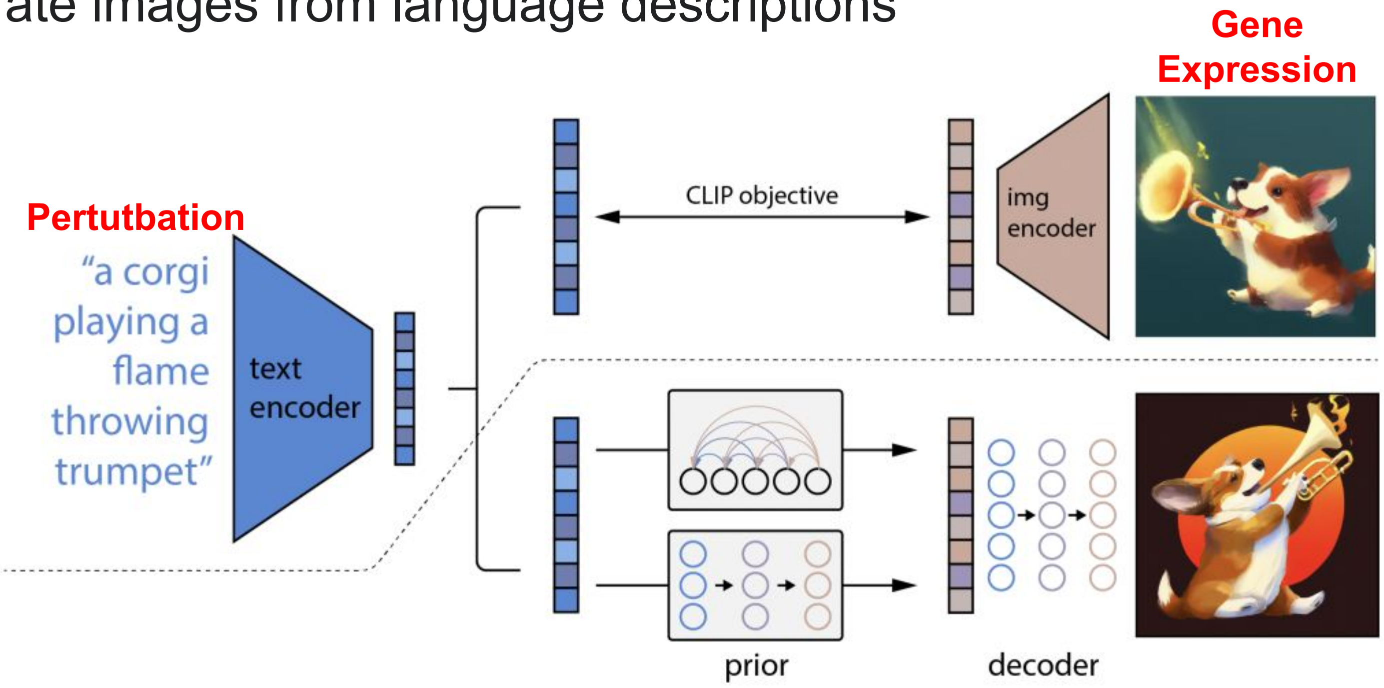
Predict Perturbation Response by adapting DALL-E2

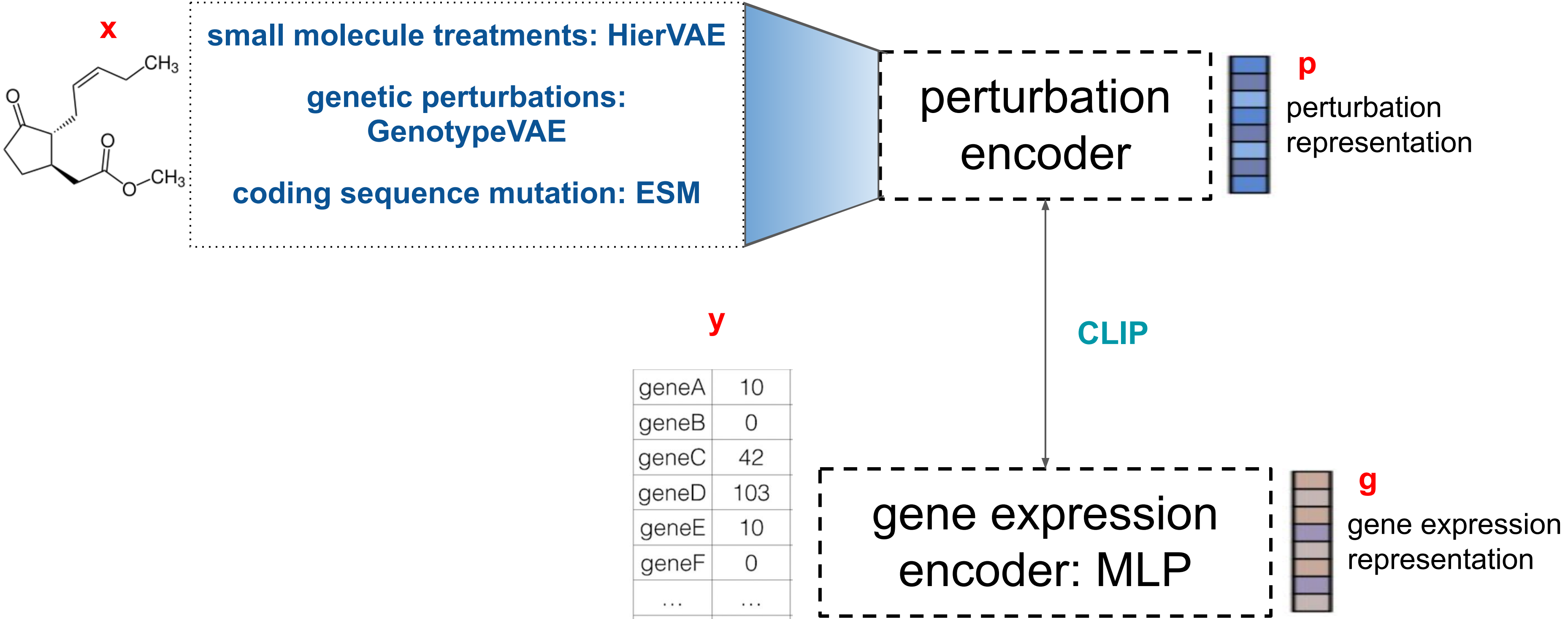
Ruoxi Gao

DALL-E2

Generate images from language descriptions

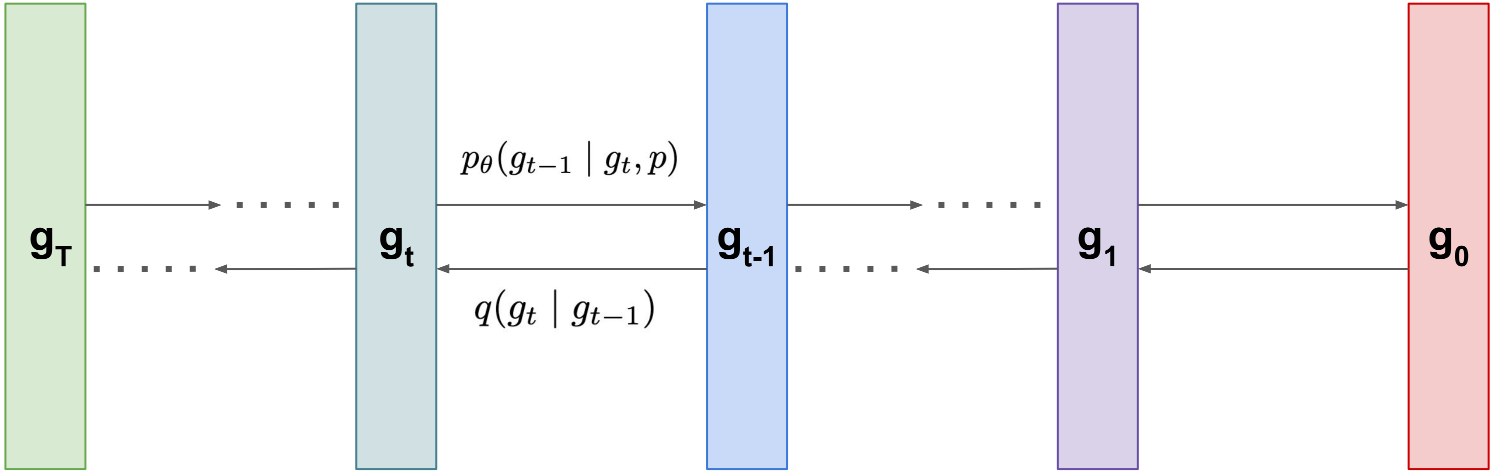


Encoders



geneA	10
geneB	0
geneC	42
geneD	103
geneE	10
geneF	0
...	...
...	...
...	...

Prior: conditoinal diffusion model

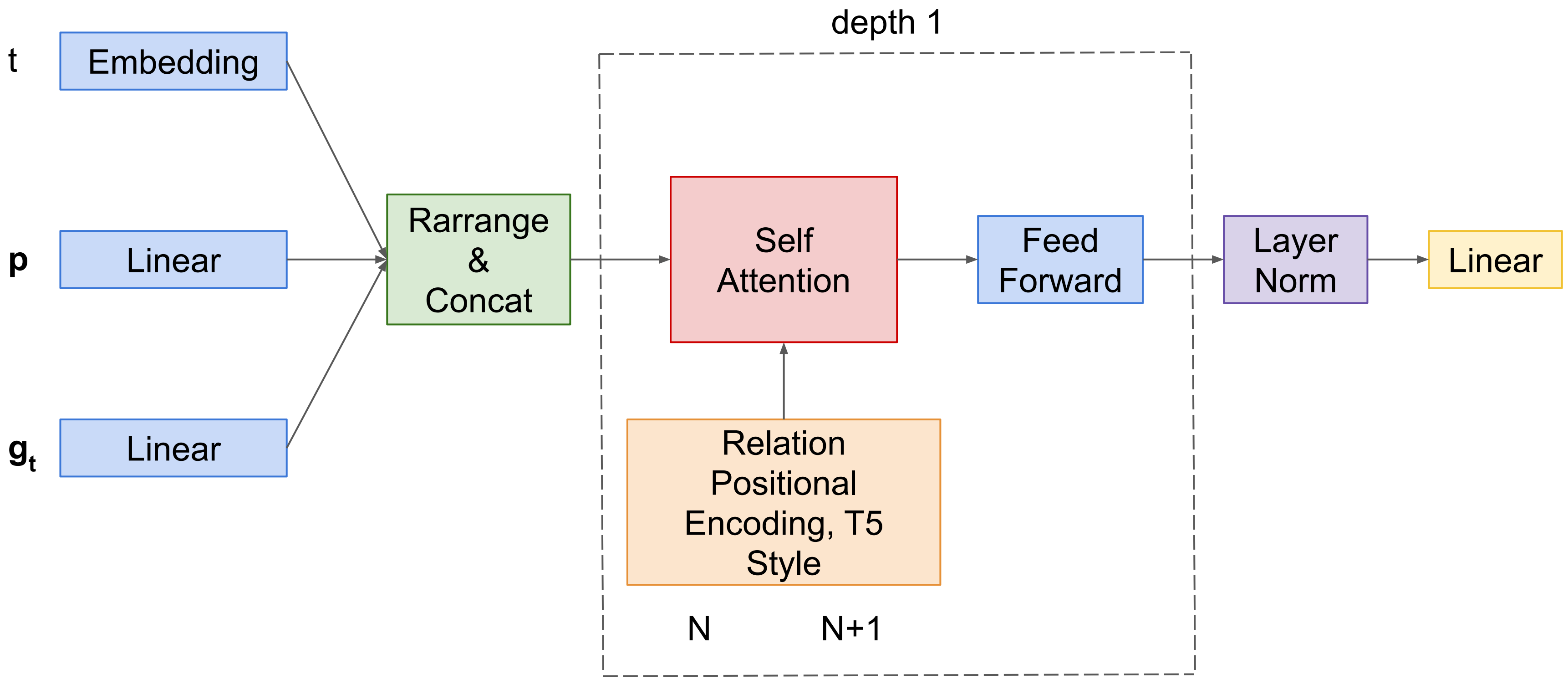


$$E[-log\ p_\theta(\mathbf{g_0} \mid \mathbf{p})] \leq E_q[-log\ \frac{p_\theta(\mathbf{g_{0:T}}|\mathbf{p})}{q(\mathbf{g_{1:T}}|\mathbf{g_0})}] = E_q[-log\ p(\mathbf{g_T}) - \sum_{t\geq 1} log\ \frac{p_\theta(\mathbf{g_{t-1}}|\mathbf{g_t},\mathbf{p})}{q(\mathbf{g_t}|\mathbf{g_{t-1}})}]$$

$$L_{simple}(\theta) = E_{\mathbf{p}}[E_{t,\mathbf{g_0(p)},\boldsymbol{\epsilon}}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(f(\mathbf{g_0},\boldsymbol{\epsilon}), t, \mathbf{p})\|^2]];$$

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}), \quad f(\mathbf{g_0}, \boldsymbol{\epsilon}) = \sqrt{\overline{\alpha}_t}\mathbf{g_0} + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}$$

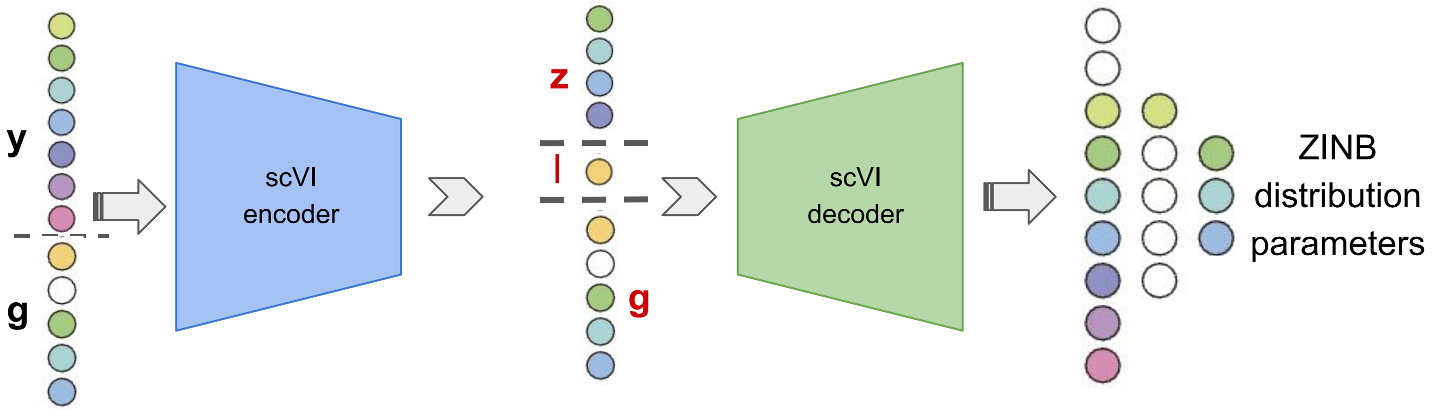
Prior: network architecture



Decoder: improved scVI

1. scVI encoder: encoder Gaussian variational posterior

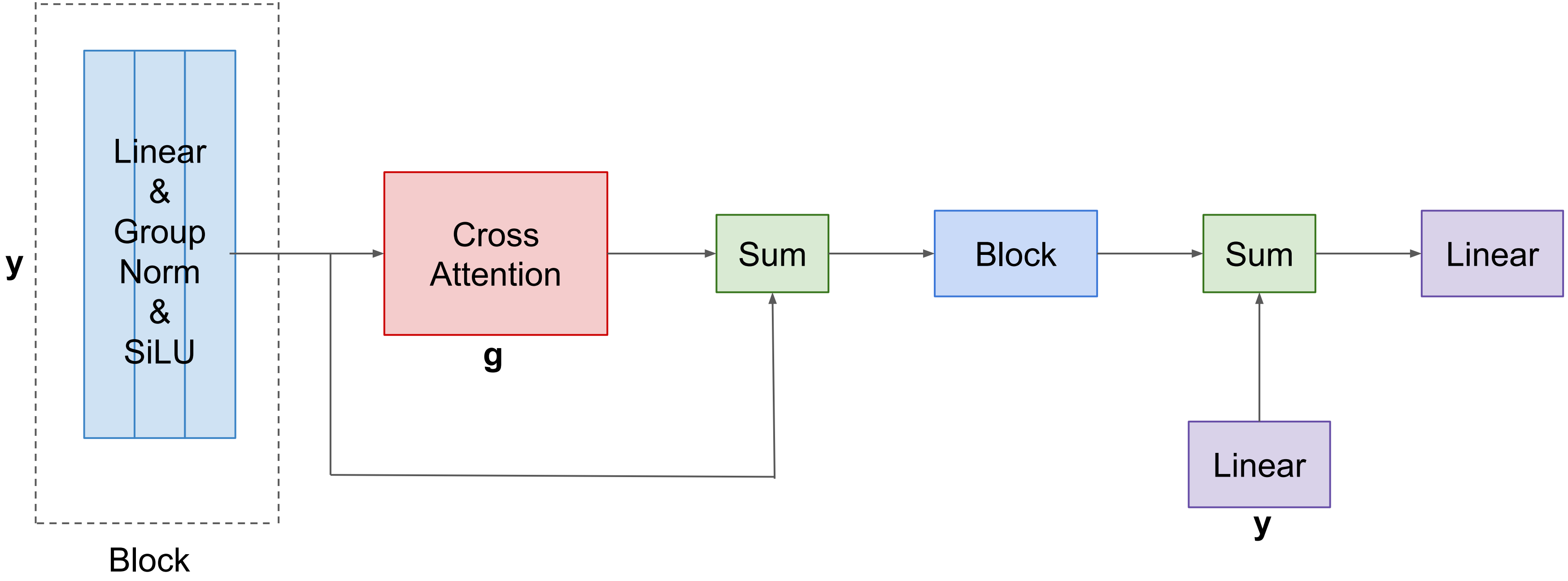
2. scVI decoder: estimate generative ZINB distribution parameters



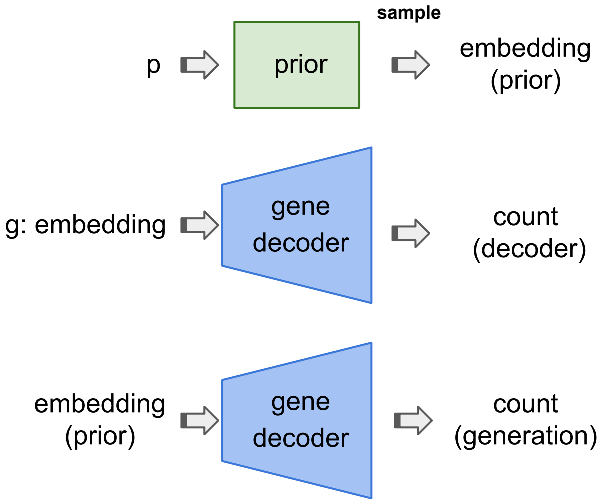
Optimize variatinoal lower bound:

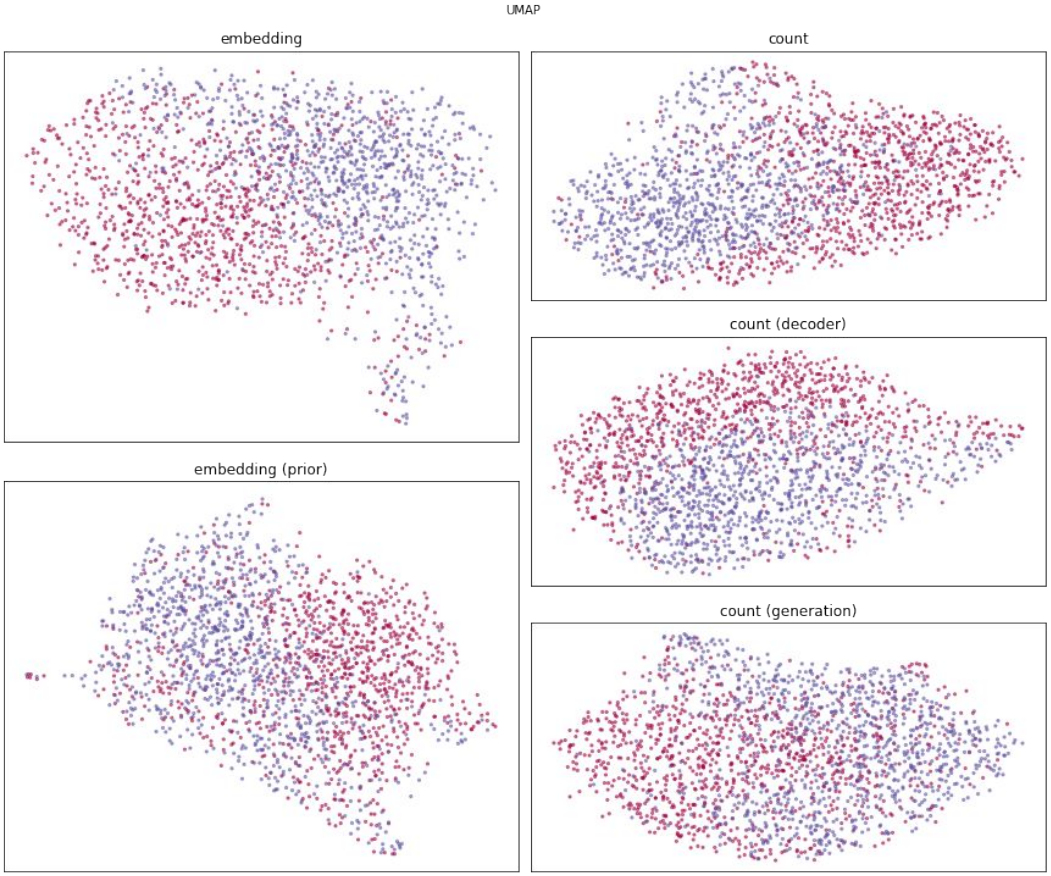
$$log\ p(\mathbf{y} \mid \mathbf{g}, \mathbf{p}) \geq E_{q(\mathbf{z},l|\mathbf{y},\mathbf{g})} log\ p(\mathbf{y} \mid \mathbf{z}, l, \mathbf{g}) - D_{KL}(q(\mathbf{z} \mid \mathbf{y}, \mathbf{g})||p(\mathbf{z})) - D_{KL}(q(l \mid \mathbf{y}, \mathbf{g})||p(l \mid \mathbf{p}))$$

Decoder: network architecture



UMAP


p
prior
sample
embedding
(prior)
g: embedding
gene
decoder
count
(decoder)
embedding
(prior)
gene
decoder
count
(generation)


UMAP
embedding
count
count (decoder)
embedding (prior)
count (generation)

Numerical Results: ursu

	perturbation	n	r2	fid
0	MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPI	10965	0.8240613246155120	0.0217970139875611
1	MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDLPSRTVDTKQAQDLATSYGIPFIETSAKTRQRVEDAFYT	982	0.8470541848581680	0.0851393905348956
2	MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDLPSRTVDTKQVQDLARSYGIPFIETSAKTRQRVEDAFYT	959	0.8298242244260930	0.0862758674351567
3	MTEYKLVVVGAGGLGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDLPSRTVDTKQAQDLARSYGIPFIETSAKTRQRVEDAFYT	957	0.8447995213222620	0.1062737460753840
4	MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGEPCLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDLPSRTVDTKQAQDLARSYGIPFIETSAKTRQRVEDAFYT	949	0.8364302108956200	0.0843765876670037
5	MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPI	947	0.8282391772762720	0.0866239887647477
6	MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPI	927	0.8233529621884760	0.0936052924820676
7	MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPI	926	0.8230960086429750	0.0922461461822123
8	MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYFPALNKMFCQLAKTCPVQLWVDSTPPI	920	0.8219451864744400	0.0842910636184122
9	MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPI	910	0.8193834322947670	0.0968988822292189
10	MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILNTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDLPSRTVDTKQAQDLARSYGIPFIETSAKTRQRVEDAFYT	902	0.8333345926365980	0.090503827106791
11	MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDRTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDLPSRTVDTKQAQDLARSYGIPFIETSAKTRQRVEDAFYT	856	0.8425824663693360	0.1003559521949740
12	MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNRCDLPSRTVDTKQAQDLARSYGIPFIETSAKTRQRVEDAFYT	829	0.846976188623999	0.1046372636504840
13	MTEYKLVVVGACGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDLPSRTVDTKQAQDLARSYGIPFIETSAKTRQRVEDAFYT	749	0.8599515098073000	0.1089276415027170
14	MTEYKLVVVGAGVVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDLPSRTVDTKQAQDLARSYGIPFIETSAKTRQRVEDAFYT	610	0.8371254801148790	0.1479446639233860
15	MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPI	505	0.8247628303550330	0.1339615763610420
16	MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPI	471	0.825372709966380	0.15550871729122
17	MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDRGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPI	323	0.8143845497155460	0.2372747957733520

Numerical Results: sciPlex

	perturbation	n	r2	fid
0	Cc1nc(-c2cccnc2)sc1C(=O)Nc1ccccc1-c1cn2c(CN3CCOCC3)csc2n1	5258	0.8778954365472620	0.39956315756656100
1	Cc1nc(C(=O)NCC(=O)O)c(O)c2ccc(Oc3ccccc3)cc12	4813	0.8443440317928	0.33967869890591400
2	CS(=O)(=O)c1ccc(-c2cnc(NCc3ccco3)n3cnnc23)cc1	4782	0.8772663311398970	0.326446751589792
3	CS(=O)(=O)OCCCCOS(C)(=O)=O	3686	0.8639891928635900	0.6194729404095090
4	Cl.Cl.c1ccc([C@@H]2C[C@H]2NC2CCNCC2)cc1	3686	0.8227544911137470	0.46706499185105300
5	CCS(=O)(=O)N1CCN(c2ccc(Nc3ncc(C(N)=O)c(NC4CC4)n3)cc2)CC1.Cl	3607	0.8675021464795920	0.3889746133996470
6	CN(C)CC(=O)Nc1ccc2[nH]c(=O)c3ccccc3c2c1	3572	0.8681040668059740	0.3501699848094620
7	CCC(=O)NCCC1CCc2ccc3c(c21)CCO3	3421	0.8725520787859180	0.36847868694635400
8	CCC1(c2ccc(N)cc2)CCC(=O)NC1=O	3418	0.8605047604251390	0.390015868204145
9	NC(=O)c1cccc(N)c1	3404	0.8625729360330590	0.374036605250879
10	CN(C)Cc1ccc(-c2nc3cccc4c3n2CCNC4=O)cc1	3400	0.8486179964854470	0.40691240400778400
11	Nc1ccccc1NC(=O)c1ccc(CNC(=O)OCc2cccnc2)cc1	3382	0.8209088169660530	1.8385339863228300
12	CN(C)Cc1ccc(S(=O)(=O)n2ccc(/C=C/C(=O)NO)c2)cc1	3318	0.8522944934950760	0.8068446362727140
13	O=C1CCC(N2C(=O)c3ccccc3C2=O)C(=O)N1	3232	0.8100994461875990	0.4106312014123500
14	N#C/C(=C(/N)Sc1ccc(N)cc1)c1ccccc1C(F)(F)F	3197	0.861683878812337	0.44054692340409600
15	N#CCNC(=O)c1ccc(-c2ccnc(Nc3ccc(N4CCOCC4)cc3)n2)cc1	3152	0.8797126028851270	0.5409651567665040
16	Cc1cn(-c2cc(NC(=O)c3ccc(C)c(Nc4nccc(-c5cccnc5)n4)c3)cc(C(F)(F)F)c2)cn1	3085	0.8526704944681460	0.719561575106713
17	CCN(CC)C(=S)SSC(=S)N(CC)CC	3084	0.8709776004306970	0.7974253250046640
18	O=C(O)[C@]1(Cc2cccc(Nc3nccs3)n2)CC[C@@H](Oc2cccc(Cl)c2F)CC1	3002	0.8860175380787680	0.4659077008894740
19	CC(C)[C@H](C(=O)Nc1ccc(C(=O)NO)cc1)c1ccccc1	2899	0.821994647220145	1.401349249891150
20	CN(C)CCC(CSc1ccccc1)Nc1ccc(S(=O)(=O)NC(=O)c2ccc(N3CCN(Cc4ccccc4-c4ccc(Cl)cc4)CC3)cc2)cc1[N+](=O)[O-]	2780	0.850507120353192	0.4921880603276240
21	N#Cc1ccnc(N2C(=O)CCC2C(=O)N(c2cncc(F)c2)C(C(=O)NC2CC(F)(F)C2)c2ccccc2Cl)c1	2495	0.8082046029733490	0.48648821683033800
22	CCS(=O)(=O)c1cccc(-c2cc(C(=O)NC3CCN(C)CC3)c(C)c3[nH]c4ncc(C)cc4c23)c1	2068	0.8628973402413010	1.2230369796520200

Numerical Results: GI

	perturbation	n	r2	fid
0	CEBPE/RUNX1T1	1198	0.9359156405053340	3.0646878155877100
1	SLC4A1/ctrl	951	0.9352522534826840	1.3929921241964200
2	ctrl/CEBPE	646	0.8784601750335980	4.226380132405520
3	LYL1/IER5L	619	0.9521373322342070	1.1865761040286100
4	BCL2L11/ctrl	575	0.9664771093759370	1.9936500503902300
5	UBASH3B/ctrl	570	0.9632920694650630	1.2598786499723300
6	S1PR2/ctrl	534	0.9423886534204050	2.8867937596243900
7	BCORL1/ctrl	526	0.9696114450565660	0.8821857700291280
8	ARRDC3/ctrl	490	0.9729711543433860	1.054288025538860
9	MAPK1/PRTG	488	0.9472136560147830	2.2308200355881100
10	FOXA3/ctrl	485	0.9722587330554140	0.7803472699061110
11	BCL2L11/TGFBR2	459	0.9707365900394600	4.2819532778329500
12	UBASH3A/ctrl	446	0.9707414748915260	1.150457806319300
13	ETS2/ctrl	444	0.936040757963672	3.042955517679860
14	IRF1/ctrl	419	0.8721488152427300	6.997562026066630
15	ATL1/ctrl	375	0.9422556603260030	3.0677738499947400
16	TMSB4X/BAK1	323	0.9679648418721080	1.8273358785956300
17	PRDM1/ctrl	294	0.9139424188441070	1.7632640411280600
18	FEV/MAP7D1	274	0.9107341690447040	5.405426049332470
19	TGFBR2/PRTG	262	0.9306394067612850	4.125999876830890
20	FOSB/IKZF3	244	0.9461364745250710	2.7425237461094900
21	CEBPE/CNN1	232	0.9348854092200420	2.2623404439187600
22	ctrl/SAMD1	211	0.9666237844958320	1.1193794960848900
23	ctrl/PTPN12	210	0.9316497796358580	1.7151786532053200
24	CEBPB/OSR2	198	0.8905173541082190	5.84315770763561
25	CBL/TGFBR2	186	0.9339413595670810	4.523567817421180
26	CEBPE/CEBPA	176	0.8140740331986840	4.844825610755920
27	FOSB/CEBPE	150	0.839744190864698	5.017764144882830
28	IGDCC3/ZBTB25	141	0.9077975968728450	3.2010085683661700
29	CDKN1B/CDKN1A	128	0.9507933360418040	2.8377133899292900
30	ZBTB10/SNAI1	97	0.8694796945973190	5.209488412586010
31	KIF18B/KIF2C	93	0.9667886900503080	1.2919243301839900
32	FOXL2/HOXB9	92	0.915615412029077	2.550168684307200