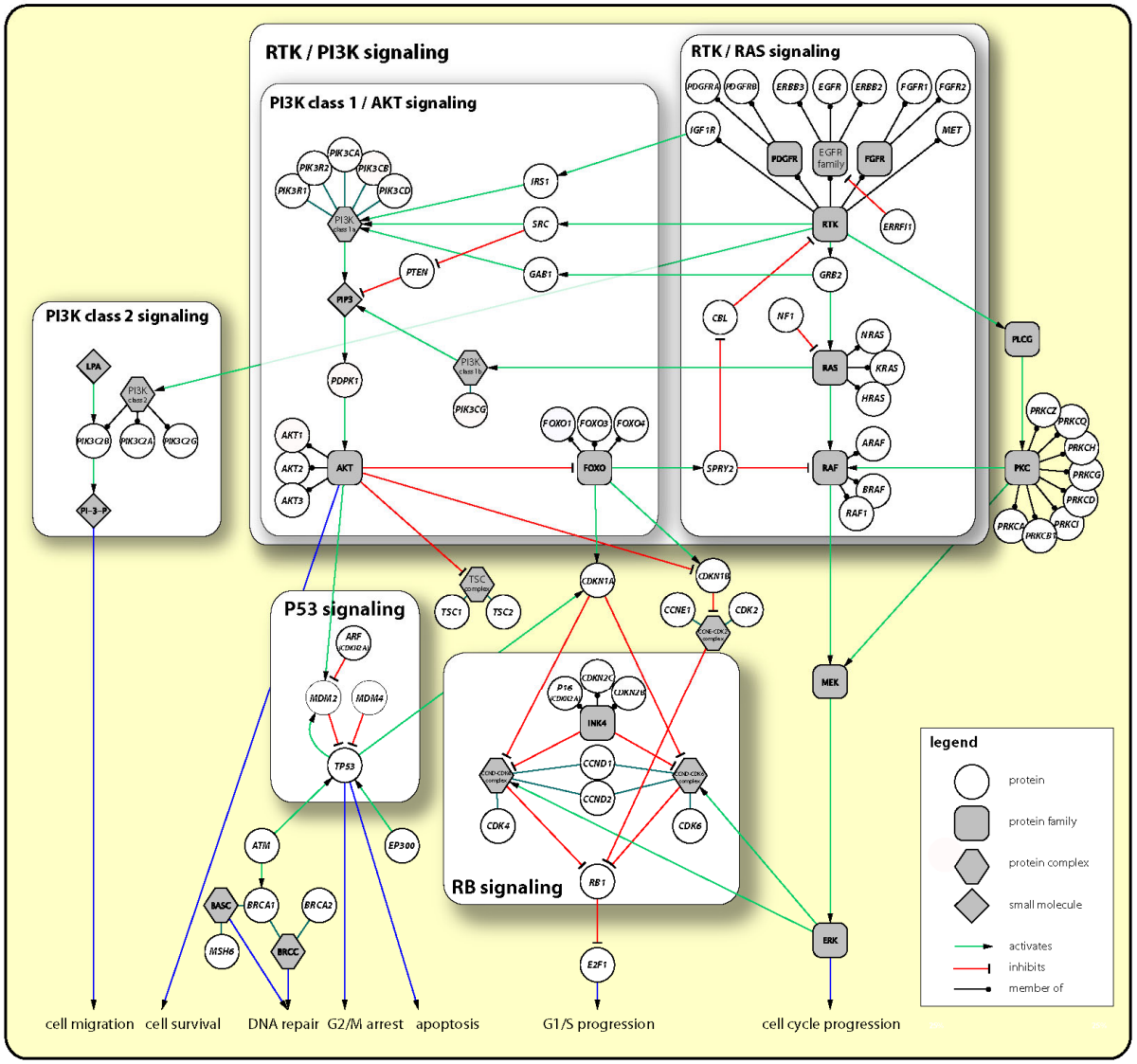
- Generation of high-dimensional -omics data such as genomics, transcriptomics, and metabolomics data offers great promises in advancing precision medicine
  - insights about heterogeneity in disease risk/prognosis and in treatment response
  - also present significant analytical challenges.
- Complex diseases are often multifactorial that may be attributed to harmful changes on multiple (omics) levels and on pathway level.
  - heterogeneity
  - aggregated signal in pathway can be considerably stronger
- Vast majority of existing statistical learning methods are entirely data driven
  - fail to incorporate biological knowledge

Biological Information/Knowledge

- Ever-deepening knowledge about biological functions and interactions of genes, gene products, and other –omics features
  - Gene regulatory networks
  - Gene co-expression networks
  - Protein-Protein interaction networks
  - Metabolic pathways

- Such biological knowledge can be represented by
  - Graph: $G = < V, E >$
  - V: the set of nodes such as genes, proteins, metabolites etc.
  - E: the set of edges representing functional interactions between nodes

Signaling pathways in glioblastoma



Biological Information/Knowledge

Database	Full Name	Knowledge
KEGG	Kyoto Encyclopedia of Genes and Genomes	Metabolic pathways
REACTOME	Reactome Pathway Database	Metabolic & signaling pathways
Mummichog	Mummichog	Metabolomic pathway
Metabolome.jp	Metabolome.jp	Metabolic pathways
MetaCyc	Metabolic Pathways From all Domains of Life	Metabolic pathways
Invitrogen iPath	Invitrogen iPath	Metabolic pathways
BioCyc	BioCyc Pathway/Genome Database Collection	Metabolic pathways
IPKB	Ingenuity Pathways Knowledge Base	Gene regulatory & signaling pathways
TRANSPATH	TRANSPATH	Gene regulatory & signaling pathways
CST	Cell Signaling Technology Pathway	Signaling pathways
TargetScan	TargetScan	gene-microRNA regulatory network
miRBase	miRBase: the microRNA database	gene-microRNA regulatory network
PicTar	Probabilistic identification of combinations of target sites	gene-microRNA regulatory network
miRDB	miRDB	gene-microRNA regulatory network
mirDIP	microRNA Data Integration Portal	gene-microRNA regulatory network
BioGRID	Biological General Repository for Interaction Datasets	Protein and genetic interactions
HuRI (Luck, 2020)	Reference interactome map of human binary protein interactions	Protein-Protein Interactions

Knowledge-guided Statistical Learning Methods

- Leverage information in $G = < V, E >$
- Improve the power of detecting weak, yet important signals:
  - signal from individual –omics feature can be weak and challenging to detect with a small/moderate sample size
  - in aggregate, signal from a pathway would be easier to detect with same sample size
- Yield biologically more interpretable and meaningful results
  - encourage selection of pathways vs individual features
  - deeper insights about molecular mechanism of complex diseases
  - heterogeneity
- Facilitate integration of multi-omics data through the incorporation of biological knowledge about functional relationships between different modalities.

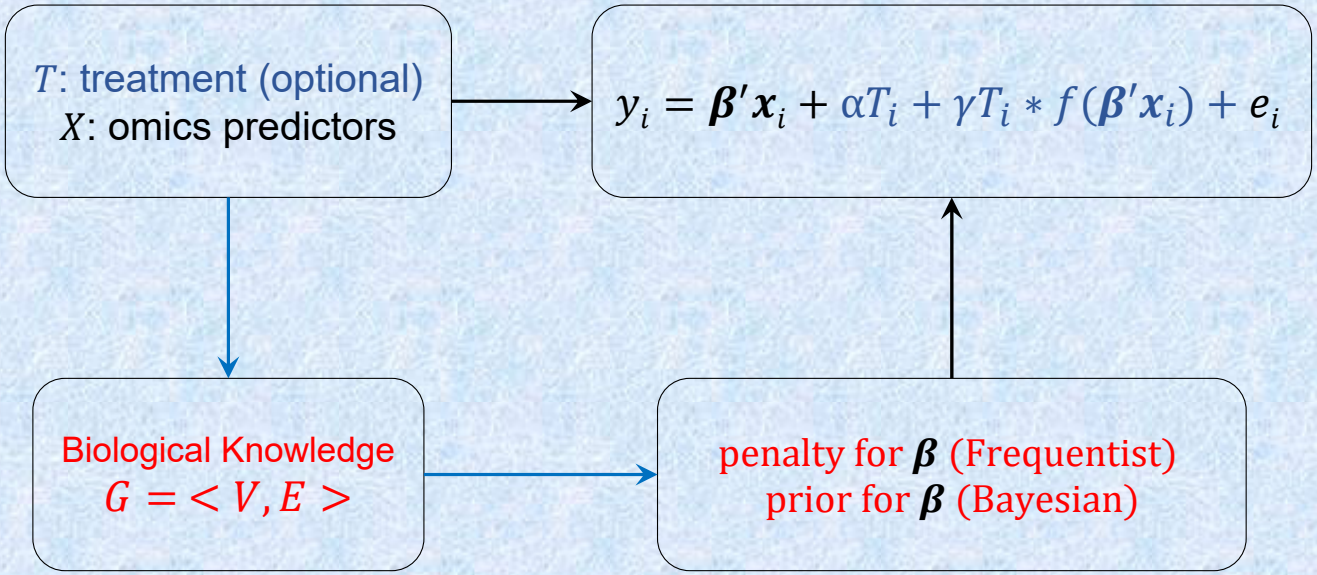
Knowledge-guided Statistical Learning Methods

- Knowledge-guided supervised learning methods
  - Construct prediction models for disease risk/progression or treatment response
    - identify higher risk group
    - tailor interventions to individual patients: heterogeneity of treatment effects
  - Uncover molecular signatures predictive of disease risk/progression, or treatment response
    - inform novel targets for therapeutic development
- Knowledge-guided unsupervised learning methods
  - Biclustering: identify disease subgroups and important pathways associated with each subgroup.
  - Identification of subgroups related to molecular differences
    - offer insights about optimizing treatment strategy for each subgroup
    - important step toward developing a precision medicine approach for complex diseases.

Knowledge-guided Supervised Statistical Learning Methods

- Knowledge-guided linear regression model
  - *Y:* clinical outcome of interest
  - ***X***: a set of high-dimensional omics features/predictors
  - *T*: treatment variable (optional)
  - *G*=<*V*, *E*>: the graph containing the biological knowledge about ***X***
    - V: the set of nodes (i.e., omics features)
    - E: the set of edges (functional relationship)
  - To incorporate *G*=<*V*, *E*>
    - Frequentist framework: penalty for $\boldsymbol{\beta}$
    - Bayesian framework: prior distribution for $\boldsymbol{\beta}$; straightforward for statistical inference and uncertainty quantification for estimation and prediction

Knowledge-guided Supervised Statistical Learning Methods



Knowledge-guided Supervised Statistical Learning Methods

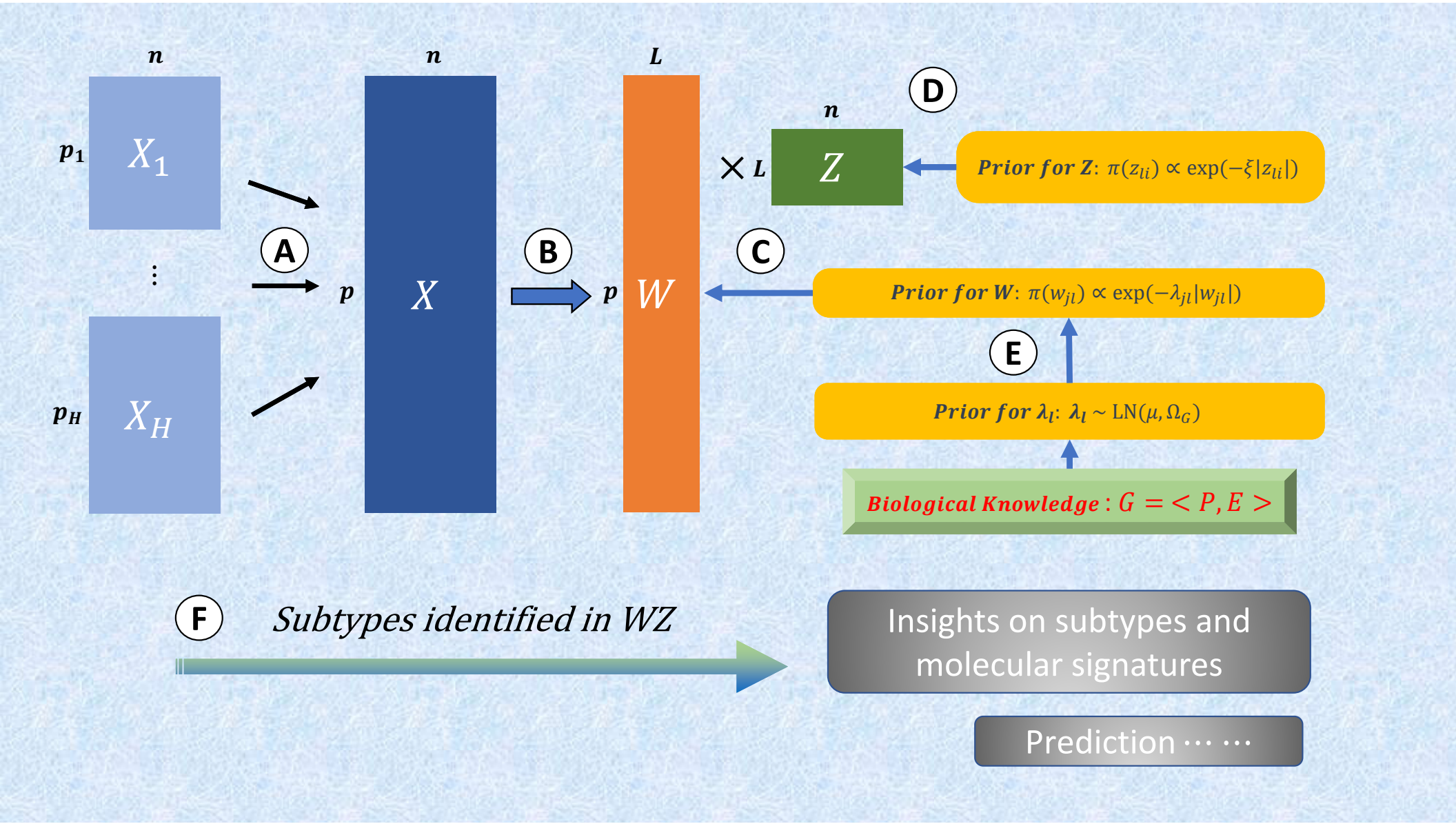
- Penalties in Frequentist Framework:
  - Network constrained penalty (Li and Li, 2009)
    - More accurate prediction for time to death among glioblastoma patients using microarray gene expression
    - Identified gene subnetworks that were highly correlated with survival time
    - Implication in precision medicine: identified gene signatures provide insights about molecular underpinning of prognosis and potential therapeutic target
  - Grouped penalty based on L_r-norm (Pan et al., 2010)
  - Hierarchical group penalty (Zhao et al., 2016)
    - More accurate prediction for prostate cancer recurrence after prostatectomy
    - Implication in precision medicine: help determine whether adjuvant therapy is needed after surgery for individual patients.

Knowledge-guided Supervised Statistical Learning Methods

- Prior Specifications in Bayesian Framework:
  - Spike and slab prior combined with Markov random field (or Ising) prior (Li and Zhang, 2010): scalability for ultra-high dimensional features (?)
  - Structured adaptive shrinkage prior (Chang et al., 2018)
    - Computationally scalable to 100,000's or millions of –omics features
    - More accurate prediction for overall survival in glioblastoma patients
    - Identified a set of risk genes along with enriched pathways
    - Implication in precision medicine: identified gene signatures provide insights about molecular underpinning of prognosis and potential therapeutic target
- Other knowledge-guided supervised learning methods:
  - SVM methods (Sun et al. 2019)
  - LDA methods (Safo et al. 2019)

Knowledge-guided Unsupervised Statistical Learning Methods

- Bayesian generalized biclustering analysis (Li et al. 2019)
  - *Step A*: Integrate data from *H* omics modalities, $\boldsymbol{X}_1$, …, $\boldsymbol{X}_H$ (continuous and discrete)
  - *Step B*: $\boldsymbol{X}$ is linked to loading $\boldsymbol{W}$ and latent factors $\boldsymbol{Z}$.
  - *Step C and D*: Prior for $\boldsymbol{W}$ and $\boldsymbol{Z}$, respectively.
  - *Step E*: Prior for incorporating biological knowledge.
  - *Step F*: Biclusters identified in the product $\boldsymbol{W}*\boldsymbol{Z}$ provides insights on disease subgroups and associated molecular signatures.
    - Subgroups can be correlated with heterogeneous treatment response


n
n
L
n
D
p_1
X_1
$\times L$
Z
Prior for Z: $\pi(z_{li}) \propto \exp(-\xi|z_{li}|)$
A
B
C
p
X
p
W
Prior for W: $\pi(w_{jl}) \propto \exp(-\lambda_{jl}|w_{jl}|)$
E
p_H
X_H
Prior for λ_l: $\lambda_l \sim \mathrm{LN}(\mu, \Omega_G)$
Biological Knowledge: $G = < P, E >$
F
Subtypes identified in WZ
Insights on subtypes and molecular signatures
Prediction ⋯ ⋯

Knowledge-guided Unsupervised Statistical Learning Methods

- Biclustering: Li et al. (2019)
  - Integrative analysis of gene expression data, DNA methylation data, and DNA copy number data from a TCGA glioblastoma dataset
    - Subgroups identified had a higher correlation with survival outcome than those by other biclustering methods that do not use biological information
  - Analysis of two -omics datasets from AMP-AD
    - This new method outperforms existing biclustering methods in terms of identifying clinical subgroups including AD, asymptomatic AD, progressive supranuclear palsy, pathologic aging, and cognitive normal (CN).
- Dimension Reduction and Feature Engineering
  - PCA (Li et al., 2017)
  - CCA (Safo et al., 2018)
  - CIA (Min et al., 2018, 2020)

Software Tools

Supervised Learning	R	Graph-constrained regularization for both sparse linear regression and sparse logistic regression	Li and Li (2008)[8] Sun et al (2014)[14]
	R	Fused lasso	Tibshirani et al (2005)[35]
	R	Incorporating Predictor Network in Penalized Regression with Application to Microarray Data	Pan et al (2010)[9]
	Matlab	Network-based penalized regression with application to genomic data	Kim et al (2013)[15]
	R	Scalable Bayesian variable selection for structured high-dimensional data	Chang et al (2018)[10]
	Matlab	Sparse knowledge-guided LDA	Safo et al (2016)[18]
	Matlab	Incorporating Biological Information into Linear Models: A Bayesian Approach to the Selection of Pathways and Genes	Stingo et al (2011)[24]
	Matlab	Joint network and node selection for pathway-based genomic data analysis	Zhe et al (2014)[11]
Unsupervised Learning	Matlab	Sparse knowledge-guided PCA	Li et al (2017)[12]
	Matlab	Sparse knowledge-guided CCA	Safo et al (2018)[36]
	R	Sparse knowledge-guided CIA	Min et al (2018)[37]
	Matlab	A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression	Liu et al (2014)[31]

Discussions

- The knowledge-guided strategy has been shown to
  - enhance analysis power
  - yield biologically more interpretable and meaningful results
- Bayesian methods enable straightforward statistical inference and uncertainty quantification for estimation and prediction
- While substantial progress has been made in the development of knowledge-guided statistical learning methods
  - promote the use of these methods in modeling treatment heterogeneity using high-dimensional –omics data
  - still much room for further methodological developments and improvements

Future Research

- Assess robustness of knowledge-guided methods to misspecification of biological knowledge
  - In practice biological knowledge from existing databases is known to be incomplete and include false edges
  - Some knowledge-guided learning methods (Chang et al. 2018) have been shown to be robust.
- Combine knowledge-guided methods with learning biological graph *G* from observed data: an integrative analysis approach
- Scalable methods for analysis of big -omics data that can have hundreds of thousands of or even millions of features
  - More research on efficient computation algorithms is needed.

References

- Zhao, Y, Chang, C and Long, Q, (2019) Knowledge-guided statistical learning methods for analysis of high-dimensional-omics data in precision oncology. *JCO Precision Oncology*, 3, pp.1-9.
  - References in Zhao et al. (2019)

- Luck, K., Kim, D.K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B. and Choi, D., 2020. A reference map of the human binary protein interactome. *Nature*, 580(7803), pp.402-408.

Thank you!