Protein-Protein Interaction Prediction using a Deep Learning Approach based on Autoencoders

Albu Alexandra
Babeș-Bolyai University

WeADL 2022 Workshop

The workshop is organized under the umbrella of WeaMyL, project funded by the EEA and Norway Grants under the number RO-NO-2019-0133.
Contract: No 26/2020.

Working together for a green, competitive and inclusive Europe

Autoencoders

- neural networks formed of an encoder and a decoder
- trained to reconstruct their input

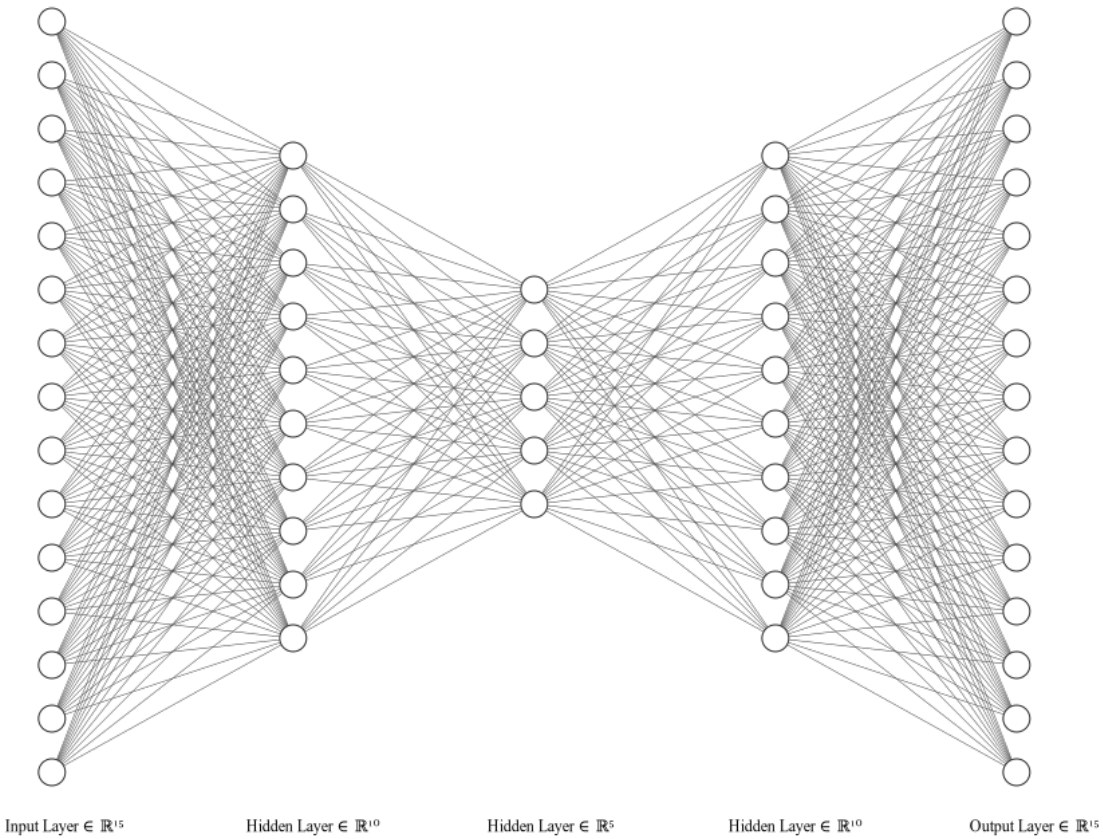
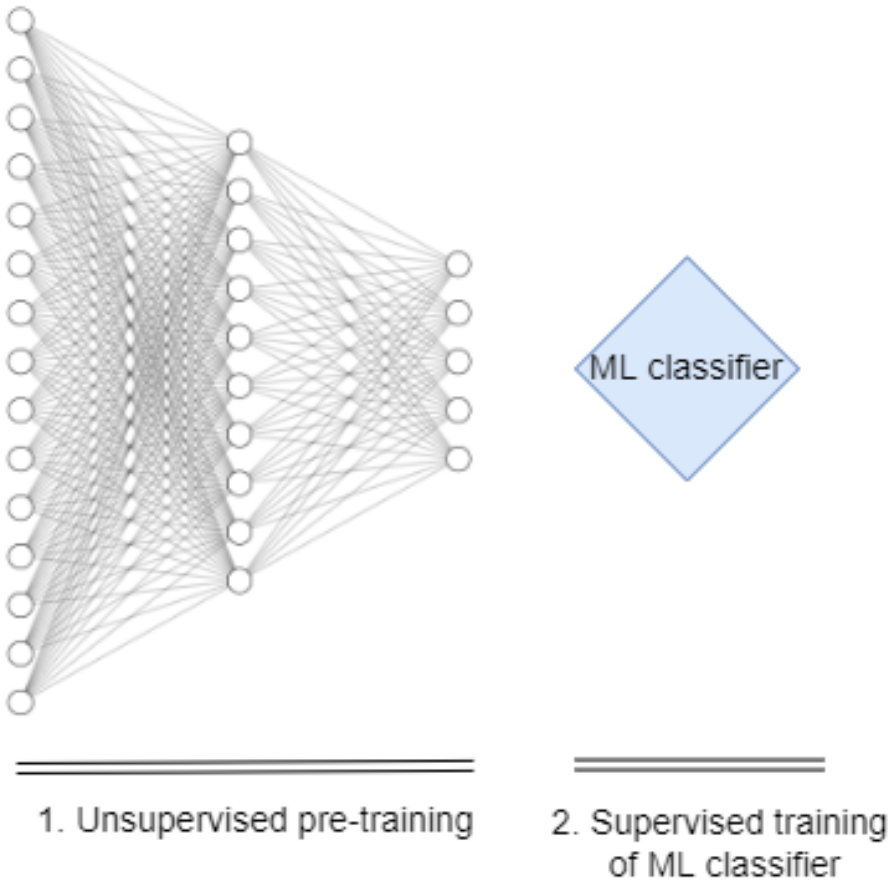


Figure: Autoencoder architecture. Created using https://alexlenail.me/NN-SVG/.

Autoencoders

Autoencoders in classification → three main directions

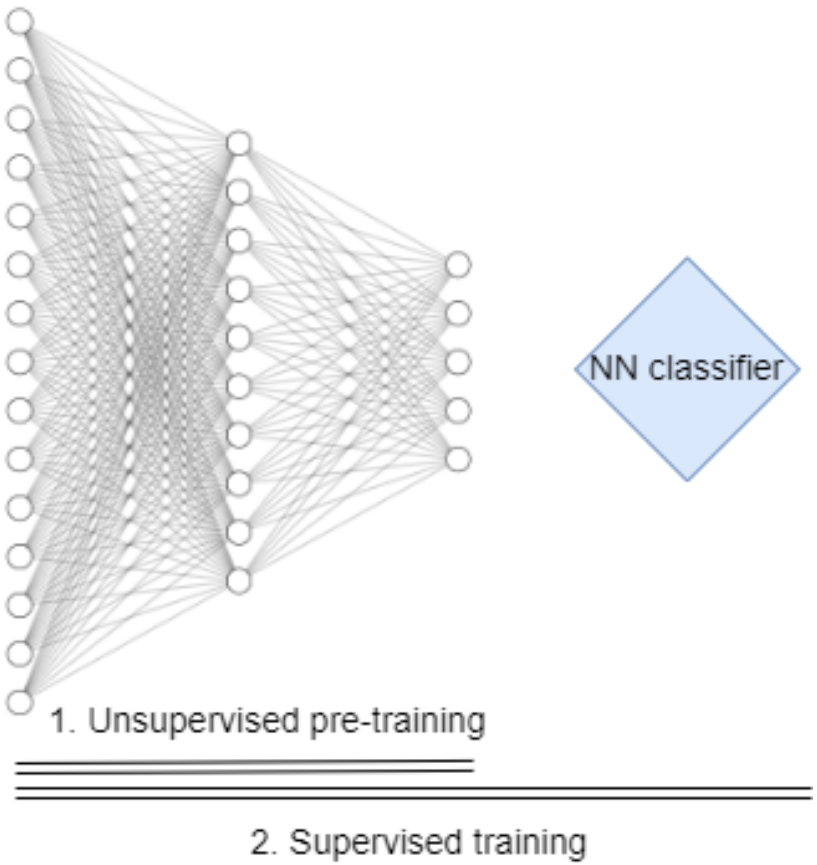
- **Feature extraction**: train a classifier on the learned autoencoder representations → challenge: embedding is performed independently from the classification stage



Autoencoders

Autoencoders in classification $\rightarrow$ three main directions

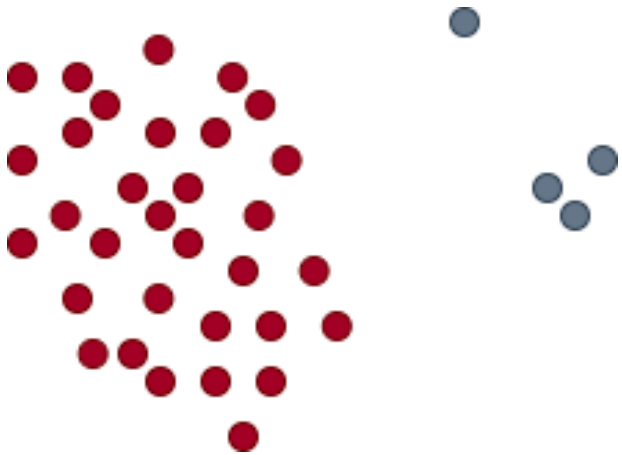
- **Fine-tuning**: fine-tuning the encoder weights together with a neural network classifier



Autoencoders

Autoencoders in classification $\rightarrow$ three main directions

- **Anomaly detection**: train an autoencoder on the majority class, detect outliers



Protein Data Analysis

- Proteins: complex macromolecules
  - important role in vital biological processes in living organisms.
- Central problem: determining protein functions

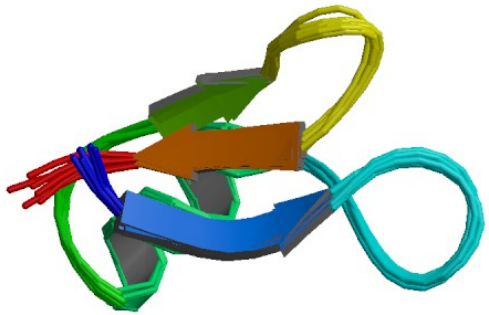


Figure: 3D view of protein 1I2U. Image obtained from the RCSB PDB[1] representing the protein with PDB ID 1I2U.

[1] Berman et al. 2000, *The Protein Data Bank*. https://www.rcsb.org/

Protein-Protein Interaction Prediction

- The majority of proteins perform their roles in complexes
- Experimental determination of PPIs - expensive and prone to false positives $\rightarrow$ computational approaches try to overcome these limitations
- Challenges: labels are noisy, small number of known interacting pairs compared to non-interacting

Literature Review. Sequence-based Protein-protein Interaction Prediction

- Machine Learning methods: SVMs[2], RFs[3], LightGBM[4]
- Ensembles of machine learning classifiers[5]and ensembles of neural networks[6]

---

[2] Guo et al., 2008, *Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences*. Nucleic acids research.

[3] Pan et al., 2010, *Large-Scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features*. Journal of proteome research.

[4] Chen et al., 2019, *Predicting protein-protein interactions through LightGBM with multi-information fusion*. Chemometrics and Intelligent Laboratory Systems.

[5] Chen et al., 2018, *Protein-protein interaction prediction using a hybrid feature representation and a stacked generalization scheme*. BMC Bioinformatics.

[6] Li et al., 2020, *Protein Interaction Network Reconstruction Through Ensemble Deep Learning With Attention Mechanism*. Frontiers in Bioengineering and Biotechnology.

Literature Review. Sequence-based Protein-Protein Interaction Prediction

- Siamese Architectures: $\rightarrow$ capture common charateristics of the two proteins in a pair
  - convolutional architecture[7]
  - residual convolutional recurrent architecture[8]
  - Inception convolutional branch and a bidirectional GRU branch[9]

[7] Hashemifar et al., 2018, *Predicting protein–protein interactions through sequence-based deep learning*. Bioinformatics

[8] Chen et al., 2019, *Multifaceted protein–protein interaction prediction based on siamese residual RCNN*. BMC Bioinformatics

[9] Zhao et al., 2020, *Conjoint feature representation of go and protein sequence for ppi prediction based on an inception rnn attention network*. Molecular Therapy-Nucleic Acids

Autoencoder-based Methods in Protein-Protein Interaction Prediction

- Autoencoders as feature extractors + probabilistic SVMs[10,11]
- Autoencoder pretraining + fine-tuning neural network classifier[12]
- Variational graph autoencoder[13] → learn nodes embeddings using the neighbours in the PPI graph

---

[10] Wang et al., 2017, *Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network.* Molecular BioSystems.

[11] Wang et al., 2018, *Predicting protein interactions using a deep learning method-stacked sparse autoencoder combined with a probabilistic classification vector machine.* Complexity.

[12] Sun et al., 2017, *Sequence-based prediction of protein protein interaction using a deep-learning algorithm*. BMC Bioinformatics.

[13] Yang et al., 2020, *Graph-based prediction of protein-protein interactions with attributed signed graph embedding*. BMC bioinformatics.

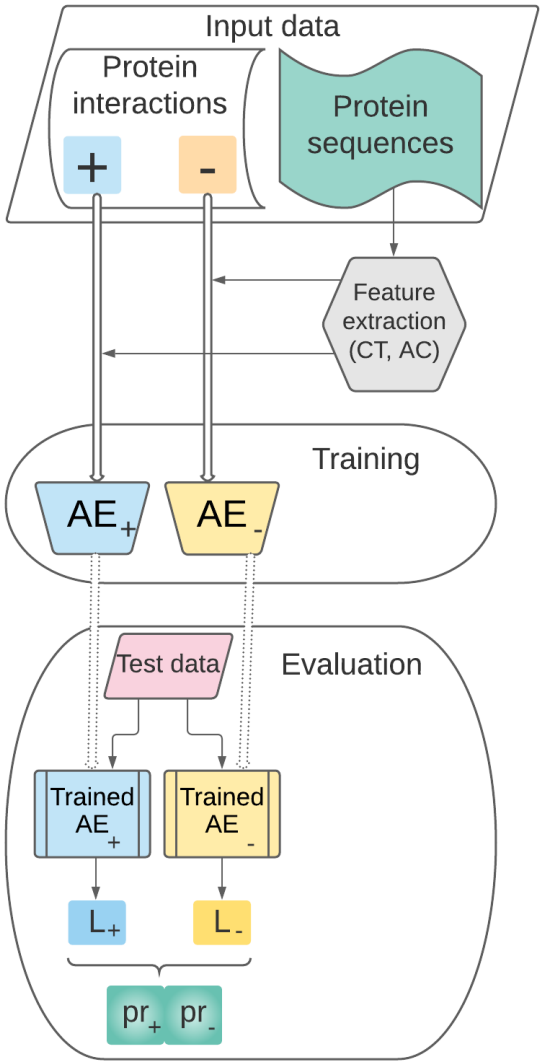
AutoPPI: Binary classification using a pair of autoencoders

Approach:

- two autoencoders trained to reconstruct instances belonging to one class
- classification stage: evaluating which of the two autoencoders is able to better reconstruct the testing data point

Czibula, G., Albu, A.I., Bocicor, M.I. and Chira, C., 2021. *AutoPPI: An Ensemble of Deep Autoencoders for Protein–Protein Interaction Prediction*. Entropy, 23(6), p.643.

AutoPPI: Binary classification using a pair of autoencoders



AutoPPI: Binary classification using a pair of autoencoders

- data samples: pairs of proteins $\rightarrow$ proposed two siamese architectures

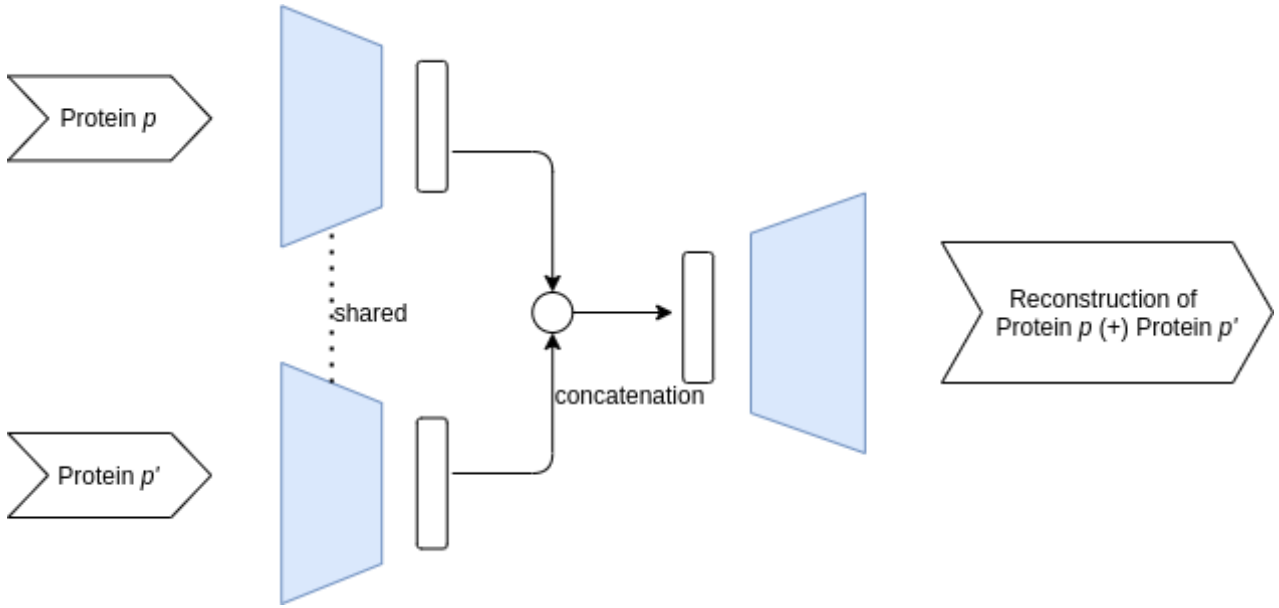


Figure: Siamese-Joint architecture.

AutoPPI: Binary classification model using a pair of autoencoders

- data samples: pairs of proteins $\rightarrow$ proposed two siamese architectures

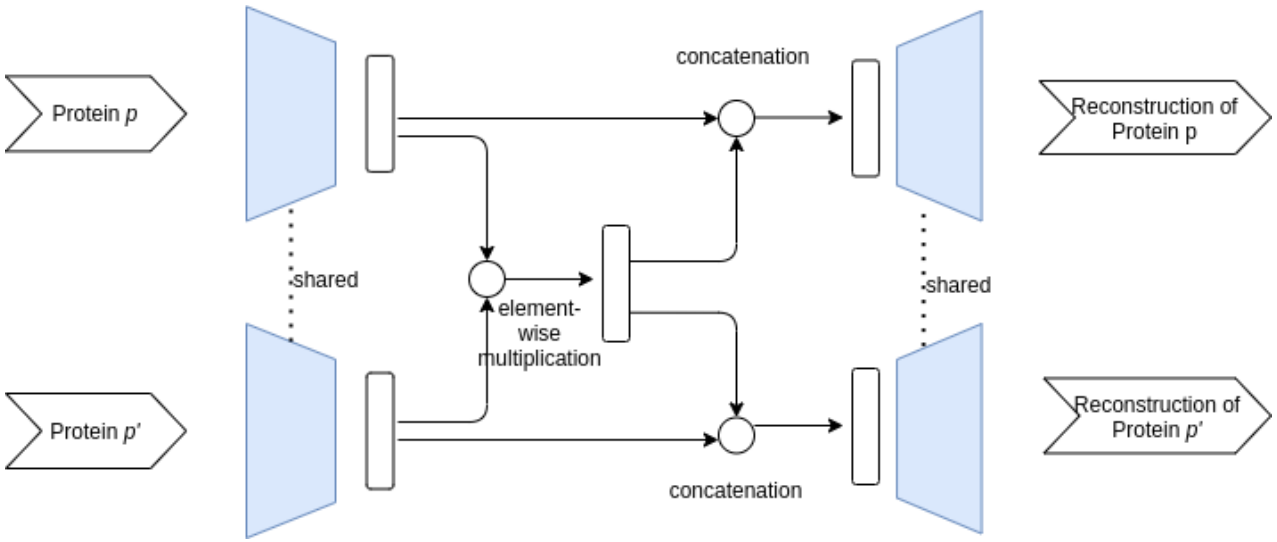


Figure: Siamese-Siamese architecture.

AutoPPI: Binary classification using a pair of autoencoders

- baseline architecture: simple concatenation of protein features (early fusion)

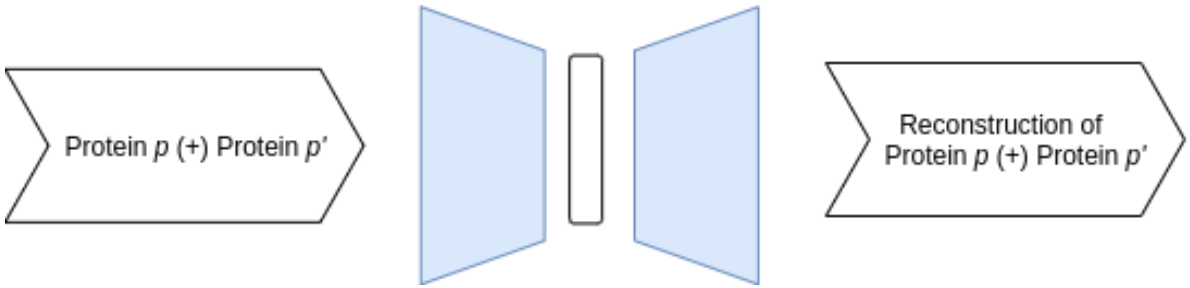


Figure: Joint-Joint architecture.

Data representation

- **Conjoint Triad (CT)** descriptors
  - group amino acids into seven classes based on their physico–chemical properties
  - compute the frequencies of possible triples of amino acid classes
- **Autocovariance (AC)** descriptors
  - define a *lag* variable
  - compute correlations between amino acids situated in the sequence at at most *lag* positions apart

$\rightarrow$ combined representation

Evaluation

Evaluation Methodology

- k-fold cross-validation - same number of folds as the related work on that data set

Evaluation metrics:

- Accuracy
- Precision
- Recall
- F1-score
- Specificity
- Area under the ROC curve

Data sets

- 4 public data sets: one human data set (HPRD) and three multi-species data sets
- Multi-species data sets: obtained by merging three data sets (Caenorhabditis elegans, Escherichia coli and Drosophila melanogaster) - all interactions, proteins filtered using 25% and 1% similarity thresholds

Data set	**Number of positive interactions**	**Number of negative interactions**
HPRD	36,630	36,480
Multi-species	32,959	32,959
Multi-species <0.25	19,458	15,827
Multi-species <0.01	10,747	8,065

Table: Data sets used in the experiments.

Results

Data set	Arch.	Accuracy	F_1-score	Precision	Recall	Specificity	AUC
HPRD	1	0.977 ± 0.0006	0.977 ± 0.0007	0.986 ± 0.0009	0.968 ± 0.001	0.986 ± 0.0009	0.977 ± 0.0006
	2	**0.979 ± 0.0007**	**0.979 ± 0.0007**	0.973 ± 0.0015	**0.985 ± 0.009**	0.973 ± 0.0015	**0.979 ± 0.0007**
	3	0.96 ± 0.0014	0.959 ± 0.0015	**0.992 ± 0.006**	0.928 ± 0.0024	**0.992 ± 0.006**	0.960 ± 0.0014
Multi-species	1	0.97 ± 0.0007	0.969 ± 0.0006	0.995 ± 0.0007	0.944 ± 0.0015	0.995 ± 0.0006	0.97 ± 0.0005
	2	0.969 ± 0.0008	0.97 ± 0.0009	0.965 ± 0.0028	**0.974 ± 0.002**	0.964 ± 0.0025	0.97 ± 0.008
	3	**0.982± 0.0008**	**0.982 ± 0.0008**	**1 ± 0**	0.964 ± 0.0016	**1 ± 0**	**0.982 ± 0.008**
Multi-species <0.25	1	0.973 ± 0.0011	0.975 ± 0.0009	0.995 ± 0.0011	0.956 ± 0.0017	0.995 ± 0.0012	0.975 ± 0.001
	2	0.976 ± 0.0007	0.978 ± 0.0008	0.974 ± 0.0011	**0.983 ± 0.0008**	0.968 ± 0.0013	0.975 ± 0.0008
	3	**0.983 ± 0.0015**	**0.984 ± 0.0014**	**1 ± 0**	0.969 ± 0.0027	**1 ± 0**	**0.985 ± 0.0013**
Multi-species <0.01	1	0.972 ± 0.0023	0.975 ± 0.0019	0.993 ± 0.001	0.958 ± 0.0035	0.991 ± 0.0015	0.975 ± 0.002
	2	0.978 ± 0.0015	0.981 ± 0.0013	0.975 ± 0.0024	**0.987 ± 0.0027**	0.966 ± 0.0031	0.976 ± 0.0015
	3	**0.981 ± 0.0016**	**0.983 ± 0.0014**	**1 ± 0**	0.966 ± 0.0027	**1 ± 0**	**0.983 ± 0.0014**

Table: Experimental results. 95% CIs are used for the results. 1 - denotes the Joint-Joint architecture, 2 - the Siamese-Joint architecture, 3 - the Siamese-Siamese architecture

- On each data set one of the siamese architectures provides the best results: Siamese-Joint architecture on HPRD and the Siamese-Siamese for the multi-species data sets

Results. Comparison with related work

Method	Accuracy	F1
AutoPPI	**0.979 ± 0.0007**	**0.979 ± 0.0007**
SAE (Sun et al., 2017)	0.9719	-
PIPR (Chen et al., 2019)	0.9811	0.9803
LDA-RF (Pan et al., 2010)	0.979 ± 0.005	-
CT-SVM (Shen et al., 2007) reported by Sun et al., 2017	0.83	-
AC-SVM (Guo et al., 2010) reported by Sun et al., 2017	0.9037	-
Parallel SVM (You et al., 2014) reported by Sun et al., 2017	0.9200–0.9740	-
ELM (You et al., 2014) reported by Sun et al., 2017	0.8480	0.8477
CS-SVM (Zhang et al., 2011)	0.941	0.937
SVM (Nanni et al., 2013)	0.942	-
DNN (Gui et al., 2020)	0.9443 ± 0.0036	-
DNN-PPI (Gui et al., 2019)	0.9726 ± 0.0018	-
DNN-CTAC (Wang et al., 2019)	**0.9837**	-
S-VGAE (Yang et al., 2020)	**0.9915 ± 0.0011**	**0.9915 ± 0.0012**

Table: Comparison between our method and related work on the HPRD data set.

Results. Comparison with related work

Data set	Method	Accuracy	F1
Multi-species	*AutoPPI*	**0.9821± 0.0008**	**0.9818 ± 0.0008**
	PIPR (Chen et al., 2019)	0.9819	0.9817
Multi-species <0.25	*AutoPPI*	**0.9829 ± 0.0015**	**0.9842 ± 0.0014**
	PIPR (Chen et al., 2019)	0.9791	0.9808
Multi-species <0.01	*AutoPPI*	**0.9808 ± 0.0016**	**0.9829 ± 0.0014**
	PIPR (Chen et al., 2019)	0.9751	0.9780

Table: Comparison between our method and related work on the Multi-species data sets.

Conclusions and future directions of research

- Introduced a procedure for binary classification of protein–protein interactions
- Proposed two new siamese architectures for the autoencoders
- Evaluated our approach on four data sets including proteins from different species
- Our approach surpassed the majority of related work approaches

Conclusions and future directions of research

Challenge: Random sampling: does not take into consideration whether the testing proteins are included in the training set $\rightarrow$ drop in performance when testing on unseen proteins[14,15]

Future directions:

- Improve generalization
- Improve performance on imbalanced data sets
- Provide interpretablity

[14] Dunham and Ganapathiraju, *Benchmark Evaluation of Protein-Protein Interaction Prediction Algorithms.*, Molecules, 2022.

[15] Park and Marcotte, *Flaws in evaluation schemes for pair-input computational predictions.*, Nature methods, 2012.

Thank you!