Frederick National Laboratory
for Cancer Research
sponsored by the National Cancer Institute

AI for Multimodal Biomedical Data

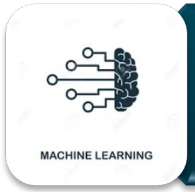
George Zaki, Ph.D. [Contractor]
Bioinformatics Manager
Biomedical Informatics and Data Science Directorate
Frederick National Lab for Cancer Research

Pinyi Lu, Ph.D. [Contractor]
Data Scientist
Biomedical Informatics and Data Science Directorate
Frederick National Lab for Cancer Research

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute

Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

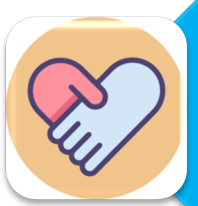
Scientific Computing Activities Program Development at Frederick National Lab for Cancer Research



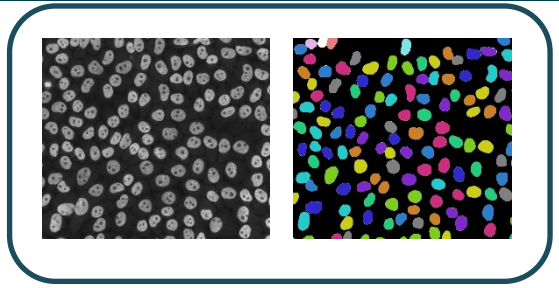
Machine Learning for Image Processing, NGS, Drug, and NLP

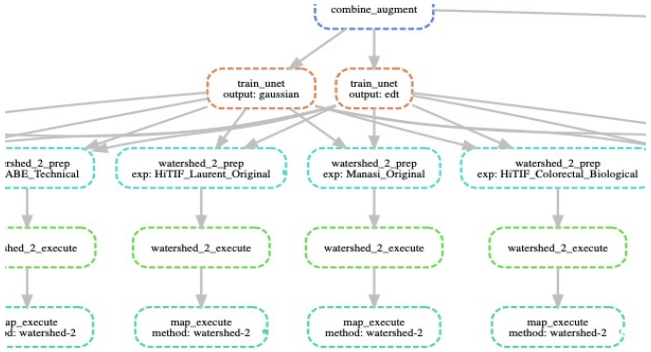
Scientific Computing Workflows Development

Custom Code Optimization & Accelerated Computing



Education, Outreach/Community Building





NIH.AI

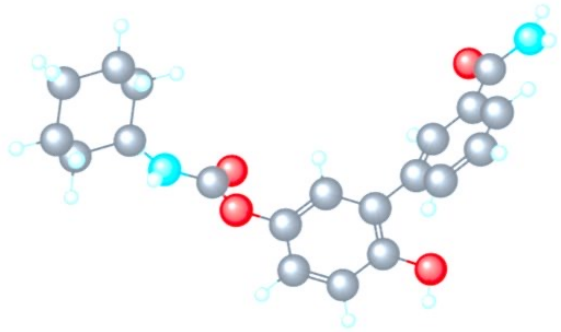
george.zaki@nih.gov

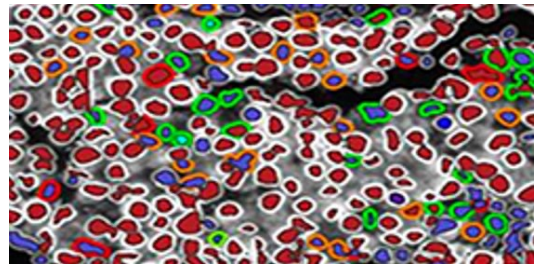
“Key questions in cancer research involve observing multiscale phenomena and collecting multimodal data from diverse sources; therefore, single datasets and most existing methods are insufficient.”

Sharpless NE, Kerlavage AR. *“The potential of AI in cancer care and research”,* Biochim Biophys Acta Rev Cancer. 2021

Data Types in Bioinformatics

- **Clinical:**
  - age, sex, race, histories, pathologies, therapeutics
- **Omics:**
  - genomics, transcriptomics, proteomics, metabolomics, etc.
- **Radiology Images:**
  - CT, CBCT, MRI, PET
- **Pathology Images:**
  - H&E, Immunohistochemistry (IHC), Multiplex immunofluorescence (MxIF)
- **Small Molecules:**
  - Mode of actions, chemical descriptors, etc.
- **Free Text:**
  - Pathology reports, abstracts, etc.
- **Other types:**
  - Cryo-EM, high content images, etc.





Examples of Applications of Multimodal Biomedical Data

- **Predict cancer prognosis and diagnosis[1]**
  - Multifaced diseases (cancer, cardiac disease, diabetes, etc.)
  - Reduce noise from a single source
  - Integrate data at different scales and organism levels
- **Generate new mechanistic insights**
  - Visualize and cluster cancer subtypes [2]
  - Understand response to treatments [3]
- **Predict Drug Sensitivity**
  - Enable precision medicine

[1] Cheerla, Anika, and Olivier Gevaert. "Deep learning with multimodal representation for pancancer prognosis prediction." Bioinformatics 35.14 (2019): i446-i454.

[2] Hoadley KA, Yau C, Wolf DM, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158(4):929-944.

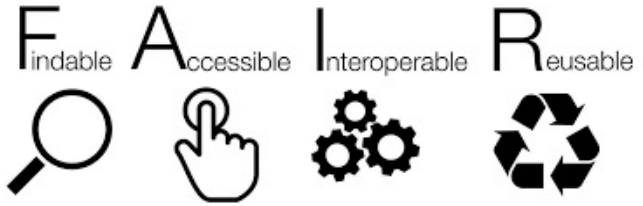
[3] De Cecco, Loris, et al. "Integrative miRNA-gene expression analysis enables refinement of associated biology and prediction of response to cetuximab in head and neck squamous cell cancer." Genes 8.1 (2017): 35.

Why Multimodal Biomedical Data in Machine Learning?

1. **Feature importance in classification:** What subset of key modalities and features is responsible for the separation of classes? Example: multiple panels in histochemistry
2. **Better predictive power:** increase classification of regression power using multimodal data. Example: Survival analysis from whole slide and gene expression
3. Same as 2 but for **unsupervised learning** (e.g., clustering). Example: Tumor subtyping
4. Study **interaction between different modalities** to understand complex biological systems from different angles:
   - Genotype-phenotype interactions
   - Drug response
5. **Missing modalities:** Perform the task with one modality when other modality is missing

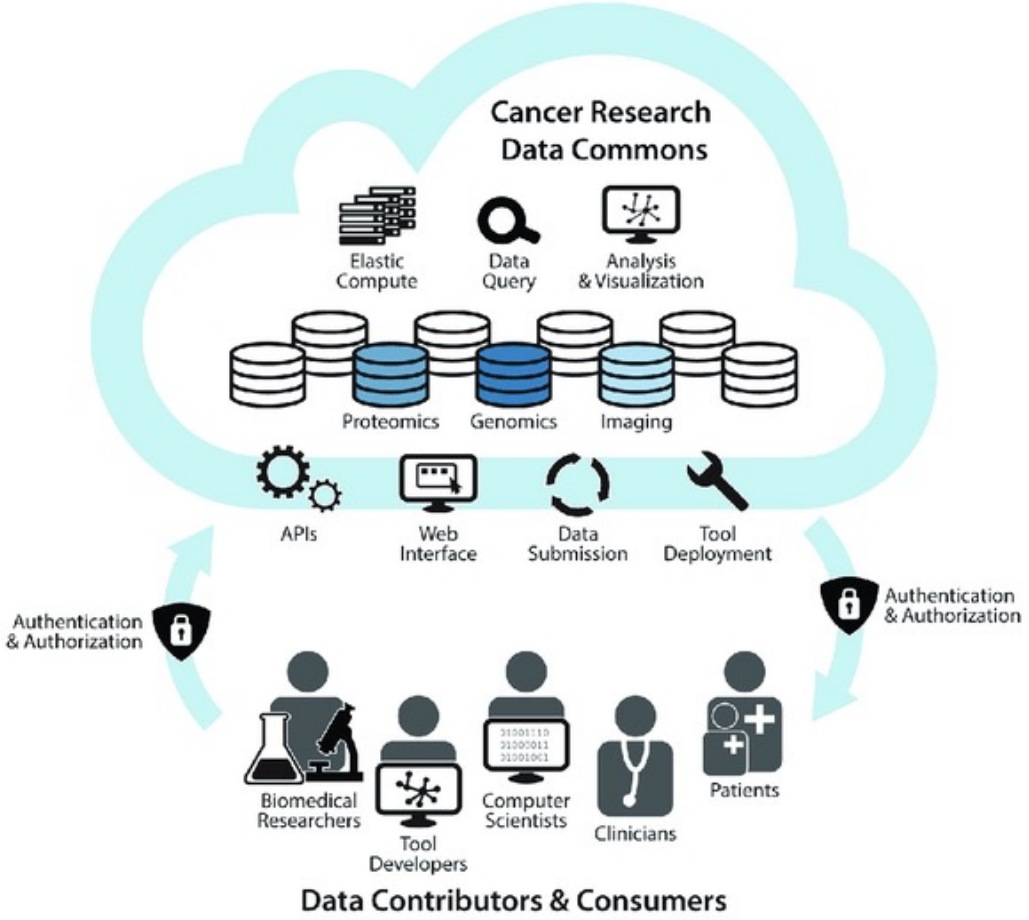
Challenges of Using Multimodal Data

- **Curse of dimensionality:** *$N_{features\ (all\ modalities)} >> N_{samples}$: 1000s of features in 100s of samples.*
- **Heterogeneous data:** scale of features, type of features, fusion, etc.
- **Missing data:** Remove, impute, bias
- **Rarity and class imbalance**
- **Big data salacity:** FAIR Data, scalable compute, etc.



Big Data Scalability

- **FAIR Data**
  - Data annotation
  - Data retrieval
- **Scalable Compute:**
  - Personal
  - Virtual machines
  - Division cluster
  - NIH High Performance Compute cluster: Biowulf
  - Cloud compute
  - Department of Energy Leadership compute

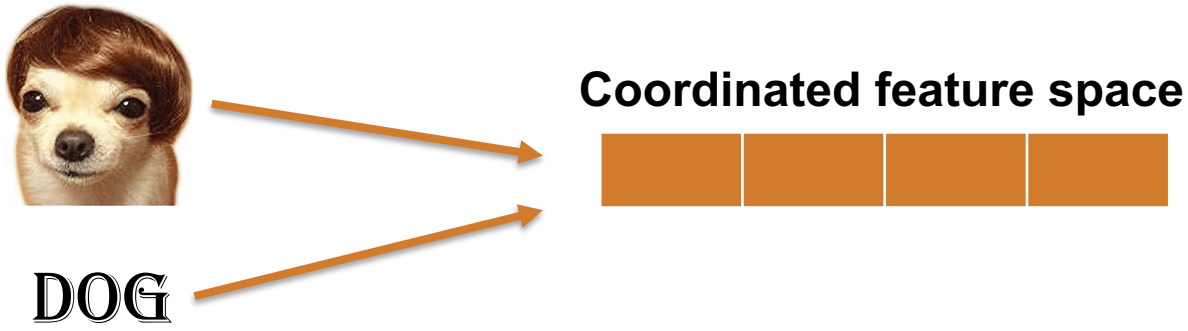


Challenges: Curse of Dimensionalities

- Every modality is represented by a features vector (e.g, pixels for images)
- $N_{features\ (all\ modalities)} >> N_{samples}$: *1000s of features versus in 100s of samples.*
- Makes machine learning prune to **over fitting**: good performance on training data, worse performance on test data
- Multimodal data makes this problem even harder
- Can be solved using:
  - **Dimensionality reduction techniques** (more later)
  - **Features selection techniques:**
    - Filter out features that do not provide much Information Gain (IG)
    - Iteratively train surrogate models with a subset of features
    - Use models that implicitly apply feature selection (LASSO regression in linear models)

Challenges: Missing Data

- **Remove samples** with missing data. Potential for large data loss
- **Impute missing data:** mean, median, regression from un-missing data, K-nearest neighbor, etc.
- **Imputation from existing modalities:** e.g., add the text based on images
  - For complementary modalities, project the two modalities into a new coordinated space, such that if one modality is missing, the other modality can be used in a given for prediction.



- Other methods include maximum likelihood estimators, Gaussian Mixture Models, denoising autoencoders for clinical and RNASeq imputation.
- **Multiple imputation:** Impute using different methods, resulting multiple new imputed samples
- The imputation methods **are prune to add bias**

Challenges: Heterogeneous Data

- **Number and types of features** in every modality:
  - Continuous, discrete, categorical, interval variables
  - Different scales, distribution, statistical properties
- **Potential preprocessing:**
  - Normalize every modality (e.g. zero mean, unit variance)
  - Scale the values in every modality by the inverse of the number of features
  - Compare every modality independently using Multiple Kernel Methods
    - Different data source has different notion of similarities
  - Dimensionality reduction for every modality separately (e.g., autoencoder) [1], [2]

[1] Zhang, Tianyu, et al. "Synergistic drug combination prediction by integrating multiomics data in deep learning models." *Translational Bioinformatics for Therapeutic Development*. Humana, New York, NY, 2021. 223-238.

[2] Cheerla, Anika, and Olivier Gevaert. "Deep learning with multimodal representation for pancancer prognosis prediction." *Bioinformatics* 35.14 (2019): i446-i454.

Challenges: Rarity and Class Imbalance

- Baseline classifier can completely ignore rare class and achieve very high accuracy by always predicting abundant class (e.g., over 99% for 10000, 100 imbalance in samples)
- This is a very common problem in biological data: enhancer in genomes, DNA methylation status, modification of amino acid residues, etc.
- Potential solutions:
  - **Data sampling:** before classification, up-sample (e.g., using SMOTE), down-sample, or mix of both
  - **Algorithm modification:** apply a higher loss weight to the minority class (e.g., SVM_Weight)
  - **Ensemble learning:** Train multiple classifier using the the minority class and a random subsample of the majority class, then combine predictions of individual classifiers

Challenges: Rarity and Class Imbalance

- Use appropriate metrics to evaluate the algorithm
- For binary class (Majority is negative, Minority is positive):
  - Specificity (accuracy of the majority class) = True Negative / Total Negative
  - Sensitivity (accuracy of the minority class) = True Positive / Total Positive
  - F_1 score = (2TP) / (2TP + FP + FN)
  - Other metrics: balanced error rates, area under the precision-recall curve, etc.
- For multiclass;
  - Micro and Macro F1 scores, balanced error rates, confusion matrix, etc.

How to Incorporate/Fuse Multi-View Data in the Learning Process?

- **Early:**
  - Concatenate the features as a single vector.
  - Features can be normalized (zero mean, unit variance)

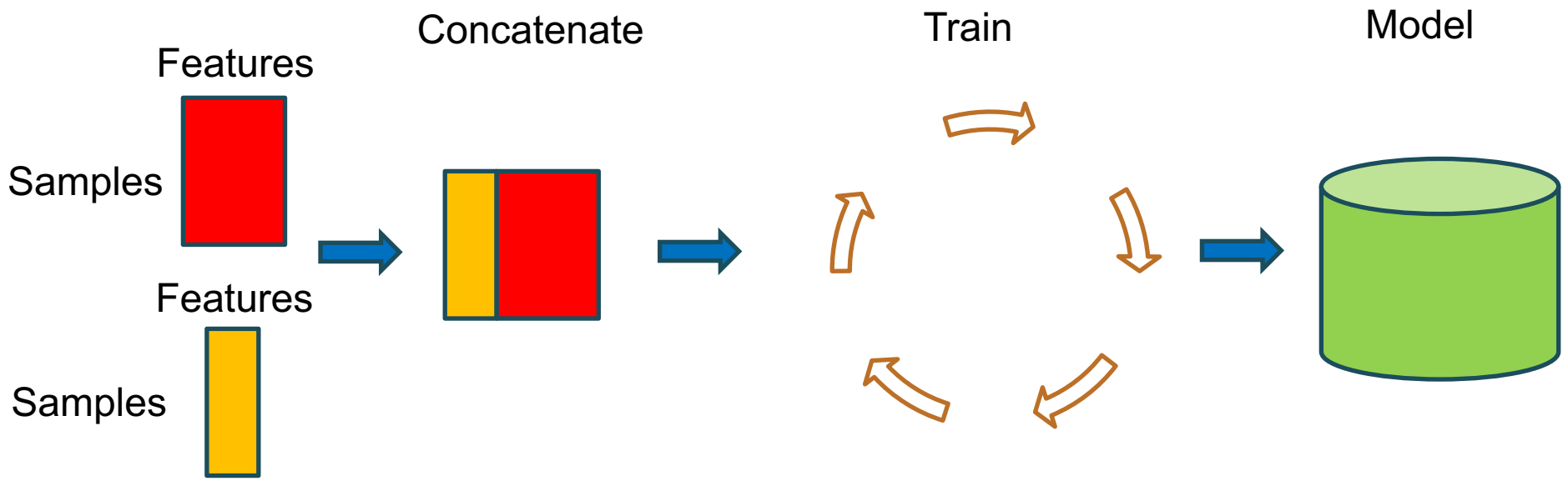


Image adapted from: Nobel W, Support vector machine applications in computational biology, 2004

Features Concatenation

- **Features come in different forms:**
  - Continuous, discrete, characters, graphical, etc.
  - A conversion would be needed:
    - Continuous to discrete (or vice versa)
    - Categorical to one hot coding (e.g., for three classes: “100”, “010” ,”001”)
- **Features come in different scales:**
  - Normalize and Standardize
- **Concatenation might not be feasible:**
  - Example (bag of words representing a document + image pixels) The semantics of the bag of words will be lost
- Concatenated features can be used in linear classification with regularization to select the most important features in achieving a task.

Trees of Mixed Data Types

- Decision trees can combine continuous and discrete data simultaneously
- There is no need for normalization because values of continuous variable can be split into ranges as part of the rules
- Decision trees are prone to noise and overfitting
  - Solution can be in and ensemble of multiple trees (e.g., random forests)
- Early and late incorporation of mixed data types can be used to build the trees

How to Incorporate Multi-view Data in the Learning Process?

- **Intermediate:**
  - First compute on every modality separately, then combine the partial computation as input to the prediction model

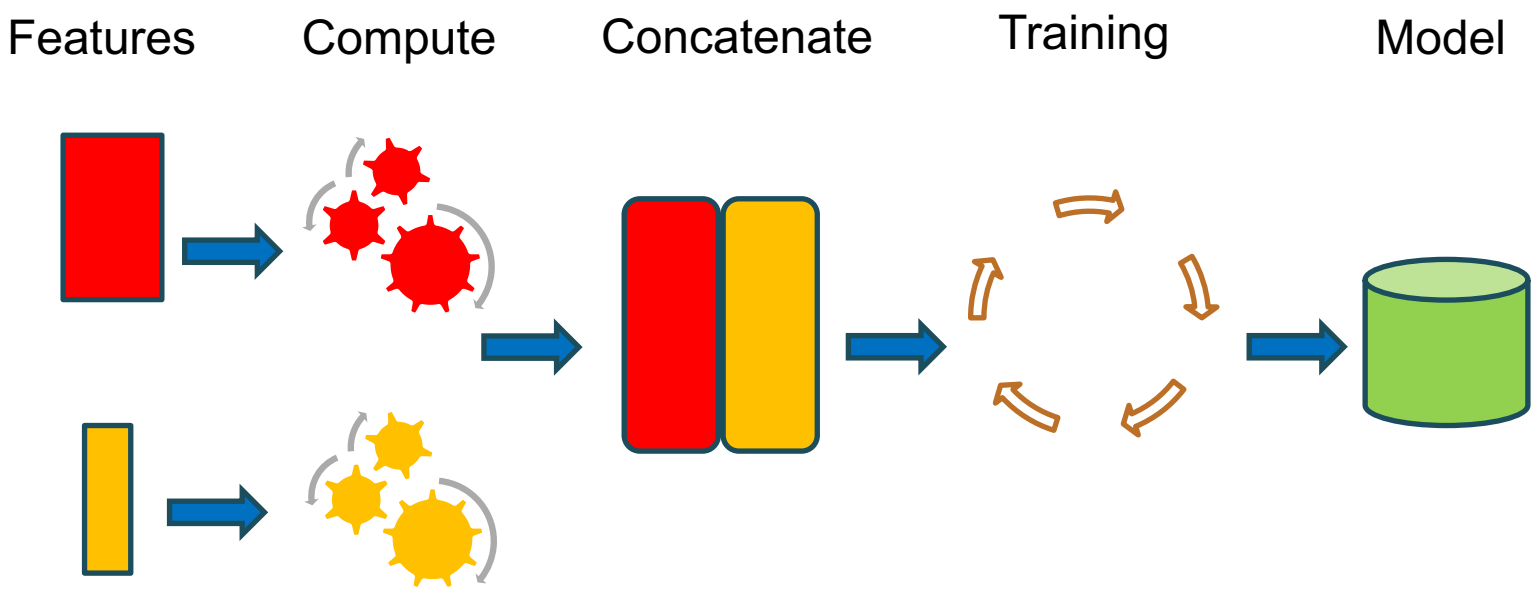
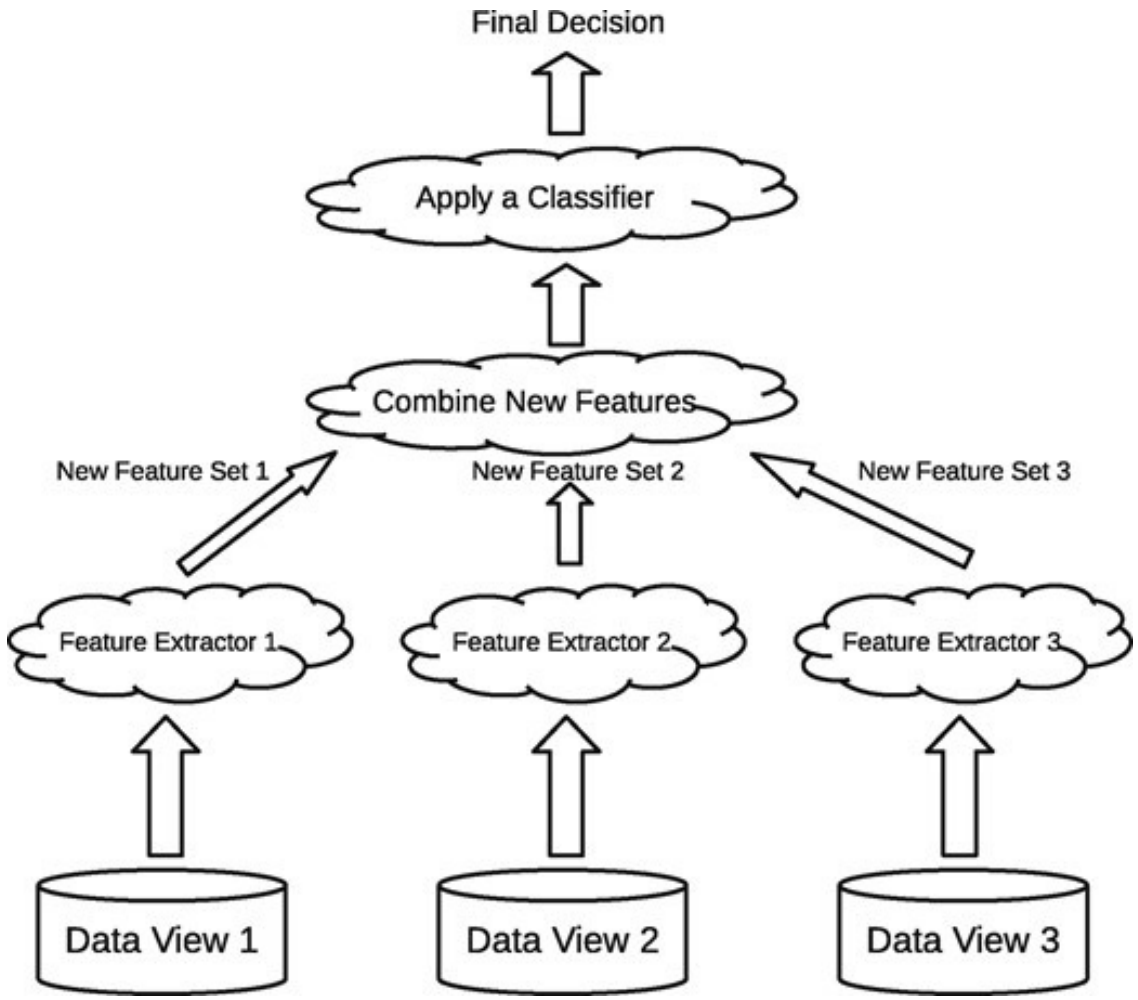


Image adapted from: Nobel W, Support vector machine applications in computational biology, 2004

Feature Extraction to View Specific Components

- Apply a feature **extraction/reduction method for every modality separately,** then concatenate these features.
- How to extract features:
  - Matrix factorization: e.g., Principal Component Analysis, Multi Omics Factor Analysis, etc.
  - Non-linear dimensionality reduction methods:
    - t-Distributed Stochastic Neighbor Embedding (t-SNE)
    - Autoencoders (neural networks)
- New features are numeric, **easy to concatenate, and have smaller dimensions**
- Interactions between features still cannot be accounted for.



How to Incorporate Multi-view Data in the Learning Process?

- Late:
  - Learn separate models from every modality, then combine the outputs to make a final prediction

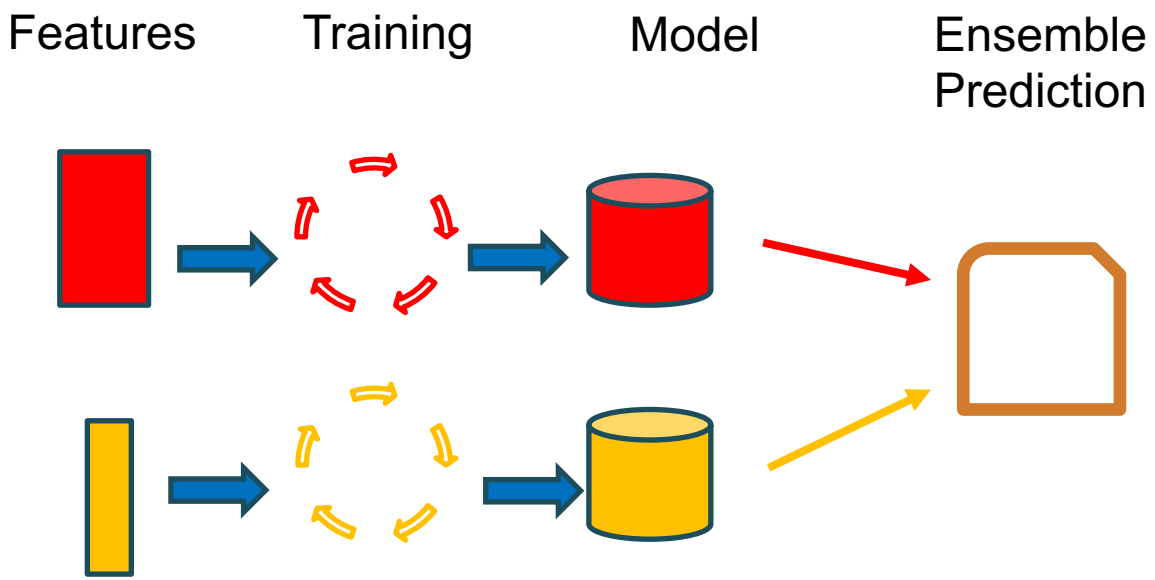


Image adapted from: Nobel W, Support vector machine applications in computational biology, 2004

Multi-Omic Clustering for 10 Cancer Types

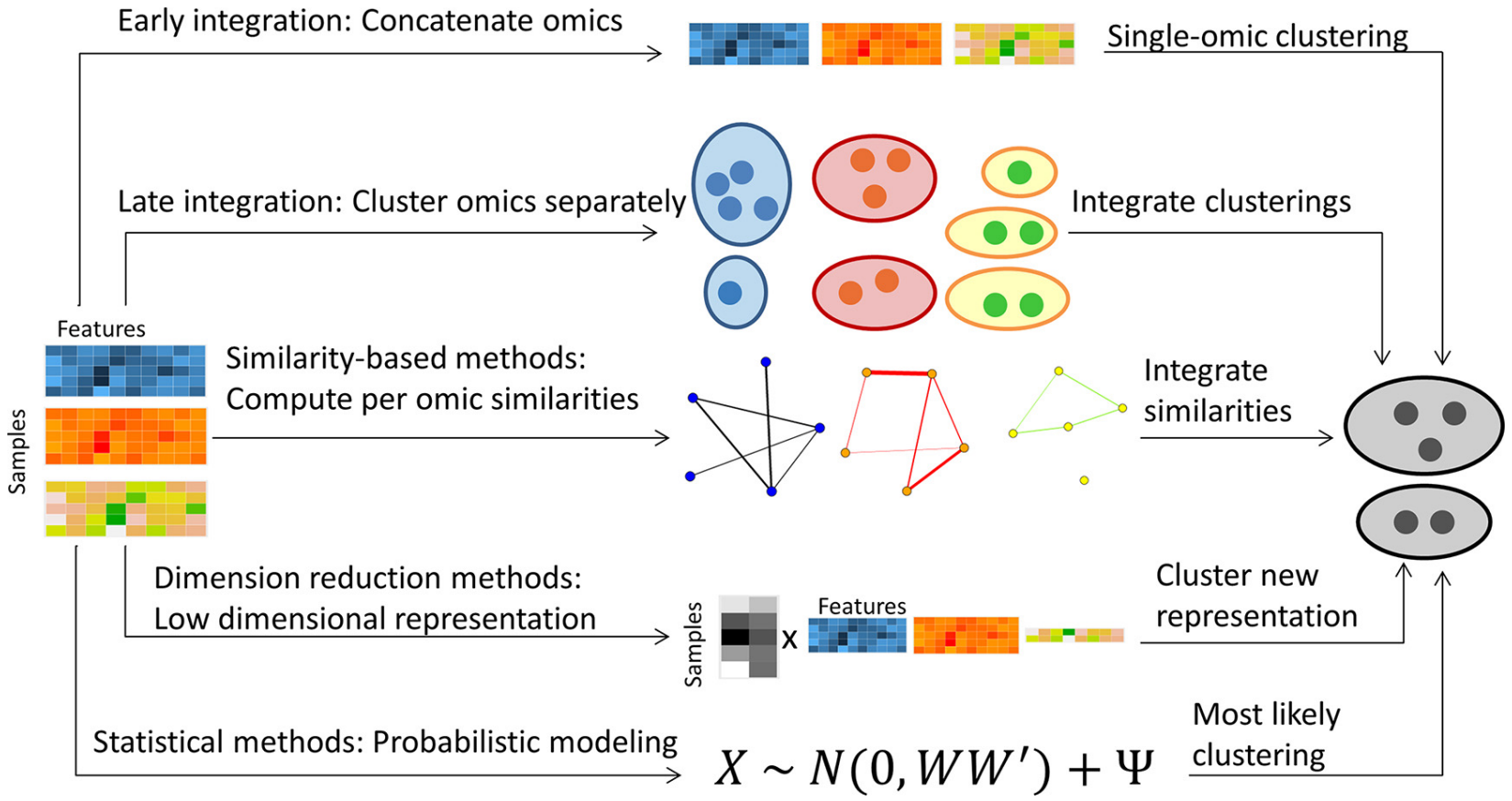
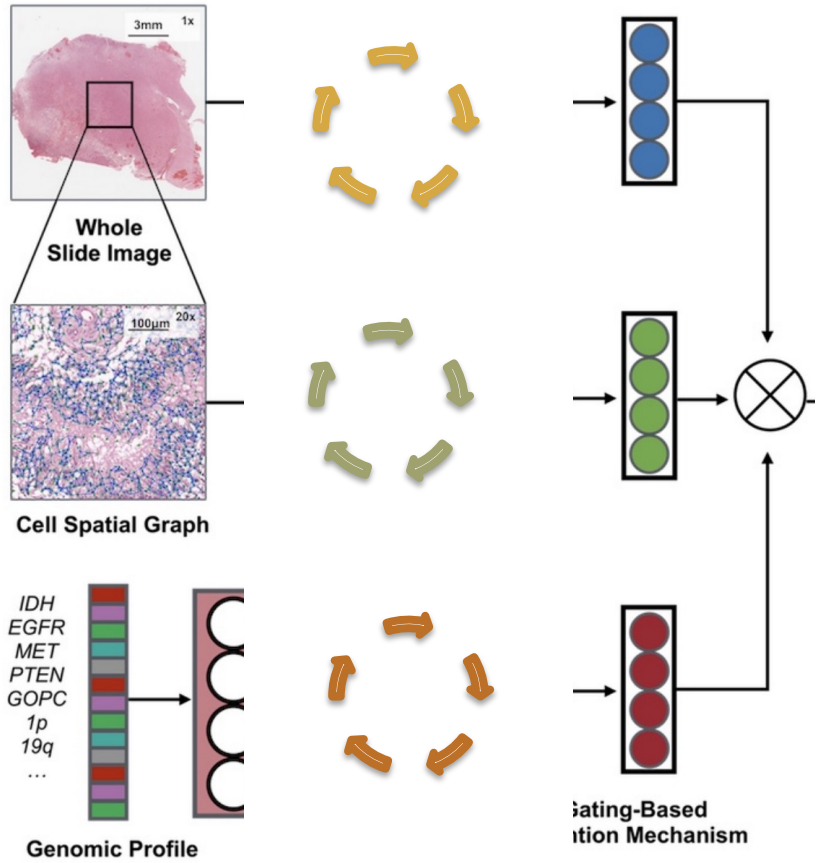
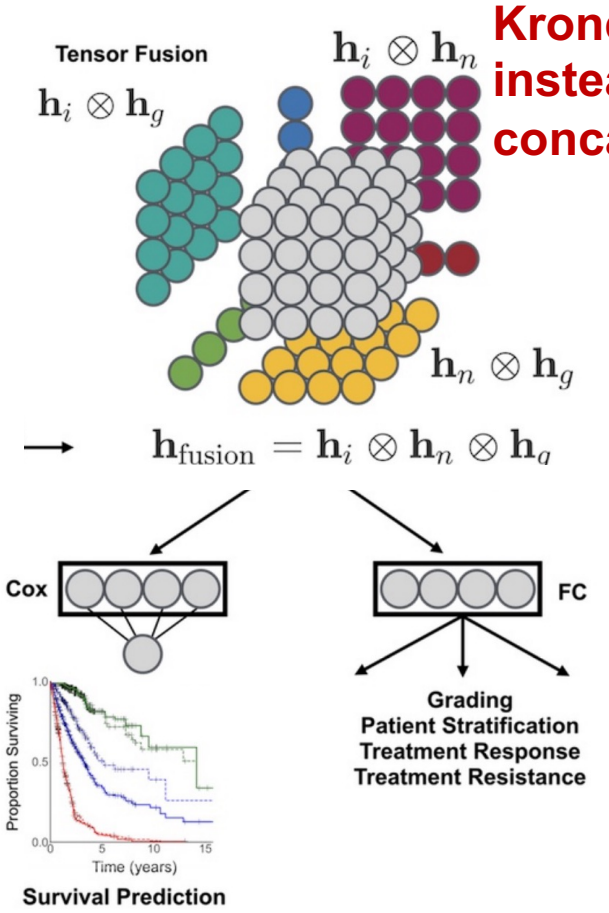


Figure source: Nimrod Rappoport, Ron Shamir, Multi-omic and multi-view clustering algorithms: review and cancer benchmark, Nucleic Acids Research, 2018

Pathomic Fusion: H&E Whole Slide + Genomic Profile

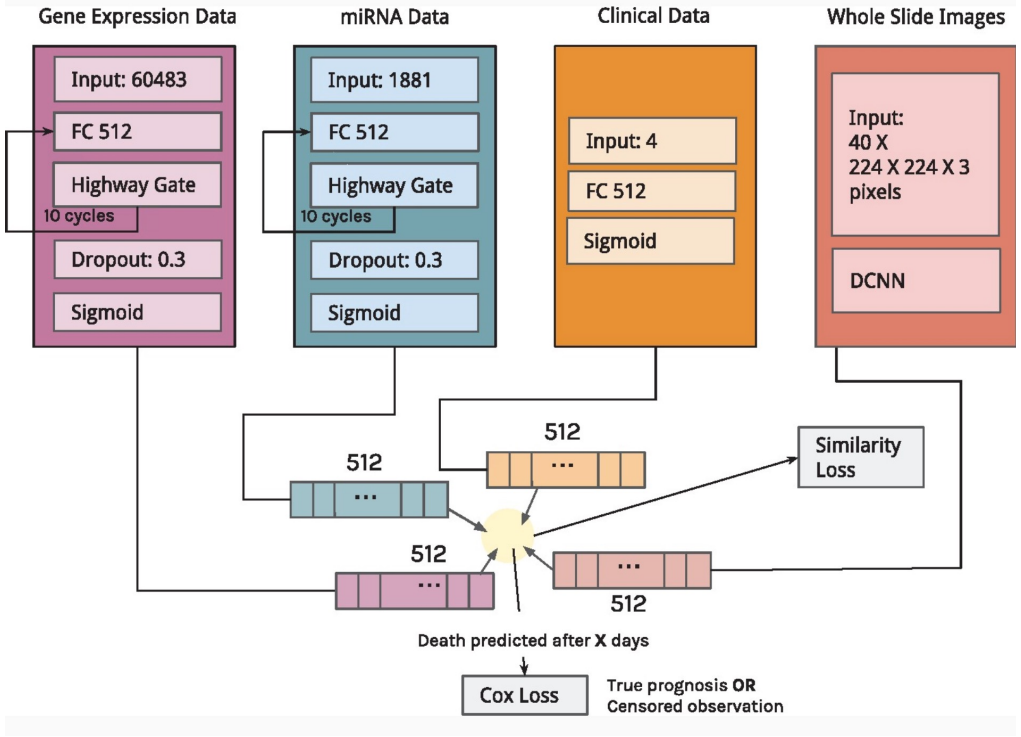




Kronecker product instead of concatenation

Figure source: Richard J. Chen, Ming Y. Lu, Jingwen Wang, Drew F. K. Williamson, Scott J. Rodig, Neal I. Lindeman, Faisal Mahmood, "Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis", IEEE Transactions on Medical Imaging 2021

Pan Cancer Prognosis Prediction using Multimodal Representation



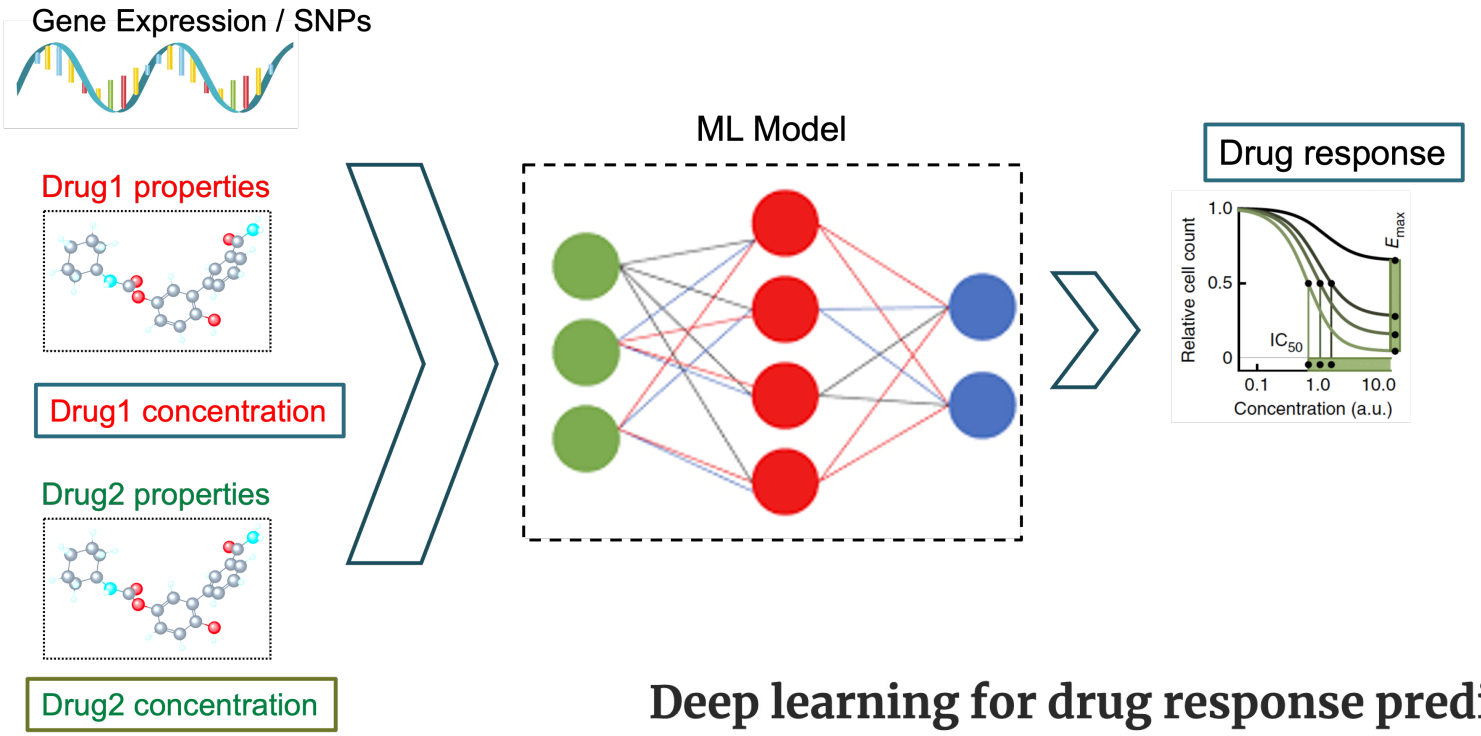
	Clin+miRNA+mRNA+WSI			Clin+miRNA			Clin+mRNA	
Cancer site	**Baseline**	**Multimodal dropout**	**Delta (%)**	**Baseline**	**Multimodal dropout**	**Delta (%)**	**Baseline**	**Multimodal dropout**
BLCA	0.65	0.73	12.6	0.66	0.69	4.4	0.60	0.58
BRCA	0.77	0.79	3.0	0.80	0.80	−0.1	0.57	0.56
CESC	0.73	0.76	4.6	0.77	0.76	−1.2	0.67	0.62
COADREAD	0.72	0.74	3.8	0.78	0.75	−4.8	0.72	0.58
HNSC	0.61	0.67	10.4	0.64	0.64	0.7	0.58	0.55
KICH	0.95	0.93	−2.0	0.82	0.85	3.0	0.80	0.84

Deep learning with multimodal representation for pancancer prognosis prediction

Anika Cheerla, Olivier Gevaert ✉

Bioinformatics, Volume 35, Issue 14, July 2019, Pages i446–i454,
https://doi.org/10.1093/bioinformatics/btz342
Published: 05 July 2019

Drug Response Prediction Using Neural Networks

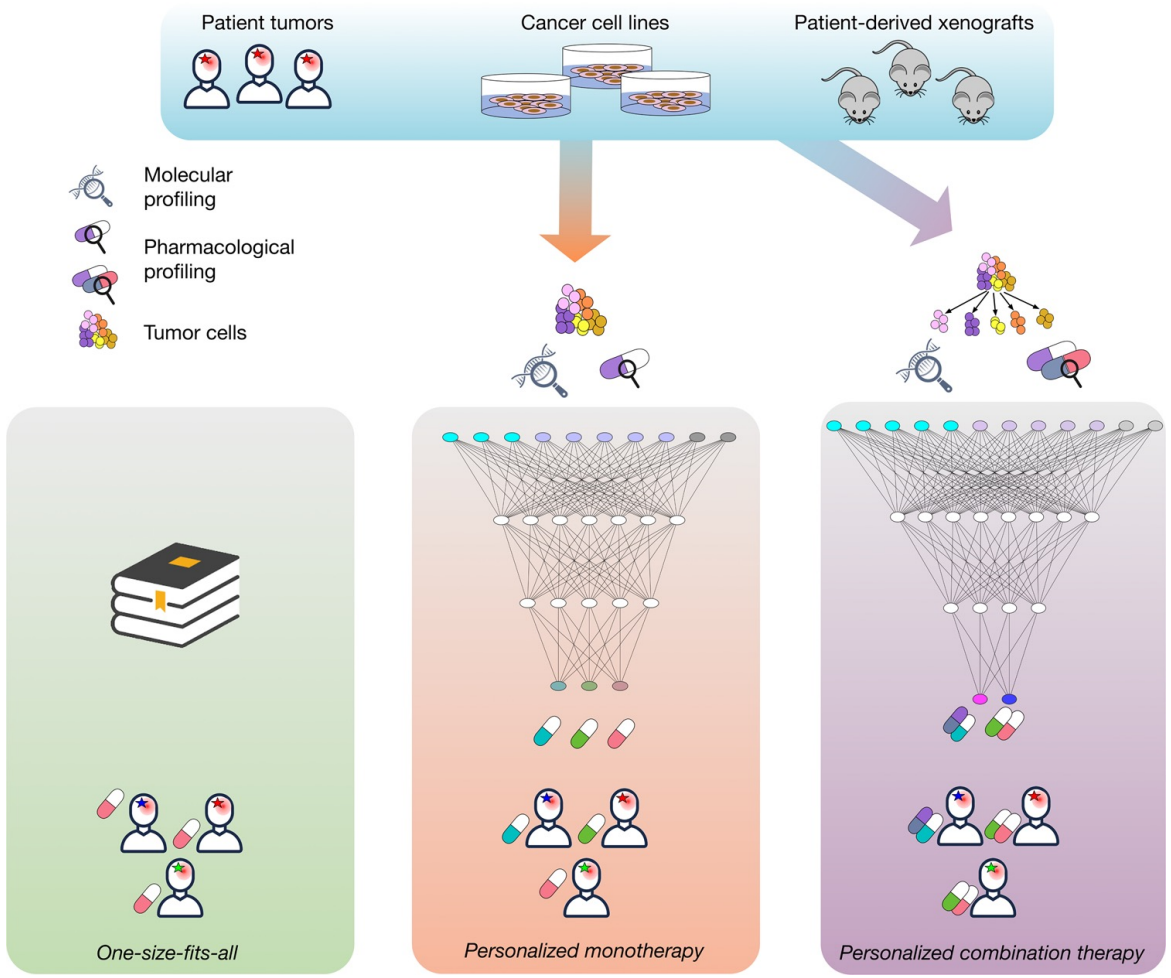


Deep learning for drug response prediction in cancer

Delora Baptista ✉, Pedro G Ferreira, Miguel Rocha

Briefings in Bioinformatics, Volume 22, Issue 1, January 2021, Pages 360–379,
https://doi.org/10.1093/bib/bbz171
Published: 17 January 2020 **Article history** ▾

Example: Predicting Tumor Cell Line Response to Drug Pairs with Deep Learning



- Most patients with cancer are still treated in a one-size-fits-all manner
- A growing number of examples of personalizing monotherapy in practice
- Monotherapies may not be effective due to tumor heterogeneity and acquired drug resistance
- A growing body of work predicting drug synergy and effective drug combinations

Figure source: George Adam et al. Machine learning approaches to drug response prediction: challenges and recent progress, NPJ Precis Oncol, 2020

Combo: Combination Drug Response Predictor

- Developed by computer scientists in the Argonne National Laboratory
- Predict tumor cell line growth to drug pairs using deep learning models (artificial neural networks)
- The workflow consists of 5 main steps:

Step 1	• Collect and preprocess data
Step 2	• Design and build the model
Step 3	• Fit the model to the training data
Step 4	• Use the model to make predictions
Step 5	• Evaluate the accuracy of the predictions

Fangfang Xia, et al. Predicting tumor cell line response to drug pairs with deep learning, BMC Bioinformatics, 2018

Combo: Data

- Data sources
  - Cell line molecular features
    - Gene expression
    - Protein abundance
    - microRNA expression
  - Drug descriptors
    - Dragon
  - Drug pair screen data
    - A subset of NCI-ALMANAC
    - 54 FDA-approved anticancer drugs

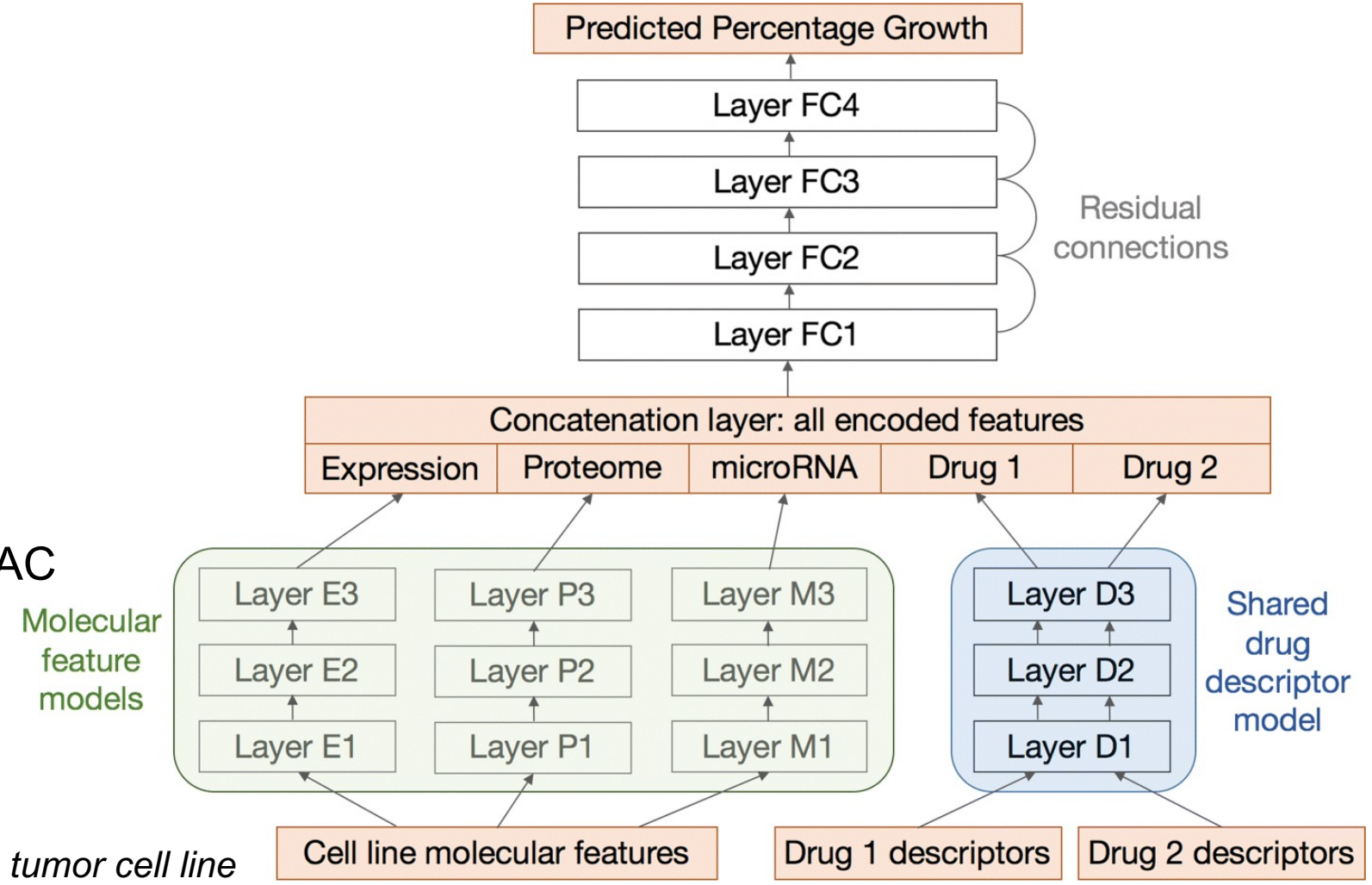


Figure source: Fangfang Xia, et al. Predicting tumor cell line response to drug pairs with deep learning, BMC Bioinformatics, 2018

Combo: Data and Data Preprocessing

NCI-ALMANAC

- Systematically examine the combination efficacy of 104 FDA-approved anticancer drugs
- Catalog in vitro screen results of their pairwise combinations against the NCI-60 cell lines
- Growth inhibition percentage converted to fraction
- ComboScore: differences in observed versus expected growth fractions

Data preprocessing

- log(x+1) transformation
- Imputation and scaling

Susan L. Holbeck, et al. The National Cancer Institute ALMANAC, Cancer Res, 2017

Combo: Design and Implementation

- Neural network Architecture
  - Feature encoding models (3 layers)
    - 3 molecular feature models
    - 1 drug descriptor model
  - Growth prediction model (4 layers)
- Implemented with Keras

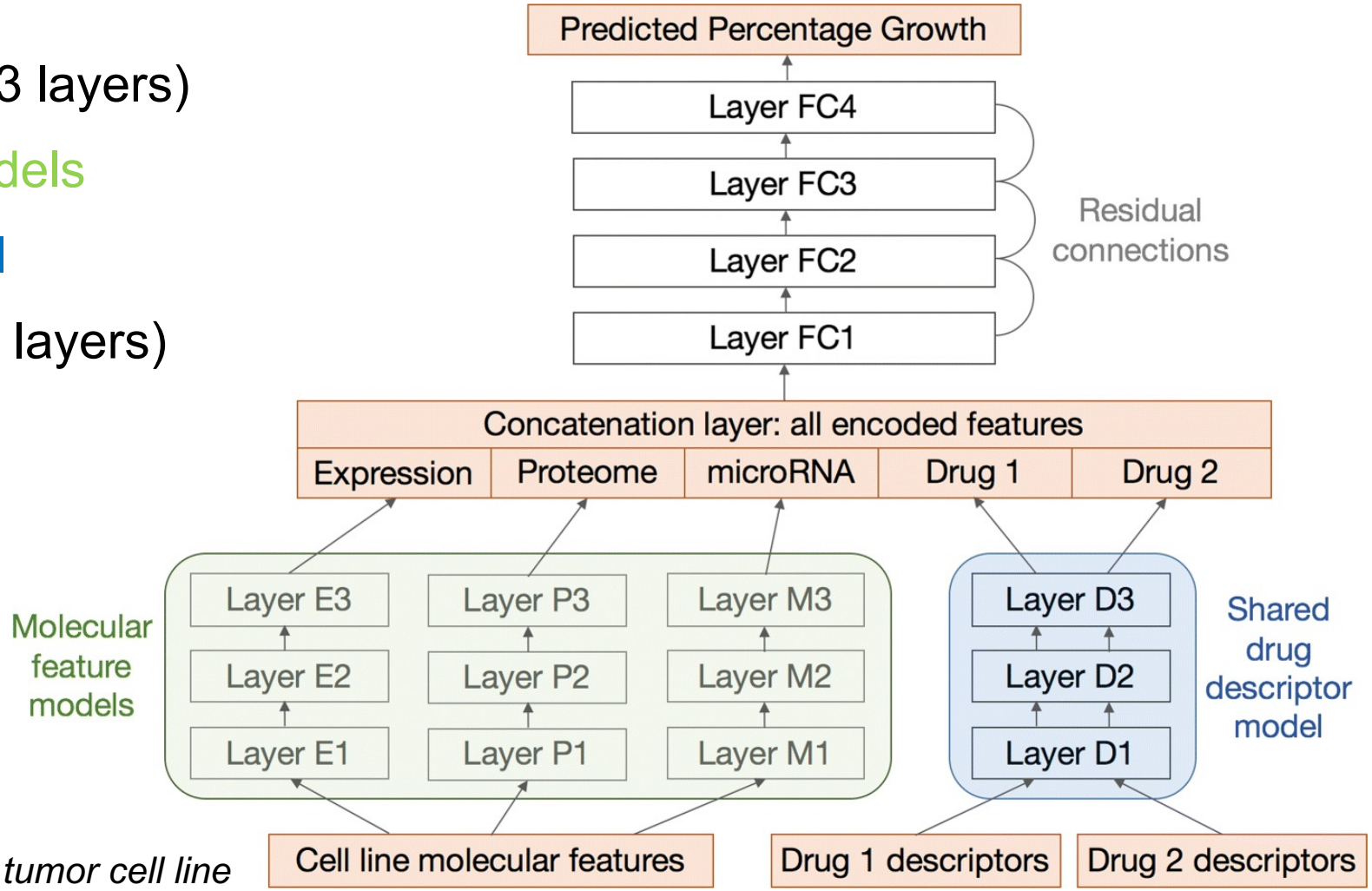


Figure source: Fangfang Xia, et al. Predicting tumor cell line response to drug pairs with deep learning, BMC Bioinformatics, 2018

Combo: Train, Test, and Performance Evaluation

- Performance of the drug pair response model measured with 5-fold cross validation
- Metrics:
  - Mean Squared Error (MSE)
  - Mean Absolute Error (MAE)
  - Coefficient of determination (R^2)
- Models tested on different combinations of feature categories to assess their relative importance

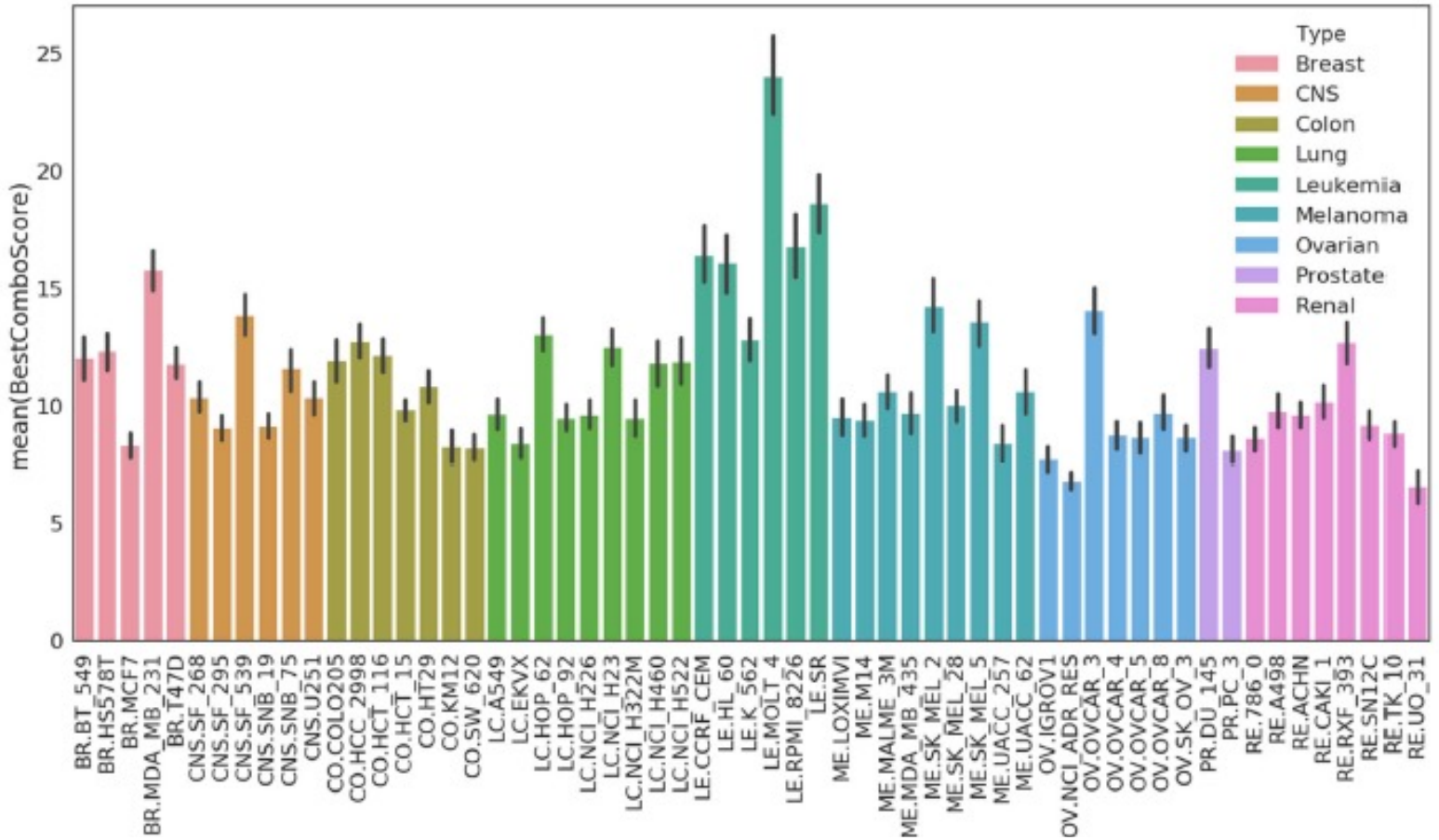
Molecular features	Drug features	MSE	MAE	R^2
Baseline	Baseline	0.5253	0.5709	-1.001
One-hot encoding	One-hot encoding	0.2448	0.3997	0.1269
Gene expression	One-hot encoding	0.2447	0.3999	0.1272
Gene expression	500-dimensional noise	0.2450	0.4008	0.1271
One-hot encoding	Dragon7 descriptors	0.0292	0.1086	0.8892
Proteome	Dragon7 descriptors	0.0303	0.1117	0.8844
microRNA	Dragon7 descriptors	0.0275	0.1050	0.8952
Gene expression	Dragon7 descriptors	0.0180	0.0906	0.9364
Gene expression, microRNA, proteome	Dragon7 descriptors	**0.0158**	**0.0833**	**0.9440**

The boldface row represents the best cross validation

Table source: Fangfang Xia, et al. Predicting tumor cell line response to drug pairs with deep learning, BMC Bioinformatics, 2018

Combo: Performance Evaluation

Cell line views of drug combination effect

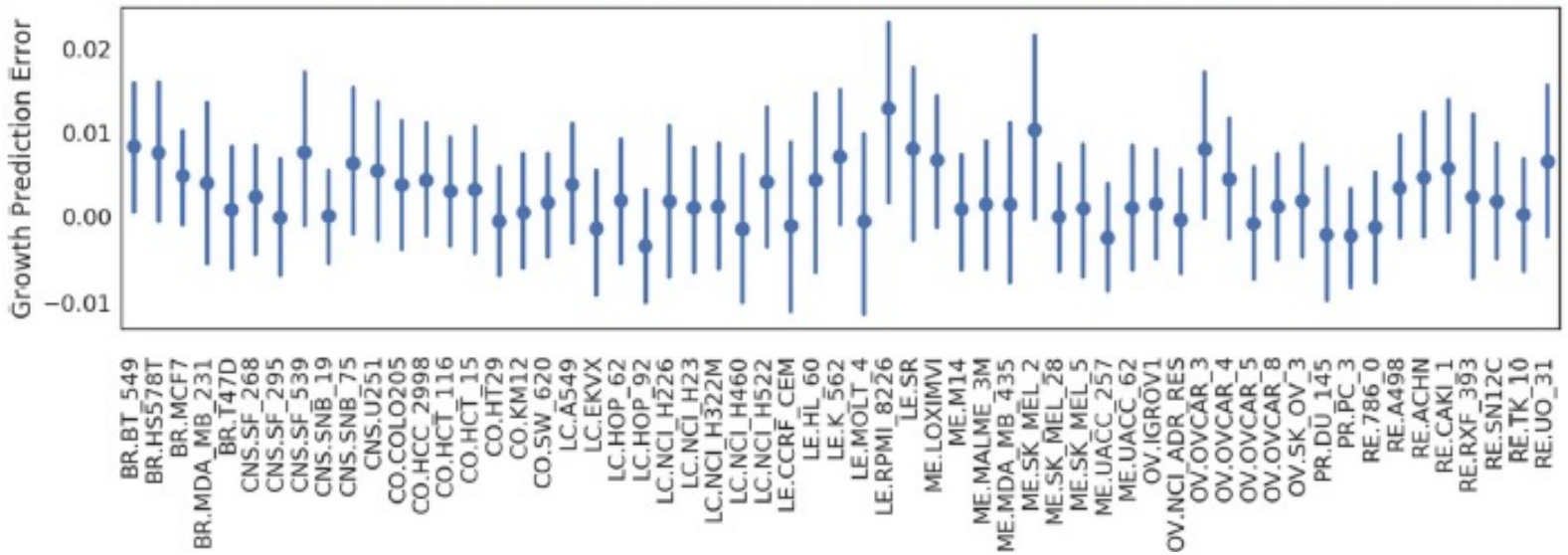


- Leukemia cell lines have drug pairs with most enhanced activity on average

Figure source: Fangfang Xia, et al. Predicting tumor cell line response to drug pairs with deep learning, BMC Bioinformatics, 2021

Combo: Performance Evaluation

Cell line views of growth prediction error



- Growth fraction prediction errors mostly cancel out near 0

Figure source: Fangfang Xia, et al. Predicting tumor cell line response to drug pairs with deep learning, BMC Bioinformatics, 2021

Combo: Performance Evaluation

Cell line views of growth ranking error

- 75% of the cell lines have the predicted top 100 list at least 75% correct

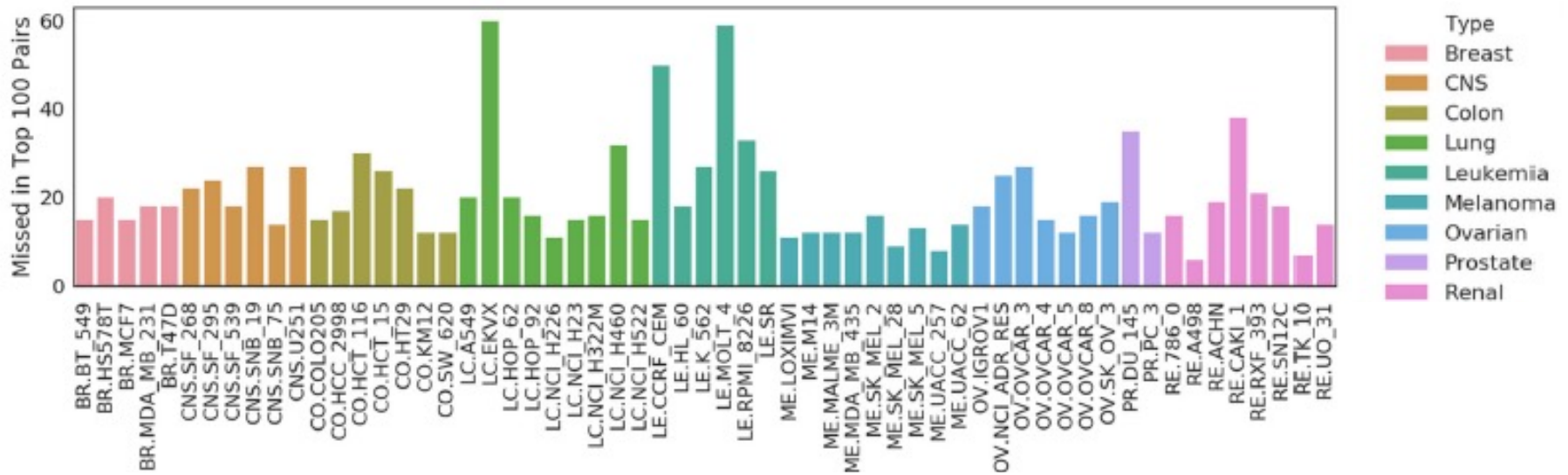


Figure source: Fangfang Xia, et al. Predicting tumor cell line response to drug pairs with deep learning, BMC Bioinformatics, 2021

Combo: Application

Drug pairs with top combination scores across cell lines

Rank	Drug pair	Predicted drug pair
1	(idarubicin, amifostine)	(idarubicin, amifostine)
2	(epirubicin, amifostine)	(epirubicin, amifostine)
3	(idarubicin, epirubicin)	(idarubicin, epirubicin)
4	(idarubicin, covidarabine)	(idarubicin, covidarabine)
5	(epirubicin, idarubicin)	(epirubicin, idarubicin)
6	(idarubicin, imiquimod)	(idarubicin, imiquimod)
7	(epirubicin, imiquimod)	(epirubicin, imiquimod)
8	(epirubicin, dexrazoxane)	(epirubicin, covidarabine)
9	(epirubicin, covidarabine)	*(epirubicin, cyclophosphamide)*
10	(idarubicin, allopurinol)	(idarubicin, allopurinol)

Table source: Fangfang Xia, et al. Predicting tumor cell line response to drug pairs with deep learning, BMC Bioinformatics, 2018

- An important use of drug response models is in high throughput virtual screening
- A list of top 10 drug pairs across cell lines ranked using the ComboScore calculated from *PREDICTED growth data*
- 80% identical, with the predicted version missing (epirubicin, dexrazoxane) and overpredicting (epirubicin, cyclophosphamide)

Summary

- Opportunities in predicting diagnosis, prognosis, clustering, and drug response that would take advantage of complementary and redundant data.
- Challenges still exist, and no single method can overcome them.
- Benefit from the integration of the multimodal data to answer key questions in biomedical research.
- A need for standard benchmarks to compare models, and systematic ways to collect machine learning ready data.

References

- Yifeng Li, Fang-Xiang Wu, Alioune Ngom, A review on machine learning principles for multi-view biological data integration, *Briefings in Bioinformatics*, Volume 19, Issue 2, March 2018, Pages 325–340, https://doi.org/10.1093/bib/bbw113
- Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal machine learning: A survey and taxonomy." *IEEE transactions on pattern analysis and machine intelligence* 41.2 (2018): 423-443
- George Adam et al. Machine learning approaches to drug response prediction: challenges and recent progress, NPJ Precis Oncol, 2020
- Fangfang Xia, et al. Predicting tumor cell line response to drug pairs with deep learning, BMC Bioinformatics, 2018
- Susan L. Holbeck, et al. The National Cancer Institute ALMANAC, Cancer Res, 2017

Acknowledgment

- **BTEP:**
  - Amy Stonelake

- **Prior speakers at the seminar series:**
  - Jonathan Allen
  - Gianluca Pegoraro
  - G Tom Brown
  - Avantika Lal