Learning in Limited Data Settings

Advancing Personalized Medicine in Cancer Treatment Planning*

Vaibhav Rajan
Google DeepMind

*work funded at National University of Singapore

One day....

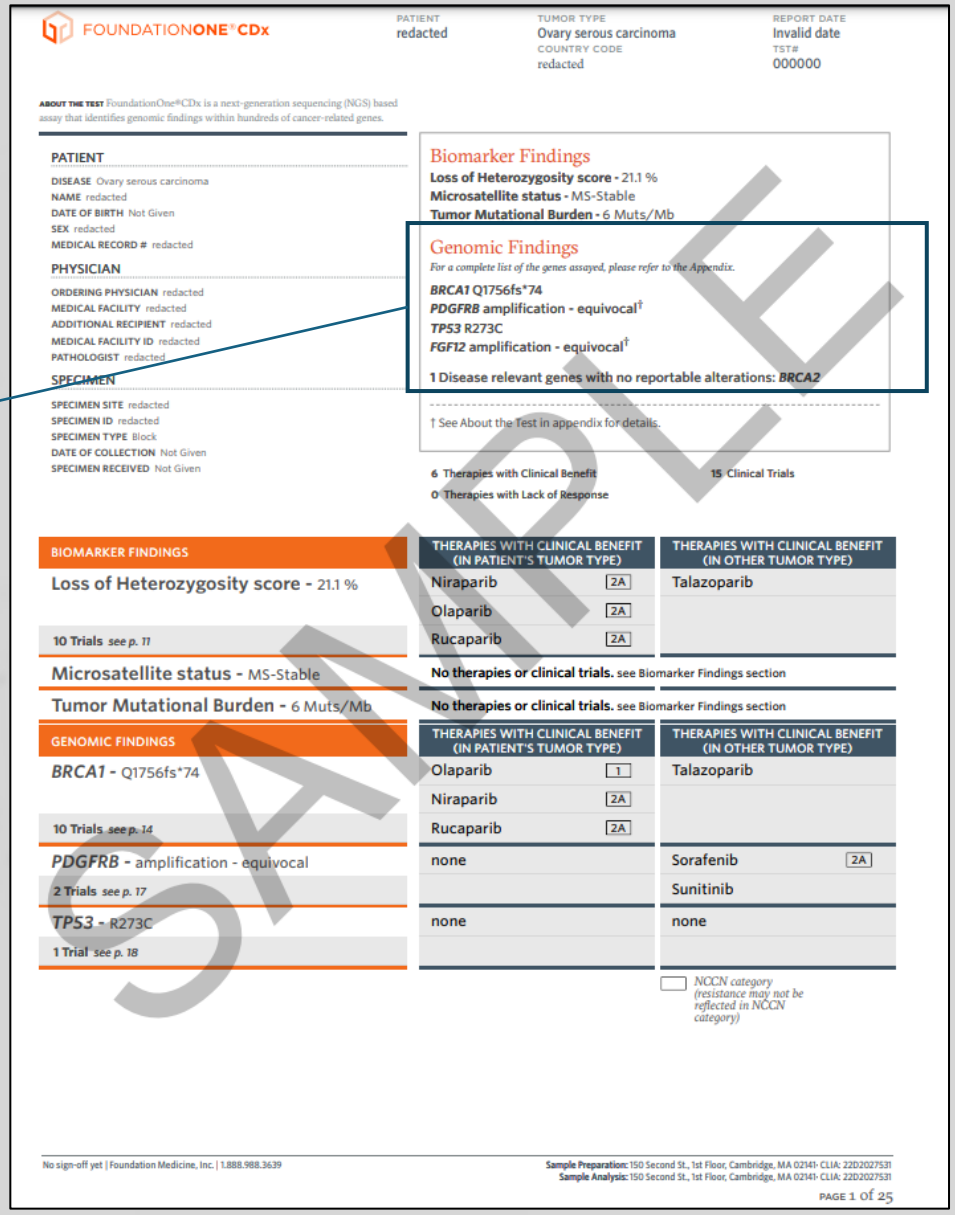
BRCA1 Q1756fs*74

PDGFRB amplification

TP53 R273C

FGF12 amplification

Mutations identified



FOUNDATIONONE®CDx

PATIENT redacted

TUMOR TYPE Ovary serous carcinoma

COUNTRY CODE redacted

REPORT DATE Invalid date

TST# 000000

ABOUT THE TEST FoundationOne®CDx is a next-generation sequencing (NGS) based assay that identifies genomic findings within hundreds of cancer-related genes.

PATIENT

DISEASE Ovary serous carcinoma
NAME redacted
DATE OF BIRTH Not Given
SEX redacted
MEDICAL RECORD # redacted

PHYSICIAN

ORDERING PHYSICIAN redacted
MEDICAL FACILITY redacted
ADDITIONAL RECIPIENT redacted
MEDICAL FACILITY ID redacted
PATHOLOGIST redacted

SPECIMEN

SPECIMEN SITE redacted
SPECIMEN ID redacted
SPECIMEN TYPE Block
DATE OF COLLECTION Not Given
SPECIMEN RECEIVED Not Given

Biomarker Findings

Loss of Heterozygosity score - 21.1 %
Microsatellite status - MS-Stable
Tumor Mutational Burden - 6 Muts/Mb

Genomic Findings

For a complete list of the genes assayed, please refer to the Appendix.

BRCA1 Q1756fs*74
PDGFRB amplification - equivocal†
TP53 R273C
FGF12 amplification - equivocal†

1 Disease relevant genes with no reportable alterations: ***BRCA2***

† See About the Test in appendix for details.

6 Therapies with Clinical Benefit
0 Therapies with Lack of Response
15 Clinical Trials

BIOMARKER FINDINGS	THERAPIES WITH CLINICAL BENEFIT (IN PATIENT'S TUMOR TYPE)	THERAPIES WITH CLINICAL BENEFIT (IN OTHER TUMOR TYPE)
Loss of Heterozygosity score - 21.1 %	Niraparib 2A Olaparib 2A Rucaparib 2A	Talazoparib
10 Trials see p. 11		
Microsatellite status - MS-Stable	No therapies or clinical trials. see Biomarker Findings section	
Tumor Mutational Burden - 6 Muts/Mb	No therapies or clinical trials. see Biomarker Findings section	
GENOMIC FINDINGS	**THERAPIES WITH CLINICAL BENEFIT (IN PATIENT'S TUMOR TYPE)**	**THERAPIES WITH CLINICAL BENEFIT (IN OTHER TUMOR TYPE)**
BRCA1 - Q1756fs*74	Olaparib 1 Niraparib 2A Rucaparib 2A	Talazoparib
10 Trials see p. 14		
PDGFRB - amplification - equivocal	none	Sorafenib 2A Sunitinib
2 Trials see p. 17		
TP53 - R273C	none	none
1 Trial see p. 18		

NCCN category (resistance may not be reflected in NCCN category)

SAMPLE

No sign-off yet | Foundation Medicine, Inc. | 1.888.988.3639

Sample Preparation: 150 Second St., 1st Floor, Cambridge, MA 02141- CLIA: 22D2027531
Sample Analysis: 150 Second St., 1st Floor, Cambridge, MA 02141- CLIA: 22D2027531

PAGE 1 of 25

Targeted therapy: “one gene, one target”

Majority of patients remain untreated by targeted therapies

Can AI help us identify the right drug for such cancer patients?

NORMAL CELL AND CANCER CELL DEVELOPMENT

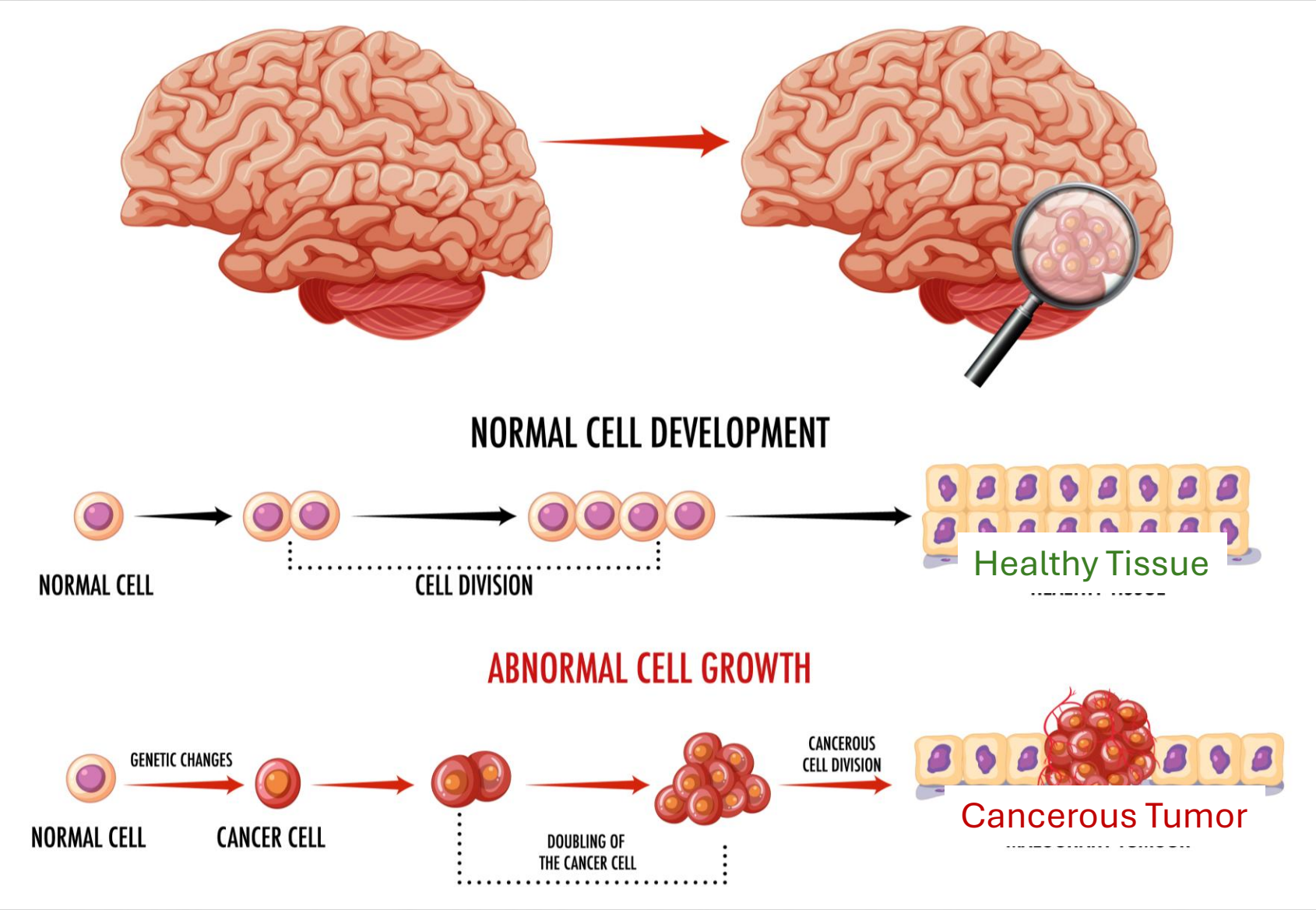
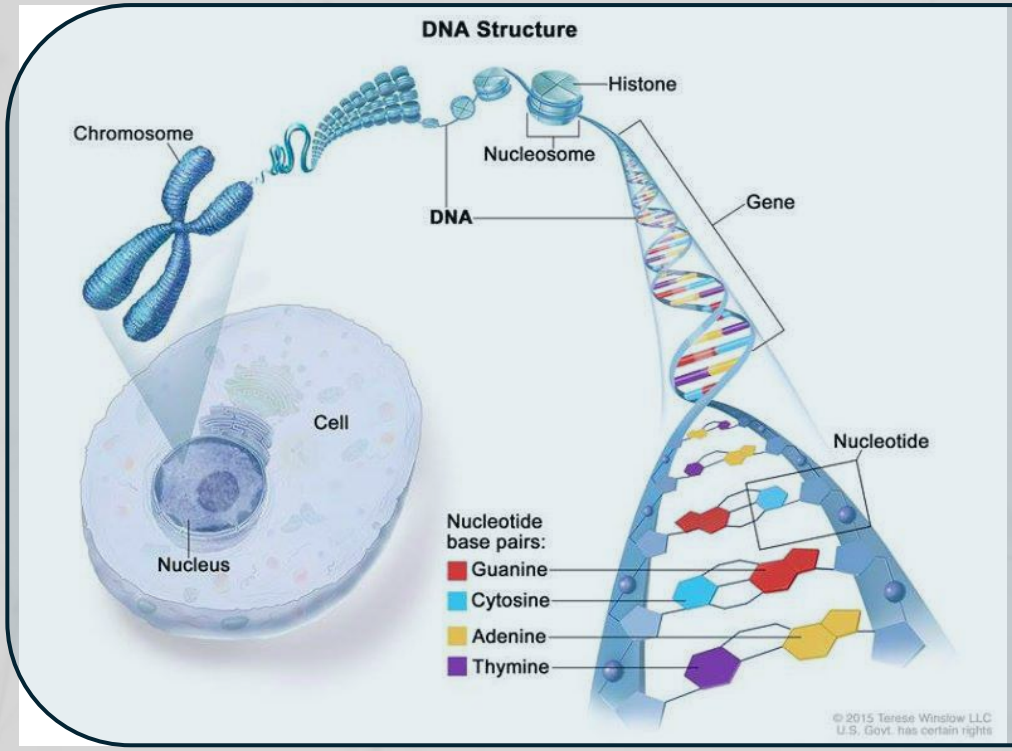


Image by brgfx on Freepik

CANCER

Cancer is a ***genetic disease***, i.e., it is caused by changes to genes (mutations)

Cancer is a leading cause of death worldwide (one-in-six deaths, 2020)



Each cell in our body contains 23 pairs of chromosomes

Each chromosome is a sequence of “base pairs” , bases are A, C, G, T

Gene: **subsequence of the chromosome** which has functional importance

~20,000 genes have been identified

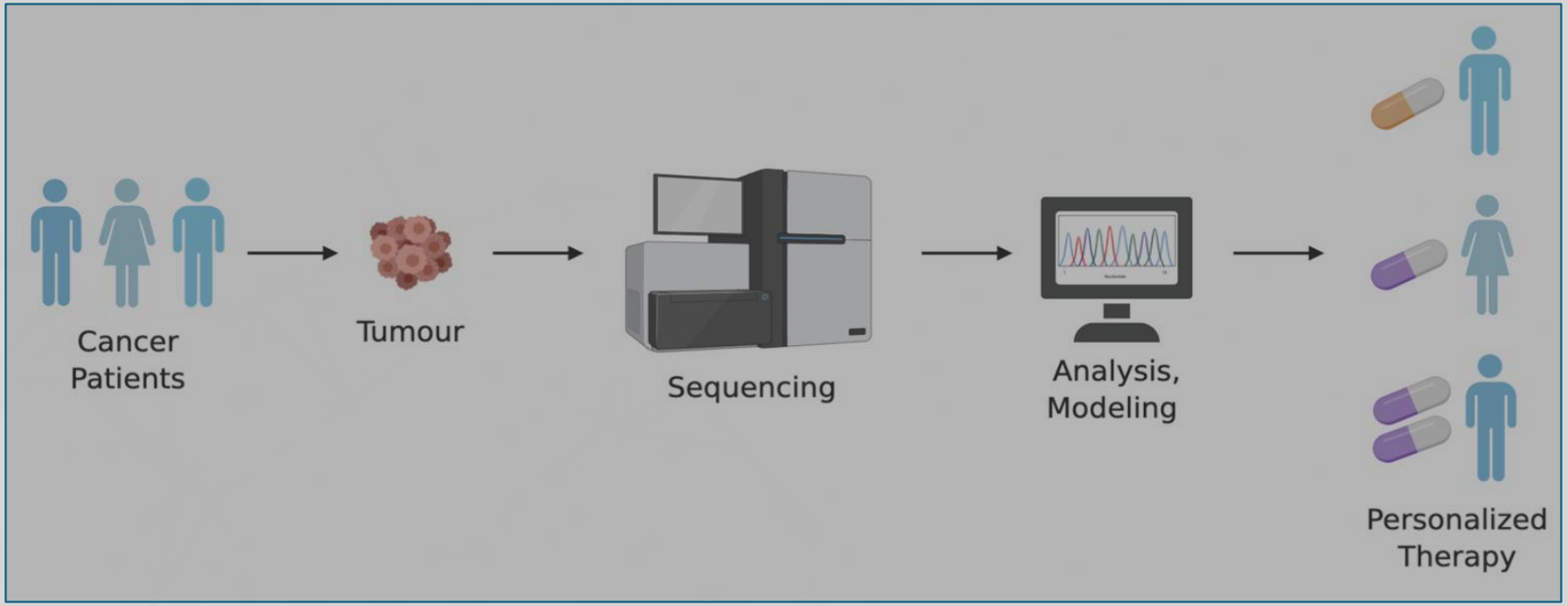
https://www.cancer.gov/about-cancer/understanding/what-is-cancer

https://www.who.int/news-room/fact-sheets/detail/cancer

CANCER TREATMENT

- Treatment remains challenging
  - Complex disease: ***Every cancer has an individual set of mutations***
  - A drug that works for one cancer patient, might have absolutely no effect on another

- Treatment must be tailored to each patient: **personalized therapy**



https://www.worldwidecancerresearch.org/news-opinion/2021/march/why-havent-we-cured-cancer-yet/
https://en.wikipedia.org/wiki/Personalized_medicine

Cancer patient

Genomic sequencing

Mutation 1
Mutation 2
...
Mutation N

Challenge in precision oncology

Clinical trials

Mutation 1

Mutation 2

Mutation 1
Mutation 2
...
Mutation N

Clinical trials intractable!

Icons from Flaticon

CANCER GENOMICS DATA

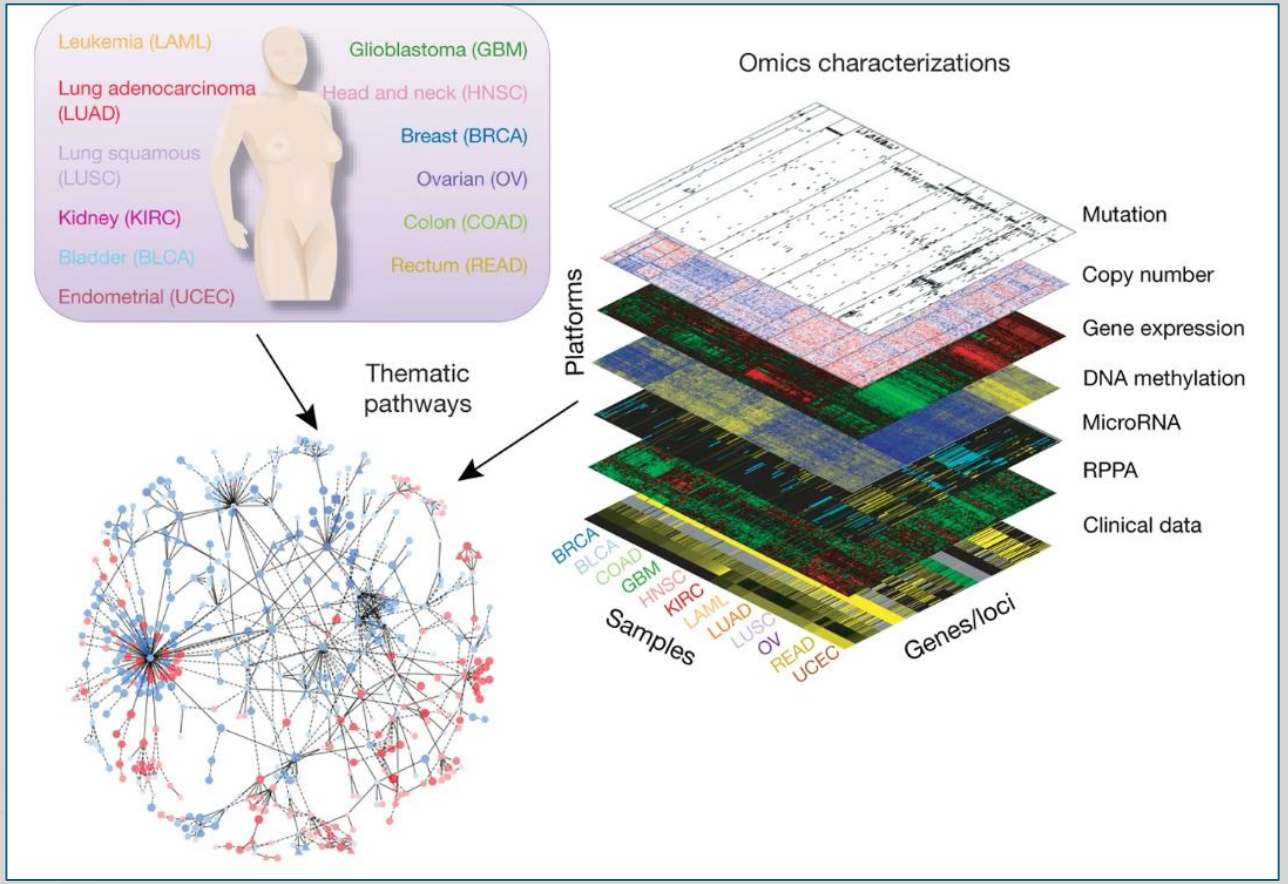
The Cancer Genome Atlas (TCGA)

Since 2006

> 11,000 patients

2.5 PetaBytes of Data

33 cancer types



- Many similar data collection efforts to understand cancer

Weinstein, J. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* 2013

REPRESENTING GENOMIC DATA

Raw sequence (rarely used)

...ACCTTTCGGCCGGACCCCC...

Mutation Vector

Genes of interest:

G1	G2	G3	G4	G5	G6	G7	G8	G9	...
1	0	1	0	0	0	0	1	0	...

Binary indicator: 1 → mutation in gene, 0 → no mutation

Gene Expression Vector

Genes of interest:

G1	G2	G3	G4	G5	G6	G7	G8	G9	...
6	2.1	3	0	0	0	0	1	0	...

Count or real value: indicates activity level of gene

Sequence of Mutations

G1: R273C, G1: S1372L, G2: L145V ...

In gene G1, at location 273 a mutation changed R to C in the protein

DRUG RESPONSE MEASUREMENTS

1. Response Evaluation Criteria In Solid Tumors (RECIST)

- Standard way to measure how well a cancer patient responds to treatment.

Good response (label +1)

Bad response (label -1)

RECIST	
CR	Complete Response
PR	Partial Response
PD	Progressive Disease
SD	Stable Disease

2. Progression-free Survival (PFS)

- The length of time during and after the treatment (days/months/years), that a patient lives without the cancer getting worse.

DRUG RESPONSE PREDICTION (DRP)

One day…

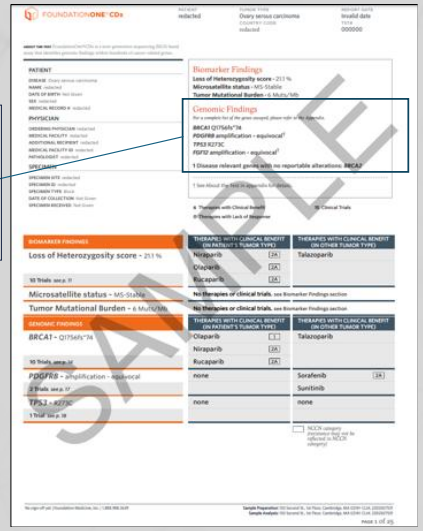
BRCA1 Q1756fs*74

PDGFRB amplification

TP53 R273C

FGF12 amplification

Mutations identified



Targeted therapy: “one gene, one target”

Majority of patients remain untreated by targeted therapies

Can AI help us identify the right drug for such cancer patients?

➢ **Given:**

- a patient’s genomic profile and
- a drug

➢ **Will the response of the patient to the drug be good?**

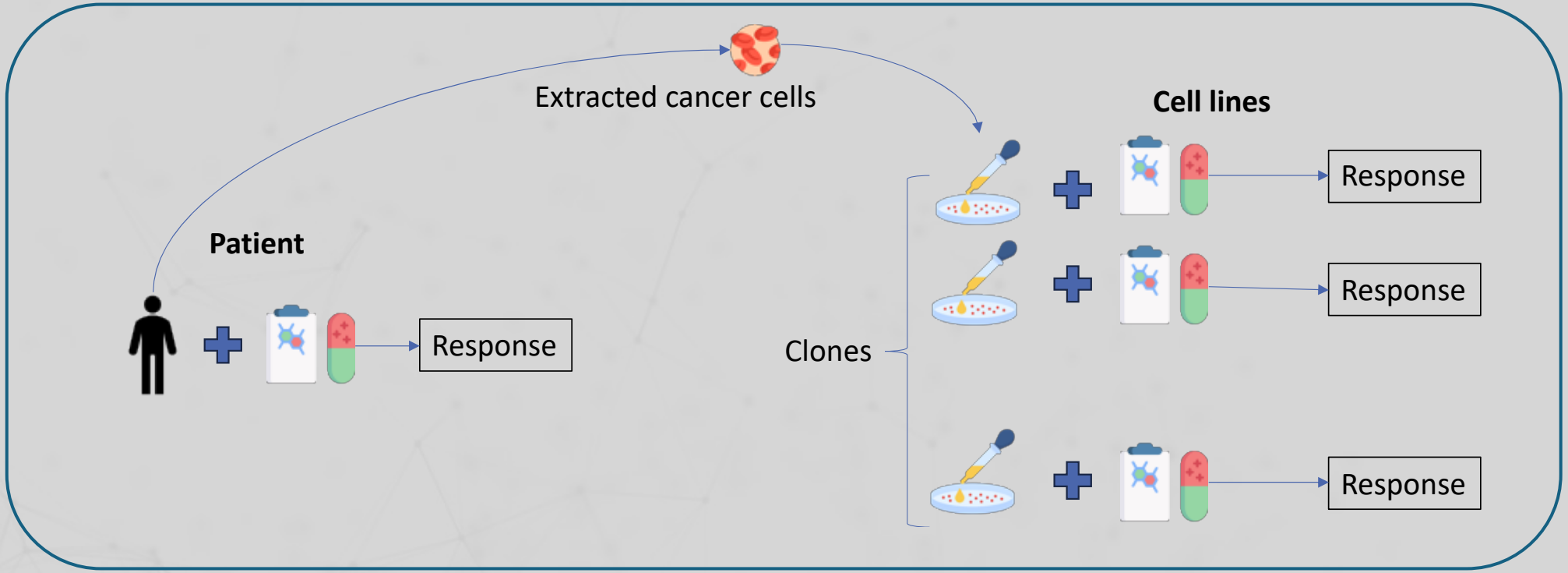
DRUG RESPONSE PREDICTION (DRP)

- X: Patient's genomic data (e.g., mutation vector or gene expression)
- Y: RECIST value after administering drug d

- $Y \sim f_d(X)$ → binary classification
- Challenge
  - X: abundant, but…
  - Y: extremely limited for any drug d

- Why?
  - Each patient is given one/few drugs, counterfactual unknown

CELL LINES: A RELATED "DOMAIN"



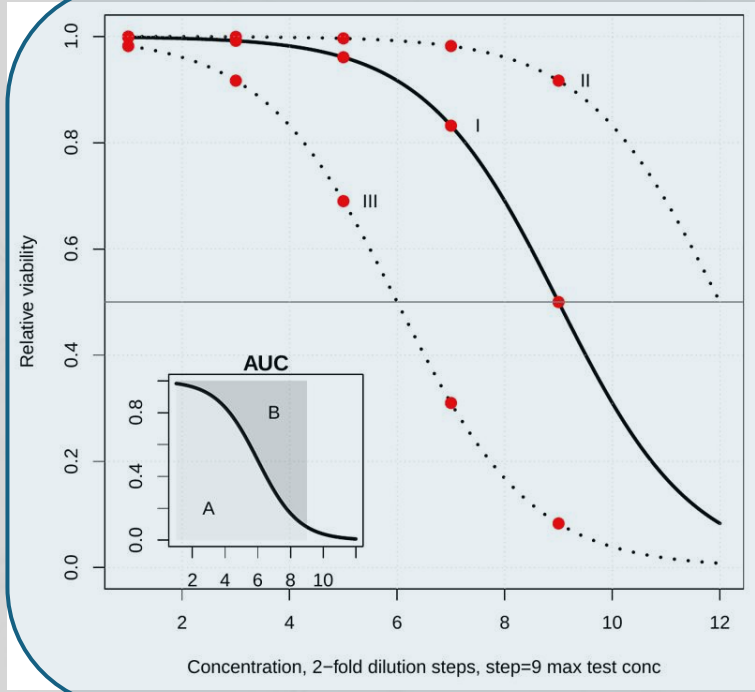
- Extract cancer cells and clone them in lab (living cells, continue growing)
  - Ensures each cell has same genomic data (X)

- Administer multiple drugs on cell lines, measure response Y

DRUG RESPONSE MEASUREMENT IN CELL LINES

Area under the Dose Response Curve (AUDRC)

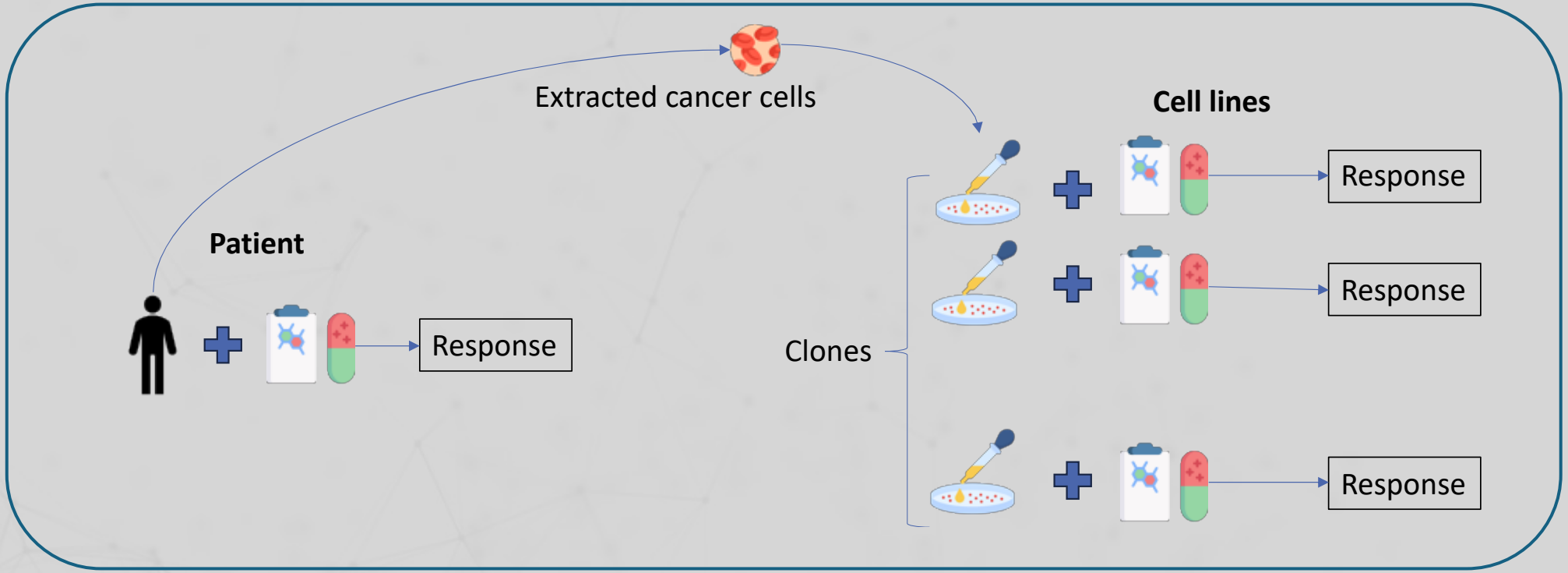
Real-valued [0,1]



- Administer progressively increasing concentration (X-axis) of drug and measure the amount of cancer cells (Y-axis) killed: Dose Response Curve (DRC)
- Lesser concentration kills more cells → more effective drug →Lower AUDRC
- E.g. efficacy of III > I > II

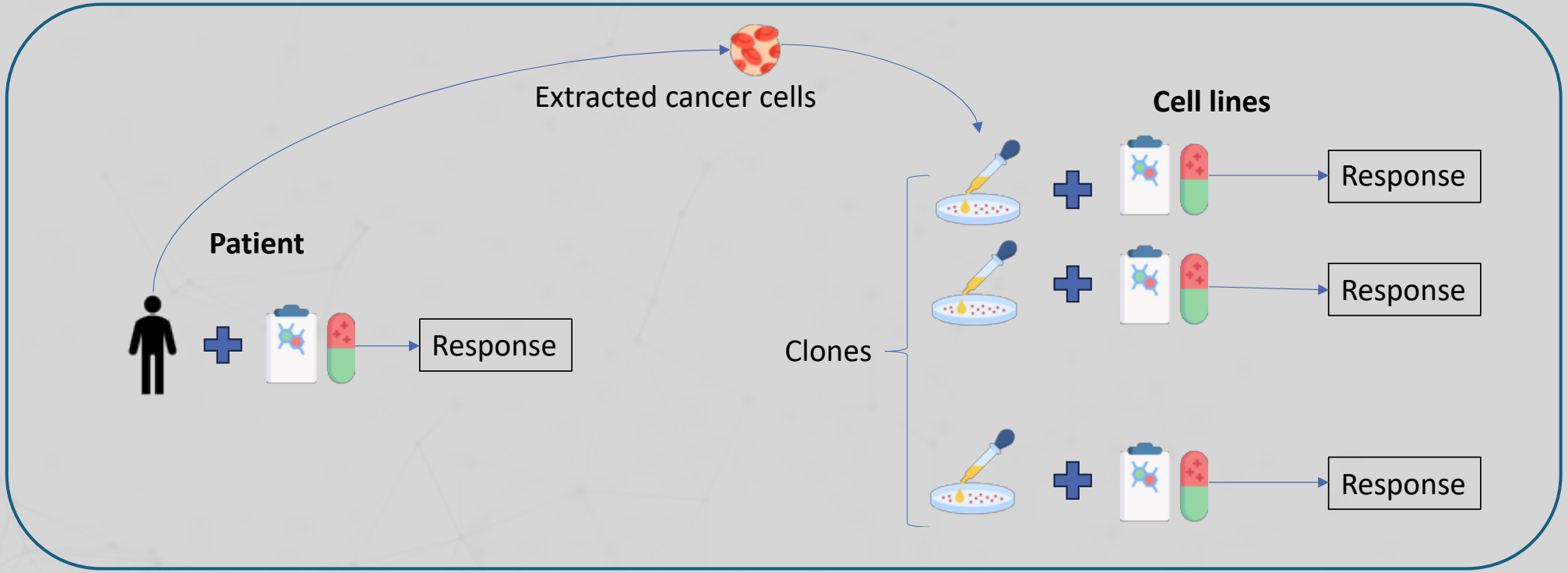
Vis, D. J. et al. Multilevel models improve precision and speed of IC50 estimates. *Pharmacogenomics 2016.*

CELL LINES: A RELATED "DOMAIN"



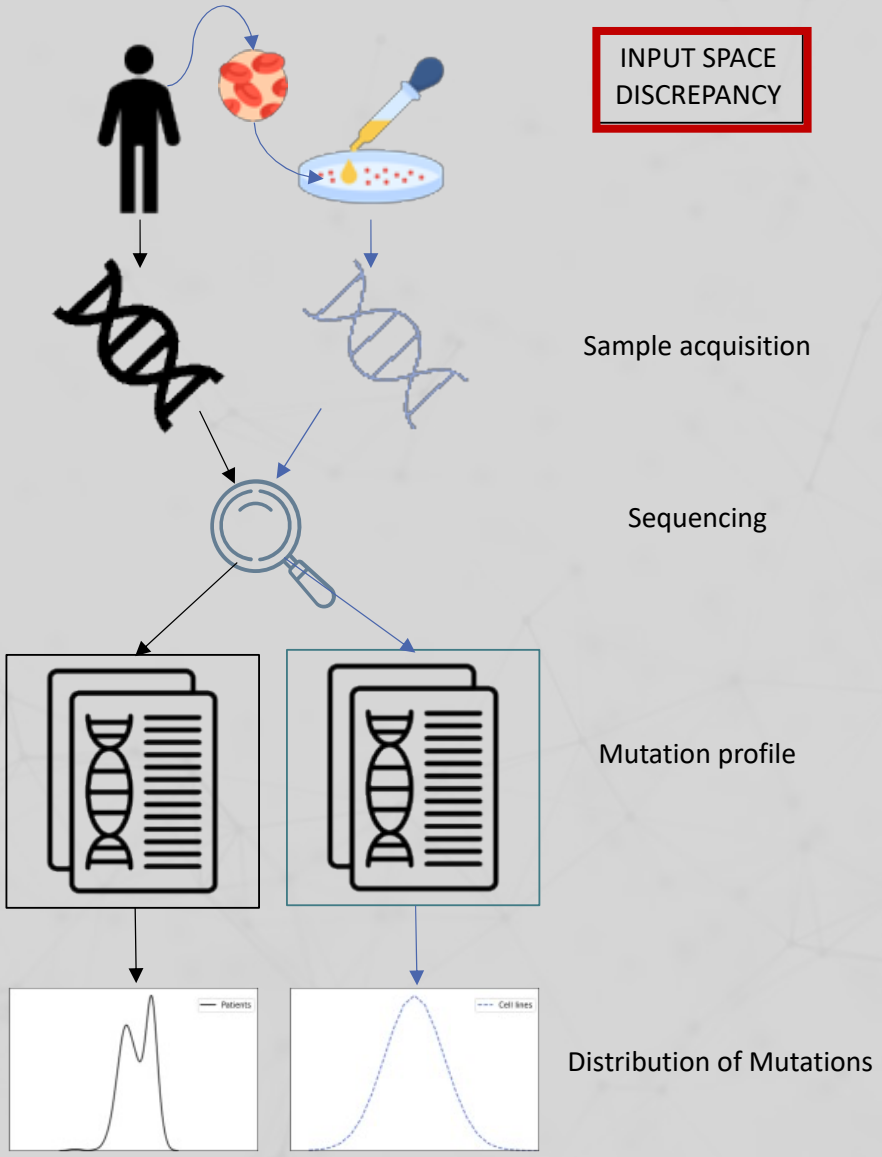
- Extract cancer cells and clone them in lab (living cells, continue growing)
  - Ensures each cell has same genomic data (X)
- Administer multiple drugs on cell lines, measure response Y
- Can a Drug Response Prediction model on cell lines $Y \sim f_d(X)$ work for patients?

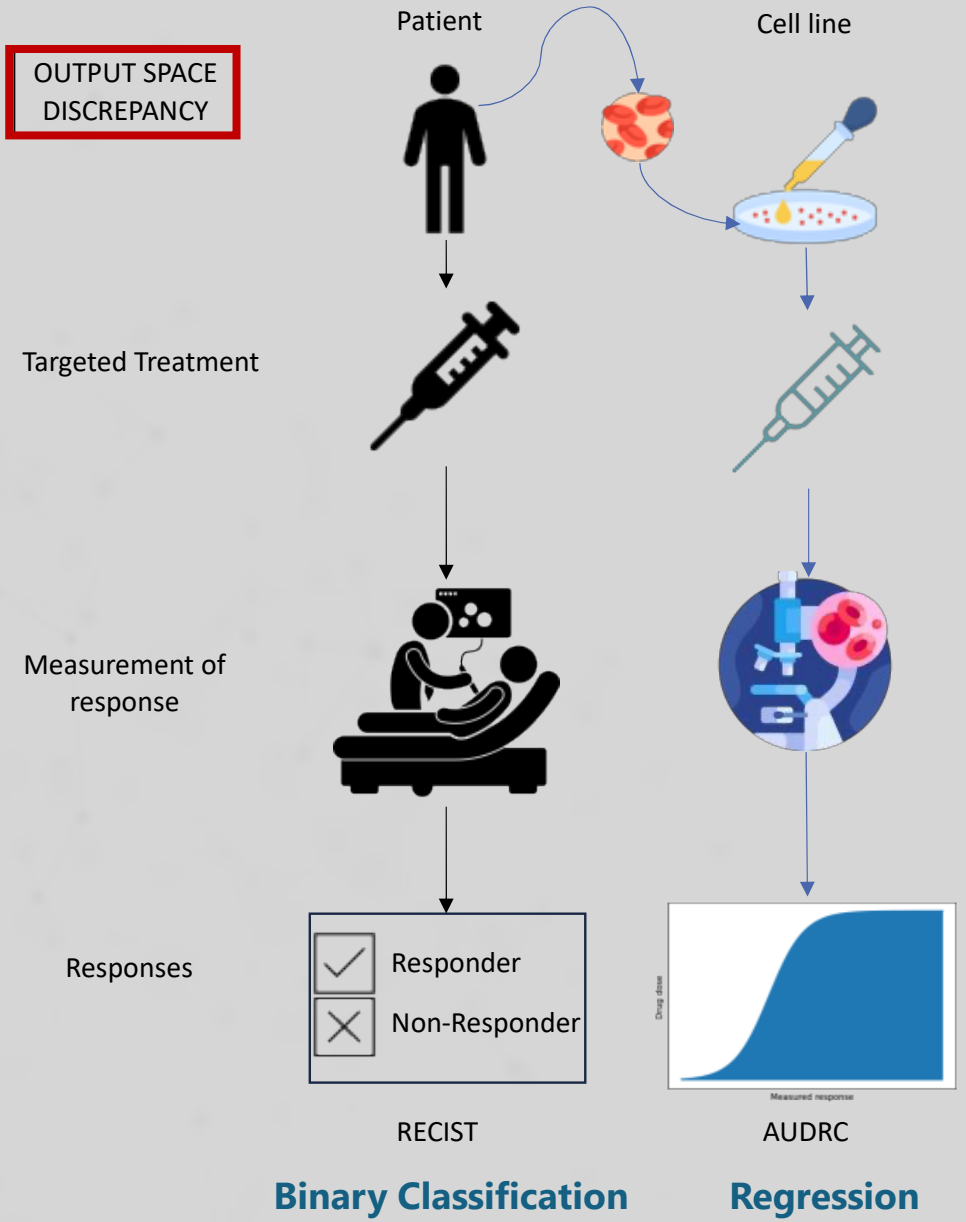
CELL LINES: A RELATED "DOMAIN"



- Extract cancer cells and clone them in lab (living cells, continue growing)
  - Ensures each cell has same genomic data (X)
- Administer multiple drugs on cell lines, measure response Y
- Can a Drug Response Prediction model on cell lines $Y \sim f_d(X)$ work for patients?

No: drug responses differ across patients and cell lines


INPUT SPACE DISCREPANCY
Sample acquisition
Sequencing
Mutation profile
Distribution of Mutations


OUTPUT SPACE DISCREPANCY
Patient
Cell line
Targeted Treatment
Measurement of response
Responses
Responder
Non-Responder
RECIST
AUDRC
Binary Classification
Regression

PROBLEM STATEMENT

- Given:

Domain	Genomic Profile	Drug Response	#samples
Cell Lines	X_c	$Y_c^d \in R$ (AADRC)	N_c labeled
Patients	X_t	$Y_t^d \in \{0,1\}$ (RECIST)	N_p labeled N_t unlabeled

$$N_p \ll N_c \ll N_t$$
$$P(X_c) \neq P(X_t)$$
$$f_c^d \neq f_t^d$$
where
$$Y_c^d \sim f_c^d(X_c), Y_t^d \sim f_t^d(X_t)$$

- Infer: Drug Response Prediction model

$$f_t^d: \; Y_t^d \sim f_t^d(X_t), \quad \forall \text{ drug } d \in \{d_1, d_2, \ldots d_n\}$$

Method	Clinical translation requirements				Transfer learning requirements		
PRECISE (2019)	?	?	?	?	Input discrepancy[1]	Output discrepancy[1]	?
AITL (2020)							
TCRP (2021)							
Velodrome (2021)							
TRANSACT (2021)							
TUGDA (2021)							
CODE-AE (2022)							
PANCDR (2024)							
Drug2tme (2024)							

[1] Sharifi-Noghabi, H., Peng, S., Zolotareva, O., Collins, C. C., and Ester, M. (2020). "AITL: adversarial inductive transfer learning with input and output space adaptation for pharmacogenomics," Bioinformatics (36:Supplement_1), pp. i380–i388.

DRP Requirements

From prior DRP literature

- Handle input discrepancy
- Handle output discrepancy
- Model patient mutation heterogeneity

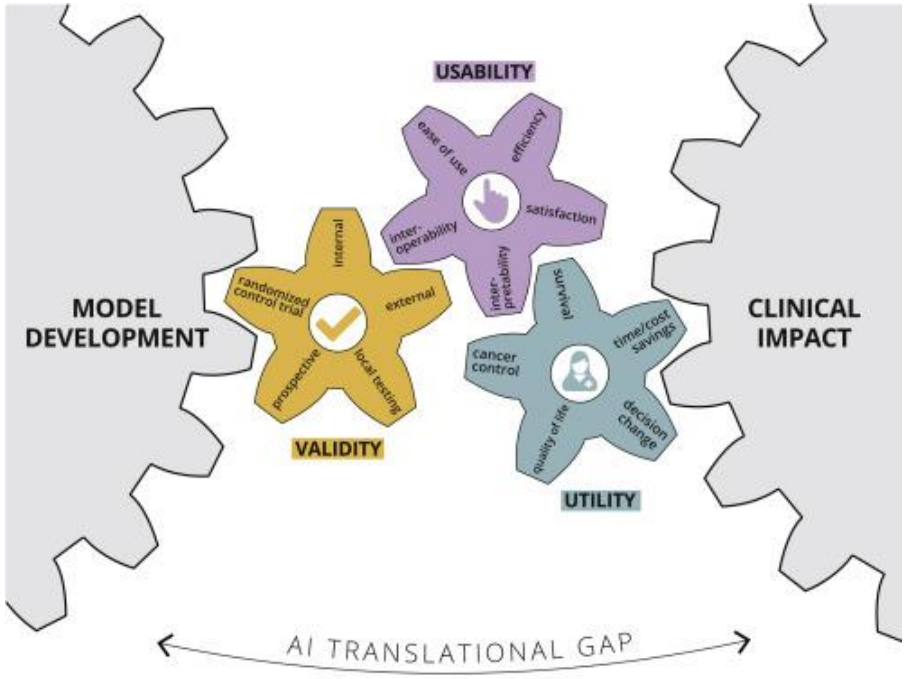
Transfer learning dimensions

Method	Clinical translation requirements				Transfer learning requirements		
PRECISE (2019)	?	?	?	?	Input discrepancy[1]	Output discrepancy[1]	Model patient heterogeneity [2]
AITL (2020)							
TCRP (2021)							
Velodrome (2021)							
TRANSACT (2021)							
TUGDA (2021)							
CODE-AE (2022)							
PANCDR (2024)							
Drug2tme (2024)							

[1] Sharifi-Noghabi, H., Peng, S., Zolotareva, O., Collins, C. C., and Ester, M. (2020). "AITL: adversarial inductive transfer learning with input and output space adaptation for pharmacogenomics," Bioinformatics (36:Supplement_1), pp. i380–i388.
[2] Zhai, J., & Liu, H. (2024). Cross-domain feature disentanglement for interpretable modeling of tumor microenvironment impact on drug response. *IEEE Journal of Biomedical and Health Informatics*.

Bridging the AI translational gap

- Clinical usability
  - Use of clinically available input data (clinical sequencing profiles)[1]
- Clinical utility
  - Use of clinically meaningful outcomes
  - Time/cost savings
- Clinical validity
  - Evaluation in real-world situations (clinical trials)



[2] Kann, B. H., Hosny, A., & Aerts, H. J. (2021). Artificial intelligence for clinical oncology. *Cancer Cell*, *39*(7), 916-927

[1] El Naqa, I., Karolak, A., Luo, Y., Folio, L., Tarhini, A. A., Rollison, D., & Parodi, K. (2023). Translation of AI into oncology clinical practice. *Oncogene*, *42*(42), 3089-3097.

Dimensions of comparison

Consideration for Clinical Translation	Requirement for clinical translation	Dimension
Clinical Usability	Use of clinically available input data	Training with mutation profiles (varying length)
Clinical Utility	Use of clinically meaningful outcomes	Utilise all patient response-related information like survival
	Time/cost savings	Enable repurposing of drugs already approved for clinical use
Clinical Validity	Evaluation in real-world situations	Clinical trials for validation

DRP Requirements

Consideration for Clinical Translation	Dimension
Clinical Usability	Training with mutation profiles (varying length)
Clinical Utility	Utilise all patient response-related information like survival
	Enable repurposing of drugs already approved for clinical use
Clinical Validity	Clinical trials for validation

For clinical translation

- Training with mutations available in clinical sequencing reports
- Model varying length mutations
- Utilise all available auxiliary patient response information (PFS)
- Predict on drugs unseen during training

After model development

DRP Requirements

For clinical translation

- Training with mutations available in clinical sequencing reports
- Predict on drugs unseen during training
- Model varying length mutations
- Utilise all available auxiliary patient response information (PFS)

From prior DRP literature

- Handle input discrepancy
- Handle output discrepancy
- Model patient mutation heterogeneity

Method	Clinical translation requirements				Transfer learning requirements		
PRECISE (2019)	Train with mutations from clinical sequencing profiles[3] (from clinicians)	Model varying length mutations[3]	Predict on drugs unseen during training[4]	Utilise available patient response data (PFS)[3]	Input discrepancy[1]	Output discrepancy[1]	Model patient heterogeneity [2]
AITL (2020)							
TCRP (2021)							
Velodrome (2021)							
TRANSACT (2021)							
TUGDA (2021)							
CODE-AE (2022)							
PANCDR (2024)							
Drug2tme (2024)							

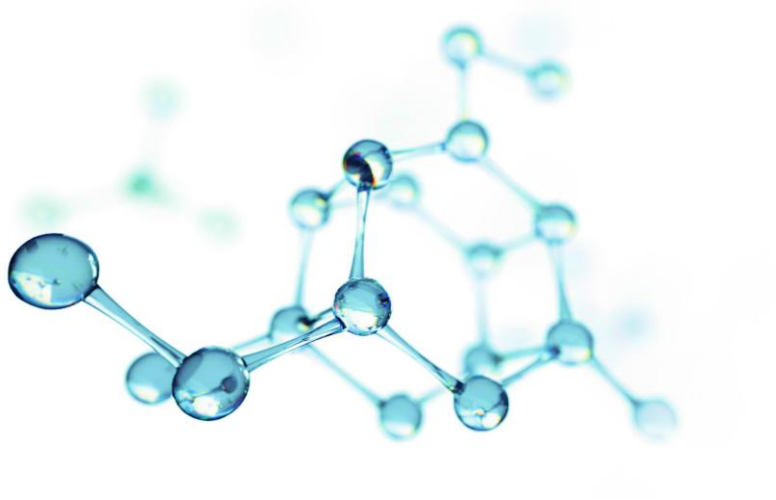
[1] Sharifi-Noghabi, H., Peng, S., Zolotareva, O., Collins, C. C., and Ester, M. (2020). "AITL: adversarial inductive transfer learning with input and output space adaptation for pharmacogenomics," Bioinformatics (36:Supplement_1), pp. i380–i388.
[2] Zhai, J., & Liu, H. (2024). Cross-domain feature disentanglement for interpretable modeling of tumor microenvironment impact on drug response. *IEEE Journal of Biomedical and Health Informatics*.
[3] Jayagopal, A., Xue, H., He, Z., Walsh, R. J., Hariprasannan, K. K., Tan, D. S. P., ... & Rajan, V. (2024, August). Personalised Drug Identifier for Cancer Treatment with Transformers using Auxiliary Information. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 5138-5149).
[4] Hua, Y., Dai, X., Xu, Y., Xing, G., Liu, H., Lu, T., ... & Zhang, Y. (2022). Drug repositioning: Progress and challenges in drug discovery for various diseases. *European Journal of Medicinal Chemistry*, *234*, 114239.

	Transfer Learning requirements			Clinical translation requirements			
Method	Input discrepancy	Output discrepancy	Model patient mutation heterogeneity in downstream DRP	Training with clinical mutations	Varying length inputs modelled	Use of auxiliary information (PFS)	Prediction on drugs unseen in training (for repurposing)
PRECISE (2019)	✓						
AITL (2020)	✓	✓					
TCRP (2021)	✓						
Velodrome (2021)	✓						
TRANSACT (2021)	✓						
TUGDA (2021)	✓						
CODE-AE (2022)	✓						
PANCDR (2024)	✓						✓
Drug2tme (2024)	✓	✓	✓				✓

Not clinically translatable!

	Transfer Learning requirements			Clinical translation requirements			
Method	Input discrepancy	Output discrepancy	Model patient mutation heterogeneity	Training with clinical mutations	Varying length inputs modelled	Use of auxiliary information (PFS)	Prediction on drugs unseen in training (for repurposing)
PRECISE (2019)	✓						
AITL (2020)	✓	✓					
TCRP (2021)	✓						
Velodrome (2021)	✓						
TRANSACT (2021)	✓						
TUGDA (2021)	✓						
CODE-AE (2022)	✓						
PANCDR (2024)	✓						✓
Drug2tme (2024)	✓	✓	✓				✓
DruID	✓	✓		✓			✓
PREDICT-AI	✓	✓		✓	✓	✓	✓
GANDALF	✓	✓	✓	✓	✓	✓	✓

DruID	PREDICT-AI	GANDALF
First approach towards predicting drug response from clinical grade mutation profiles	Utilises all available patient response related data	Data augmentation approach to handle data scarcity
Cell iScience	KDD'24	ICLR'25



Methods

DRP Requirements

For clinical translation

- Training with mutations available in clinical sequencing reports
- Predict on drugs unseen during training
- Model varying length mutations
- Utilise all available auxiliary patient response information (PFS)

From prior DRP literature

- Handle input discrepancy
- Handle output discrepancy
- Model patient mutation heterogeneity

DruID: Drug IDentifier

Jayagopal, A., Walsh, R.J., Hariprasannan, K.K., Mariappan, R., Mahapatra, D., Jaynes, P.W., Lim, D., Tan, D.S.P., Tan, T.Z., Pitt, J.J., Jeyasekharan, A.D. and Rajan, V, "A multi-task domain-adapted model to predict chemotherapy response from mutations in recurrently altered cancer genes." *iScience* 28.3 (2025).

DRP Requirements

For clinical translation

- Training with mutations available in clinical sequencing reports
- Predict on drugs unseen during training
- Model varying length mutations
- Utilise all available auxiliary patient response information (PFS)

From prior DRP literature

- Handle input discrepancy
- Handle output discrepancy
- Model patient mutation heterogeneity

Model Design

For clinical translation

- Training with mutations available in clinical sequencing reports
- Predict on drugs unseen during training
- Model varying length mutations
- Utilise all available auxiliary patient response information (PFS)

From prior DRP literature

- Handle input discrepancy
- Handle output discrepancy
- Model patient mutation heterogeneity

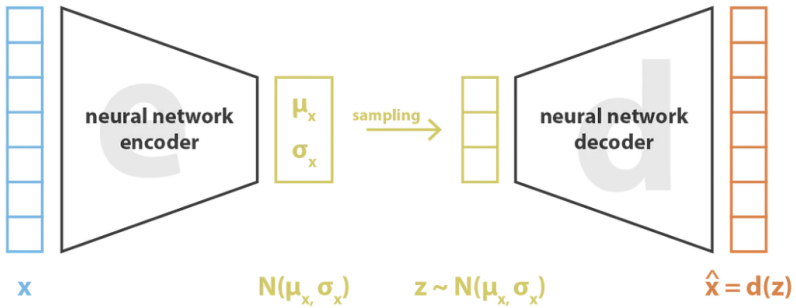
Requirement
Training with mutations available in clinical sequencing reports
Predict on drugs unseen during training
Handle input space discrepancy
Handle output space discrepancy

Model Design

Requirement	Considerations	Design Choice
Training with mutations available in clinical sequencing reports	Sparse, high dimensional nature of mutations	Use VAEs, with zero inflated distributions to model sparsity and high dimensional data

Variational Autoencoders

- A variational autoencoder tries to find a latent representation z that increases the probability of reconstructing the original input from it. In the encoder, variational probability Q(z|Y) is used to approximate the posterior P(z|Y). Neural networks are used as encoders and decoders to obtain the lower dimensional representation.
- Basic idea is to learn the probability distribution



$$\text{loss} = ||x - \hat{x}||^2 + KL[N(\mu_x, \sigma_x), N(0, I)] = ||x - d(z)||^2 + KL[N(\mu_x, \sigma_x), N(0, I)]$$

In variational autoencoders, the loss function is composed of a reconstruction term (that makes the encoding-decoding scheme efficient) and a regularisation term (that makes the latent space regular).

From https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

Zero inflated distributions

Zero Inflated Distribution

- Used for sparse datasets.
- Has a point mass at 0, for 0 values and NB for ordinal values/Normal for real.

$$ZINB(Y;\Pi,\Omega,\Theta) = \Pi\Delta_0(Y) + (1-\Pi)NB(Y;\Omega,\Theta)$$

$$\Pi = sigmoid(Y.W_\Pi);\ \Omega = \exp(Y.W_\Omega)\ ;\ \Theta = \exp(Y.W_\Theta)$$

$$NB(Y;\Omega,\Theta) = \frac{\Gamma(Y+\Theta)}{Y!\,\Gamma(\Theta)}\left(\frac{\Theta}{\Theta+\mu}\right)^{\Theta}\left(\frac{\mu}{\Theta+\mu}\right)^{Y}$$

Parameter estimation:

$$L_{Re} = NLL_{ZI}(X_e;\ \Pi_e,\Omega_e,\Theta_e) + \lambda\left|\left|\Pi_e\right|\right|^2$$

NLL ZINB

- For non-zero values
  $-\log[(1-\pi)NB(x;\mu,\theta)]$
  $= log\Gamma(x+1) + (\theta+x)\log(\theta+\mu) - xlog\mu\ -\ \theta log\theta + log\Gamma(\theta)$
  $- log\Gamma(x+\theta) - \log(1-\pi)$
- For zero case

$$-log\left[\pi + (1-\pi)\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\right]$$

- Add regularizing term with above value

$$\lambda\left|\left|\Pi_e\right|\right|^2$$

Model Design

Requirement	Considerations	Design Choice
Training with mutations available in clinical sequencing reports	Sparse, high dimensional nature of mutations	Use VAEs, with zero inflated distributions to model sparsity and high dimensional data
Predict on drugs unseen during training	Include drug information as a model input	Morgan fingerprint (binary) to encode drug information

Model Design

Requirement	Considerations	Design Choice
Training with mutations available in clinical sequencing reports	Sparse, high dimensional nature of mutations	Use VAEs, with zero inflated distributions to model sparsity and high dimensional data
Predict on drugs unseen during training	Include drug information as a model input	Morgan fingerprint (binary) to encode drug information
Handle input space discrepancy	Model shared characteristics common to cell lines and patients	Learn shared embedding by aligning domain representations, CORAL loss

CORAL loss

- Unsupervised domain adaptation loss
- Aligns source and target domains - minimizes covariance of input feature distributions
- Useful when distributions of domains are different

$$L_{CORAL} = ||cov_n(z_c) - cov(z_p)||$$

$$cov(z) = \frac{1}{n}\sum_{i=1}(z_i - \bar{z}_i)(z_i - \bar{z}_i)'$$

Model Design

Requirement	Considerations	Design Choice
Training with mutations available in clinical sequencing reports	Sparse, high dimensional nature of mutations	Use VAEs, with zero inflated distributions to model sparsity and high dimensional data
Predict on drugs unseen during training	Include drug information as a model input	Morgan fingerprint (binary) to encode drug information
Handle input space discrepancy	Model shared characteristics common to cell lines and patients	Learn shared embedding by aligning domain representations, CORAL loss
Handle output space discrepancy	Model differences in responses in both domains	Use multi-task learning to model the outputs – regression for cell lines and classification for patients.

Incorporate biological information available on each mutation

DruID: Pretraining

- Pretraining loss is a combination of reconstruction loss and alignment loss

$$L_{Re} = NLL_{ZI}(X_e;\ \Pi_e, \Omega_e, \Theta_e) + \ \lambda\left|\left|\Pi_e\right|\right|^2$$

$$L_{KLDe} = -0.5\ * \sum(1 + \log({\sigma_e}^2) - \mu_e^2 - \sigma_e^2)$$

$$L_{CORAL} = ||\underset{n}{cov}(z_c) - cov(z_p)||$$

$$cov(z) = \frac{1}{n}\sum_{i=1}(z_i - \bar{z}_i)(z_i - \bar{z}_i)'$$

$$L_{pretraining} = L_{Rc} + L_{KLDc} + L_{Rp} + L_{KLDp} + L_{CORAL}$$

DruID: MTL

- Training loss is a multi objective optimization of MSE and BCE logit loss

$$L_{BCE} = -[y_{RECIST} \log(sigmoid(\bar{y}_{RECIST})) + (1 - y_{RECIST})\log(1 - sigmoid(\bar{y}_{RECIST}))]$$

$$L_{MSE} = (y_{AUDRC} - \bar{y}_{AUDRC})^2$$

$$L_{MTL} = \max(\lambda_c L_{MSE}, \lambda_p L_{BCE})$$

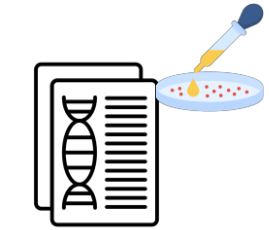
Model Training

Stage 1:
Pretraining VAEs

Stage 2:
Training the Multi-task learning (MTL) model

- Handles input discrepancy
- Training on mutation data

- Handles output space discrepancy
- Allows drug repurposing



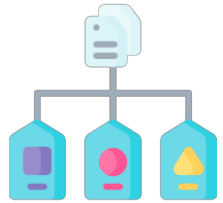
Cancer Cell Line Mutations

Clinical NGS report

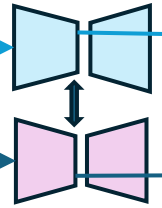
Extraction of mutations from report

DruID

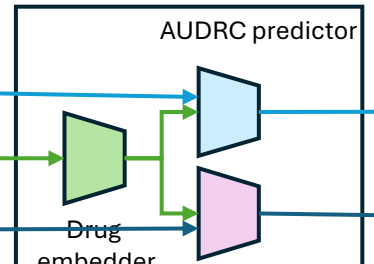
Drug Morgan fingerprint



(I) Variant Annotation



(II) Unsupervised Domain-Invariant Representation Learning



(III) Multi-Task Drug Response Prediction

Predicted AUDRC

0.8

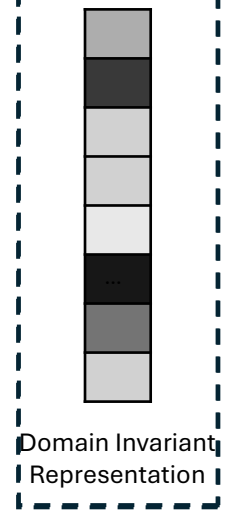
Responder

Predicted RECIST Category

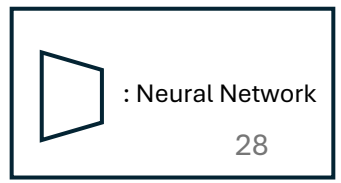
Mutations
ABL1 G259G
ABL1 N96S
XRCC2 T194T
AKT1 E17K

Mutation	GPD	ClinVar	Annovar based score
ABL1 G259G	NCU	VUS	0
ABL1 N96S	PIU	VUS	0.5
XRCC2 T194T	PIU	VUS	0
AKT1 E17K	PIU	VUS	1

Annotated Mutation Vector
(7776 dimensions
324 genes * 24 features)



Domain Invariant Representation



Extract patient identifier and set of point mutations from report

For each point mutation, obtain genomic coordinates, HGVSc, Consequence using **TransVar**

Derive **ClinVar annotations** from the genomic coordinates using filter-based Annovar annotations

Group Clinical Significance returned into 3 – pathogenic, benign or variants of unknown significance(VUS)

ClinVar Annotation

Derive dbNSFP categorical annotations from the genomic coordinates using filter-based **Annovar annotations**

Encode "Deleterious" as 1 and "Tolerated" as 0 to obtain a 17-dimensional binary vector per point mutation

AnnoVar Annotation

Obtain Ensembl ID for the gene associated with each point mutation

Use **GPD** to identify where each mutation lies – protein information unit (PIU), linker unit(LU) or non-coding unit(NCU)

GPD Annotation

Mutation	Clinical Significance	SIFT	LRT	...	FATHMM	Row-wise mean (d-score)	GPD Annotation
ABL1 G259G	VUS	0	0	...	0	0	NCU
XRCC2 T194T	Benign	0	0	...	0	0	NCU
AKT1 E17K	Pathogenic	1	1	...	1	15/17	PIU

Variant Annotation

Point mutation	GPD annotation	ClinVar annotation	d-score
ABL1 G259G	NCU	VUS	0
ABL1 N96S	PIU	VUS	0.5
ABL1 P622L	PIU	VUS	0
ABL1 R483W	PIU	VUS	1

Gene	ABL1
PIU mean	0.5
PIU sum	1.5
PIU max	1
PIU count	3
NCU mean	0
NCU sum	0
NCU max	0
NCU count	1
LU mean	0
LU sum	0
LU max	0
LU count	0

Gene	ABL1
Pathogenic mean	0
Pathogenic sum	0
Pathogenic max	0
Pathogenic count	0
Benign mean	0
Benign sum	0
Benign max	0
Benign count	0
VUS mean	0.375
VUS sum	2.5
VUS max	1
VUS count	4

Gene level aggregation example

Mutation 1

...

Mutation N

ClinVar Annotation

AnnoVar Annotation

GPD Annotation

...

23-dimensional vector

23-dimensional vector

24-dimensional vector/gene

Gene 1

Gene 2

Gene 324

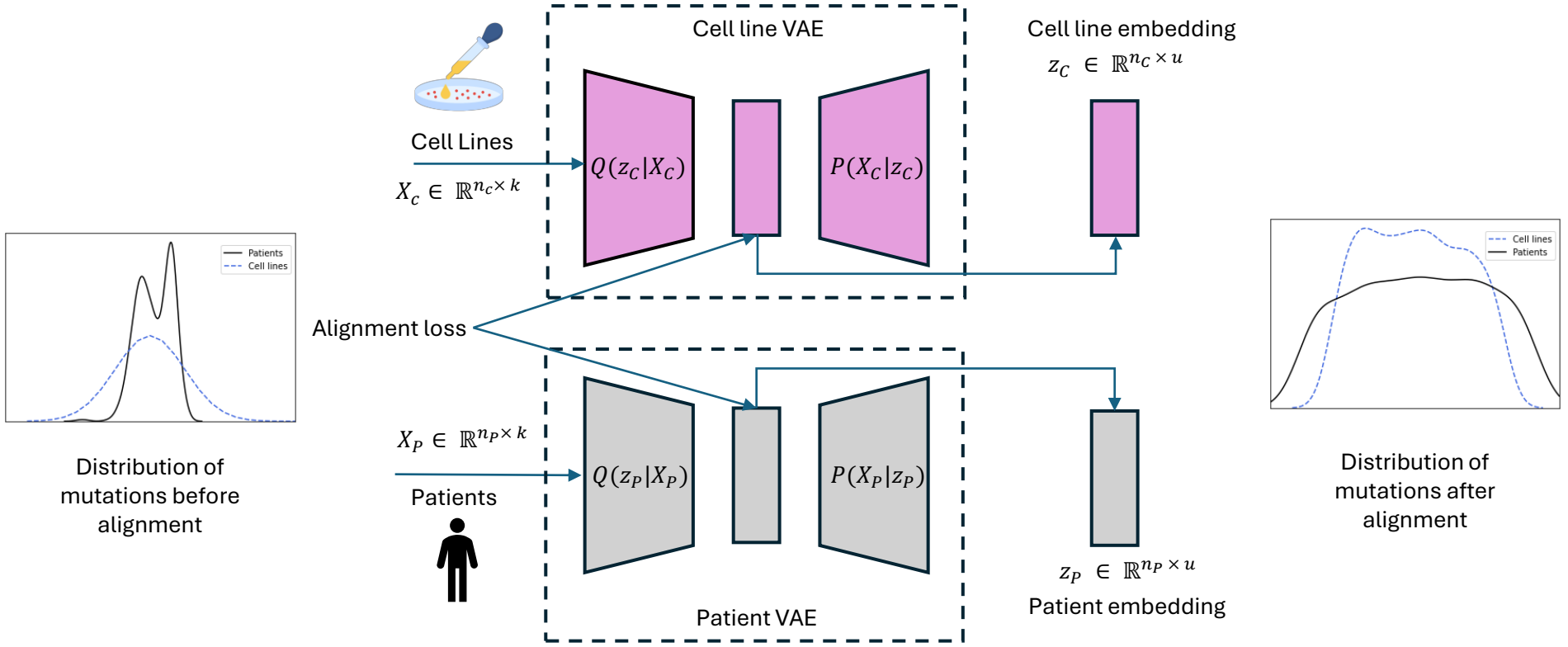
24-dimensional vector

24-dimensional vector

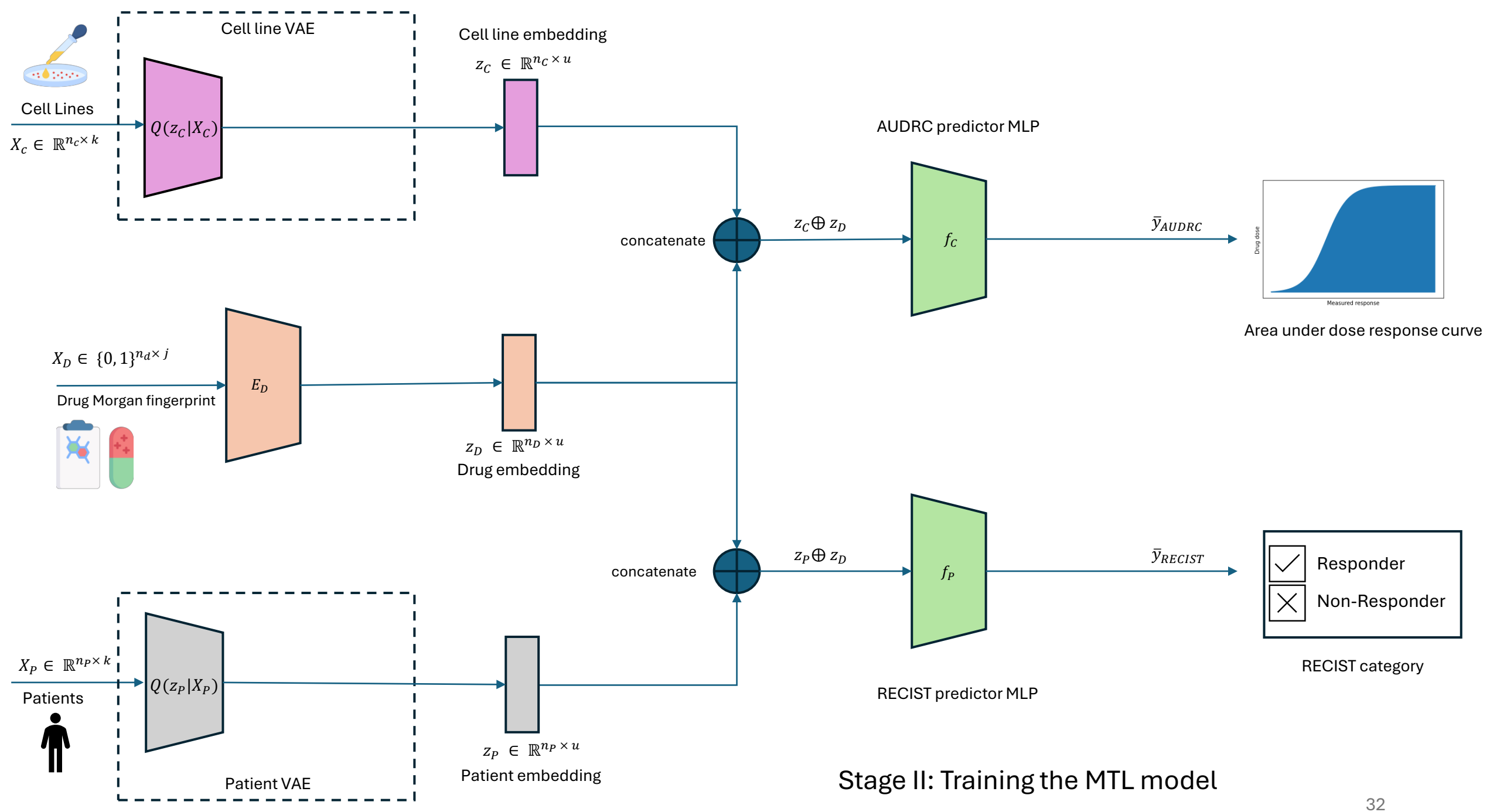
...

24-dimensional vector

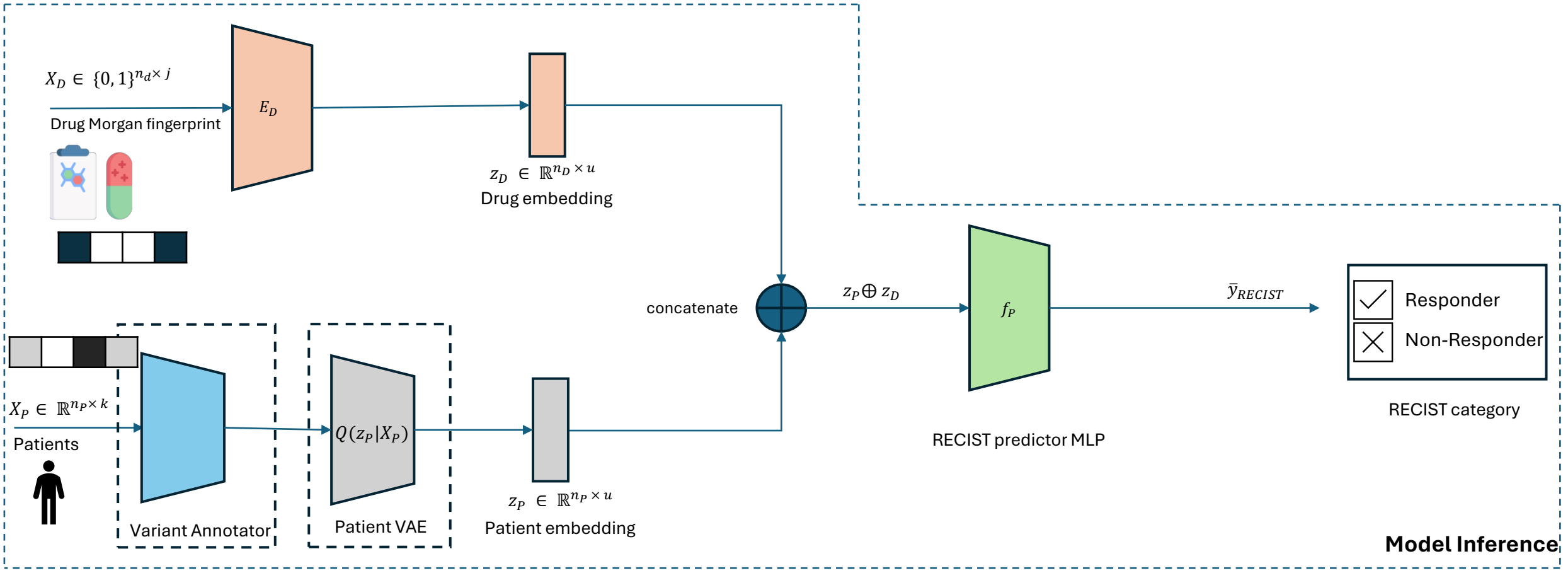
7776-dimensional vector/patient



Stage I: Pretraining VAEs



Stage II: Training the MTL model


$X_D \in \{0,1\}^{n_d \times j}$
Drug Morgan fingerprint
E_D
$z_D \in \mathbb{R}^{n_D \times u}$
Drug embedding
concatenate
$z_P \oplus z_D$
f_P
RECIST predictor MLP
$\bar{y}_{RECIST}$
Responder
Non-Responder
RECIST category
$X_P \in \mathbb{R}^{n_P \times k}$
Patients
Variant Annotator
$Q(z_P|X_P)$
Patient VAE
$z_P \in \mathbb{R}^{n_P \times u}$
Patient embedding
Model Inference

Experiment: Data

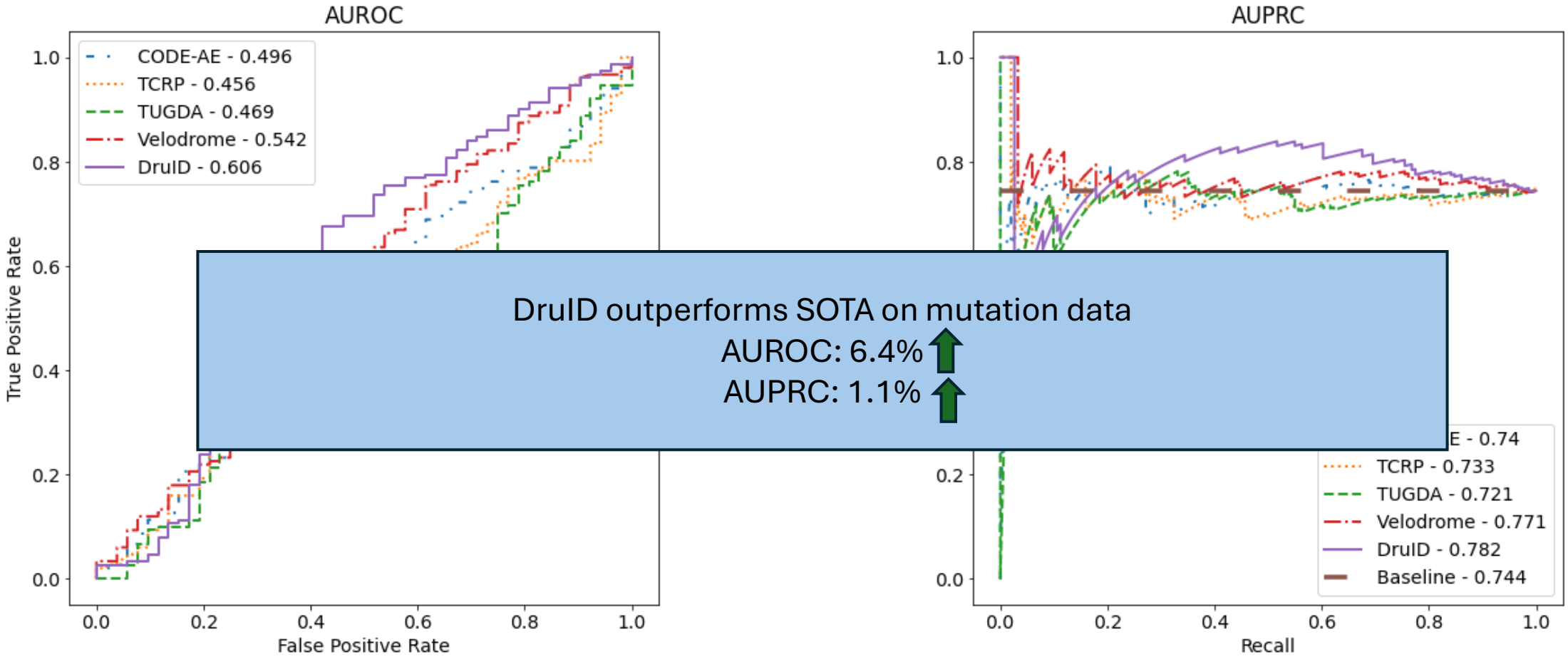
- Cancer cell lines
  - Mutation profiles: CCLE DepMap portal[1]
  - AUDRC labels: GDSC portal[2]
- Patients
  - Pan cancer TCGA patient mutation profiles and RECIST labels: TCGA GDC portal[3]
- Only 324 genes from FoundationOne CDx report retained for use

[1] https://depmap.org/portal/data_page/?tab=allData
[2] https://www.cancerrxgene.org/
[3] https://portal.gdc.cancer.gov/

Experimental Settings

- Split cell line and patient data into train and test splits (80:20)
  - Generate 3 different train-test splits
  - TCGA test split – 90
- Train model on train splits of cell lines and patients
  - Cell line – TCGA dataset: 689 cell line-drug pairs, 444 patient-drug pairs
- Evaluation using AUROC and AUPRC on test splits of patients (RECIST)
- Comparison of baselines against most recent work – CODE-AE, TUGDA, TCRP, Velodrome

Results: Comparison against SOTA



DruID outperforms SOTA on mutation data

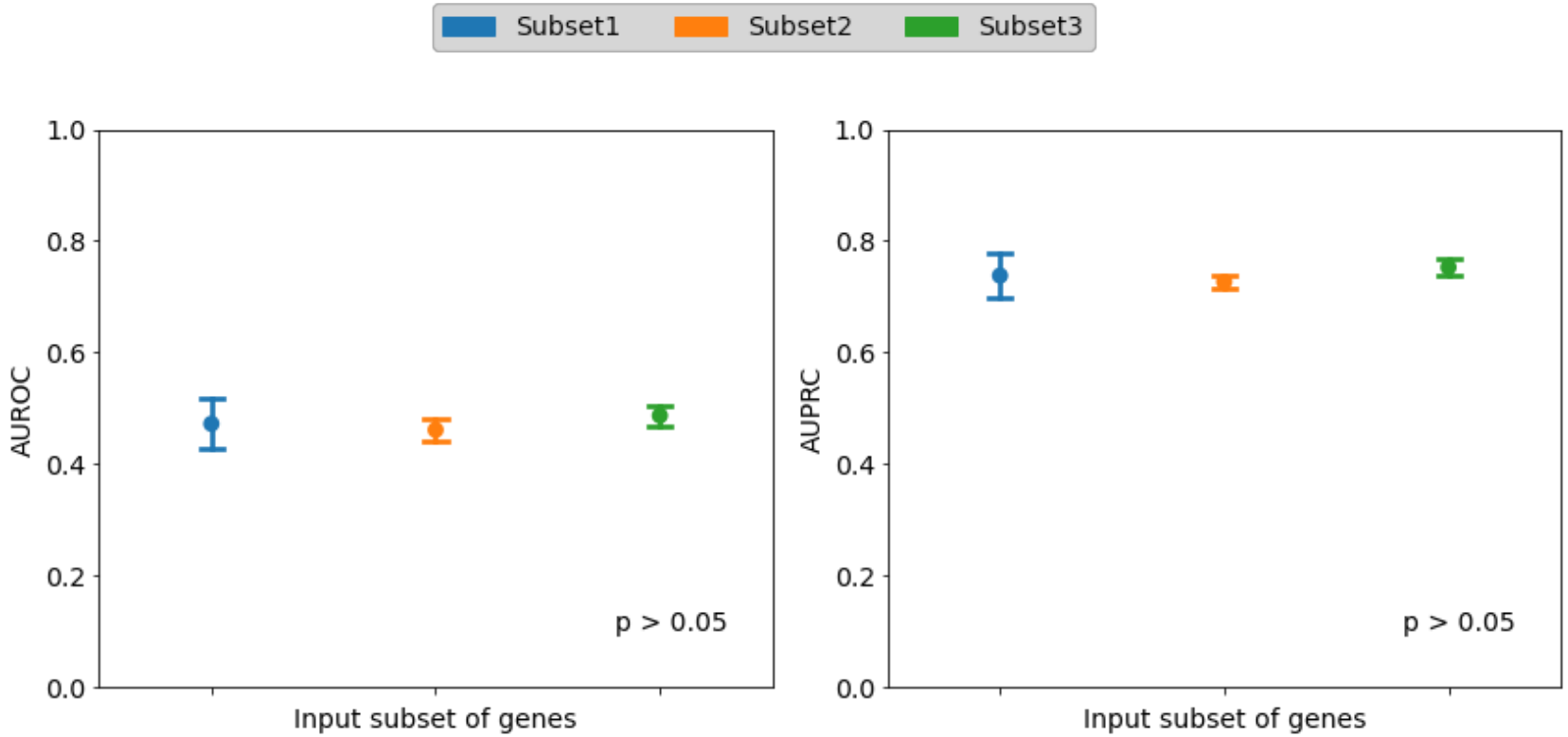
AUROC: 6.4% ↑

AUPRC: 1.1% ↑

Additional experiments

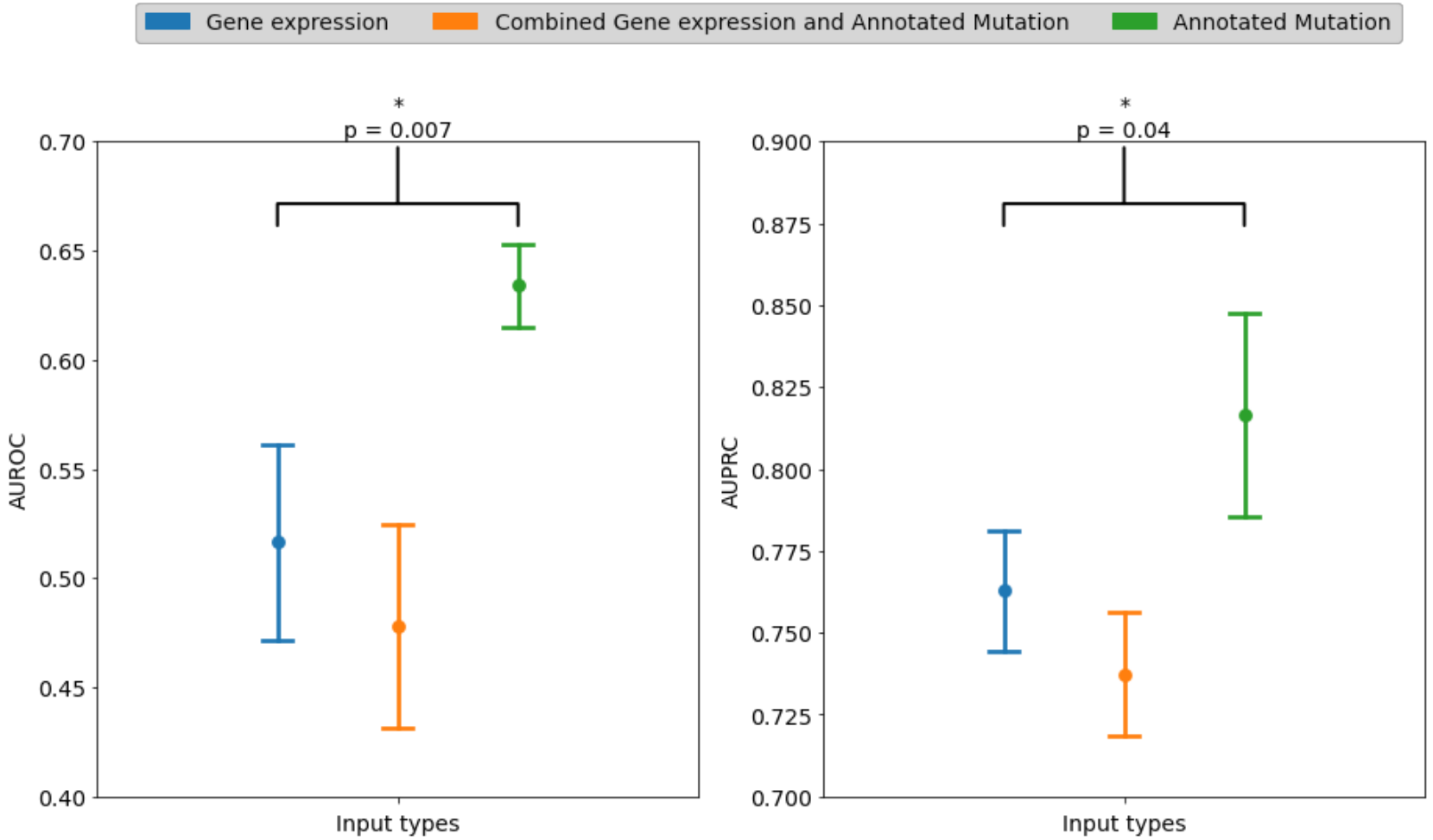
- Performance comparison of mutation against gene expression (& other data types) data from genes in clinical sequencing panels
- Validation of DruID on real-world data from NUH, Singapore
- Comparison of DruID against SOTA on gene expression
- Ablation study

Does the use of clinical NGS work as well as WES?

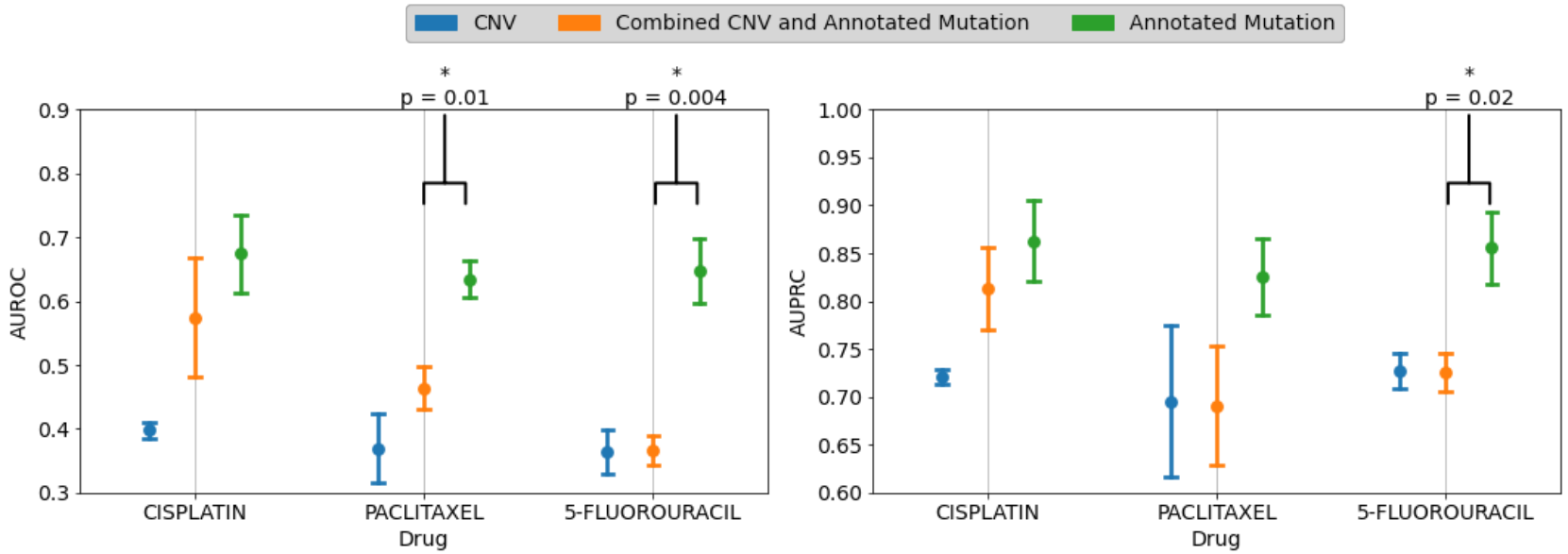


Subset1: 324 genes included in FoundationOne CDx.
Subset2: 285 genes common across FoundationOne CDx, TruSight Oncology 500 and Tempus xF+ cNGS panels.
Subset3: 19,536 genes, nearly all those available from WES.

DruID: Comparison of input types



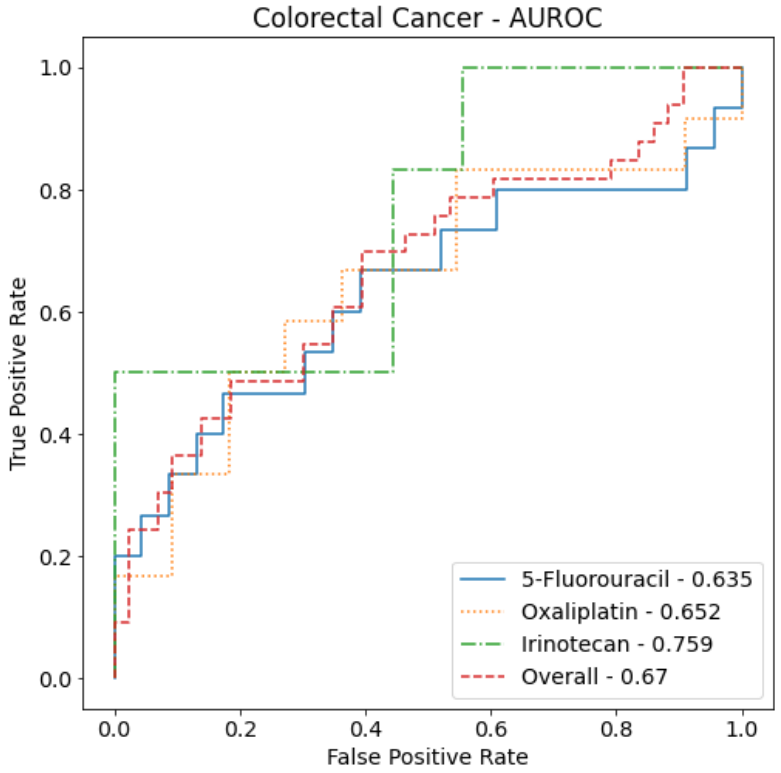
DruID: Comparison of input types

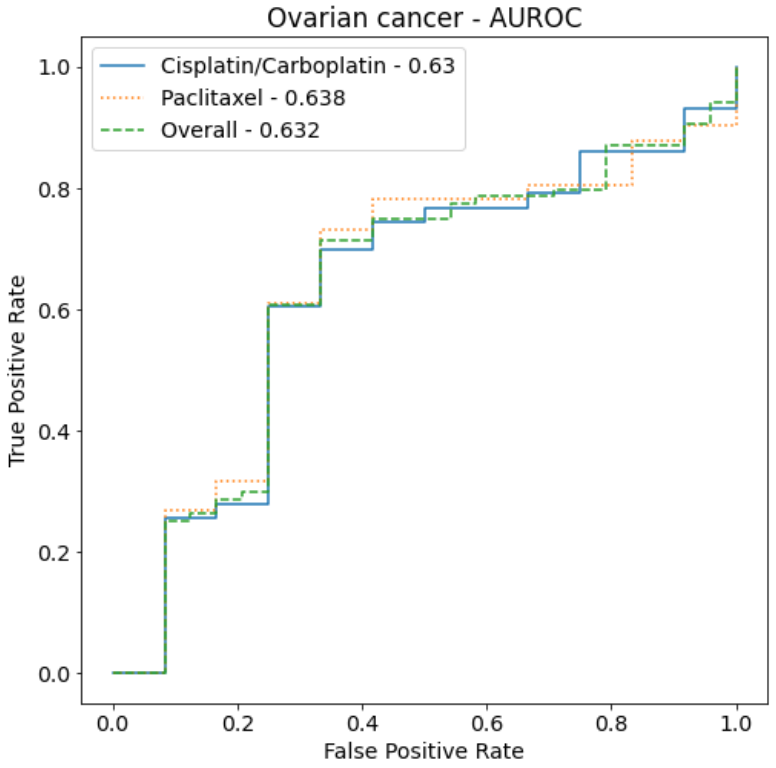


Additional experiments

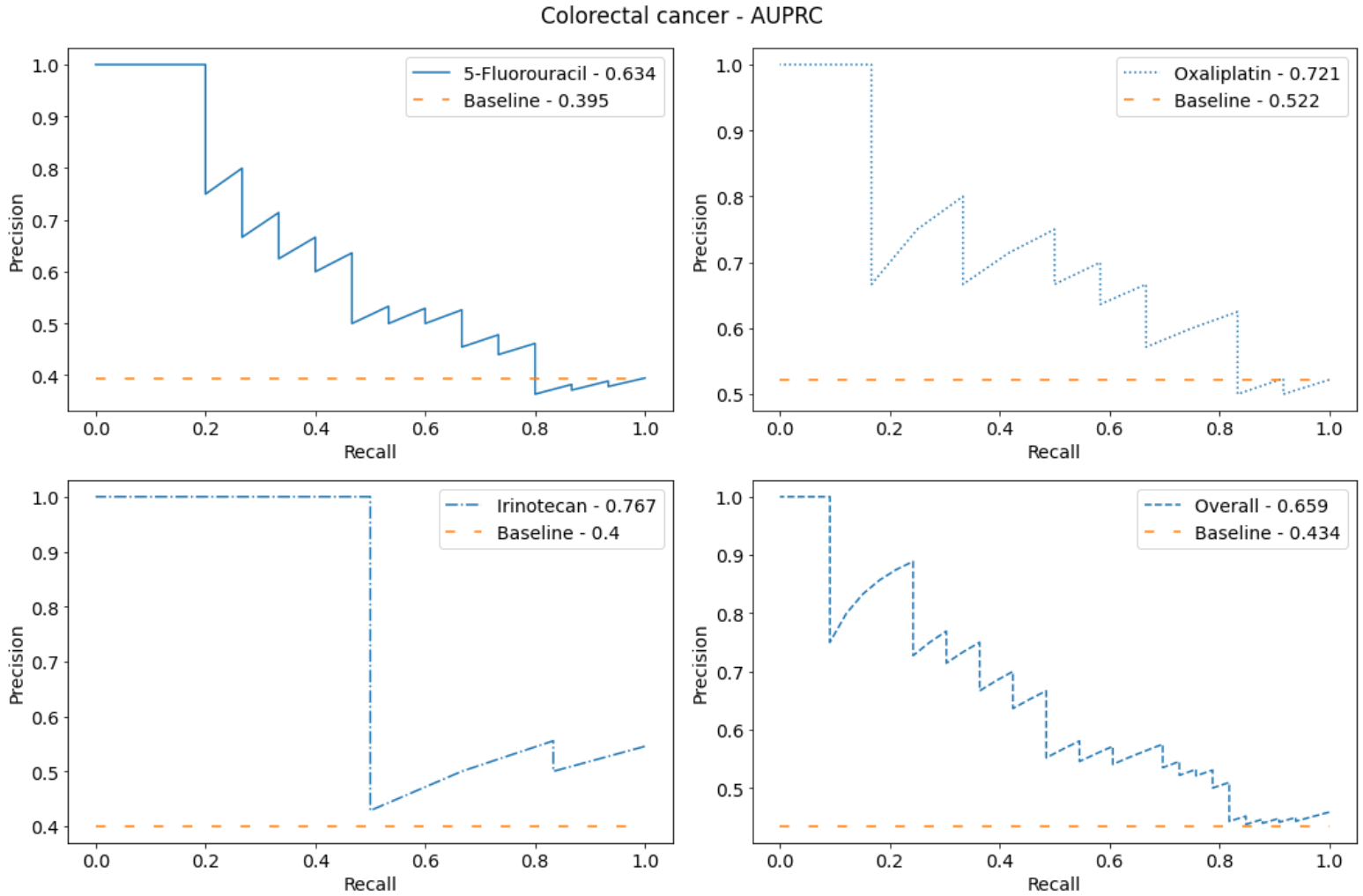
- Performance comparison of mutation against gene expression data from genes in clinical sequencing panels
- Validation of DruID on real-world data from NUH, Singapore
- Comparison of DruID against SOTA on gene expression
- Ablation study

DruID: On real world datasets

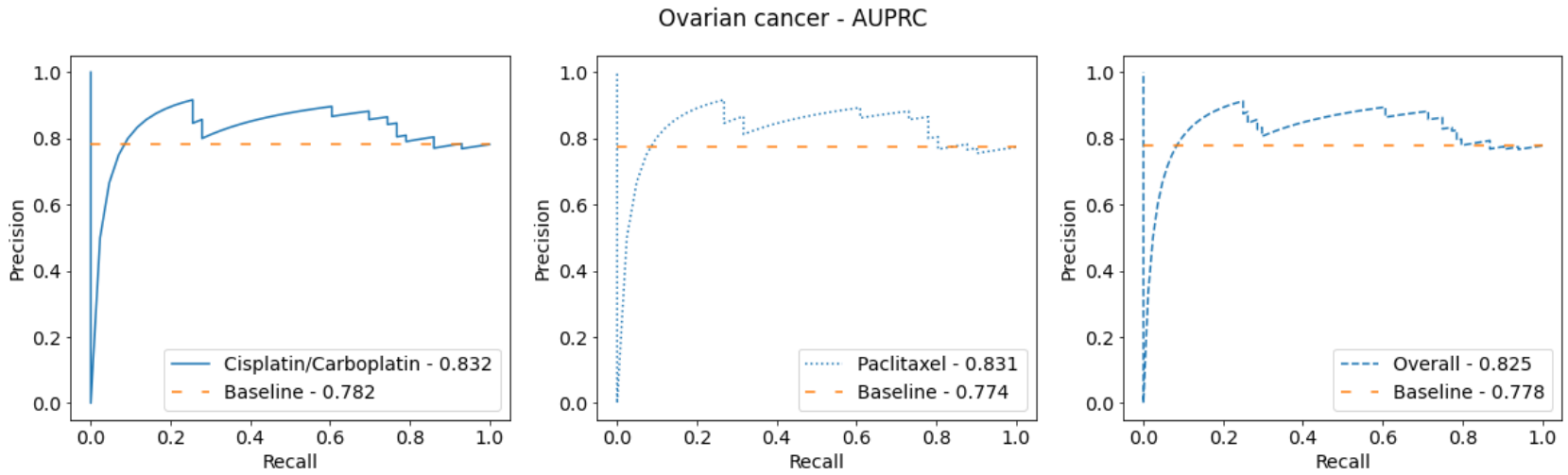




DruID: On real world datasets



DruID: On real world datasets



Additional experiments

- Performance comparison of mutation against gene expression data from genes in clinical sequencing panels
- Validation of DruID on real-world data from NUH, Singapore
- Comparison of DruID against SOTA on gene expression
- Ablation study

DruID: Comparison against SOTA on gene expression

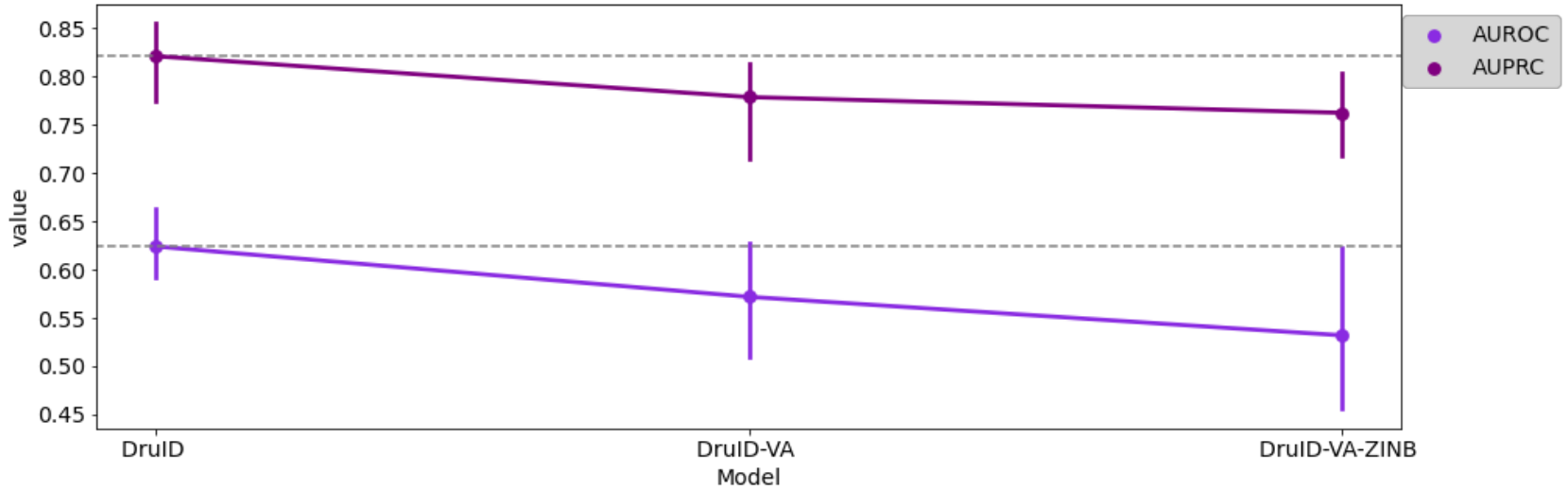
AUROC scores					
Drug	TCRP	TUGDA	Velodrome	CODE-AE	DruID
SORAFENIB	0.5482 +- 0.3445	0.4786 +- 0.3203	0.5482 +- 0.3066	0.3704 +- 0.3208	**0.6889 +- 0.3006**
CISPLATIN	0.6222 +- 0.4018	0.512 +- 0.2038	0.2984 +- 0.1507	0.4127 +- 0.1915	**0.8222 +- 0.1678**
GEMCITABINE	0.5347 +- 0.1185	0.432 +- 0.0944	0.4216 +- 0.1713	0.4474 +- 0.2252	**0.6984 +- 0.2374**
TEMOZOLOMIDE	0.6984 +- 0.0999	0.4716 +- 0.1375	0.7401 +- 0.0653	**0.9127 +- 0.0422**	0.6548 +- 0.2407
5-FLUOROURACIL	0.7222 +- 0.347	0.3684 +- 0.1255	0.3889 +- 0.2546	0.6111 +- 0.2546	**0.7778 +- 0.0962**

AUPRC scores					
Drug	TCRP	TUGDA	Velodrome	CODE-AE	DruID
SORAFENIB	0.613 +- 0.3706	0.086 +- 0.0258	0.6333 +- 0.3756	0.537 +- 0.3207	**0.6944 +- 0.3938**
CISPLATIN	0.6909 +- 0.2677	0.0984 +- 0.0144	0.3555 +- 0.1869	0.4219 +- 0.2401	**0.9056 +- 0.1055**
GEMCITABINE	0.7391 +- 0.2081	0.2104 +- 0.1009	0.65 +- 0.2173	0.6551 +- 0.1409	**0.8119 +- 0.1544**
TEMOZOLOMIDE	0.7714 +- 0.1482	0.204 +- 0.1348	0.7579 +- 0.1835	**0.9222 +- 0.0592**	0.7818 +- 0.0589
5-FLUOROURACIL	0.7611 +- 0.282	0.0609 +- 0.012	0.4556 +- 0.1134	0.6111 +- 0.2097	**0.8055 +- 0.0481**

Additional experiments

- Performance comparison of mutation against gene expression data from genes in clinical sequencing panels
- Validation of DruID on real-world data from NUH, Singapore
- Comparison of DruID against SOTA on gene expression
- Ablation study

DruID: Ablation



- Addition of variant annotations improves performance
- Using zero inflated distributions help, but is not significant