Single Cell Perturbation scPerturb

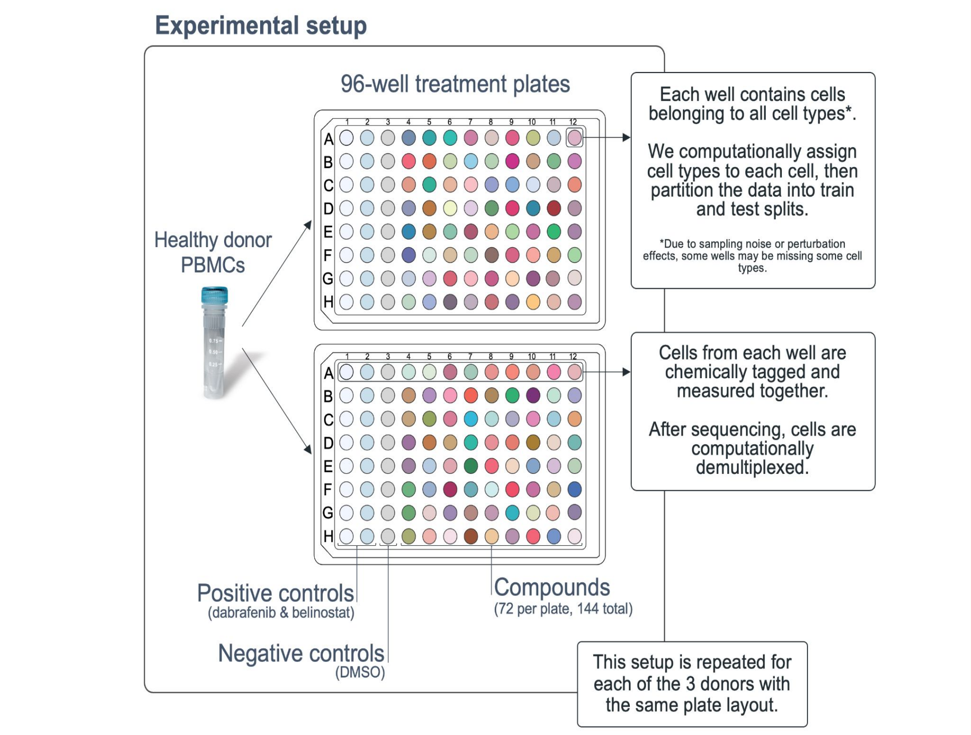
DS542 DL4DS
Sicheng Yi (Tiger Yi)
tigeryi@bu.edu

Table of Contents

1. Background
2. Data
3. Augmentation
4. Model
5. Result

NeurIPS 2023 Kaggle competition

Background



Experimental setup

96-well treatment plates

Healthy donor PBMCs

Each well contains cells belonging to all cell types*.

We computationally assign cell types to each cell, then partition the data into train and test splits.

*Due to sampling noise or perturbation effects, some wells may be missing some cell types.

Cells from each well are chemically tagged and measured together.

After sequencing, cells are computationally demultiplexed.

Positive controls (dabrafenib & belinostat)

Negative controls (DMSO)

Compounds (72 per plate, 144 total)

This setup is repeated for each of the 3 donors with the same plate layout.

PBMC: peripheral blood molecular cell

- T cells, NK cells, B cells, myeloid cells

144 compounds (drug, sm_names)

16 positive control: dabrafenib, belinostat

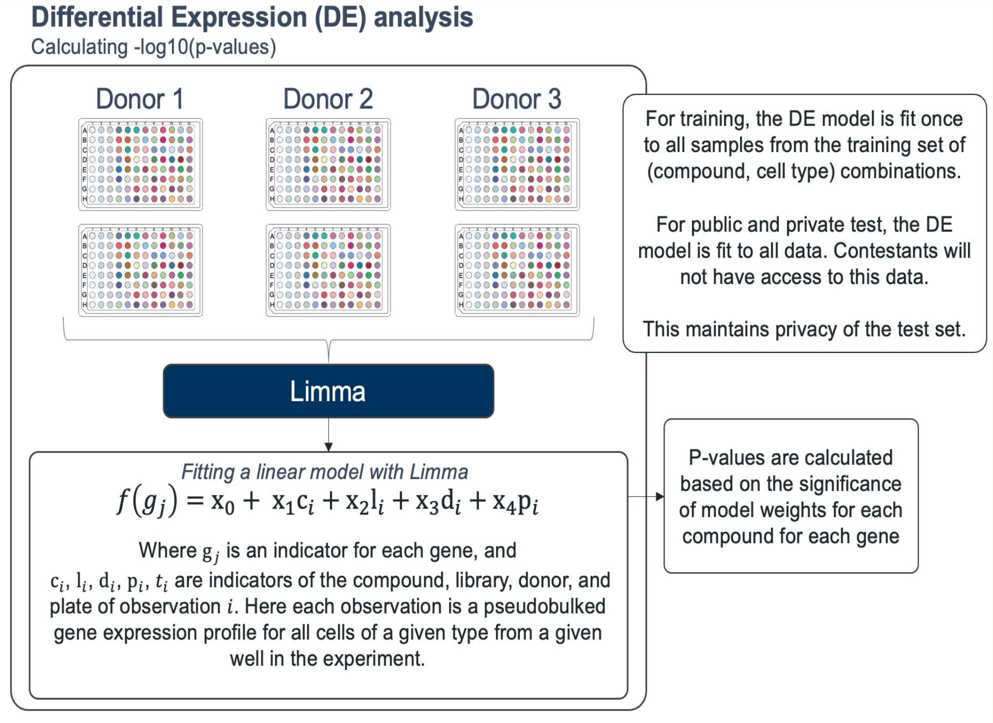
8 negative control: DMSO

scRNA-seq: raw RNA counts

scATAC-seq: chromatin peaks

DE differential expression

Background



Differential Expression (DE) analysis

Calculating -log10(p-values)

For training, the DE model is fit once to all samples from the training set of (compound, cell type) combinations.

For public and private test, the DE model is fit to all data. Contestants will not have access to this data.

This maintains privacy of the test set.

Fitting a linear model with Limma

$$f(g_j) = x_0 + x_1 c_i + x_2 l_i + x_3 d_i + x_4 p_i$$

Where g_j is an indicator for each gene, and c_i, l_i, d_i, p_i, t_i are indicators of the compound, library, donor, and plate of observation i. Here each observation is a pseudobulked gene expression profile for all cells of a given type from a given well in the experiment.

P-values are calculated based on the significance of model weights for each compound for each gene

DE differential expression:

Identify change in gene expression level

Between control and perturbed cells

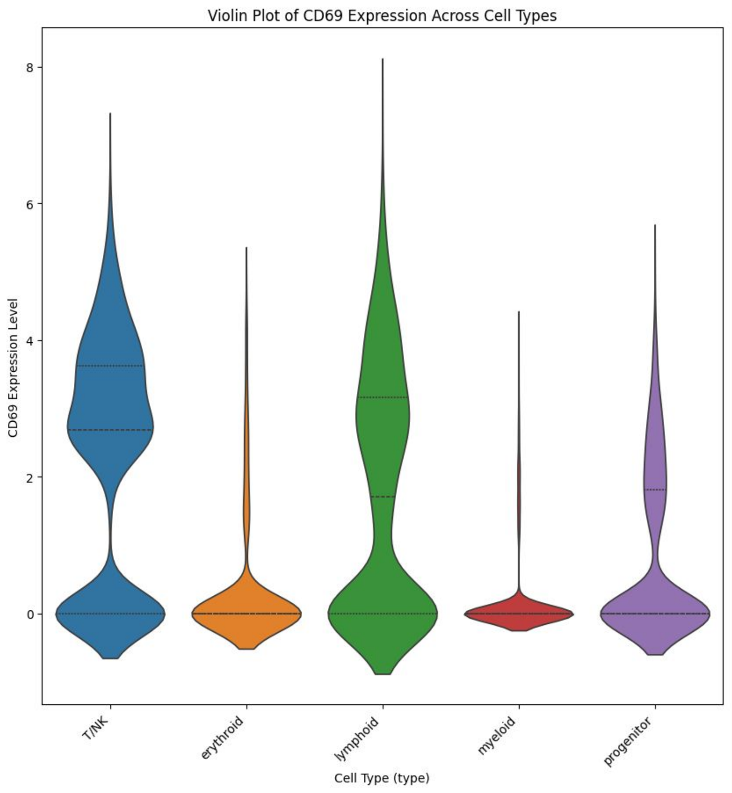
LIMMA: linear model

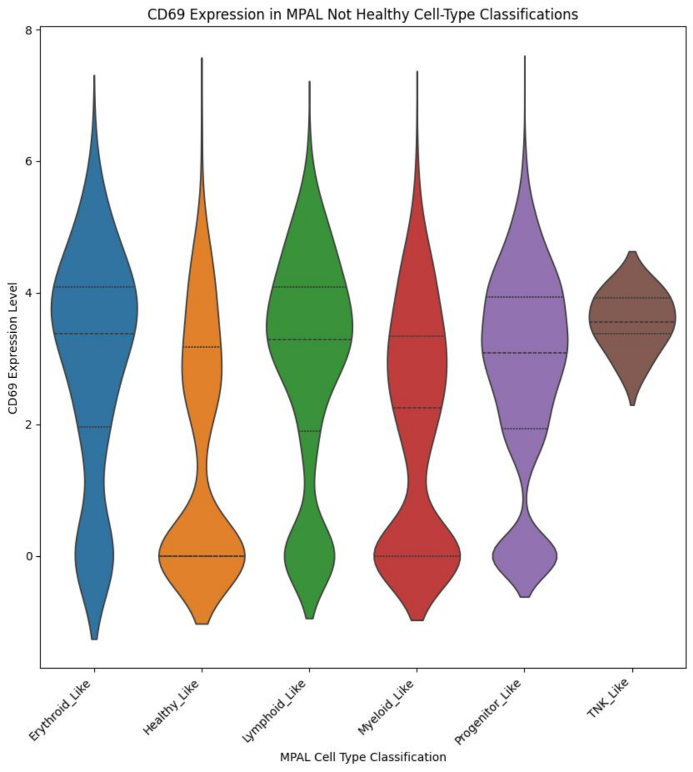
Calculate DE value

-log10(p-value) * sign(Log Fold Change)

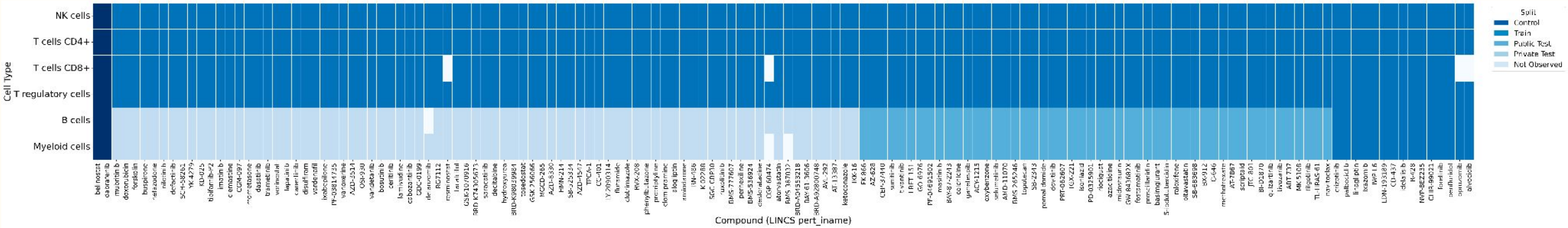
144 drug compounds, 18000+ genes

EDA


Violin Plot of CD69 Expression Across Cell Types
CD69 Expression Level
0
2
4
6
8
T/NK
erythroid
lymphoid
myeloid
progenitor
Cell Type (type)


CD69 Expression in MPAL Not Healthy Cell-Type Classifications
CD69 Expression Level
0
2
4
6
8
Erythroid_Like
Healthy_Like
Lymphoid_Like
Myeloid_Like
Progenitor_Like
TNK_Like
MPAL Cell Type Classification

Data

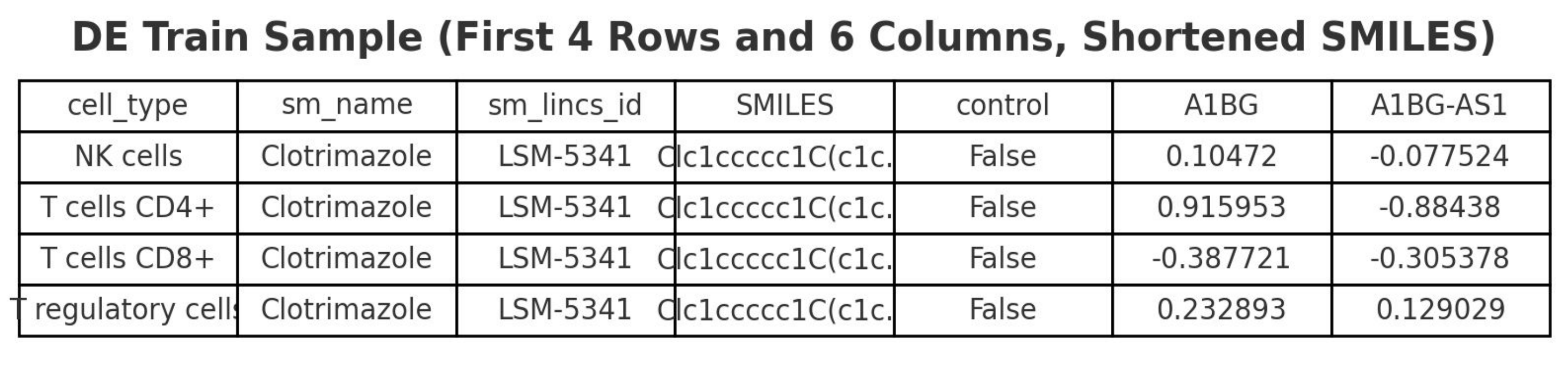


Test: Predict differential expression for majority of Myeloid cells and B cells

Train: 144 compounds in T, NK cells. 15 compounds + controls in Myeloid, B cells.

T cells: CD4+ , CD8+, regulatory

Data - de_train



DE Train Sample (First 4 Rows and 6 Columns, Shortened SMILES)

cell_type	sm_name	sm_lincs_id	SMILES	control	A1BG	A1BG-AS1
NK cells	Clotrimazole	LSM-5341	Clc1ccccc1C(c1c.	False	0.10472	-0.077524
T cells CD4+	Clotrimazole	LSM-5341	Clc1ccccc1C(c1c.	False	0.915953	-0.88438
T cells CD8+	Clotrimazole	LSM-5341	Clc1ccccc1C(c1c.	False	-0.387721	-0.305378
T regulatory cells	Clotrimazole	LSM-5341	Clc1ccccc1C(c1c.	False	0.232893	0.129029

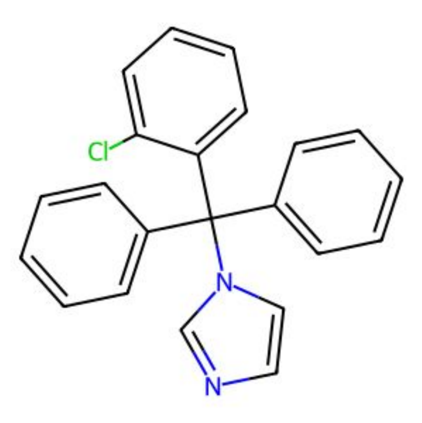
Row (614): cell type, sm compound pairs

Column (18216):

18211 genes of DE value: -log10(p-value)

control: true or false

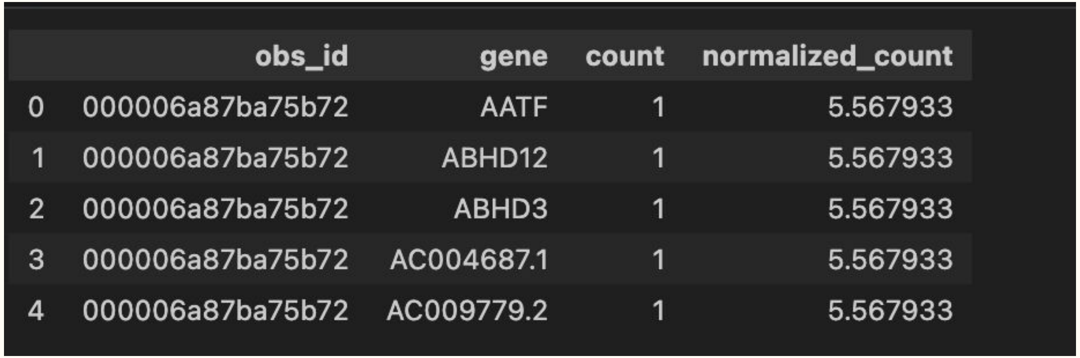
SMILES: single line 1D molecular structure



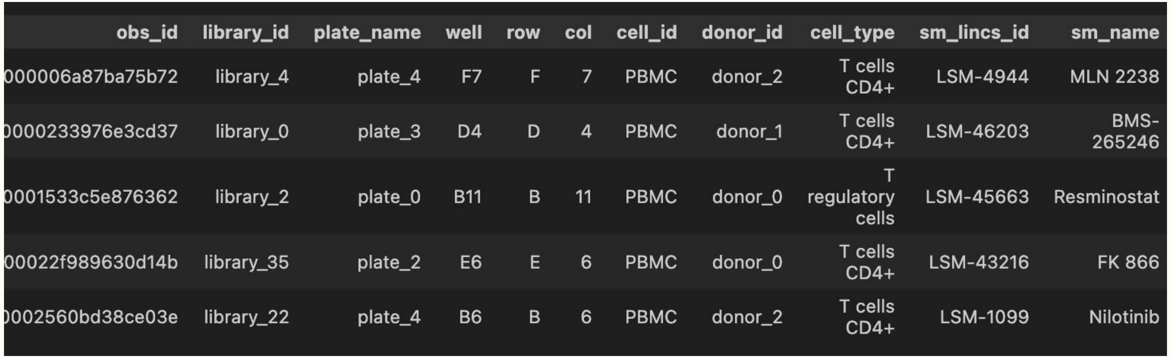
+ up regulated
- down regulated

number: -log10(p)
Higher, more significant

Data - adata



	obs_id	gene	count	normalized_count
0	000006a87ba75b72	AATF	1	5.567933
1	000006a87ba75b72	ABHD12	1	5.567933
2	000006a87ba75b72	ABHD3	1	5.567933
3	000006a87ba75b72	AC004687.1	1	5.567933
4	000006a87ba75b72	AC009779.2	1	5.567933



obs_id	library_id	plate_name	well	row	col	cell_id	donor_id	cell_type	sm_lincs_id	sm_name
000006a87ba75b72	library_4	plate_4	F7	F	7	PBMC	donor_2	T cells CD4+	LSM-4944	MLN 2238
0000233976e3cd37	library_0	plate_3	D4	D	4	PBMC	donor_1	T cells CD4+	LSM-46203	BMS-265246
0001533c5e876362	library_2	plate_0	B11	B	11	PBMC	donor_0	T regulatory cells	LSM-45663	Resminostat
00022f989630d14b	library_35	plate_2	E6	E	6	PBMC	donor_0	T cells CD4+	LSM-43216	FK 866
0002560bd38ce03e	library_22	plate_4	B6	B	6	PBMC	donor_2	T cells CD4+	LSM-1099	Nilotinib

scRNA-seq

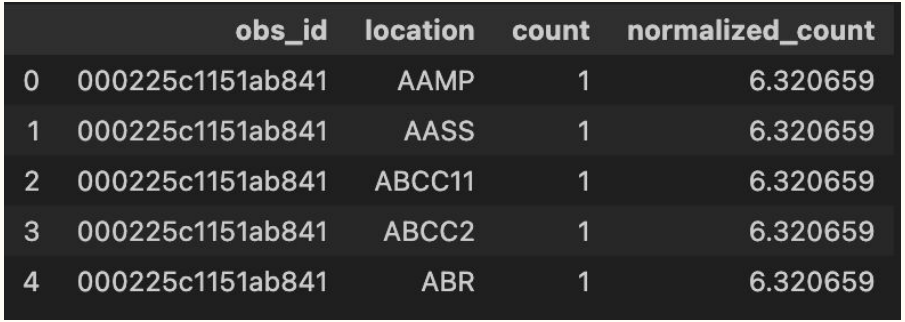
obs: individual cell

gene: column detrain

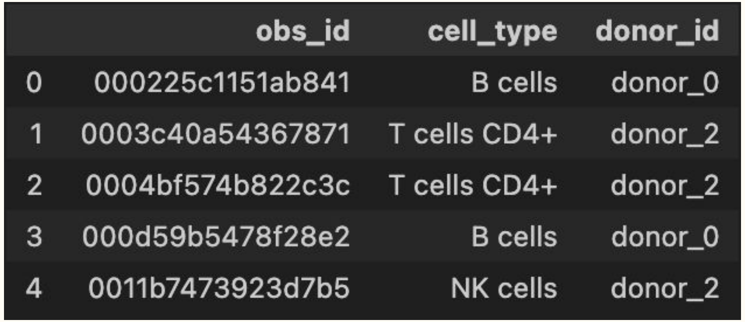
count: raw molecular counts of gene in cell

Norm: log(X+1)

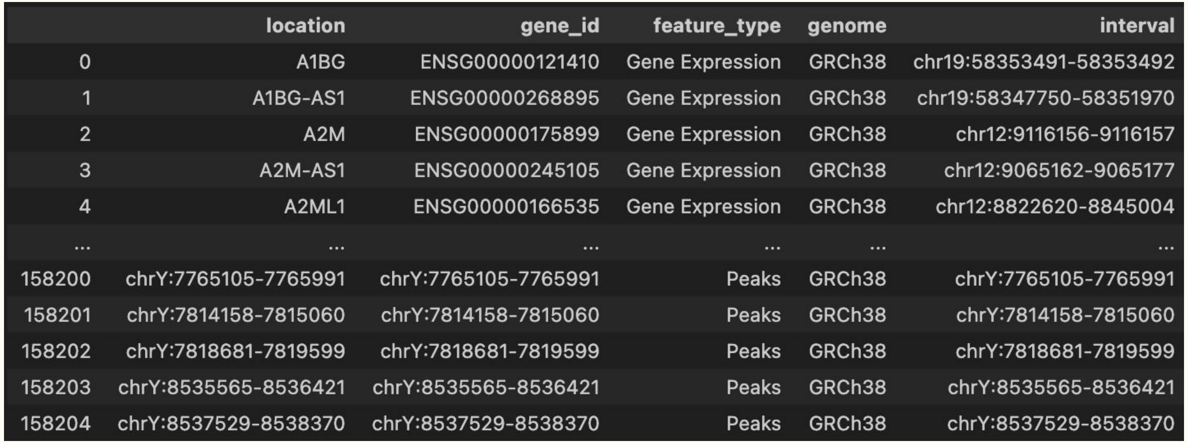
Data - multiome



	obs_id	location	count	normalized_count
0	000225c1151ab841	AAMP	1	6.320659
1	000225c1151ab841	AASS	1	6.320659
2	000225c1151ab841	ABCC11	1	6.320659
3	000225c1151ab841	ABCC2	1	6.320659
4	000225c1151ab841	ABR	1	6.320659



	obs_id	cell_type	donor_id
0	000225c1151ab841	B cells	donor_0
1	0003c40a54367871	T cells CD4+	donor_2
2	0004bf574b822c3c	T cells CD4+	donor_2
3	000d59b5478f28e2	B cells	donor_0
4	0011b7473923d7b5	NK cells	donor_2

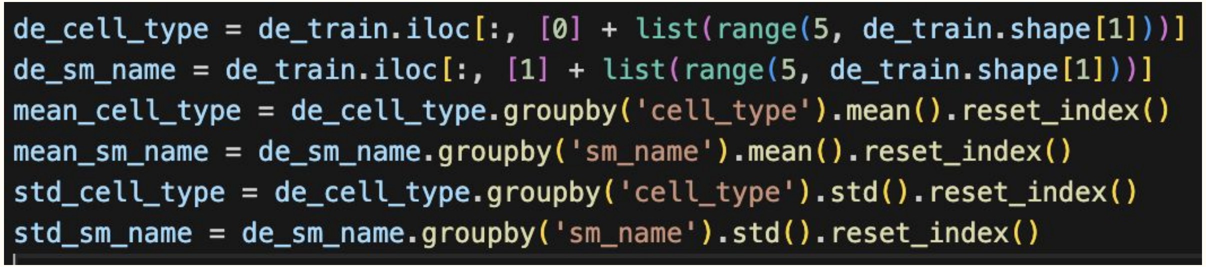


	location	gene_id	feature_type	genome	interval
0	A1BG	ENSG00000121410	Gene Expression	GRCh38	chr19:58353491-58353492
1	A1BG-AS1	ENSG00000268895	Gene Expression	GRCh38	chr19:58347750-58351970
2	A2M	ENSG00000175899	Gene Expression	GRCh38	chr12:9116156-9116157
3	A2M-AS1	ENSG00000245105	Gene Expression	GRCh38	chr12:9065162-9065177
4	A2ML1	ENSG00000166535	Gene Expression	GRCh38	chr12:8822620-8845004
...	...	...	...	...	...
158200	chrY:7765105-7765991	chrY:7765105-7765991	Peaks	GRCh38	chrY:7765105-7765991
158201	chrY:7814158-7815060	chrY:7814158-7815060	Peaks	GRCh38	chrY:7814158-7815060
158202	chrY:7818681-7819599	chrY:7818681-7819599	Peaks	GRCh38	chrY:7818681-7819599
158203	chrY:8535565-8536421	chrY:8535565-8536421	Peaks	GRCh38	chrY:8535565-8536421
158204	chrY:8537529-8538370	chrY:8537529-8538370	Peaks	GRCh38	chrY:8537529-8538370

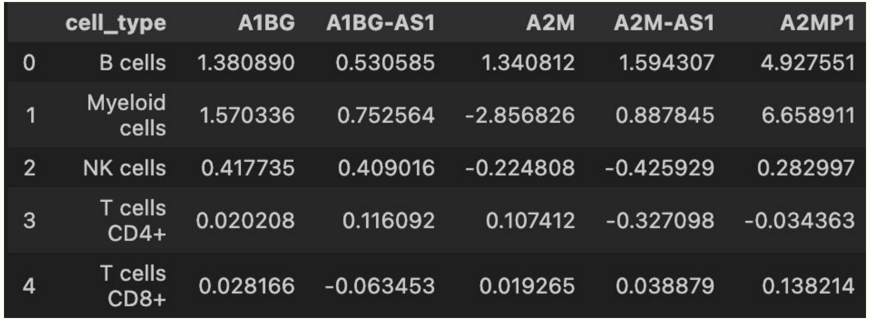
scRNA-seq:
gene expression

scATAC-seq:
chormation peaks

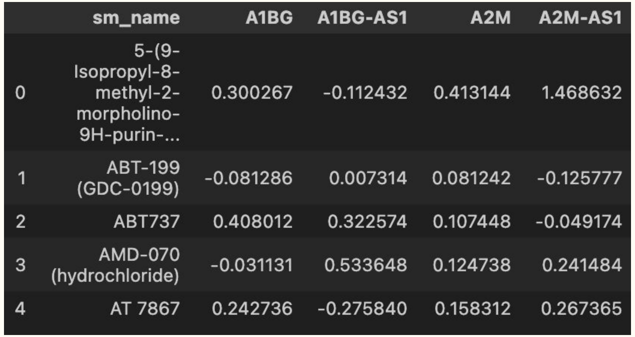
Augmentation - mean, std of cell types, sm_name



```
de_cell_type = de_train.iloc[:, [0] + list(range(5, de_train.shape[1]))]
de_sm_name = de_train.iloc[:, [1] + list(range(5, de_train.shape[1]))]
mean_cell_type = de_cell_type.groupby('cell_type').mean().reset_index()
mean_sm_name = de_sm_name.groupby('sm_name').mean().reset_index()
std_cell_type = de_cell_type.groupby('cell_type').std().reset_index()
std_sm_name = de_sm_name.groupby('sm_name').std().reset_index()
```

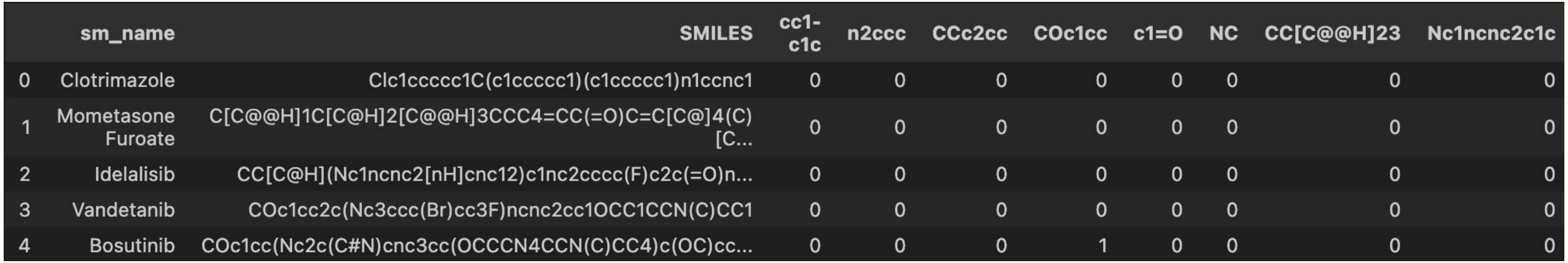


	cell_type	A1BG	A1BG-AS1	A2M	A2M-AS1	A2MP1
0	B cells	1.380890	0.530585	1.340812	1.594307	4.927551
1	Myeloid cells	1.570336	0.752564	-2.856826	0.887845	6.658911
2	NK cells	0.417735	0.409016	-0.224808	-0.425929	0.282997
3	T cells CD4+	0.020208	0.116092	0.107412	-0.327098	-0.034363
4	T cells CD8+	0.028166	-0.063453	0.019265	0.038879	0.138214

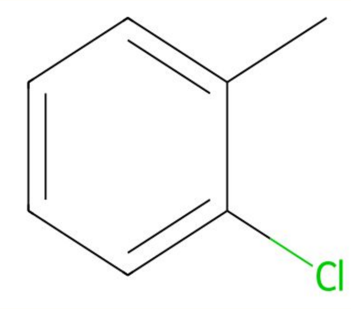


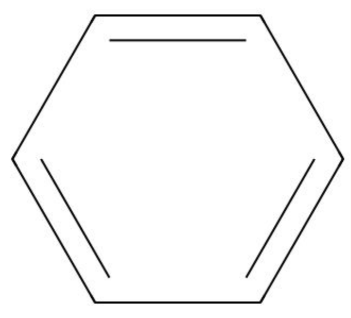
	sm_name	A1BG	A1BG-AS1	A2M	A2M-AS1
0	5-(9-Isopropyl-8-methyl-2-morpholino-9H-purin-...	0.300267	-0.112432	0.413144	1.468632
1	ABT-199 (GDC-0199)	-0.081286	0.007314	0.081242	-0.125777
2	ABT737	0.408012	0.322574	0.107448	-0.049174
3	AMD-070 (hydrochloride)	-0.031131	0.533648	0.124738	0.241484
4	AT 7867	0.242736	-0.275840	0.158312	0.267365

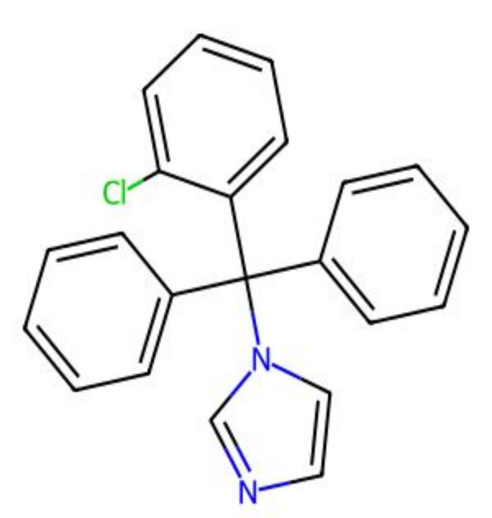
Augmentation - SMILES



	sm_name	SMILES	cc1-c1c	n2ccc	CCc2cc	COc1cc	c1=O	NC	CC[C@@H]23	Nc1ncnc2c1c
0	Clotrimazole	Clc1ccccc1C(c1ccccc1)(c1ccccc1)n1ccnc1	0	0	0	0	0	0	0	0
1	Mometasone Furoate	C[C@@H]1C[C@H]2[C@@H]3CCC4=CC(=O)C=C[C@]4(C)[C...	0	0	0	0	0	0	0	0
2	Idelalisib	CC[C@H](Nc1ncnc2[nH]cnc12)c1nc2cccc(F)c2c(=O)n...	0	0	0	0	0	0	0	0
3	Vandetanib	COc1cc2c(Nc3ccc(Br)cc3F)ncnc2cc1OCC1CCN(C)CC1	0	0	0	0	0	0	0	0
4	Bosutinib	COc1cc(Nc2c(C#N)cnc3cc(OCCCN4CCN(C)CC4)c(OC)cc...	0	0	0	1	0	0	0	0

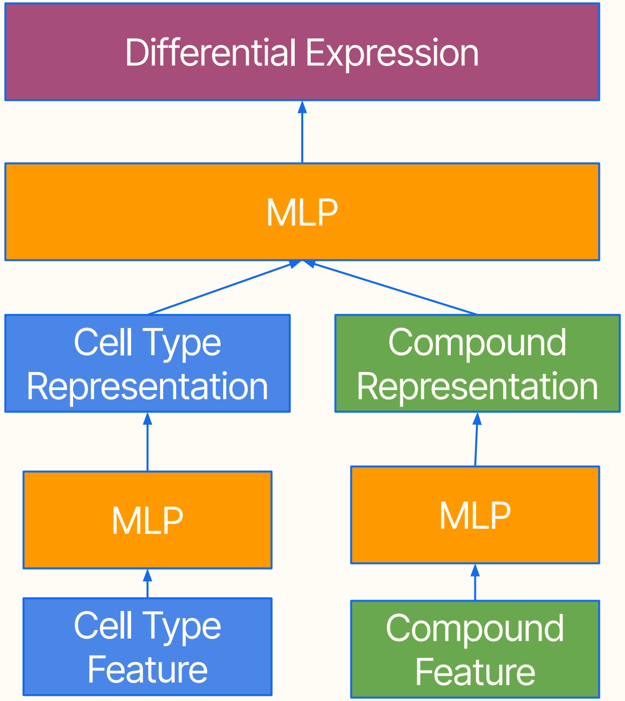






Clc1ccccc1C(c1ccccc1)(c1ccccc1)n1ccnc1
Element_Count: {'Cl': 1, 'C': 22, 'N': 2}

Model



Differential Expression

MLP

Cell Type Representation

Compound Representation

MLP

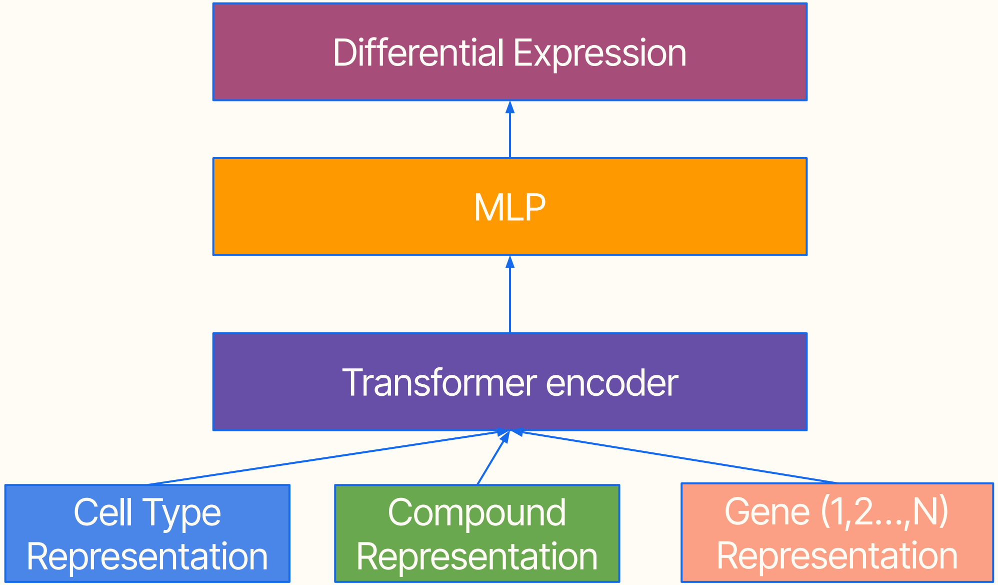
MLP

Cell Type Feature

Compound Feature

RNA-seq
ATAC-seq

sm: SMILES
controls



Differential Expression

MLP

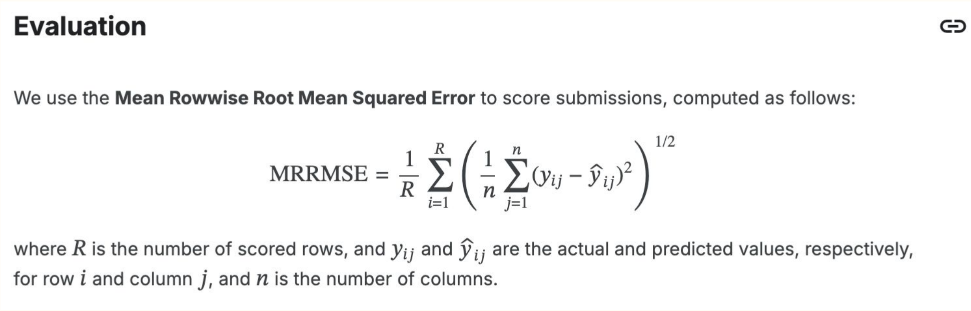
Transformer encoder

Cell Type Representation

Compound Representation

Gene (1,2...,N) Representation

Results



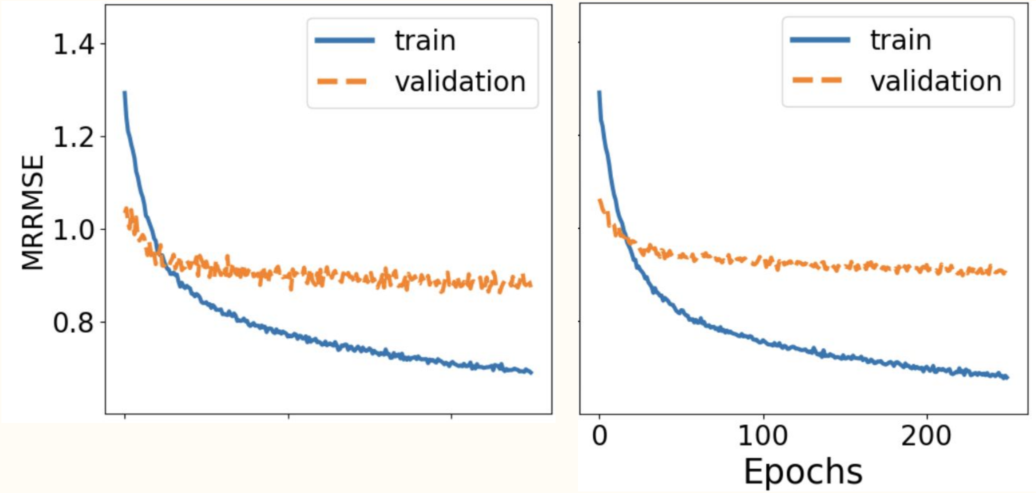
Evaluation

We use the **Mean Rowwise Root Mean Squared Error** to score submissions, computed as follows:

$$\mathrm{MRRMSE} = \frac{1}{R}\sum_{i=1}^{R}\left(\frac{1}{n}\sum_{j=1}^{n}(y_{ij} - \widehat{y}_{ij})^2\right)^{1/2}$$

where R is the number of scored rows, and y_{ij} and $\widehat{y}_{ij}$ are the actual and predicted values, respectively, for row i and column j, and n is the number of columns.

	MLP	Transformer
validation	0.86	0.90
test	0.93	0.87



Transformer seems to generalize better on the unseen test set.

Questions?