Baler analysis

Gene expression dataset

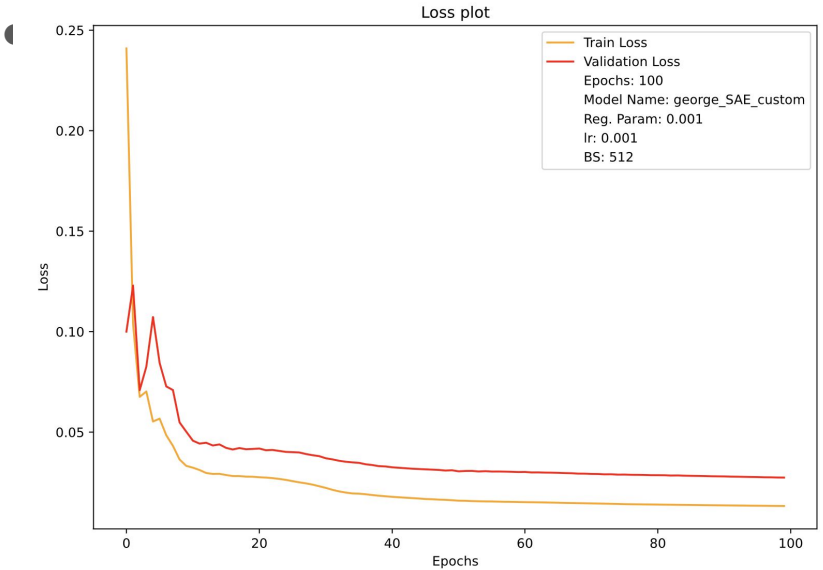
Data

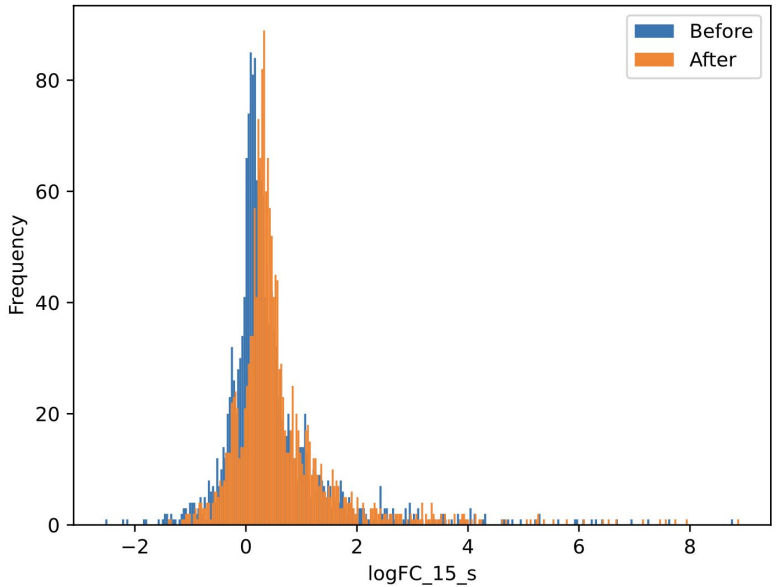
- Data obtained from github.com/sara-venkatraman/Bayesian-Gene-Dynamics
- Gene expression dataset - 1736 genes * 18 time points

	logFC_1_s	logFC_2_s	logFC_3_s	logFC_5_s	logFC_6_s
28SrRNA-Psi:CR45853	0	-0.720451849432708	-0.661040947385393	-0.60567784290551	1.43509786930533
AANATL2	0	-0.210578854070007	-0.338596705612285	-0.561582298011931	-0.438849931569944
Aatf	0	0.156728178765932	0.192964471460985	0.442685566154213	0.533390767767294
aay	0	1.35176128209764	1.2649155789955	1.07729869615486	0.591920784803151
ACC	0	0.190123680601836	-0.352465705243686	-0.892768241206817	-1.08485050598154
AcCoAS	0	0.219786313160162	-0.202753223892627	-1.15746316920644	-1.38274284761186
Achl	0	-1.27030859344197	-1.45862786899483	-0.557136295101593	-0.816807862055875
Acn	0	-0.0706638664544972	-0.0134962233397653	0.0210467758061252	-0.067349848004417
Acon	0	-0.0482951553135909	-0.0594588910829152	-0.199721890950762	-0.201450117096522
Acp1	0	-0.840462132385083	-1.61375384093499	-1.57527909367271	-0.652934724502326
Act42A	0	0.096426243500904	0.307881083156005	0.193475872487467	0.0115930995883042
Act57B	0	-0.169575537681476	-0.19548871852084	-0.317005863213994	-0.281675291248552
Act5C	0	0.117663698105465	0.319746558833284	0.0217693107921093	0.163718723119564
Act79B	0	-0.485784124588861	-0.50323019950044	-0.508261848153372	-0.361785287393477

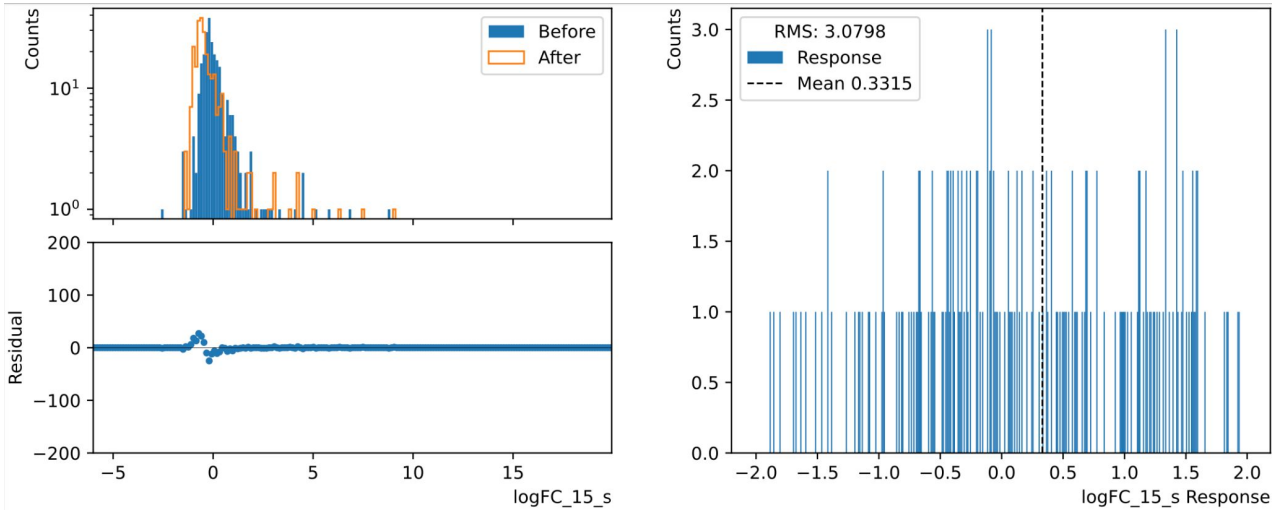
Results

- Min-Max normalization, nan values dropped
- Same `george_SAE_custom` model, training process as HEP data, 2x compression





Results (contd.)



- While it is evident that the model learns the structure of the data to some extent, the reconstruction is still not as good as the HEP data case. This is because of the much smaller size of the dataset and the fact that there has been no hyperparameter search for the model architecture, loss functions and preprocessing techniques.

Discussion

- Using this dataset with the Baler framework was quite simple and required appropriate modifications in the config, preprocessing and analysis files.
- As in the HEP case, improvement in performance would involve systematically studying modifications in the architecture, loss functions and training strategy.
- Nowadays, as models are trained on extremely large datasets (e.g. LLMs like GPT/LLaMA, stable diffusion models), storage can quickly become a limiting factor and be a bottleneck even with ample memory and processing power. Studying lossy compression techniques like autoencoders is an important direction in attempting to remedy this.