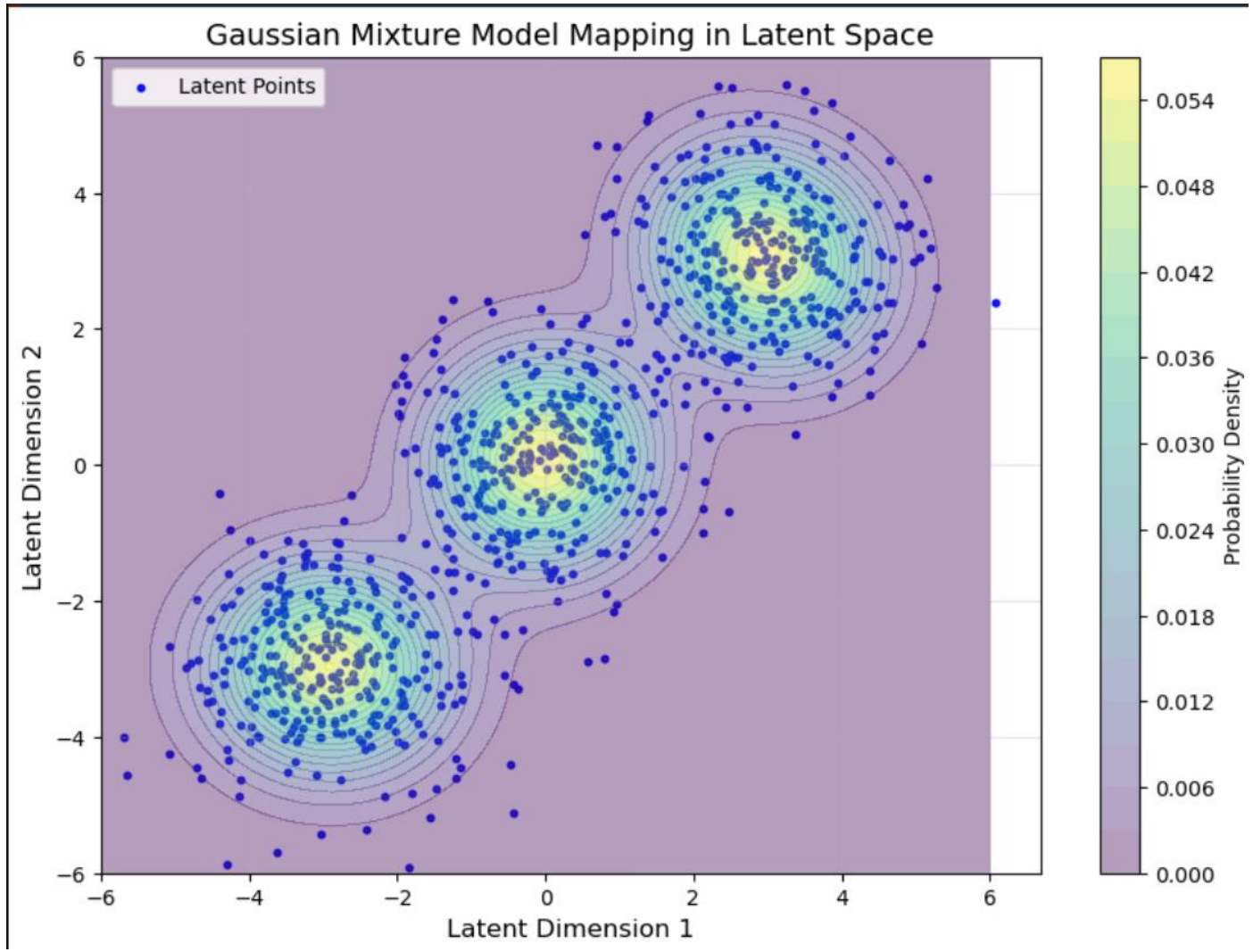
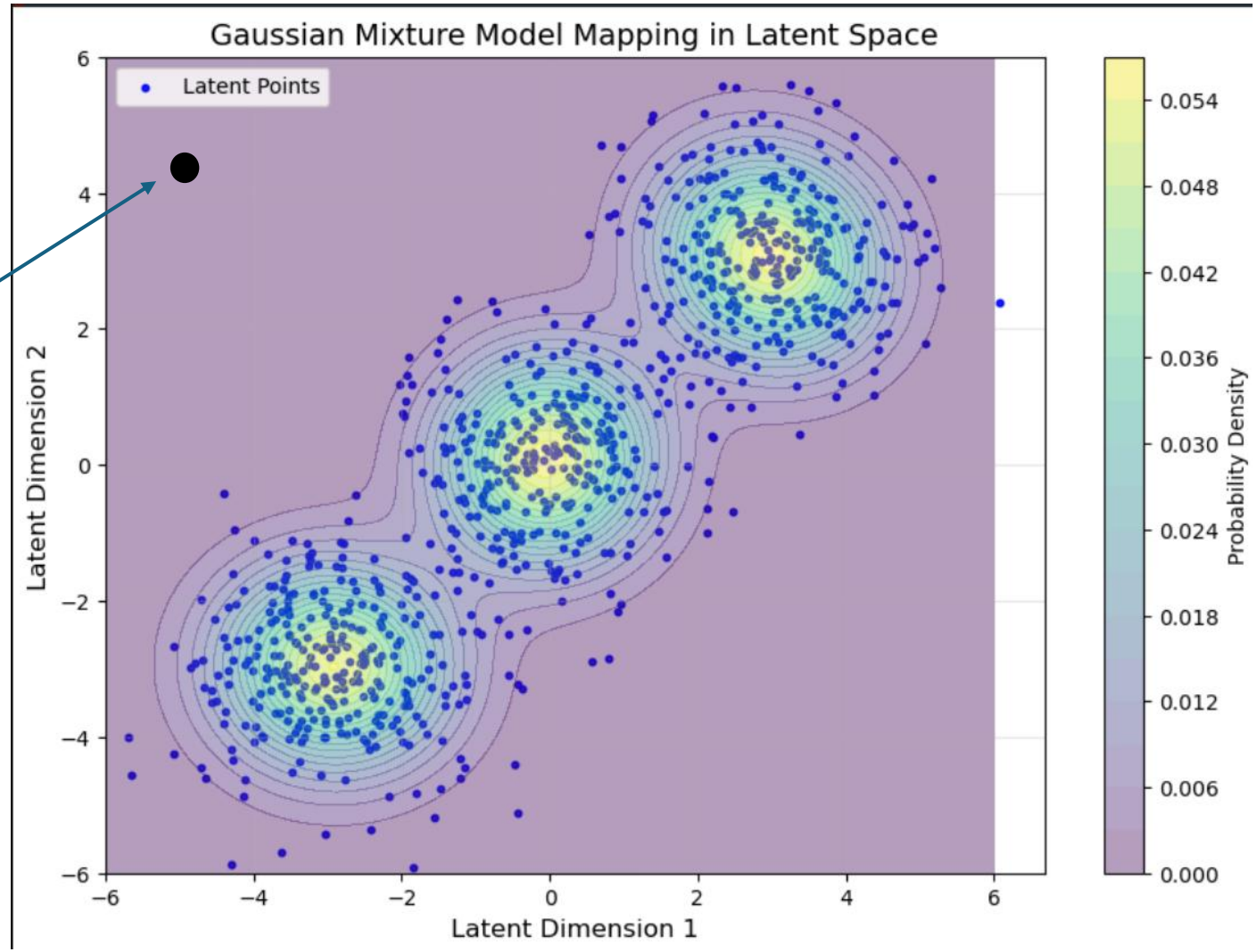
EGNN-GMM: Reliability of Molecular Property Predictions

Imagine your training set in latent space in latent sapce looks like this:



Source: https://github.com/cuauhtemocnv/EGNN-GMMFit

Blue points = training set
You Want to predict this point.
Do you think your dataset is okay?
Do you think that you will get a good prediction?



Gaussian Mixture Model

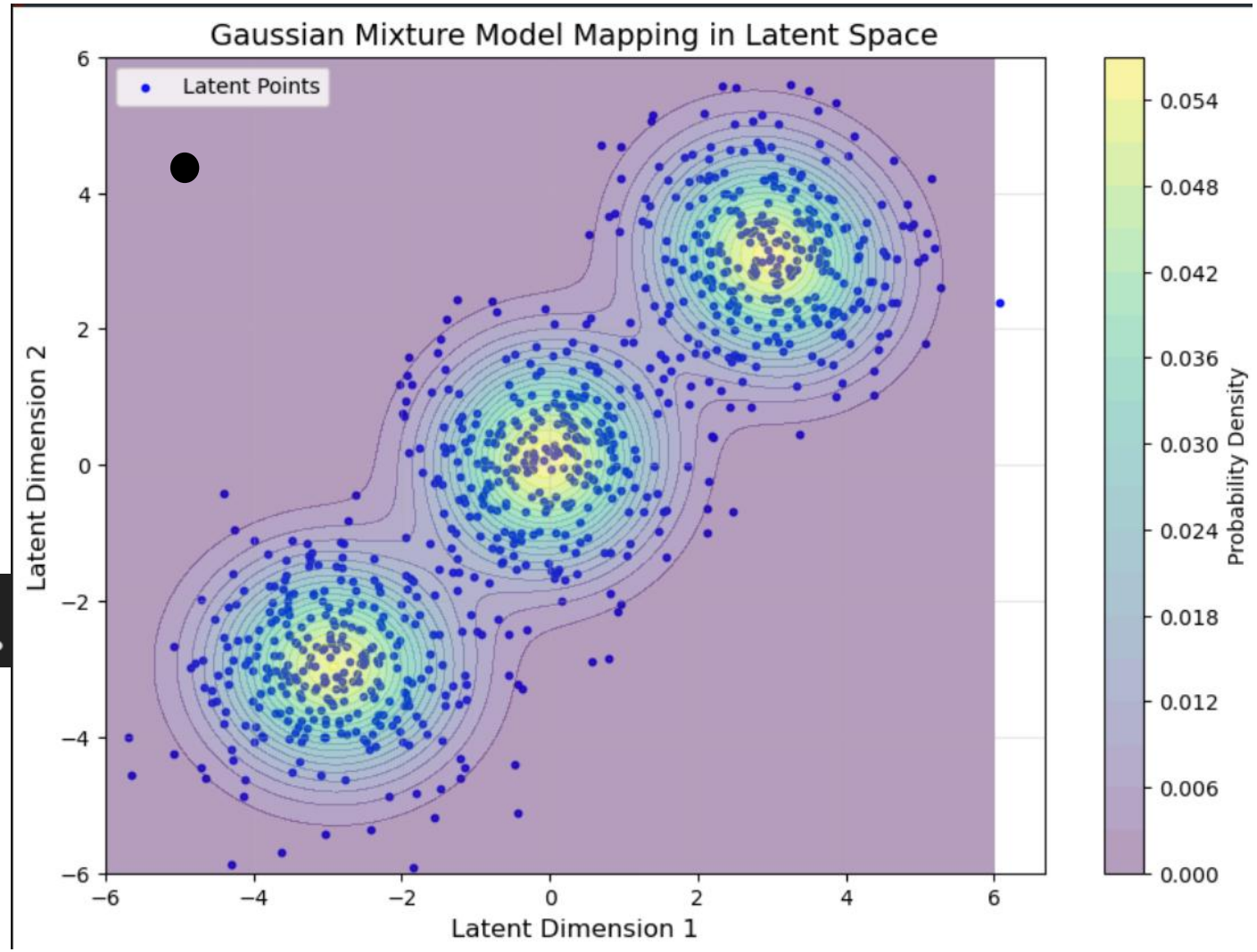
Model the distribution of the training points in latent space with 3 gaussians (K=3)
“Probability” density p(x) your point x lies into the distribution:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

We define the negative log likelihood

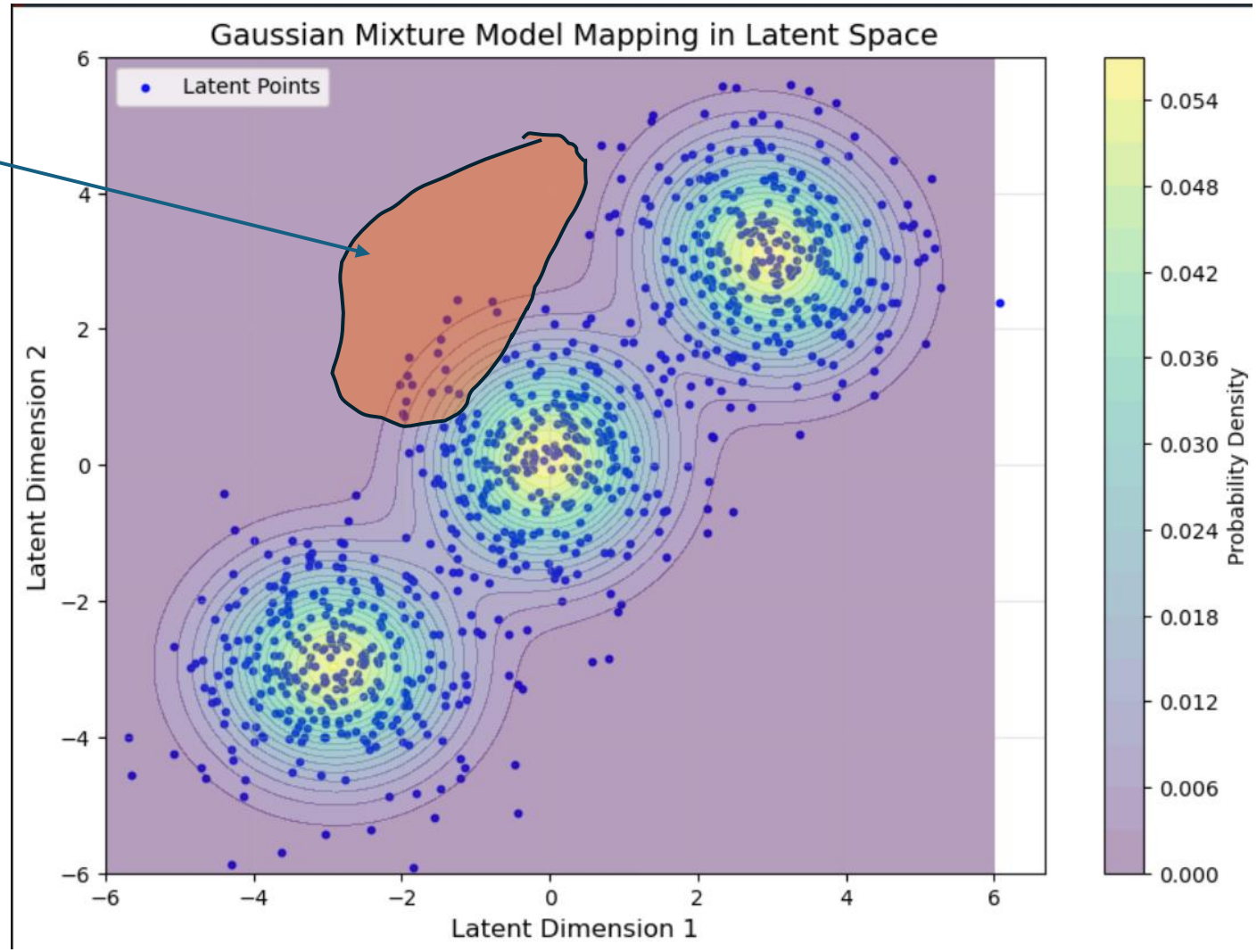
$$\mathrm{NLL}(\mathbf{x}) = -\log\left(p(\mathbf{x})\right).$$

The closer to –inf the closer we are in to distribution and if NLL>>1 we are very far from distribution



Robustness enhancement using GMM

The datapoints you want to predict lie here



Bias your NN by taking those datapoints with high NLL and low error inside the red region and add them many times into your training set.