2.5.3 Classification within feature in Latent Space

It is also important to understand whether the features are disentangled well not only with each other but within itself. A dimension in the latent space that disentangles feature y from others, should also disentangle the different values of y well. For example, if a dimension that disentangles cell type well from other features, should also be able to distinguish different cell types from each other within that dimension. Therefore a second level disentanglement score was created that quantified this aspect of the model: For a particular feature, y a batch of $\mathbf{B}$ vectors z_m^b for every m value of y , to be fed as inputs to a linear classifier was created as follows:

1. Choose a feature value m, for example CD14 Mono in cell type.

2. For L samples in a batch:

   (a) 2 samples, x_1^l and x_2^l from feature space were picked with the value m of the feature y to be the same.

   (b) They were mapped to the latent space using the encoder to get z_1^l, z_2^l.

   (c) The L1 difference between the two points in latent space was calculated: $z_m^l = |z_1^l - z_2^l|$

3. The average L1 norm for the L samples is calculated: $z_m^b = \frac{1}{L}\sum_l z_m^l$. Steps 1-3 were repeated for every value m of y.

Therefore z_m^b is used as an input to predict the fixed feature value m of y. The classifier's goal is to predict $p(m|z_m^b)$. The accuracy of this predictor is used as the disentanglement metric score. Multiple dimensions could represent a disentangled view of the same feature y. To understand which one disentangles better, the classifier for these scores was also trained dimension-wise.

3 Results

3.1 Comparing Model Architectures for Disentanglement

For all the versions of the model, three loss components were looked at: KL loss, Reconstruction Loss and total VAE loss. For VAE with dHSIC, the value of the kernel is also compared. Additionally the disentanglement metric is also accounted for in deciding the relevant latent spaces.

3.1.1 β-VAE Model Comparison

β-VAE without C: Loss Components

It can be observed that for both the datsets, the reconstruction and VAE loss increases with the value of *beta* in the β-VAE model. However, as *beta* increases, the KL loss is penalized more and is pushed towards 0 (Figures: 1,2). With KL loss becoming 0, the approximated posterior is more and more similar to the actual posterior, which was assumed to be a factorised unit Gaussian distribution. However, the reconstruction loss, which is also dependent on the approximated posterior becomes poorer.

Figure 1: β-VAE Loss Components for Dentate Gyrus dataset

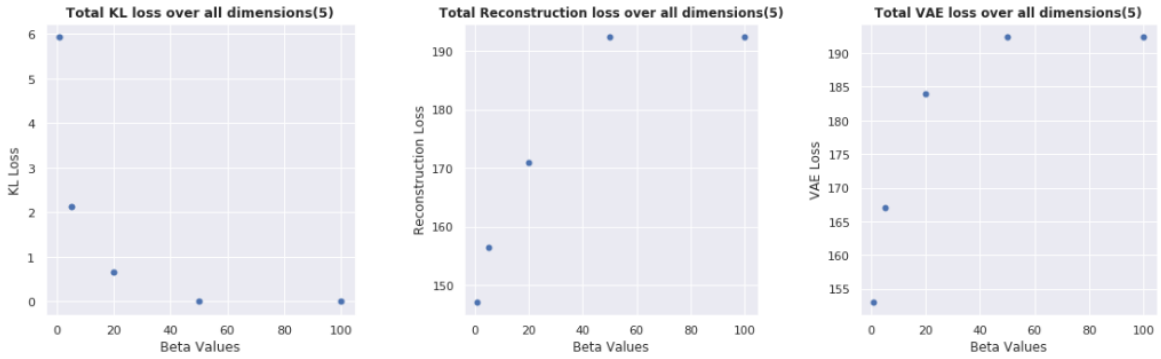

Figure 2: β-VAE Loss Components for Kang dataset

β-VAE without C: Disentanglement Metric

In the β-VAE model, for both the datasets, with increase in β, the average and the maximum Disentanglement Scores reduce (Figures: 3,4).

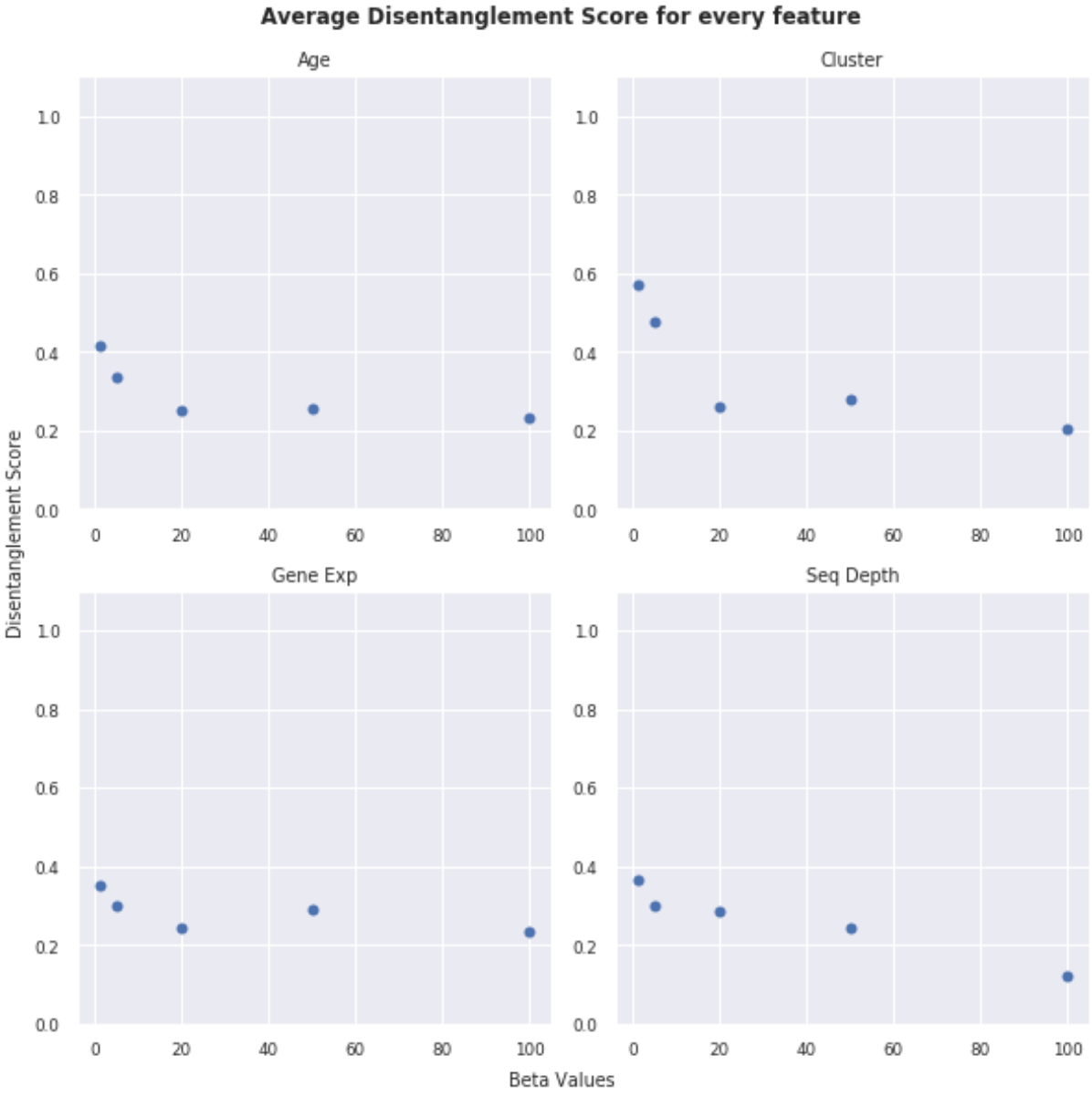

Figure 3: β-VAE Average Disentanglement score for every feature for Dentate Gyrus dataset

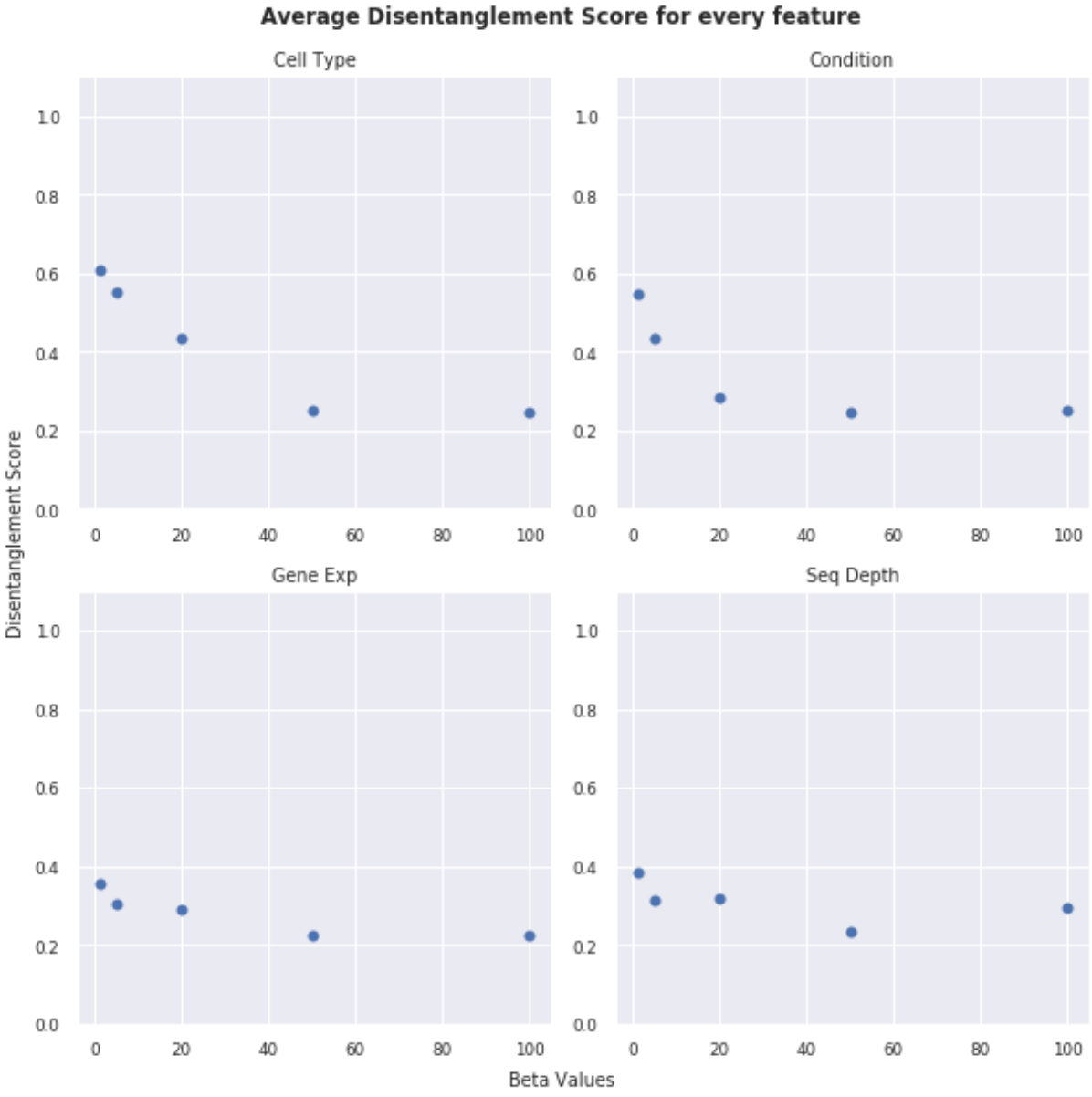

Figure 4: β-VAE Average Disentanglement score for every feature for Kang dataset

β-VAE with C Dentate Gyrus: Loss Components

For the extended β-VAE model with 'controlling' capacity C, different models with various parameters were trained. The models were trained for 5 dimensions in the latent space. This was decided keeping in mind the number of features to disentangle. Models with ¿5 dimensions in the latent space do not encode more information or improve disentanglement. Loss components are dependent on the choice of β and C. Recalling equation 40, to minimize the loss function, the KL loss should be close to the value of C. Therefore, the KL loss does not change with the increasing value of β. It however, changes only with the value of C (Figures: 5). Additionally, the Reconstruction loss,

increases with β (Figures: 6). The relation with C is not very distinct, although there is a slight drop as C increases. Lastly, the total VAE loss is also increasing with β (Figures: 7). For smaller values of C ($\leq$50), even with a high β, the VAE loss does not change a lot. However, for higher values of C, the loss increases drastically.

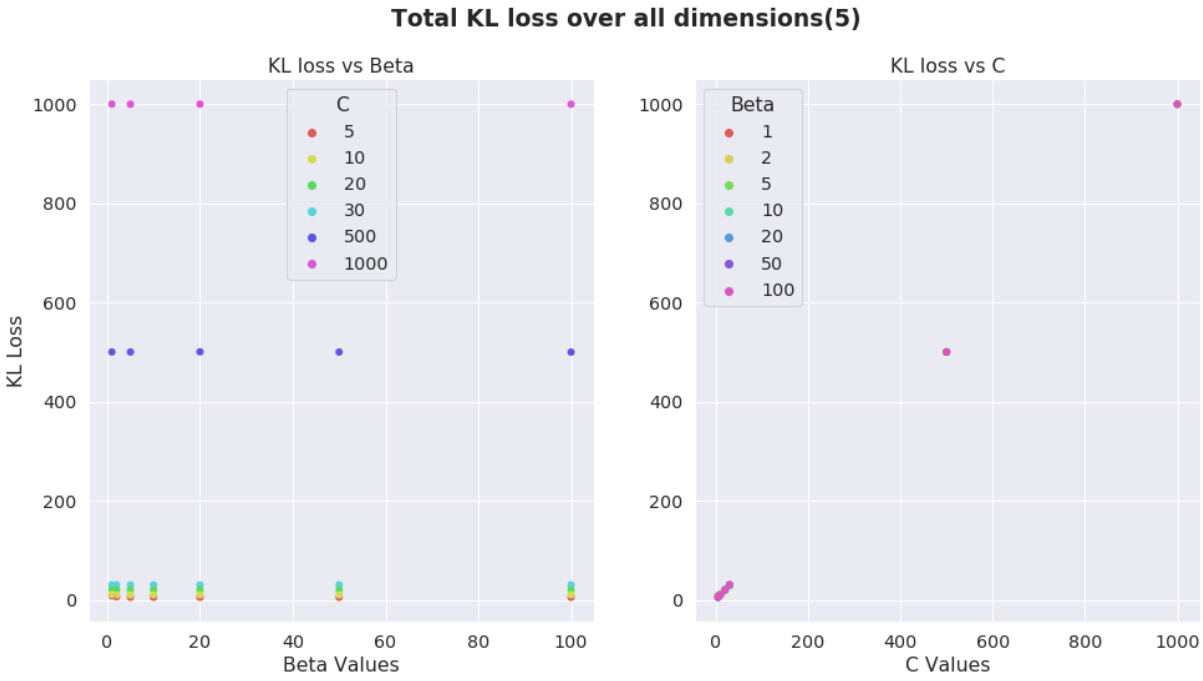

Figure 5: KL Loss for Dentate Gyrus dataset in the β-VAE with C model. The latent space units in this figure are 5

Figure 6: Reconstruction Loss for Dentate Gyrus dataset in the β-VAE with C model. The latent space units in this figure are 5

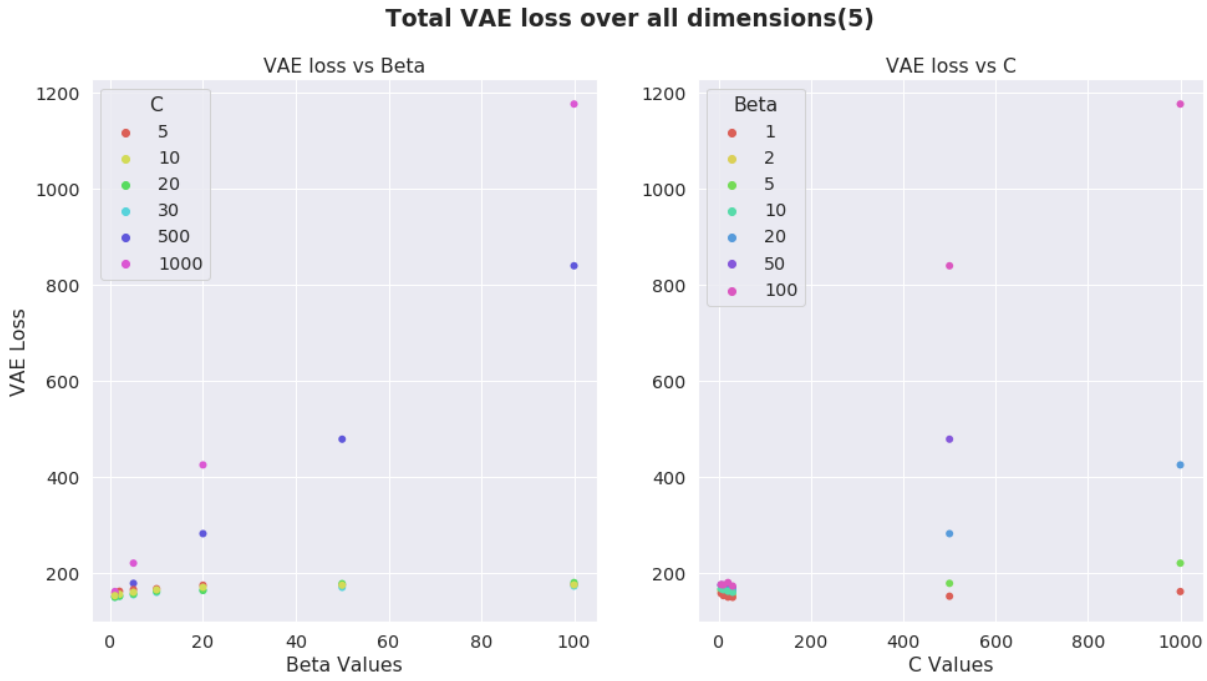

Figure 7: VAE Loss for Dentate Gyrus dataset in the β-VAE with C model. The latent space units in this figure are 5

β-VAE with C Dentate Gyrus: Disentanglement Metrics

In the extended β-VAE model with C, the disentanglement scores varied with both the values of β and C (Figures: 8). The scores for **Seq Depth** and **Gene Exp** are low ($\sim$0.5) and do not vary with β or C. The average score for **Age(days)** is low, however the maximum score is high for a few models. Again, no clear trend can be determined. It can be seen, that both maximum and average scores for the **Cluster** feature is more than all other features. This means that this feature is being disentangled well. The maximum scores increase with β. The trend with C is not strong although it can be seen that it decreases slightly (Figure: 9).

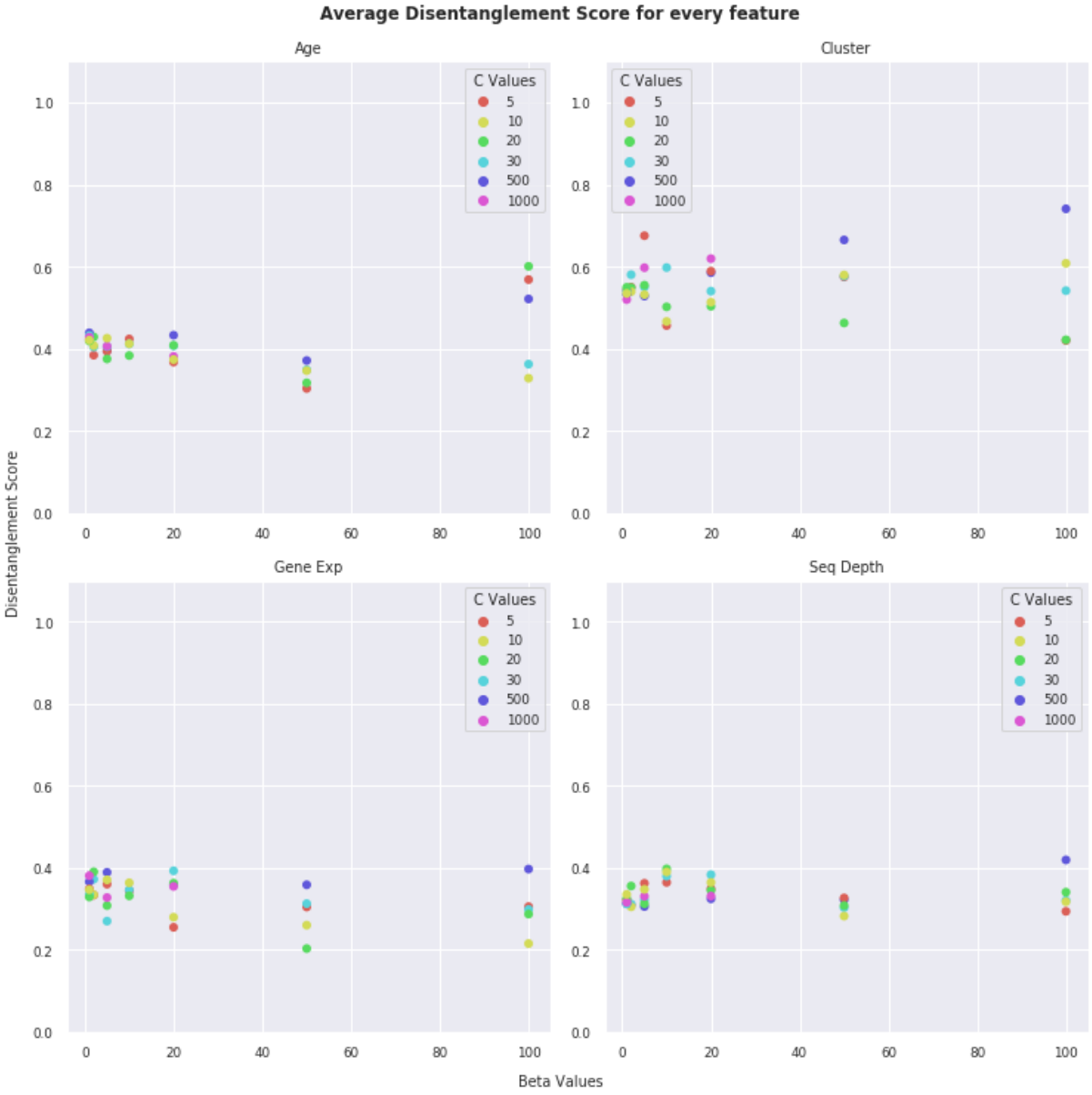

Figure 8: Average Disentanglement score for every feature for Dentate Gyrus dataset in the β-VAE with C model. The latent space units in this figure are 5.

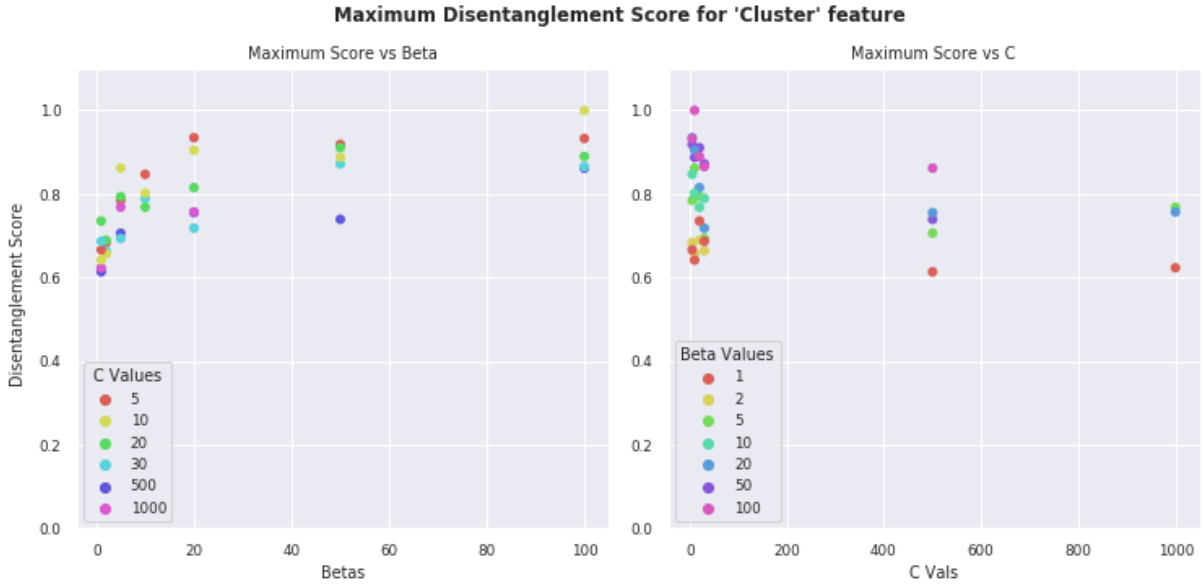

Figure 9: Maximum Disentanglement score for 'Cluster' feature for Dentate Gyrus dataset in the β-VAE with C model. The latent space units in this figure are 5.

β-VAE with C Kang: Loss Components

For the Kang dataset, 5 dimensional latent space models were trained. The results for the kang dataset are similar to that of the Dentate Gyrus set when comparing loss components. As before, they are dependent on the choice of β and C. The KL loss does not change with the increasing value of β. It however, changes only with the value of C (Figures: 10). The Reconstruction loss and VAE loss, increase with β (Figures: 11,12). The relation with C is not very distinct, although there is a slight drop as C increases. For smaller values of C ($\leq$50), even with a high β, the VAE loss does not change a lot. However, for higher values of C, the loss increases drastically.

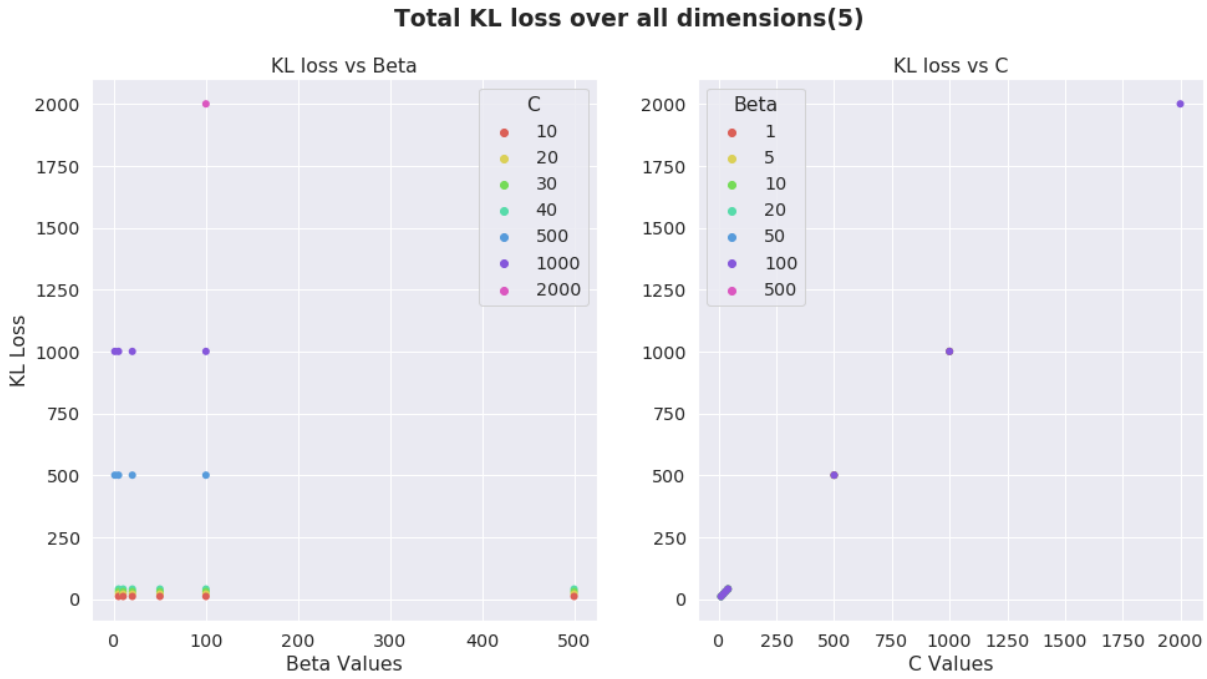

Figure 10: KL Loss for Kang dataset in the β-VAE with C model. The latent space units in this figure are 5

Figure 11: Reconstruction Loss for Kang dataset in the β-VAE with C model. The latent space units in this figure are 5

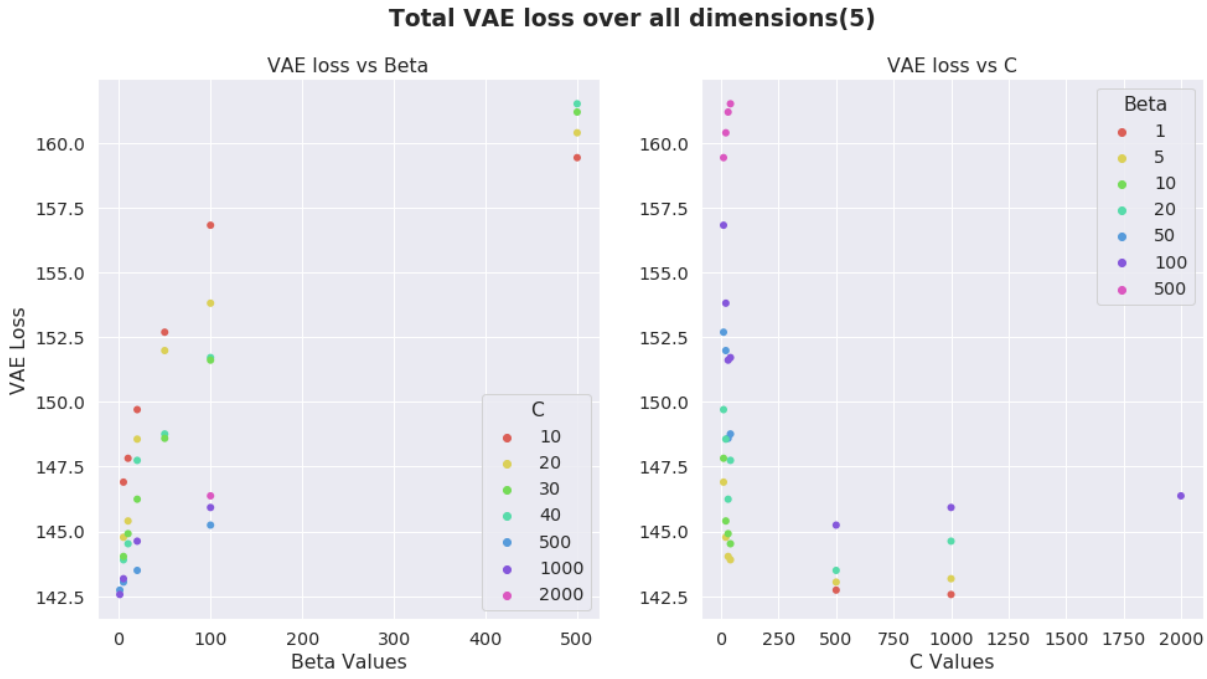

Figure 12: VAE Loss for Kang dataset in the β-VAE with C model. The latent space units in this figure are 5

β-VAE with C Kang: Disentanglement Metrics

The two features that are represented well are **condition** and **cell type**. Both averages and maximum scores for these features decrease with β. The change is more drastic for **condition**. Scores for **Gene Exp** and **Seq Depth** are generally low with some minor exceptions (Figures: 13).

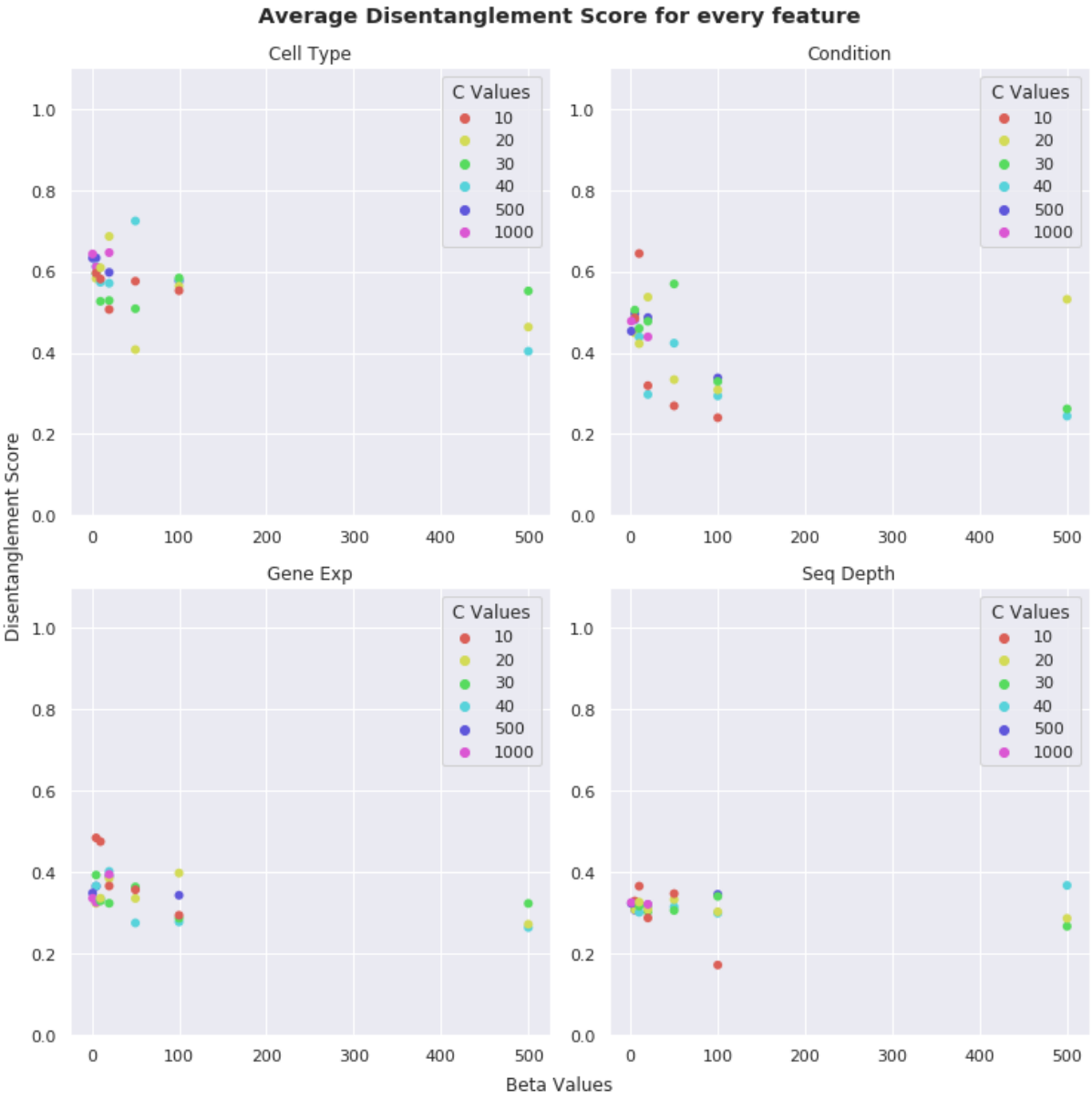

Figure 13: Average Disentanglement score for every feature for Kang dataset in the β-VAE with C model. The latent space units in this figure are 5.

For maximum scores in **condition** and **cell type**, the relation between β and C can be understood from Figures: 14,15. All scores decrease with increase in β and slightly increase with C.

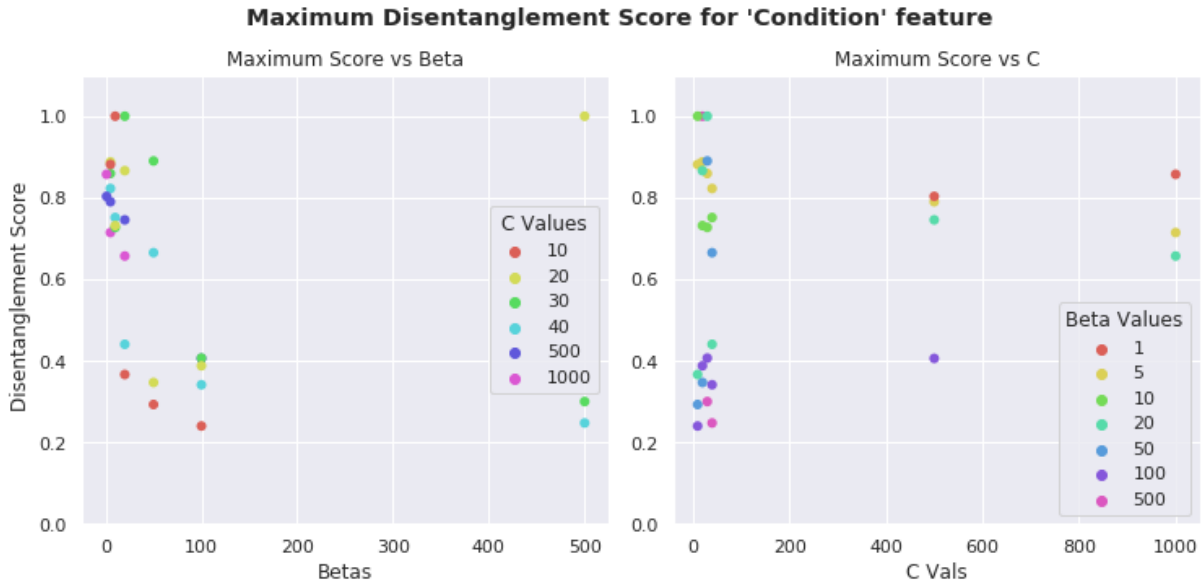

Figure 14: Maximum Disentanglement score for 'Condition' feature for Kang dataset in the β-VAE with C model. The latent space units in this figure are 5.

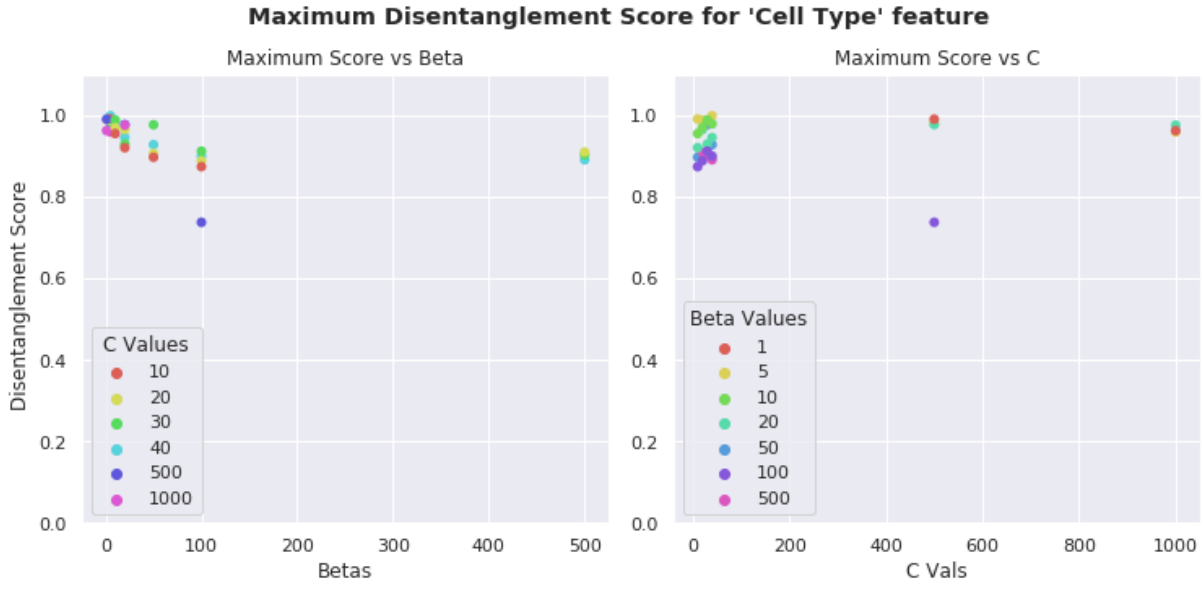

Figure 15: Maximum Disentanglement score for 'Cell Type' feature for Kang dataset in the β-VAE with C model. The latent space units in this figure are 5.

3.1.2 dHSIC Model Comparison

dHSIC Dentate Gyrus: Loss Components

Now, the last model that was used to rain on the datasets is the dHSIC model (equation

44. The two parameters in this model are β and γ. In Dentate Gyrus dataset, the KL loss naturally decreases with β as in the β-VAE model (Figures: 16). The reconstruction and VAE loss increases with β and γ. Although strong trends are not visible (Figures: 17,18). The last addition to this model, the dHSIC kernel value, decreases with increase in β (Figure: 19). This could be because as β increases, the approximated posterior value is pushed towards a factorised unit Gaussian distribution which also induces independence between the dimensions. With increase in γ, the kernel value drops initially. For very high values of γ, the kernel value increases. This could be because after a certain threshold, changes in the value of the kernel will not reduce the total vae loss. Therefore, the optimal value is achieved by focussing on reconstruction loss.

Figure 16: KL Loss for Dentate Gyrus dataset in the dHSIC model. The latent space units in this figure are 5

Figure 17: Reconstruction Loss for Dentate Gyrus dataset in the dHSIC model. The latent space units in this figure are 5

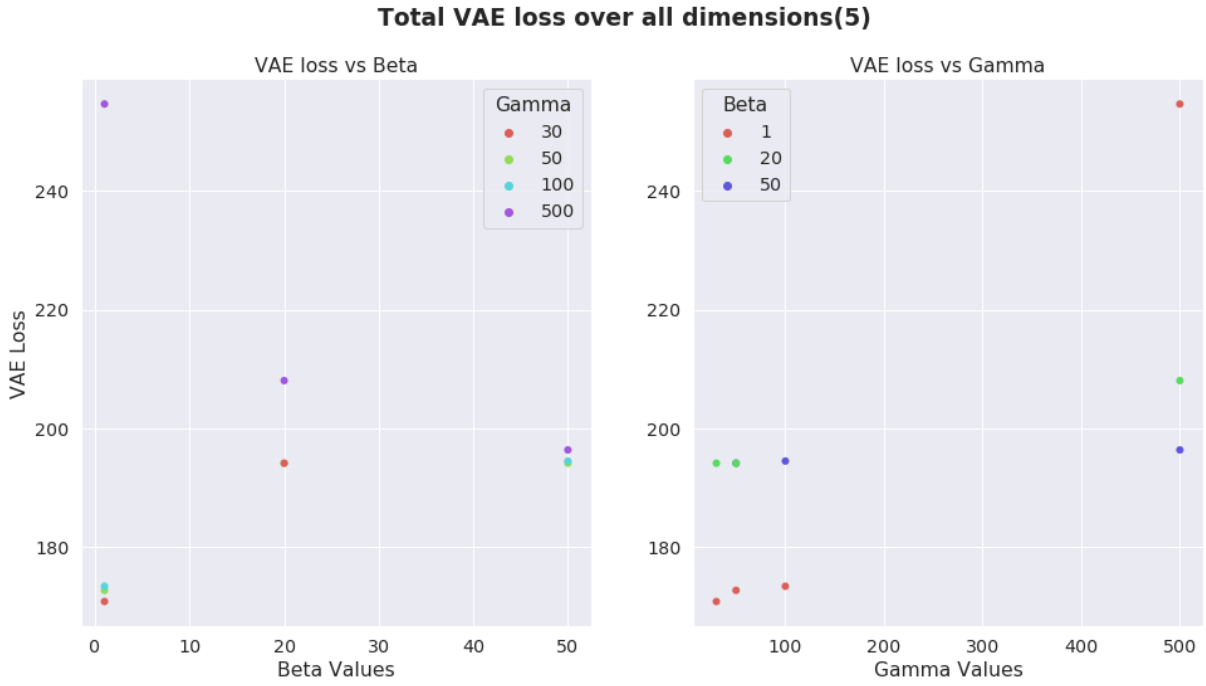

Figure 18: VAE Loss for Dentate Gyrus dataset in the dHSIC model. The latent space units in this figure are 5

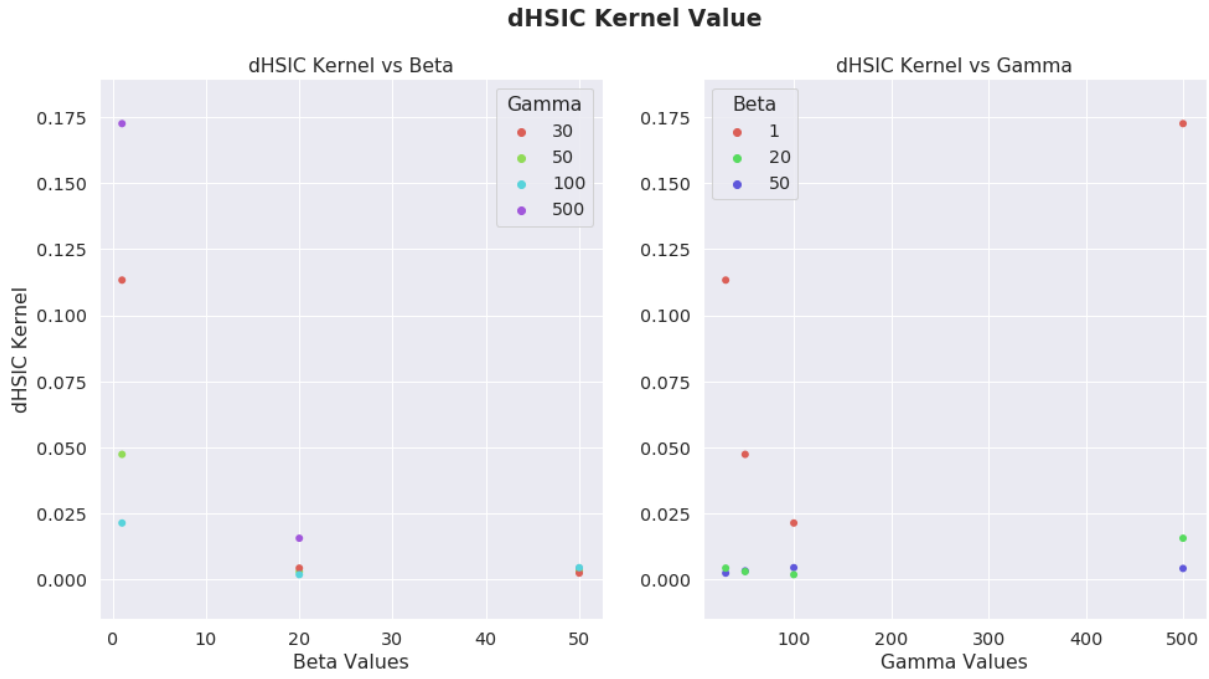

Figure 19: dHSIC kernel value for Dentate Gyrus dataset in the dHSIC model. The latent space units in this figure are 5

dHSIC Dentate Gyrus: Disentanglement Metrics

In the dHSIC model, the average scores for every feature decreases, with increase in β (Figures: 20). For the maximum disentanglement score, the values decrease too with β. The trend with γ is not very strong (Figures: 21).The disentanglement scores for **Seq Depth** and **Gene Exp** are low ($\sim$0.4) and do not vary with Γ.

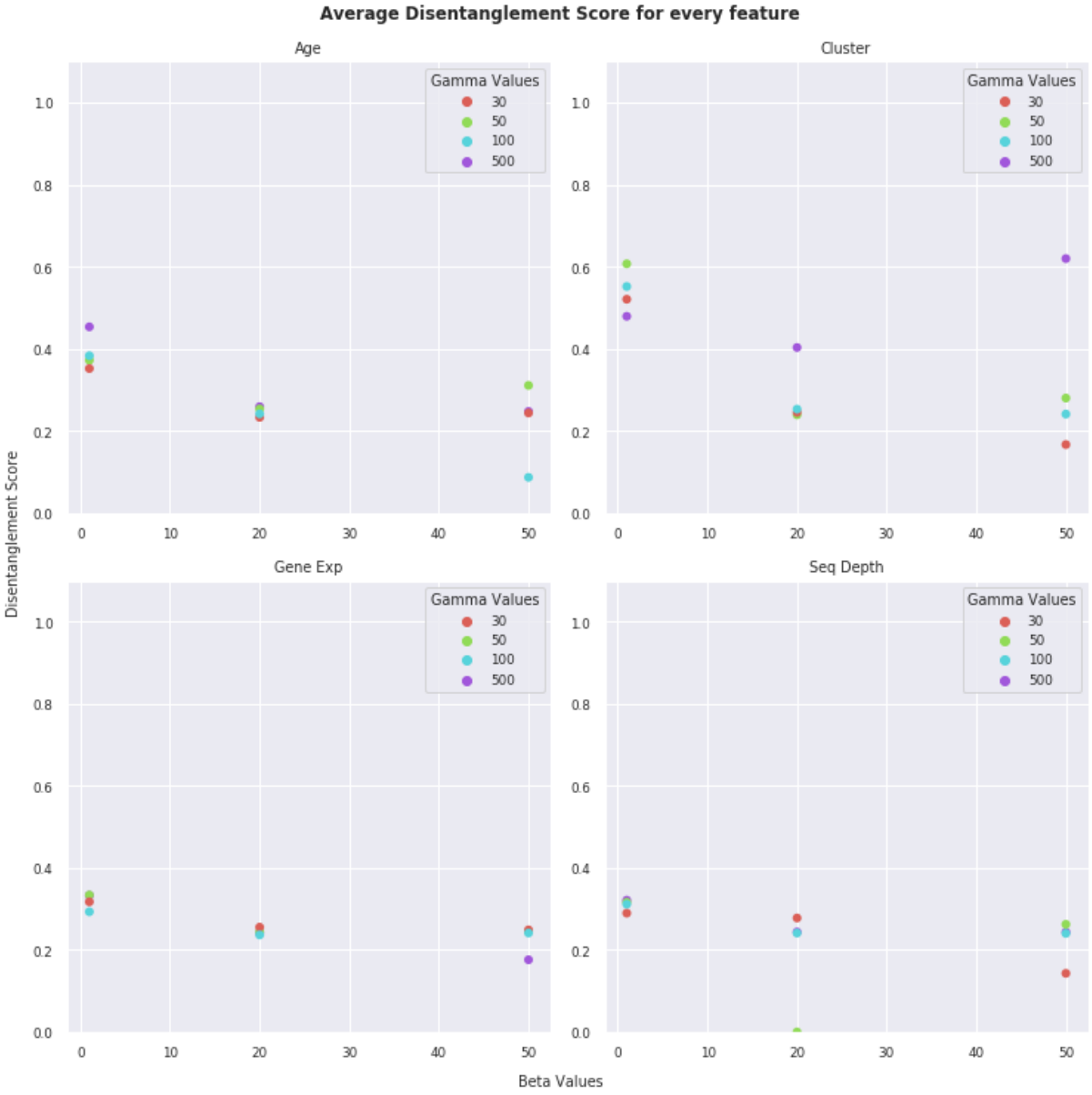

Figure 20: Average Disentanglement score for every feature for Dentate Gyrus dataset in the dHSIC model. The latent space units in this figure are 5.

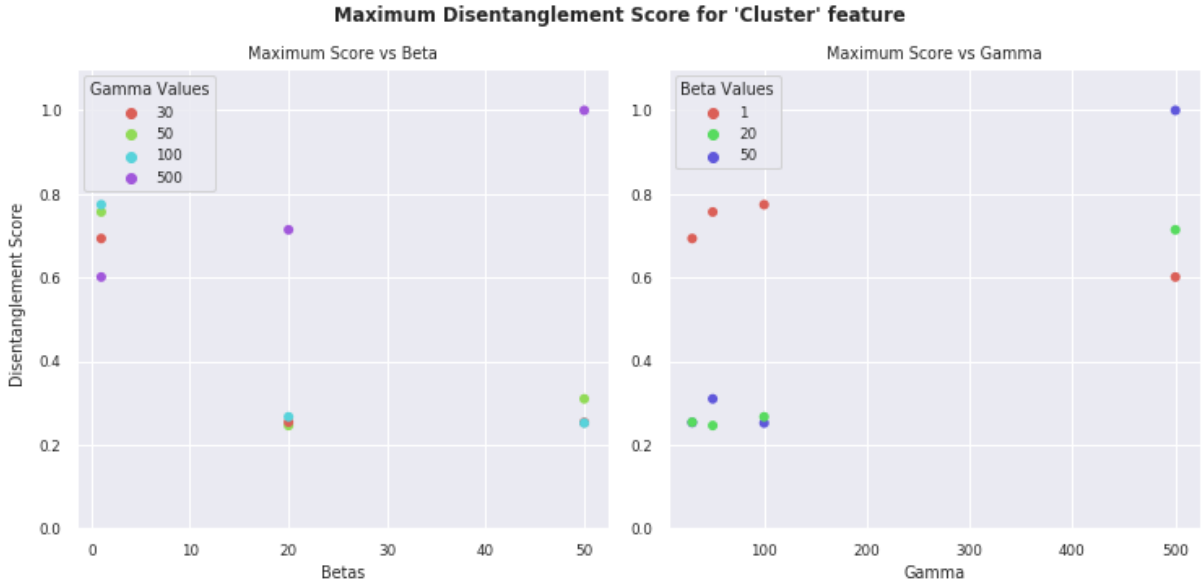

Figure 21: Maximum Disentanglement score for 'Cluster' feature for Dentate Gyrus dataset in the dHSIC model. The latent space units in this figure are 5.

dHSIC Kang: Loss Components

Similar trends as Dentate Gyrus can be seen in the Kang dataset (Figures: 22,23,24,25).

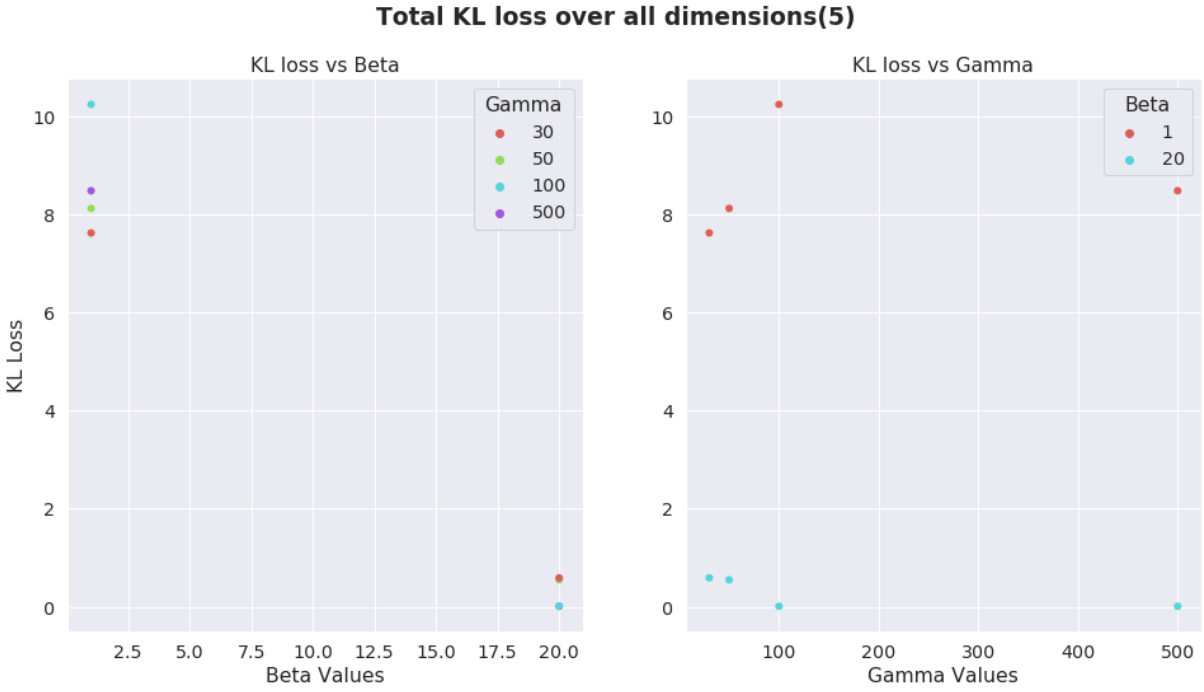

Figure 22: KL Loss for Kang dataset in the dHSIC model. The latent space units in this figure are 5

Figure 23: Reconstruction Loss for Kang dataset in the dHSIC model. The latent space units in this figure are 5

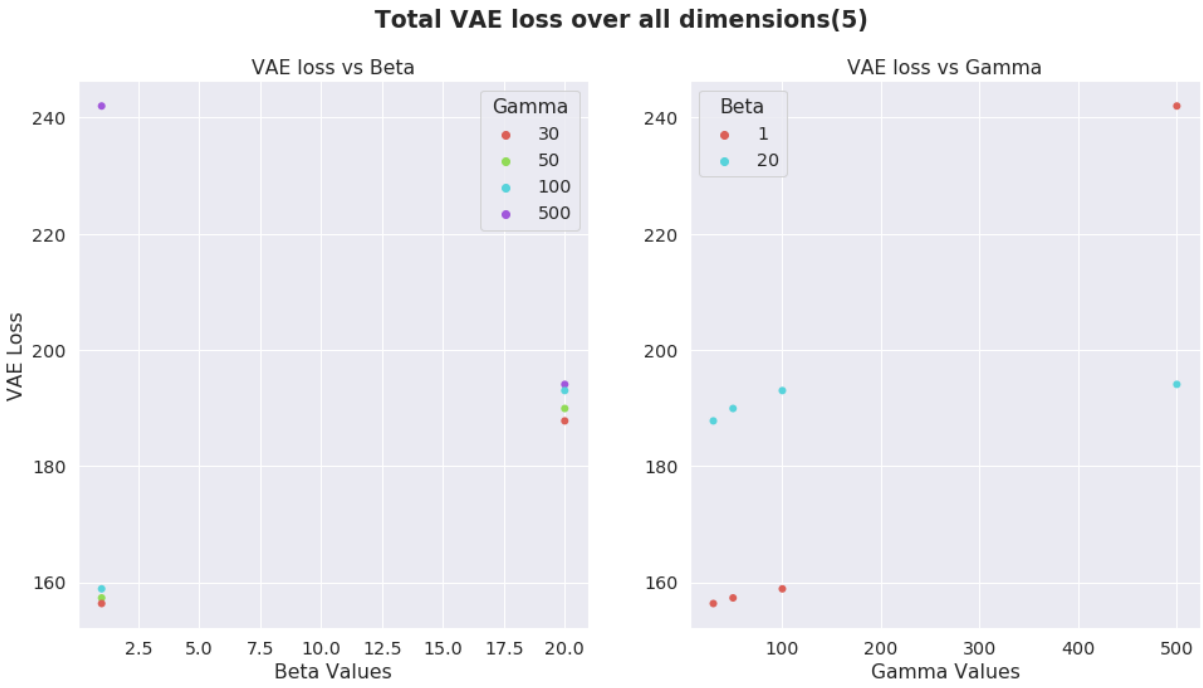

Figure 24: VAE Loss for Kang dataset in the dHSIC model. The latent space units in this figure are 5

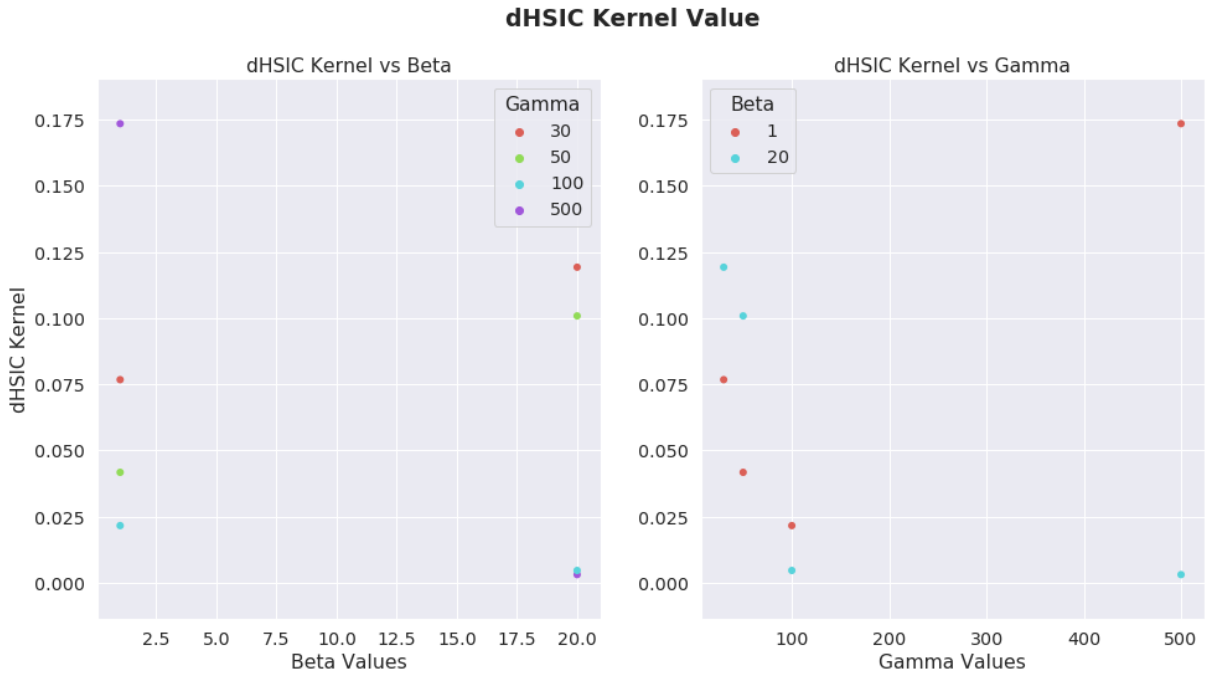

Figure 25: dHSIC kernel value for Kang dataset in the dHSIC model. The latent space units in this figure are 5

dHSIC Kang: Disentanglement Metrics

In the dHSIC model, the average scores for every feature decreases, with increase in β (Figures: 26). For the maximum disentanglement score, the values decrease too with β. The trend with γ is not very strong although a slight decrease is visible (Figures: 27,28). For both **condition** and **cell type** features, $\beta = 1$ gives the best results. Additionally, the scores increase slightly with γ. However for high values of γ, they are low. The disentanglement scores for **Seq Depth** and **Gene Exp** are low ($\sim$0.4) and do not vary with γ.

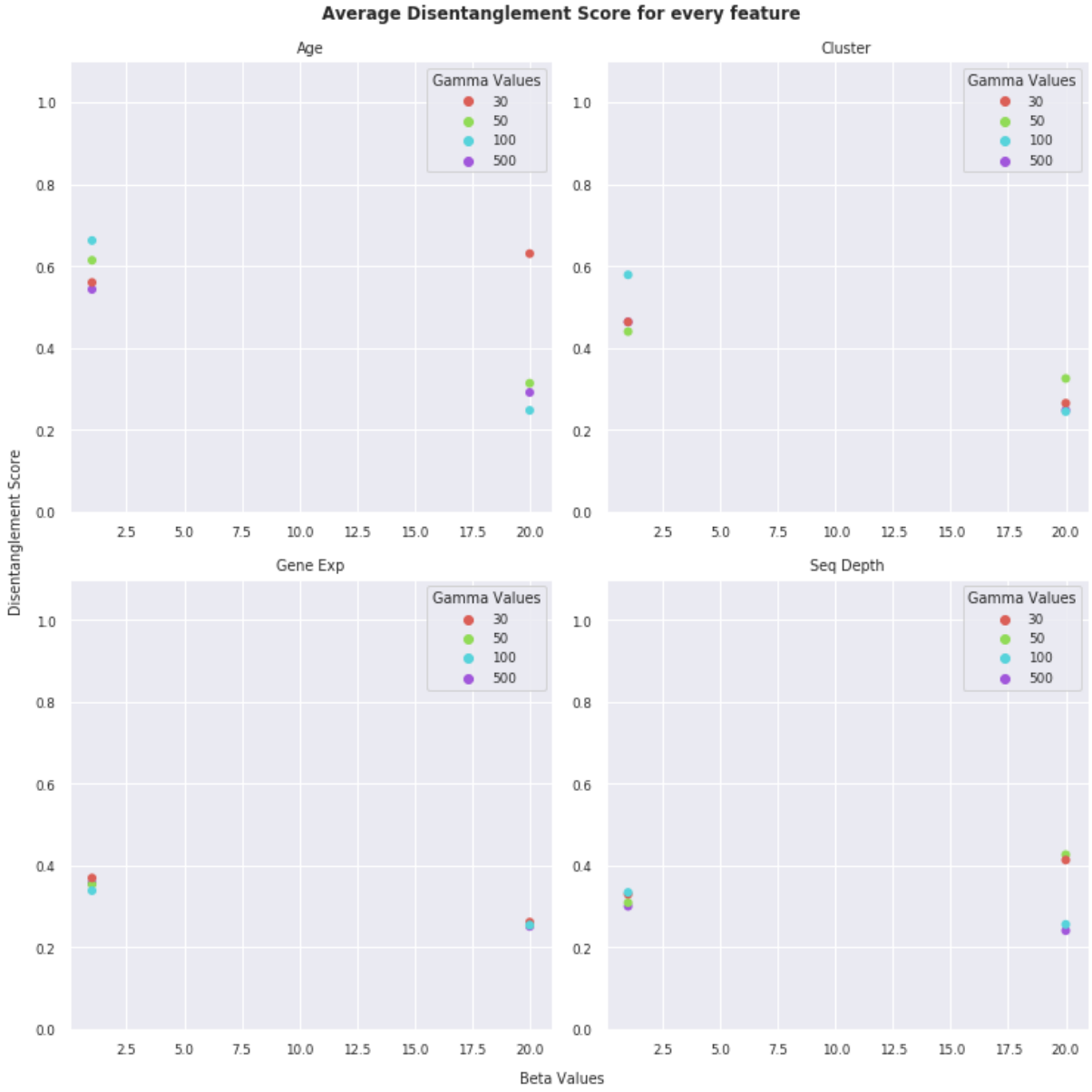

Figure 26: Average Disentanglement score for every feature for Kang dataset in the dHSIC model. The latent space units in this figure are 5.

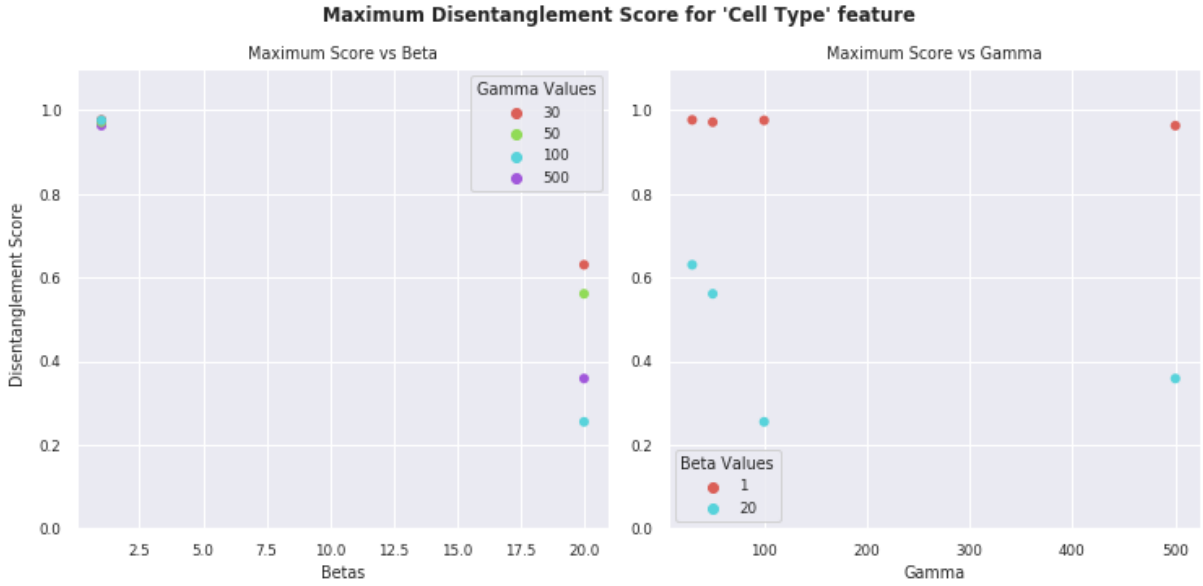

Figure 27: Maximum Disentanglement score for 'Cell Type' feature for Kang dataset in the dHSIC model. The latent space units in this figure are 5.

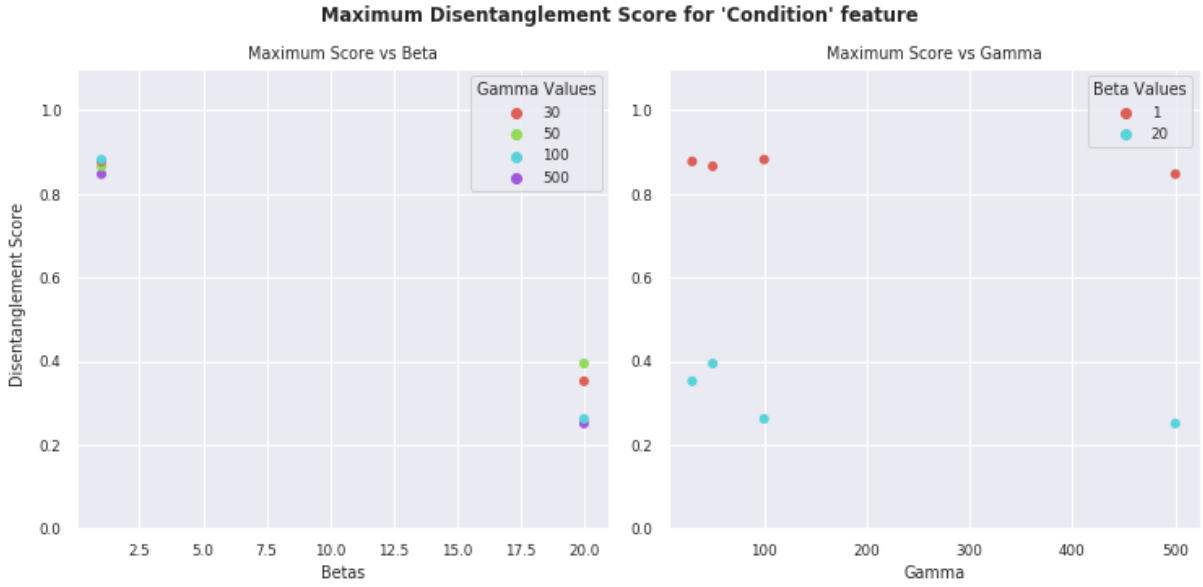

Figure 28: Maximum Disentanglement score for 'Condition' feature for Kang dataset in the dHSIC model. The latent space units in this figure are 5.

3.2 Visualizing Disentanglement

From the results in section (3.1), it was decided to to choose the following models for further analysis in the β-VAE model. This is was decided because a balance between

different loss components and disentanglement score was required. Moreover, the choice of the parameters, is a combination of high and low values of the parameters. All models had a latent space dimension of 5 as the number of features being studied are 4.

- Dentate Gyrus dataset: Alpha $\in$ [50,100] and $C \in$ [30,500]
- Kang dataset: Alpha $\in$ [5,20,100] and $C \in$ [30,500]

For the dHSIC models, the models used were:

- Dentate Gyrus dataset: Alpha $\in$ [50,100] and $\gamma \in$ [50,100]
- Kang dataset: Alpha = 1 and $\gamma \in$ [50,100]

3.2.1 Disentanglement of Features

β-VAE model with C: Dentate Gyrus

For the models mentioned above, the disentanglement scores for every feature for every dimension gives a larger overview of which dimension is associated with which feature. In Table (1), the disentanglement scores for all except **cluster** is poor. For **cluster**, only dimension 5 promises disentanglement ($\geq$0.7).

dimension	age(days)	cluster	exp_gene	seq_depth
1	0.00	0.23	0.00	0.22
2	0.38	0.63	0.00	0.30
3	0.25	0.00	0.00	0.25
4	0.24	0.00	0.25	0.00
5	0.53	0.87	0.37	0.43

Table 1: Disentanglement Scores for features in Dentate Gyrus dataset. The model used had specifications of β=50 and C=30.

In Table (2), by increasing the value of C there are now 2 dimensions (2 and 5) that promise disentanglement ($\geq$0.7).

dimension	age(days)	cluster	exp_gene	seq_depth
1	0.33	0.64	0.41	0.33
2	0.36	0.74	0.33	0.33

Continued on next page

dimension	age(days)	cluster	exp_gene	seq_depth
3	0.37	0.64	0.00	0.33
4	0.35	0.59	0.00	0.29
5	0.45	0.73	0.34	0.33

Table 2: Disentanglement Scores for features in Dentate Gyrus dataset. The model used had specifications of β=50 and C=500.

In Table (3), by increasing the value of β but keeping C low, the disentanglement does not improve compared to β=50. Only dimension 2 promises disentanglement ($\geq$0.7).

dimension	age(days)	cluster	exp_gene	seq_depth
1	0.00	0.24	0.23	0.24
2	0.51	0.87	0.37	0.46
3	0.24	0.00	0.36	0.00
4	0.44	0.52	0.26	0.26
5	0.26	0.00	0.27	0.00

Table 3: Disentanglement Scores for features in Dentate Gyrus dataset. The model used had specifications of β=100 and C=30.

In Table (4), higher β and C improve disentanglement score. More dimensions have a score of $\geq$0.7. However, at the same time, the features are less disentangled from each other. For example, in simension 4, both **cluster** and **age(days)** have a good score. This could denote that the disentanglement from each other is not strong.

dimension	age(days)	cluster	exp_gene	seq_depth
1	0.40	0.67	0.00	0.32
2	0.35	0.73	0.00	0.34
3	0.47	0.74	0.32	0.41
4	0.71	0.71	0.43	0.47
5	0.67	0.86	0.44	0.56

Continued on next page

dimension	age(days)	cluster	exp_gene	seq_depth

Table 4: Disentanglement Scores for features in Dentate Gyrus dataset. The model used had specifications of β=100 and C=500.

β-VAE model with C: Kang

The representation and scores for **condition** feature reduces with increase in β and there is no strong relation with C. This can be seen from the Tables 5,6,7,8,9. For the **cell type** feature, there are no significant changes with C, however with increase in β more dimensions encode disentanglement for **cell type**.

dimension	cell_type	condition	exp_gene	seq_depth
1	0.41	0.33	0.00	0.27
2	0.98	0.45	0.46	0.35
3	0.48	0.00	0.35	0.31
4	0.76	0.38	0.39	0.31
5	0.34	0.86	0.37	0.33

Table 5: Disentanglement Scores for features in Kang dataset. The model used had specifications of β=5 and C=30.

dimension	cell_type	condition	exp_gene	seq_depth
1	0.35	1.00	0.27	0.27
2	0.35	0.33	0.21	0.24
3	0.58	0.38	0.34	0.30
4	0.93	0.39	0.40	0.36
5	0.43	0.28	0.39	0.34

Table 6: Disentanglement Scores for features in Kang dataset. The model used had specifications of β=20 and C=30.

dimension	cell_type	condition	exp_gene	seq_depth
1	0.69	0.00	0.34	0.29
2	0.54	0.00	0.35	0.31
3	0.44	0.27	0.34	0.00
4	0.34	0.75	0.44	0.34
5	0.98	0.44	0.49	0.34

Table 7: Disentanglement Scores for features in Kang dataset. The model used had specifications of β=20 and C=500.

dimension	cell_type	condition	exp_gene	seq_depth
1	0.30	0.00	0.00	0.24
2	0.88	0.41	0.35	0.36
3	0.00	0.24	0.17	0.00
4	0.91	0.34	0.37	0.42
5	0.25	0.00	0.26	0.00

Table 8: Disentanglement Scores for features in Kang dataset. The model used had specifications of β=100 and C=30.

dimension	cell_type	condition	exp_gene	seq_depth
1	0.74	0.00	0.38	0.33
2	0.63	0.33	0.33	0.40
3	0.71	0.36	0.32	0.00
4	0.28	0.41	0.28	0.36
5	0.51	0.26	0.39	0.29

Table 9: Disentanglement Scores for features in Kang dataset. The model used had specifications of β=100 and C=500.

dHSIC models: Dentate Gyrus

From the Table (10) it can be seen that the model disentangles **cluster** well from other features in 2 and 5. Although the disentanglement in 5 is better as the score is small for

other features. By increasing γ to 100, the scores actually perform slightly poorer (Table: 11.

dimension	age(days)	cluster	exp_gene	seq_depth
1	0.00	0.29	0.00	0.25
2	0.58	0.76	0.40	0.45
3	0.37	0.66	0.30	0.31
4	0.28	0.00	0.27	0.25
5	0.26	0.72	0.36	0.33

Table 10: Disentanglement Scores for features in Dentate Gyrus dataset. The dHSIC model used had specifications of β=1 and γ=50.

dimension	age(days)	cluster	exp_gene	seq_depth
1	0.39	0.77	0.00	0.33
2	0.48	0.61	0.35	0.37
3	0.43	0.60	0.29	0.30
4	0.00	0.00	0.25	0.24
5	0.24	0.22	0.00	0.00

Table 11: Disentanglement Scores for features in Dentate Gyrus dataset. The dHSIC model used had specifications of β=1 and γ=100.

dHSIC models: Kang

From the Table (12) it can be seen that the model disentangles **cell type** and **condition** features very well from each other and others too. Dimension 2 and 4 represent **cell type** and dimension 3 **condition**. Additionally the scores for **cell type** is low in dimension 3 and vice versa. In Table (13) by increasing γ the disentanglement score improves significantly for both **cell type** and **condition**. It is encoded in dimension 1 and 4 respectively.

dimension	cell_type	condition	exp_gene	seq_depth
1	0.00	0.25	0.40	0.24
2	0.78	0.38	0.29	0.30
3	0.34	0.87	0.35	0.35
4	0.97	0.44	0.40	0.34
5	0.36	0.27	0.33	0.00

Table 12: Disentanglement Scores for features in Kang dataset. The dHSIC model used had specifications of β=1 and γ=50.

dimension	cell_type	condition	exp_gene	seq_depth
1	0.98	0.47	0.41	0.36
2	0.63	0.38	0.35	0.33
3	0.00	0.00	0.24	0.00
4	0.36	0.88	0.00	0.34
5	0.68	0.00	0.35	0.31

Table 13: Disentanglement Scores for features in Kang dataset. The dHSIC model used had specifications of β=1 and γ=100.

3.2.2 Disentanglement within Features

β-VAE model with C: Dentate Gyrus

As disentanglement in Dentate Gyrus, was strong for **cluster** feature in all models, within features scoring is shown for **cluster**. The scores for other models is included in Appendix (ADD REF).

dimension	1	2	3	4	5
Astrocytes	0.14	0.67	0.00	0.00	0.93
Cajal Retzius	0.32	0.59	0.00	0.00	1.00
Cck-Tox	0.00	0.64	0.25	0.00	0.61
Endothelial	0.00	0.72	0.00	0.00	1.00

Continued on next page

dimension	1	2	3	4	5
GABA	0.04	0.84	0.00	0.00	1.00
Granule immature	0.00	0.51	0.00	0.00	0.00
Granule mature	0.04	0.57	0.09	0.00	0.51
Microglia	0.00	1.00	0.00	0.00	0.51
Mossy	0.00	0.83	0.00	0.00	1.00
Neuroblast	0.00	0.00	0.00	0.07	1.00
OL	0.00	1.00	0.16	0.00	0.92
OPC	0.00	0.68	0.07	0.00	0.98
Radial Glia-like	0.00	0.73	0.11	0.00	0.97
nIPC	0.00	0.53	0.00	0.00	0.95

Table 14: Disentanglement Scores within 'cluster' feature in Dentate Gyrus dataset. The model used had specifications of β=50 and C=30. From Table (1), dimension 5 showed strong disentanglement. It can be seen here as well. Dimension 5 disentangles within the different clusters very well. Even, dimension 2 can be seen to disentangle.

dimension	1	2	3	4	5
Astrocytes	0.93	0.96	0.00	0.43	0.85
Cajal Retzius	0.46	0.00	0.60	0.68	0.29
Cck-Tox	0.95	0.64	0.63	1.00	0.81
Endothelial	0.93	0.96	0.91	0.47	1.00
GABA	0.50	0.55	0.69	0.83	0.60
Granule immature	0.00	0.55	0.49	0.17	0.00
Granule mature	0.76	0.55	0.55	0.56	0.35
Microglia	0.98	0.97	1.00	1.00	1.00
Mossy	0.39	0.75	0.31	0.00	0.52
Neuroblast	0.98	0.97	0.70	0.80	0.85
OL	0.84	0.00	1.00	1.00	1.00
OPC	0.91	0.69	0.61	0.79	0.74
Radial Glia-like	0.94	0.46	0.46	0.55	1.00
nIPC	0.56	0.47	0.84	0.00	0.95

Continued on next page

dimension	1	2	3	4	5

Table 15: Disentanglement Scores within 'cluster' feature in Dentate Gyrus dataset. The model used had specifications of β=50 and C=500. From Table (2), dimension 2 and 5 showed strong disentanglement. It can be seen here as well. The different clusters are being disentangled from each other very well. Even, the other dimensions show some good results.

dimension	1	2	3	4	5
Astrocytes	0.00	0.96	0.00	0.62	0.00
Cajal Retzius	0.00	1.00	0.00	0.93	0.00
Cck-Tox	0.20	0.00	0.00	0.64	0.00
Endothelial	0.00	0.94	0.00	0.92	0.00
GABA	0.00	0.76	0.07	0.59	0.00
Granule immature	0.00	0.45	0.00	0.47	0.00
Granule mature	0.00	0.26	0.00	0.51	0.00
Microglia	0.17	0.70	0.00	0.63	0.00
Mossy	0.13	0.39	0.00	0.90	0.00
Neuroblast	0.00	1.00	0.00	0.43	0.07
OL	0.20	0.83	0.00	1.00	0.00
OPC	0.00	1.00	0.00	0.90	0.00
Radial Glia-like	0.00	0.81	0.00	0.71	0.00
nIPC	0.00	0.98	0.00	0.56	0.00

Table 16: Disentanglement Scores within 'cluster' feature in Dentate Gyrus dataset. The model used had specifications of β=100 and C=30. From Table (3), dimension 2 showed strong disentanglement. It can be seen here as well. The different clusters are being disentangled from each other very well. Along with this, dimension 4 also disentangles a few clusters.

dimension	1	2	3	4	5
Astrocytes	0.81	0.64	0.00	1.00	0.97
Cajal Retzius	0.75	0.84	0.92	0.76	1.00
Cck-Tox	0.75	0.97	0.98	0.85	0.82
Endothelial	0.74	0.98	1.00	0.93	0.77
GABA	0.43	0.98	1.00	1.00	0.85
Granule immature	0.75	0.97	0.80	0.98	0.98
Granule mature	0.71	1.00	0.58	0.97	0.95
Microglia	1.00	0.91	1.00	0.85	1.00
Mossy	0.83	0.64	1.00	1.00	1.00
Neuroblast	0.96	0.92	1.00	0.99	0.87
OL	0.65	1.00	0.58	0.83	0.75
OPC	0.54	0.82	0.57	0.92	0.94
Radial Glia-like	0.90	0.87	1.00	0.96	1.00
nIPC	0.63	0.76	0.80	0.60	0.96

Table 17: Disentanglement Scores within 'cluster' feature in Dentate Gyrus dataset. The model used had specifications of β=100 and C=500. All dimensions disentangle the clusters. Every dimension, disentangles one cluster type very well from others.

β-VAE model with C: Kang

For the Kang dataset, the disentanglement accuracy has to be determined as a two-fold view. The dimensions that disentangle features from each other, might not be able to differentiate the values within the feature. This is dependent on the value of the parameters that were selected. Increasing C in the model, does not necessarily improve disentanglement within features. However, it reduces the variance in the latent space significantly. The range of Latent Space values that represent the feature, reduces. It becomes more deterministic. This would be more evident in the Section (3.2.3) where the latent spaces are directly visualized and give a platform to interpret disentanglement. Tables 20,21,22,23 summarize the scores for $\beta = 20$. A similar trend is visible for β=100, which is summarized in Appendix[APPP REF]

dimension	1	2	3	4	5
B	0.18	0.67	0.34	0.93	0.27

Continued on next page

dimension	1	2	3	4	5
B Activated	0.61	0.84	0.89	0.76	0.21
CD14 Mono	0.30	0.99	0.17	0.52	0.18
CD16 Mono	0.89	0.98	0.00	0.64	0.24
CD4 Memory T	0.19	0.65	0.36	0.30	0.47
CD4 Naive T	0.53	0.51	0.58	0.88	0.17
CD8 T	0.21	0.57	0.79	0.35	0.17
DC	0.09	0.95	0.41	0.84	0.00
Mk	0.00	0.79	0.16	0.36	0.25
NK	0.00	0.43	0.53	0.30	0.32
T activated	0.71	0.41	0.13	1.00	0.00
pDC	0.50	0.94	0.58	0.63	0.54

Table 18: Disentanglement Scores within 'cell type' feature in Kang dataset. The model used had specifications of β=5 and C=30. In the Table (5) 'cell type' is disentangled from other features in dimension 2 and 4. Here, too the 2nd and 4th dimensions show disentanglement of different cell types from each other.

dimension	1	2	3	4	5
CTRL	0.62	0.60	0.95	0.74	1.0
STIM	0.67	0.61	0.87	0.70	1.0

Table 19: Disentanglement Scores within 'condition' feature in Kang dataset. The model used had specifications of β=5 and C=30. In the Table (5) 'condition' is disentangled from other features in dimension 5. Here, too the 5th dimension shows a complete disentanglement of 'CTRL' from 'STIM'.

dimension	1	2	3	4	5
B	0.24	0.51	0.81	0.59	0.42
B Activated	0.30	0.62	0.30	0.83	0.81

Continued on next page

dimension	1	2	3	4	5
CD14 Mono	0.00	0.00	0.43	0.84	0.49
CD16 Mono	0.19	0.16	0.48	0.68	0.70
CD4 Memory T	0.00	0.79	0.40	0.59	0.23
CD4 Naive T	0.50	0.22	0.41	0.51	0.37
CD8 T	0.26	0.25	0.71	0.52	0.31
DC	0.26	0.46	0.85	0.84	0.72
Mk	0.15	0.15	0.31	0.85	0.00
NK	0.68	0.18	0.76	0.50	0.22
T activated	0.55	1.00	0.81	0.62	0.64
pDC	0.14	1.00	0.55	0.68	0.19

Table 20: Disentanglement Scores within 'cell type' feature in Kang dataset. The model used had specifications of β=20 and C=30. In the Table (6) 'cell type' is disentangled from other features in dimension 4. Here, too the 4th dimension also shows disentanglement of different cell types from each other.

dimension	1	2	3	4	5
CTRL	0.61	0.69	0.5	0.72	1.0
STIM	0.76	0.63	0.0	0.91	1.0

Table 21: Disentanglement Scores within 'condition' feature in Kang dataset. The model used had specifications of β=20 and C=30. In the Table (6) 'condition' is disentangled from other features in dimension 5. Here, too the 5th dimension shows a complete disentanglement of 'CTRL' from 'STIM'.

dimension	1	2	3	4	5
B	0.00	0.00	0.00	0.00	0.00
B Activated	0.00	0.00	0.00	0.00	0.08
CD14 Mono	0.00	0.00	0.00	0.00	0.00

Continued on next page

dimension	1	2	3	4	5
CD16 Mono	0.00	0.00	0.00	0.00	0.00
CD4 Memory T	0.08	0.00	0.00	0.00	0.00
CD4 Naive T	0.00	0.00	0.00	0.00	0.00
CD8 T	0.00	0.00	0.00	0.08	0.00
DC	0.00	0.00	0.08	0.00	0.00
Mk	0.00	0.00	0.00	0.00	0.00
NK	0.00	0.08	0.00	0.00	0.00
T activated	0.00	0.00	0.00	0.00	0.00
pDC	0.00	0.00	0.00	0.00	0.00

Table 22: Disentanglement Scores within 'cell type' feature in Kang dataset. The model used had specifications of β=20 and C=500. In the Table (7) 'cell type' is disentangled from other features in dimension 5. However, the disentanglement within the feature is very poor.

dimension	1	2	3	4	5
CTRL	0.48	0.47	0.5	0.5	0.00
STIM	0.00	0.00	0.0	0.0	0.48

Table 23: Disentanglement Scores within 'condition' feature in Kang dataset. The model used had specifications of β=20 and C=500. In the Table (7) 'condition' is disentangled from other features in dimension 4. However, the disentanglement within the feature is very poor.

dHSIC models: Dentate Gyrus

dimension	1	2	3	4	5
Astrocytes	0.56	0.72	1.00	0.17	0.00
Cajal Retzius	0.31	0.76	0.50	0.22	0.93
Cck-Tox	0.33	0.55	0.64	0.14	1.00

Continued on next page

dimension	1	2	3	4	5
Endothelial	0.28	0.82	0.78	0.00	0.96
GABA	0.00	0.46	0.75	0.21	1.00
Granule immature	0.19	0.89	0.67	0.44	0.61
Granule mature	0.14	0.66	0.81	0.00	0.86
Microglia	0.31	0.89	1.00	0.00	1.00
Mossy	0.00	1.00	0.74	0.00	0.97
Neuroblast	0.00	0.90	0.81	0.00	0.97
OL	0.39	0.68	0.76	0.00	1.00
OPC	1.00	0.73	0.38	0.00	0.43
Radial Glia-like	0.98	0.65	0.00	0.13	0.51
nIPC	0.90	0.62	0.80	0.57	0.73

Table 24: Disentanglement Scores within 'cluster' feature in Dentate Gyrus dataset. The model used had specifications of β=1 and γ=50. From Table (10), dimensions 2 and 5 showed strong disentanglement. It can be seen here as well. Dimension 5 disentangles within the different clusters better than dimension 2.

dimension	1	2	3	4	5
Astrocytes	1.00	0.47	1.00	0.27	0.00
Cajal Retzius	0.54	0.82	0.52	0.15	0.11
Cck-Tox	0.95	0.77	0.33	0.17	0.16
Endothelial	0.98	1.00	0.29	0.16	0.00
GABA	0.64	0.72	0.27	0.00	0.14
Granule immature	0.52	0.68	0.62	0.00	0.09
Granule mature	0.52	0.22	0.31	0.38	0.00
Microglia	1.00	1.00	1.00	0.00	0.19
Mossy	0.56	0.72	0.48	0.00	0.23
Neuroblast	0.62	0.62	0.72	0.00	0.00
OL	0.95	1.00	0.29	0.18	0.16
OPC	0.86	1.00	1.00	0.00	0.00
Radial Glia-like	0.97	0.44	0.75	0.27	0.00
nIPC	0.66	0.96	0.55	0.75	0.58

Continued on next page

dimension	1	2	3	4	5

Table 25: Disentanglement Scores within 'cluster' feature in Dentate Gyrus dataset. The model used had specifications of β=1 and γ=50. From Table (11), dimensions 1,2 and 3 showed strong disentanglement for 'cluster'.

dHSIC models: Kang

dimension	1	2	3	4	5
B	0.00	0.94	0.00	0.42	0.25
B Activated	0.15	0.48	0.17	0.75	0.25
CD14 Mono	0.14	0.32	0.31	0.92	0.00
CD16 Mono	0.12	0.96	0.18	0.58	0.00
CD4 Memory T	0.07	0.00	0.31	0.49	0.18
CD4 Naive T	0.18	0.31	0.15	0.78	0.16
CD8 T	0.00	0.30	0.12	0.30	0.54
DC	0.00	0.66	0.85	0.56	0.55
Mk	0.15	0.00	0.26	0.74	0.33
NK	0.00	0.60	0.47	0.37	0.56
T activated	0.00	1.00	0.33	0.94	0.00
pDC	0.13	0.64	0.00	0.36	0.58

Table 26: Disentanglement Scores within 'cell type' feature in Kang dataset. The model used had specifications of β=1 and γ=50. From Table (12), dimension 4 showed strong disentanglement for 'cell type'. As evident from the table above, disentanglement within the feature is also

dimension	1	2	3	4	5
B	0.49	0.85	0.18	0.24	0.47
B Activated	0.69	0.83	0.00	0.17	0.54

Continued on next page

dimension	1	2	3	4	5
CD14 Mono	0.56	0.25	0.00	0.15	0.41
CD16 Mono	0.55	0.35	0.14	0.00	0.94
CD4 Memory T	0.73	0.85	0.12	0.25	0.45
CD4 Naive T	0.55	0.25	0.00	0.00	0.58
CD8 T	0.40	0.44	0.11	0.15	0.64
DC	0.81	0.90	0.10	0.67	0.54
Mk	0.00	0.17	0.00	0.00	0.42
NK	0.65	0.88	0.00	0.19	0.88
T activated	0.88	0.69	0.13	0.49	0.91
pDC	0.76	0.95	0.14	0.14	0.32

Table 27: Disentanglement Scores within 'cell type' feature in Kang dataset. The model used had specifications of β=1 and γ=100. From Table (13), dimension 1 showed strong disentanglement for 'cell type'. As evident from the table above, disentanglement within the feature is also good.

dimension	1	2	3	4	5
CTRL	0.64	0.62	1.0	0.66	0.71
STIM	0.62	0.56	1.0	0.61	0.62

Table 28: Disentanglement Scores within 'condition' feature in Kang dataset. The model used had specifications of β=1 and γ=50. From Table (12), dimension 3 showed strong disentanglement for 'condition'. As evident from the table above, disentanglement within the feature is completely accurate.

dimension	1	2	3	4	5
CTRL	0.59	0.64	0.65	1.0	0.59
STIM	0.60	0.55	0.62	1.0	0.54

Continued on next page

dimension	1	2	3	4	5

Table 29: Disentanglement Scores within 'condition' feature in Kang dataset. The model used had specifications of β=1 and γ=100. From Table (13), dimension 4 showed strong disentanglement for 'condition'. As evident from the table above, disentanglement within the feature is completely accurate.

3.2.3 Disentanglement through Latent Space

β-VAE model with C: Dentate Gyrus Disentanglement can be visualised by their distributions in the latent space.

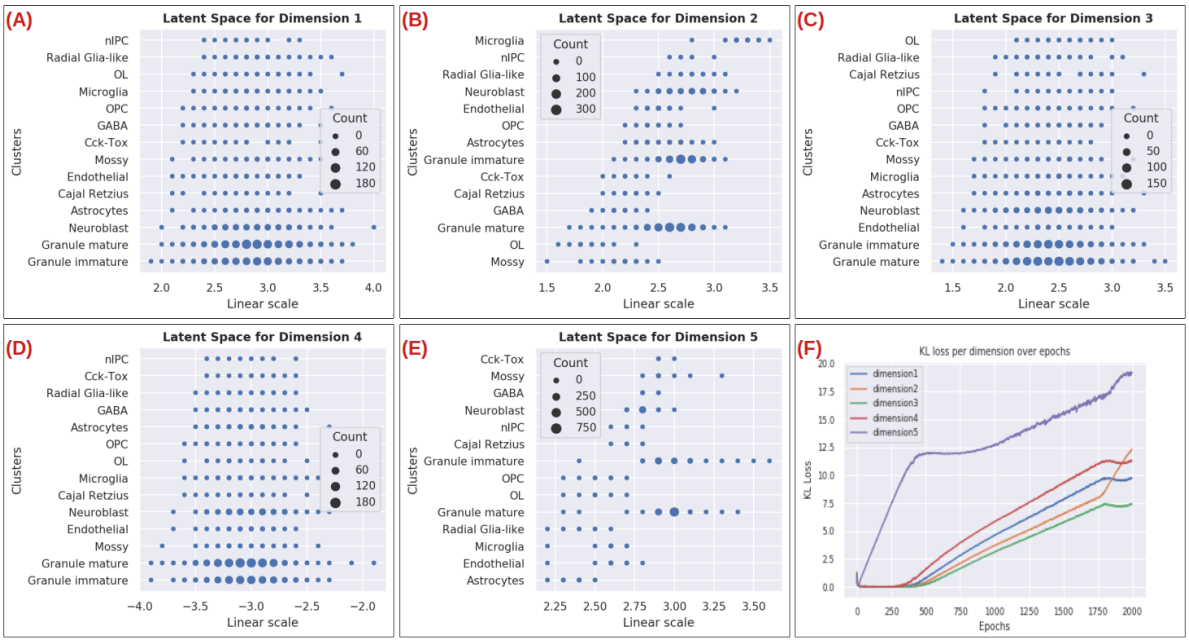

Figure 29: (A)-(E) are the approximated posterior distribution of the latent space, per dimension. It also shows how the 'cluster' feature is distributed. (F) shows how the KL loss component varies over epochs and the linearly increasing controlling capacity, C. The KL loss for the 5th dimension is the highest. This corresponds to the disentanglement in dimension 5 for the different clusters. The KL loss for the 2nd dimension increases towards the end when more capacity is allowed. This also improves disentanglement in the 2nd dimension. The model used has specifications of β=50 and C=30.

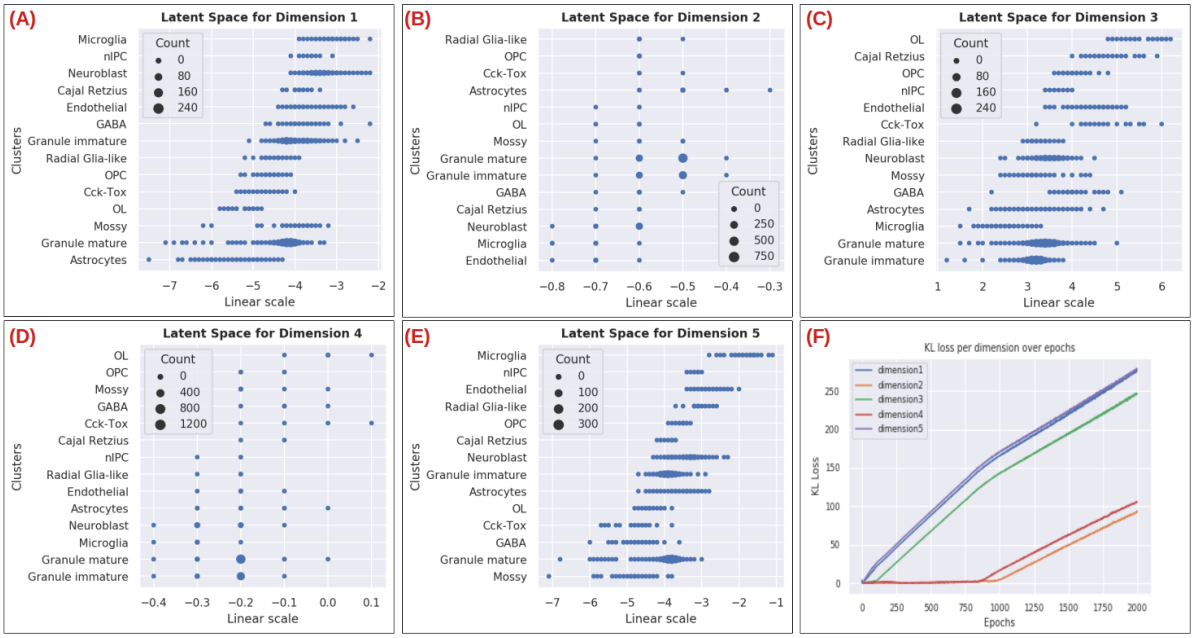

Figure 30: (A)-(E) are the approximated posterior distribution of the latent space, per dimension. It also shows how the 'cluster' feature is distributed. (F) shows how the KL loss component varies over epochs and the linearly increasing controlling capacity, C. The KL loss steadily increases with epochs for the 1st, 3rd and 5th dimensions. The KL loss for 2nd and 4th dimensions increases after 1000 epochs. The latent space plots for these two dimensions, does not disentangle different cluster values well. However the variance is less. The range of values in the latent space plot are more deterministic. The model used has specifications of β=50 and C=500.

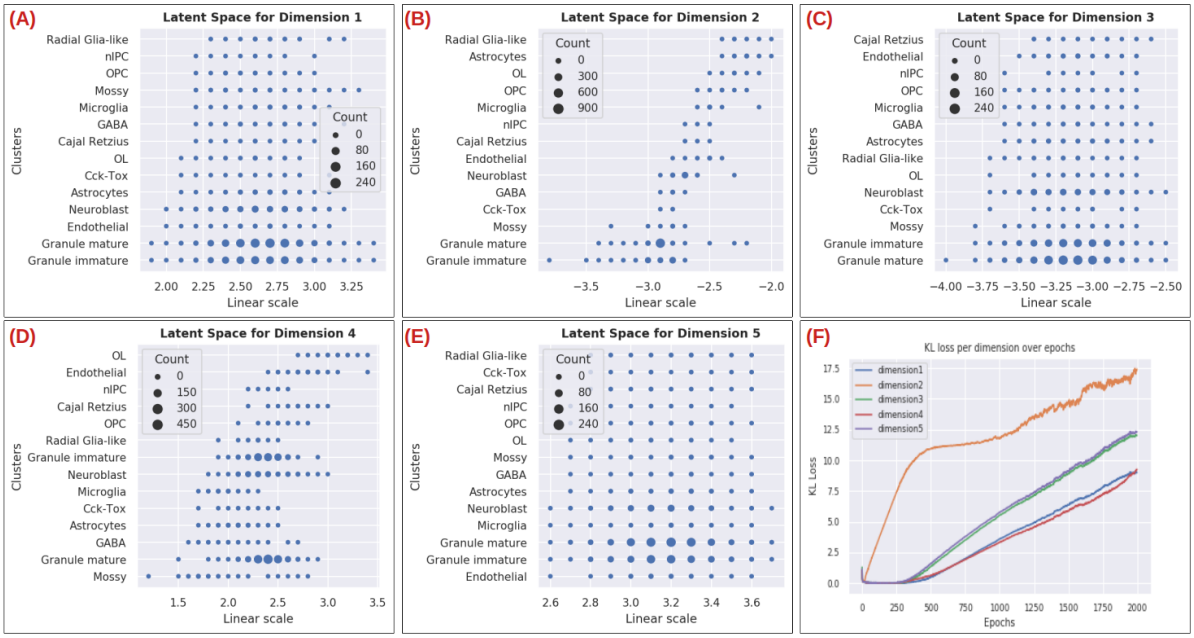

Figure 31: (A)-(E) are the approximated posterior distribution of the latent space, per dimension. It shows also shows how the 'cluster' feature is distributed. (F) shows how the KL loss component varies over epochs and the linearly increasing controlling capacity, C. The KL loss is highest for 2nd dimension which also shows strong disentanglement of values. 1st, 3rd and 5th dimensions do not vary for different clusters. Although the KL loss for 4th dimension was small, a slight disentanglement can be observed. The model used has specifications of β=100 and C=30.

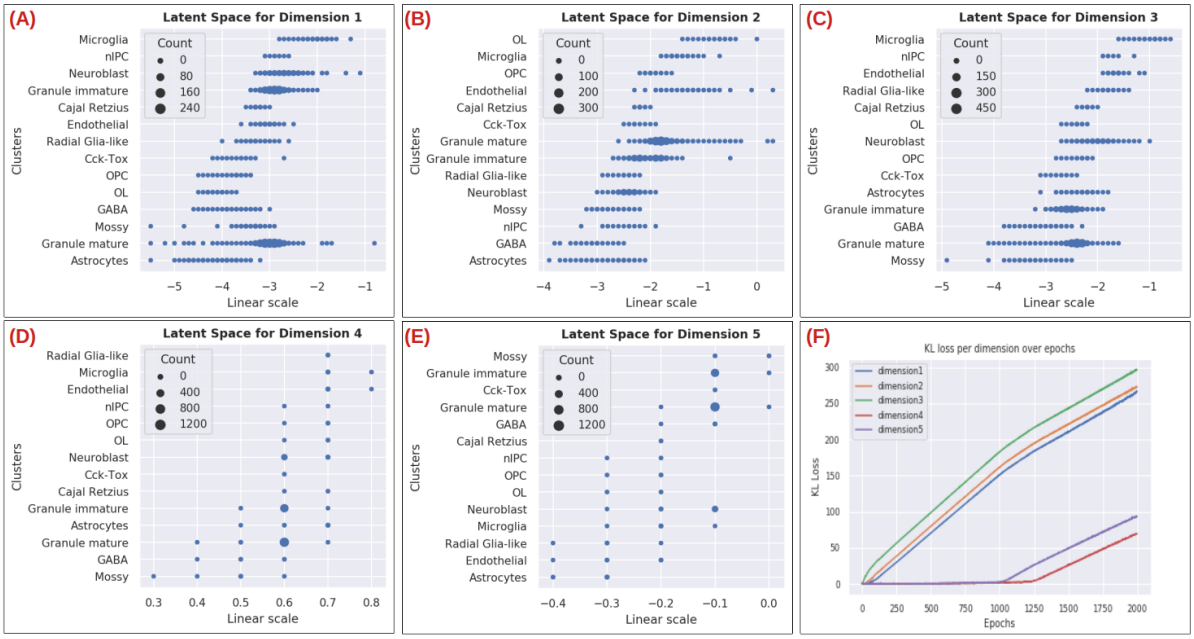

Figure 32: (A)-(E) are the approximated posterior distribution of the latent space, per dimension. It shows also shows how the 'cluster' feature is distributed. (F) shows how the KL loss component varies over epochs and the linearly increasing controlling capacity, C. The KL loss steadily increases with epochs for the 1st, 2nd and 3rd dimensions. They also show disentanglement for different clusters. The values for 4th and 5th dimensions are more deterministic than smaller C values models 31. These two dimensions, also disentangle better than the small β value models 30. The model used has specifications of β=100 and C=500.

β-VAE model with C: Kang

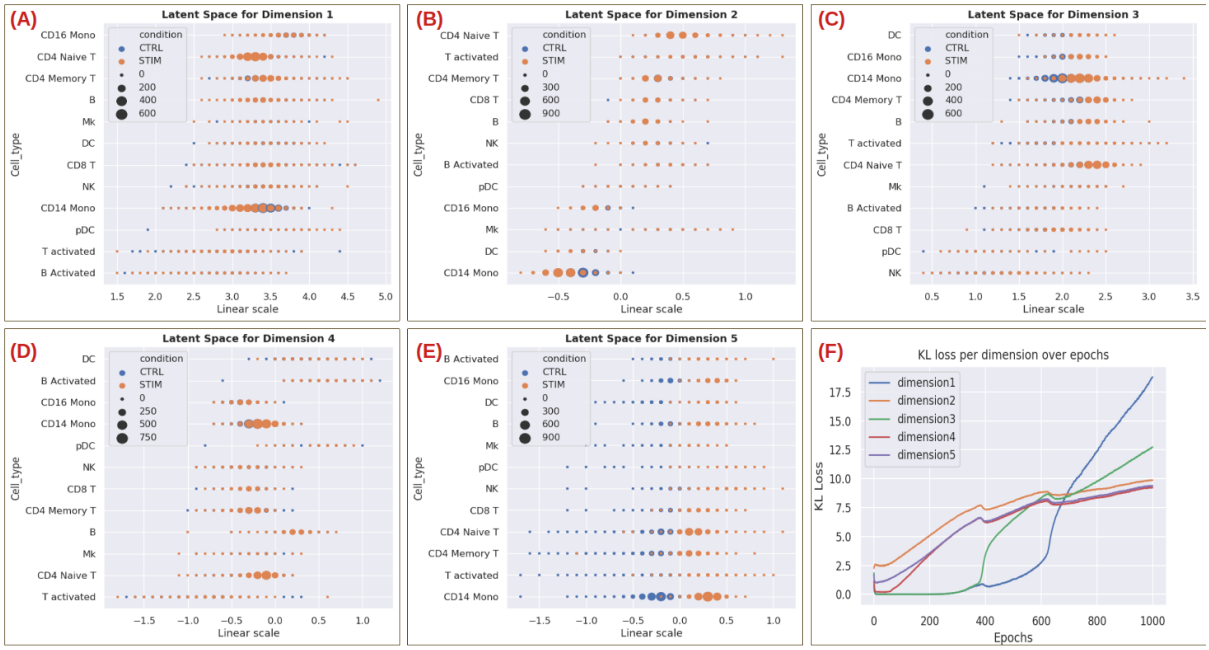

Figure 33: (A)-(E) are the approximated posterior distribution of the latent space, per dimension. It shows how the features, 'condition' and 'cell type' are distributed. (F) shows how the KL loss component varies over epochs and the linearly increasing controlling capacity, C. The KL loss slowly increases with epochs for 2nd, 4th and 5th dimensions. Dimensions 2 and 4 disentangle 'cell type' and dimension 5 disentangles 'condition'. The KL loss for 1st and 3rd increase after a few epochs, however they do not show any disentanglement. The model used has specifications of β=5 and C=30.

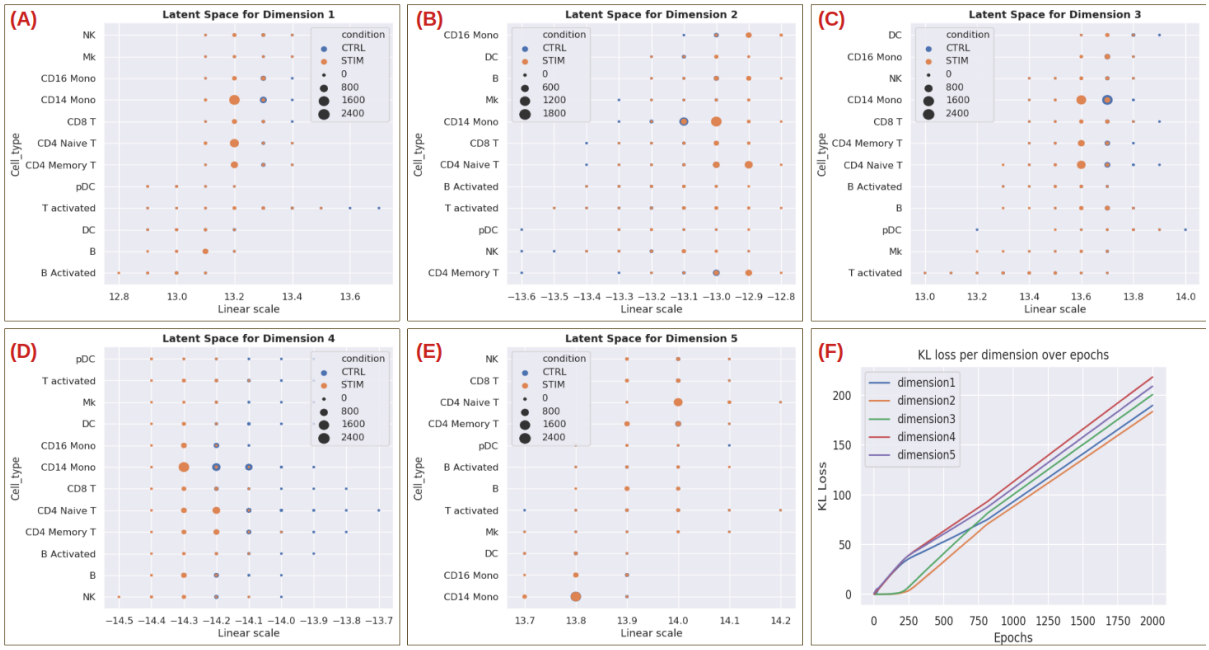

Figure 34: (A)-(E) are the approximated posterior distribution of the latent space, per dimension. It shows how the features, 'condition' and 'cell type' are distributed. (F) shows how the KL loss component varies over epochs and the linearly increasing controlling capacity, C. The KL loss increases with epochs for dimensions 1, 4 and 5. Dimensions 1 and 5 disentangle 'cell type' and dimension 4 disentangles 'condition'. The KL loss for 2nd and 3rd increases after a few epochs, however they do not show any disentanglement. It can also be seen that the values in the latent space are more deterministic and the variance is less. The model used has specifications of β=20 and C=500.

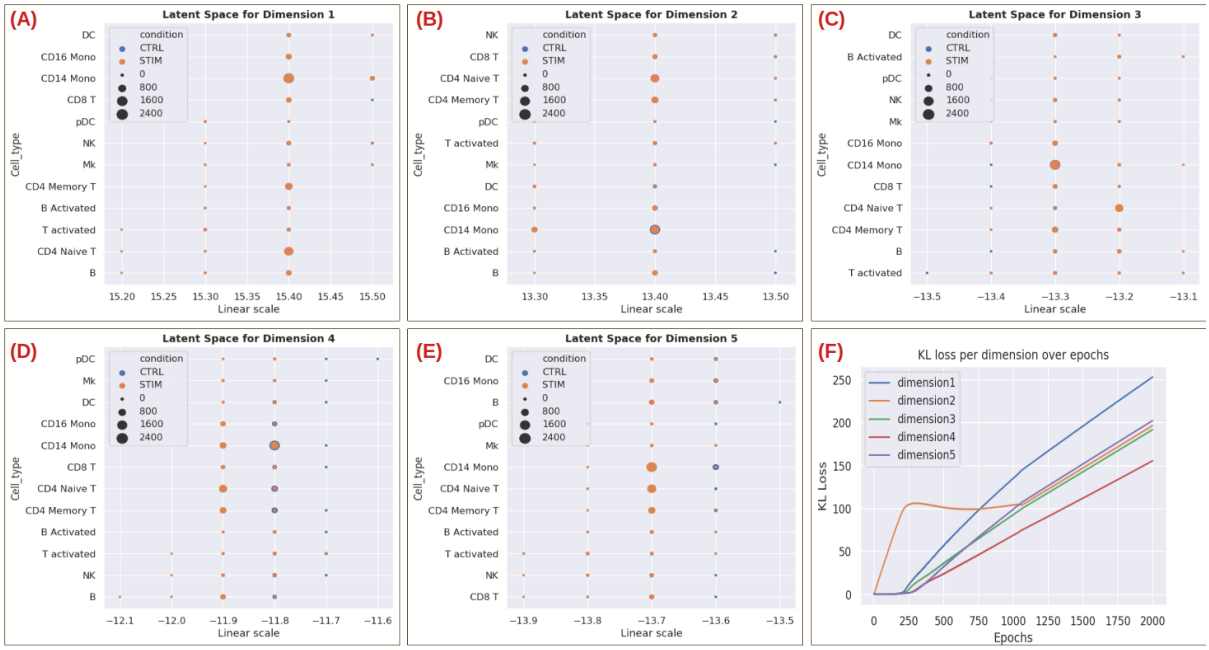

Figure 35: (A)-(E) are the approximated posterior distribution of the latent space, per dimension. It shows how the features, 'condition' and 'cell type' are distributed. (F) shows how the KL loss component varies over epochs and the linearly increasing controlling capacity, C. The KL loss rapidly increase for dimension 2 and then saturates. Later it increases again. This dimension does not show disentanglement. All other dimensions show some disentanglement and also have more deterministic values. Dimension 4 is differentiating between conditions to some extent. The model used has specifications of β=100 and C=500.

dHSIC model: Dentate Gyrus

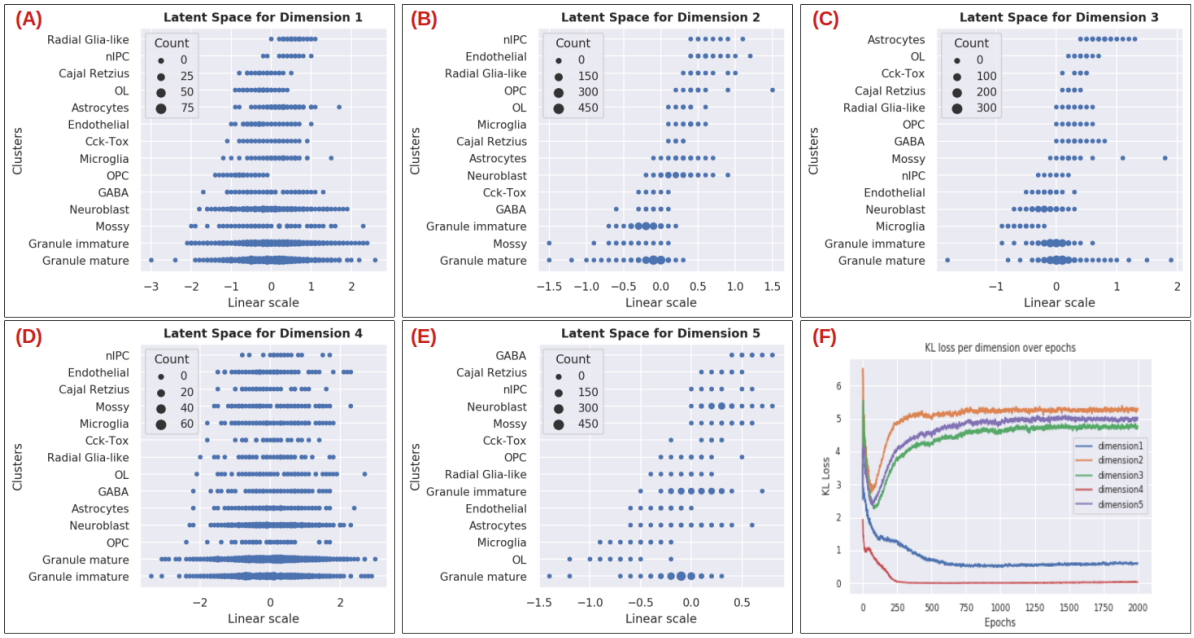

Figure 36: (A)-(E) are the approximated posterior distribution of the latent space, per dimension. It shows also shows how the 'cluster' feature is distributed. (F) shows how the KL loss component varies over epochs. The KL loss for 1st and 4th dimensions steadily decreases with epochs. They do not show any disentanglement within different clusters. For 2nd, 3rd and 5th dimensions, the KL loss first falls and then increases till it reaches a saturated value of about 5. These dimensions show disentanglement as well. The model used has specifications of β=1 and γ=50.

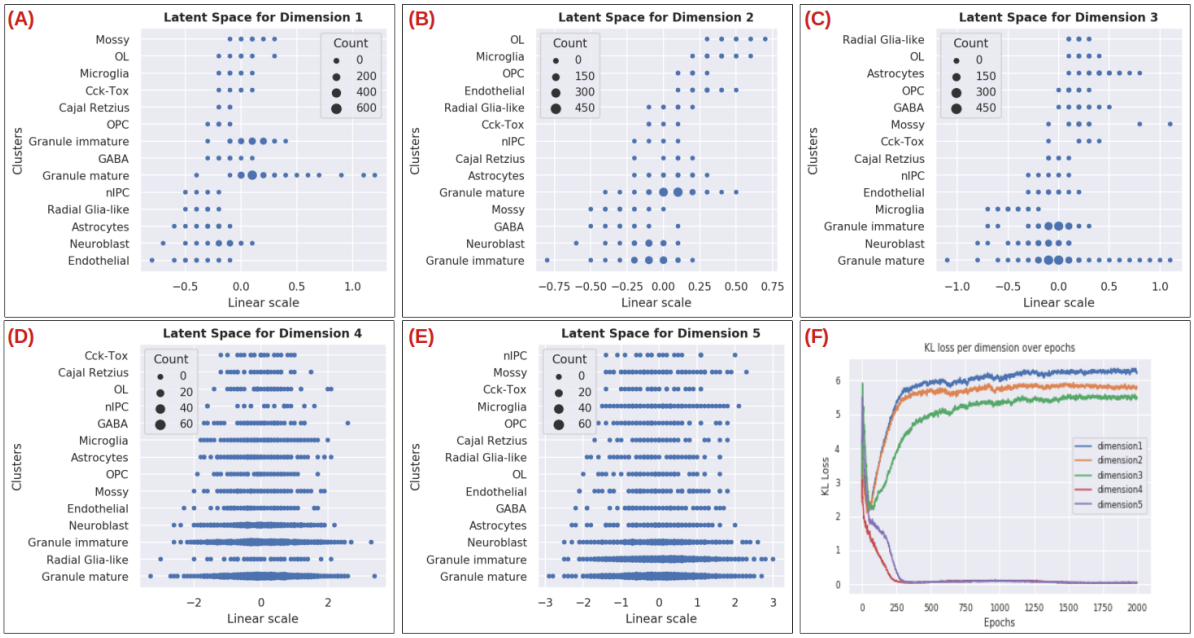

Figure 37: (A)-(E) are the approximated posterior distribution of the latent space, per dimension. It shows also shows how the 'cluster' feature is distributed. (F) shows how the KL loss component varies over epochs. The KL loss for 4th and 5th dimensions steadily decreases with epochs. They do not show any disentanglement within different clusters. For 1st, 2nd and 3rd dimensions, the KL loss first falls and then increases till it reaches a saturated value of about 6. These dimensions show disentanglement as well. The model used has specifications of β=1 and γ=100. Compared to the KL loss in Figure 36(F) where the value of the highest KL loss is 5, the disentanglement score is also smaller.

dHSIC model: Kang

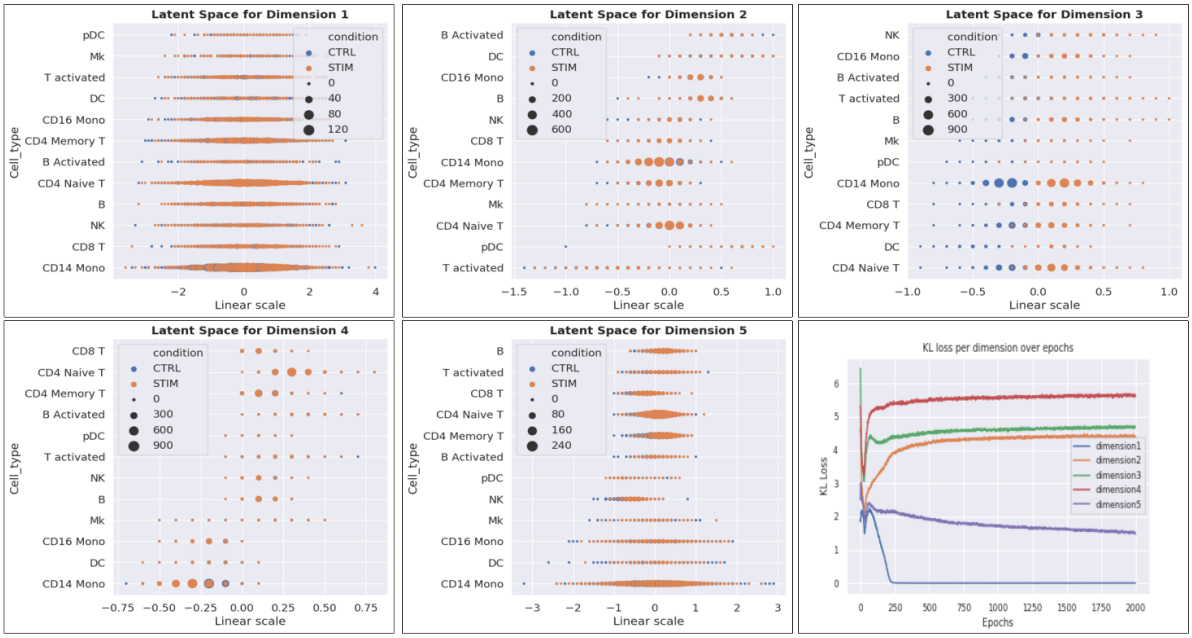

Figure 38: (A)-(E) are the approximated posterior distribution of the latent space, per dimension. It shows also shows how 'cell type' and 'condition' features are distributed. (F) shows how the KL loss component varies over epochs. Dimensions with highest KL loss are 3 and 4. They encode disentanglement for 'condition' and 'cell type' values. Although KL loss for 2nd dimension is also more than other, the disentanglement is not evident. The highest KL loss value is approximately 5.5. The model used has specifications of β=1 and γ=50.

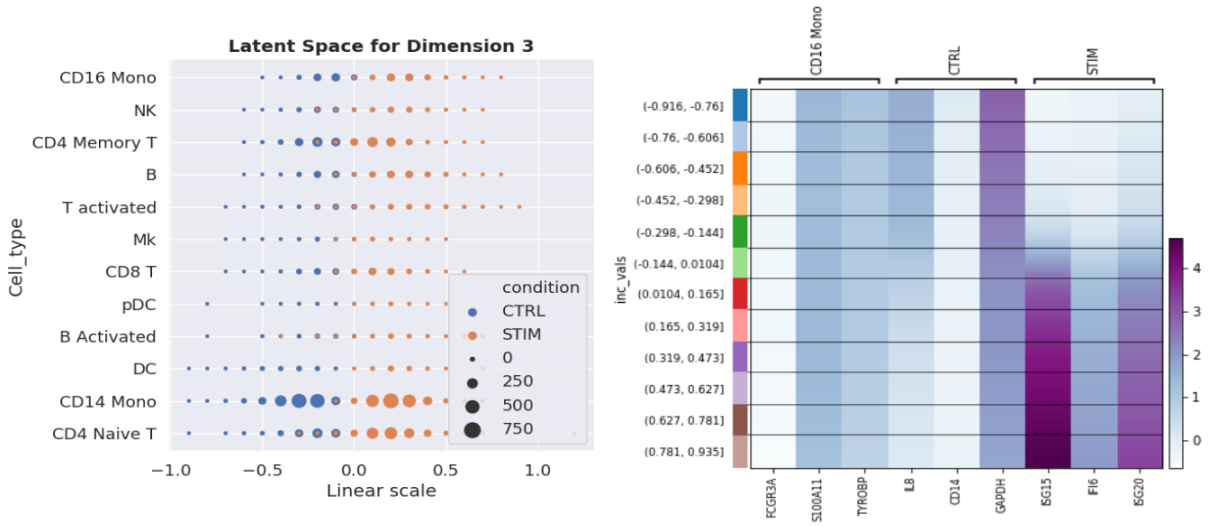

Figure 39: In this figure, the first plot shows the latent space for the 3rd dimension for the same dHSIC model(β=1,γ=50) as in Figure: 38. It can be seen that this dimension disentangles the 'condition' feature. The second plot shows how by manipulating the 3rd dimension, only the 'condition' marker genes are changing. The cell marker gene for 'CD16 Mono' stays the same. This shows that the model disentangles the two features from each other.

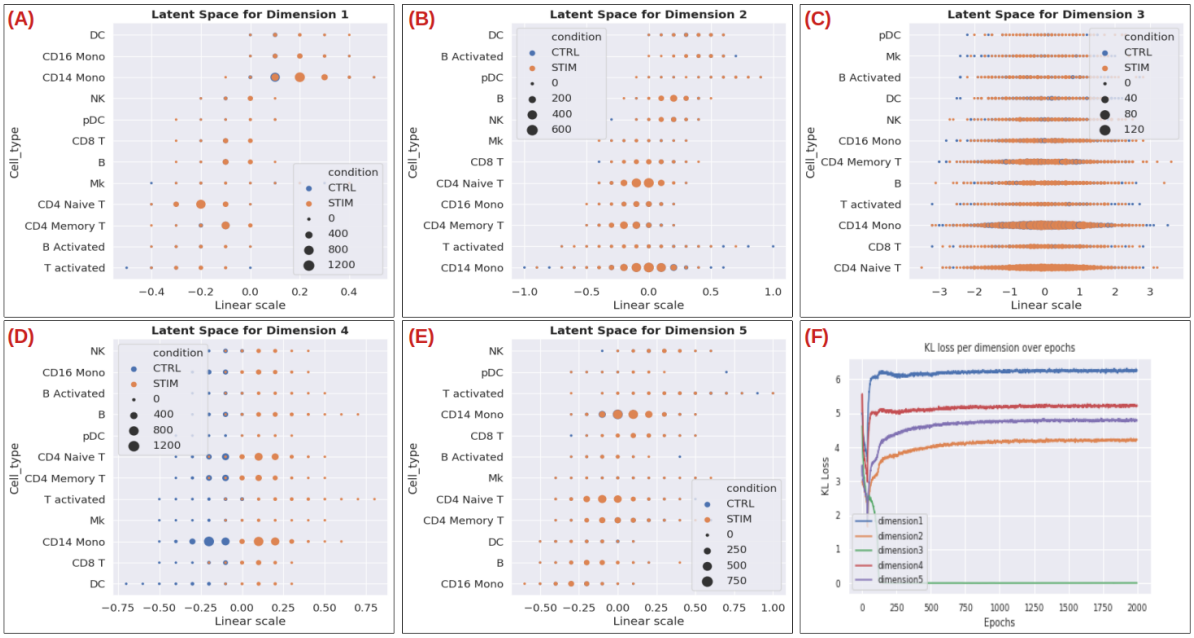

Figure 40: (A)-(E) are the approximated posterior distribution of the latent space, per dimension. It shows also shows how 'cell type' and 'condition' features are distributed. (F) shows how the KL loss component varies over epochs. In this model, there are 4 dimensions that have a high KL loss. These dimensions also show disentanglement. Additionally, the value of the highest KL loss is 6. The model used has specifications of β=1 and γ=100.

3.3 Out-of Sample Prediction

As an application of disentanglement, 'out-of-sample' prediction was carried for all 'cluster' type in Dentate Gyrus dataset. Every cluster was dropped once before training and was recovered by simulating values in the latent space starting from another cluster. On average the β-VAE with C with β=50 and C=30 performed best for this task.

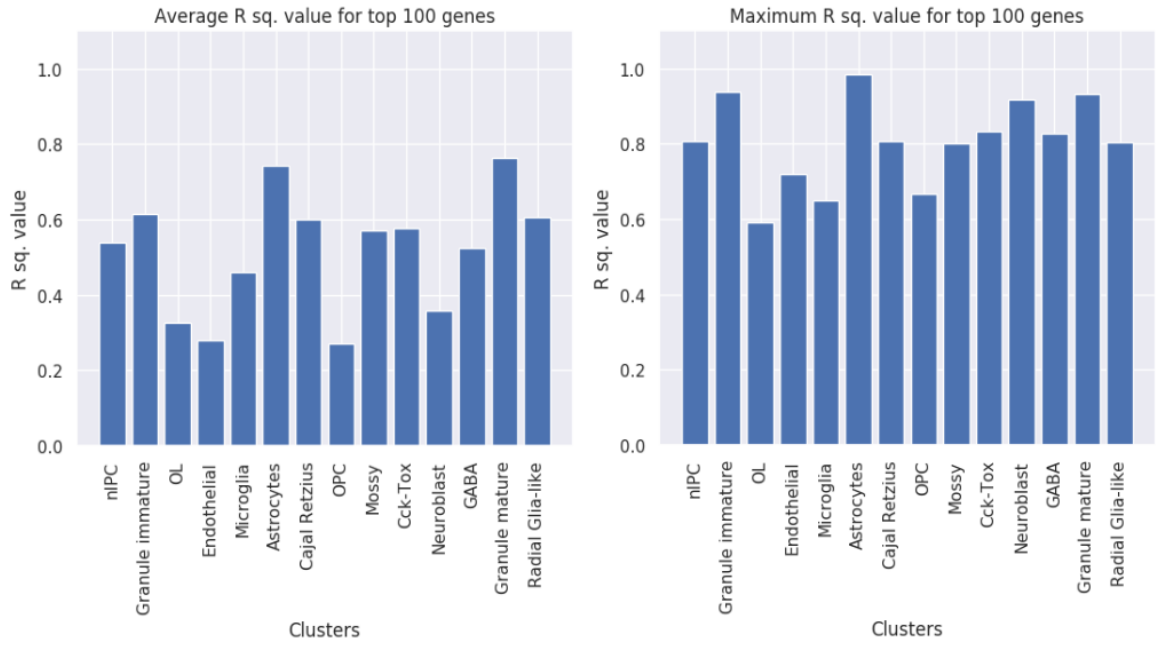

Figure 41: This figure shows the performance for out-of-sample prediction for β-VAE model with C with β=50 and C=30. This is for the Dentate Gyrus dataset.

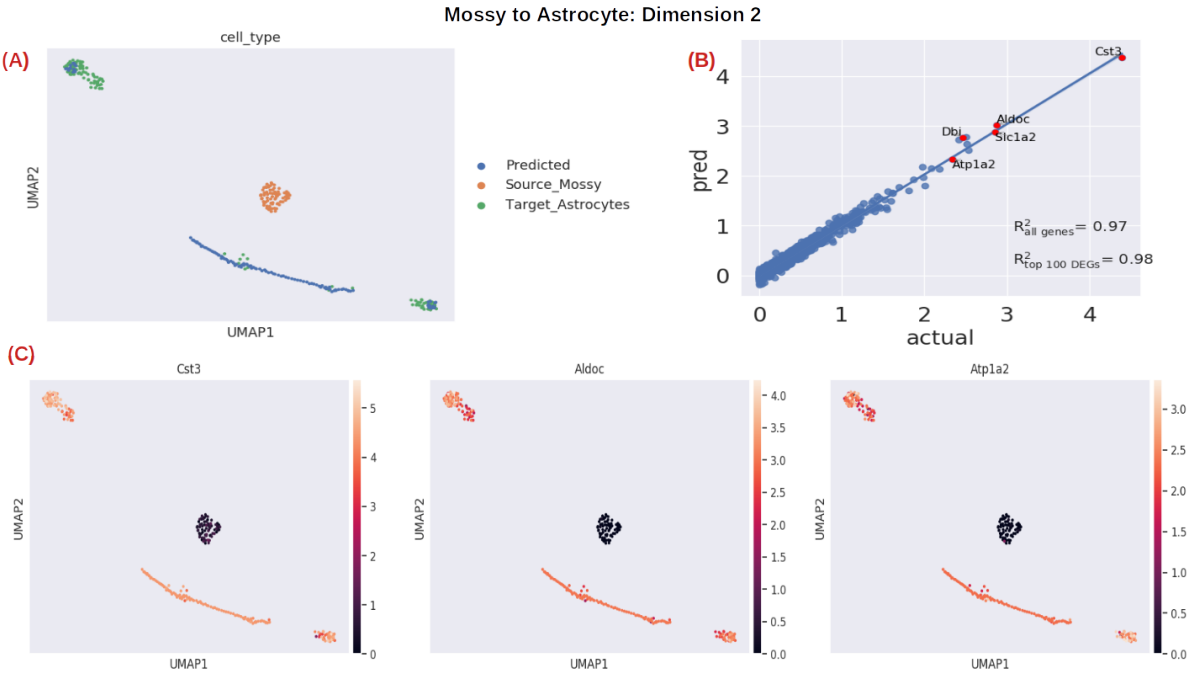

Figure 42: In this example, the cluster 'Mossy' recovered 'Astrocyte' by manipulating values in the 2nd dimension. (A) shows the UMAP of the predicted, source and target cells. (C) shows the UMAP for the top 3 marker genes for 'Astrocyte'. It can be seen that 'Mossy' cluster recovers the 'Astrocyte' cluster and it's genes. (B) shows a linear regression plot between the predicted genes and the actual 'Astrocyte' genes. The R squared value for the top 100 genes for 'Astrocyte' is 0.98. This also proves that 'Astrocyte' is recovered from 'Mossy' in dimension 2 although the model did not train in 'Astrocyte' at all.

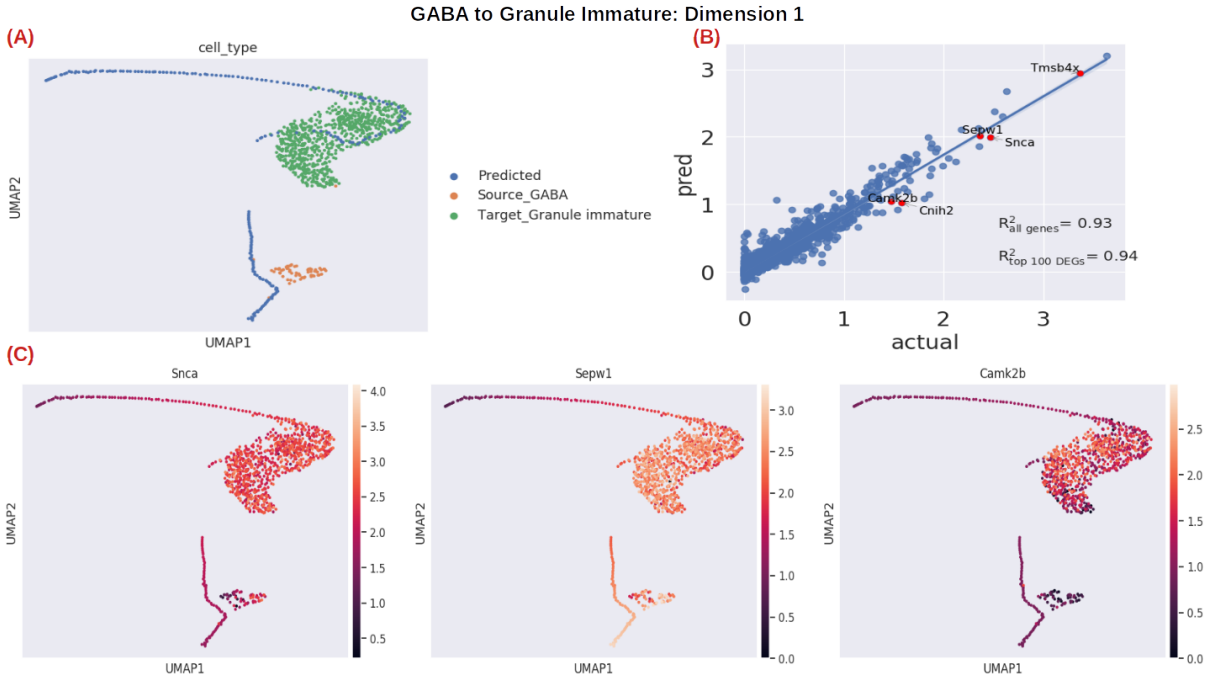

Figure 43: In this example, the cluster 'GABA' recovered 'Granule Immature' by manipulating values in the 1st dimension. (A) shows the UMAP of the predicted, source and target cells. (C) shows the UMAP for the top 3 marker genes for 'Granule Immature'. It can be seen that 'GABA' cluster recovers the 'Granule Immature' cluster and it's genes. (B) shows a linear regression plot between the predicted genes and the actual 'Granule Immature' genes. The R-squared value for the top 100 genes for 'Astrocyte' is 0.94. This also proves that 'Granule Immature' is recovered from 'GABA' in dimension 1 although the model did not train in 'Granule Immature' at all.

Out of sample prediction for Kang dataset was also done by dropping one cell type and training the model. For β-VAE model with C, $\beta = 100$, C=500, on average gave the best R-squared value for top 100 genes, 0.79. The worst was $\beta = 5$, C=30 where the average was 0.51.

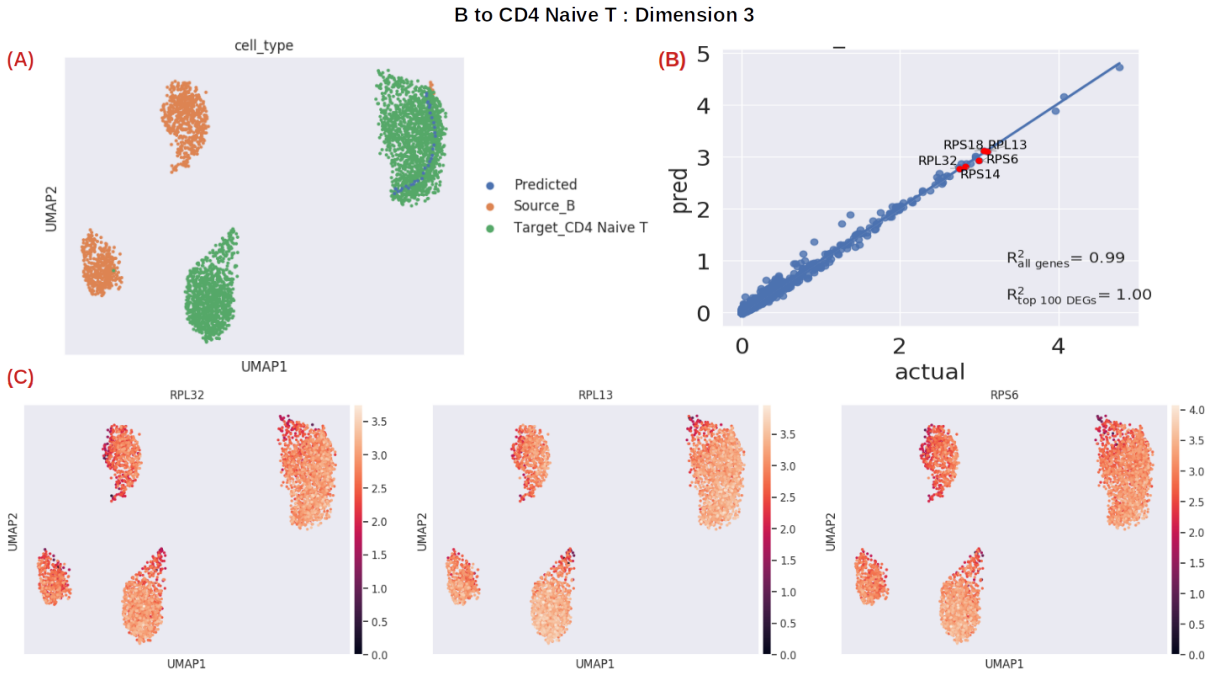

Figure 44: In this example, the cell type 'B' recovered 'CD4 Naive T' by manipulating values in the 3rd dimension. (A) shows the UMAP of the predicted, source and target cells. (C) shows the UMAP for the top 3 marker genes for 'CD4 Naive T'. It can be seen that 'B' cell recovers the 'CD4 Naive T' cell and it's genes. (B) shows a linear regression plot between the predicted genes and the actual 'CD4 Naive T' genes. The R-squared value for the top 100 genes for 'CD4 Naive T' is 1. This also proves that 'CD4 Naive T' is recovered from 'B' in dimension 3 although the model did not train in 'CD4 Naive T' at all.

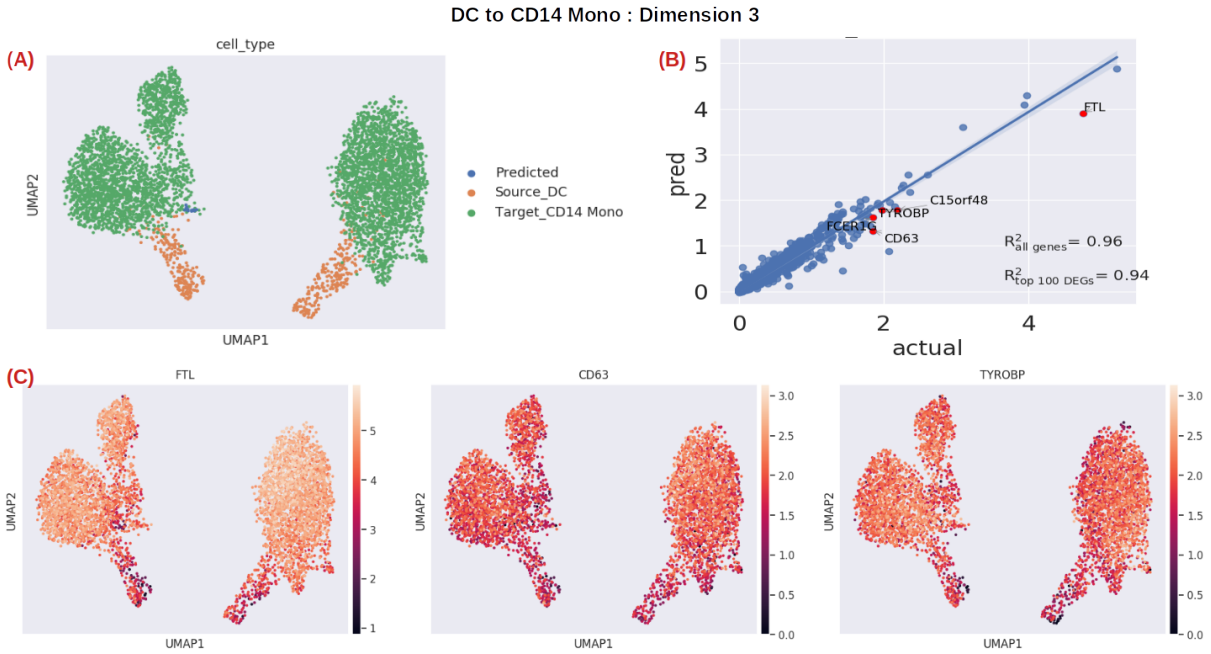

Figure 45: In this example, the cell type 'DC' recovered 'CD14 Mono' by manipulating values in the 3rd dimension. (A) shows the UMAP of the predicted, source and target cells. (C) shows the UMAP for the top 3 marker genes for 'CD14 Mono'. It can be seen that 'DC cell recovers the 'CD14 Mono' cell and it's genes. (B) shows a linear regression plot between the predicted genes and the actual 'CD14 Mono' genes. The R-squared value for the top 100 genes for 'CD14 Mono' is 0.94. This also proves that 'CD14 Mono' is recovered from 'DC' in dimension 3 although the model did not train in 'CD14 Mono' at all.

dHSIC Out of sample prediction did not perform well for the dHSIC model in the Dentate Gyrus dataset. The average R-squared value for top 100 genes was only 0.23. However, the results for the Kang dataset were relatively better. The out-of-sample prediction worked for β=1 and γ=50. The average R-squared value for top 100 genes was 0.64.

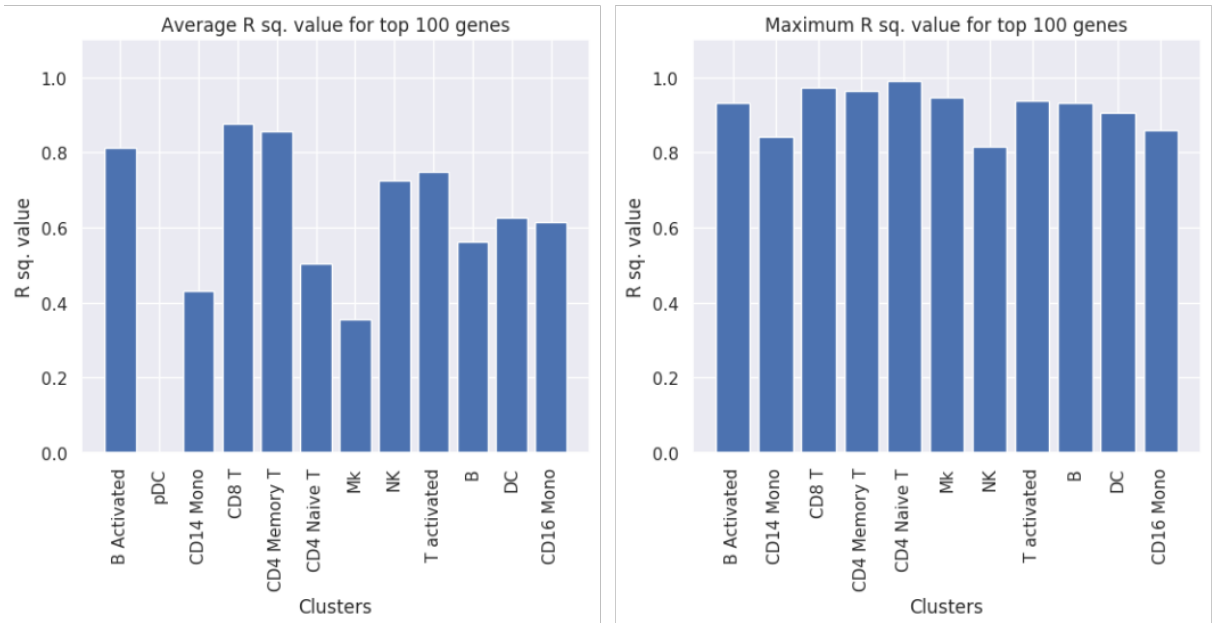

Figure 46: This figure shows the performance for out-of-sample prediction for dHSIC with β=1 and γ=50. This is for the Kang dataset.

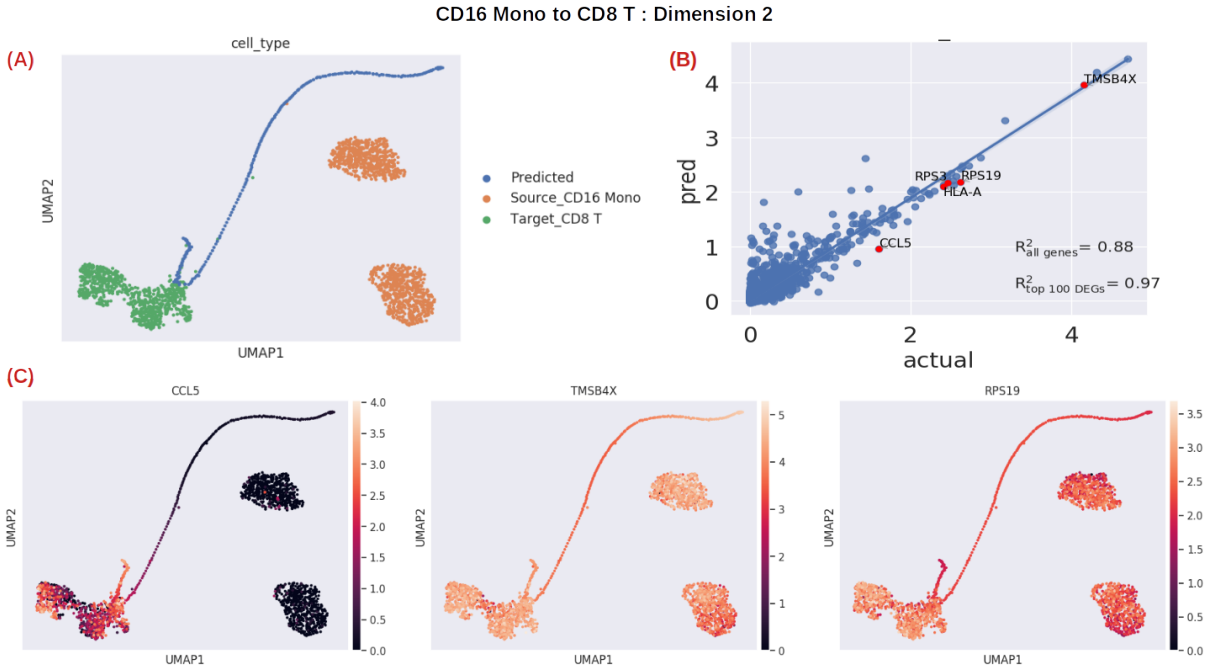

Figure 47: In this example, the cell type 'CD16 Mono' recovered 'CD8 T' by manipulating values in the 2nd dimension. (A) shows the UMAP of the predicted, source and target cells. (C) shows the UMAP for the top 3 marker genes for 'CD8 T'. It can be seen that 'CD16 Mono' cell recovers the 'CD8 T' cell and it's genes. (B) shows a linear regression plot between the predicted genes and the actual 'CD8 T' genes. The R-squared value for the top 100 genes for 'CD8 T' is 0.97. This also proves that 'CD8 T' is recovered from 'CD16 Mono' in dimension 2 although the model did not train in 'CD8 T' at all.

4 Discussion

The goal of this project was to find Latent Spaces that would be biologically interpretable and relevant for scRNA-sequences. As discussed in Section (1.3), learning a disentangled representation in an unsupervised way is an innovative and promising approach. It essentially helps in the downstream tasks which include classification, regression, visualization, and policy learning in reinforcement learning [20]. The models used in this project were based on unsupervised generative models. The modification made in the loss function, encouraged disentanglement of the latent space. By having such a representation, the latent space becomes more interpretable. This is even more important for massive datasets like scRNA-seq, where the challenge is to find biologically comprehensive lower dimensional spaces. The work done in this project is a step closer to disentangling generative features of scRNA-seq.

It was challenging to find generative features in the datasets. Two additional features