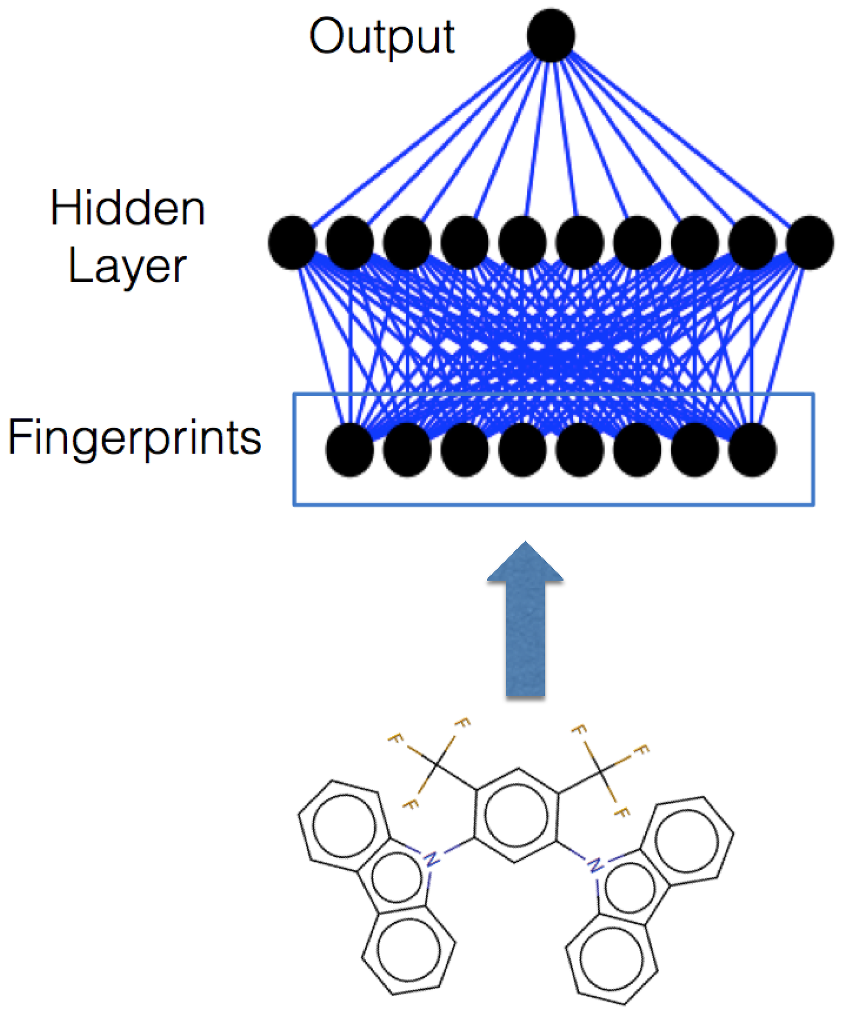
Auto-Encoding Variational Bayes

Paweł Budzianowski, Thomas F. W. Nicholson,

Problem Definition

- Input can be any size or shape
- Hard to turn into fixed-length vector
- In our case, graphs represent molecules
- Applications to photovoltaics, organic LEDS, flow batteries and pharmaceuticals

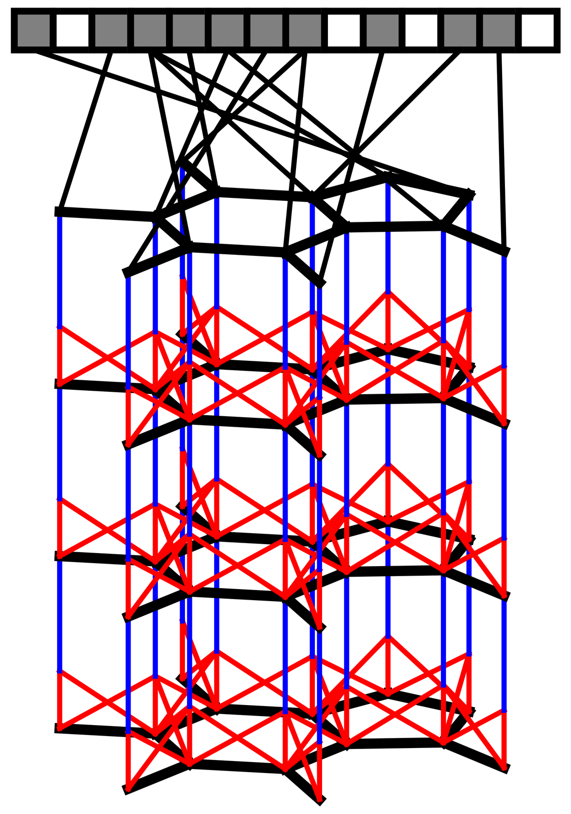


SGVB

- Maps variable-sized molecular graph to fixed-length binary vector
- Binary features indicate presence of substructures

Can be efficiently computed using local operations:

- At each layer, hash the features of each atom and its neighbors/bonds
- More layers correspond to increasing radius of substructures
- Interpret each hash as integer and set that entry to one



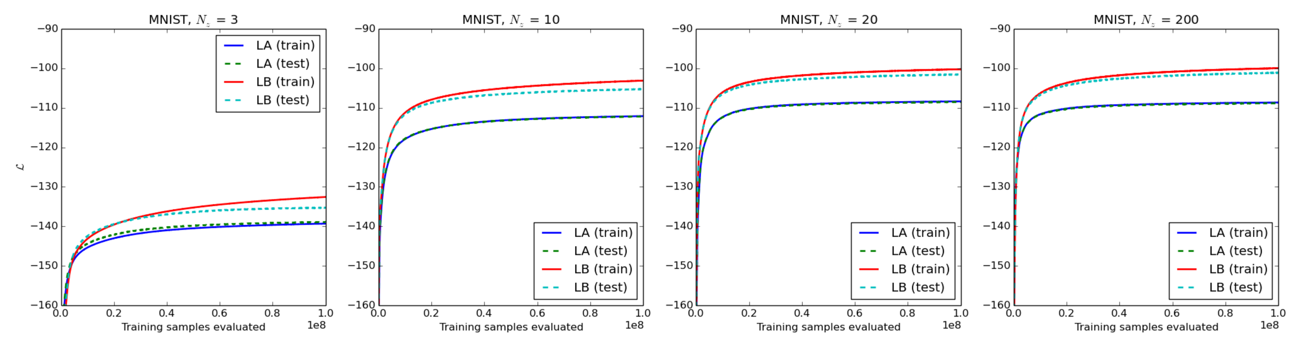
AEVB

Generic estimator versus default SGVB one

In the case of the non-Gaussian distributions, it is often impossible to obtain closed-form expression for the KL-divergence term which also requires estimation by sampling. This yields more generic estimator of the form:

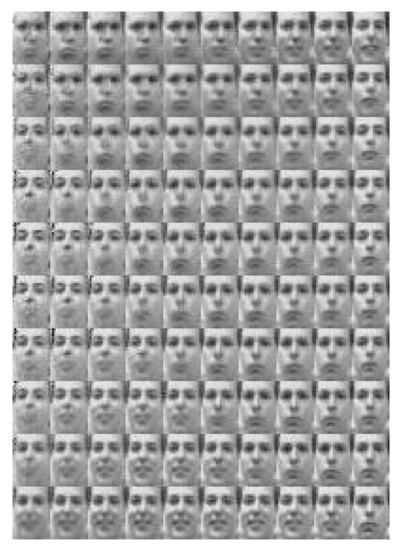
$$\widetilde{\mathcal{L}}^A(\boldsymbol{\theta},\boldsymbol{\phi};\mathbf{x}^{(i)}) = \frac{1}{L}\sum_{l=1}^{L}\left(\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)},\mathbf{z}^{(i,l)}) - \log q_{\boldsymbol{\phi}}(\mathbf{z}^{(i,l)}|\mathbf{x}^{(i)})\right).$$

We decided that it will be informative to compare the performance of both estimators using only one sample i.e. $L = 1$.



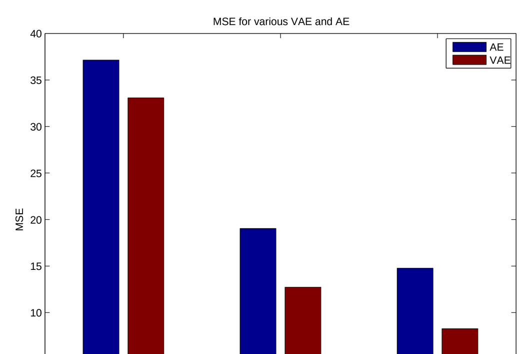
Visualisation of learned manifolds

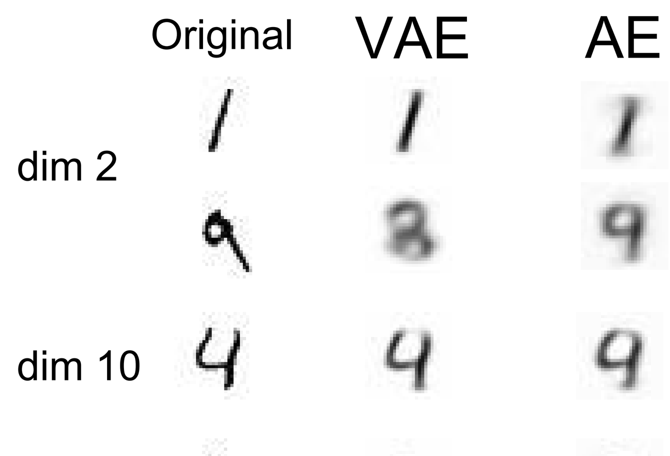
It is possible to observe what the encoder learnt during training if we choose a low-dimensional latent space e.g. $2D$. The linearly spaced grid of coordinates over the unit square is mapped through the inverse CDF of the Gaussian to obtain the value of $\mathbf{z}$ which can be used to sample from $p_\theta(\mathbf{x}|\mathbf{z})$ with the estimated parameters $\boldsymbol{\theta}$.



Baysian: is it really all that?

Compare reconstruction to Auto-encoder





Full VB

Possible to perform full VB on parameters:

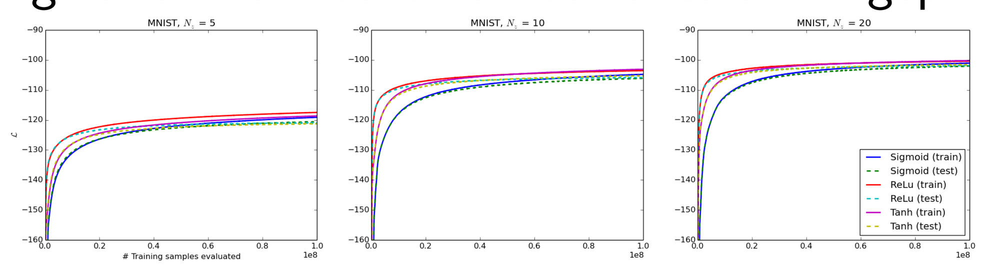
$$\mathcal{L}(\phi,\mathbf{X}) = \int q_\phi(\theta)(log\ p_\theta(X) + log\ p_\alpha(\theta) - log\ q_\phi(\theta))d\theta$$

Can once again use Monte Carlo estimate to approximate, and differentiate to perform SGVI. Yields a distribution over parameters rather than a point estimate.

Implementation showed a decrease of variational lower bound, but no evidence of learning, possibly due to strict Gaussian assumptions of variational approximate posteriors.

Large random weights give similar behavior to circular fingerprints:

- Different activation functions.

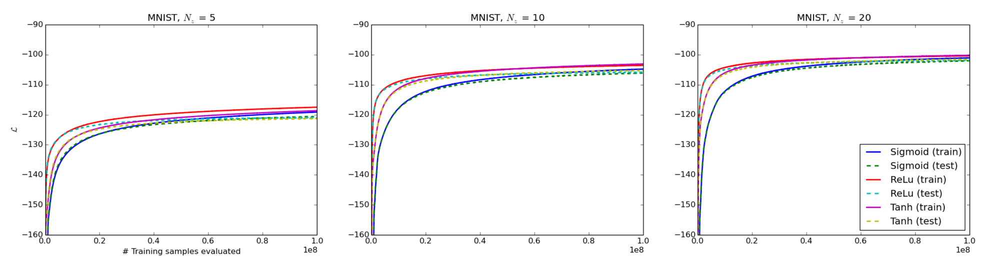


Small random weights already much better than circular fingerprints! Can do even better by optimizing for given task.

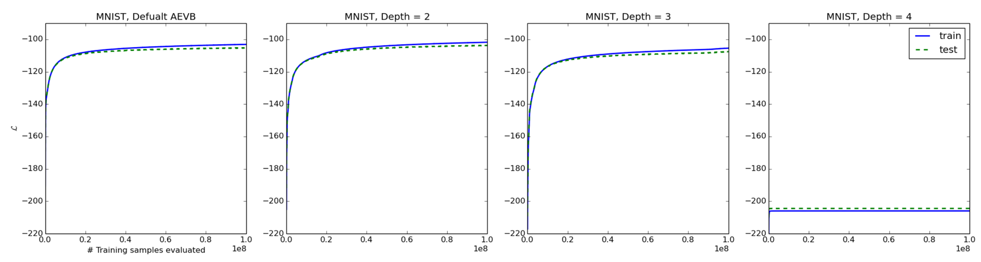
Architecture experiments

We examined various changes to the original architecture of the auto-encoder to test the robustness and flexibility of the model which lead to improvement in terms of optimising the lower bound and computational efficiency.

- Different activation functions.



- Increasing the depth of the encoder.



Future works

I. Scheduled training of VAEB [2].
II. Direct parameterization of differentiable transform
III.

References