HC method builds upon an AE network. During training, an input data point $x \in \mathcal{R}^m$ is encoded into embedding $z \in \mathcal{R}^h$ and then decoded back into $\widehat{x} \in R^m$. From z a subnetwork learns clustering probability $c \in \mathcal{R}^k$. HC requires of three loss terms. The first loss ensures that the pairwise feature distances are preserved in the learned embeddings. Let δ be the pairwise cosine similarity and γ be the pairwise euclidean distance. For a minibatch X_B, we define:

$$\mathcal{L}_{x,z} = \lambda_{x,z,1}\frac{1}{B}\|\delta_B(Z_B) - \delta_B(X_B)\|_2^2 + \lambda_{x,z,2}\frac{1}{B}\|\delta_B(Z_B) - \delta_B(X_B)\|_2^2 \quad (1)$$

A soft norm penalty is required to enhance injectivity as it penalizes embeddings far from the unit norm hypersphere:

$$\mathcal{L}_{\|.\|} = \max(\|Z_B\|_2, (1 - \log(\|Z_B\|_2 + 1e^{-10}))) \quad (2)$$

The learned embeddings preserve the angular representation from input data, so that the unit vectors representing angles have direct correspondences to cluster probabilities by element-wise square root. The loss $\mathcal{L}_{z,c}$ that encourages matching angles between embeddings and clustering memberships is defined as:

$$\mathcal{L}_{z,c} = \frac{1}{B}\|\delta_B(C_B^{1/2}) - \max(\delta_B(Z_B), 0)\|_2^2 \quad (3)$$