Development and Tuning of Auto-Encoder Approaches for Cancer Prognosis from Gene Expression Data

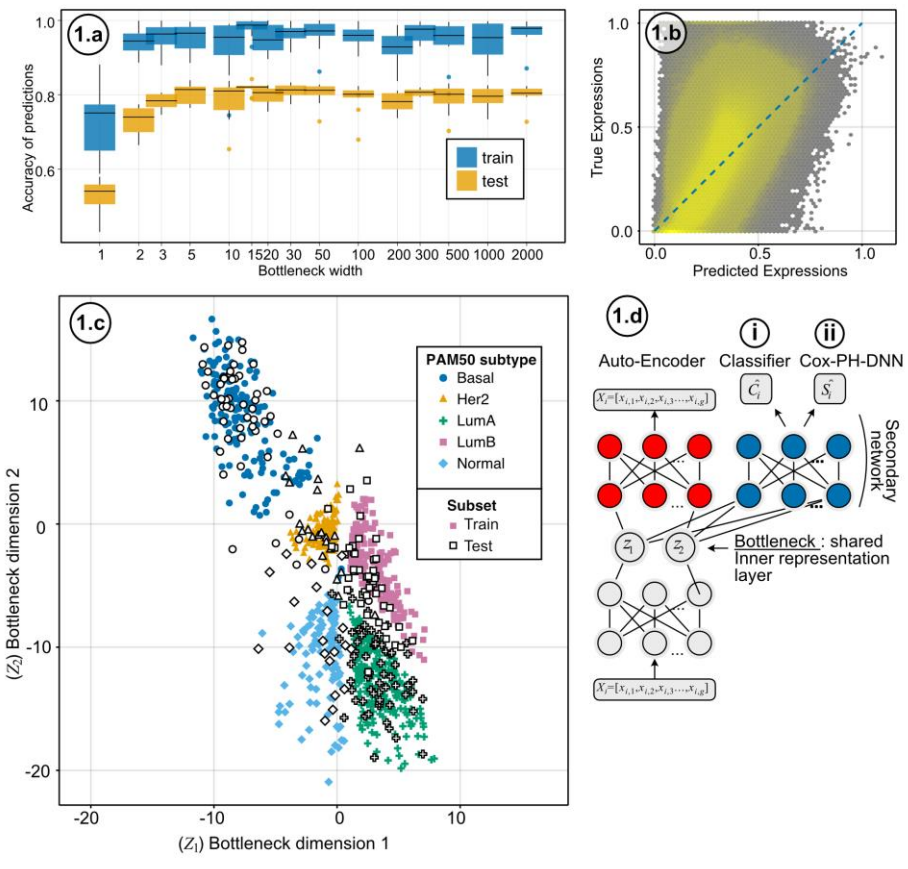
Léonard Sauvé[1] and Sébastien Lemieux[1,2]

1. Institute for Research in Immunology and Cancer (IRIC), Université de Montréal.

2. Department of Biochemistry and Molecular Medicine, Université de Montréal.

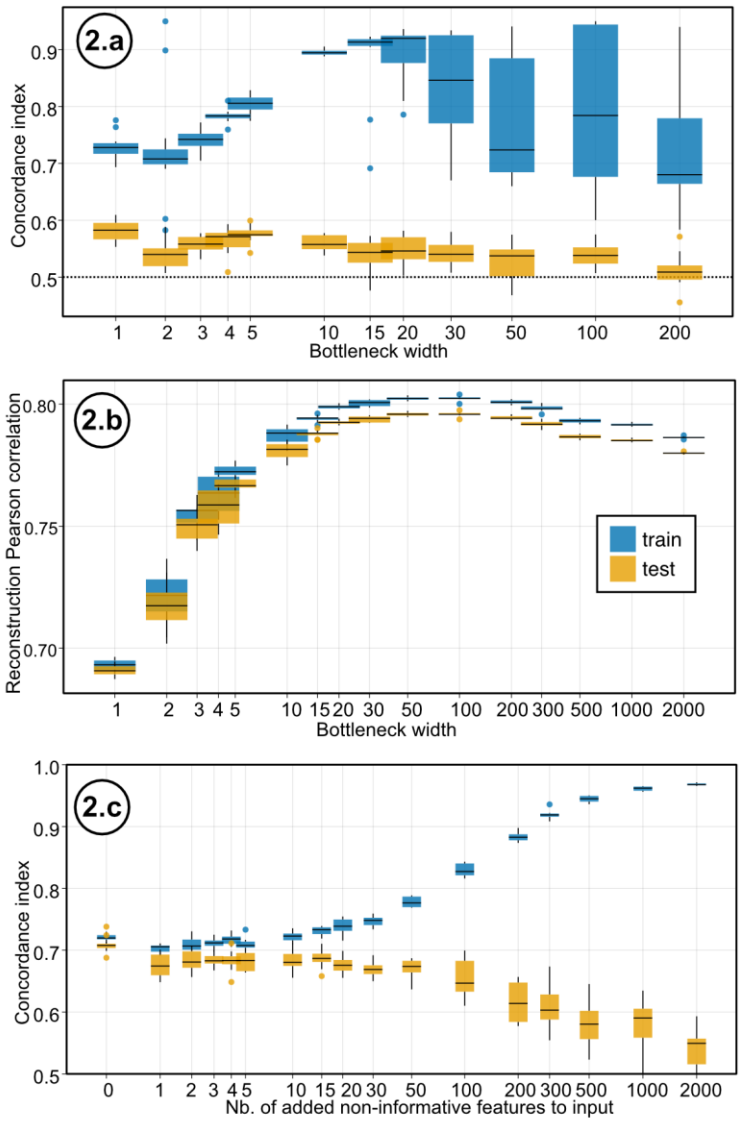
Introduction: Machine-learning mediated cancer prognosis from gene expression profiling data could lead to more accurate risk stratification and treatment selection, and better survival. Gene expression data has a very high dimensionality and easily leads to overfitting for survival prediction. Dimensionality reduction is therefore necessary. Here, we investigate an alternative strategy to the usual combination of deep neural networks[1-4] (DNN), using dual objectives DNN. In this study, we investigate models that share inner representations (bottleneck) and optimize two objectives in alternance. Consequently, we built and tested two variants of this approach as described in (**1.d**); (**i**) an Auto-Encoder Classifier-Deep-Neural network (AE+C) and (**ii**) an Auto-Encoder Cox-Proportional-Hazard Deep-Neural-Network (AE+CPH). Both these architectures provide dimensionality reductions, an auto-encoder, and a secondary predictor. For performance comparisons, a baseline survival Cox-PH-DNN (CPH) was trained on clinical features only (age & cancer stage) and evaluated with a c-index of 0.715. All empirical data presented here was generated with 10 replicates in 5-fold cross-validation. Test and train set performances are reported to investigate overfitting.

Data: TCGA breast cancer dataset (n=1050) patient profiles comprising RNA-seq gene expression profiles & clinical/demographic features (age and cancer stage). Samples are also annotated according to their PAM50 molecular subtypes: Her2 enriched (n=133), Basal-like (n=218), Luminal A (n=296), Luminal B (n=242), normal-like (n=134), unknown (n=27). Gene expression values in FPKM were min-max normalized and all genes that displayed a variance under 0.02 were rejected which resulted in 11,802 gene expression features.



Results: To verify AE+C's ability to reconstruct input profiles and perform molecular subtype classification, we tested (**1.a**) AE+C with increasing bottleneck width on the TCGA breast cancer dataset by 5-fold cross-validation. Accuracy of predictions (correct / total) were collected, and the experiment was run 10 times. Accuracy of the train and test set are reported to investigate overfitting. We noticed increasing accuracies of prediction with bottleneck width up until 5 nodes, where the models ceased to perform better on the test nor training set. When the bottleneck layer was set to only two nodes, trained models can reconstruct the input profiles as well as generate 2-dimensional representations of the data. (**1.b**) Reconstruction of the input gene expression profiles is represented by a 2D histogram with a Pearson correlation coefficient of 0.67. (**1.c**) 2-dimensional inner representations of AE+C driven by the reconstruction of the input gene expression profiles and PAM50 molecular subtypes classification. Train (filled) and test (empty) samples are represented with markers corresponding to their PAM50 subtype. Prediction accuracy of the PAM50 subgroups was evaluated at 77.5% on the test set and 98.4% on training set. Prediction accuracy of a regular classifier DNN trained on the gene expression profile was evaluated to 83.1%. This performance is higher than the tested

model, but expected since we used a 2D representation. As tested in **1.a**, this classification performance can be increased with a larger bottleneck size. In the next section we discuss experiments involving the second setup for survival prediction.



(**2.a**) Performance (c-index) of the AE+CPH with respect to bottleneck width. We notice that the best performing model (c-index=0.59) has only one node in its bottleneck layer and performs worse than the benchmark model using only CPH with clinical features (c=0.715), signifying that the model is not performing as expected, probably due to overfitting. We observe that overfitting is occurring at every bottleneck size but is at its peak at around 15 nodes. To identify the cause of this under-performance, we dissected the AE+CPH model in its two individual components: the Auto-Encoder and the CPH. First, we looked at the effect of bottleneck width on the Auto-Encoder reconstruction correlation. (**2.b**) shows Auto-Encoder Pearson correlation between input and reconstruction vs. the bottleneck width. This time the Auto-encoder was trained without a secondary network. We observe that the model displayed a reconstruction correlation of 0.67 with only 1 bottleneck node, which is surprisingly high. This is suspected to be the recognition of the mean expression by each gene. As the bottleneck width increases, correlations go up to 0.77 with 20 nodes, which indicates that more information is learned about each gene expression pattern than the mean level only. We also notice very little overfitting between train and test correlations. Additionally, it seems that the training correlations drop slightly with over 100 nodes, potentially indicating that optimization requires more steps for higher number of learned parameters. Suggesting that the Auto-Encoder bottleneck width can be tuned for better reconstruction. Next, we wanted to investigate the impact of dimensionality on the CPH model, taken separately. We specifically wanted to verify that the CPH model is still able to retrieve informative features among an input with additional, non-informative nodes. (**2.c**) The CPH performance trained on clinical features (age, stage) with an increasing number of extra noisy input features (uniform [0,1]). We can see that the test c-index falls below 0.72 with only a single non-informative input node and stays stable until 15 extra input features are added, after which it consistently drops back to 0.5 due to overfitting.

Conclusion: In this study, we were able to propose AE-based architectures that optimize dual objectives simultaneously. These models provide reconstruction of the input via the AE and use the bottleneck layer (a dimensional reduction of the input gene expression profiles) as input to predictive methods for both classification and survival regression problems. Using the TCGA breast cancer dataset, the first type of network seems to display good performance for PAM50 classification. Although we only tested AE+C and AE+CPH combinations, we propose that this methodology can be applied to other AE+X networks. However, for survival regression AE+CPH still underperforms vs the standard approach due to overfitting. To alleviate these issues, we intend to explore the potential benefits of using a variational auto-encoder and to expand the methodology to other cancer datasets.

References

1. Ching, T., Zhu, X., & Garmire, L. X. (2018). Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLOS Computational Biology*, *14*(4), e1006076. https://doi.org/10.1371/journal.pcbi.1006076
2. Faraggi, D., & Simon, R. (1995). A neural network model for survival data. *Statistics in Medicine*, *14*(1), 73–82. https://doi.org/10.1002/sim.4780140108
3. Jiang, L., Xu, C., Bai, Y., Liu, A., Gong, Y., Wang, Y.-P., & Deng, H.-W. (2023). AUTOSurv: Interpretable Deep Learning Framework for Cancer Survival Analysis Incorporating Clinical and Multi-omics Data. *Research Square*, rs.3.rs-2486756. https://doi.org/10.21203/rs.3.rs-2486756/v1
4. Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, *18*(1), 24. https://doi.org/10.1186/s12874-018-0482-1