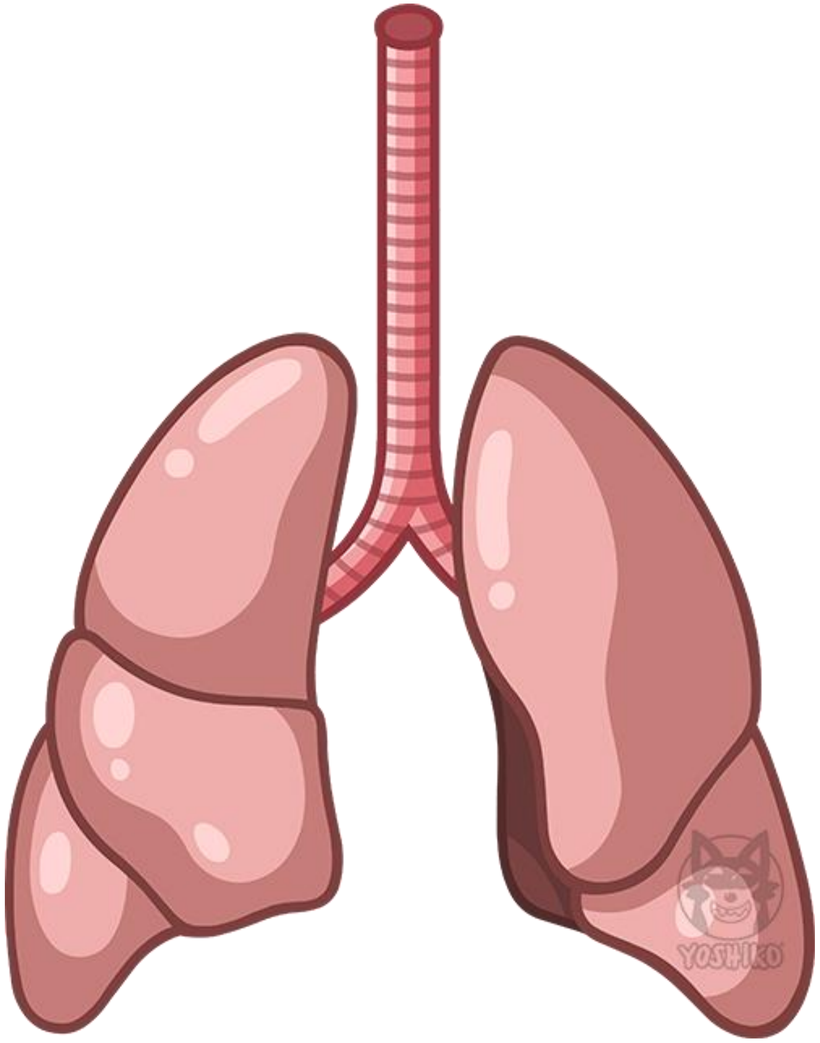
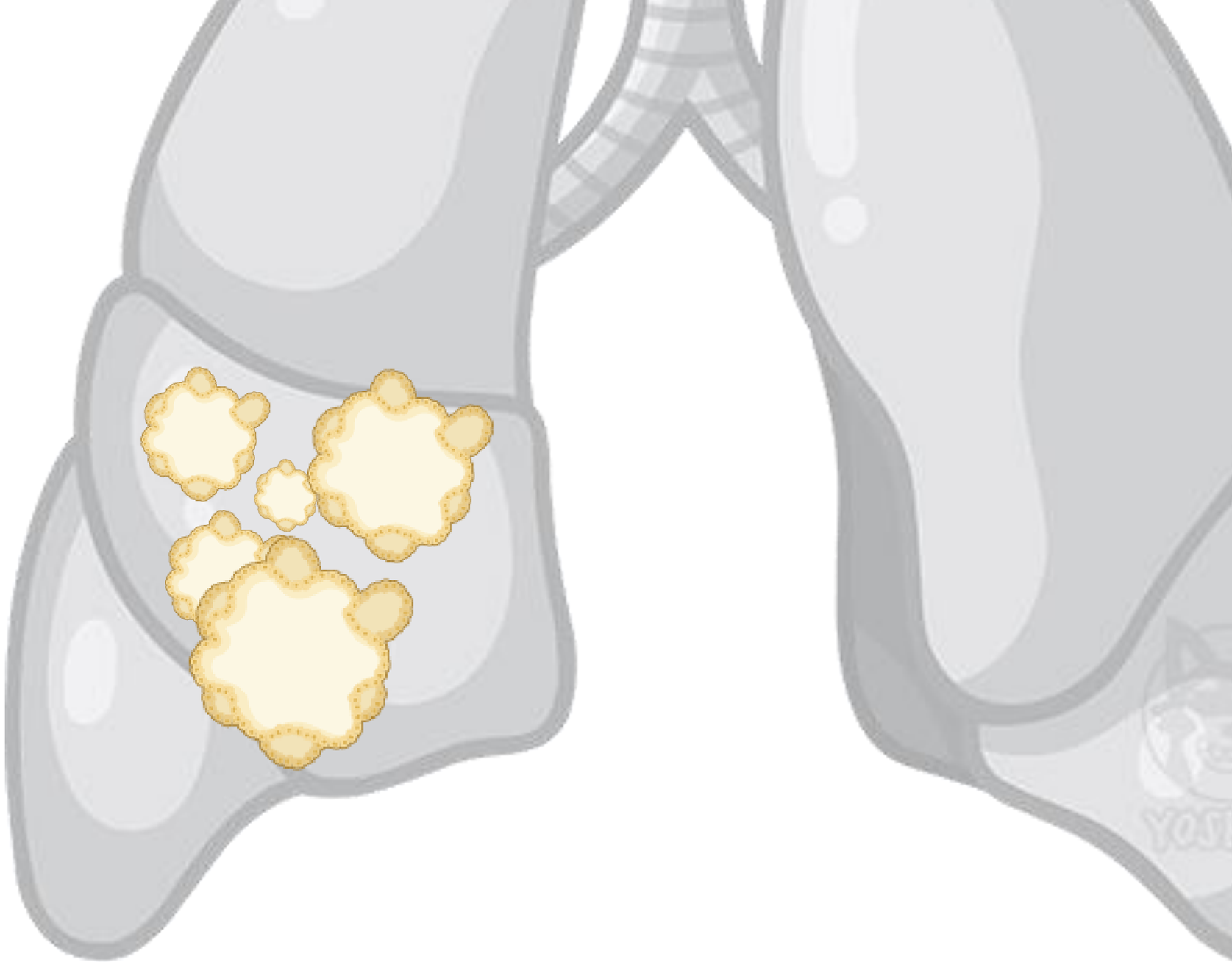
Predicting molecular biomarkers for cancer therapeutic response

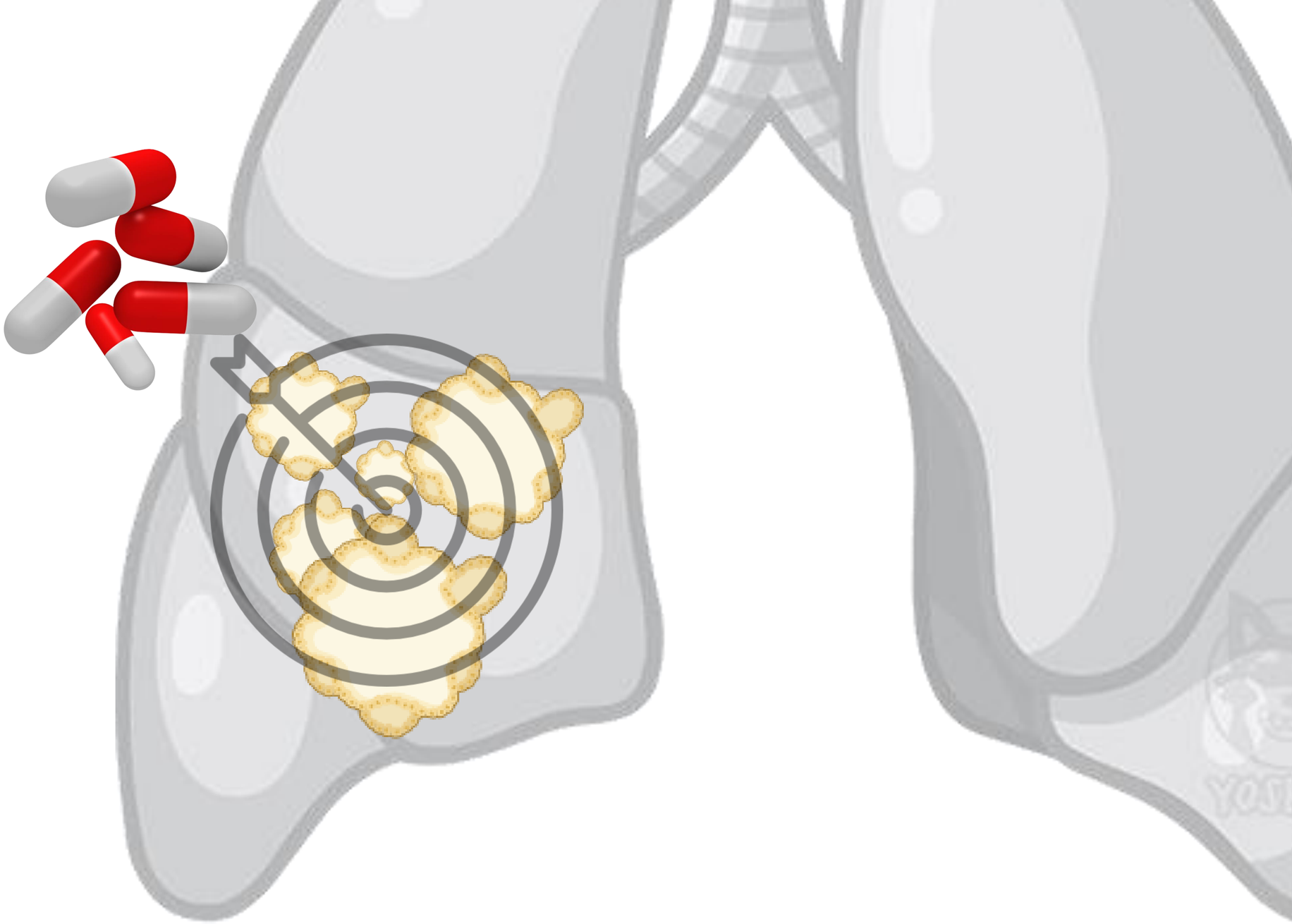
Aubhishek Zaman, PhD


YOSHIKO

Lung Cancer

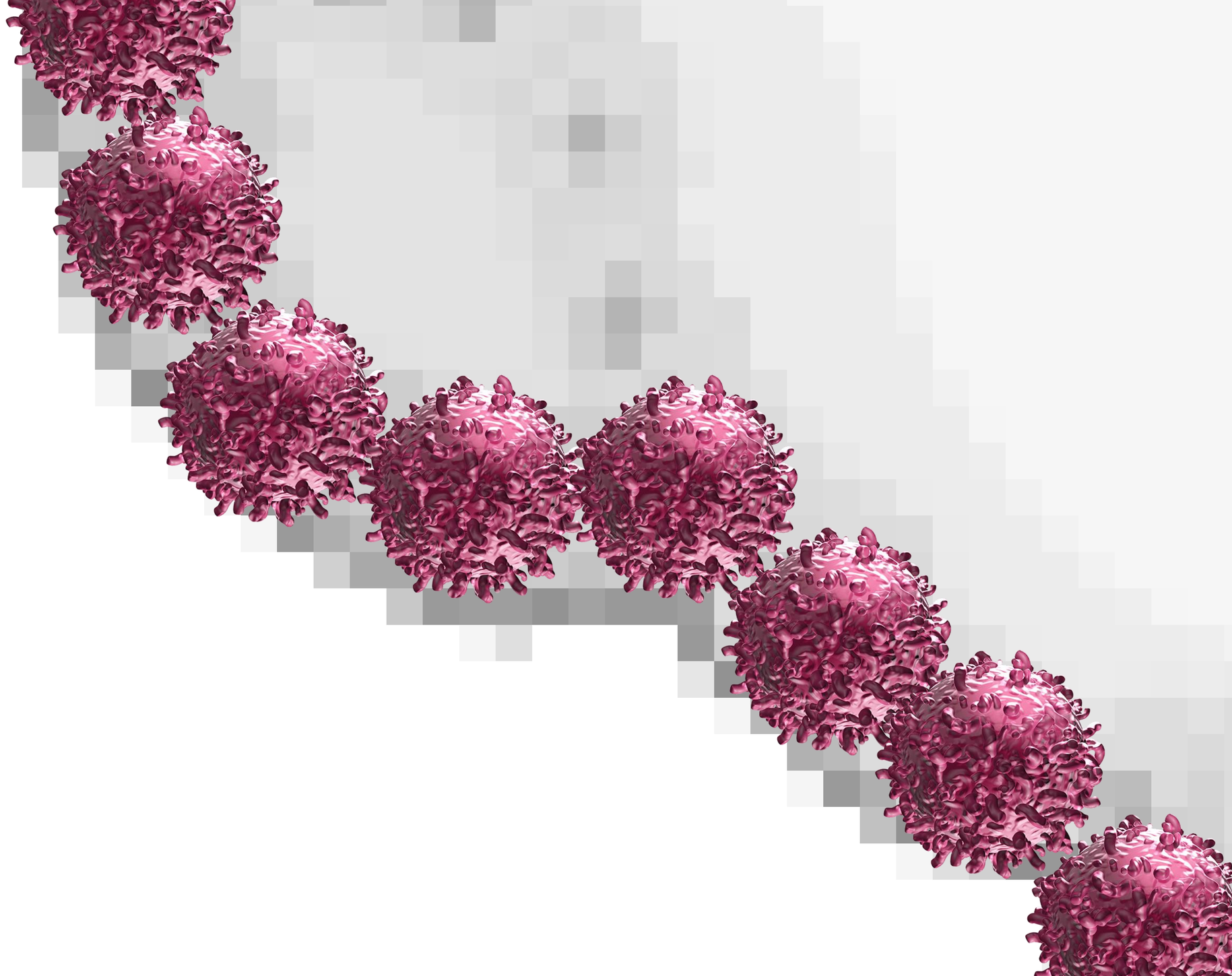
- leading cause of Cancer related death in the United States
- aggressive solid with molecular difference from normal cells
- cancer specific molecular difference is targetable

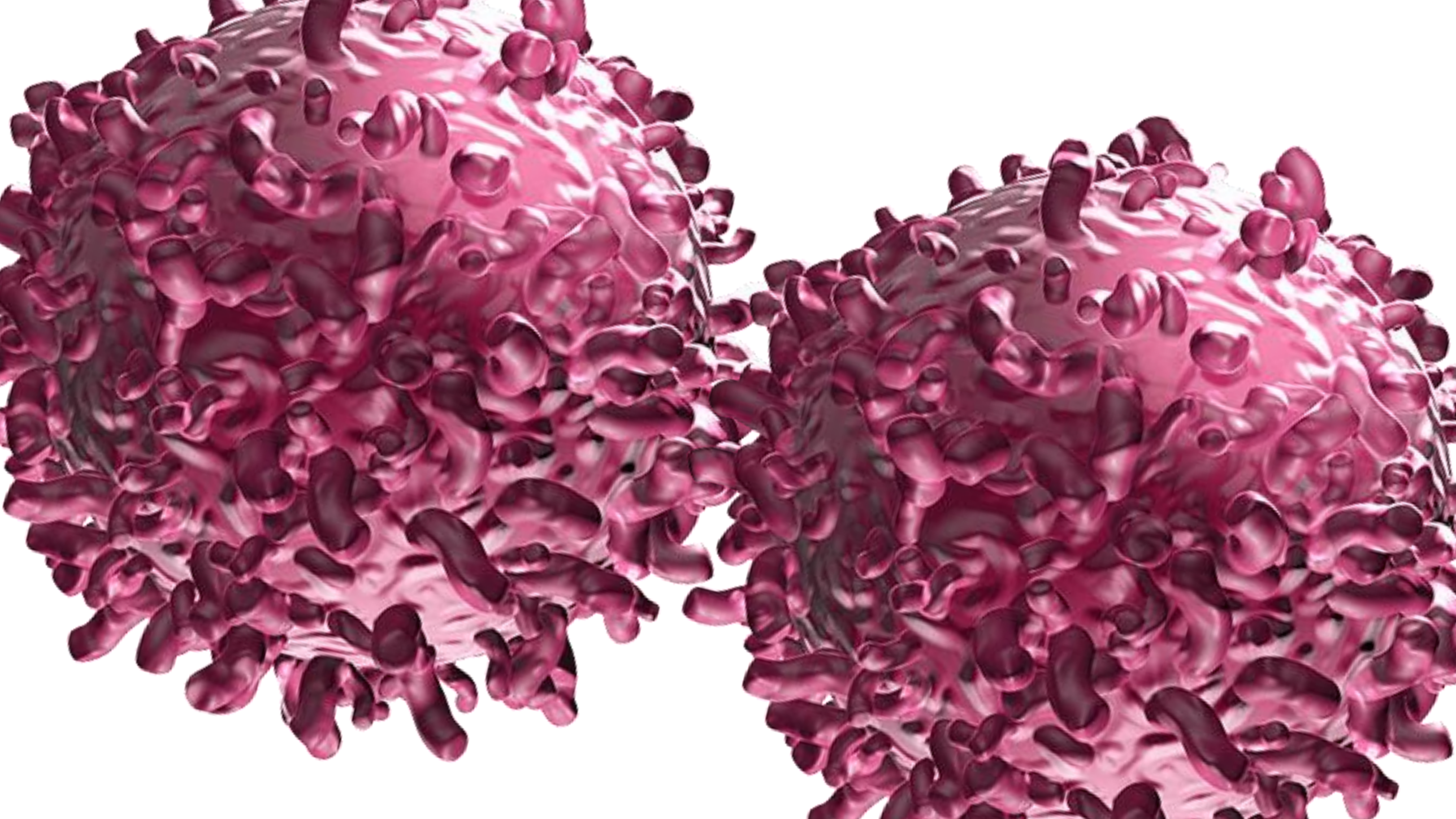


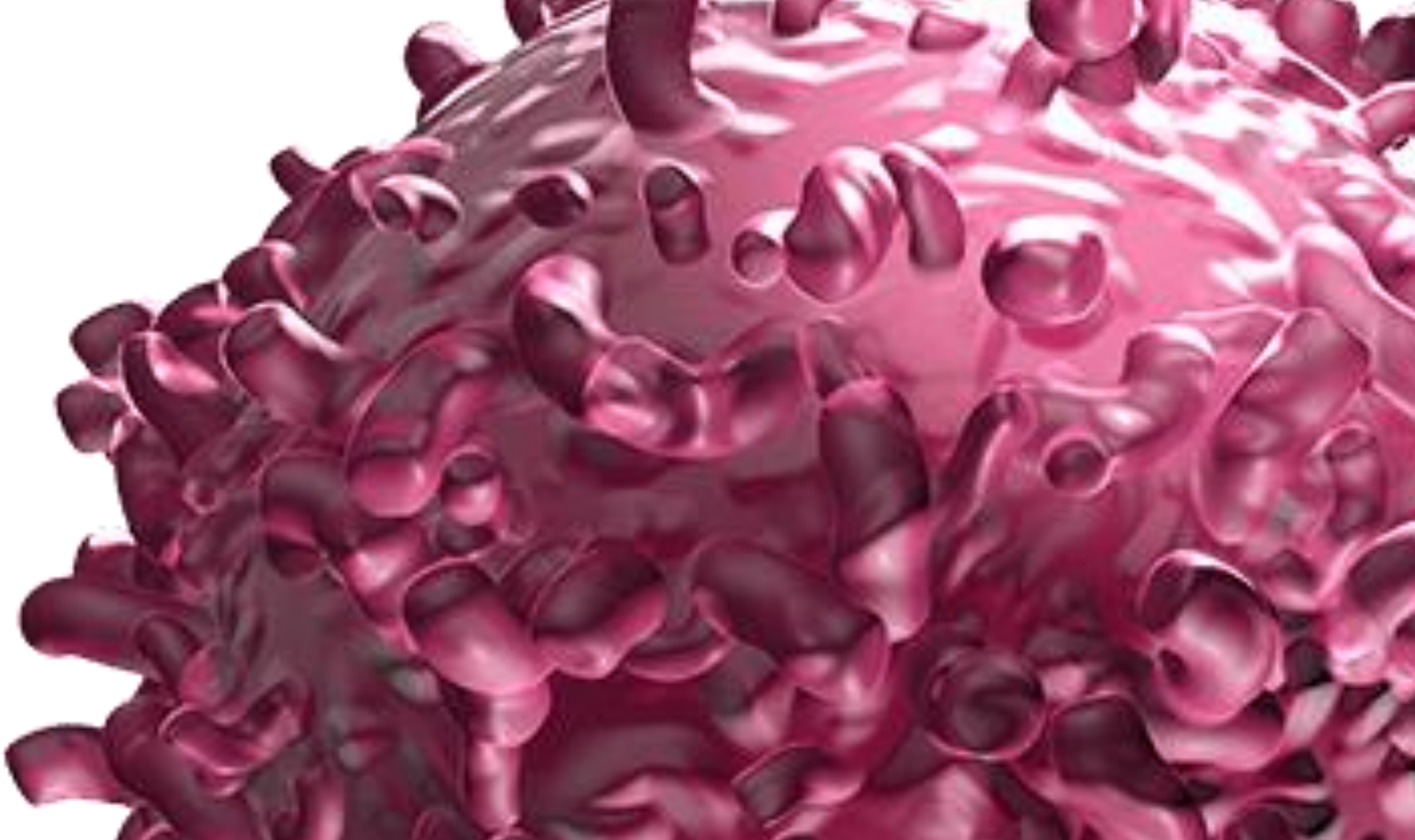


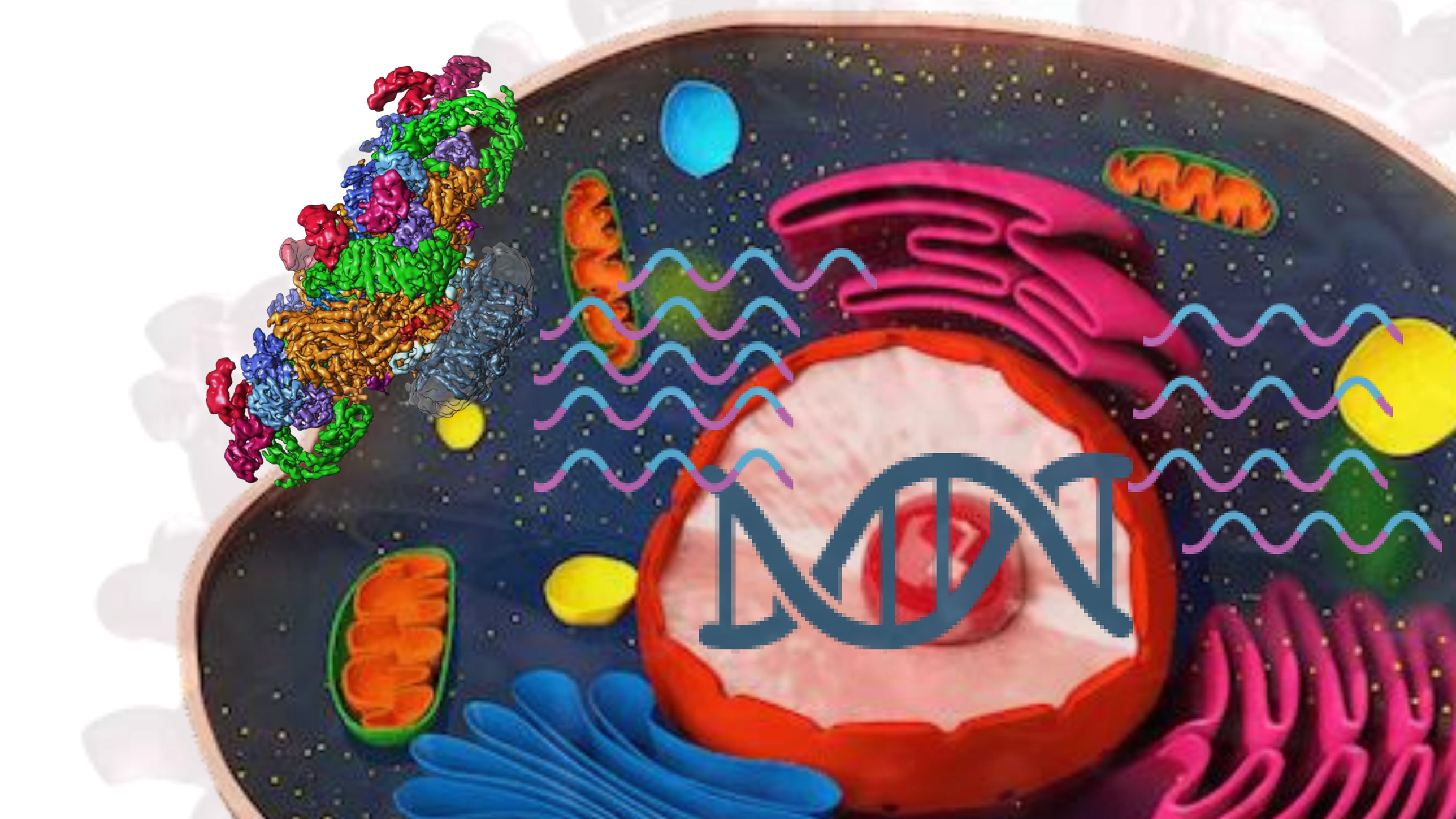
Molecular Features

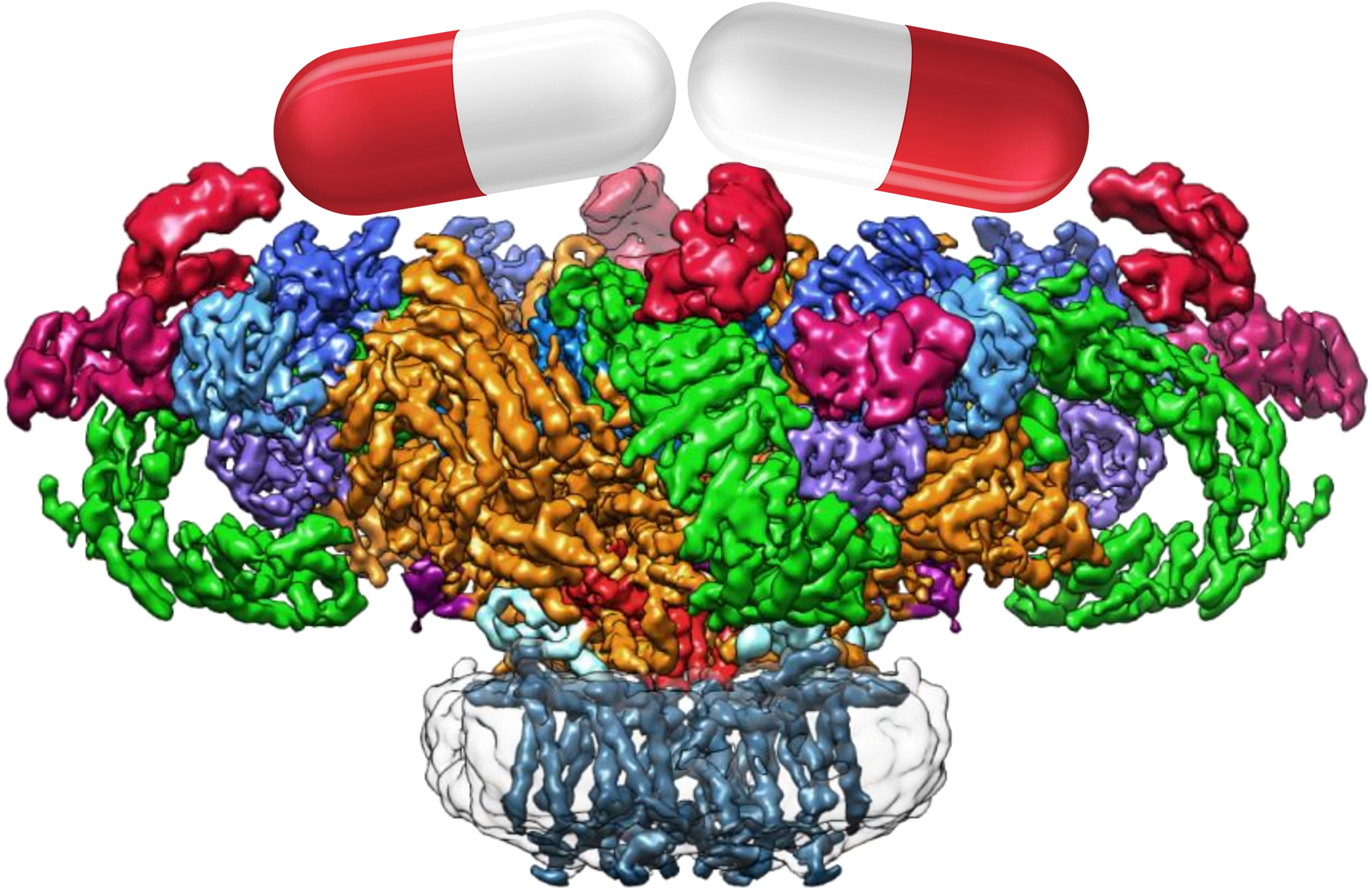
- Human cells (also cancer cells) have DNA, RNA, Portein, Carbohydrate and Lipid intracellular macromolecule
- Proteins are cells' functional coins

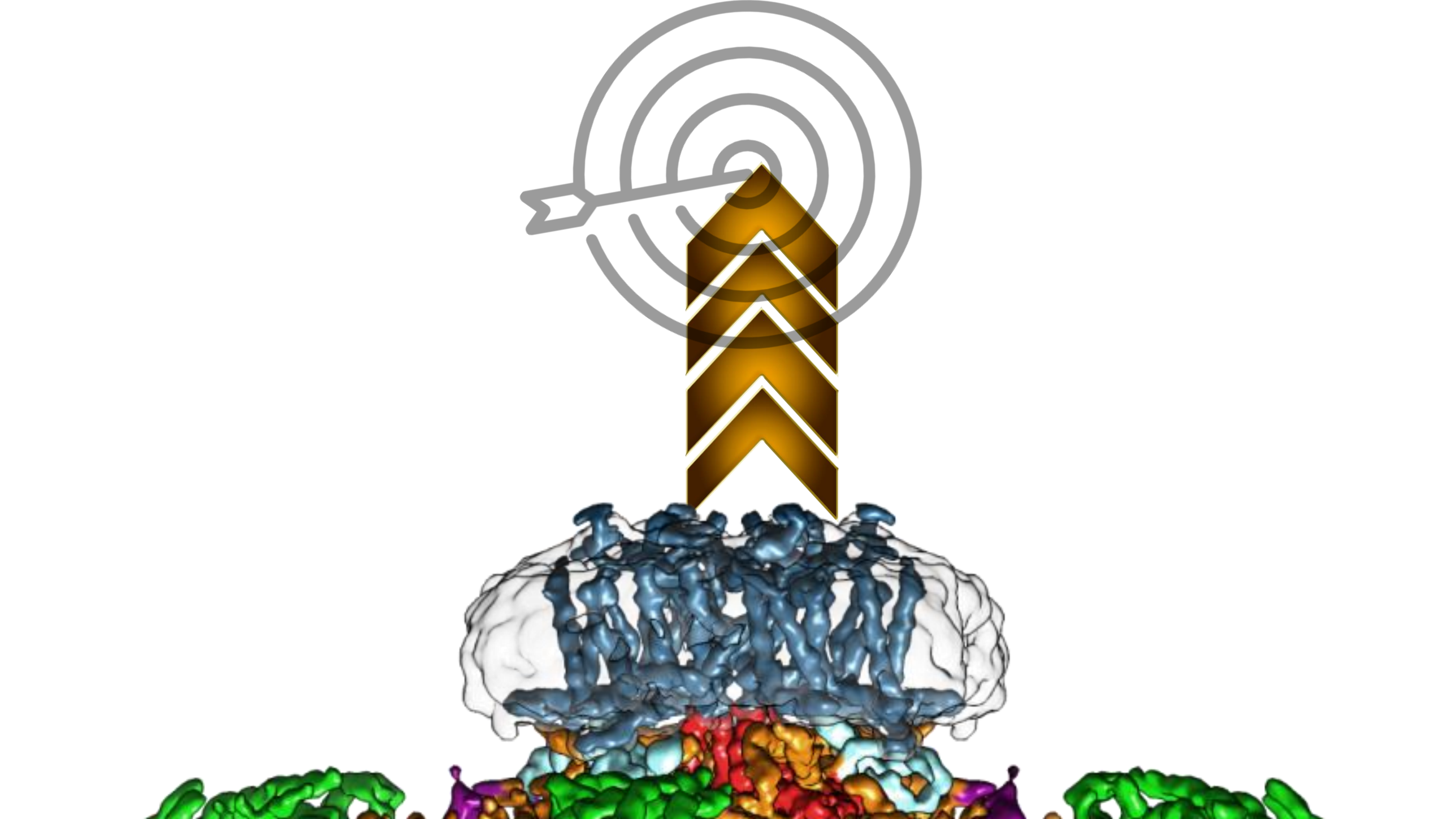






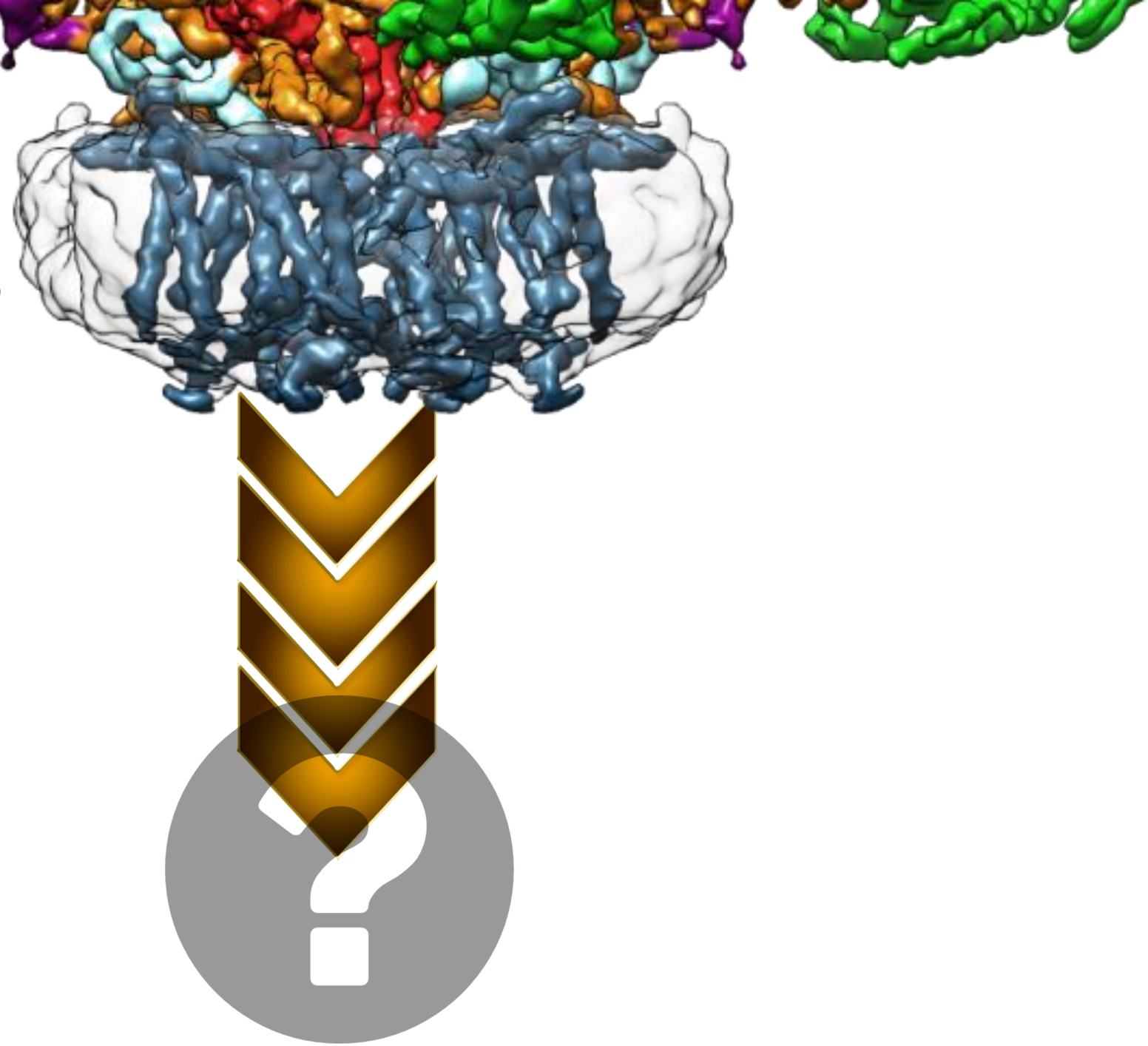






Targeted therapeutics

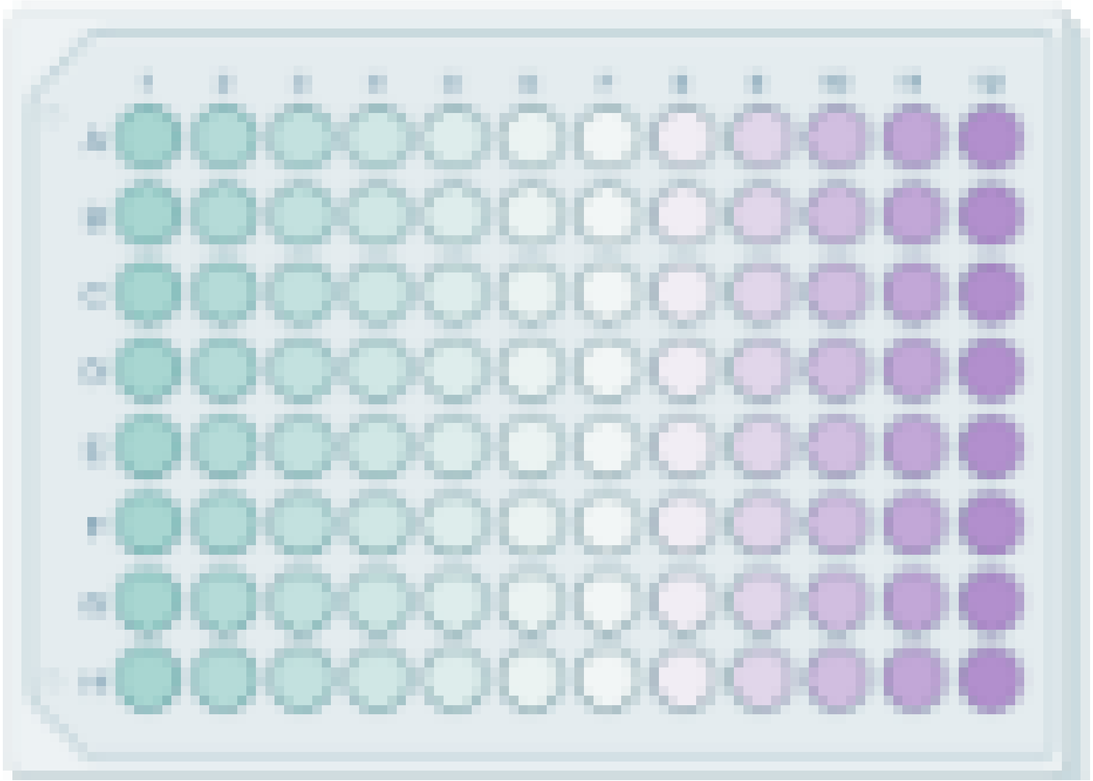
- Specific small molecules can be toxic in a cancer targeted manner
- Candidate small molecule to drug approval requires understanding of exact mechanism of action (MOA) for toxicity



Methodologies

1. Bioactive compound screenings:

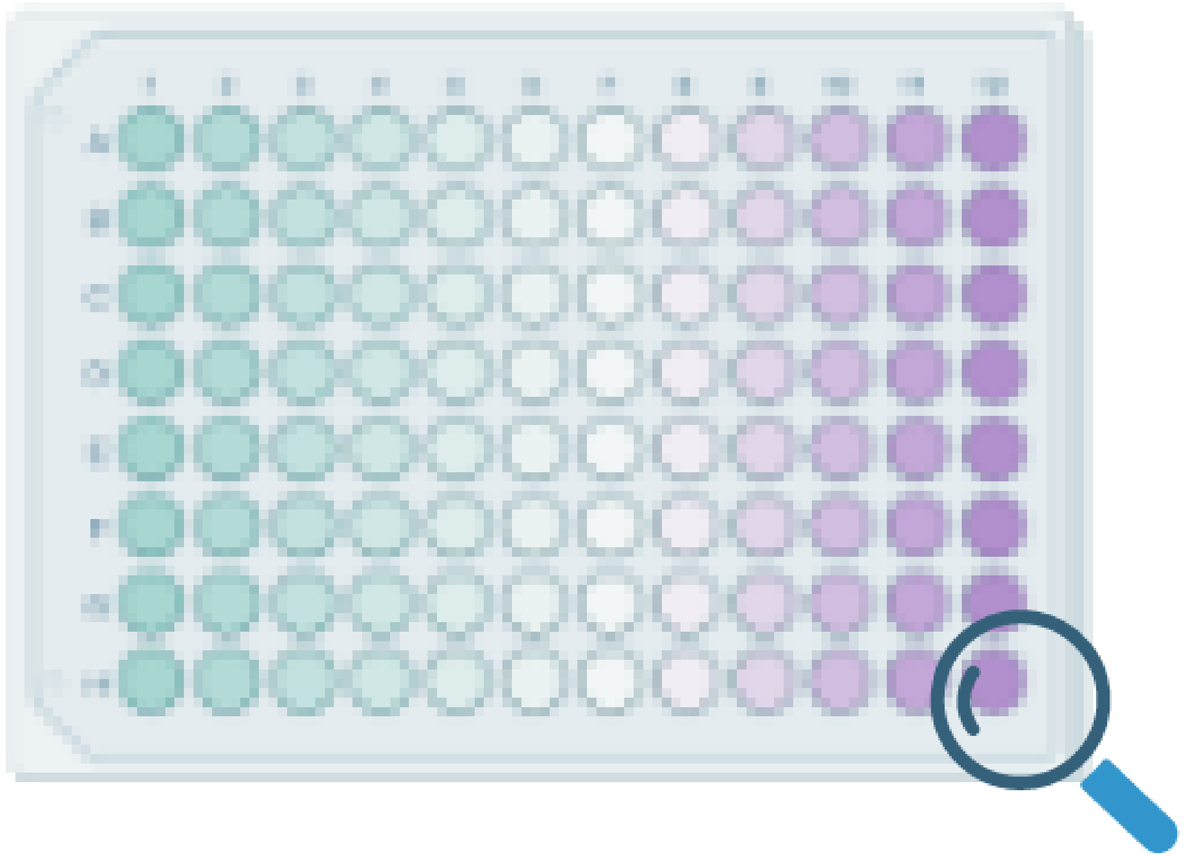
- Easy to perform
- Combinatorial compound libraries are easy to scale



Methodologies

1. Bioactive compound screenings:

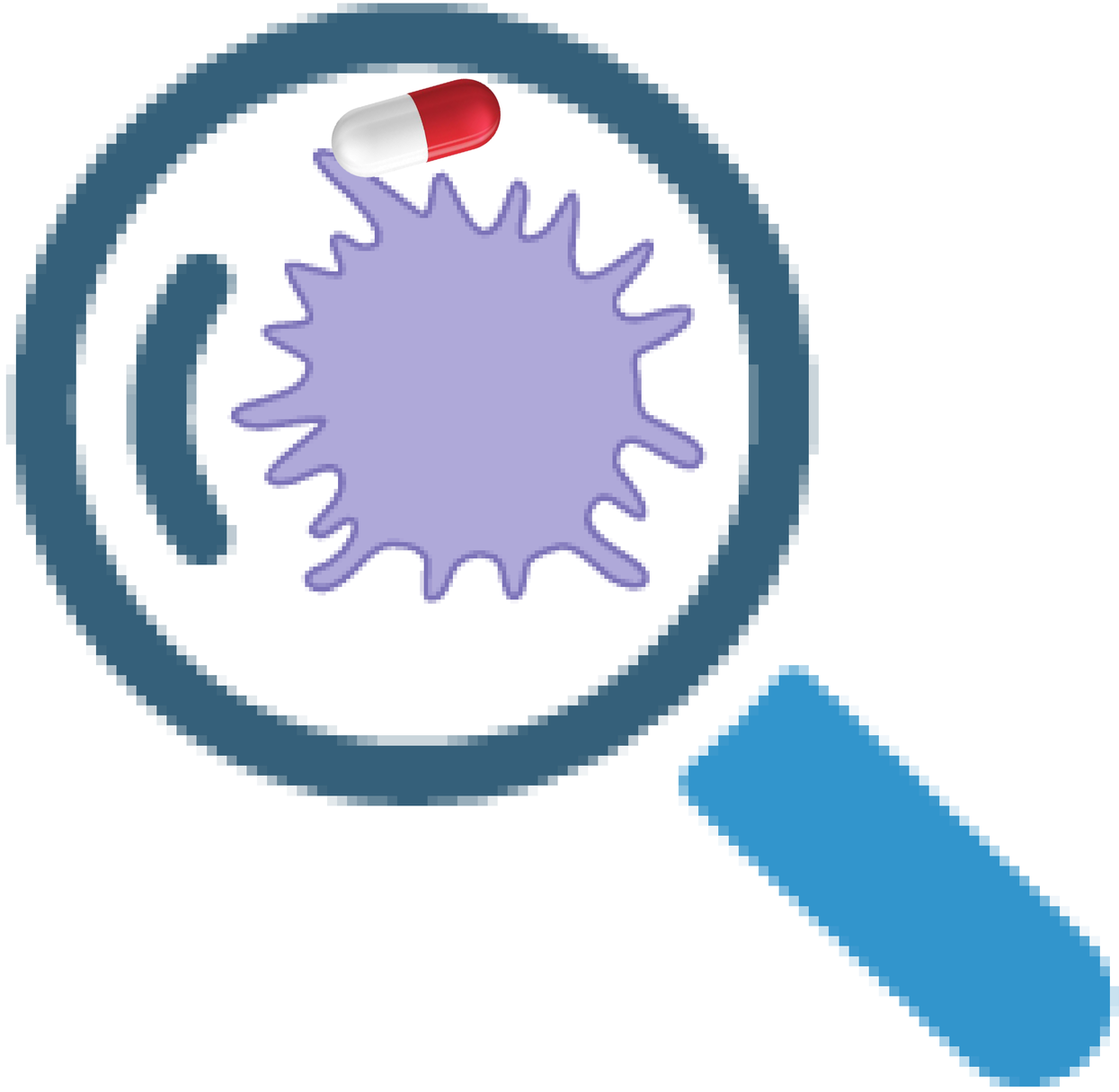
- Easy to perform
- Combinatorial compound libraries are easy to scale



Methodologies

1. Bioactive compound screenings:

- Easy to perform
- Combinatorial compound libraries are easy to scale



Methodologies

1. Bioactive compound screenings:

- Easy to perform
- Combinatorial compound libraries are easy to scale

Methodologies

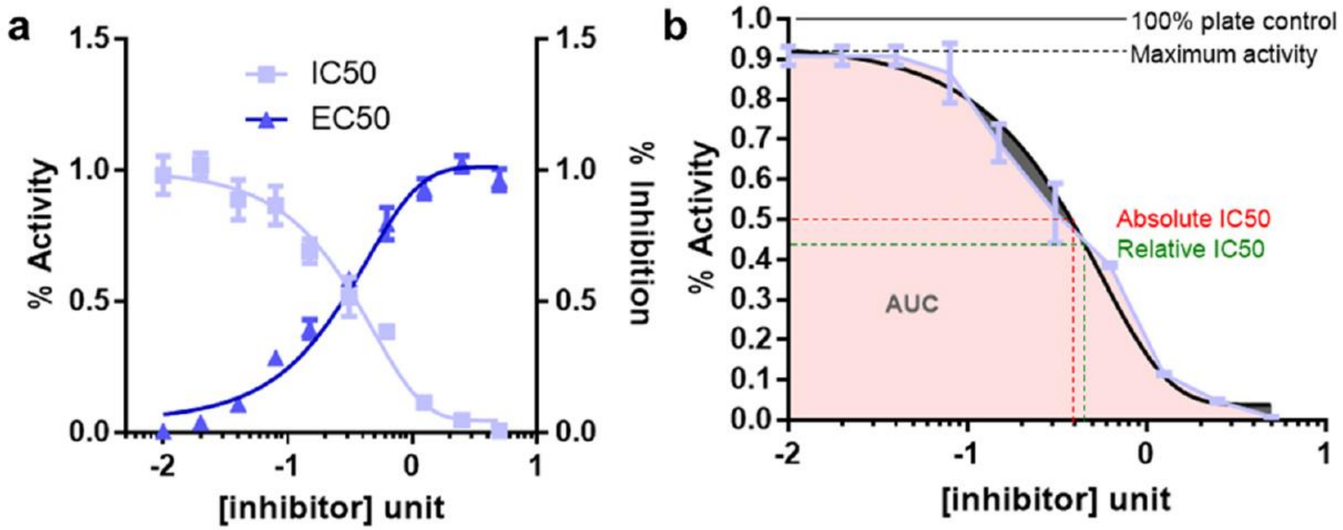
1. Bioactive compound screenings:

- Easy to perform
- Combinatorial compound libraries are easy to scale
- MOA studies answer how cells are sensitive to the drugs

Methodologies

Compare sensitive vs resistant cells:

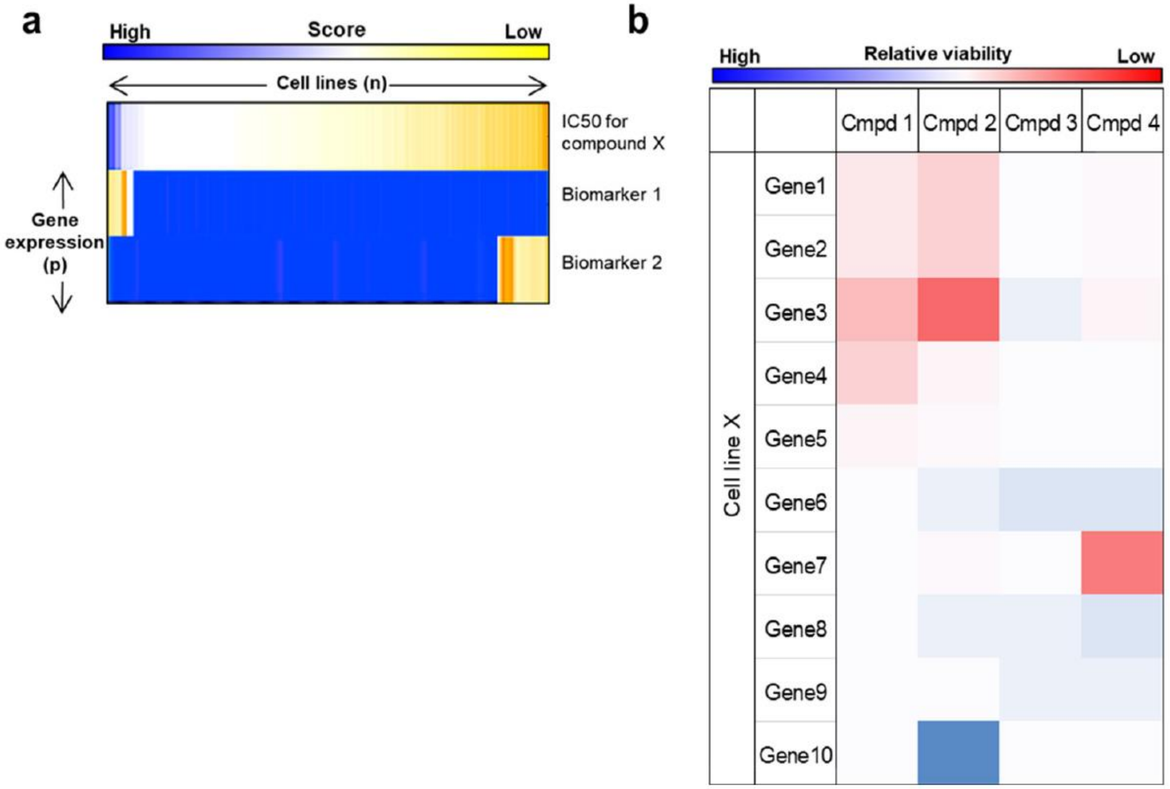
- Gene set enrichment analysis
- Weights are much less comprehensive
- Less quantitative more qualitative
- A group of gene prediction



Methodologies

Multiple elastic net regression (MLR):

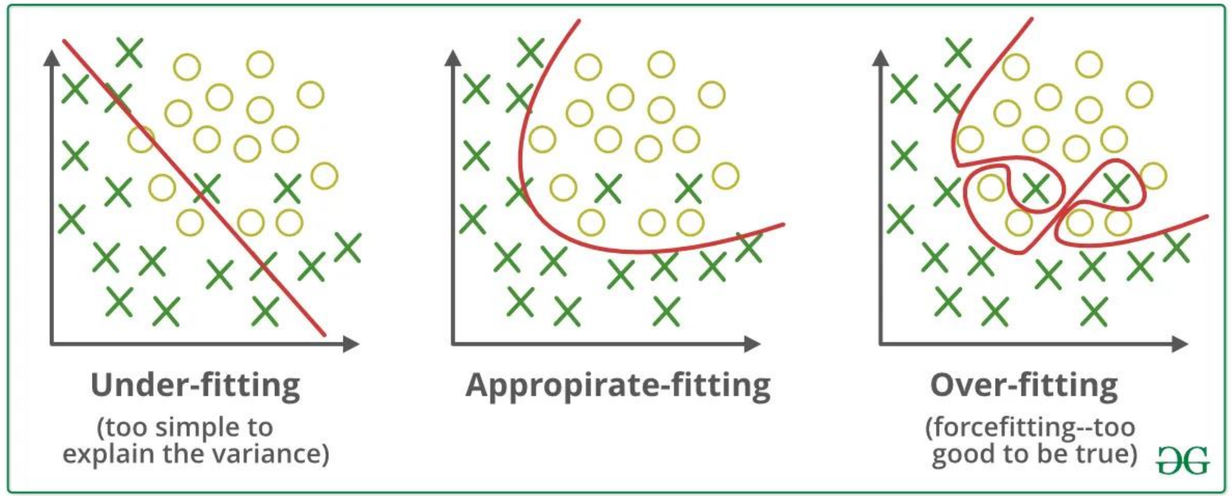
- Quantitative
- Deep learning methodology
- Train and test needs to be optimized often
- Individual biomarker nomination
- Compared to OMICS features
- N/P ratio problem
- Results in overfitting: the solution is unique for the dataset and fails on unknown test data

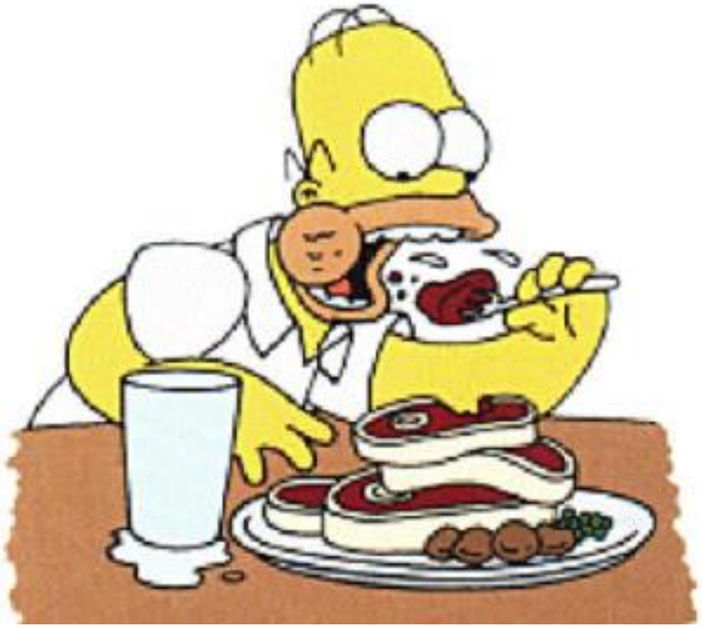


Elastic net regularization curbs overfitting

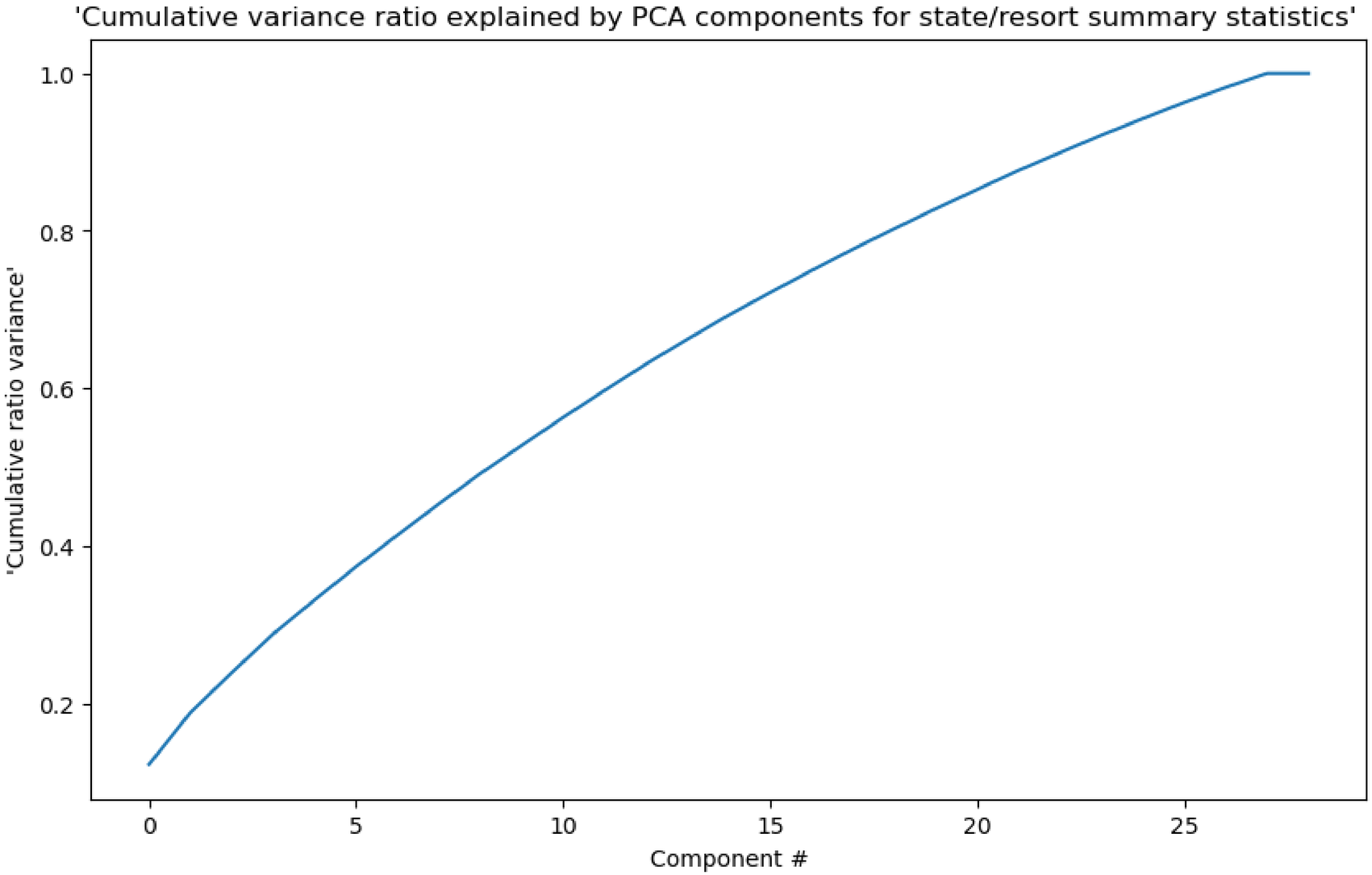
- L1 (Lasso) regularization: penalty term that lowers the complexity by minimizing the number of non-zero coefficients (note the w term being zero nullifies wx)

- L2 (Ridge) regularization: penalty term that lowers the complexity by making the weights more homogenous (note the squared term for weights, which are fractions)

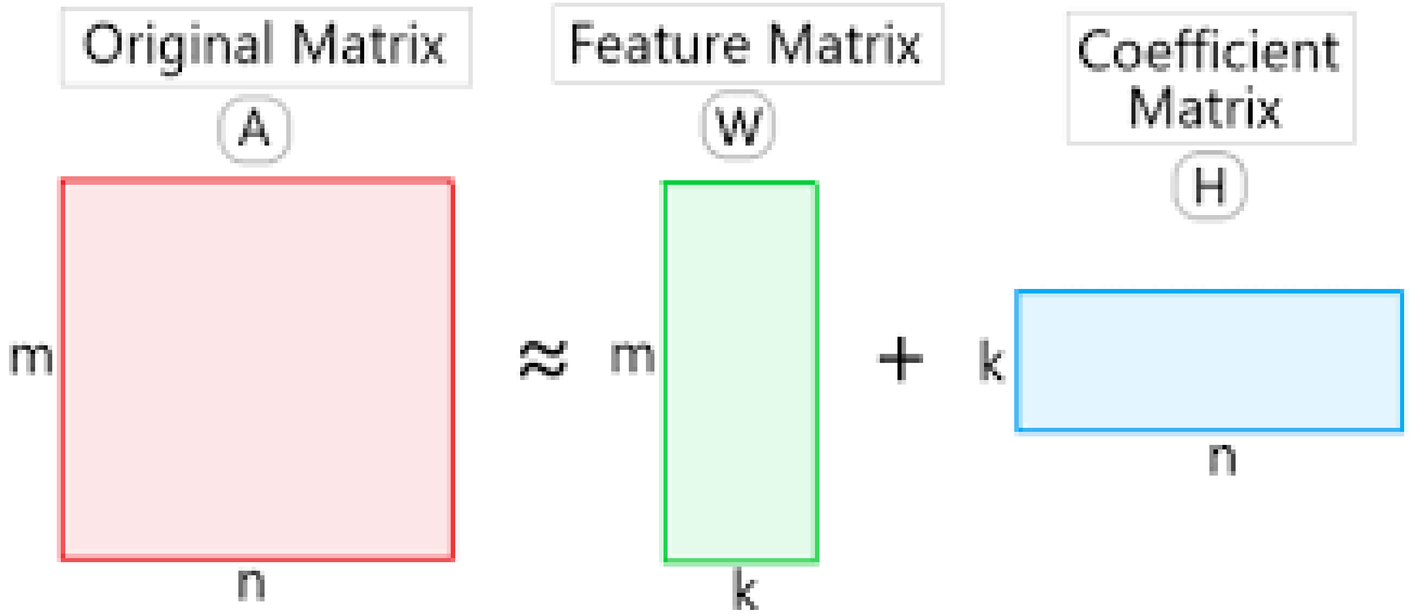




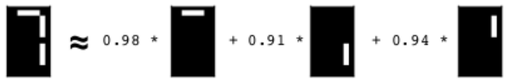
Data complexity is not contributed by a handful of exempler components/ outliers



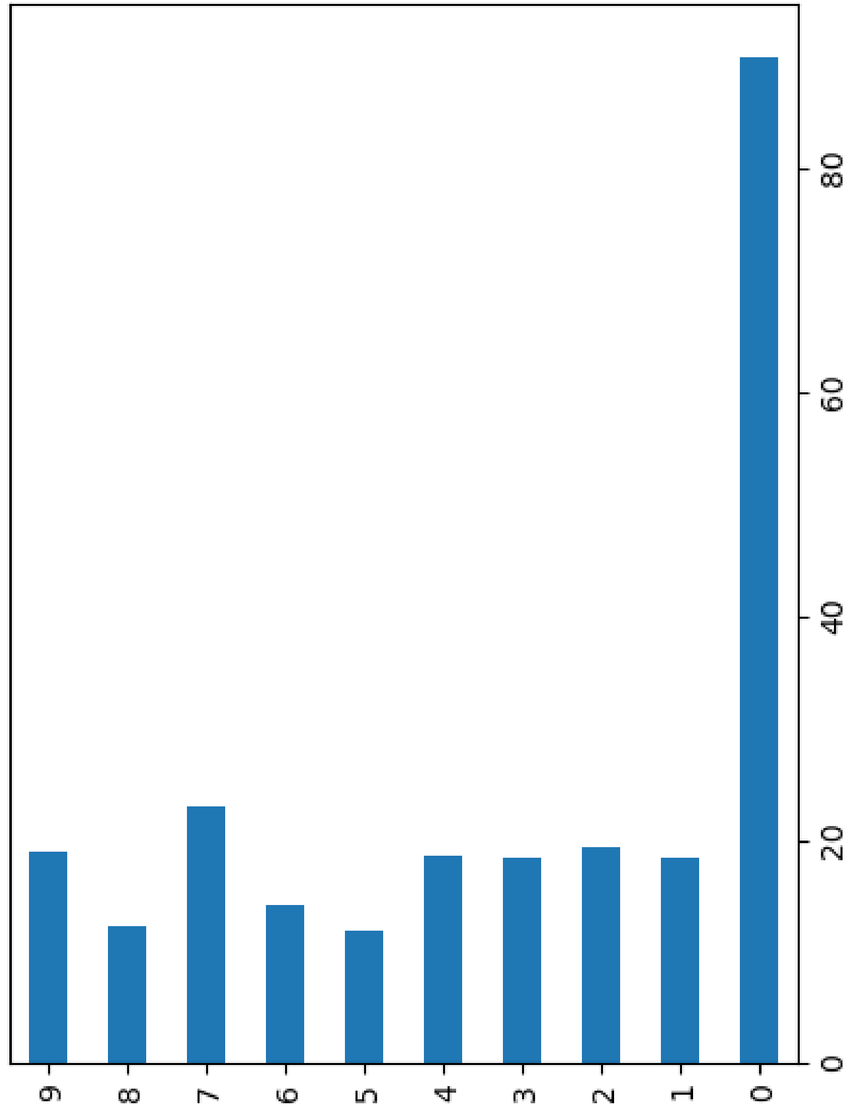
Alternative modeling: NMF (non-negative matrix factorization)



- NMF expresses images as combinations of patterns

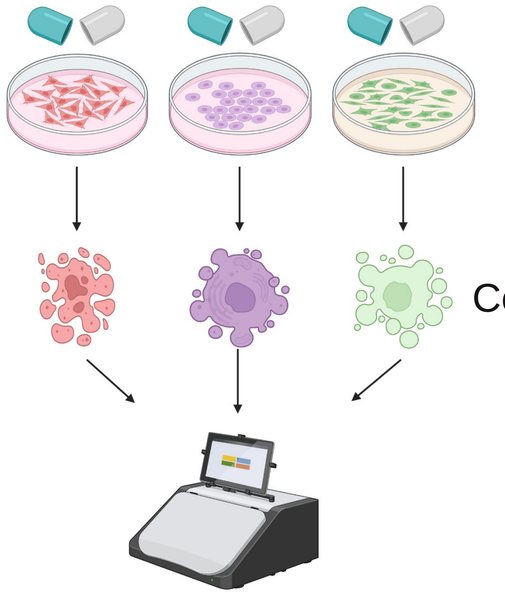


- Advantages: multiple OMICS data can be integrated



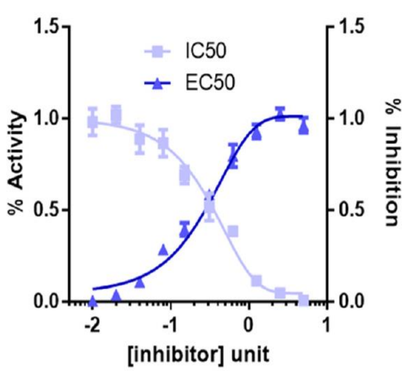
Data wrangling: dataset merging, cleanup and structuring

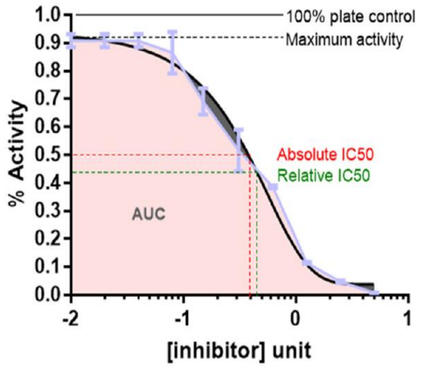
Drug response matrix



Cell death

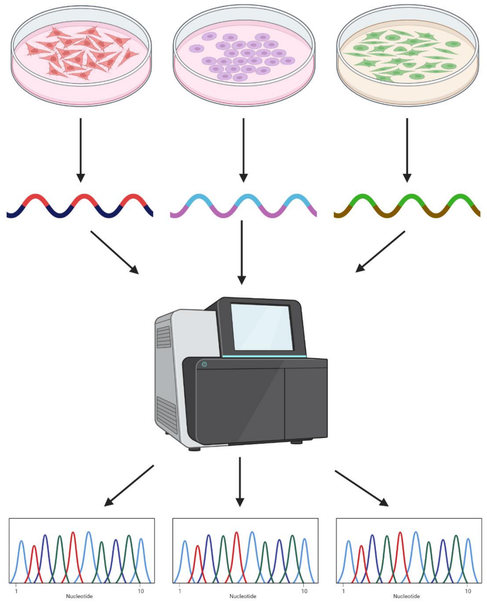
Drug response





∩

Feature matrix

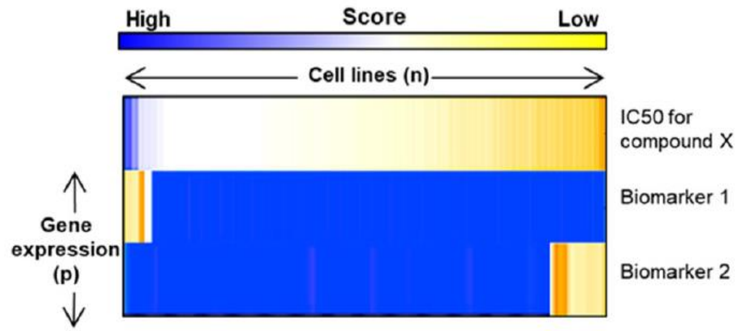


RNAs

Corresponding transcriptomic features

=

Best feature-response association



Modeling

Train-test split:

- 75-25
- No validation set

Hyperparameter tuning:

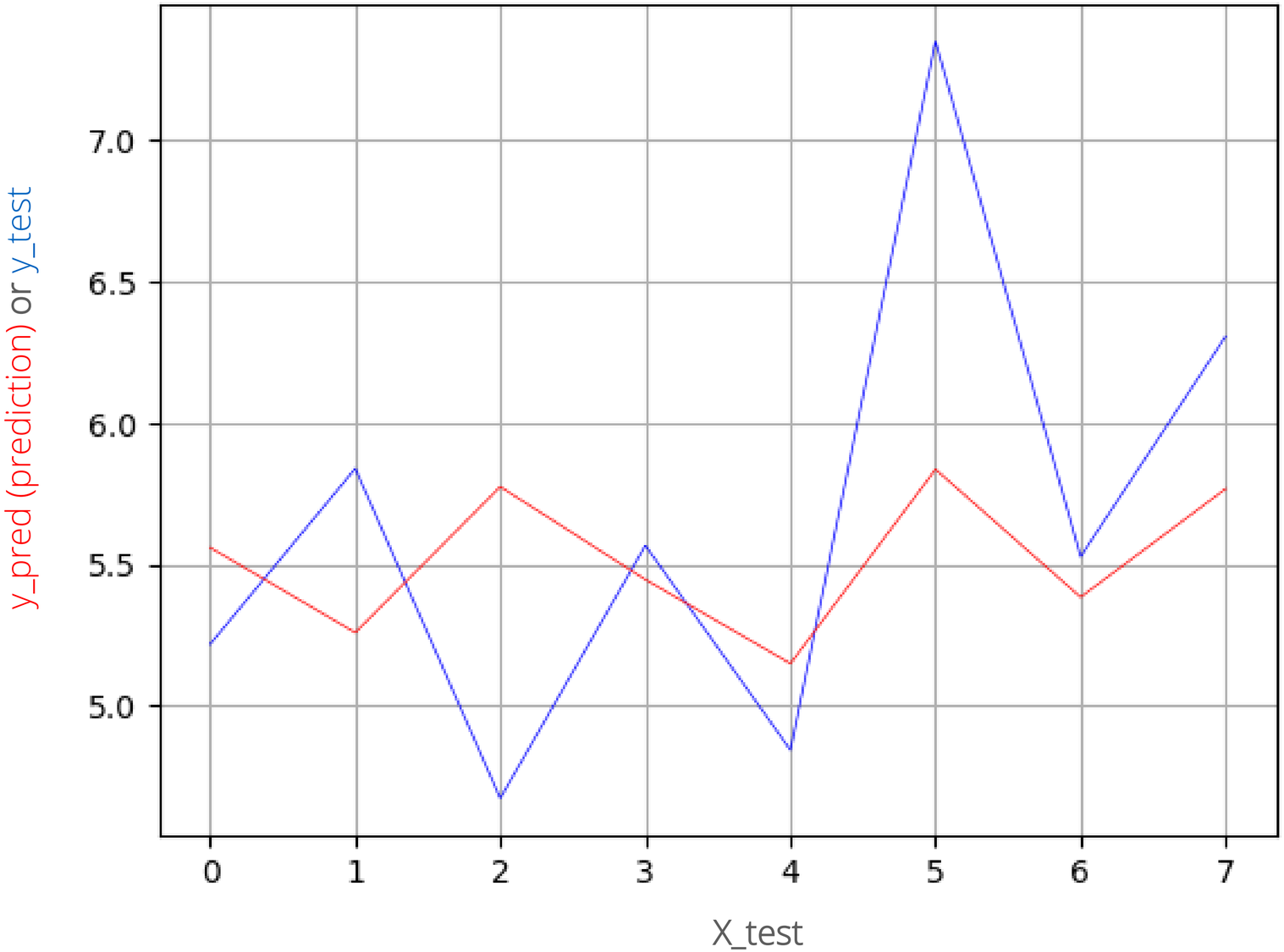
- Gridsearch:
- "max_iter" {1, 5, 10, 100, 500, 1000, 5000, 10000}
- "alpha" {0.0001, 0.001, 0.01, 0.1, 1, 10, 100}
- "l1_ratio" {0.0, 1.0, 0.1

K-fold cross-validation:

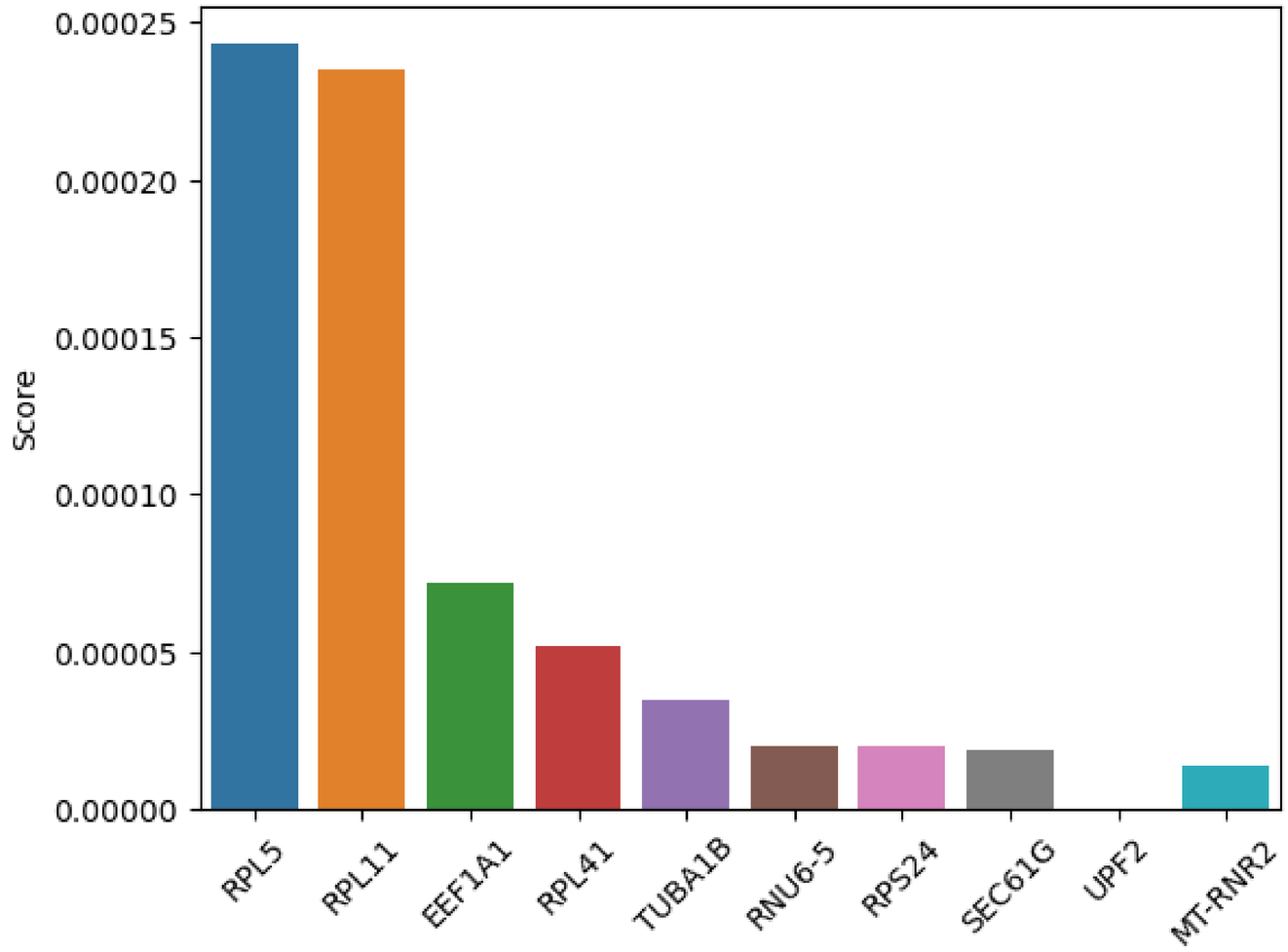
- Randomize (with seed)
- Repeat train test split and subsequent analysis

RMSE and MAE:

- Accuracy check: ~0.7

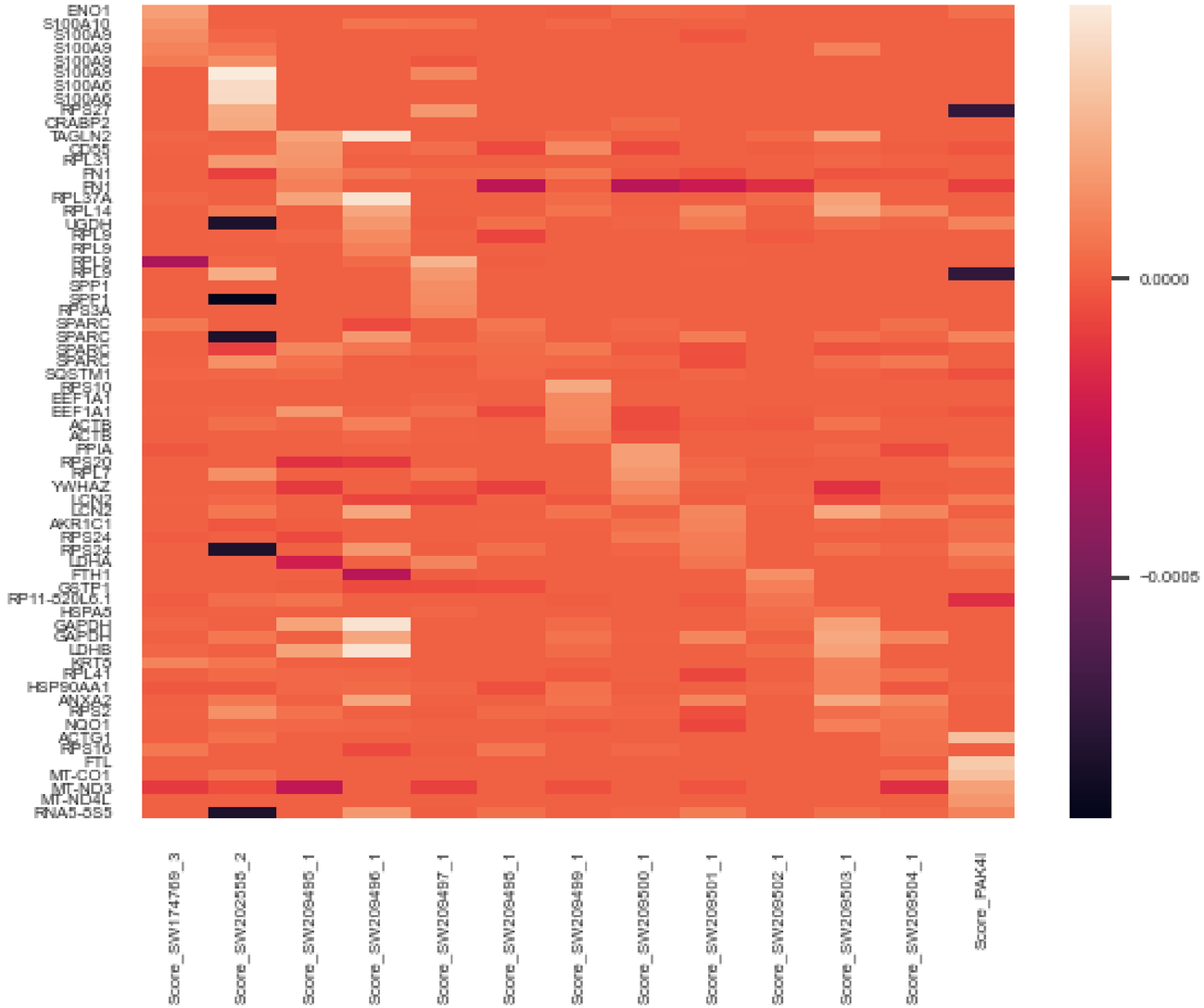


Result for feature importance for chemotherapeutic agent PAK4i

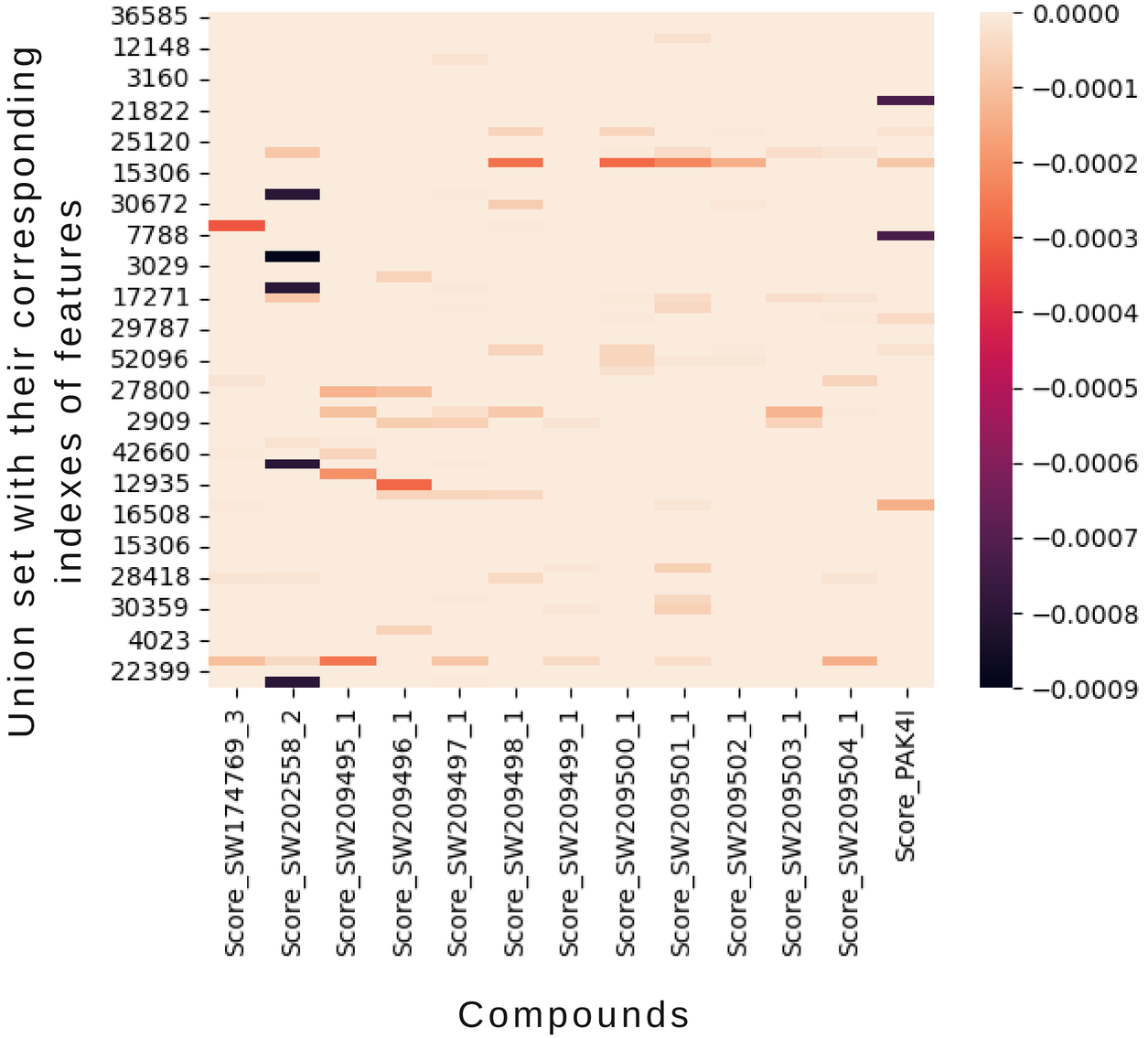


Heatmap for the result: positive correlators

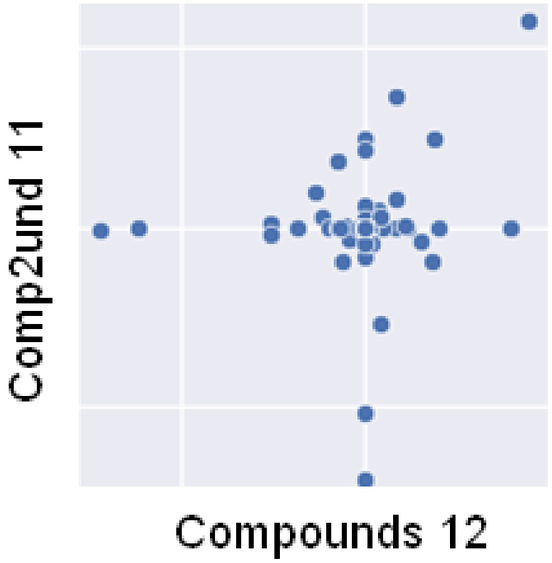
- For PAK4 inhibitor, the best biomarker was ACTG1 which has similar functionality to PAK4 (drug target)
- For the 2nd compound S100 identified as a biomarkers
- For the 4th compound GAPDH was identified. The 3rd, 11th and 12th compounds also were predicting this as somewhat of a candidate biomarker

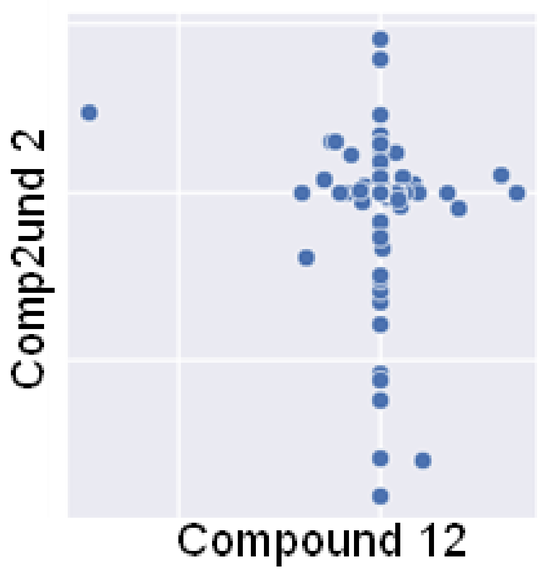


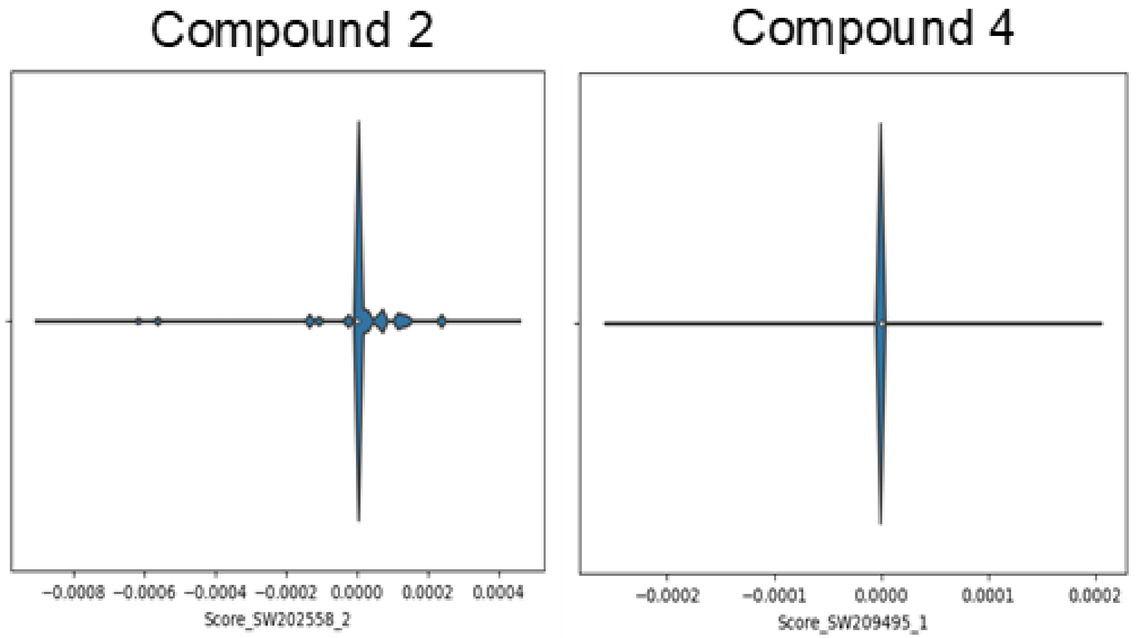
Heatmap for the result: negative correlators



Nonspecific vs specific hits







Future directions

Hybrid model: Classification followed by regression

Improved datasets: Dataset with validation set/gold standard and bigger datapoints should. SHAP and LIME methods to compare model accuracy

Multi-OMICS analysis: Couple transcriptomics with genomics and proteomics