Multifaceted protein–protein interaction prediction based on Siamese residual RCNN

Muhao Chen, Chelsea J.-T. Ju, Guangyu Zhou, Xuelu Chen, Tianran Zhang, Kai-Wei Chang, Carlo Zaniolo and Wei Wang

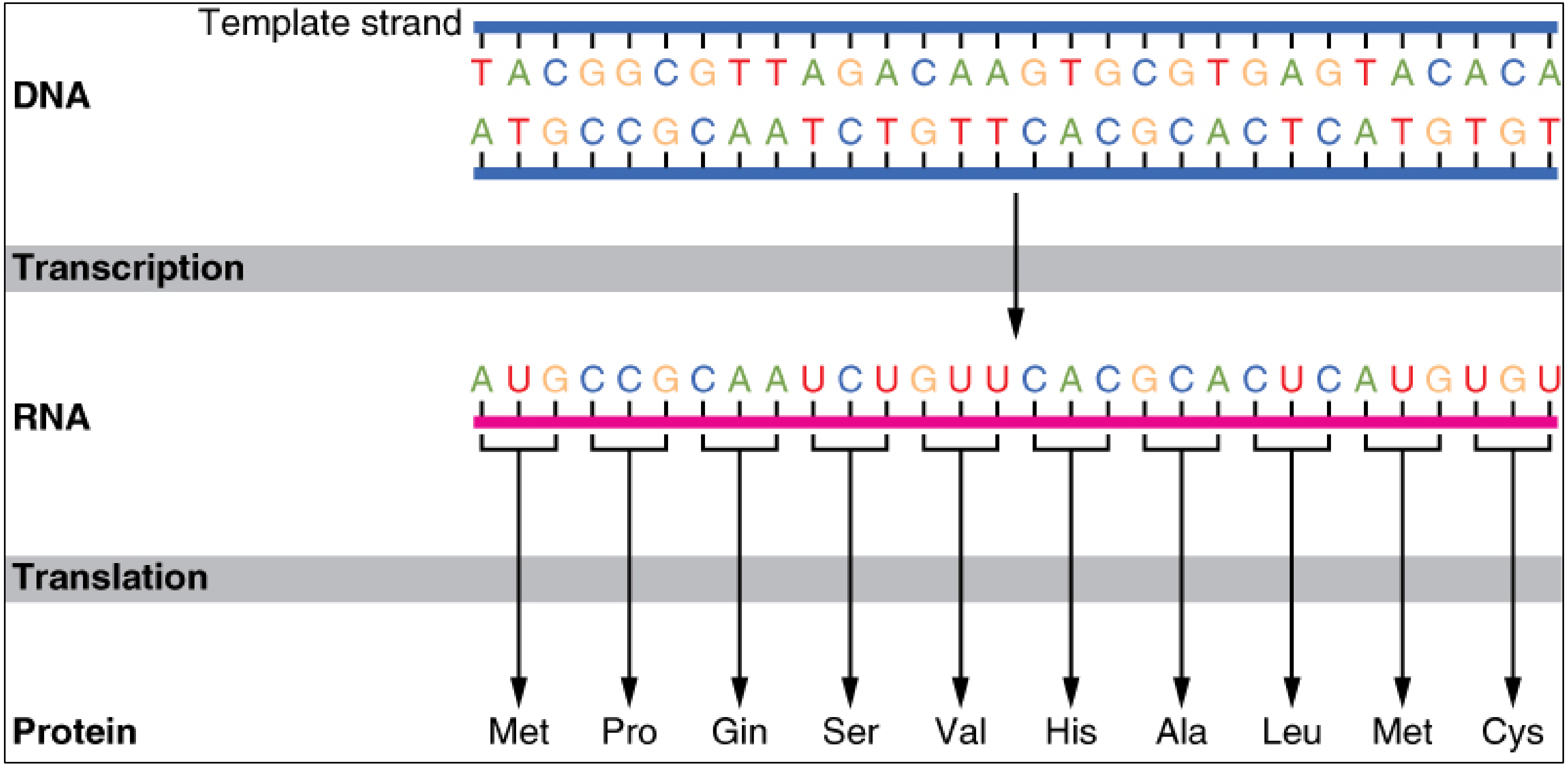
Gian Favero

October 19th, 2023

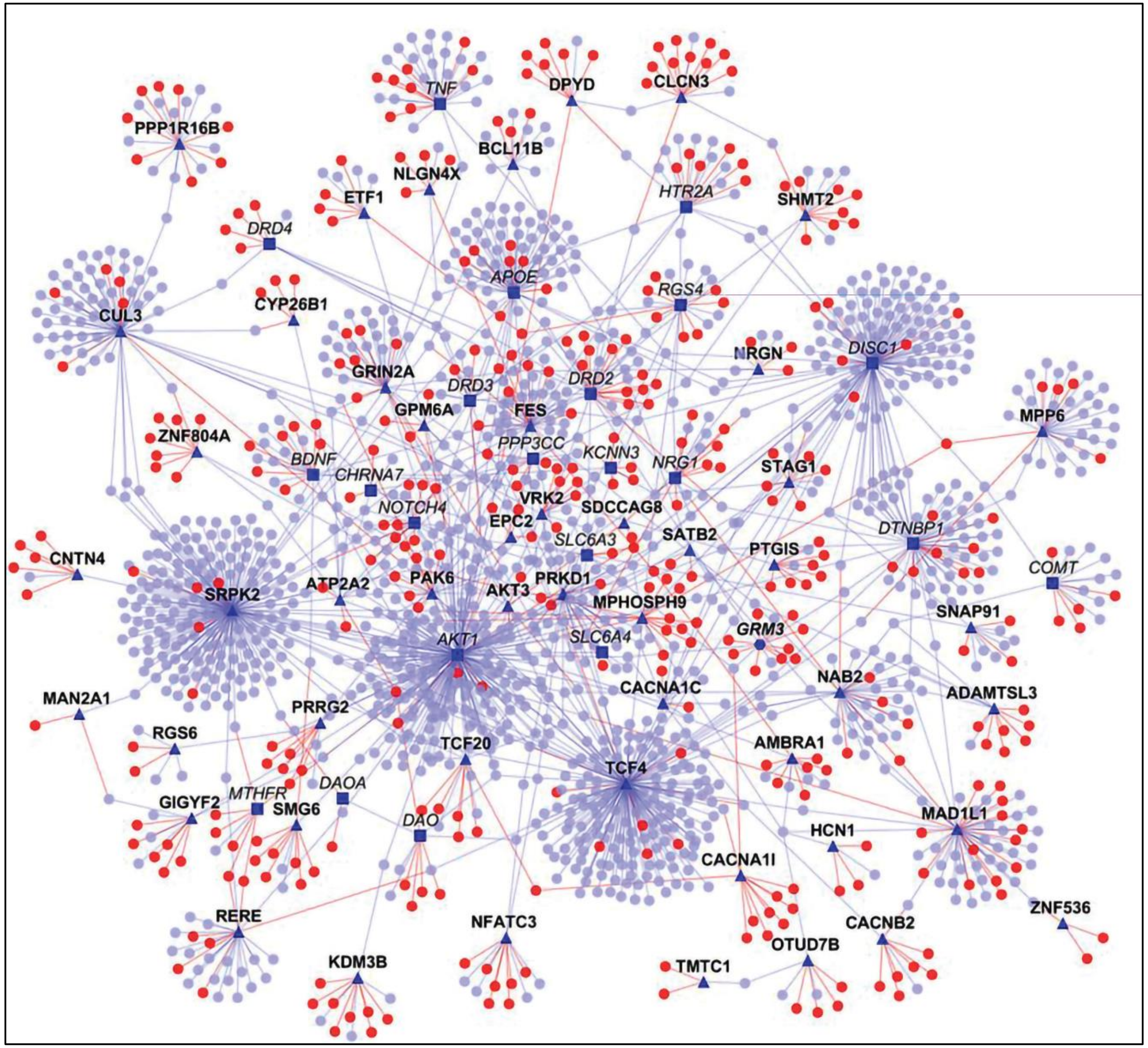
Table of Contents

1. Background
2. Introduction/Motivation
3. Problem Formulation
4. Approach
5. Results
6. Shortcomings

Background


Template strand
DNA
T A C G G C G T T A G A C A A G T G C G T G A G T A C A C A
A T G C C G C A A T C T G T T C A C G C A C T C A T G T G T
Transcription
RNA
A U G C C G C A A U C U G U U C A C G C A C U C A U G U G U
Translation
Protein
Met
Pro
Gin
Ser
Val
His
Ala
Leu
Met
Cys

https://courses.lumenlearning.com/suny-ap1/chapter/3-4-protein-synthesis/

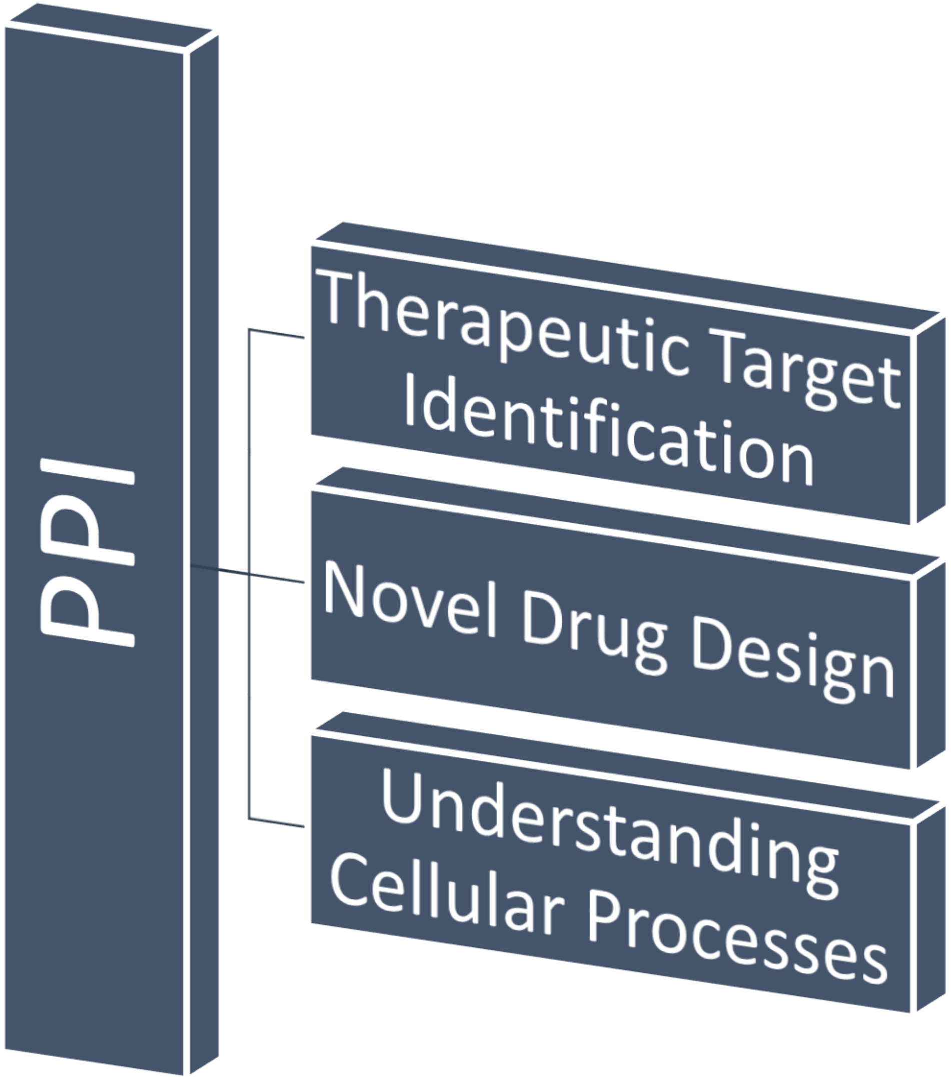


https://www.genengnews.com/insights/protein-protein-interactions-get-a-new-groove-on/

- In a similar fashion to genes, proteins interact with each other

Introduction/Motivation

> Detecting protein–protein interactions (PPIs) and characterizing the interaction types are essential toward understanding cellular biological processes in normal and disease states [1].



PPI

Therapeutic Target Identification

Novel Drug Design

Understanding Cellular Processes

Existing Methods - Experimental

- Yeast Two-Hybrid Screens (Fields and Song, 1989)
- Tandem Affinity Purification (Gavin et Al., 2002)
- Mass Spectrometric Protein Complex Identification (Ho et Al., 2002)

- Expensive
- Labor-intensive
- Time-consuming
- High false positives

Existing Methods – Statistical

- SVM (Guo et al., 2008; You et al., 2014)
- kNN (Yang et al., 2010)
- Random Forest (Wong et al., 2015)
- Multi-layer perceptron (MLP) (Du et al., 2017)
- Ensemble ELM (EELM) (You et al., 2013)

- Rely on extracted features
- Limited coverage on interaction information
- Dedicated to specific protein profiles

Existing Methods – Deep Learning

- DPPI (Hashemifar et al., 2018)
  - CNN based
- DNN-PPI (Li et al., 2018)
  - Two CNN encoders

- Require pre-processing
- Limited to binary prediction
- Do not consider contextual and sequential information

Problem Formulation

> Evidently, there is an immense need for reliable computational approaches to identify and characterize PPIs [1].

Constraints

☑ Process large-scale data and automatically extract useful features without preprocessing

☑ Consider contextualized and sequential information when modelling PPIs

☑ Generalize to multiple PPI prediction tasks

Approach

Addressing Constraint 1

☑ Process large-scale data and automatically extract useful features without preprocessing

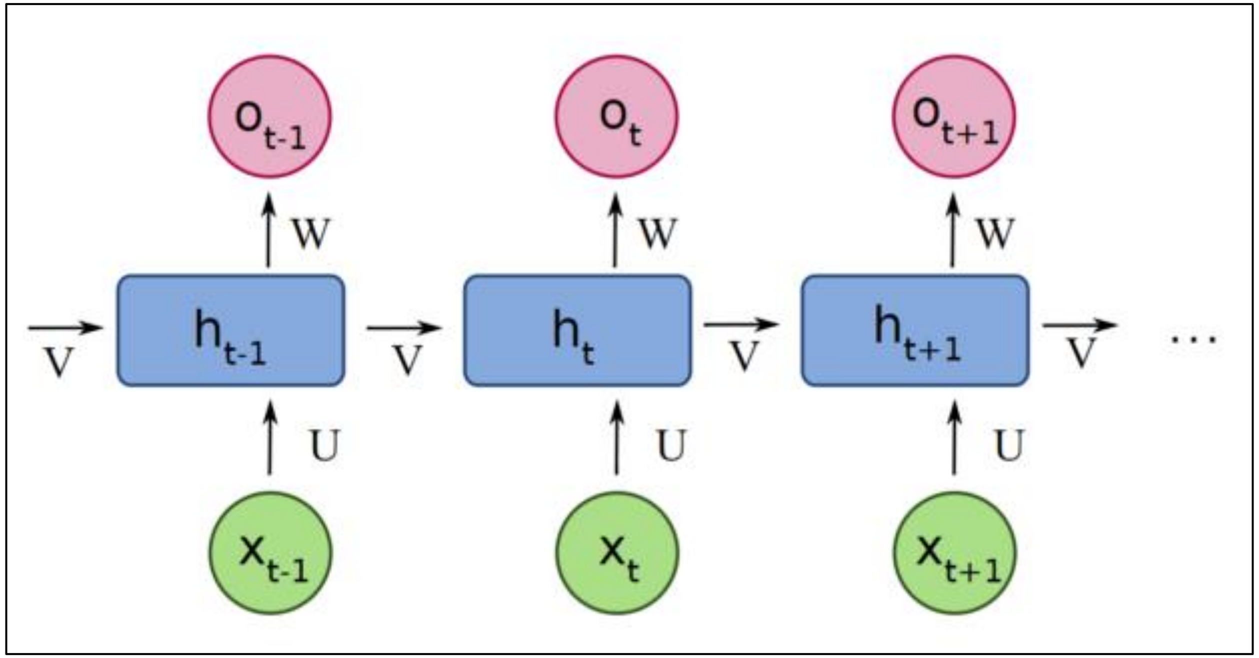
- Deep learning is powerful enough to process large-scale data
  - Convolutional neural nets (CNNs)

- CNNs used in similar bioinformatics problems to select features
  - Genetic variants detection (Anderson, 2018)
  - RNA-binding site prediction (Zhang et al., 2016)

- NLP-like sequence modeling can prevent need for pre-processing

Addressing Constraint 2

☑ Consider contextualized and sequential information when modelling PPIs



https://www.analyticsvidhya.com/blog/2022/03/a-brief-overview-of-recurrent-neural-networks-rnn/

- Recurrent neural nets (RNNs) aim at preserving contextualized and long-term ordering information

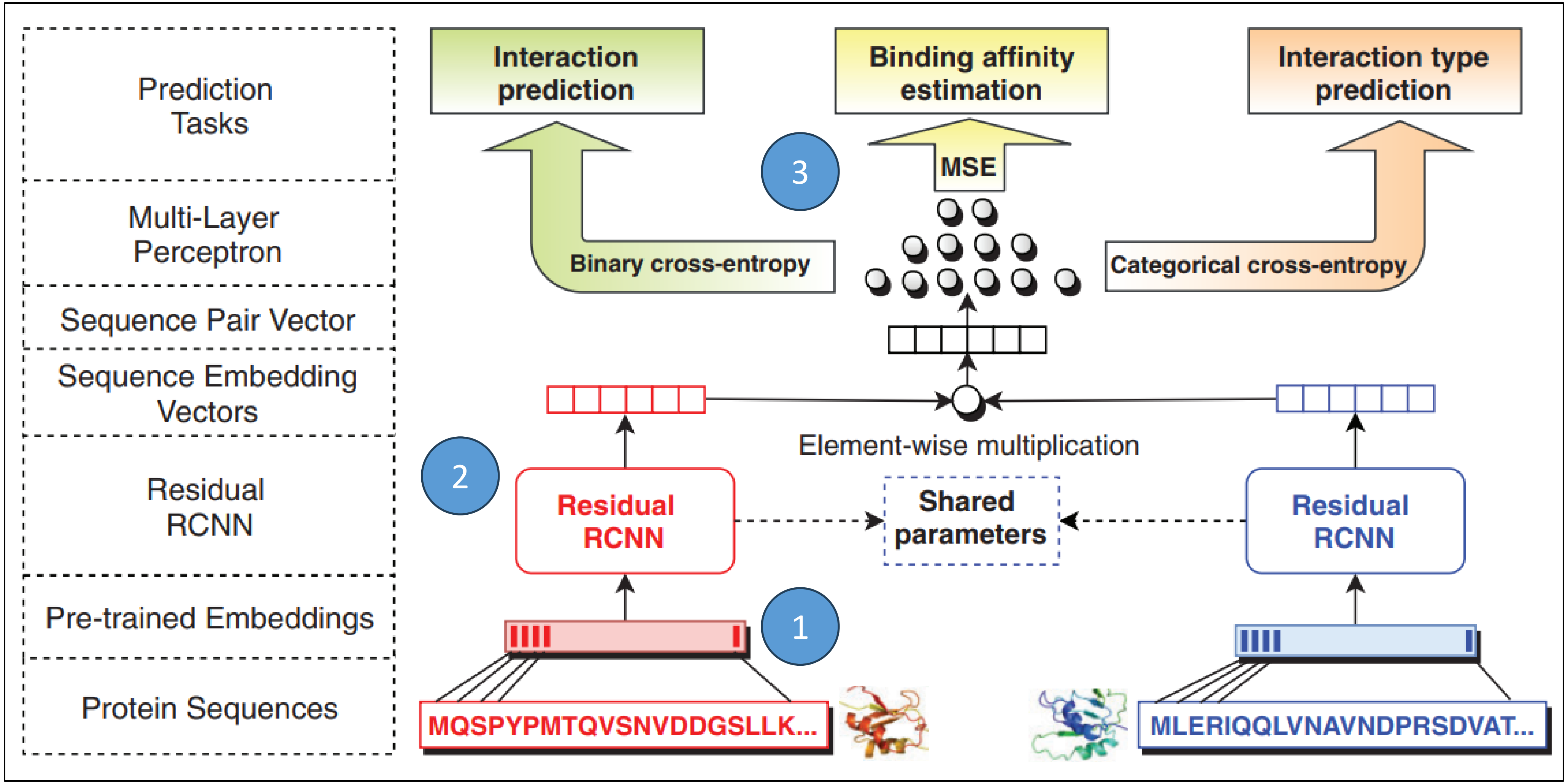
- RNNs used in similar problems
  - DNA function classification (Quang and Xie, 2016)

Addressing Constraint 3

☑ Generalize to different PPI prediction tasks

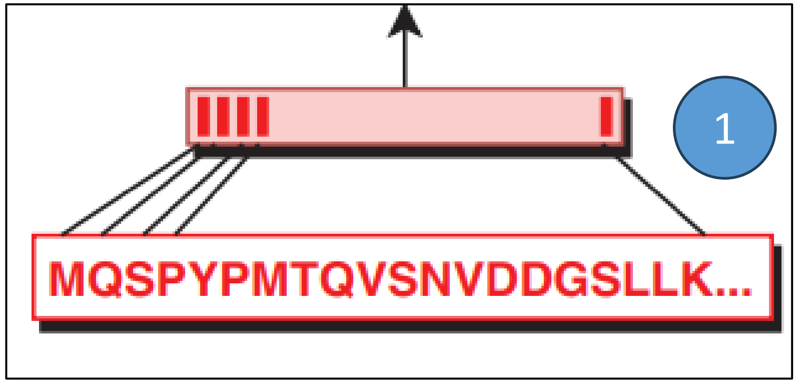
- Multi-output Siamese network captures the mutual influence of a protein sequence pair

- MLP used to predict:
  - Interaction (binary classification)
  - Binding affinity estimation (regression)
  - Interaction type (multiclass classification)



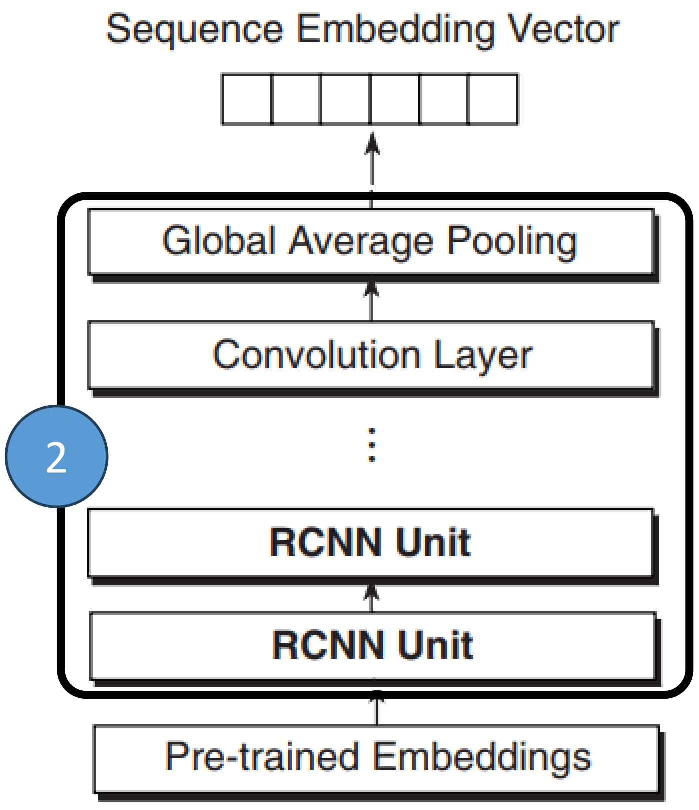
[1]

Multifaceted PPI based on Siamese Residual RCNN (PIPR)



[1]

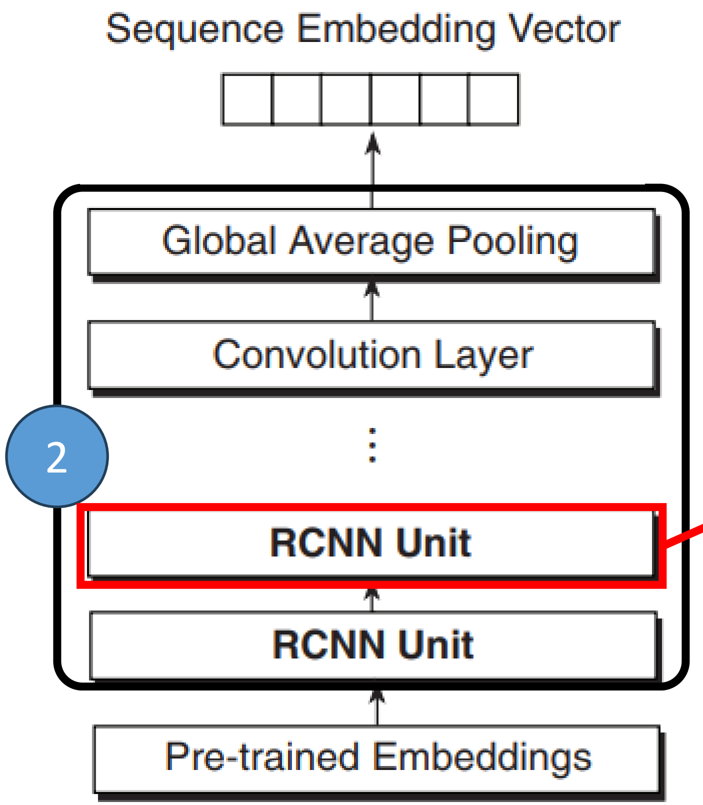
- Inputs are amino acid sequences ☑
- Comparable to sentence pair modeling tasks in NLP
- A pre-trained encoder (Skip-Gram) reduces input to a latent vector



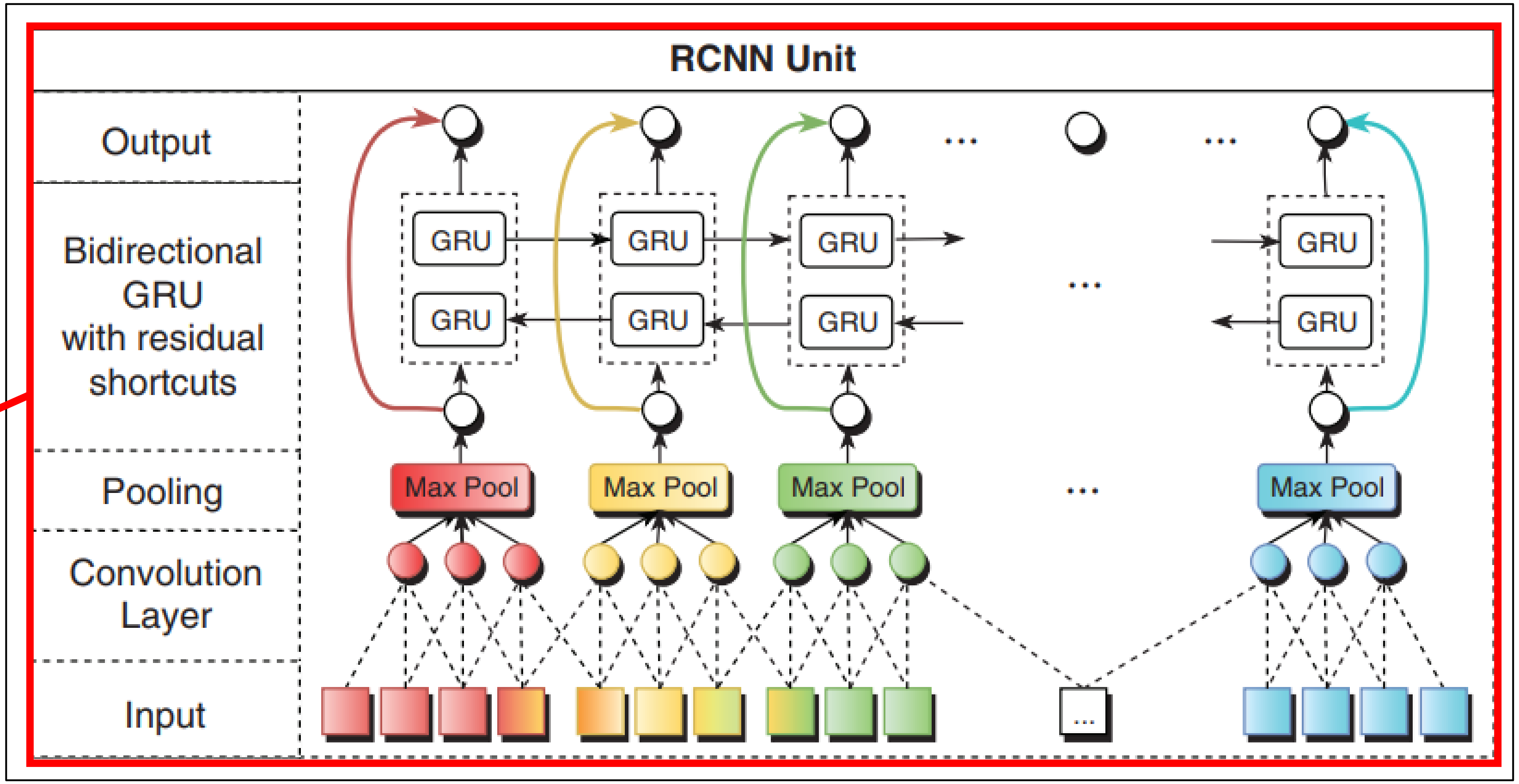
[1]

Residual RCNN

- Stacks multiple layers of RCNN units
- Seeks to leverage global sequential information and local features ☑
- Outputs a sequence embedding vector of a lower-dimension

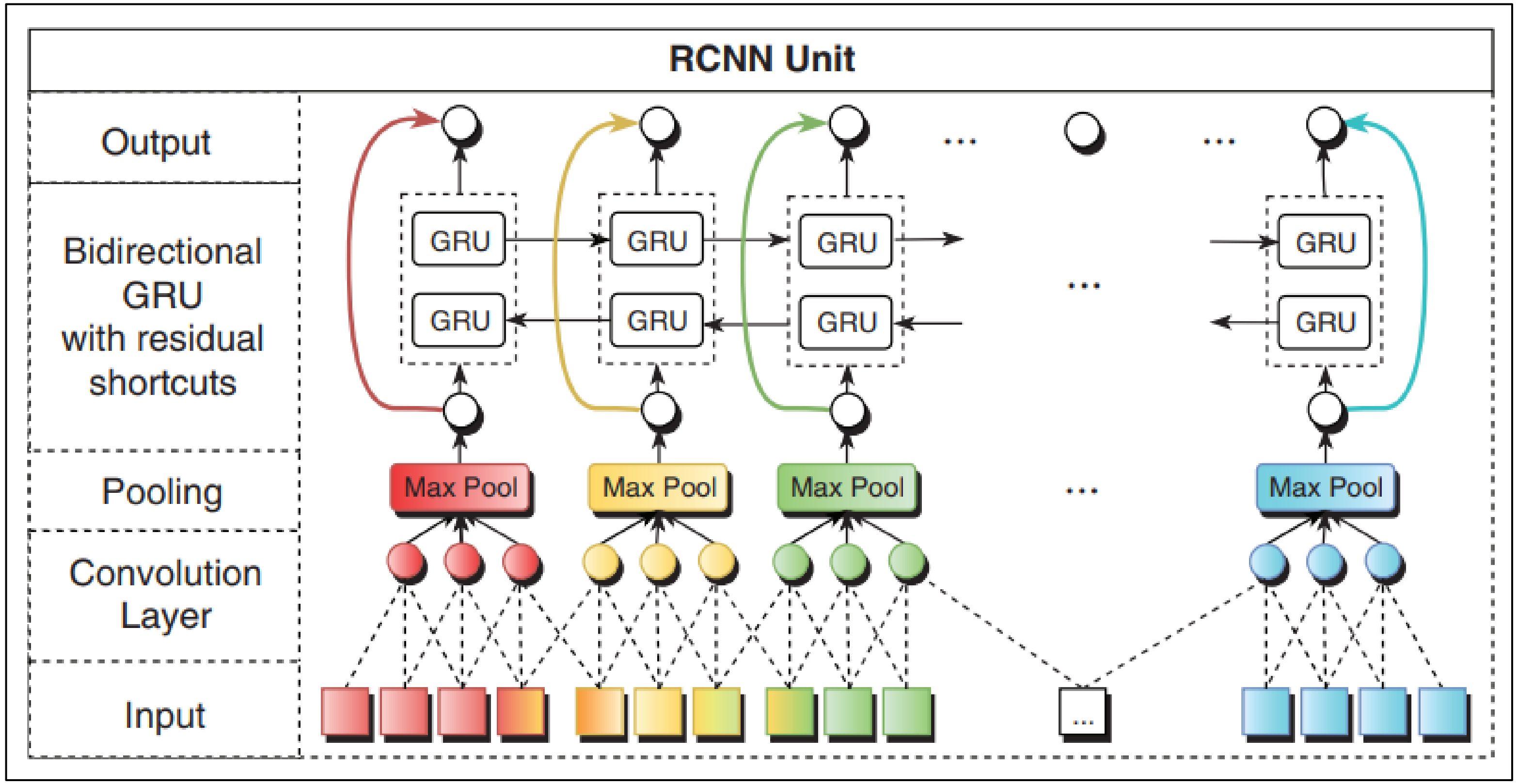


Residual RCNN

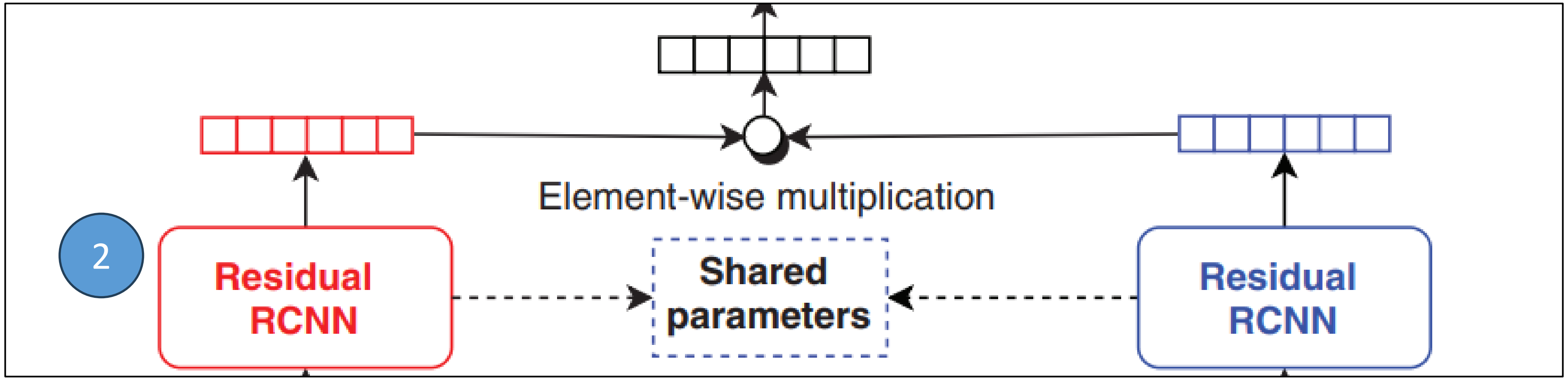


[1]

- Convolution extracts local features
- Pooling preserves important local features
- Bidirectional GRU layer preserves sequential information
- Residual connections improves training

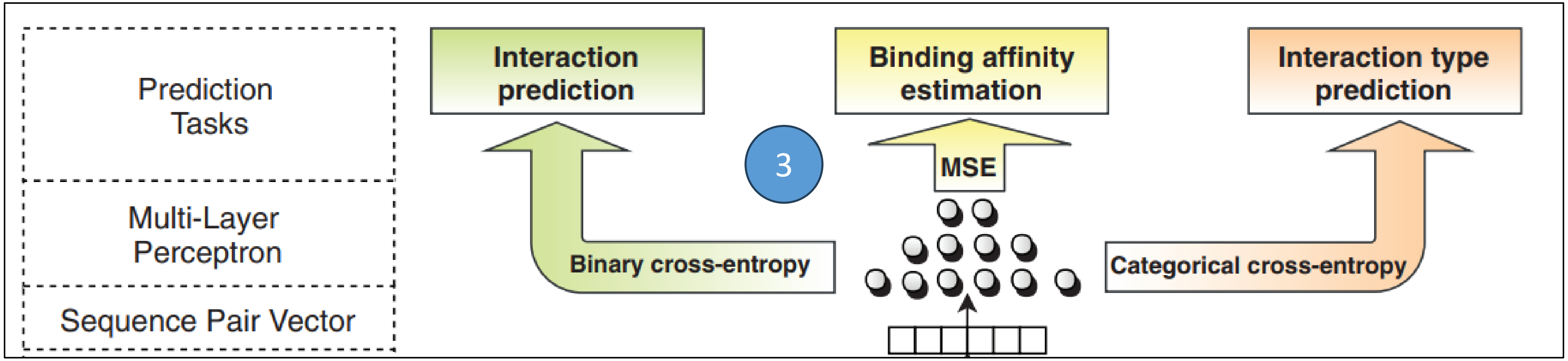


[1]



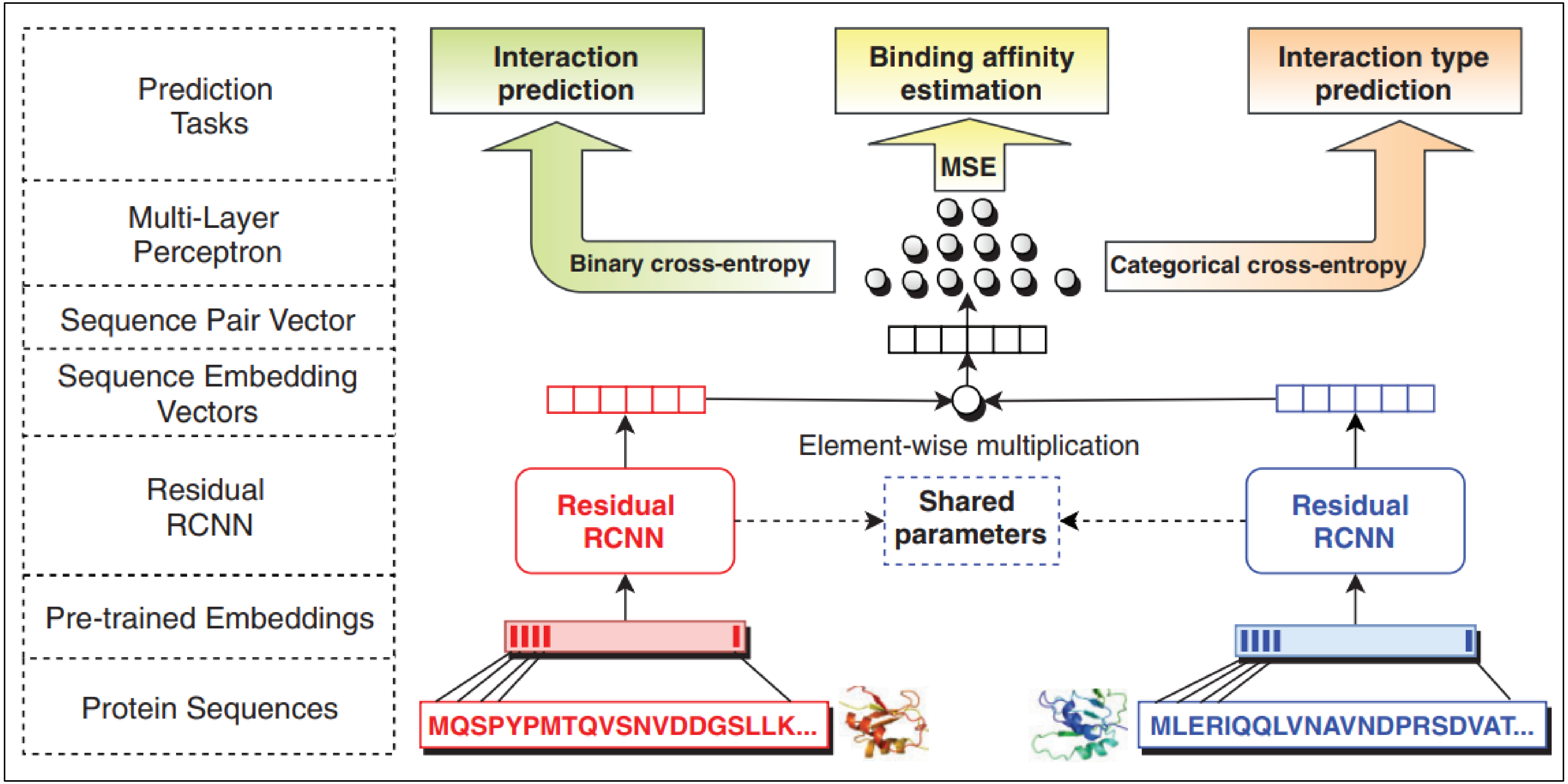
[1]

- Residual RCNNs are trained together with shared parameters
- Sequence embeddings are combined via element-wise multiplication (shown to be better than concatenation)



[1]

- Sequence pair vector fed into an MLP
- MLP optimizes loss function based on task: ☑
  - Cross-entropy for interaction/type prediction
  - MSE for binding affinity estimation



[1]

Recap of PIPR

Results

Binary PPI Prediction

- Used the Yeast dataset for benchmarking

- Compared against statistical and deep learning baselines

- Also compared against 2 ablations of PIPR
  - SRGRU
    - All convolution layers in PIPR discarded
    - Shows value of contextualized and sequential information
  - SCNN
    - Removal of residual GRU in PIPR
    - Shows value of significant local features

Binary PPI Prediction

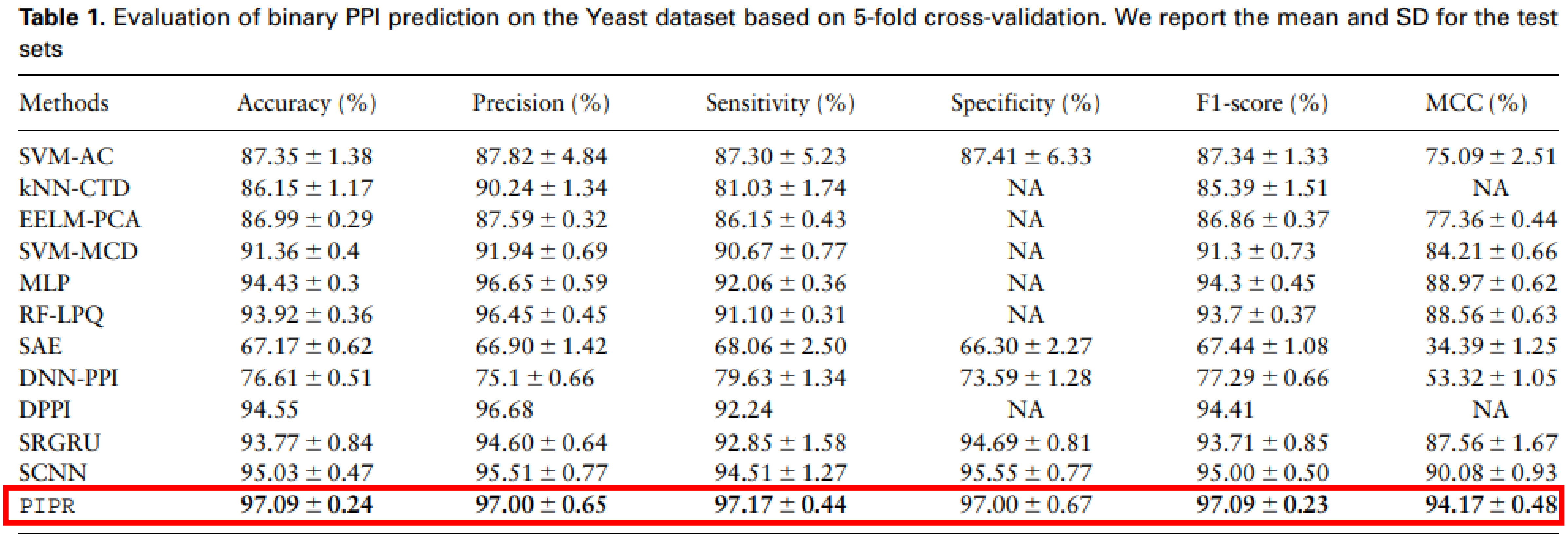


Table 1. Evaluation of binary PPI prediction on the Yeast dataset based on 5-fold cross-validation. We report the mean and SD for the test sets

Methods	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	MCC (%)
SVM-AC	87.35 ± 1.38	87.82 ± 4.84	87.30 ± 5.23	87.41 ± 6.33	87.34 ± 1.33	75.09 ± 2.51
kNN-CTD	86.15 ± 1.17	90.24 ± 1.34	81.03 ± 1.74	NA	85.39 ± 1.51	NA
EELM-PCA	86.99 ± 0.29	87.59 ± 0.32	86.15 ± 0.43	NA	86.86 ± 0.37	77.36 ± 0.44
SVM-MCD	91.36 ± 0.4	91.94 ± 0.69	90.67 ± 0.77	NA	91.3 ± 0.73	84.21 ± 0.66
MLP	94.43 ± 0.3	96.65 ± 0.59	92.06 ± 0.36	NA	94.3 ± 0.45	88.97 ± 0.62
RF-LPQ	93.92 ± 0.36	96.45 ± 0.45	91.10 ± 0.31	NA	93.7 ± 0.37	88.56 ± 0.63
SAE	67.17 ± 0.62	66.90 ± 1.42	68.06 ± 2.50	66.30 ± 2.27	67.44 ± 1.08	34.39 ± 1.25
DNN-PPI	76.61 ± 0.51	75.1 ± 0.66	79.63 ± 1.34	73.59 ± 1.28	77.29 ± 0.66	53.32 ± 1.05
DPPI	94.55	96.68	92.24	NA	94.41	NA
SRGRU	93.77 ± 0.84	94.60 ± 0.64	92.85 ± 1.58	94.69 ± 0.81	93.71 ± 0.85	87.56 ± 1.67
SCNN	95.03 ± 0.47	95.51 ± 0.77	94.51 ± 1.27	95.55 ± 0.77	95.00 ± 0.50	90.08 ± 0.93
PIPR	**97.09 ± 0.24**	**97.00 ± 0.65**	**97.17 ± 0.44**	97.00 ± 0.67	**97.09 ± 0.23**	**94.17 ± 0.48**

[1]

Binary PPI Prediction

- P-values of PIPR and ablations against baselines on Yeast dataset

- P-value < 0.01 are considered significant

- DPPI not included as it has no standard deviation measure

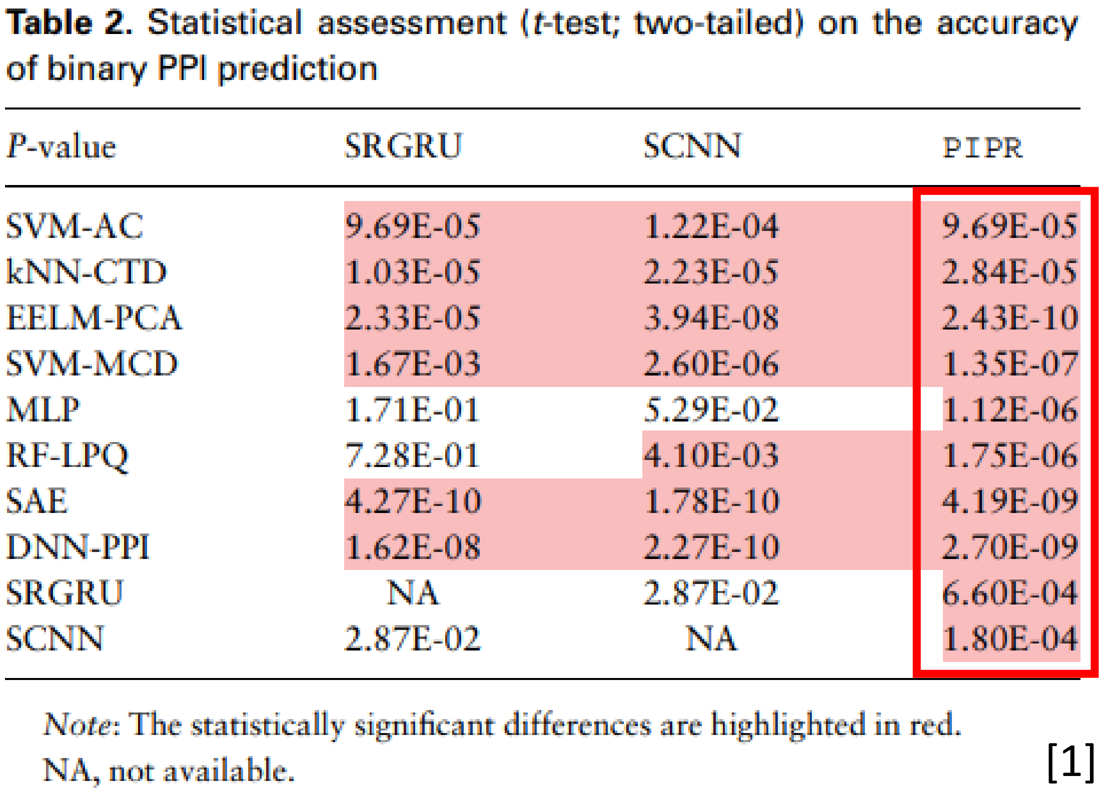


Table 2. Statistical assessment (*t*-test; two-tailed) on the accuracy of binary PPI prediction

P-value	SRGRU	SCNN	PIPR
SVM-AC	9.69E-05	1.22E-04	9.69E-05
kNN-CTD	1.03E-05	2.23E-05	2.84E-05
EELM-PCA	2.33E-05	3.94E-08	2.43E-10
SVM-MCD	1.67E-03	2.60E-06	1.35E-07
MLP	1.71E-01	5.29E-02	1.12E-06
RF-LPQ	7.28E-01	4.10E-03	1.75E-06
SAE	4.27E-10	1.78E-10	4.19E-09
DNN-PPI	1.62E-08	2.27E-10	2.70E-09
SRGRU	NA	2.87E-02	6.60E-04
SCNN	2.87E-02	NA	1.80E-04

Note: The statistically significant differences are highlighted in red.
NA, not available.

[1]

Binary PPI Prediction

- Performance of PIPR analyzed on a multi-species dataset
- Accuracy and F1-score reported on a 5-fold CV
- Performance remained high

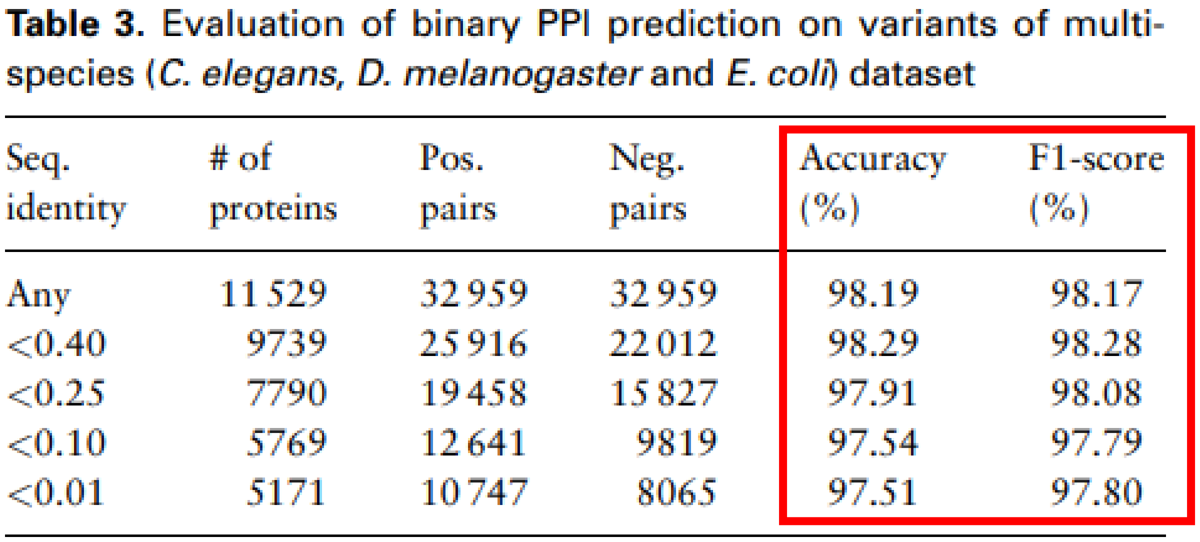


Table 3. Evaluation of binary PPI prediction on variants of multi-species (*C. elegans, D. melanogaster* and *E. coli*) dataset

Seq. identity	# of proteins	Pos. pairs	Neg. pairs	Accuracy (%)	F1-score (%)
Any	11 529	32 959	32 959	98.19	98.17
<0.40	9739	25 916	22 012	98.29	98.28
<0.25	7790	19 458	15 827	97.91	98.08
<0.10	5769	12 641	9819	97.54	97.79
<0.01	5171	10 747	8065	97.51	97.80

[1]

Interaction Type Prediction

- Evaluated based on STRING PPI datasets for benchmarking
  - SHS27k and SHS148k

- Interaction types (7):
  - Activation, binding, catalysis, expression, inhibition, posttranslational modification and reaction

- 10-fold CV used to calculate metrics for each baseline
  - Accuracy
  - Fold changes over zero (more is better)

Interaction Type Prediction

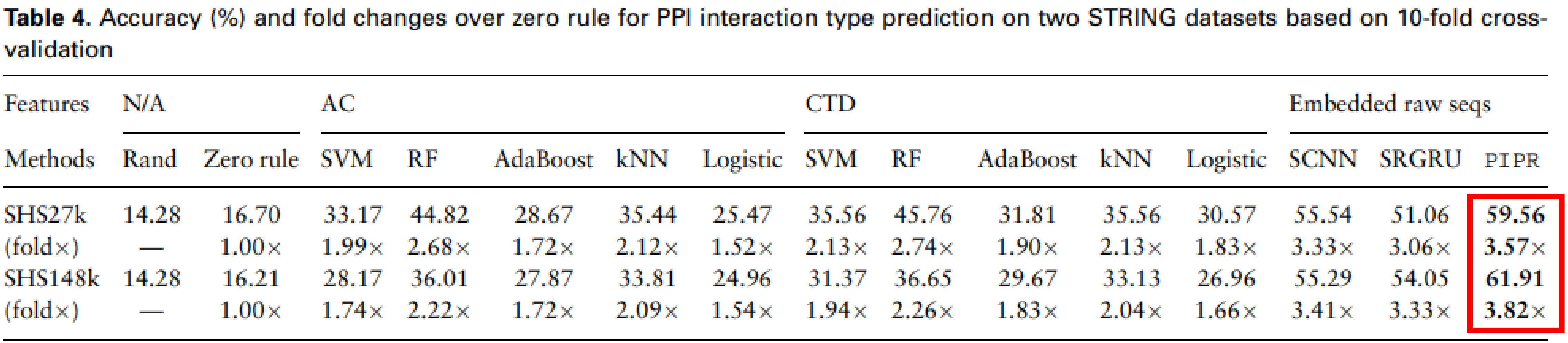


Table 4. Accuracy (%) and fold changes over zero rule for PPI interaction type prediction on two STRING datasets based on 10-fold cross-validation

Features	N/A		AC					CTD					Embedded raw seqs		
Methods	Rand	Zero rule	SVM	RF	AdaBoost	kNN	Logistic	SVM	RF	AdaBoost	kNN	Logistic	SCNN	SRGRU	PIPR
SHS27k	14.28	16.70	33.17	44.82	28.67	35.44	25.47	35.56	45.76	31.81	35.56	30.57	55.54	51.06	**59.56**
(fold×)	—	1.00×	1.99×	2.68×	1.72×	2.12×	1.52×	2.13×	2.74×	1.90×	2.13×	1.83×	3.33×	3.06×	**3.57×**
SHS148k	14.28	16.21	28.17	36.01	27.87	33.81	24.96	31.37	36.65	29.67	33.13	26.96	55.29	54.05	**61.91**
(fold×)	—	1.00×	1.74×	2.22×	1.72×	2.09×	1.54×	1.94×	2.26×	1.83×	2.04×	1.66×	3.41×	3.33×	**3.82×**

[1]

- Accuracy is much lower as this is a much harder task, but PIPR still outperforms the next highest method by nearly 4%

Binding Affinity Estimation

- Evaluated based on SKEMPI dataset
- Compared against several regression models as baselines
- Evaluation based on MSE, MAE, and Pearson’s Correlation Coefficient

Binding Affinity Estimation

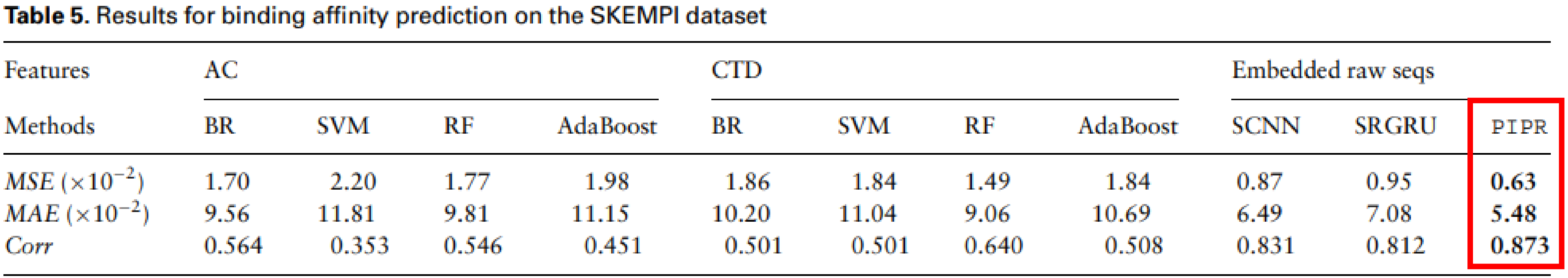


Table 5. Results for binding affinity prediction on the SKEMPI dataset

Features	AC				CTD				Embedded raw seqs		
Methods	BR	SVM	RF	AdaBoost	BR	SVM	RF	AdaBoost	SCNN	SRGRU	PIPR
MSE ($\times 10^{-2}$)	1.70	2.20	1.77	1.98	1.86	1.84	1.49	1.84	0.87	0.95	**0.63**
MAE ($\times 10^{-2}$)	9.56	11.81	9.81	11.15	10.20	11.04	9.06	10.69	6.49	7.08	**5.48**
Corr	0.564	0.353	0.546	0.451	0.501	0.501	0.640	0.508	0.831	0.812	**0.873**

[1]

- PIPR and its ablations performed much better than all baselines in binding affinity estimation

Amino Acid Embeddings

- Embeddings describe physicochemical properties
  - Co-occurrence similarity of amino acids, a_c
  - Categorization of electrostaticity, a_{ph}

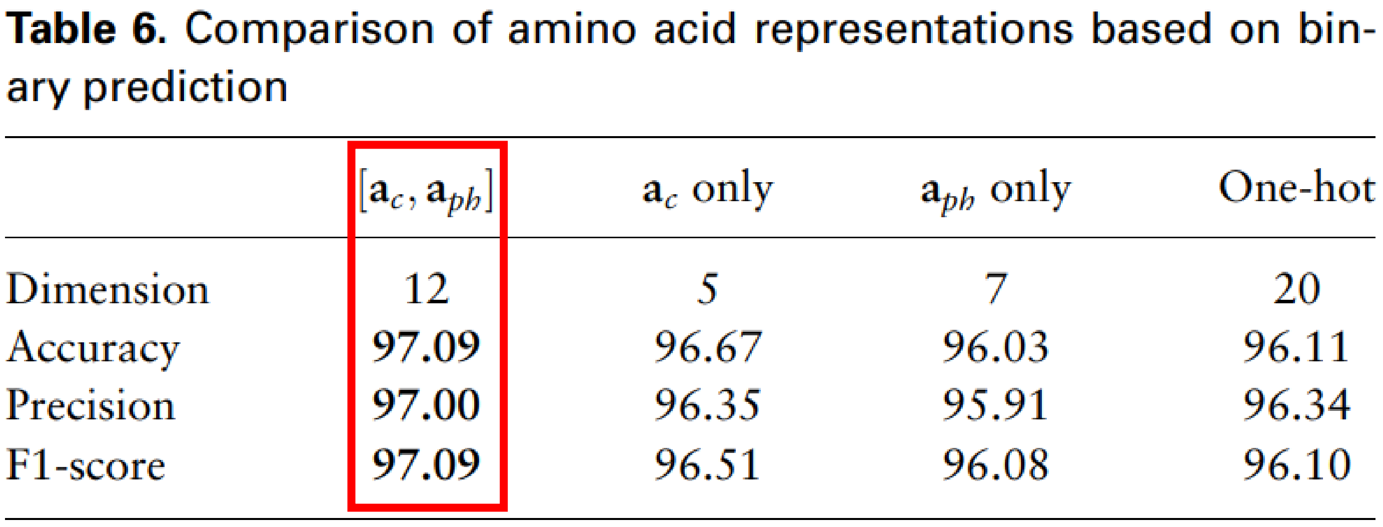


Table 6. Comparison of amino acid representations based on binary prediction

	$[\mathbf{a}_c, \mathbf{a}_{ph}]$	$\mathbf{a}_c$ only	$\mathbf{a}_{ph}$ only	One-hot
Dimension	12	5	7	20
Accuracy	**97.09**	96.67	96.03	96.11
Precision	**97.00**	96.35	95.91	96.34
F1-score	**97.09**	96.51	96.08	96.10

[1]

Runtime

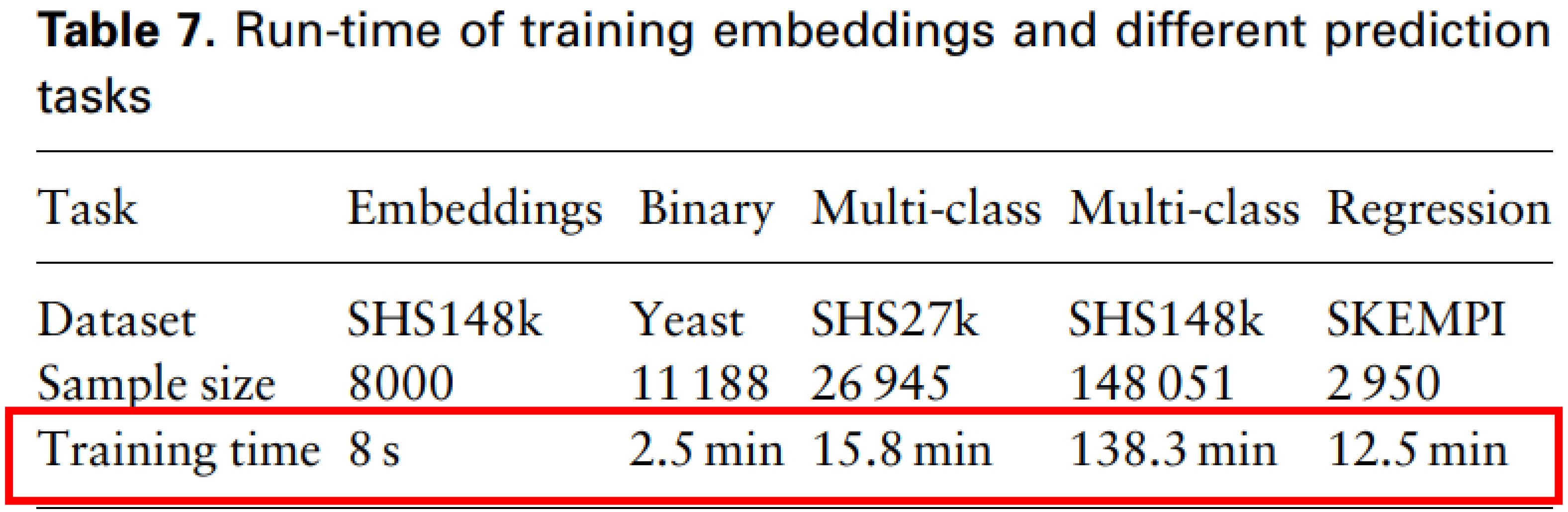


Table 7. Run-time of training embeddings and different prediction tasks

Task	Embeddings	Binary	Multi-class	Multi-class	Regression
Dataset	SHS148k	Yeast	SHS27k	SHS148k	SKEMPI
Sample size	8000	11 188	26 945	148 051	2 950
Training time	8 s	2.5 min	15.8 min	138.3 min	12.5 min

[1]

- DPPI requires extensive resources for pre-processing
  - Estimated 26 days for the Yeast dataset

Conclusions and Shortcomings

Conclusions

- Local and sequential information captured by PIPR shown to be effective in finding mutual influence of proteins
- Framework is adaptable to other PPI tasks
- Extensive evaluation on five datasets shows that the framework is superior to previous statistical and deep learning methods
- Pre-defined features and data preprocessing can be avoided in PPI prediction tasks

Shortcomings

- Baseline models for interaction type were developed for binary interaction prediction
  - Performance increase not genuine

- All three ablations performed similarly
  - Sequential and contextualized information not as pronounced as authors made it seem
  - Siamese architecture seems to be the real boon

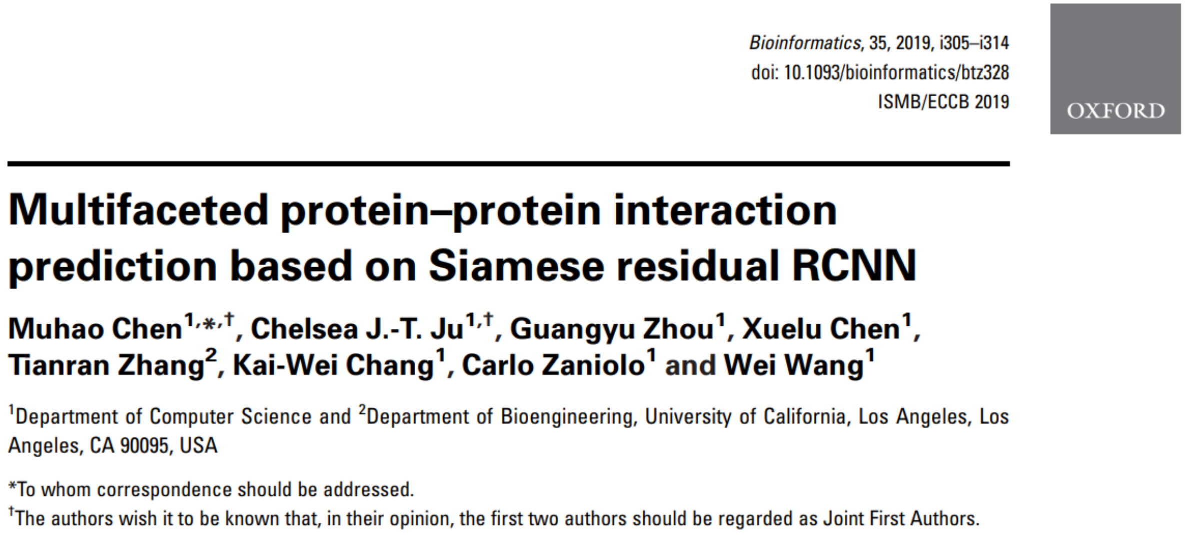
- MLP layer of methodology slightly unclear

References

McGill

References

[1] Muhao Chen, Chelsea J -T Ju, Guangyu Zhou, Xuelu Chen, Tianran Zhang, Kai-Wei Chang, Carlo Zaniolo, Wei Wang, Multifaceted protein–protein interaction prediction based on Siamese residual RCNN, Bioinformatics, Volume 35, Issue 14, July 2019, Pages i305–i314, https://doi.org/10.1093/bioinformatics/btz328



Bioinformatics, 35, 2019, i305–i314
doi: 10.1093/bioinformatics/btz328
ISMB/ECCB 2019

Multifaceted protein–protein interaction prediction based on Siamese residual RCNN

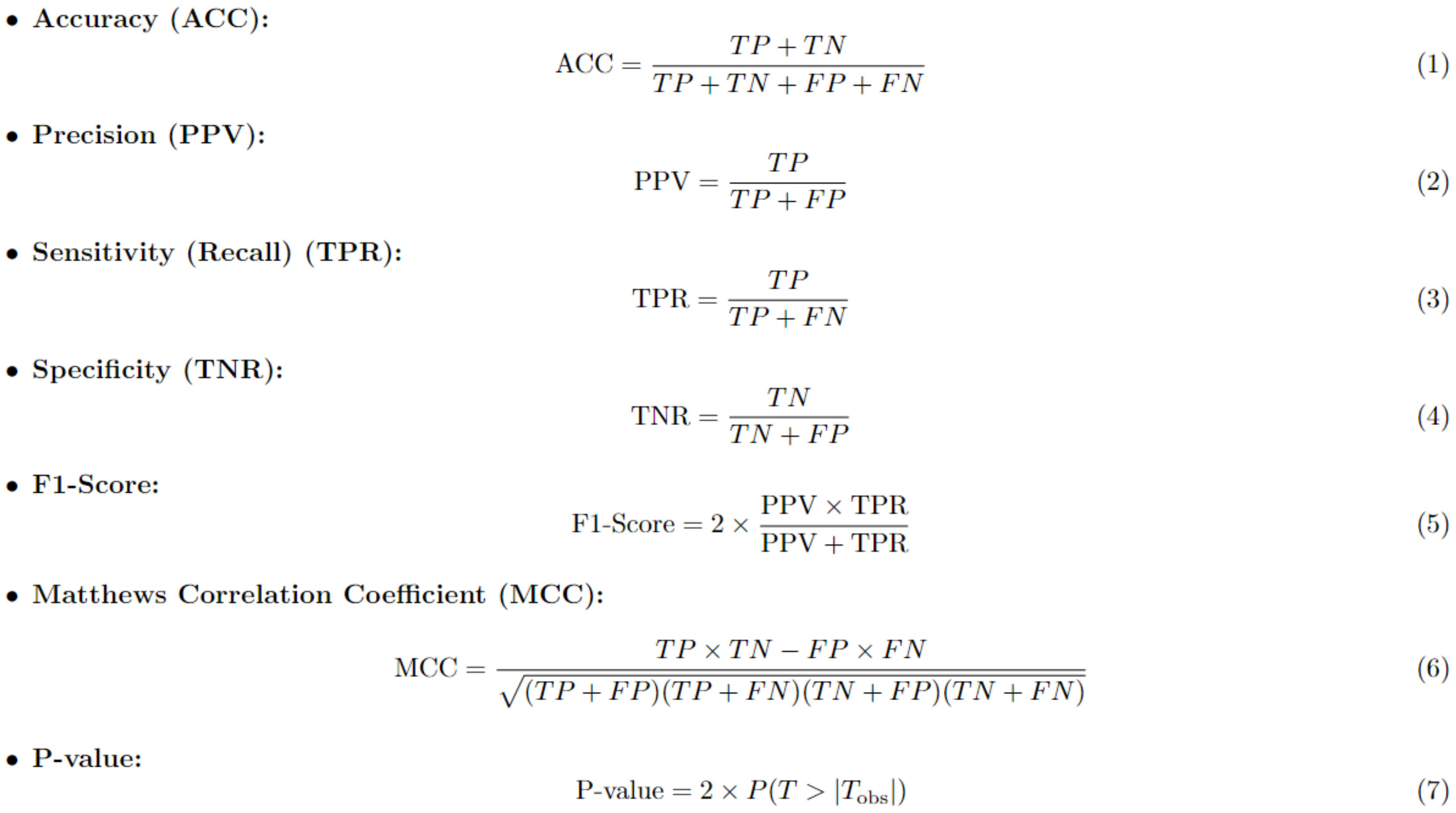
Muhao Chen[1,*,†], Chelsea J.-T. Ju[1,†], Guangyu Zhou[1], Xuelu Chen[1], Tianran Zhang[2], Kai-Wei Chang[1], Carlo Zaniolo[1] and Wei Wang[1]

[1]Department of Computer Science and [2]Department of Bioengineering, University of California, Los Angeles, Los Angeles, CA 90095, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Appendix

Metric Formulas



- **Accuracy (ACC):**

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

- **Precision (PPV):**

$$\text{PPV} = \frac{TP}{TP + FP} \quad (2)$$

- **Sensitivity (Recall) (TPR):**

$$\text{TPR} = \frac{TP}{TP + FN} \quad (3)$$

- **Specificity (TNR):**

$$\text{TNR} = \frac{TN}{TN + FP} \quad (4)$$

- **F1-Score:**

$$\text{F1-Score} = 2 \times \frac{\text{PPV} \times \text{TPR}}{\text{PPV} + \text{TPR}} \quad (5)$$

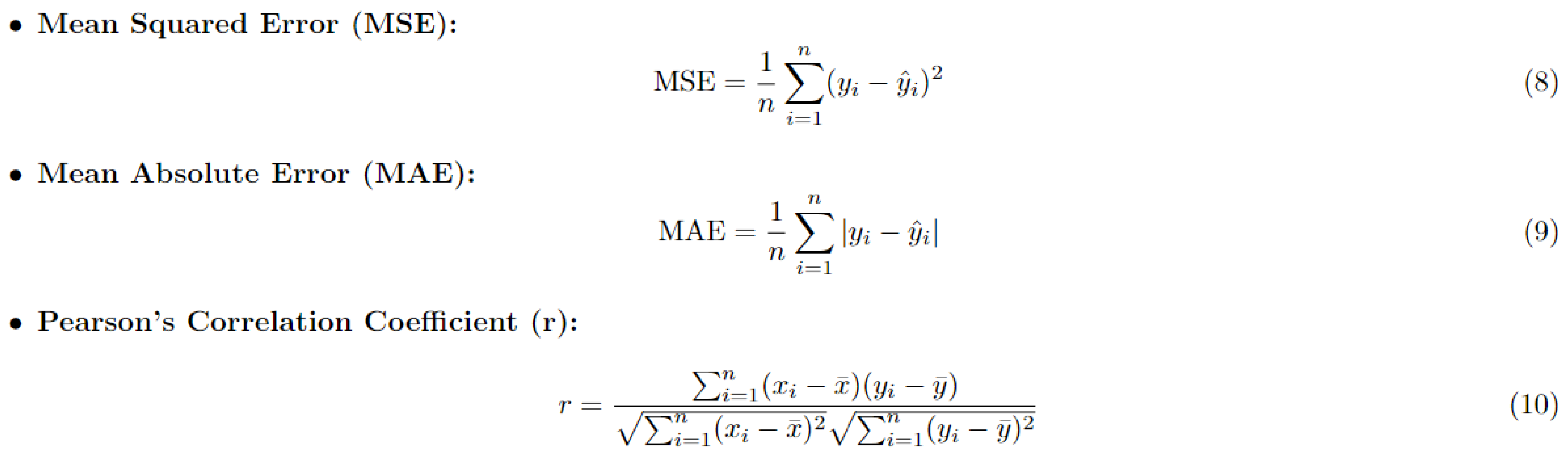
- **Matthews Correlation Coefficient (MCC):**

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (6)$$

- **P-value:**

$$\text{P-value} = 2 \times P(T > |T_{\text{obs}}|) \quad (7)$$

Metric Formulas



- **Mean Squared Error (MSE):**

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

- **Mean Absolute Error (MAE):**

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9}$$

- **Pearson's Correlation Coefficient (r):**

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{10}$$