Who I am?

Ziying Yang

- Fresh PhD candidate in Bioinformatics, University of Zurich (defending June 2025)
- 7+ years experience in bioinformatics, modeling, and pipeline development
- Led independent ML projects on cancer genomics and biological heterogeneity
- Passionate about real-world applications of interpretable AI in healthcare

Autonomous Chemical Research with LLMs

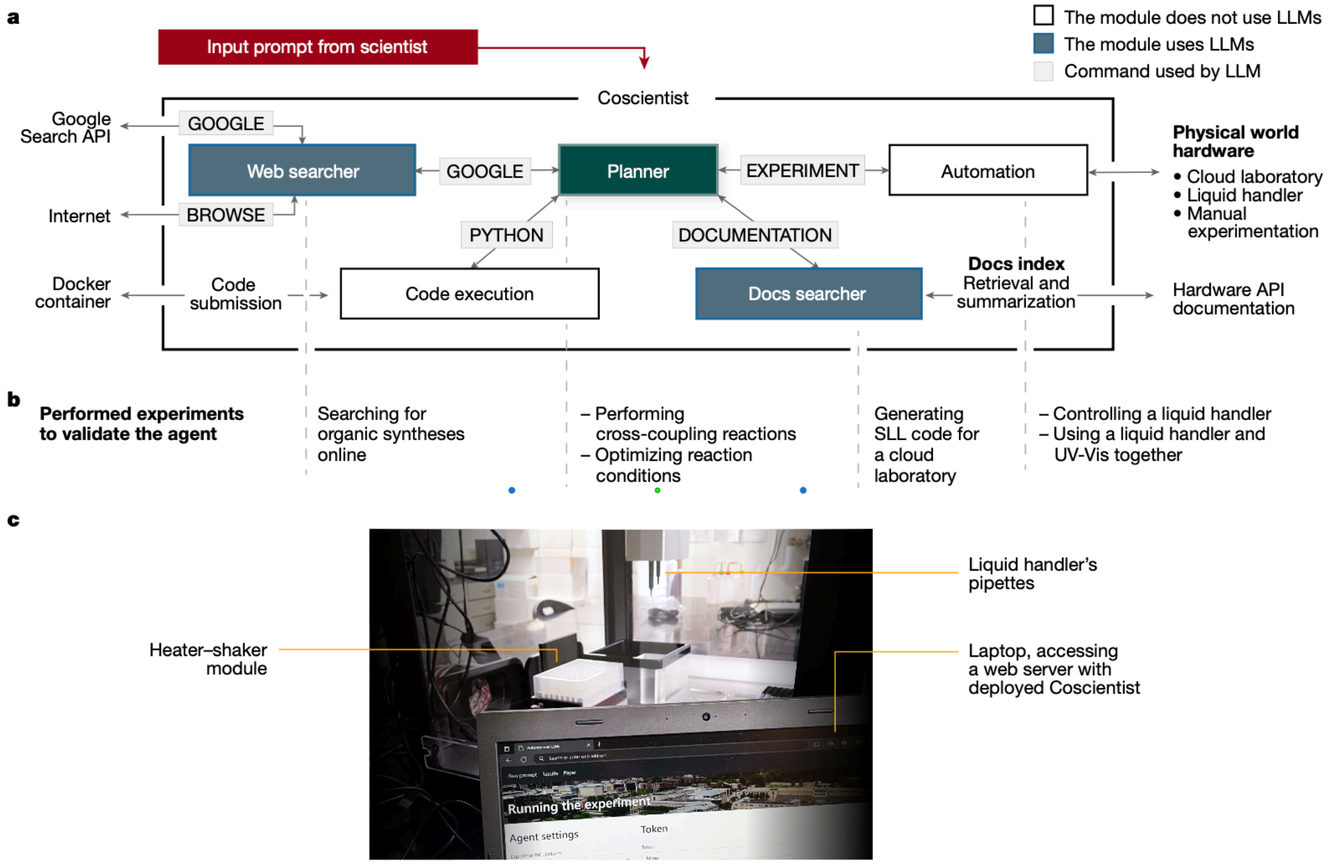
Coscientist - Nature, Vol 624, 2023

Ziying Yang

Introduction

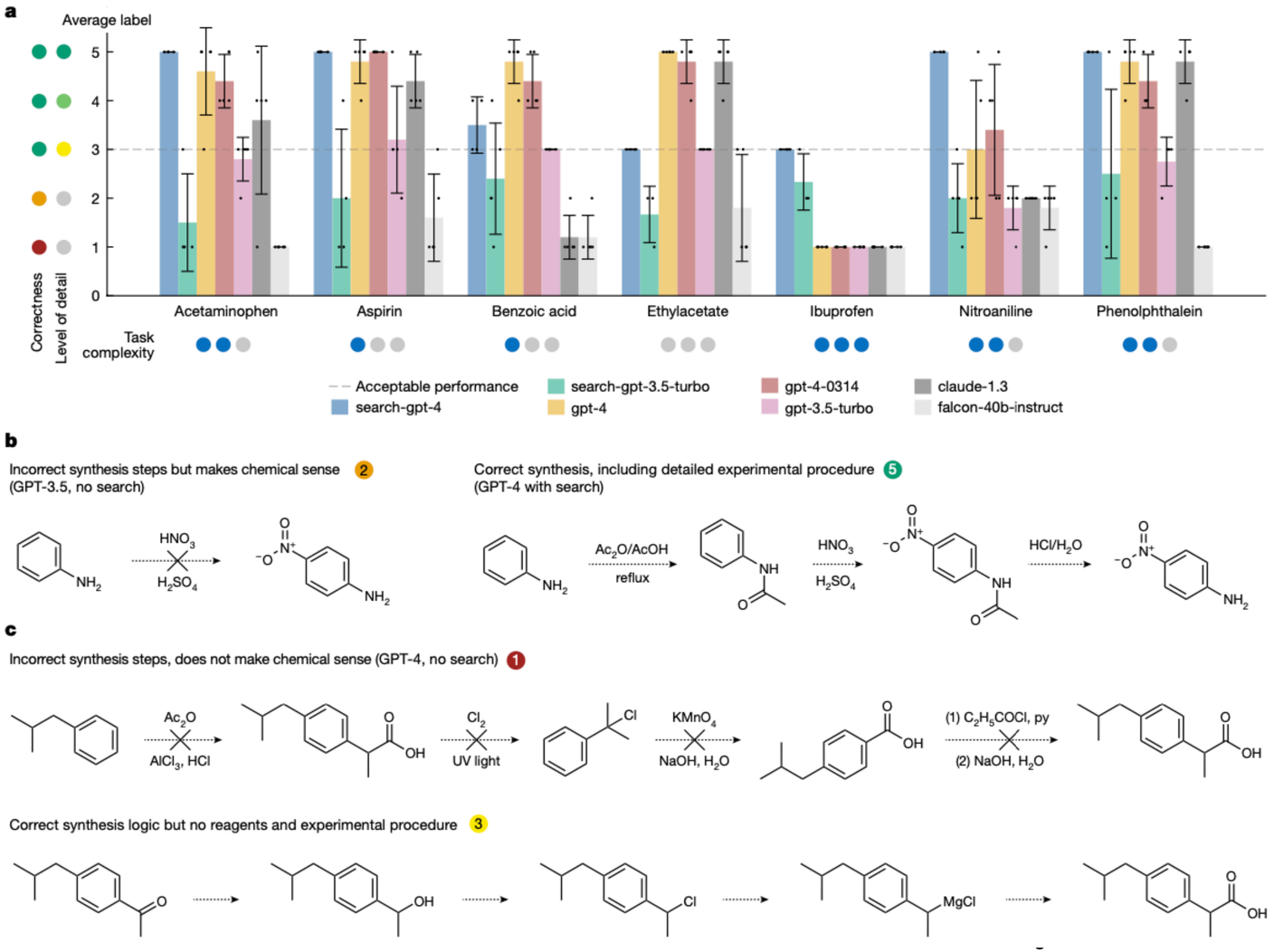
- The emergence of Large Language Models (LLMs) in scientific research
- Rising automation in chemistry labs
- Objective: Create an autonomous agent for real-world experimentation
- Introducing **Coscientist**: GPT-4-powered scientific reasoning and execution system

System Architecture



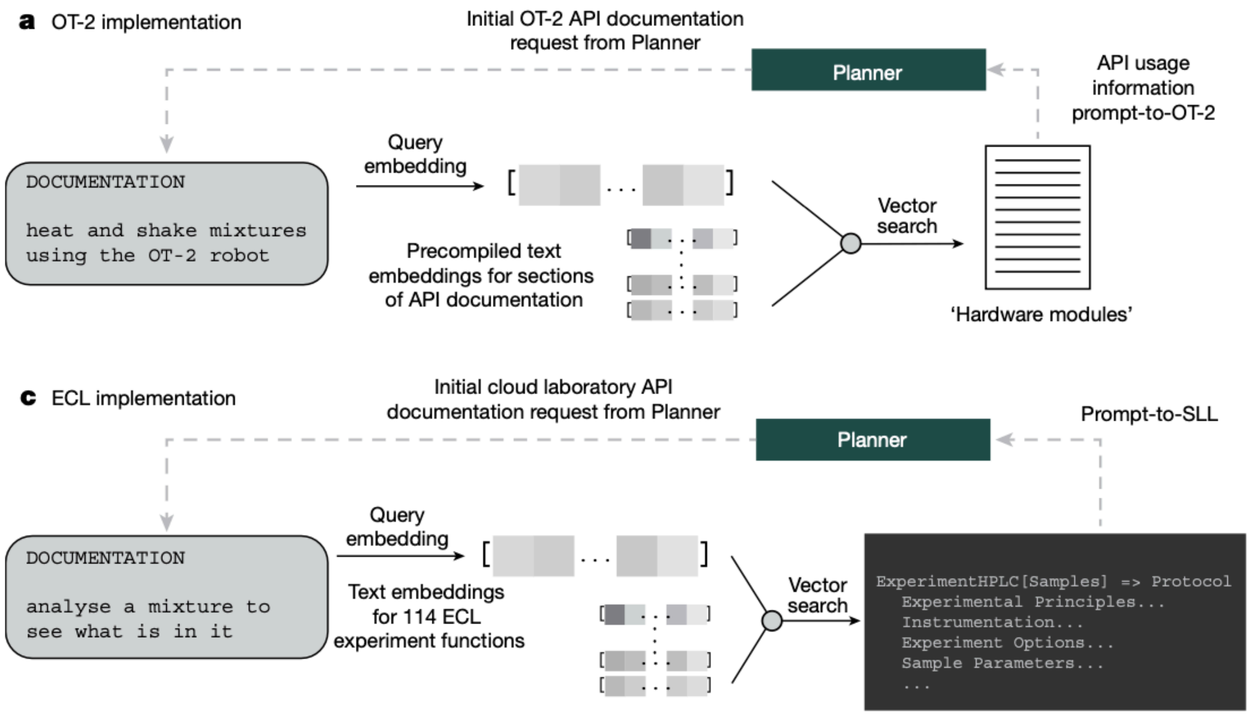
- Central planner: GPT-4 (Coscientist)
- Modular tool use via 4 commands:
  - GOOGLE → web search
  - DOCUMENTATION → reads technical docs
  - PYTHON → runs custom code
  - EXPERIMENT → controls lab devices
- Enables full cycle: Plan → Execute → Analyze

Synthesis Planning Performance



- Synthesizing 7 drug-like molecules from scratch
- GPT-4 outperformed other models in synthesis planning
- Accessed documentation and generated working API code
- Controlled robots and analysis devices using natural language

Document & API Understanding



b Valid OT-2 API code

```python
# Heat and shake the reaction
hs_mod.set_target_temperature(75)
hs_mod.wait_for_temperature()
hs_mod.set_and_wait_for_shake_speed(500)

# Deactivate heater and shaker
hs_mod.deactivate_heater()
hs_mod.deactivate_shaker()
hs_mod.open_labware_latch()
```

Proper usage of heater–shaker module

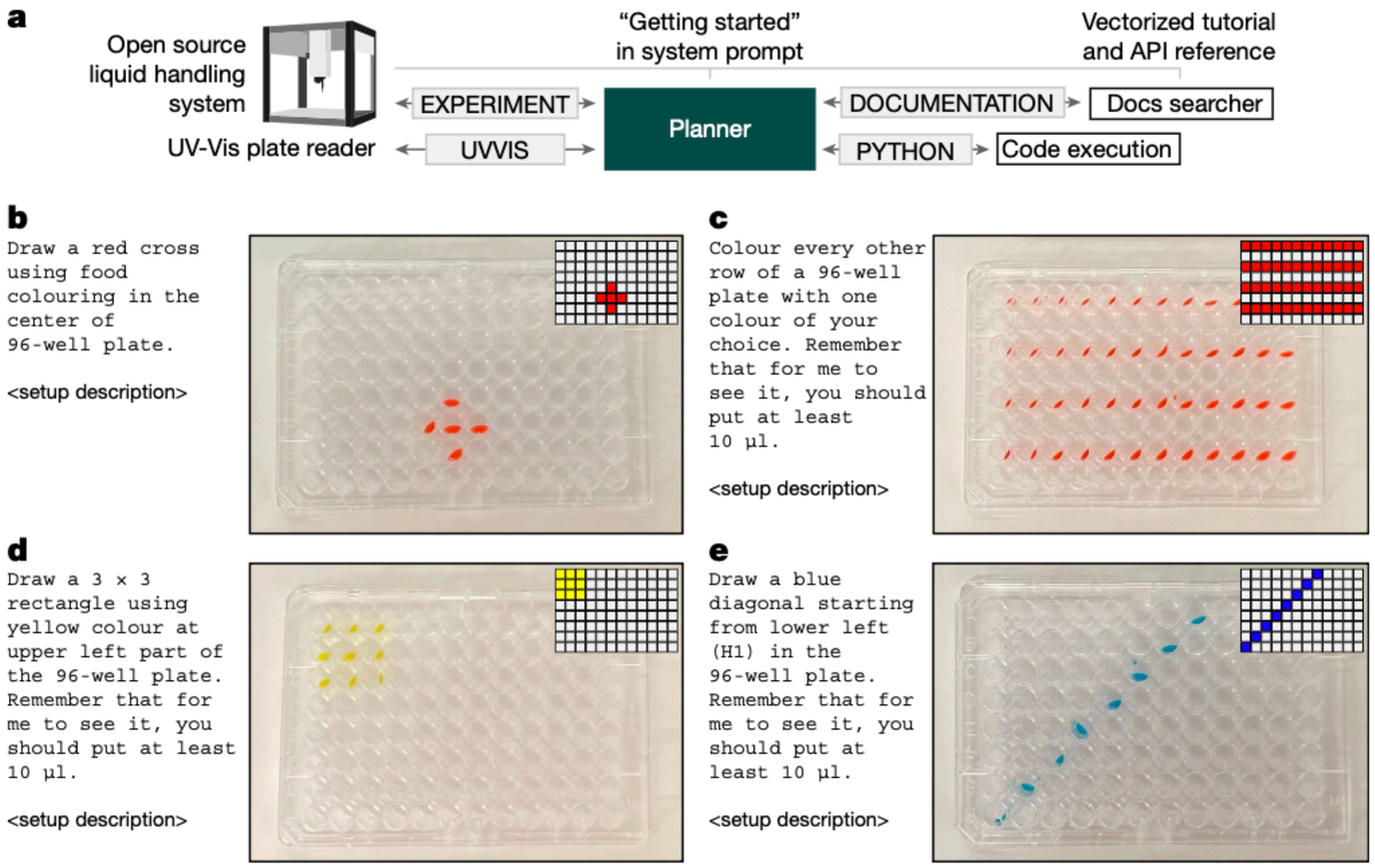
d Valid ECL SLL code

```
# Generated HPLC Experiment SLL Function Call
ExperimentHPLC[
    Object[Sample, ...],
    Instrument -> Model[Instrument, ...]
]
```

Targeted experiment options are set by the Planner

- Extracts necessary parameters and commands from raw documentation
- Automatically writes Python code to interface with lab platforms:
  - Opentrons OT-2
  - Emerald Cloud Lab (ECL)
- Shows reasoning and self-correction abilities

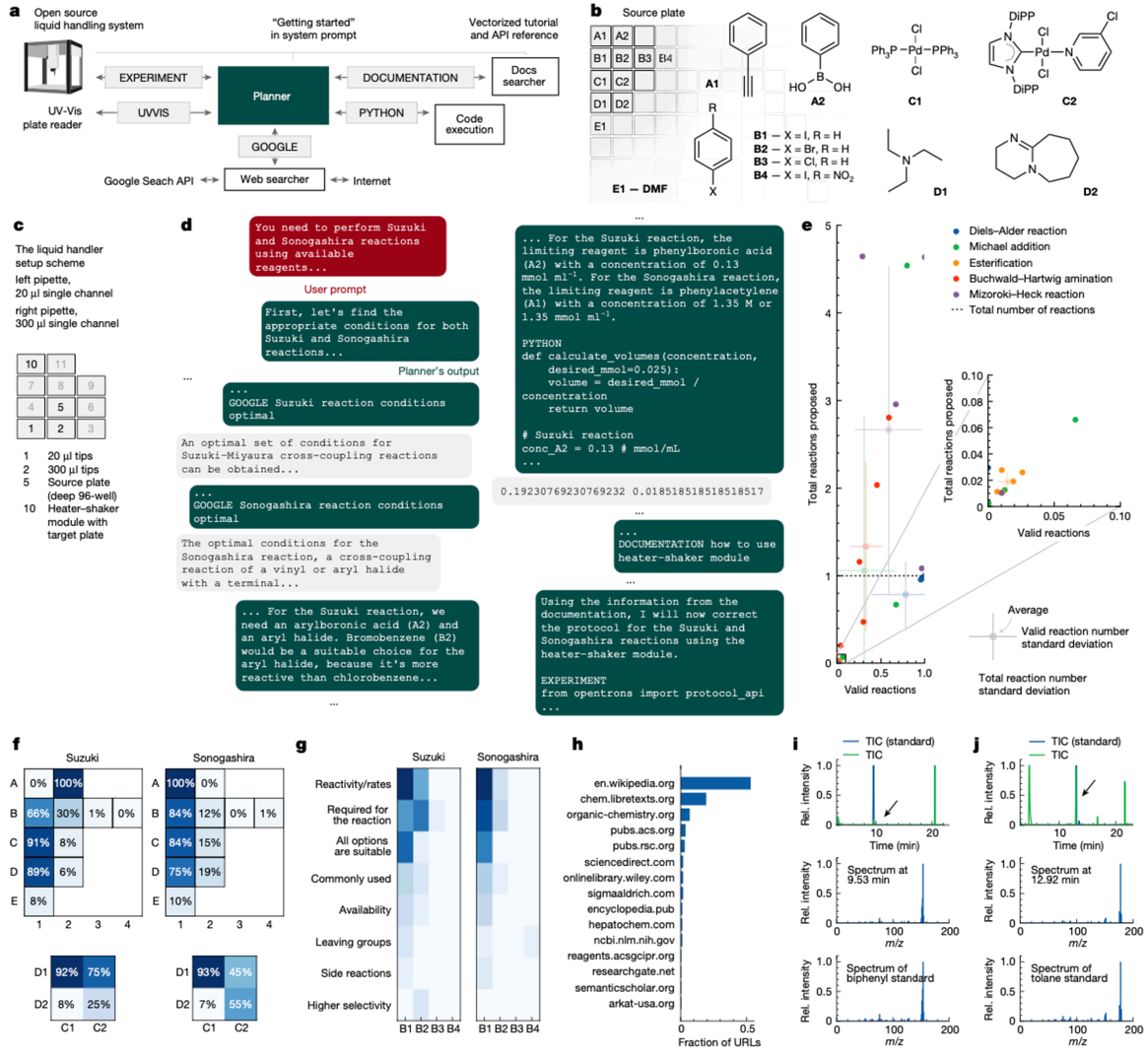
Executing Experiments



- Conducts operations like pipetting, mixing, plotting
- Controlled colorimetric assays via UV-Vis spectroscopy
- Interprets graphs and draws conclusions

Full Reaction Execution

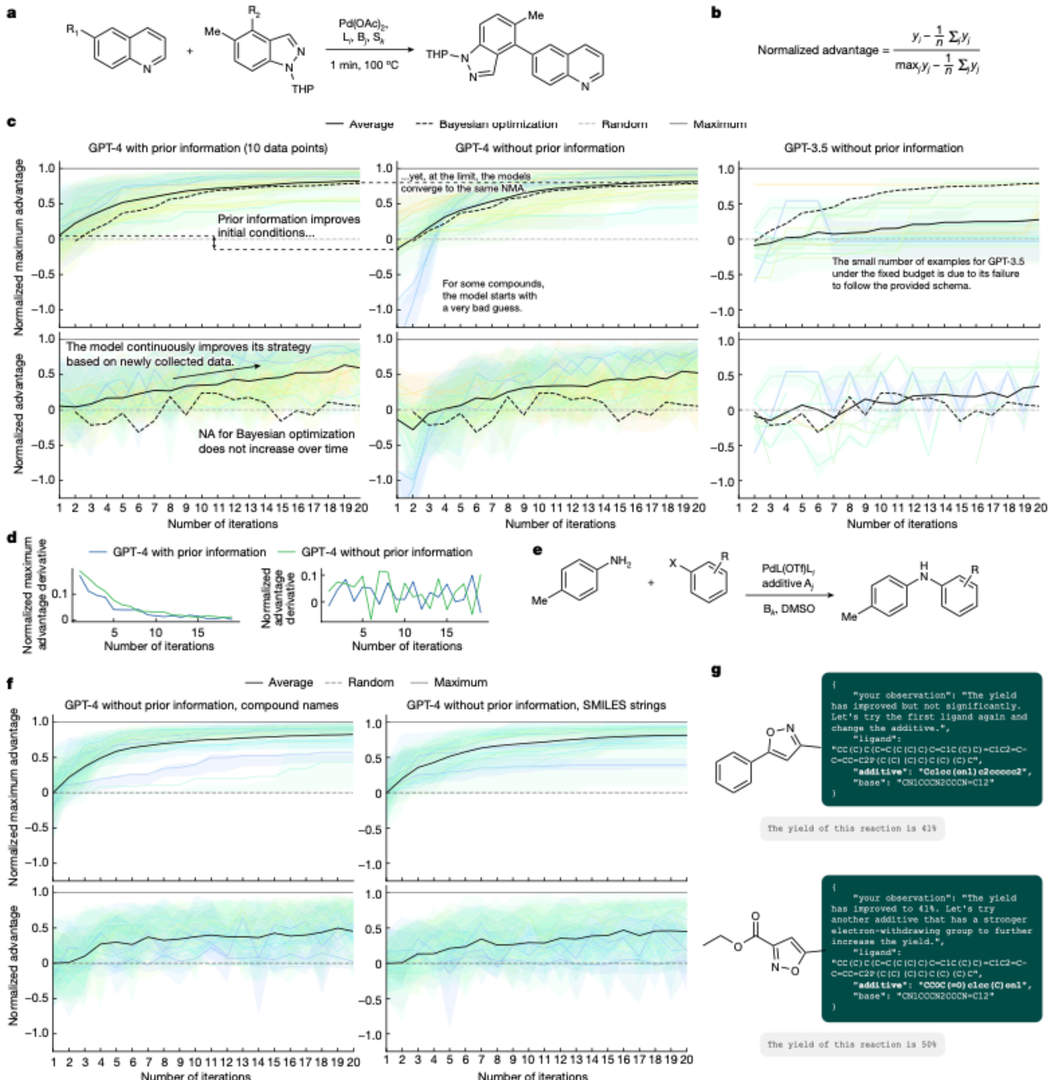
Integrated Reaction Design



- Performed Suzuki and Sonogashira coupling reactions
- Selected reagents, wrote code, fixed execution errors
- Verified output via GC-MS analysis

Optimization Strategy

Reaction Optimization Game



- Used GPT-4 to optimize experimental yield in minimal trials
- Compared with traditional Bayesian optimization
- GPT-4 learned patterns and improved performance over rounds

Discussion

- GPT-4 shows reasoning, planning, error correction, and adaptation
- Key step toward autonomous science agents
- Limitations: hallucination, safety, dual-use risks
- Future work: integration with databases, reinforcement learning, multimodal I/O

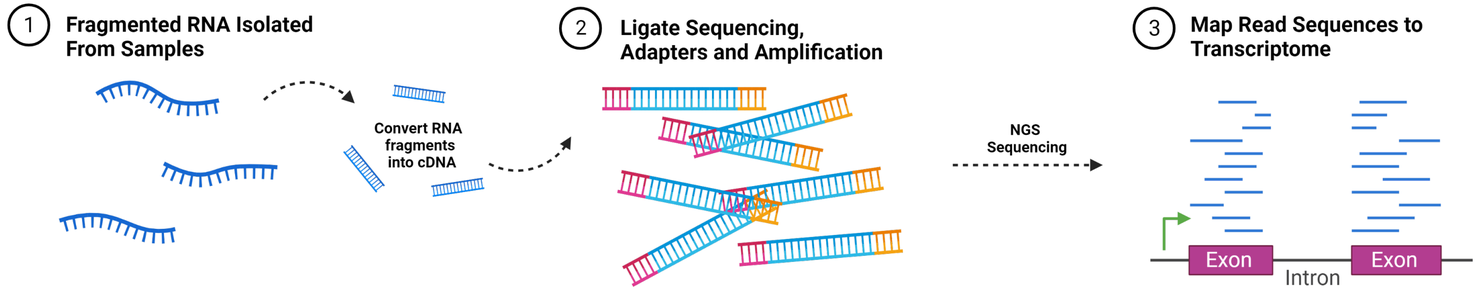
Analysis of Gene Expression on Breast Cancer

Ziying Yang

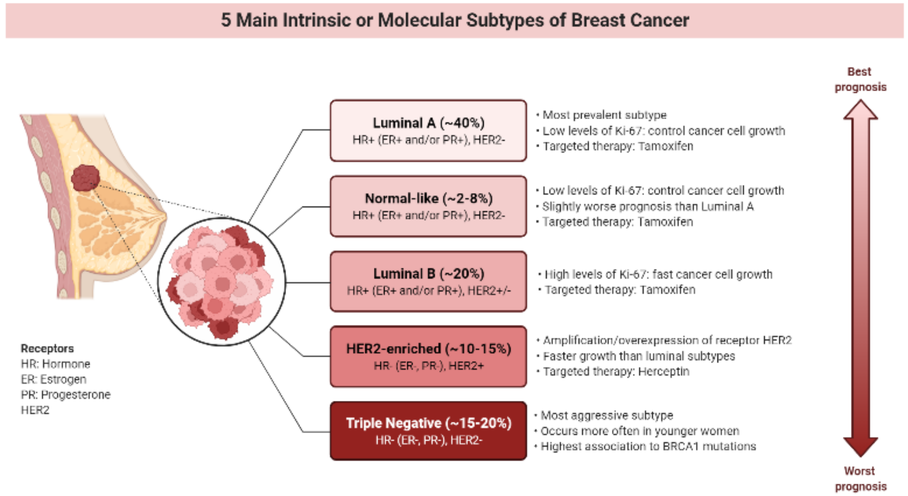
Background

RNA-seq and breast cancer

RNA Sequencing (RNA-Seq)



- Converts RNA to cDNA, then sequences using platforms like Illumina.
- Generates **counts** of transcripts per gene, which reflect gene expression levels.
- Allows differential expression analysis between different groups



Breast cancer is a **highly heterogeneous** disease. It includes multiple **molecular subtypes**, each with different:

- Gene expression patterns
- Prognosis
- Response to treatment

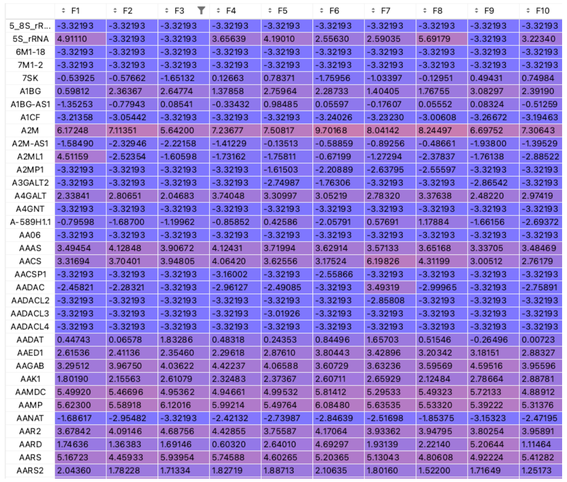
RNA-Seq allows to:

- Classify tumors into subtypes (e.g., Luminal A, Luminal B, HER2-enriched, Basal-like) based on gene expression signatures like PAM50.
- Identify biomarkers related to prognosis or drug response.
- Understand the tumor microenvironment, including immune activity, stromal cells, and signaling pathways.
- Discover potential therapeutic targets by analyzing differentially expressed genes.

Datasets

30,865 genes

Gene expression data



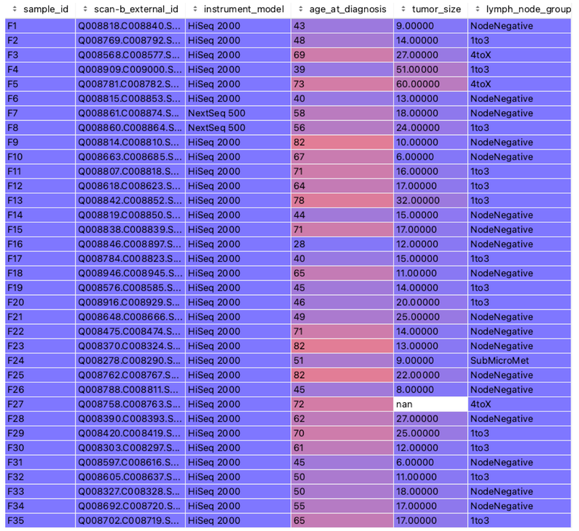
	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
5_8S_rR...	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193
5S_rRNA	4.91110	-3.32193	-3.32193	3.65639	4.19010	2.55630	2.59035	5.69179	-3.32193	3.22340
6M1-18	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193
7M1-2	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193
7SK	-0.53925	-0.57662	-1.65132	0.12663	0.78371	-1.75956	-1.03397	-0.12951	0.49431	0.74984
A1BG	0.59812	2.36367	2.64774	1.37858	2.75964	2.28733	1.40405	1.76755	3.08297	2.39190
A1BG-AS1	-1.35253	-0.77943	0.08541	-0.33432	0.98485	0.05597	-0.17607	0.05552	0.08324	-0.51259
A1CF	-3.21358	-3.05442	-3.32193	-3.32193	-3.32193	-3.24026	-3.23230	-3.00608	-3.26672	-3.19463
A2M	6.17248	7.11351	5.64200	7.23677	7.50817	9.70168	8.04142	8.24497	6.69752	7.30643
A2M-AS1	-1.58490	-2.32946	-2.22158	-1.41229	-0.13513	-0.58859	-0.89256	-0.48661	-1.93800	-1.39529
A2ML1	4.51159	-2.52354	-1.60598	-1.73162	-1.75811	-0.67199	-1.27294	-2.37837	-1.76138	-2.88522
A2MP1	-3.32193	-3.32193	-3.32193	-3.32193	-1.61503	-2.20889	-2.63795	-2.55597	-3.32193	-3.32193
A3GALT2	-3.32193	-3.32193	-3.32193	-3.32193	-2.74987	-1.76306	-3.32193	-3.32193	-2.86542	-3.32193
A4GALT	2.33841	2.80651	2.04683	3.74048	3.30997	3.05219	2.78320	3.37638	2.48220	2.97419
A4GNT	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193
A-589H1.1	-0.79598	-1.68700	-1.19962	-0.85852	0.42586	-2.05791	0.57691	1.17884	-1.66156	-2.69372
AA06	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193
AAAS	3.49454	4.12848	3.90672	4.12431	3.71994	3.62914	3.57133	3.65168	3.33705	3.48469
AACS	3.31694	3.70401	3.94805	4.06420	3.62556	3.17524	6.19826	4.31199	3.00512	2.76179
AACSP1	-3.32193	-3.32193	-3.32193	-3.16002	-3.32193	-2.55866	-3.32193	-3.32193	-3.32193	-3.32193
AADAC	-2.45821	-2.28321	-3.32193	-2.96127	-2.49085	-3.32193	3.49319	-2.99965	-3.32193	-2.75891
AADACL2	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-2.85808	-3.32193	-3.32193	-3.32193
AADACL3	-3.32193	-3.32193	-3.32193	-3.32193	-3.01926	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193
AADACL4	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193	-3.32193
AADAT	0.44743	0.06578	1.83286	0.48318	0.24353	0.84496	1.65703	0.51546	-0.26496	0.00723
AAED1	2.61536	2.41136	2.35460	2.29618	2.87610	3.80443	3.42896	3.20342	3.18151	2.88327
AAGAB	3.29512	3.96750	4.03622	4.42237	4.06588	3.60729	3.63236	3.59569	4.59516	3.95596
AAK1	1.80190	2.15563	2.61079	2.32483	2.37367	2.60711	2.65929	2.12484	2.78664	2.88781
AAMDC	5.49920	5.46696	4.95362	4.94661	4.99532	5.81412	5.29533	5.49323	5.72133	4.88912
AAMP	5.62300	5.58918	6.12016	5.99214	5.49764	6.08480	5.63535	5.53320	5.39222	5.31376
AANAT	-1.68617	-2.95482	-3.32193	-2.42132	-2.73987	-2.84639	-2.51698	-1.85375	-3.15323	-2.47195
AAR2	3.67842	4.09146	4.68756	4.42855	3.75587	4.17064	3.93362	3.94795	3.80254	3.95891
AARD	1.74636	1.36383	1.69146	0.60320	2.64010	4.69297	1.93139	2.22140	5.20644	1.11464
AARS	5.16723	4.45933	5.93954	5.74588	4.60265	5.20365	5.13043	4.80608	4.92224	5.41282
AARS2	2.04360	1.78228	1.71334	1.82719	1.88713	2.10635	1.80160	1.52200	1.71649	1.25173

3,273 samples (136 replicated)

Clinical data

```
^SAMPLE = GSM2528081
!Sample_title = F3
!Sample_geo_accession = GSM2528081
!Sample_status = Public on Mar 12 2018
!Sample_submission_date = Mar 09 2017
!Sample_last_update_date = May 04 2022
!Sample_type = SRA
!Sample_channel_count = 1
!Sample_source_name_ch1 = Primary breast tumor
!Sample_organism_ch1 = Homo sapiens
!Sample_taxid_ch1 = 9606
!Sample_characteristics_ch1 = scan-b external id: Q008568.C008577.S000424.l.r.m3.c.l
!Sample_characteristics_ch1 = instrument model: HiSeq 2000
!Sample_characteristics_ch1 = age at diagnosis: 69
!Sample_characteristics_ch1 = tumor size: 27
!Sample_characteristics_ch1 = lymph node group: 4toX
!Sample_characteristics_ch1 = lymph node status: NodePositive
!Sample_characteristics_ch1 = er status: 1
!Sample_characteristics_ch1 = pgr status: 1
!Sample_characteristics_ch1 = her2 status: 0
!Sample_characteristics_ch1 = ki67 status: NA
!Sample_characteristics_ch1 = nhg: G3
!Sample_characteristics_ch1 = er prediction mgc: 1
!Sample_characteristics_ch1 = pgr prediction mgc: 1
!Sample_characteristics_ch1 = her2 prediction mgc: 0
!Sample_characteristics_ch1 = ki67 prediction mgc: 1
!Sample_characteristics_ch1 = nhg prediction mgc: G3
!Sample_characteristics_ch1 = er prediction sgc: 1
!Sample_characteristics_ch1 = pgr prediction sgc: 1
!Sample_characteristics_ch1 = her2 prediction sgc: 0
```

3,273 samples

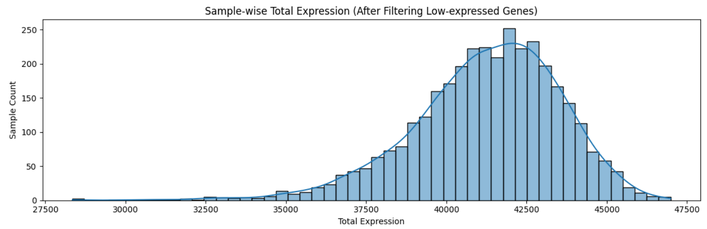


sample_id	scan-b_external_id	instrument_model	age_at_diagnosis	tumor_size	lymph_node_group
F1	Q008818.C008840.S...	HiSeq 2000	43	9.00000	NodeNegative
F2	Q008769.C008792.S...	HiSeq 2000	48	14.00000	1to3
F3	Q008568.C008577.S...	HiSeq 2000	69	27.00000	4toX
F4	Q008909.C009000.S...	HiSeq 2000	39	51.00000	1to3
F5	Q008781.C008782.S...	HiSeq 2000	73	60.00000	4toX
F6	Q008815.C008853.S...	HiSeq 2000	40	13.00000	NodeNegative
F7	Q008861.C008874.S...	NextSeq 500	58	18.00000	NodeNegative
F8	Q008860.C008864.S...	NextSeq 500	56	24.00000	1to3
F9	Q008814.C008810.S...	HiSeq 2000	82	10.00000	NodeNegative
F10	Q008663.C008685.S...	HiSeq 2000	67	6.00000	NodeNegative
F11	Q008807.C008818.S...	HiSeq 2000	71	16.00000	1to3
F12	Q008618.C008623.S...	HiSeq 2000	64	17.00000	1to3
F13	Q008842.C008852.S...	HiSeq 2000	78	32.00000	1to3
F14	Q008819.C008850.S...	HiSeq 2000	44	15.00000	NodeNegative
F15	Q008838.C008839.S...	HiSeq 2000	71	17.00000	NodeNegative
F16	Q008846.C008897.S...	HiSeq 2000	28	12.00000	NodeNegative
F17	Q008784.C008823.S...	HiSeq 2000	40	15.00000	1to3
F18	Q008946.C008945.S...	HiSeq 2000	65	11.00000	NodeNegative
F19	Q008576.C008585.S...	HiSeq 2000	45	14.00000	1to3
F20	Q008916.C008929.S...	HiSeq 2000	46	20.00000	1to3
F21	Q008648.C008666.S...	HiSeq 2000	49	25.00000	NodeNegative
F22	Q008475.C008474.S...	HiSeq 2000	71	14.00000	NodeNegative
F23	Q008370.C008324.S...	HiSeq 2000	82	13.00000	NodeNegative
F24	Q008278.C008290.S...	HiSeq 2000	51	9.00000	SubMicroMet
F25	Q008762.C008767.S...	HiSeq 2000	82	22.00000	NodeNegative
F26	Q008788.C008811.S...	HiSeq 2000	45	8.00000	NodeNegative
F27	Q008758.C008763.S...	HiSeq 2000	72	nan	4toX
F28	Q008390.C008393.S...	HiSeq 2000	62	27.00000	NodeNegative
F29	Q008420.C008419.S...	HiSeq 2000	70	25.00000	1to3
F30	Q008303.C008297.S...	HiSeq 2000	61	12.00000	1to3
F31	Q008597.C008616.S...	HiSeq 2000	45	6.00000	NodeNegative
F32	Q008605.C008637.S...	HiSeq 2000	50	11.00000	1to3
F33	Q008327.C008328.S...	HiSeq 2000	50	18.00000	NodeNegative
F34	Q008692.C008720.S...	HiSeq 2000	55	17.00000	NodeNegative
F35	Q008702.C008719.S...	HiSeq 2000	65	17.00000	1to3

24 clinical information types

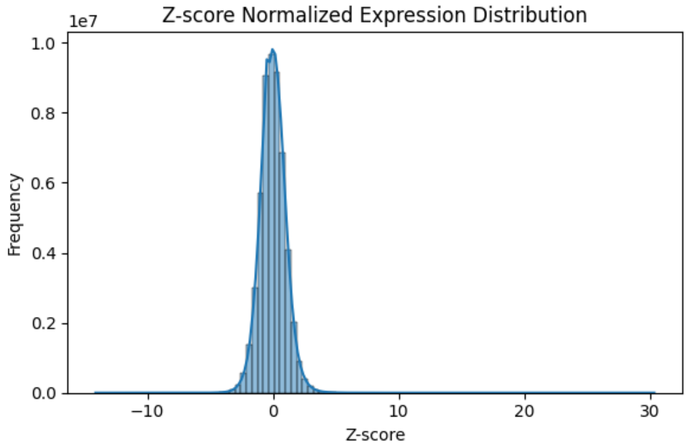
Data Processing on Expression Data

- Quality Control (QC)
  - Remove low expression genes on the majority (over 90%) of samples
  - Delete samples with extreme low expression (total expression less than 5%)
- Normalization
  - Z-score on each gene



30865 genes -> 16531 genes

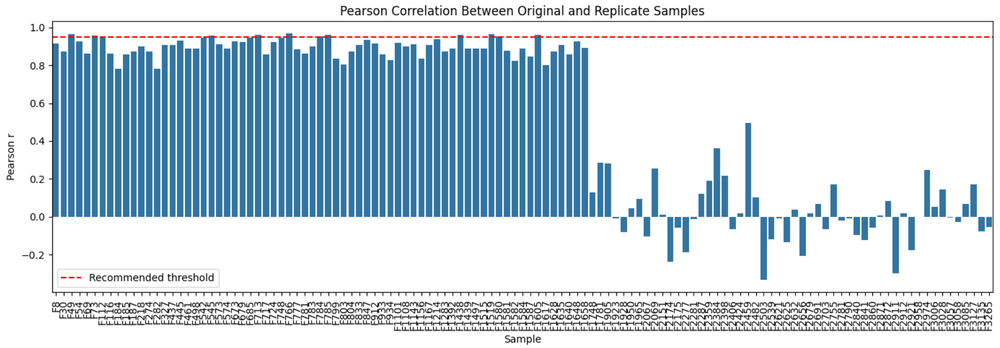
3409 samples -> 3238 samples



Data Processing on Expression Data

Replicates processing

- Identify replicate pairs
  - 121 replicate pairs found
- Remove replicate pairs with Pearson correlation lower than 0.9 -> 87 pairs deleted

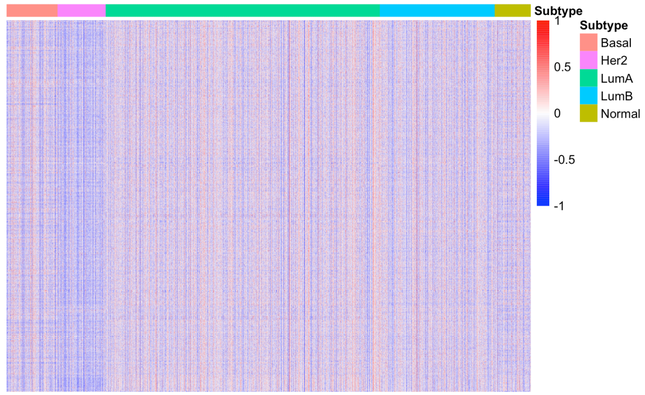


- For left pairs compute average expression -> 34 pairs combined

16531 genes * 3117 samples

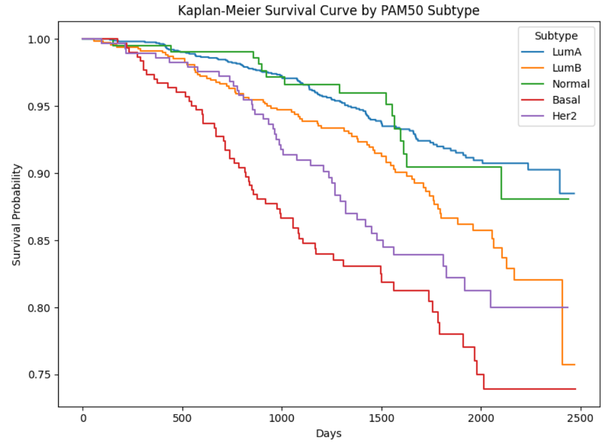
Data analysis on gene expression

Use-case: cancer subtype classification



No clear gene patterns related to subtypes

Subtype	Sample Count
LumA	1631
LumB	684
Basal	303
Her2	286
Normal	213

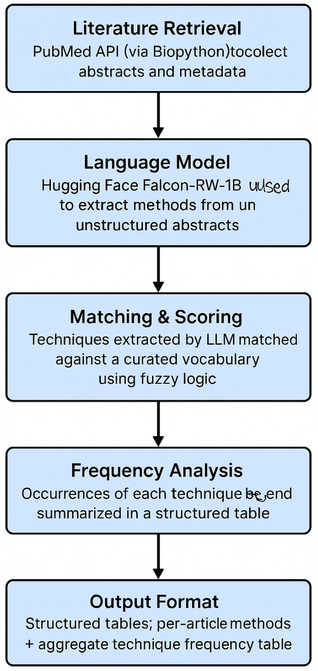


No clear boundaries for different subtypes on PCA space

Literature review of classification on subtypes

LLM integration

Module	Description
Literature Retrieval	PubMed API (via Biopython) to collect abstracts and metadata
Language Model	Hugging Face `Falcon-RW-1B` used for extracting methods from unstructured abstracts
Hybrid Extraction	Combined LLM outputs with keyword dictionary for better coverage and standardization
Matching & Scoring	Techniques extracted by LLM are matched against a curated vocabulary using fuzzy logic
Frequency Analysis	Occurrences of each technique are counted and summarized in a structured table
Output Format	Structured tables: per-article methods + aggregate technique frequency table



PMID	Title	Matched_Techniques
40660340	RASGEF1C methylation for the distinguishment and classification of benign and	support vector machine, random forest, svm, decision tree, k-means
40659967	ESE and Transfer Learning for Breast Tumor Classification.	
40656955	Clinical significance and heterogeneity of circulating tumor cells and clusters	
40656144	Epidemiological and molecular profile of breast cancer: a retrospective study in	
40647433	Using [(18)F]FDG PET/CT to Identify Optimal Responders to Neoadjuvant Therapy in	support vector machine, decision tree
40638258	Artificial Intelligence-powered copilots for precision diagnosis and surgical	support vector machine, naive bayes, random forest, decision tree, k-means
40633212	Antigen-presenting CAFs orchestrate immunosuppressive niches via CXCL13-mediated	
40630982	Normalization and Selecting Non-Differentially Expressed Genes Improve Machine	
40629347	Artificial intelligence-driven discovery of YH395A: A novel TGFbetaR1 inhibitor with	k-nearest neighbor, k-means
40628151	Combined model-driven and dual-cycle interactive strategy few-shot learning	support vector machine, lasso, logistic regression, random forest, deep learning, autoencoder, decision tree

Technique	Count
support vector machine	4
decision tree	4
random forest	3
k-means	3
svm	1
naive bayes	1
k-nearest neighbor	1
lasso	1
logistic regression	1
deep learning	1
autoencoder	1

Data analysis on gene expression

Unsupervised learning use-case: cancer subtype classification

- In many cancer types, **labels are incomplete or unknown**, or existing classifications are **too coarse**.
- Expression data (e.g., RNA-seq) is high-dimensional and noisy — unsupervised learning helps to:
  - Reduce dimensionality
  - **Reveal latent structure** (e.g., biological subtypes, molecular patterns)
  - Guide new subtype definitions

Dimension reduction

Method	Purpose
PCA	Captures global variance (linear)
t-SNE / UMAP	Preserves local structure; great for visualization
Autoencoders	Learns compressed, **non-linear** representations

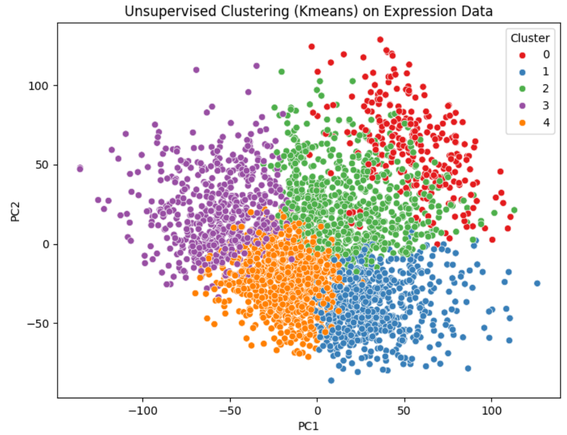
Clustering

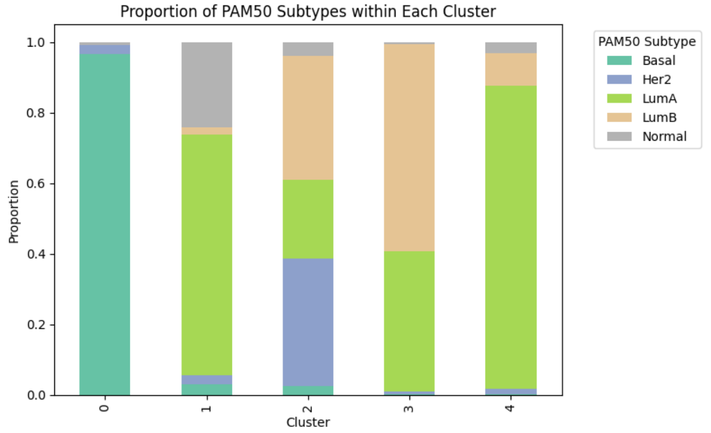
Method	Highlights
K-Means / Mini-Batch	Fast, simple; assumes spherical clusters; must predefine `k`
Hierarchical Clustering	Builds a dendrogram; linkage method (ward/average) affects structure
DBSCAN / OPTICS	Density-based; finds arbitrary shapes; no need to set `k`
Gaussian Mixture Models (GMM)	Probabilistic; allows soft (fractional) assignments
Spectral Clustering	Graph-based; captures non-linear relationships

Data analysis on gene expression

Unsupervised learning use-case: cancer subtype classification

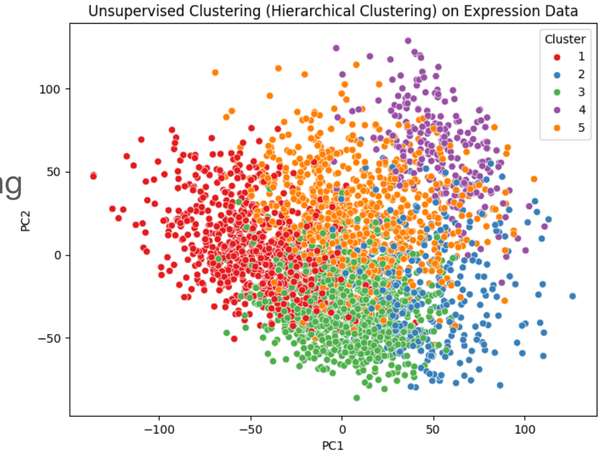
Kmeans

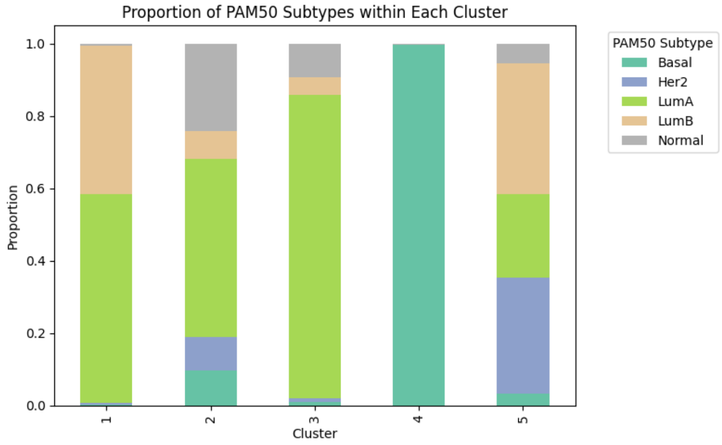




Cluster Accuracy (Hungarian aligned): 0.585

Hierarchical Clustering





Cluster Accuracy (Hungarian aligned): 0.551

Data analysis on gene expression

Unsupervised learning use-case: cancer subtype classification

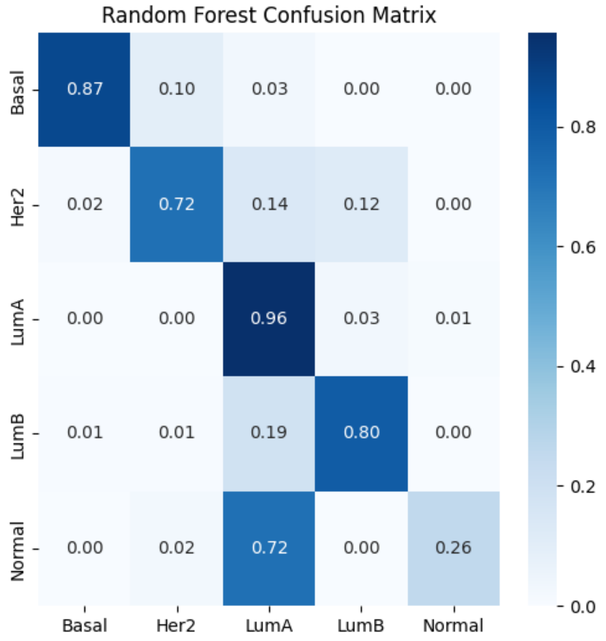
Dimension reduction method	Clustering method	ARI	NMI	AMI	Homogeneity	Completeness	Cluster Accuracy (Hungarian aligned)
PCA	Kmeans	0.236	0.351	0.350	0.383	0.324	0.504
PCA	Hierarchical Clustering	0.191	0.312	0.311	0.333	0.294	0.530
PCA	GMM	0.245	0.377	0.376	0.413	0.346	0.576
PCA	Spectral Clustering	0.249	0.377	0.376	0.376	0.378	0.519
Autoencoders	Kmeans	0.205	0.293	0.292	0.325	0.267	0.480
Autoencoders	Hierarchical Clustering	0.255	0.323	0.322	0.353	0.298	0.544
Autoencoders	GMM	0.262	0.339	0.337	0.367	0.314	0.555
Autoencoders	Spectral Clustering	0.354	0.242	0.240	0.236	0.248	0.570
Contrastive Learning	Kmeans	0.098	0.143	0.142	0.157	0.132	0.395
Contrastive Learning	Hierarchical Clustering	0.128	0.159	0.158	0.172	0.148	0.385
Contrastive Learning	GMM	0.104	0.139	0.138	0.153	0.128	0.412
Contrastive Learning	Spectral Clustering	0.125	0.166	0.164	0.183	0.152	0.378
Graph-based Methods	Kmeans	0.212	0.318	0.316	0.316	0.319	0.490
Graph-based Methods	Hierarchical Clustering	0.170	0.320	0.319	0.306	0.336	0.534
Graph-based Methods	GMM	0.296	0.334	0.332	0.326	0.342	0.577
Graph-based Methods	Spectral Clustering	0.153	0.298	0.296	0.283	0.315	0.445

- **High dimensionality** of gene expression → distance metrics become unreliable
- Non-spherical and overlapping clusters → violates KMeans assumptions
- Strong noise and batch effects → obscure biological structure
- **Imbalanced subtype sizes** → small subtypes get ignored or absorbed
- **Linear structure assumption** → fails to capture complex gene expression manifolds
- **No guidance from biological labels** → clusters reflect variance, not true subtypes
- **Global similarity dominance** → misses small subtype-specific gene patterns

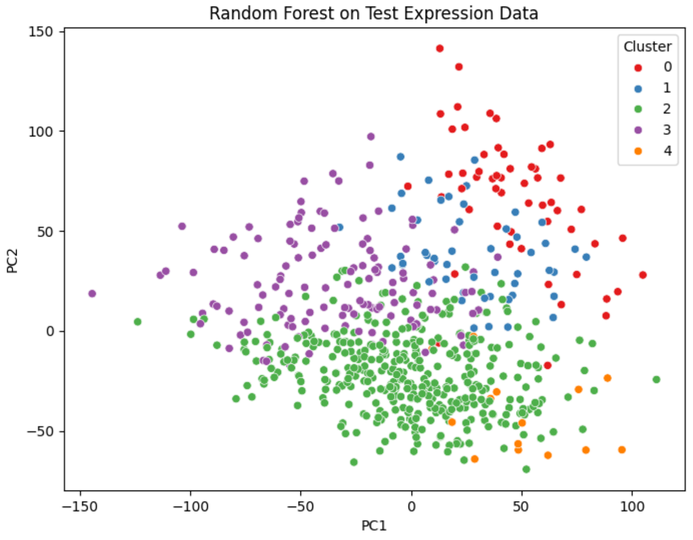
Basically cannot see a reasonable classification on unsupervised method, even with multiple combination of dimension reduction and clustering methods.

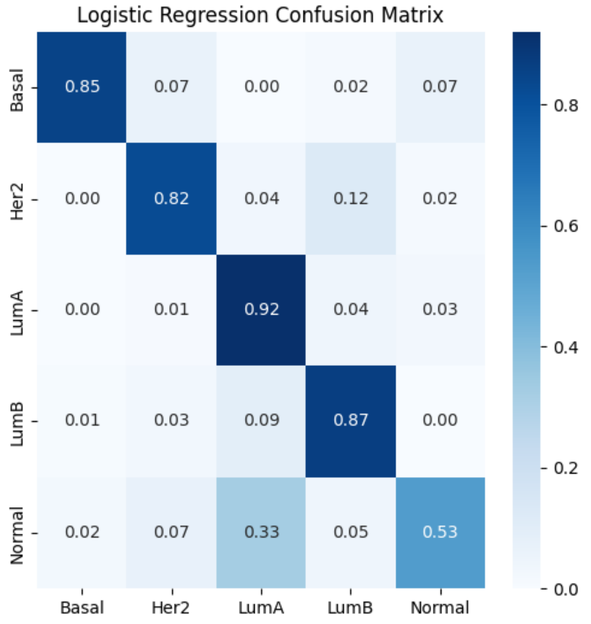
Data analysis on gene expression

Supervised learning use-case: cancer subtype classification



Average accuracy: 0.72



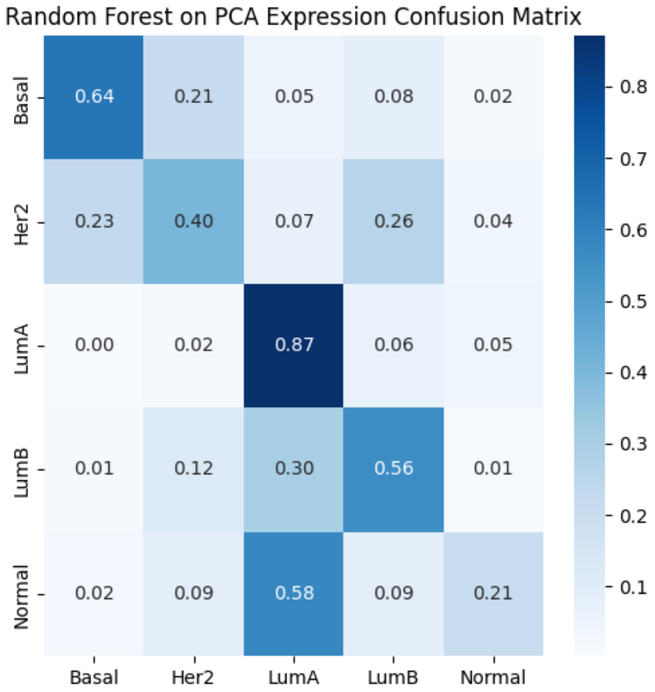


Average accuracy: 0.80

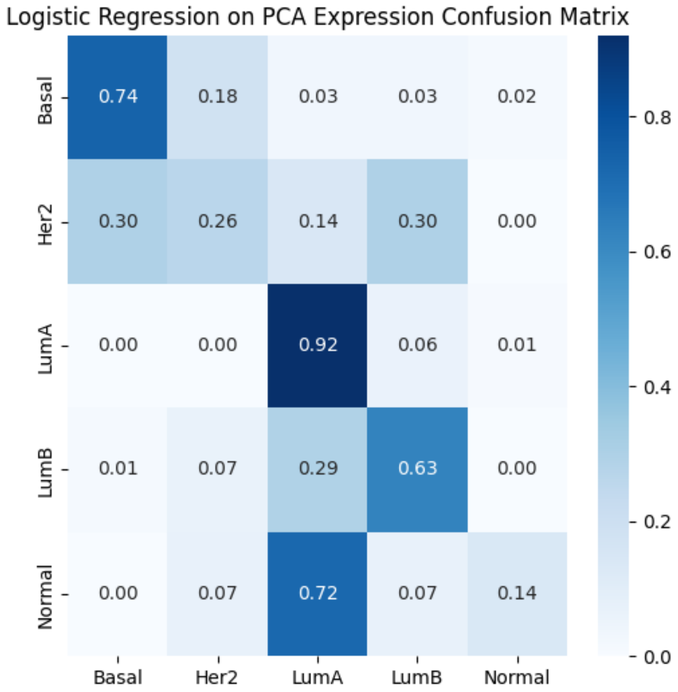
Obvious increased accuracy on supervised methods, compared to unsupervised methods.

Data analysis on gene expression

Supervised learning + dimension reduction use-case: cancer subtype classification



↓Average accuracy: 0.53



↓Average accuracy: 0.54

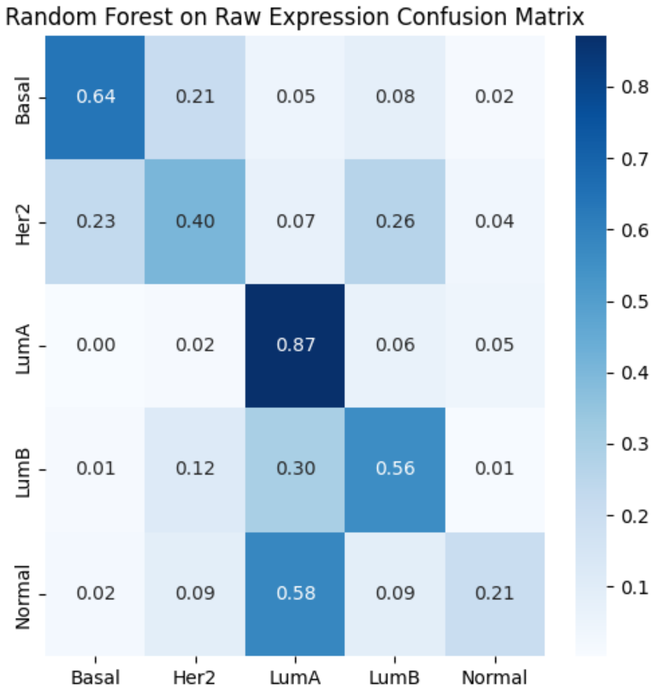
↓Average accuracy: 0.65

We may say, in this subtype classification problem, the subtype patterns cannot be captured well in the PCA/phate space

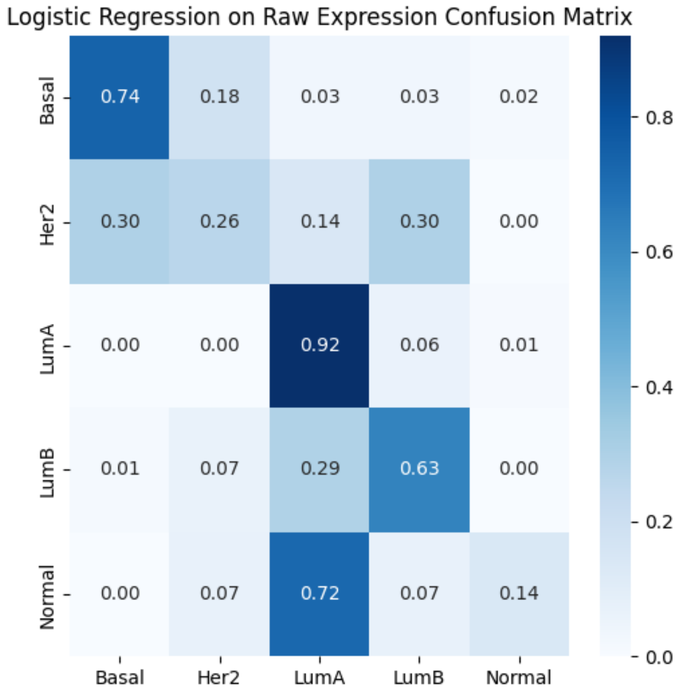
Data analysis on gene expression

Supervised learning on raw data use-case: cancer subtype classification

Does the data processing work?



↓Average accuracy: 0.53



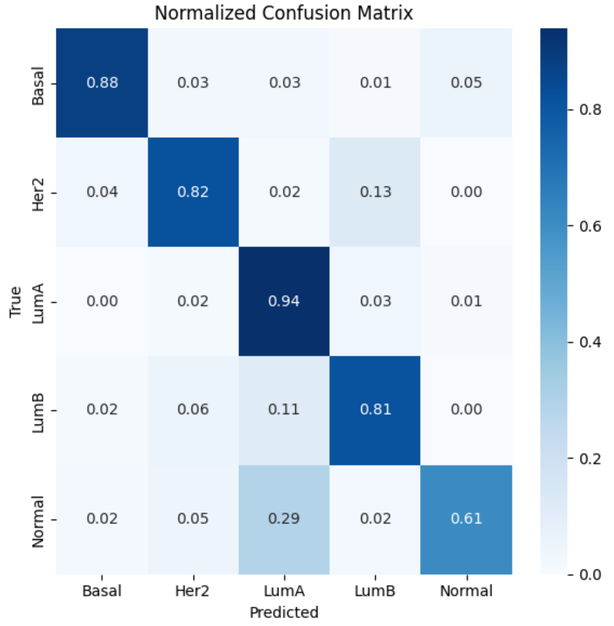
↓Average accuracy: 0.54

The efficiency of expression data processing

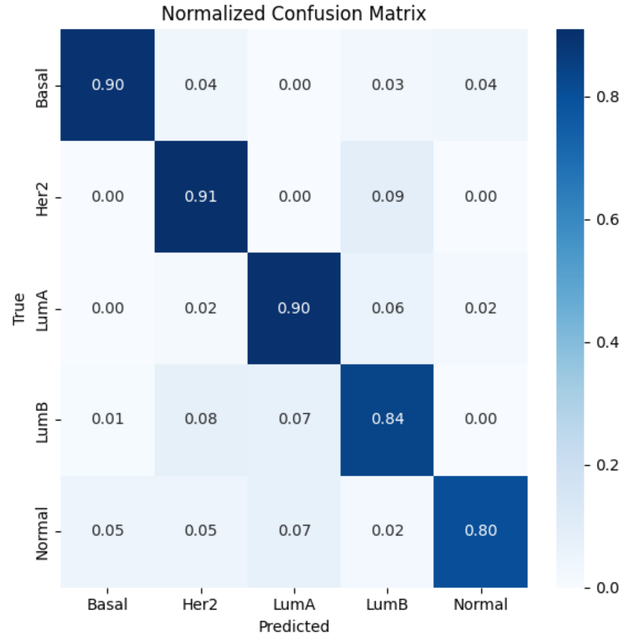
Data analysis on gene expression

Supervised learning **FFNN** use-case: cancer subtype classification

- Feedforward Neural Network, FFNN
  - **Simple feedforward architecture** with fully connected layers and dropout for regularization
  - **Class imbalance handled via class weights**:
    - Gives higher importance to minority classes during training



Average accuracy: 0.81



Average accuracy: ↑0.87

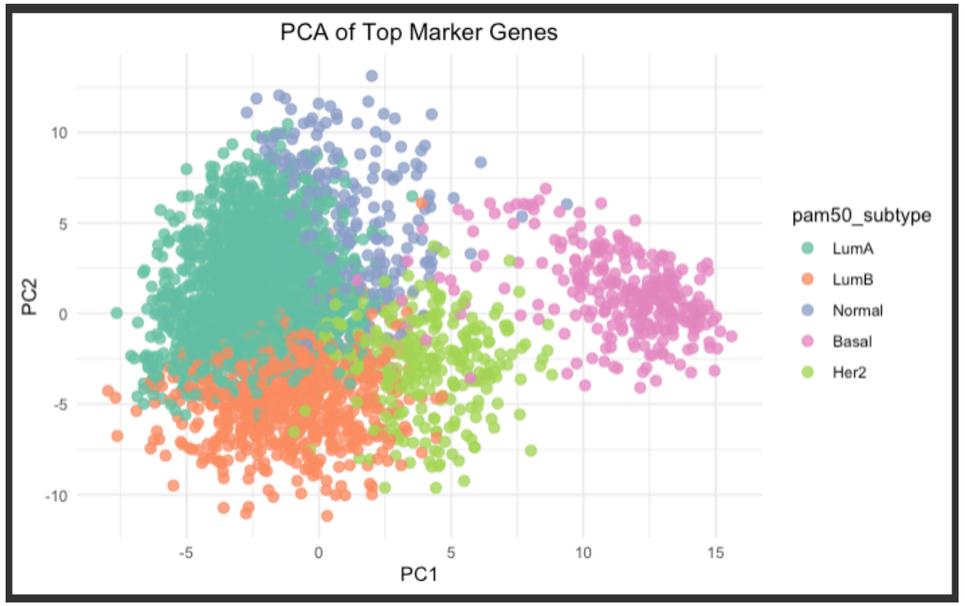
Data analysis on gene expression

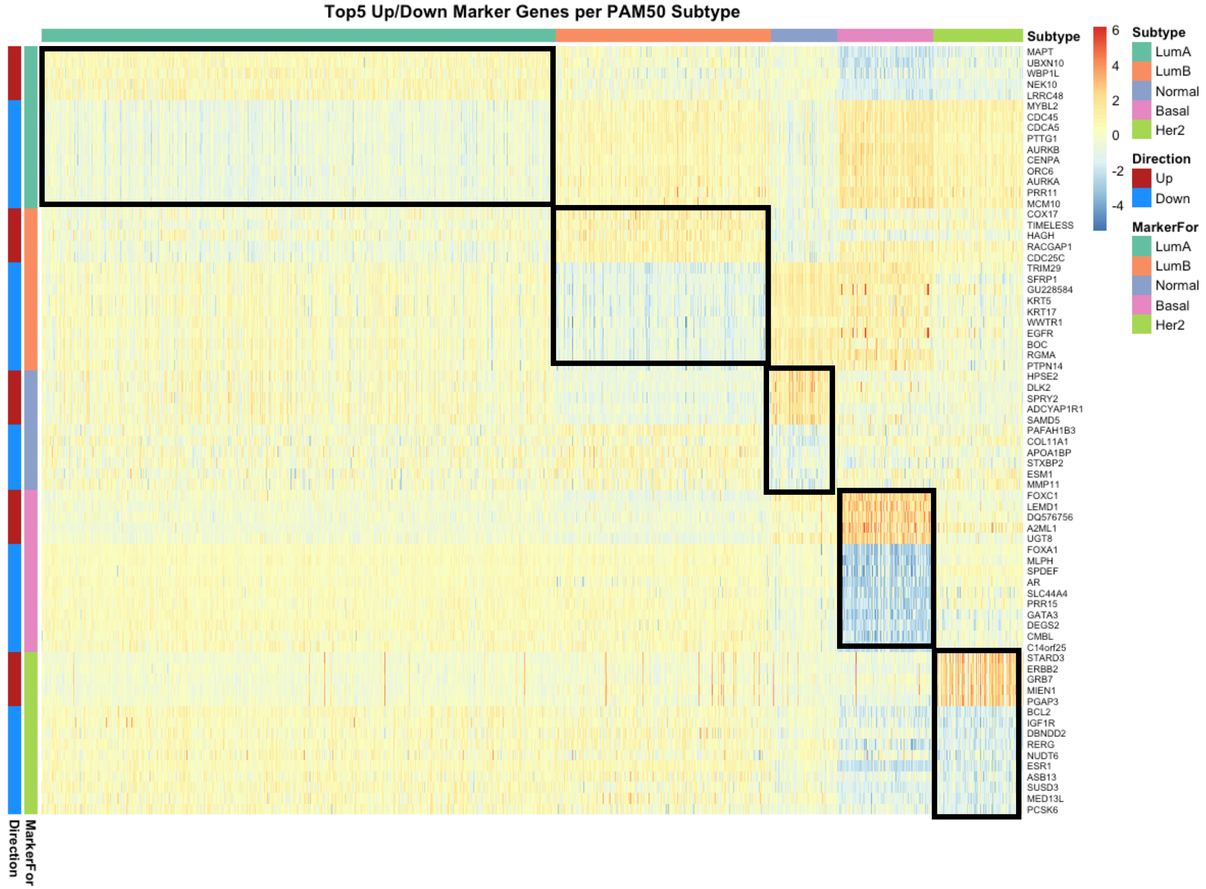
Subtype V.S. gene pattern use-case: limma

limma:

Linear Modeling Framework: Fits gene-wise linear models for differential expression analysis.

Empirical Bayes Moderation: Shrinks variance estimates across genes, improving stability with small sample sizes.





Data analysis on gene expression

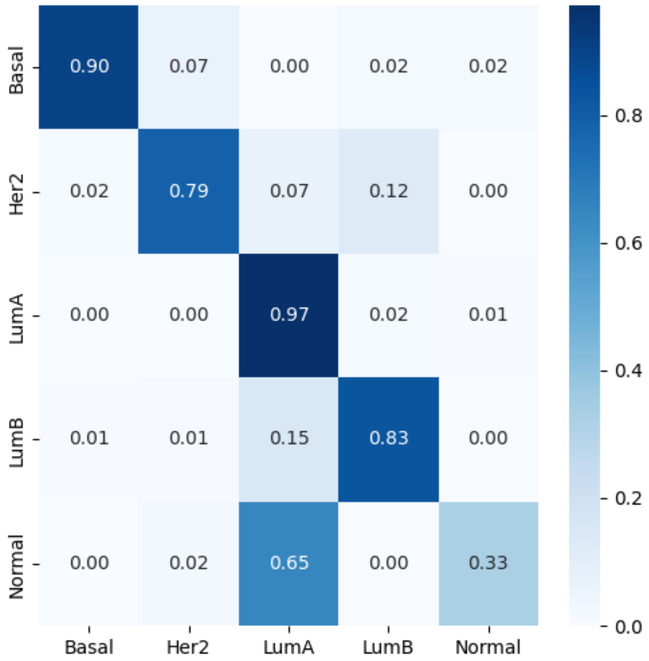
Subtype V.S. gene pattern use-case: limma

Subtype	Upregulated Genes	Biological Meaning (↑)	Downregulated Genes	Biological Meaning (↓)	Conclusion
LumA	MAPT, UBXN10, WBP1L, NEK10, LRRC48	✅ *MAPT* stabilizes microtubules; *NEK10* involved in cell cycle regulation. Indicates differentiated status.	MYBL2, CDC45, CDCA5, AURKA, CENPA, MCM10	✅ Cell cycle and proliferation genes – downregulation reflects low proliferation activity.	✅ Consistent with LumA's well-differentiated, low-proliferation phenotype.
LumB	CDC25C, CCNB1, E2F1, BRCA1, ESPL1	✅ Genes involved in G2/M transition, mitosis, and DNA repair.	KRT5, KRT17, EGFR, SFRP1	✅ *KRT5*, *EGFR*: Basal markers; *SFRP1*: WNT pathway inhibitor often silenced in LumB.	✅ High proliferation, distinct from Basal; very reasonable pattern.
Normal-like	HPSE2, SPRY2, TSHZ2, PAMR1, SYNM	✅ Associated with epithelial structure and differentiation.	COL11A1, ESM1, MMP11	✅ Invasion-related genes suppressed – consistent with low malignancy.	✅ Expression resembles normal/less malignant breast tissue.
Basal	FOXC1, LEMD1, UGT8, PSAT1, SFT2D2	✅ Basal-specific markers; associated with aggressiveness and proliferation.	FOXA1, GATA3, AR, SPDEF	✅ Luminal (hormone receptor) genes suppressed – supports triple-negative profile.	✅ Strong match with Basal's aggressive and hormone-negative profile.
Her2	ERBB2, GRB7, MIEN1, STARD3, PSMD3	✅ Classic Her2-amplified genes in 17q12 region.	BCL2, ESR1, IGF1R, RERG, SUSD3	✅ Estrogen signaling and survival-related genes suppressed.	✅ Matches the canonical Her2-enriched subtype profile.

Data analysis on gene expression

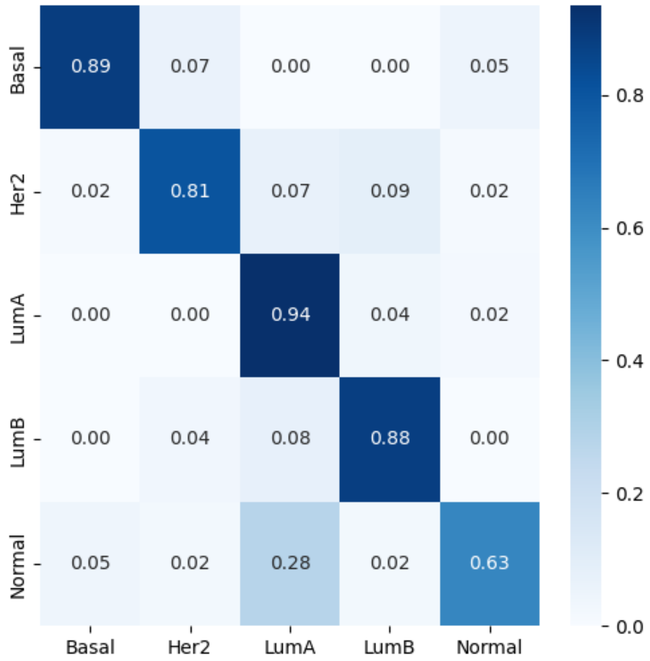
Supervised classification using only feature genes

Random Forest



Average accuracy: ↑ 0.76

Logistic Regression



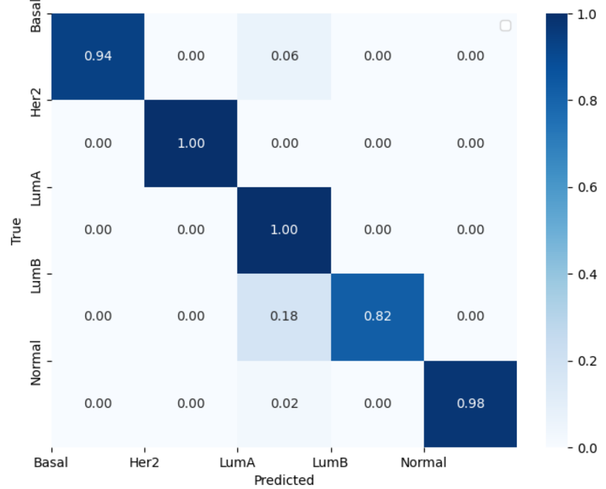
Average accuracy: ↑ 0.83

Using only top 10 selected genes of each subtype selected by limma, the classification performance of RF and LR both increase, indicating the efficiency of feature genes selected by limma, in both biological and computational meaning.

Data analysis on gene expression

CNAttention extending use-case

CNAttention



Average accuracy: 0.95

Subtype	Representative Marker Genes	Supporting Sources
Her2	GRB7, ERBB2, MIEN1, PSMD3, PGAP3	PAM50, TCGA, METABRIC
Luminal A	ESR1, GATA3, AGR3, TBC1D9, ZMYND10	PAM50, METABRIC, Luminal literature
Luminal B	CDC45, CDCA8, POLQ, HJURP, CDCA5, MYBL2, UBE2C, NCAPH	PAM50, proliferation studies
Normal-like	FAM3D, C2orf88, IGSF10, FREM1, EPHB1	METABRIC, expression atlas

The extending use-case of CNAttention in the breast cancer subtype classification on gene expression data, indicating the efficiency of MIL in considering the heterogeneity of gene expression and also the attention mechanism on catching the gene patterns on different breast cancer subtypes.

https://www.biorxiv.org/content/10.1101/2025.05.26.656180v1

Analysis of gene patterns on subtypes

LLM integration

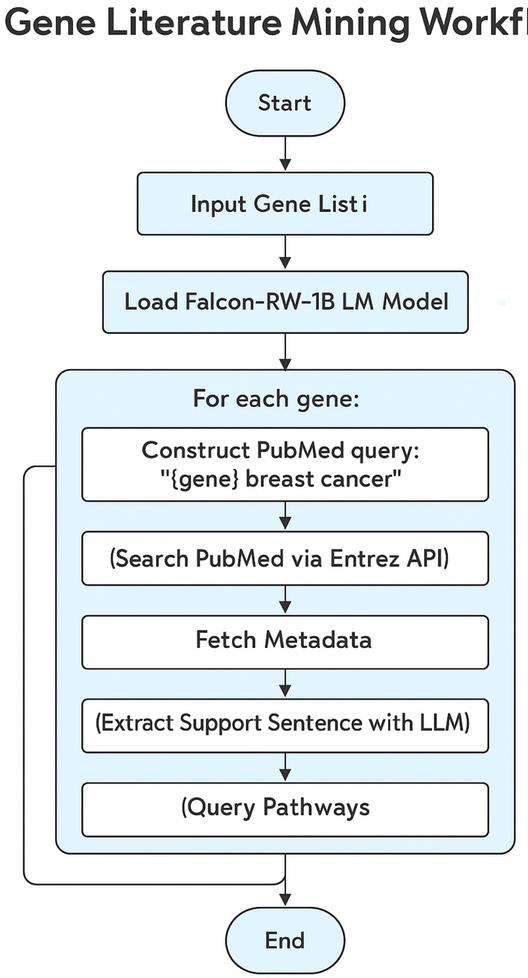
Input:

[					
{gene:	"FOXA1",	subtype:	"LumA",	direction:	"UP"},
{gene:	"SOX11",	subtype:	"Her2",	direction:	"DOWN"},
{gene:	"ERBB2",	subtype:	"Her2"}		
]					

- **Goal of the Table:**
  Help researchers quickly understand the functional roles of each gene in breast cancer via:
  - Evidence-backed literature (PubMed),
  - Mechanistic insight (supportive sentence),
  - Pathway-level context.

Output:

Gene	Subtype	Direction	PMID	Support Sentence	Pathway(s)
FOXA1	LumA	UP	12345678	FOXA1 is upregulated in LumA and promotes ER signaling.	Estrogen signaling pathway
SOX11	Her2	DOWN	23456789	SOX11 is associated with poor prognosis in Her2 tumors.	NA



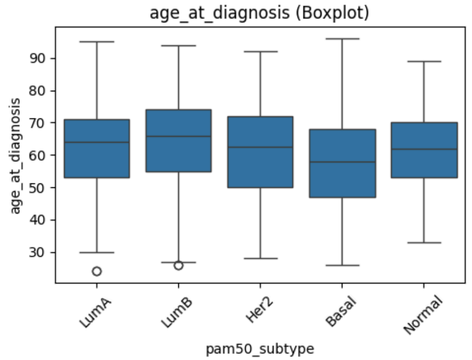
Data analysis on clinical information

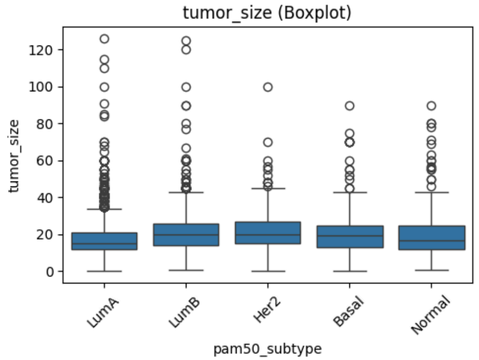
Basic data preview

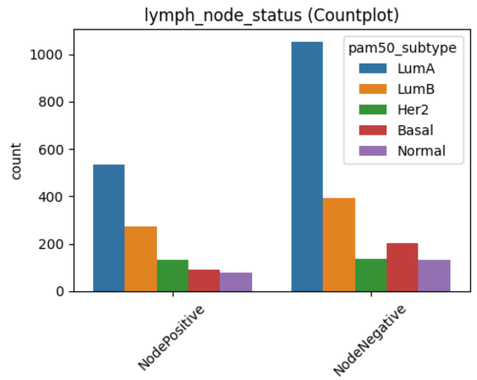
Variable Group	Key Columns (examples)	What it Captures / Why It Matters	Why It's Important
Basic / Technical	`sample_id`, `instrument_model`	Identifiers and sequencing platform or batch information.	*Controls for batch effects* → avoid confounding when comparing expression or outcomes.
Patient & Tumor Characteristics	`age_at_diagnosis`, `tumor_size`, `lymph_node_status`, `lymph_node_group`, `nhg`	Clinical staging, tumour burden and grading.	**Direct prognostic factors** – larger size, positive nodes, and high grade predict poorer survival.
Receptor / Proliferation Status	`er_status`, `pgr_status`, `her2_status`, `ki67_status`	IHC-based hormone-receptor and HER2 status; Ki-67 indicates proliferation.	**Defines standard clinical subtypes** and guides therapy (e.g. endocrine therapy for ER+; HER2-targeted therapy).
Algorithm-Predicted Scores	`*_prediction_mgc`, `*_prediction_sgc` for ER, PR, HER2, Ki67, NHG	Machine-generated surrogate markers from image or multi-gene classifiers.	Can be **cross-validated against true IHC**; useful when lab data missing or noisy.
Outcomes	`overall_survival_days`, `overall_survival_event`	Time-to-event information for survival analyses.	Gold-standard endpoint for prognostic modelling (Kaplan-Meier, Cox).
Therapy Indicators	`endocrine_treated`, `chemo_treated`	Whether patient received endocrine or chemotherapy.	Enables **treatment-response studies** and interaction analyses with subtype.
Molecular Subtype (Label)	`pam50_subtype`	PAM50 expression-based subtype ground truth (LumA, LumB, Basal, Her2, Normal).	**Target variable** for prediction models; baseline for differential-gene and clinical comparisons.

Data analysis on clinical information

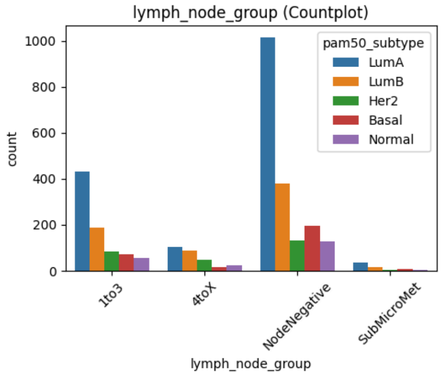
Basic data preview

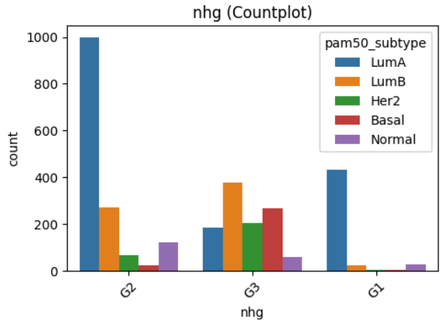


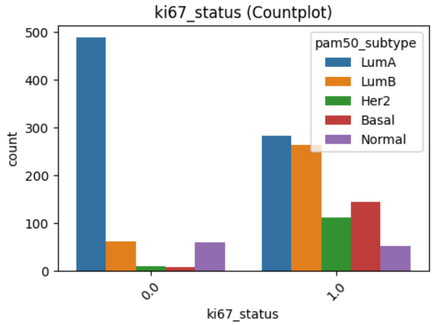




- **Clinical Parameters** (age_at_diagnosis, tumor_size, lymph_node_status):
  - **Her2** patients appear younger
  - **Basal** subtype may show larger tumor sizes
  - **LumA** has fewer positive lymph nodes
    ➤ Suggests different progression profiles—evaluate via boxplots or ANOVA.

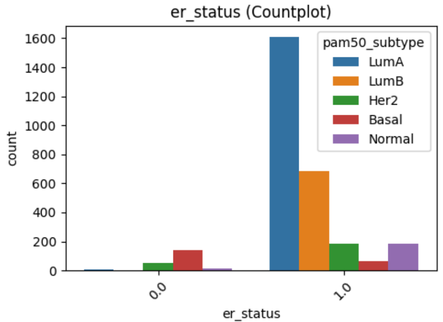


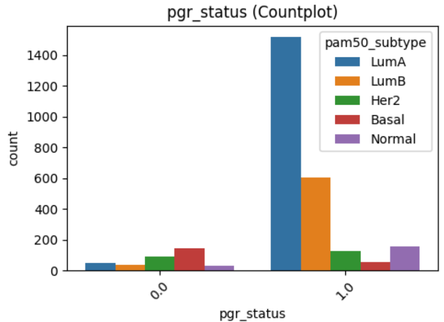


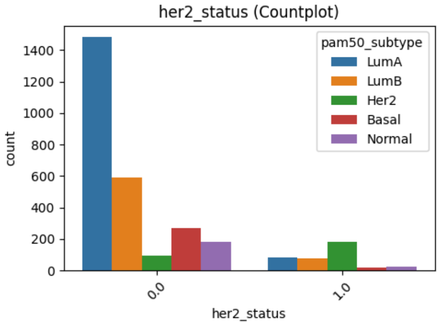


Proliferation Markers (ki67_status, nhg) differ across subtypes:

- **Basal & LumB** show higher proliferation (Ki67+, NHG=G3)
  ➤ Indicates more aggressive tumor behavior—consider modeling this association.





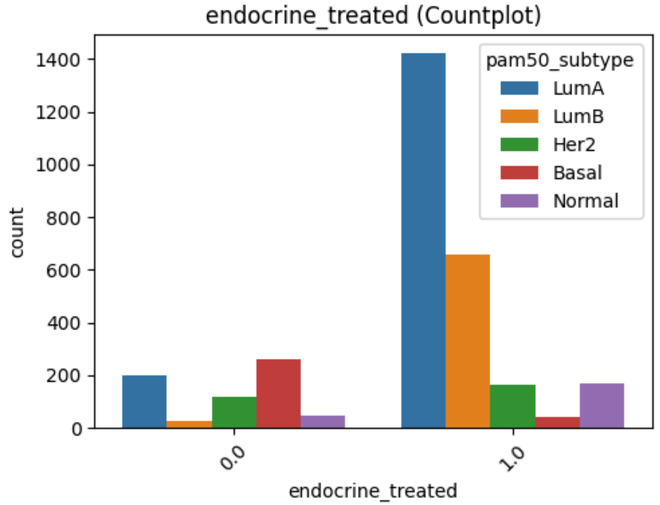


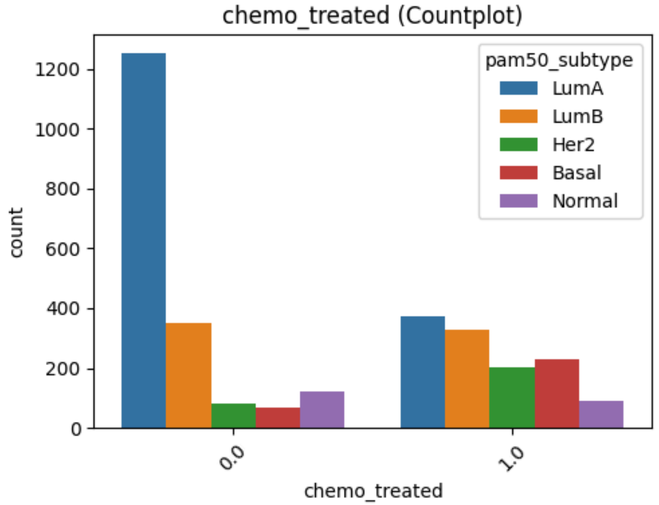
Hormone Receptor Status (er_status, pgr_status, her2_status) shows strong subtype-specific patterns:

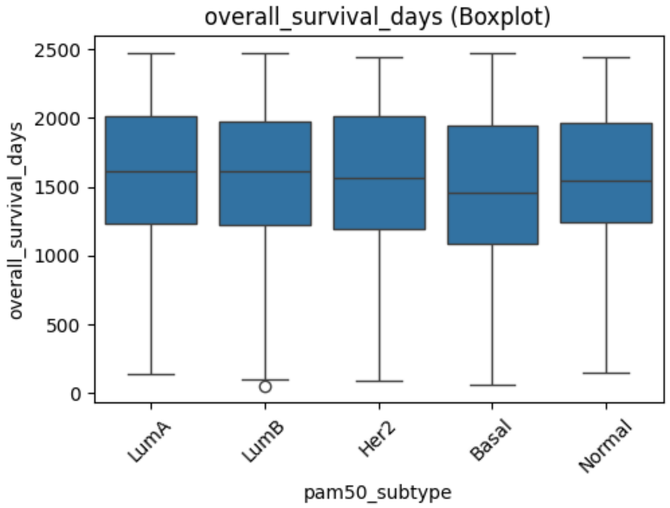
- **LumA/LumB**: High ER/PGR positivity
- **Her2-enriched**: High HER2 positivity
- **Basal-like**: Predominantly triple-negative
  ➤ Suggests a clear biological distinction—worth confirming with chi-square tests.

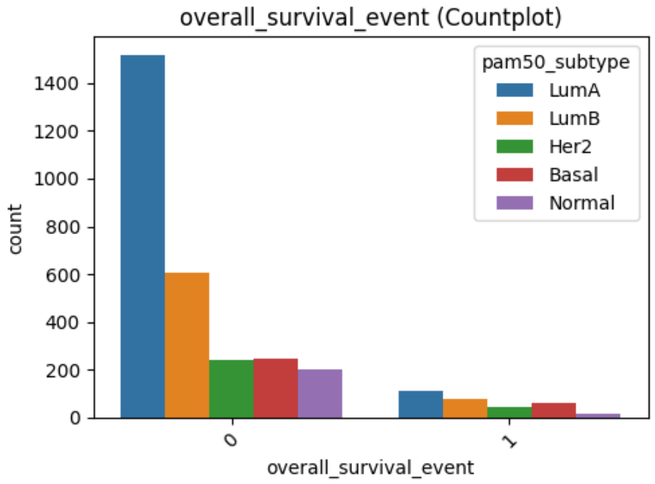
Data analysis on clinical information

Basic data preview









Treatment Response Differences:

- **Endocrine therapy** is primarily seen in **LumA/LumB** patients
- **Chemotherapy** is more common in **Basal** and **Her2** subtypes
  ➤ Highlights personalized treatment strategies—quantify proportions and response.

Survival Trends:

- **LumA** patients have the longest overall survival
- **Basal/Her2** subtypes tend toward shorter survival
  ➤ Worth analyzing with Kaplan-Meier curves + log-rank tests.

Data analysis on clinical information

Categorical/continuous variables vs. PAM50 subtype

Chi-square Tests (Categorical variables vs. PAM50 subtype)

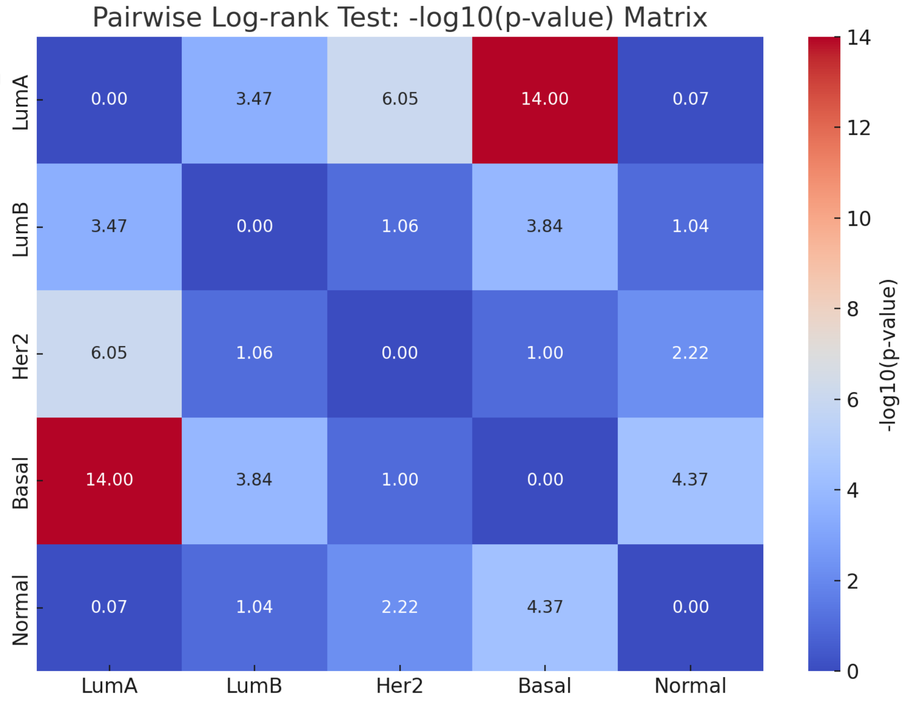
Variable	p-value	Significance
ER Status	2.46×10^{-289}	*** Extremely significant association
PR Status	3.35×10^{-203}	*** Extremely significant association
HER2 Status	1.67×10^{-167}	*** Extremely significant association
Ki67 Status	7.13×10^{-77}	*** Strong association
Lymph Node Status	4.30×10^{-7}	** Significant association

All listed categorical clinical variables show statistically significant associations with PAM50 subtypes. This suggests that hormone receptor status (ER, PR), HER2, proliferation index (Ki67), and nodal status are strongly subtype-dependent.

Kruskal-Wallis / ANOVA (Continuous variables vs. PAM50 subtype)

Variable	Kruskal p-value	ANOVA p-value	Significance
Tumor Size	1.23×10^{-28}	5.29×10^{-18}	*** Highly significant
Age at Diagnosis	3.13×10^{-11}	9.67×10^{-13}	*** Significant difference
Overall Survival Days	2.85×10^{-3}	2.60×10^{-4}	** Moderate significance

All continuous variables also show significant differences across PAM50 subtypes, with tumor size having the strongest effect. This highlights distinct clinical behavior and prognosis across molecular subtypes.



LumA shows significantly better survival than LumB, Her2, and Basal (all $p < 0.001$).

No significant survival difference between **LumA and Normal** — supports similarity.

Basal subtype has significantly **worse survival** than LumA, LumB, and Normal.

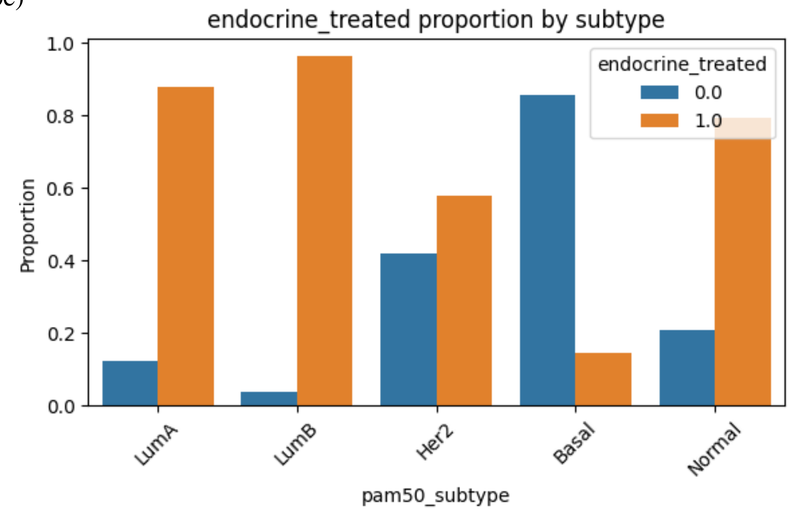
Her2 vs LumB and **Her2 vs Basal** are **not significantly different**, suggesting closer prognosis.

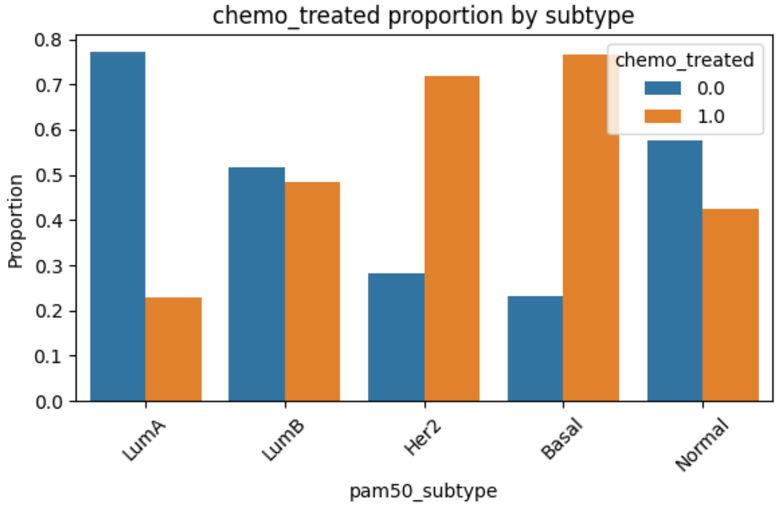
Data analysis on clinical information

Treatment V.S. subtypes

Logistic regression (chemotherapy received vs. PAM50 subtype)
LumA as baseline

Subtype	Coefficient	Odds Ratio (exp(coef))	Significance
LumB	1.151	~3.16× higher	***
Her2	2.147	~8.56× higher	***
Basal	2.409	~11.12× higher	***
Normal	0.911	~2.49× higher	***





Compared to LumA, all other subtypes were **significantly more likely** to receive chemotherapy, especially **Basal** and **Her2**.

This aligns with clinical understanding: more aggressive subtypes tend to receive more intensive treatment.

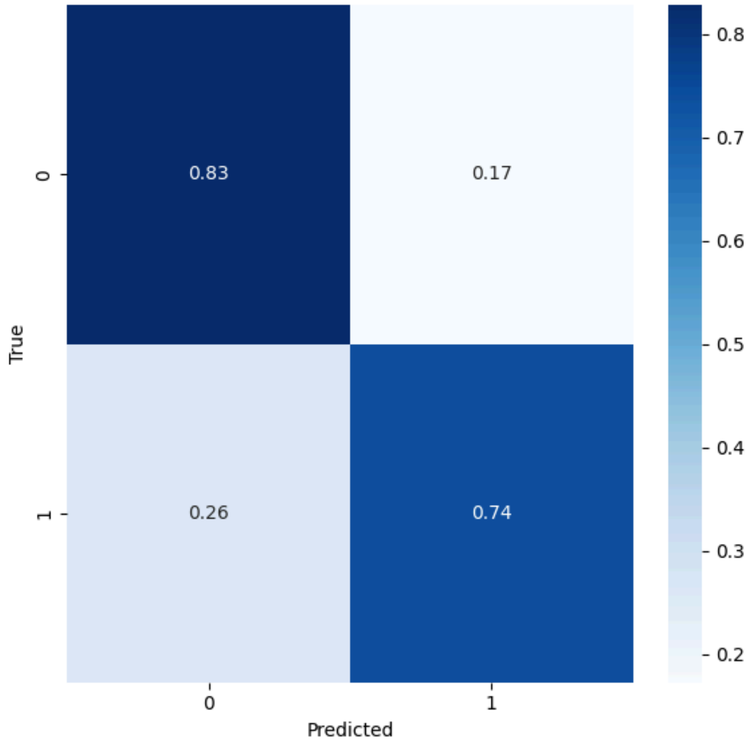
Treatment Strategy Varies by Subtype:

- LumA/LumB: High endocrine therapy usage, low chemo
- Basal/Her2: High chemotherapy, low endocrine therapy
- Treatment decisions align with biological features of subtypes

Data analysis on clinical information

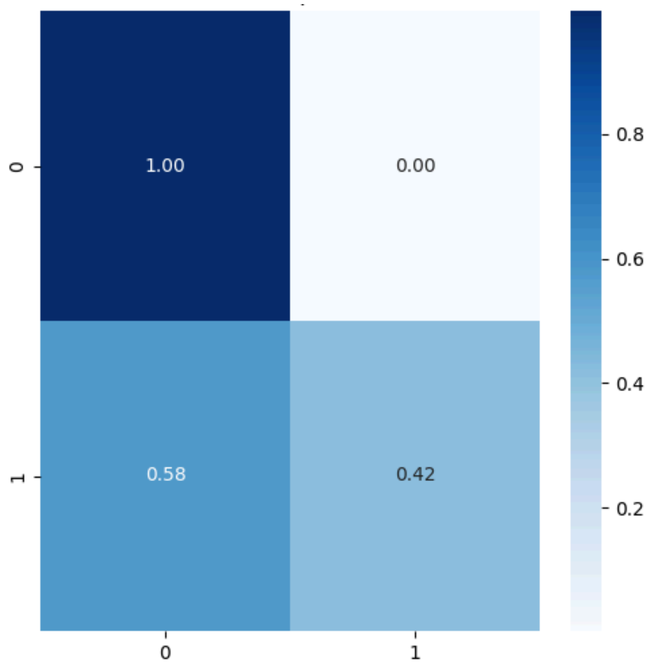
Survival event prediction

FFNN



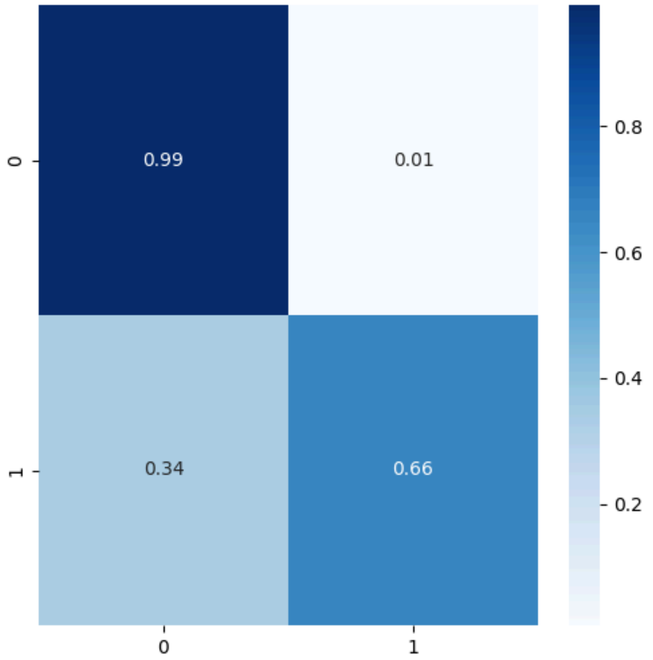
Accuracy: 0.78

Random Forest



Accuracy: 0.71

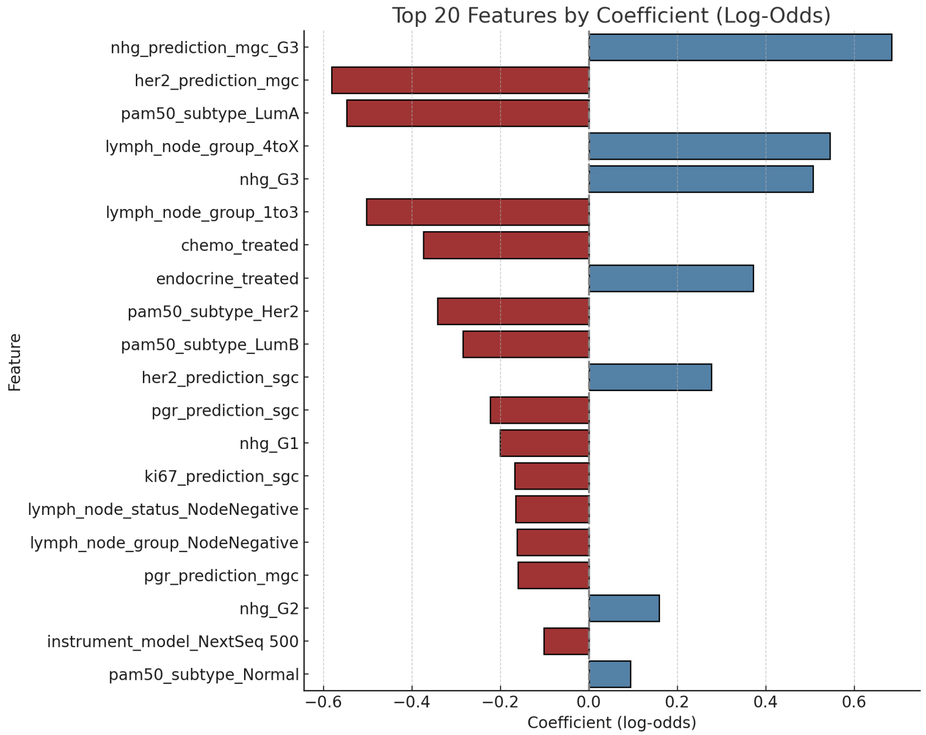
Logistic Regression



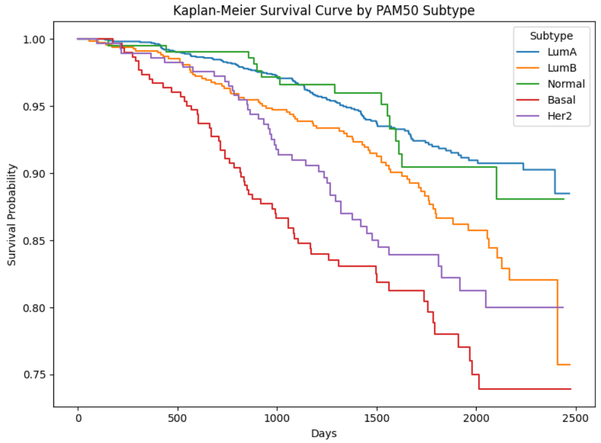
Accuracy: 0.83

Data analysis on clinical information

Logistics Regression Coefficient V.S. Survival Risk



Model Insight	Clinical Validity
High-grade tumors (G3) strongly ↑ risk	✅ Widely established
LumA subtype → lower risk	✅ Confirmed in survival studies
Node-positive → higher risk	✅ Key part of TNM staging
Molecular predictions (e.g., Her2) dominate over raw status	✅ Reflect deeper signals
Some expected features (age, size) contribute little	⚠️ Check model structure & collinearity



Summary

Gene Expression & Subtype Analysis

- Extensive exploration of gene expression data (3,200+ samples, 16k+ genes) across breast cancer subtypes (PAM50).
- Applied dimensionality reduction, clustering, and supervised learning:
  - Unsupervised: No clear separation across subtypes.
  - Supervised: Accuracy ↑ to 0.87 (e.g., FFNN, logistic regression).
  - Feature selection (limma) improved both biological interpretability and classifier performance.

LLM Integration for Literature Mining

- Built a lightweight pipeline with GPT-based models:
  - PubMed retrieval → Abstract extraction → Keyword/technique mining.
  - Compared BioGPT, Mistral, and Falcon: Falcon provides balance between cost and control.
- Automatically identified techniques used in subtype classification tasks.
- Efficiently matched genes with biological pathways and literature support.

Clinical Information Modeling

- Statistical tests showed strong associations between clinical markers (ER, PR, HER2, Ki67, etc.) and subtypes.
- Models (FFNN, RF, LR) achieved up to 0.83 accuracy in survival prediction.
- Logistic regression coefficients matched clinical expectations (e.g., grade, nodal status).

Outlook

Subtype classification benefits significantly from **supervised learning** and **domain-informed feature selection**.

LLMs (even smaller open models) can automate literature mining and provide real-time biological reasoning support.

Integrating multi-source data (expression + clinical + literature) enhances both biological insight and predictive power.