Domen Mohorčič
Tbilisijska ulica 30, 1000 Ljubljana, Slovenija
Study programme: Data Science Masters in Computer and Information Science
Enrollment number: 63180210

Committee for Student Affairs
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Večna pot 113, 1000 Ljubljana

The master's thesis topic proposal
Candidate: Domen Mohorčič

I, Domen Mohorčič, a student of the 2nd cycle study programme at the Faculty of computer and information science, am submitting a thesis topic proposal to be considered by the Committee for Student Affairs with the following title:

Slovenian: **Učenje s prenosom pri napovedovanju fenotipa na majhnih podatkovnih naborih o izraženosti genov**
English: **Transfer learning for phenotype prediction from small gene expression data sets**

This topic was already approved last year: ***NO***

I declare that the mentors listed below have approved the submission of the thesis topic proposal described in the remainder of this document.

I would like to write the thesis in English as I am part of the Data Science program, where English is our primary learning language.

I propose the following mentor:

Name and surname, title: Blaž Zupan, prof. dr.
Institution: Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
E-mail: blaz.zupan@fri.uni-lj.si

Ljubljana, 22. maj 2023.

Proposal of the masters thesis topic

1 The narrow field of the thesis topic

English: bioinformatics, machine learning transfer learning

2 Key words

English: gene expression data, data integration, transfer learning, data embedding, knowledge-based data analysis

3 Detailed thesis proposal

Past approvements of the proposed thesis topic:
The proposed thesis has not been submitted nor approved in previous years.

3.1 Introduction and problem formulation

Understanding biological data is an integral part of medical research. But usually, we deal with vast amounts of data, more than a single person or even a team can manage. Analyzing it by hand is nearly impossible, and that is where machine learning comes in. Machine learning in bioinformatics is used to analyze and extract results from data and can offer helpful insight [1]. Its prominent uses are in genomics, where researchers use it to identify candidate genes behind certain diseases or unique biological properties, and proteomics for protein folding tasks [2].

In many cases, machine learning is used for prediction purposes [3]. An example would be the likelihood of a patient developing cancer. We have many large databases, suitable for such learning. The Cancer Genome Atlas [4] is a cancer genomics program where more than 20 thousand samples across 33 cancer types are molecularly characterized. Gene Expression Omnibus (GEO) database [5] is an international public repository that contains microarray, next-generation sequencing, and other forms of genomics data submitted by the community. The GEO database has around 5.4 million samples which are organized into more than 4 thousand data sets. But we also have many smaller data sets where we have a few instances and a large number of attributes. GEO database has around 1800

data sets with less than 10 instances. Learning about those comes with higher uncertainty as there is a lack of samples.

A problem of machine learning in bioinformatics is explainability [6]. Deep models [7] such as neural networks or deep belief models [8] are usually very hard to interpret and require large amounts of data to train but are very successful [9, 10]. Simpler models such as linear or logistic regression and decision trees are preferred due to their simplicity, but even their interpretability fails when dealing with thousands of attributes.

The problem we propose to solve in this thesis is learning from small data sets and producing accurate and explainable models. The smaller data sets are difficult for explainable models to learn from due to high uncertainty. But integrating them, producing latent and generalizable features with autoencoders, and using transfer learning to model smaller data sets could result in accurate and explainable models.

3.2 Related work

Autoencoders [11] are neural networks used to produce efficient embedding of the provided data. They are composed of two models: an encoder, which encodes the data into embedding, and a decoder, which can decode the embedding. They have been successfully used for anomaly detection [12], image denoising [13], and dimensionality reduction.

Transfer learning [14] is a machine learning problem of gaining knowledge in one domain and applying it to another. It is a popular approach to use pre-trained models as a starting point, especially in image classification [15] and natural language processing [16], where there are vast amounts of data.

Hanson et. at. in 2019 [17] used transfer learning for predicting molecular recognition features (MoRFs) from the identification of intrinsically disordered protein regions. They used the model trained on disordered protein regions for protein representation and build new models on top of them. They showed that this new model performs better than a model trained only on MoRFs data. One drawback is that both tasks are highly correlated, and their approach is not generalizable.

Rao et. al. in 2019 [18] introduced tasks assessing protein embeddings (TAPE), which are a set of 5 biologically relevant semi-supervised learning tasks and are used in evaluating protein embeddings. Their model of learned protein representation was then later used by Ekvall et. al. in 2022 [19], where they predict mass spectrometry intensities of different proteins. They used transformer architecture and obtained better results than deep models, trained on raw data, which is encouraging.

In genomics, transfer learning has been used in the reconstruction of gene regulatory

networks (GRNs). Mignone et. al. in 2019 [20] used knowledge from the GRN of one organism and used it to reconstruct the GRN of another organism. GRNs help understand the regulatory mechanisms involved in diseases. Their method works without negative samples, which is usually the case with regulatory connections.

Liu et. al. in 2022 [21] used neural networks for evaluating the predictive power of grouped genome data. They then discarded the least accurate models and used the output of the remaining models as input to another neural network, which tackles the disease risk prediction task. Since inputs are measures of predictive power and only the best were kept, the authors argue that their final deep neural network framework is biologically interpretable.

Chen et. al. in 2022 [22] used transfer learning to predict cancer drug responses to RNA sequences. They used denoising autoencoders to learn a compact representation of the gene expressions in certain cells and to learn the correlation between drug responses and gene expressions. Another autoencoder was trained to extract features from single-cell RNA sequences. Transfer learning is done using another deep model to adapt extracted gene features to have similar distributions. The final prediction is done by taking a single-cell RNA sequence along with adapted features and returning whether a certain cell is sensitive or resistant to a certain drug.

3.3 Expected contributions

We will explore whether transfer learning can be used to enhance learning on a collection of small gene expression data sets and whether explainable models can be built. The contribution of the proposed thesis will be a database of small gene expression data sets, a framework for training and explaining models, and a framework for qualitative assessment of used methods.

3.4 Methodology

We propose to implement the following work plan:

Collecting relevant data sets. Our target is at least 50 data sets. The number of samples per data set would ideally be between 10 and 50. Initially, we will focus only on binary classification tasks, but we plan to later include tasks with a higher number of classes as well.

Data preprocessing and cleaning. We will inspect the data and check whether their attributes are comparable between the data sets. We expect data sets to have diffe-

rent attributes and we will focus only on 978 landmark genes described in a profiling method called L1000 [23]. We will use them because the authors demonstrated they are suitable for the inference of the expression levels of 81% of other genes.

Developing a transfer learning framework. The framework will integrate the data sets and train an autoencoder. Training will be done on larger data sets to infer the autoencoder that captures cross-domain patterns and can improve both explainability and accuracy on smaller data sets.

Testing the encodings. We will test the encodings of smaller data sets by training a classification model such as logistic regression and using the log loss metric and classification accuracy. Encodings will be compared to models, trained on all features, and we expect to get comparable results. We plan to research how the architecture of the autoencoder influences these results. Since genes are phenotypically related, we anticipate that a sparse autoencoder with only local connections will perform the same as a fully connected autoencoder.

Explainability. We will try to explain the models and their predictions from encodings with explainability methods and the domain knowledge of an expert.

3.5 References

[1] P. Larrañaga, B. Calvo, R. Santana, et. al., Machine learning in bioinformatics, Briefings in Bioinformatics 7 (1) (2006) 86–112.

[2] J. Jumper, R. Evans, A. Pritzel, et. al., Highly accurate protein structure prediction with alphafold, Nature 596 (7873) (2021) 583–589.

[3] R. E. Soria-Guerra, R. Nieto-Gomez, D. O. Govea-Alonso, et. al., An overview of bioinformatics tools for epitope prediction: implications on vaccine development, Journal of biomedical informatics 53 (2015) 405–414.

[4] The cancer genome atlas program (2006).
URL `https://www.cancer.gov/tcga`

[5] R. Edgar, M. Domrachev, A. E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, Nucleic Acids Research 30 (1) (2002) 207–210.

[6] H. Han, X. Liu, The challenges of explainable ai in biomedical data science, BMC Bioinformatics 22 (12) (2022) 443.

[7] B. Tang, Z. Pan, K. Yin, et. al., Recent advances of deep learning in bioinformatics and computational biology, Frontiers in Genetics 10 (2019).

[8] G. E. Hinton, S. Osindero, Y.-W. Teh, A Fast Learning Algorithm for Deep Belief Nets, Neural Computation 18 (7) (2006) 1527–1554.

[9] J. Zhou, O. G. Troyanskaya, Predicting effects of noncoding variants with deep learning–based sequence model, Nature methods 12 (10) (2015) 931–934.

[10] J. J. Almagro Armenteros, C. K. Sønderby, S. K. Sønderby, et. al., Deeploc: prediction of protein subcellular localization using deep learning, Bioinformatics 33 (21) (2017) 3387–3395.

[11] R. S. Zemel, G. Hinton, Autoencoders, minimum description length and helmholtz free energy, in: Proceedings of the Neural Information Processing Systems, Citeseer, 1994.

[12] C. Zhou, R. C. Paffenroth, Anomaly detection with robust deep autoencoders, in: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017, pp. 665–674.

[13] L. Gondara, Medical image denoising using convolutional denoising autoencoders, in: 2016 IEEE 16th international conference on data mining workshops (ICDMW), IEEE, 2016, pp. 241–246.

[14] S. Bozinovski, A. Fulgosi, The influence of pattern similarity and transfer learning upon training of a base perceptron b2, in: Proceedings of Symposium Informatica, Vol. 3, 1976, pp. 121–126.

[15] M. Hussain, J. J. Bird, D. R. Faria, A study on cnn transfer learning for image classification, in: UK Workshop on computational Intelligence, Springer, 2018, pp. 191–202.

[16] J. Howard, S. Ruder, Universal language model fine-tuning for text classification (2018).

[17] J. Hanson, T. Litfin, K. Paliwal, et. al., Identifying molecular recognition features in intrinsically disordered regions of proteins by transfer learning, Bioinformatics 36 (4) (2019) 1107–1113.

[18] R. Rao, N. Bhattacharya, N. Thomas, et. al., Evaluating Protein Transfer Learning with TAPE, Curran Associates Inc., Red Hook, NY, USA, 2019.

[19] M. Ekvall, P. Truong, W. Gabriel, et. al., Prosit transformer: A transformer for prediction of ms2 spectrum intensities, Journal of Proteome Research 21 (5) (2022) 1359–1364, pMID: 35413196.

[20] P. Mignone, G. Pio, D. D'Elia, et. al., Exploiting transfer learning for the reconstruction of the human gene regulatory network, Bioinformatics 36 (5) (2019) 1553–1561.

[21] L. Liu, Q. Meng, C. Weng, et. al., Explainable deep transfer learning model for disease risk prediction using high-dimensional genomic data, PLOS Comput. Biol. 18 (7) (2022) e1010328.

[22] J. Chen, X. Wang, A. Ma, et. al., Deep transfer learning of cancer drug responses by integrating bulk and single-cell rna-seq data, Nature Communications 13 (1) (2022) 6494.

[23] A. Subramanian, R. Narayan, S. M. Corsello, et. al., A next generation connectivity map: L1000 platform and the first 1,000,000 profiles, Cell 171 (6) (2017) 1437–1452.e17.