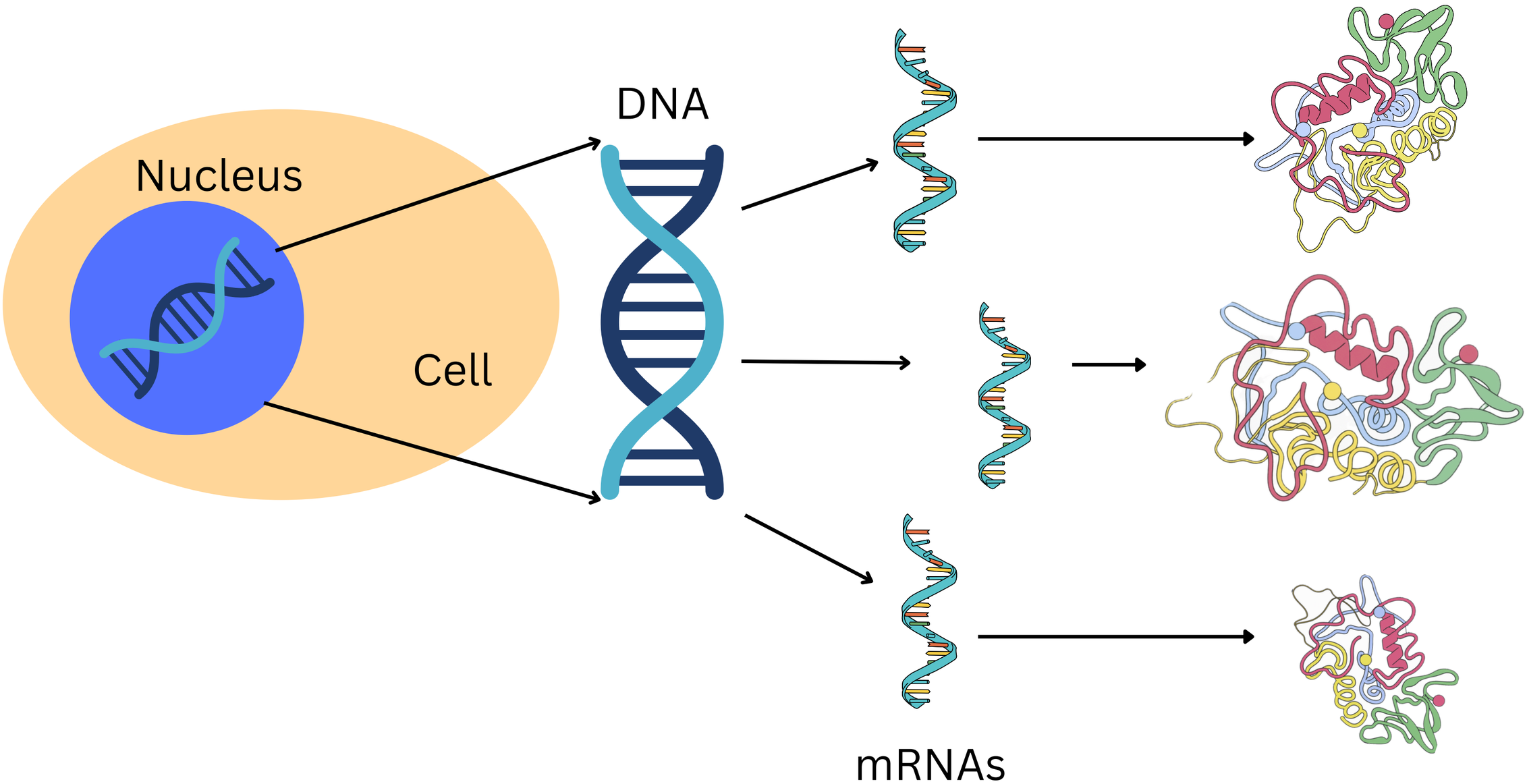
Genotype guided de novo molecule generation

ML-Aided Anti-Cancer Drug Discovery

- **Cancer** : Tumor that moves across the body (metastasis).
- Traditional approaches use target (protein) specific information to generate new molecules.
- But, **“biological context”** of any type of cancer lies in the broader picture of cell biology.
- Cancer cell biology takes into account the **transcriptomic signatures**, **metabolic pathways**, **mutations** and similar genetic parameters to complete the picture.

Central Dogma of Molecular Biology

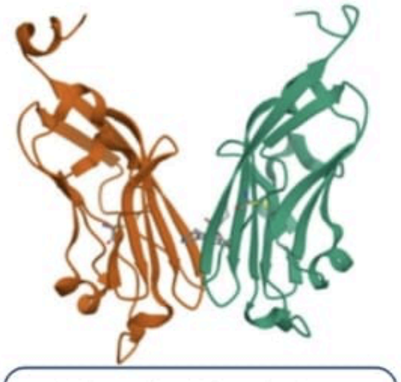


Cancer and its working

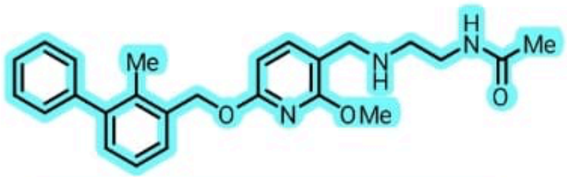
DNA ⟶ RNA ⟶ Protein ⟶ Functions

Omics	Normal Cell	Cancer Cell
Genomics	DNA	DNA (mutations)
Transcriptomics	↓	↓
Proteomics	RNA	RNA (shift in transcriptome)
Systems biology	↓	↓
	Protein	Protein (abnormal functions)
	↓	↓
	System level	System level (misregulation of pathways)

Prediction of IC50 values from PD-L1 Inhibitors



PD-L1 Protein
(Cancer Target)



Reported inhibitors
Total inhibitors = 300

PDL1
Inhibitors
(300 Drugs)

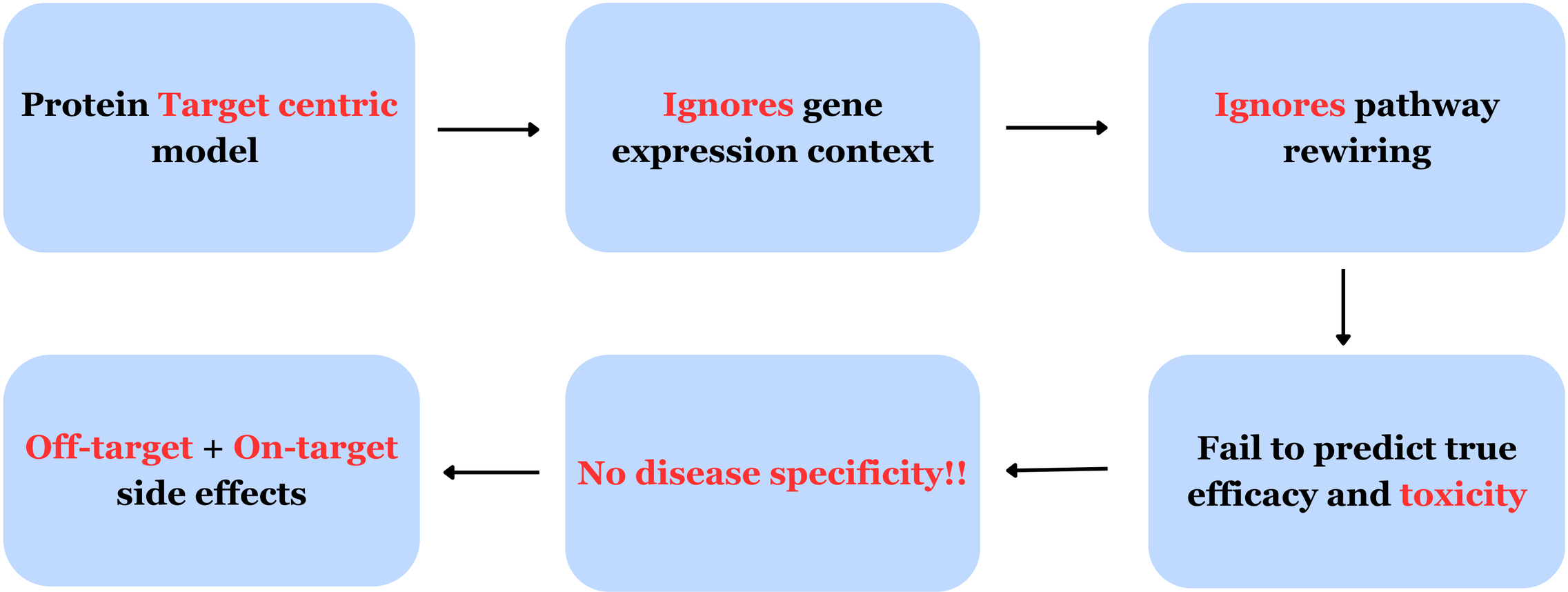
Biological
Activity
(IC50)

Gen AI

New molecule

Synthesis & Bio
evaluation of candidate
Molecules

Challenges faced by Target-based models



Problem Statement

Target based drug discovery overlooks the molecular biology of cancer which leads to high rejection rate, toxicity and side-effects

Hypothesis

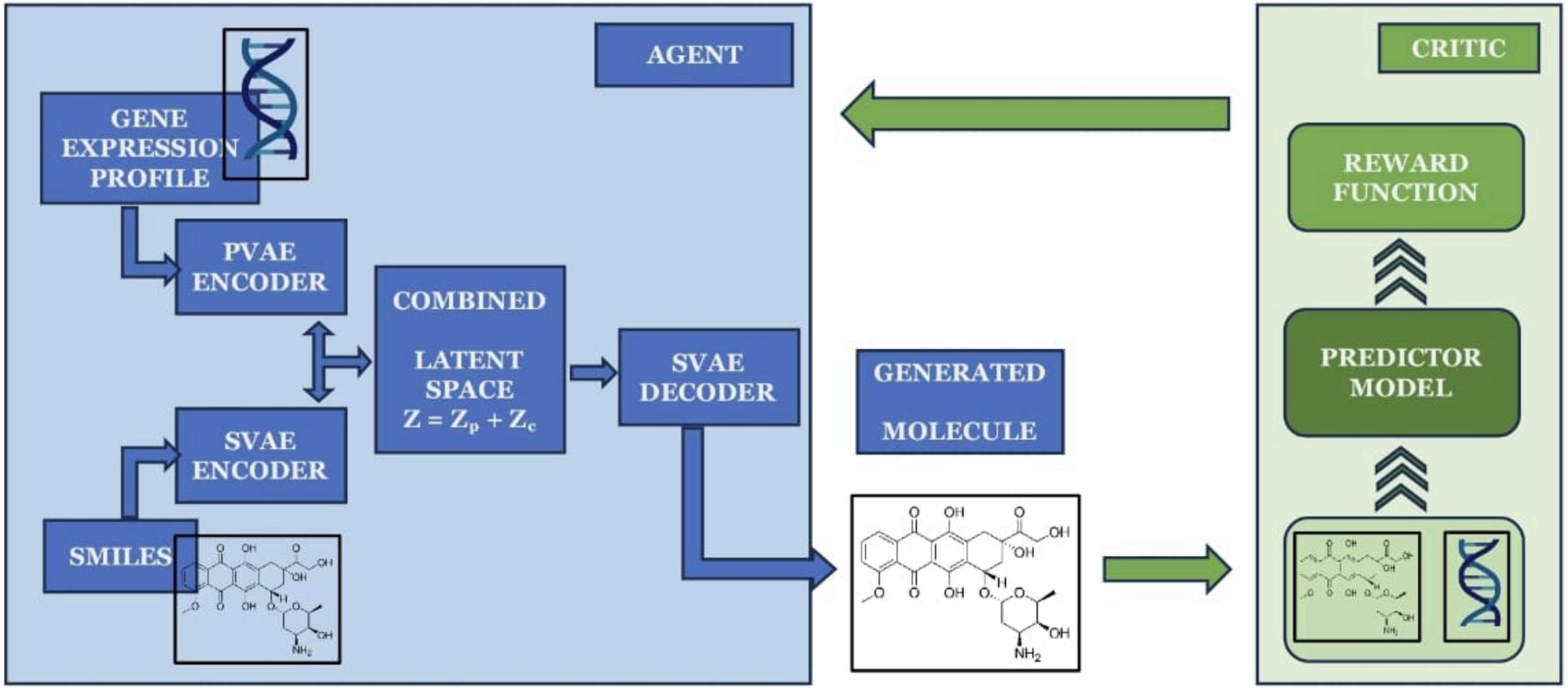
If we condition molecule generation on transcriptomic profiles, the model can learn to design molecules specifically effective for that cancer profile

Objectives

- **To build a generative model to generate new anti-cancer molecules based only on gene expression profiles for a particular cancer type.**

- **To use reinforcement learning in a manner that the generated molecules have high efficacy (low IC50)**

Pipeline


AGENT
GENE EXPRESSION PROFILE
PVAE ENCODER
COMBINED LATENT SPACE $Z = Z_p + Z_c$
SVAE DECODER
SVAE ENCODER
SMILES
GENERATED MOLECULE
CRITIC
REWARD FUNCTION
PREDICTOR MODEL

H21

	A	B
1	SMILES	canonical
2	CC(=O)OC1=C(CC2CCCCC2)C(=O)c2ccccc2C1=O	yes
3	COc1ccc([N+](=O)[O-])cc1NC(=O)c1ccco1	yes
4	CC(C)C[C@@H]1NC(=O)[C@H](CC(C)C)NC(=O)[C@H](Cc2ccccc2)NC(=O)[C@@H](CO)NC(=O)[C@H](CC(C)C)NC1=O	yes
5	Cc1ccc(F)cc1S(=O)(=O)N[C@@H]1CCN(Cc2ccc3[nH]ccc3c2)C1	yes
6	CC(=O)N[C@@H](CSC(=O)Nc1ccc(C(F)(F)F)cc1[N+](=O)[O-])C(=O)O	yes
7	N#C[C@@H]1C[C@@H]2C[C@@H]2N1C(=O)[C@@H](N)C12CC3CC(CC(OS(=O)(=O)O)(C3)C1)C2	yes
8	COc1ccccc1NS(=O)(=O)c1cc(-c2cnc(C3CC3)o2)ccc1C	yes
9	CCCCCCCCCC[C@@H](O)[C@@H]1CC[C@@H]([C@@H]2CC[C@H]([C@H](O)CCCCCCCCCC[C@@H](O)CC3=C[C@H](C)OC3=O)O2)O1	yes
10	CC(C)c1ccc(OC(C)(Cc2ccc3ccccc3c2)C(=O)O)cc1	yes
11	CSc1nc2cc(C)nn2c(-c2ccccc2Cl)c1C#N	yes
12	CC(=O)c1cc(C(=O)NC2(c3ccc(Br)cc3)CC2)n(C)c1	yes
13	CCCCCC(C)NC(=O)c1ccoc1C	yes
14	COc1ccc(OC)c(/C=C(/C#N)c2nc(-c3ccc(-c4ccccc4)cc3)cs2)c1	yes
15	c1ccc(Nc2ncnc3c2nc2n3Cc3ccccc3N2CCN2CCOCC2)cc1	yes
16	C[C@H](NC1=NS(=O)(=O)c2sc(Cl)cc2N1)c1ccc(Br)cc1	yes
17	C[N+](C)(C)CCOP(=O)([O-])OCCNC(=O)c1ccc2ccccc2c1	yes
18	O=S(=O)(CCC1CCc2ccccc2N1S(=O)(=O)c1ccc(Cl)cc1)N1CCC(NCc2cccs2)CC1	yes
19	Cc1ccc(N2CCN(c3ccc4c(ncc5c4c(=O)c(C(=O)O)cn5C)c3F)CC2)cc1F	yes
20	CC(C)CC#Cc1cc(-c2nn(CCCN3CCOCC3)c3c2CN(S(C)(=O)=O)CC3)ccc1Cl	yes
21	COc1cccc(-c2c(C#N)c(=O)oc3c2ccc2c3ccn2C)c1	yes

ChemBL

TCGA
(The Cancer Genome Atlas)

lusc-tcga-rnaseq_gene-expression.csv

gene_expression · Updated 30 Oct 2019 by Jannis Born

	ST6GALNAC5	ENSA	C21orf62	COL7A1
LUSC-TCGA-18-3406-01	6.237959233758621	9.501716408936643	0.0	5.762407745481894
LUSC-TCGA-18-3407-01	6.2455277977315555	8.834452572583226	0.45937232387854676	8.837772537471375
LUSC-TCGA-18-3408-01	4.824961138544715	9.367482272392754	0.5629852073373928	5.784074001087402
LUSC-TCGA-18-3409-01	5.578291732106161	8.492690354834286	3.322459467696414	7.793005278694553
LUSC-TCGA-18-3410-01	5.079612837664114	8.755875352014533	0.7231132852067131	7.972666114835727
LUSC-TCGA-18-3411-01	5.124942943850648	9.42566910544004	0.0	8.562899174110344
LUSC-TCGA-18-3412-01	5.077751950905447	9.273924815674034	0.0	7.75371102305417
LUSC-TCGA-18-3414-01	4.673573779677196	9.040661484989585	1.3725645673979932	7.688344703674084
LUSC-TCGA-18-3415-01	5.728532489975753	9.89728337890642	0.0	8.888716904011728
LUSC-TCGA-18-3416-01	2.6792401670378627	9.053340831646604	1.7496914841547744	9.441334774897314

Patient ID

Gene name/ ID

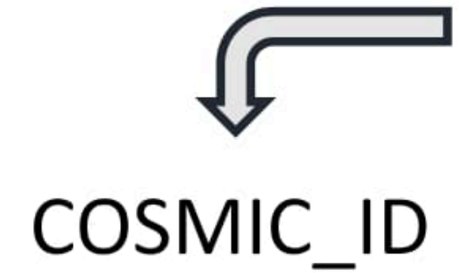
Log2 normalized abundance

Pretraining Datasets

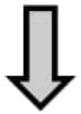
Training Dataset : GDSC (Genomics of Drug Sensitivity in Cancer)

- Built by Sanger/MGH ; tight cross-refs to **COSMIC** cell-line records
- All baseline cancer transcriptomic data, **Pre-treatment**
- **Drug response :** viability after ~ 72 h treatment; dose-response modeling - IC50, AUC.
- **~1000 cell lines** across many tissues
  - **GDSC1:970 cell lines** screened with **403 compounds** (333,292 IC50s)
  - **GDSC2:969 cell lines** screened with **297 compounds** (243,466 IC50s)

Training Dataset : GDSC (Genomics of Drug Sensitivity in Cancer)



COSMIC_ID



Cell_line_name

	drug	cell_line	IC50
0	SGC0946	1240121	2.06379275
1	SGC0946	1240122	1.9901499
2	SGC0946	1240123	1.27559563
3	SGC0946	1240124	2.6672847
4	SGC0946	1240125	0.99012768
5	SGC0946	1240127	2.55048192
6	SGC0946	1240128	1.48934683
7	SGC0946	1240129	2.4270105

	YWHAB	YWHAE	YWHAH
143b	6.761574108393271	7.407318078670741	5.886111747410234
201T	6.086779892161543	6.144190239078273	4.605270170991424
451Lu	6.177948418356412	6.881412357741531	5.7991018370705
A388	5.991470797049389	6.419997580684626	5.298342365610589
A427	6.400260204648842	6.126873951342351	5.087634437002271
A431	6.759256616476298	6.598510884737168	5.863639246264856
ACN	6.779923198792065	6.625394126968762	5.438098211978882

GDSC Drug Sensitivity

- Similarly, the internal drug_id for GDSC links to the drug molecule completing the picture

	A	B
1	CC(C)CC(=O)NC1=NNC2=C1CN(C2(C)C)C(=O)C3CCN(CC3)C	PHA-793887
2	CC(=C(C#N)C(=O)NC1=C(C=CC(=C1)Br)Br)O	LFM-A13
3	CCN1CCN(CC1)CC2=C(C=C(C=C2)NC(=O)C3=CC(=C(C=C3)C)C=CC4=CN=C5C(=C4OC)C=CN5	HG6-64-1
4	COC1=C(C=C2C(=C1)N=CN2C3=CC(=C(S3)C(=O)N)OCC4=CC=CC=C4C(F)(F)F)OC	GW843682X

	drug	cell_line	IC50
0	Erlotinib	MC-CAR	2.453524
1	Erlotinib	ES3	3.376592
2	Erlotinib	ES5	3.614664
3	Erlotinib	ES7	3.223394
4	Erlotinib	EW-11	2.486405
5	Erlotinib	SK-ES-1	2.048918

Agent (Molecule Generator)

- Dual VAE architecture - profile VAE (PVAE) and SMILES VAE (SVAE)
- Both VAEs are first pre-trained individually on large transcriptomic and molecular datasets (TCGA and ChemBL respectively) and then jointly fine-tuned on GDSC dataset
- Genetic embedding and SMILES embedding are combined together using gaussian addition and then decoded using SMILES decoder for generation of new molecules.

Critic : Methodology

Let $q_\theta(x'|z_p + z_c)$ represent the decoder
The decoder constructs the molecule from the combined embedding in a token-by-token manner.

At each point in the molecule generation, let the state of the be S_t

$$S_T = tuple(C_T, x_c)$$

where t = T represents the stage of complete molecule generation

Critic : Methodology

$$p(S_T) = \prod_{t=1}^{T} p(a_t|S_{t-1})$$

Once the molecule is generated, the reward function assigns a reward score

$$R(S_T) = exp(-f(C_T, x_c)/\alpha)$$

f is the regression model trained on GDSC dataset to estimate the IC50 value for a drug molecule

Critic : Methodology

The objective is to maximize the expected value of the reward function

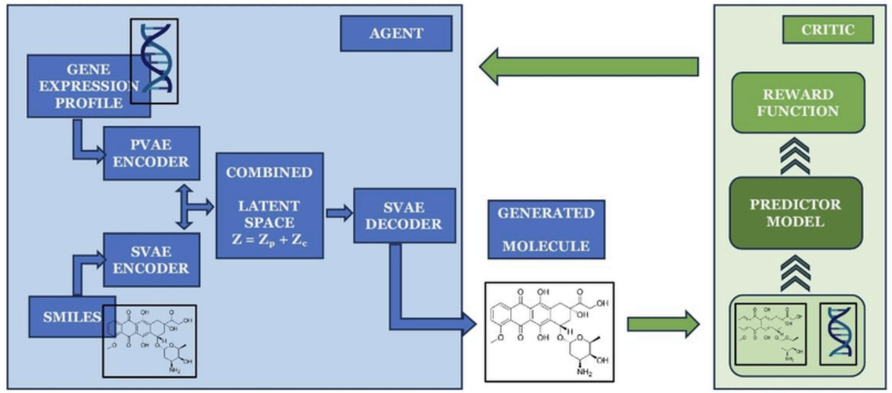
$$E(R(S_T)) = \sum_{T \in M} p(S_T) R(S_T)$$

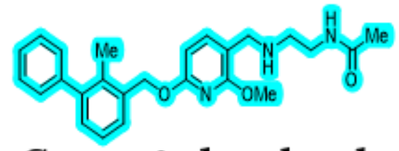
M is the set of all possible molecules that can be generated

$$\frac{\partial(E(R(S_T)))}{\partial \theta} = 0$$

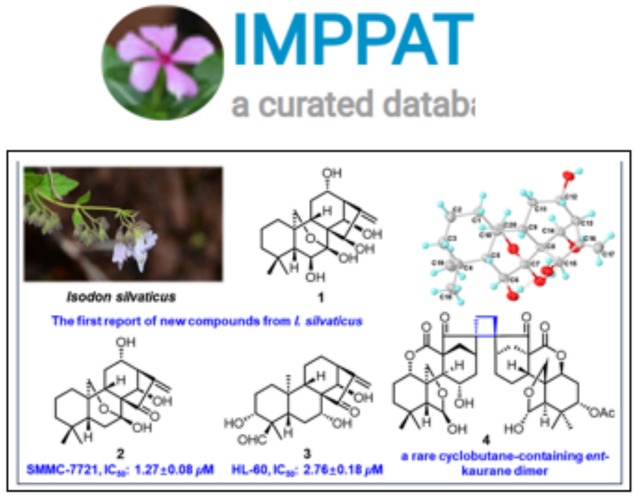
Hence we get the optimal parameters for the decoder

Plant extract based approach toward anti-cancer drug





Generated molecule with high Predicted efficacy against a cell line



Similarity index between IMPPAT and generated molecules

Sorted Phyto-molecules proceed for biological evaluation