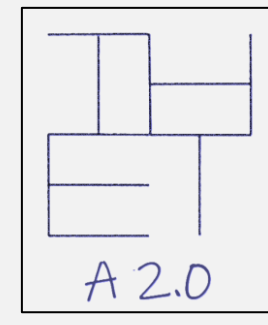


Prada versus Versace – Deep-Learning-based SNP imputation on the genomewide scale

Joanna Szyda



Joanna.Szyda@pwr.edu.pl



THETA 2.0 Biostatistics Group, Department of Biomedical Engineering, Wroclaw University of Science and Technology

Conclusions

- Extremely simplified models surprisingly provide good imputation metrics
- Subset (500 SNPs) $\neq$ full chromosome: models successful on subsets do not necessarily scale to full chromosome, HPO results differ, HPO on full data - problematic

Data

- BTA28 $\rightarrow$ 1000 Bull Genomes
- 742 HF bulls
- 297 605 SNPs

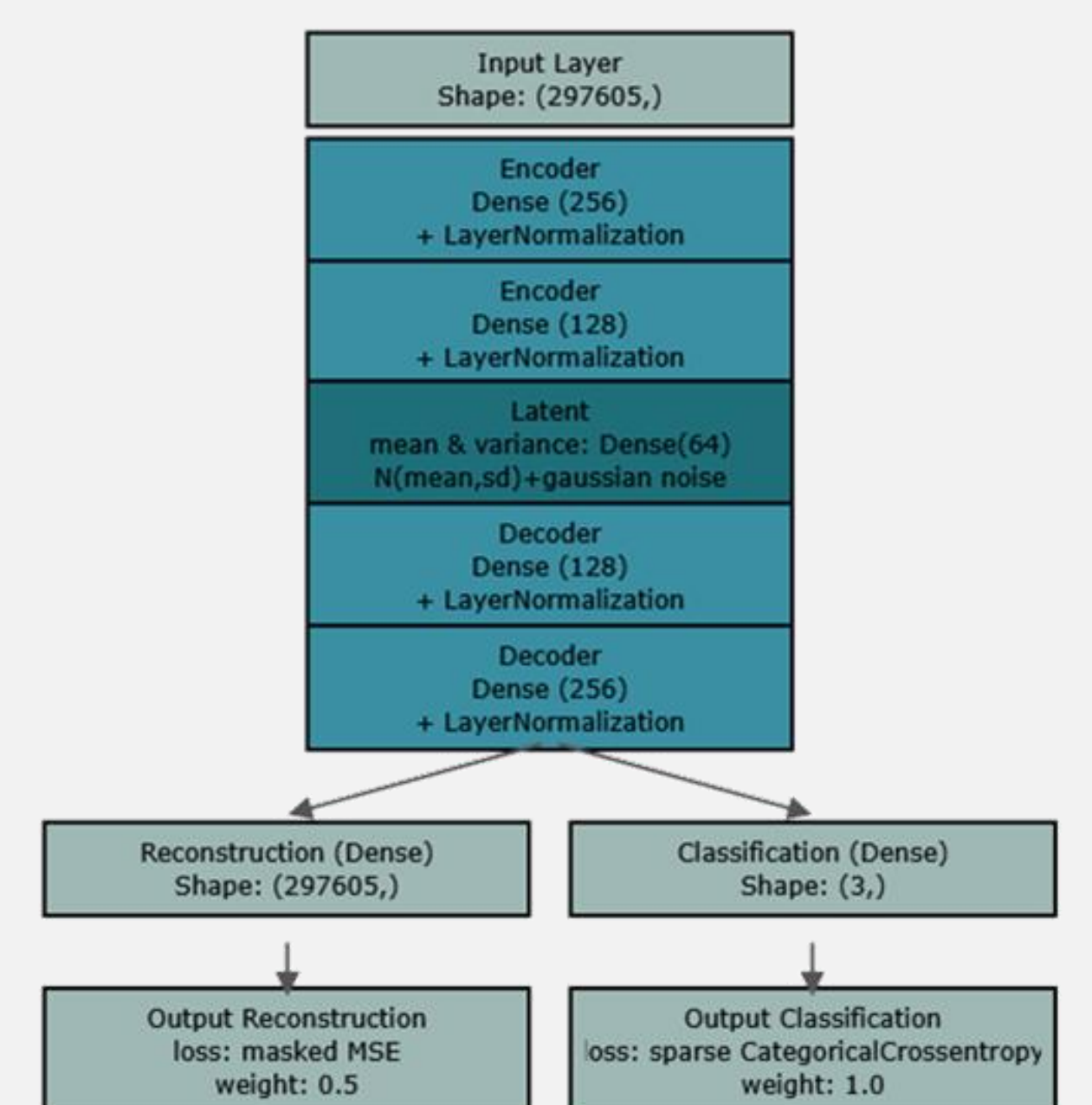
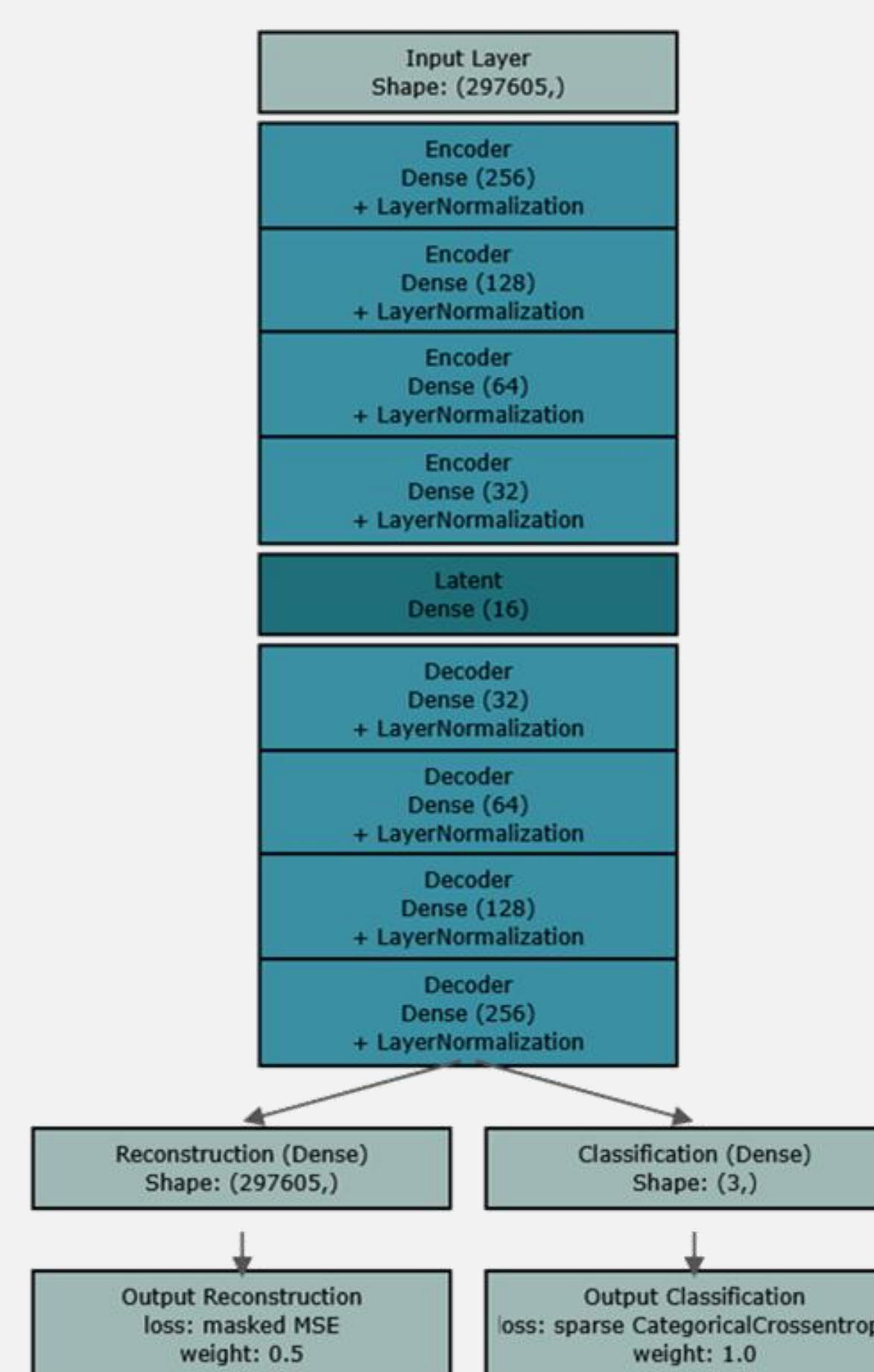
- 186 HF bulls
- 296 822 missing SNPs
- 99.7 % missing

- each bull has the same missing
- SNPs **not** on Illumina 50K chip
- missing streak 5 – 8164 SNPs

AutoEncoders

Näive linear (🧠) LAE

VAE

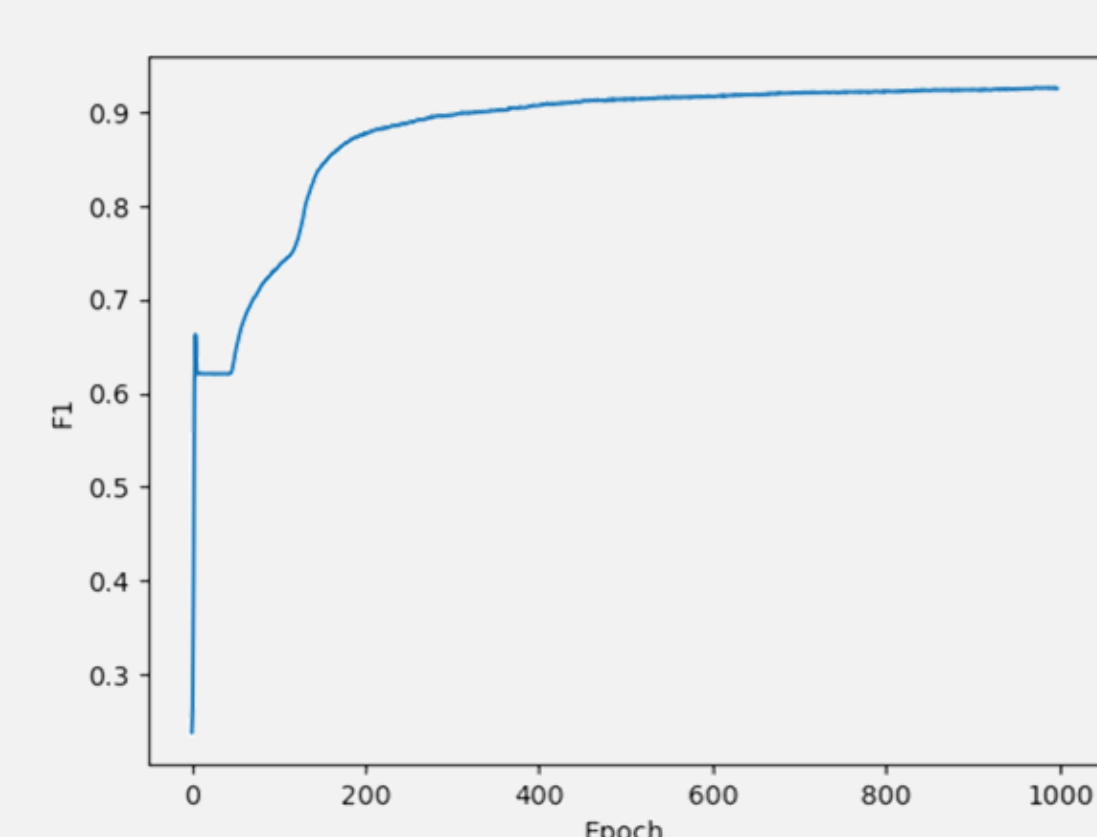
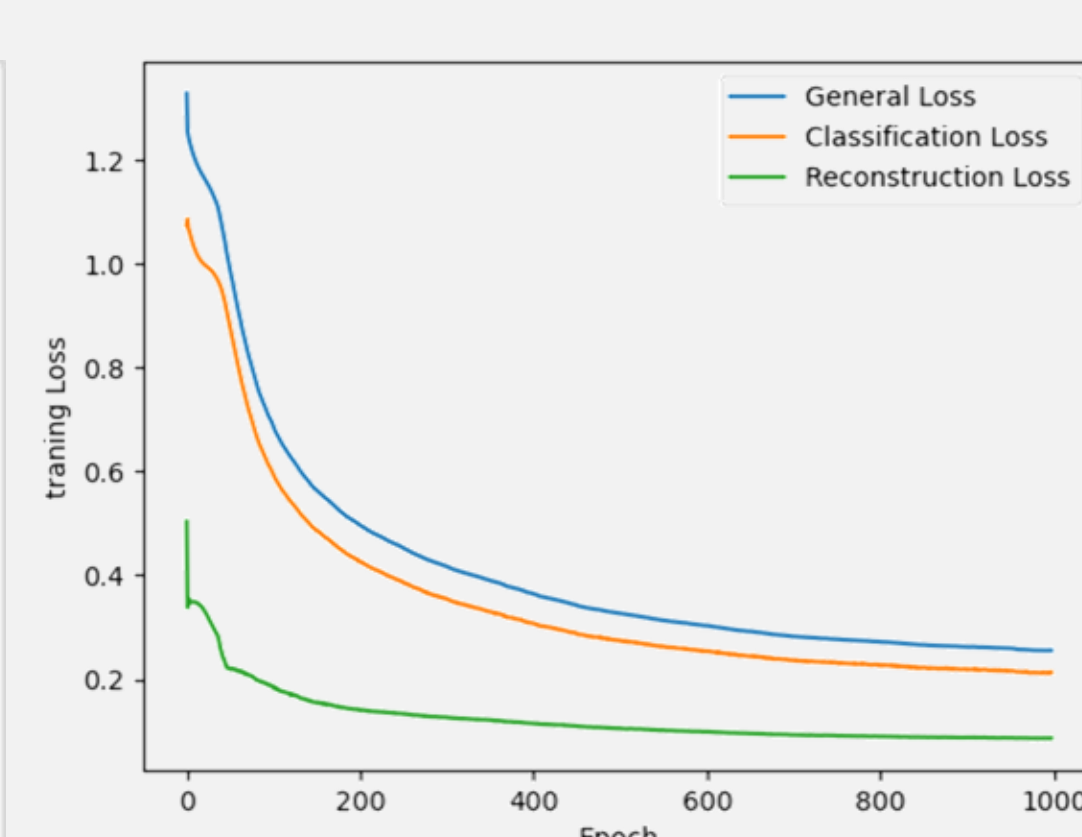
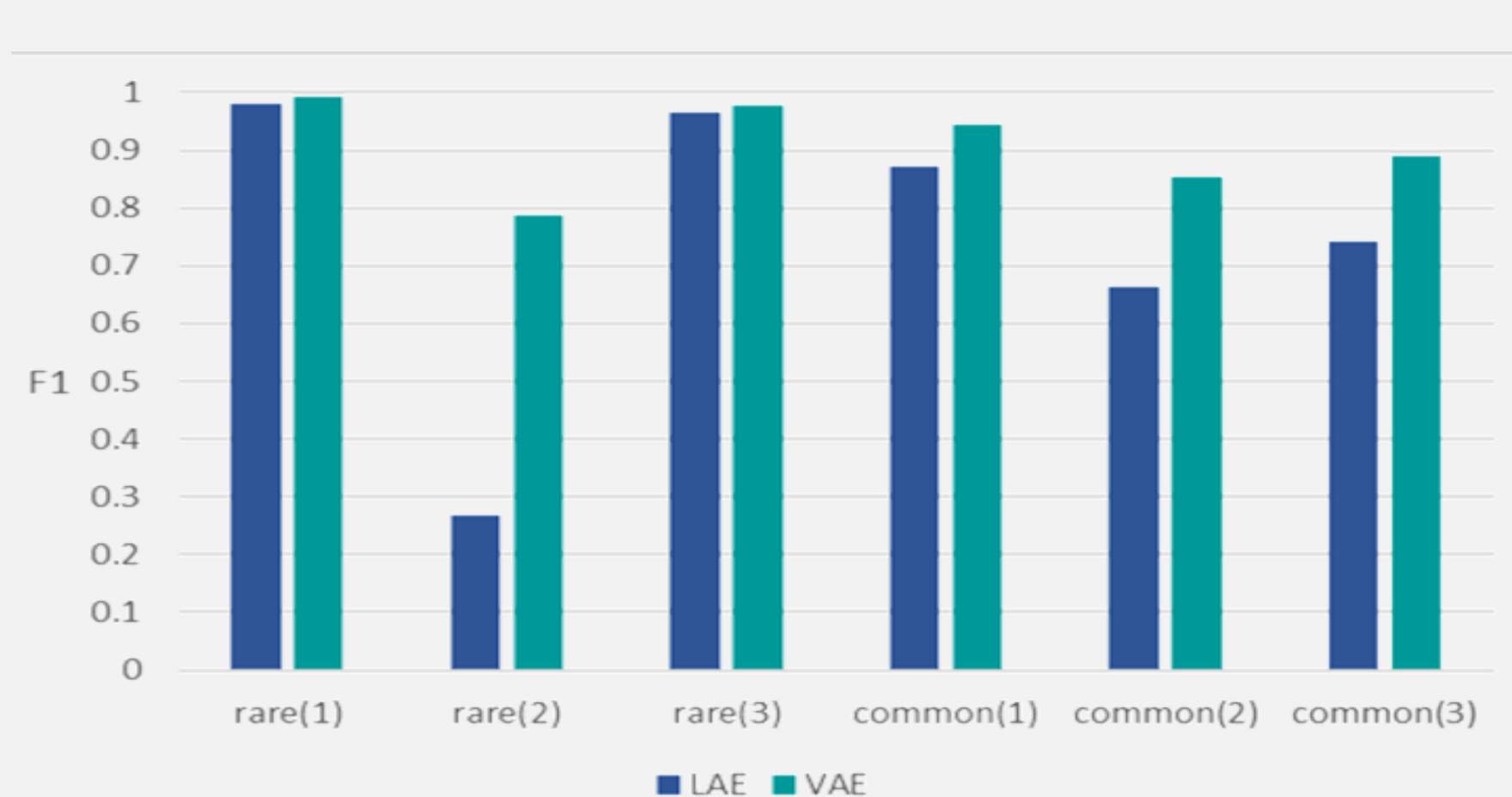
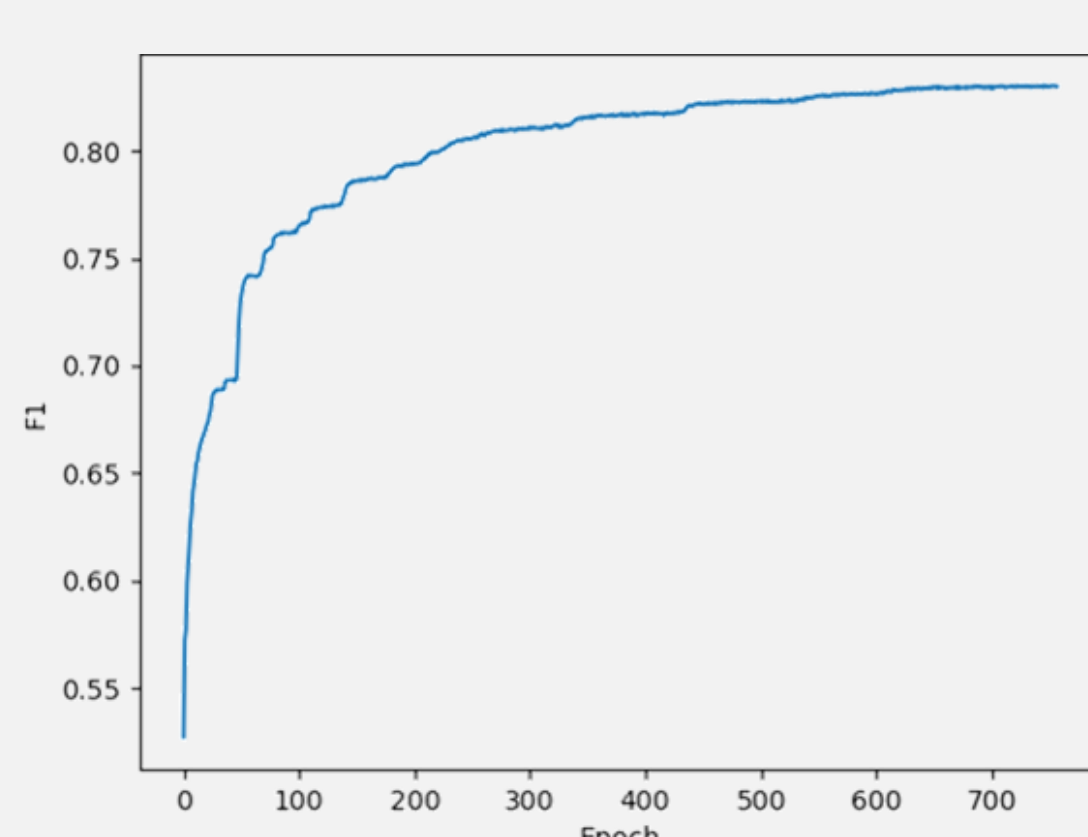
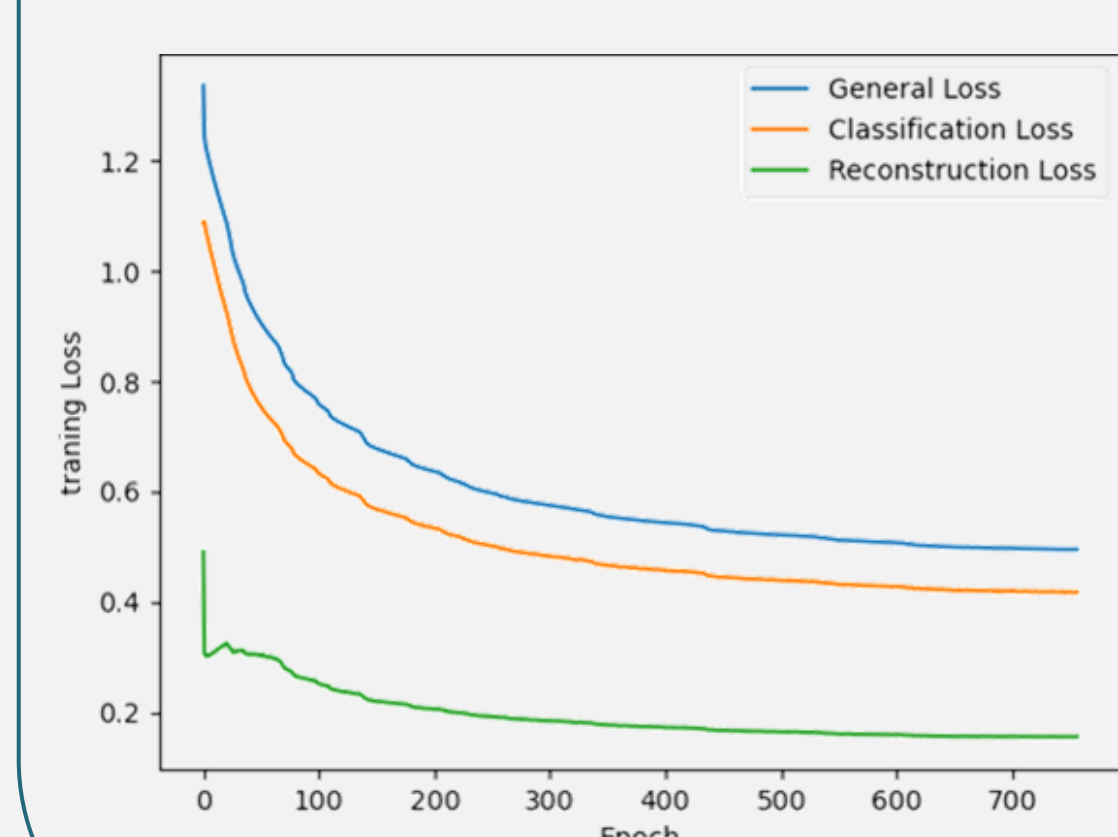


- 1000 epochs with early stopping, Adam
- TensorFlow with Keras
- 32 CPU cores, Nvidia V100, 190G RAM

Results

LAE

VAE



More work

- Positional encoding in Dense based AEs
- Partial nonlinearity $\rightarrow$ nonsymmetric AE
- LLM
- two-stage imputation