

GANDALF: Generative Attention based Data Augmentation and predictive modeLing Framework for personalized cancer treatment

Aishwarya Jayagopal¹, Yanrong Zhang¹, Robert J. Walsh², Tuan Zea Tan³, Anand D. Jeyasekharan³, Vaibhav Rajan¹

¹ National University of Singapore, ² National University Cancer Institute,

³ Cancer Science Institute of Singapore

Research funded by National Research Foundation, Singapore and Cancer Science Institute of Singapore.





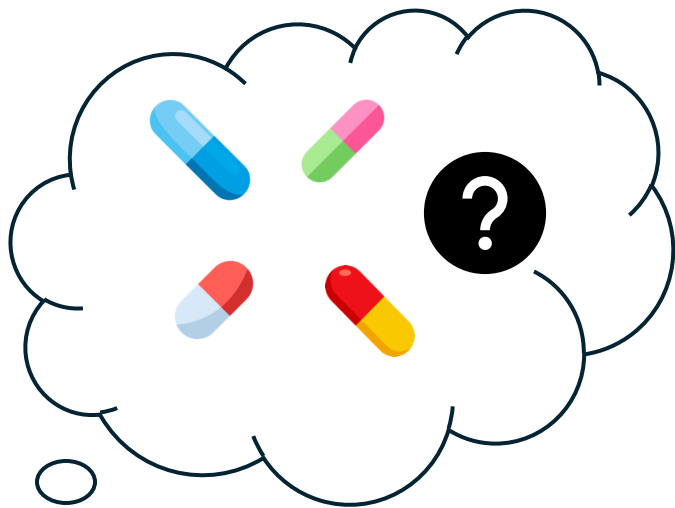
Cancer patient



Genomic sequencing



Mutation 1
Mutation 2
...
Mutation N



Challenge in precision oncology



Clinical trials

Mutation 1



Mutation 2

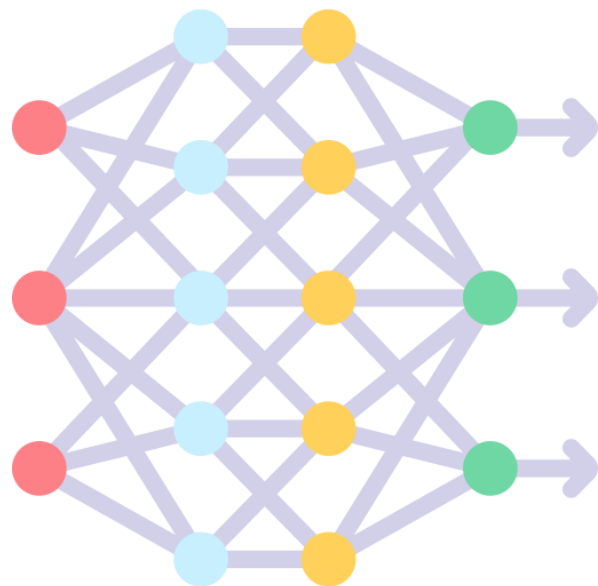


Mutation 1
Mutation 2
...
Mutation N



Clinical trials intractable!

Mutation 1
Mutation 2
...
Mutation N



Deep learning methods

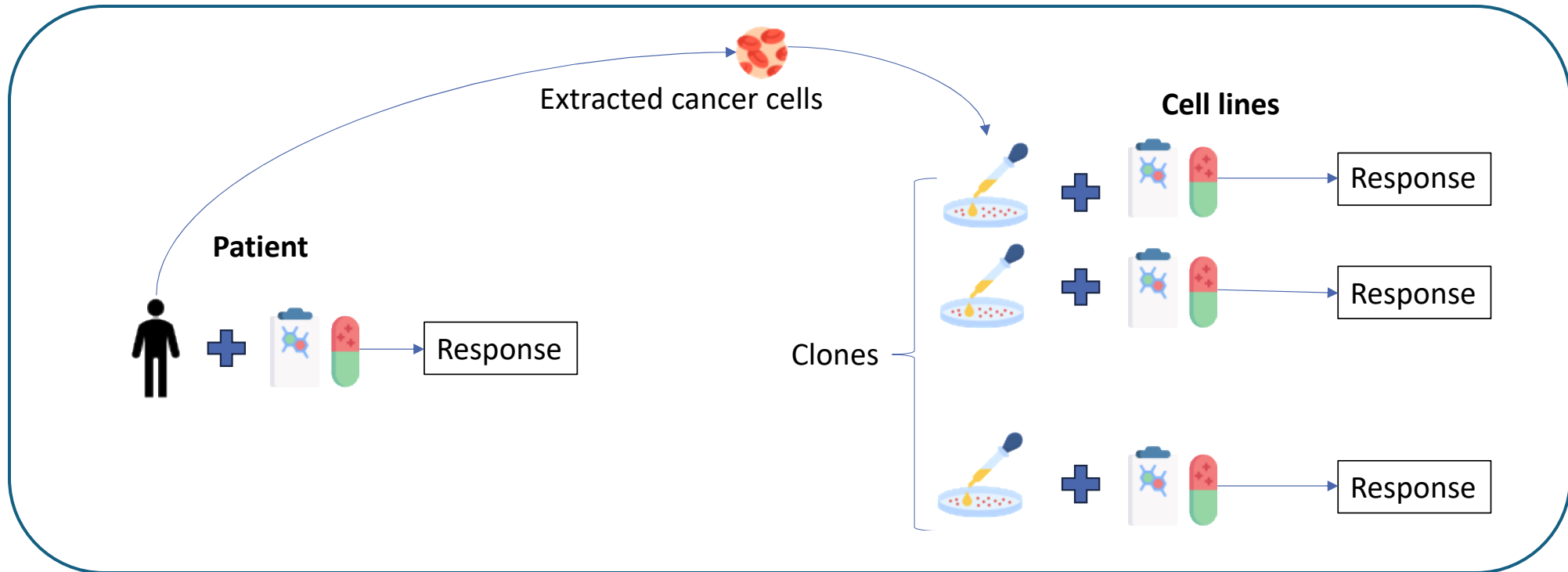
Recommendations



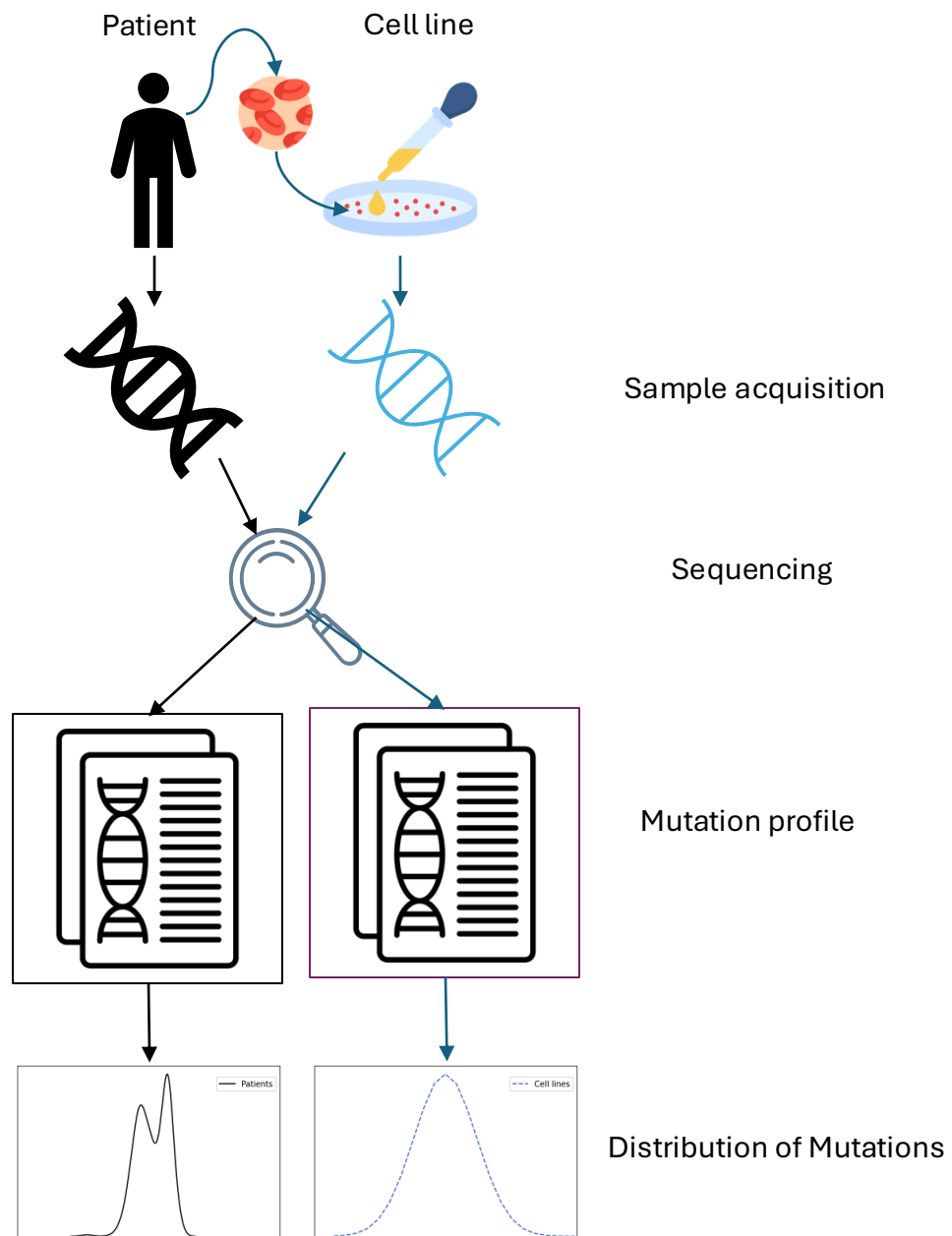
Limited labelled
patient data

Clinical standard
of care
treatment uses
only subset of
drugs

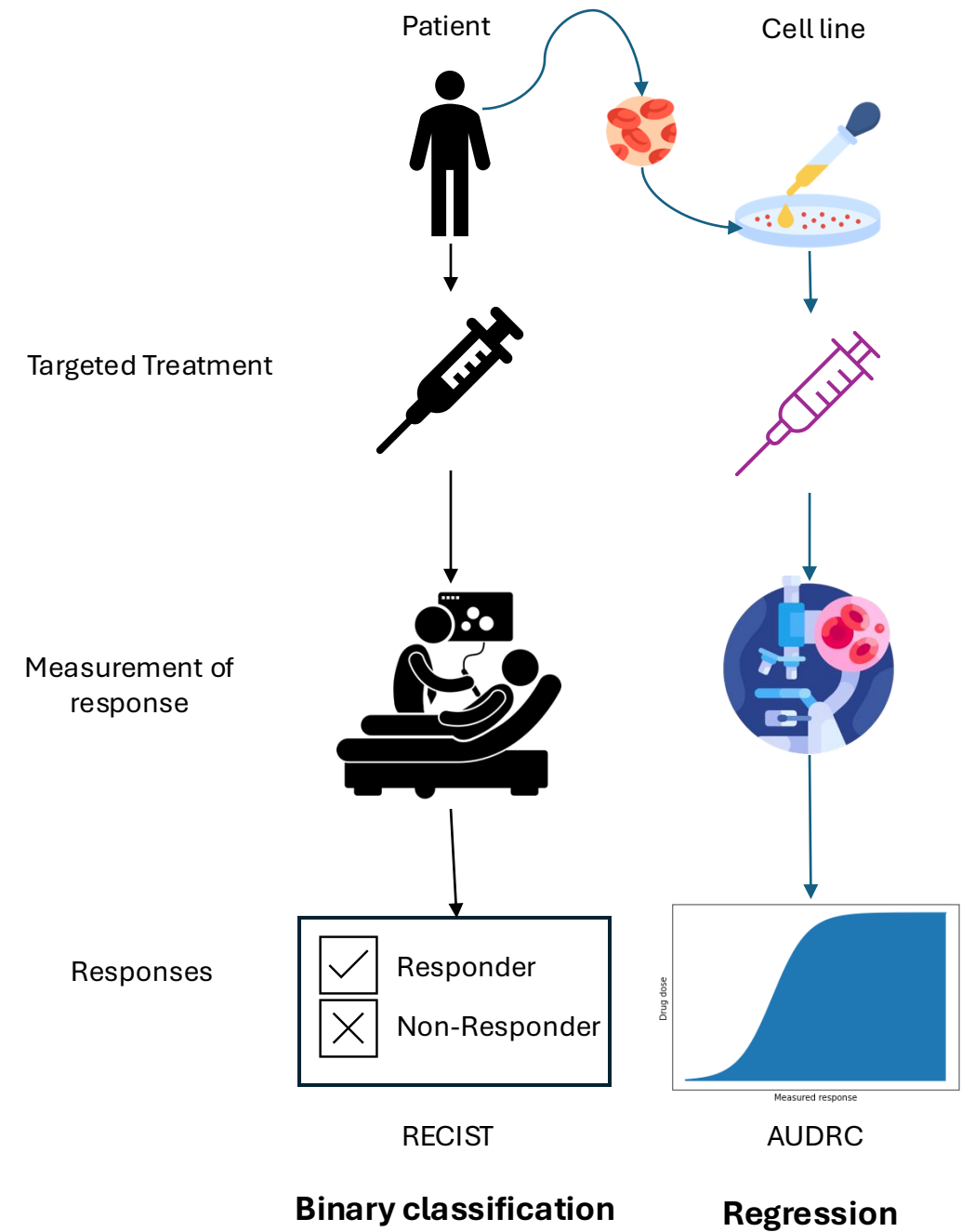
Cell lines: A Related “Domain”

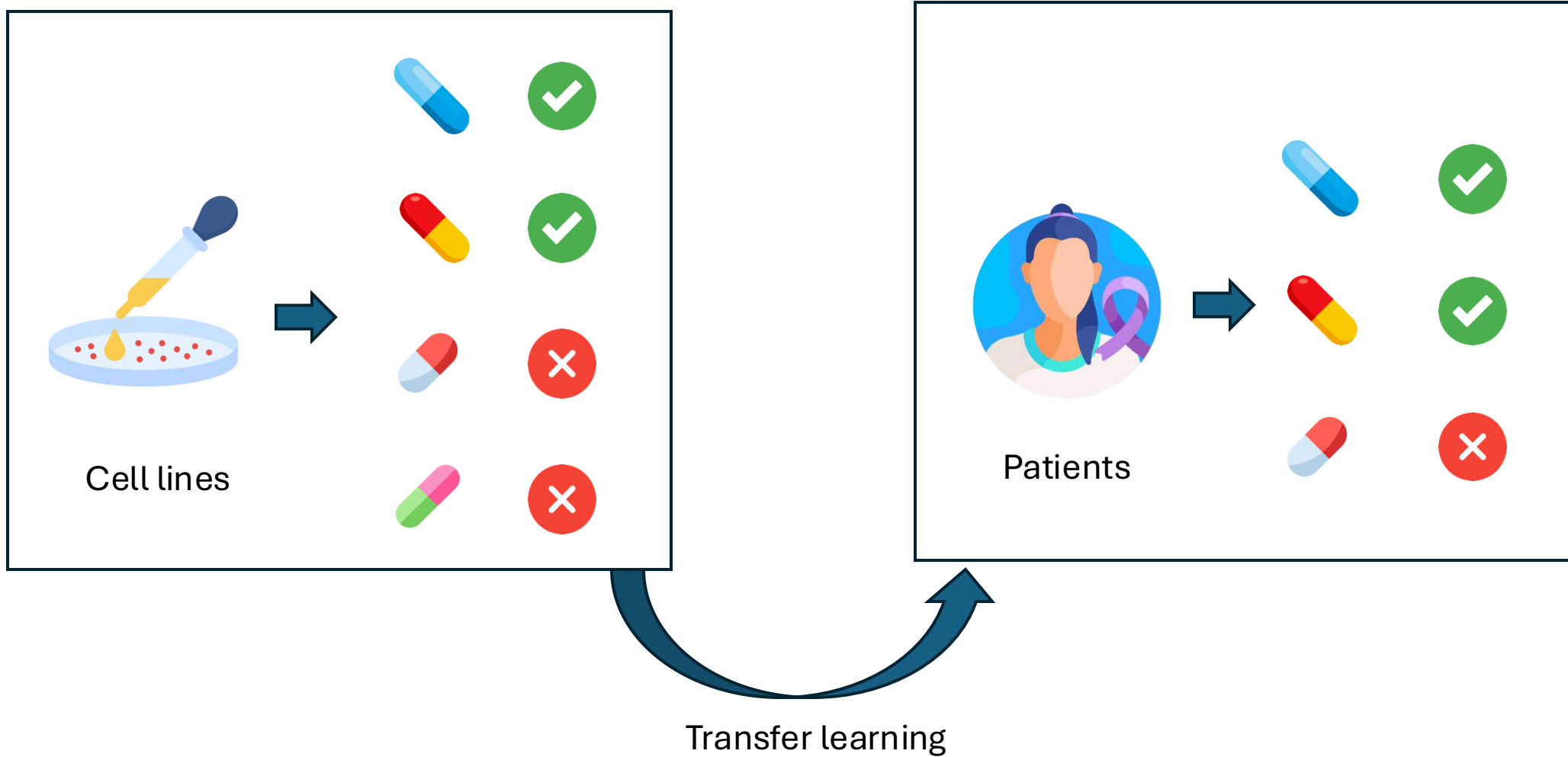


- Extract cancer cells and clone them in lab (living cells, continue growing)
 - Ensures each cell has same genomic data (X)
- Administer multiple drugs on cell lines, measure response (Y)

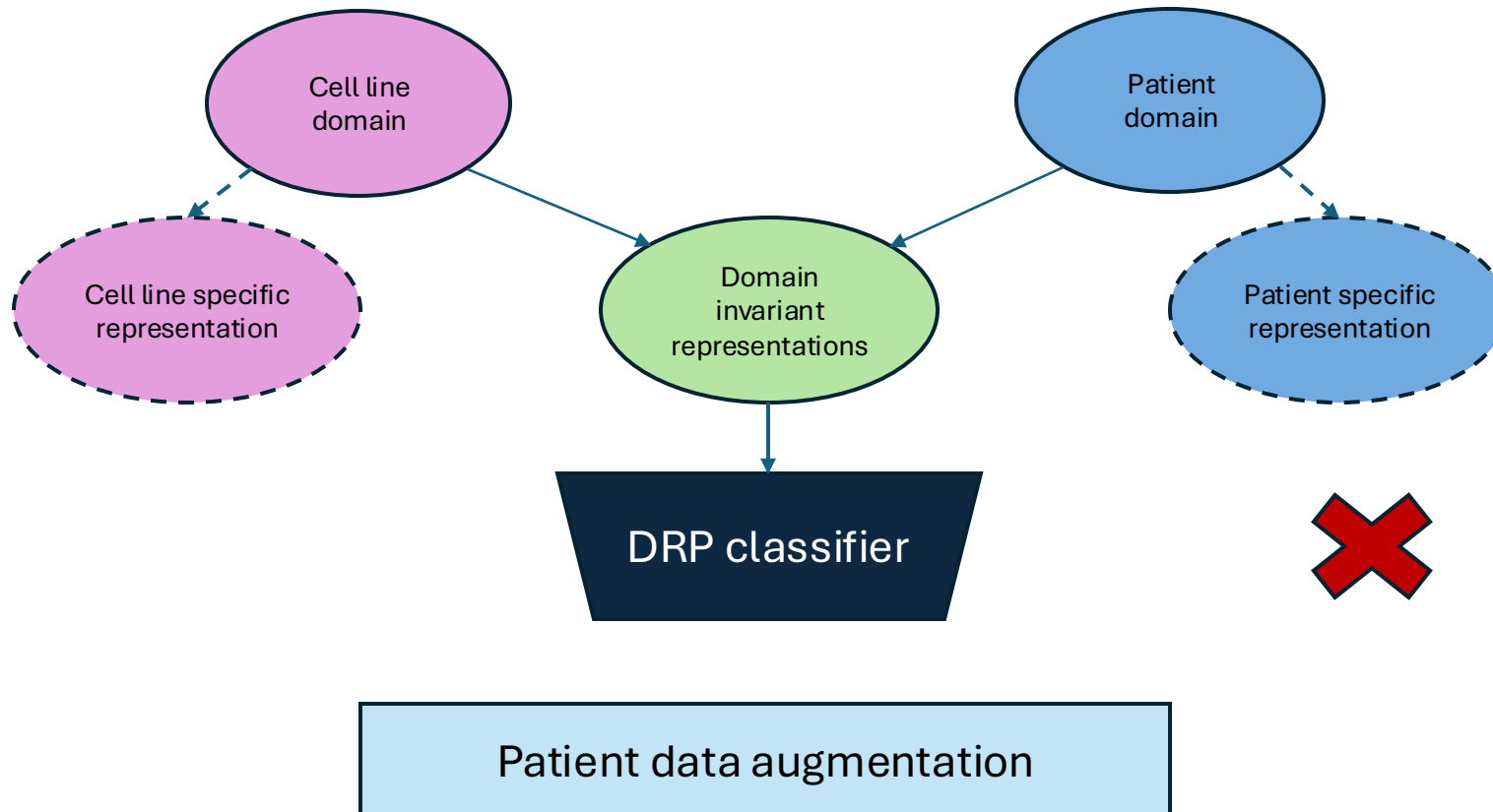


Icons from Flaticon and Freepik



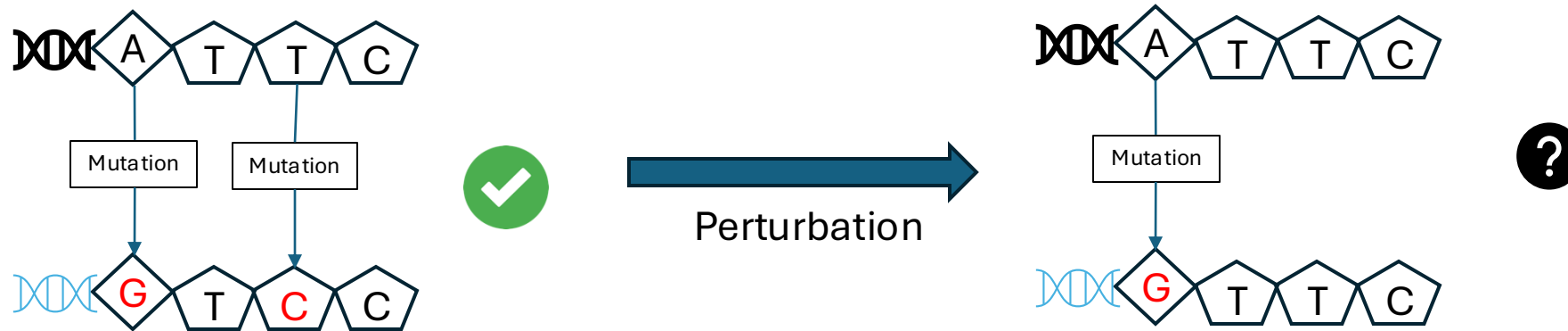


Modeling patient specific representations



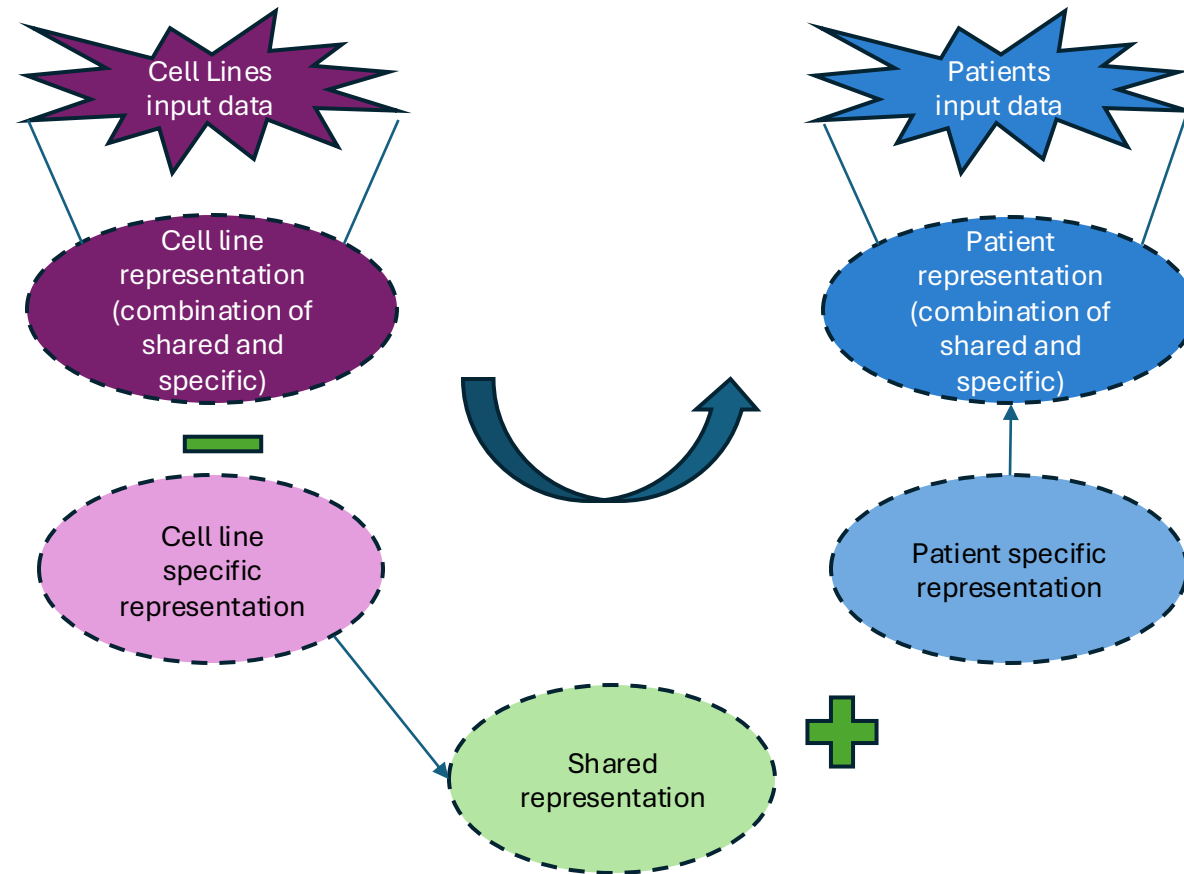
Genomic Data Augmentation

- No ‘label-invariance’ guarantee
- “Nearby” data points can respond differently to drugs
 - Perturbation \Rightarrow addition/removal of mutation

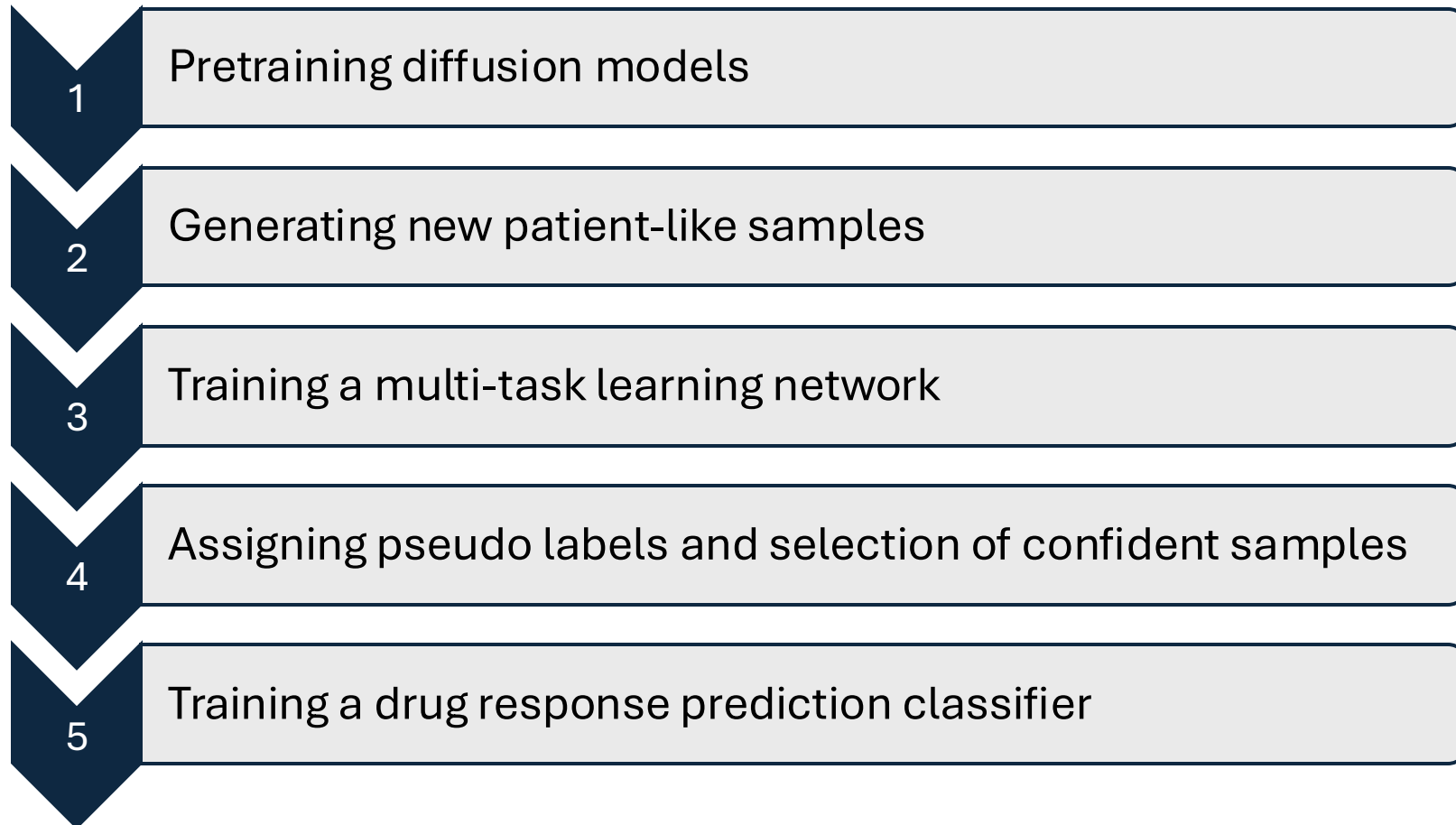


Use labelled cell lines to generate patient-like samples and assign labels to them.

Intuition



Model Training



Results: Comparison against SOTA

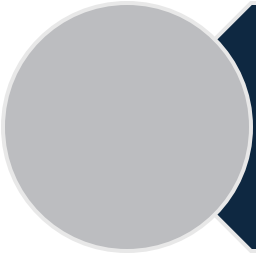
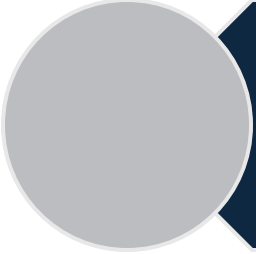

AUROC (Mean / Standard deviation)					
Method	Cis	Flu	Gem	Pac	Tem
GANDALF	<u>0.6343</u> ± 0.0306	0.7309 0.0664	0.6188 0.0674	0.7728 0.1253	0.6451 0.0776
DruID	0.6764 ± 0.1447	0.6071 ± 0.1988	0.5092 ± 0.1005	0.5119 ± 0.2324	0.6194 ± 0.0420
PANCDR	0.6278 ± 0.0308	0.4762 ± 0.1798	0.4429 ± 0.2268	0.4236 ± 0.4168	0.6436 ± 0.2310
PREDICT-AI	0.5072 ± 0.0331	0.3869 ± 0.0372	0.5046 ± 0.1181	0.6815 ± 0.1786	0.5350 ± 0.0606
drug2tme	0.5243 ± 0.1301	0.7167 ± 0.1957	0.4568 ± 0.0857	0.3194 ± 0.3127	0.5951 ± 0.2541
CODE-AE	0.6322 ± 0.1872	0.5381 ± 0.1606	0.5085 ± 0.0503	0.3611 ± 0.3155	0.4332 ± 0.3123

AUPRC (Mean / Standard deviation)					
Method	Cis	Flu	Gem	Pac	Tem
GANDALF	<u>0.9093</u> ± 0.0355	0.8483 0.0933	0.5874 ± 0.175	0.9558 ± 0.024	0.2535 ± 0.1108
DruID	0.9176 ± 0.0671	0.7588 ± 0.1484	0.4515 ± 0.1297	0.8897 ± 0.0223	0.3104 ± 0.1039
PANCDR	0.9018 ± 0.0324	0.6951 ± 0.1530	0.4562 ± 0.2270	0.8561 ± 0.1019	<u>0.3049</u> <u>0.2653</u>
PREDICT-AI	0.8622 ± 0.0189	0.5885 ± 0.0581	0.3873 ± 0.0489	0.8687 ± 0.1090	0.1373 ± 0.0050
drug2tme	0.8754 ± 0.0523	0.8092 ± 0.1722	0.4826 ± 0.0947	0.7824 ± 0.1023	0.3058 0.1327
CODE-AE	0.9059 ± 0.0521	0.6665 ± 0.1435	0.4735 ± 0.0701	0.8208 ± 0.0574	0.1756 ± 0.0929

Max Improvement in AUROC: 10.96%

Max Improvement in AUPRC: 3.91%

Advantages of GANDALF

-  Model is generative
-  Allows use of prior information from cell lines
-  Avoiding overfitting when using augmented patient data

Thank you!

Paper



Code repo

