CENTER FOR BIOSYSTEMS & BIOTECH DATA SCIENCE

Yunseol Park, Espoir Kabanga, Jasper Zuallaert, Hyunjin Shim, Arnout Van Messem, Wesley De Neve

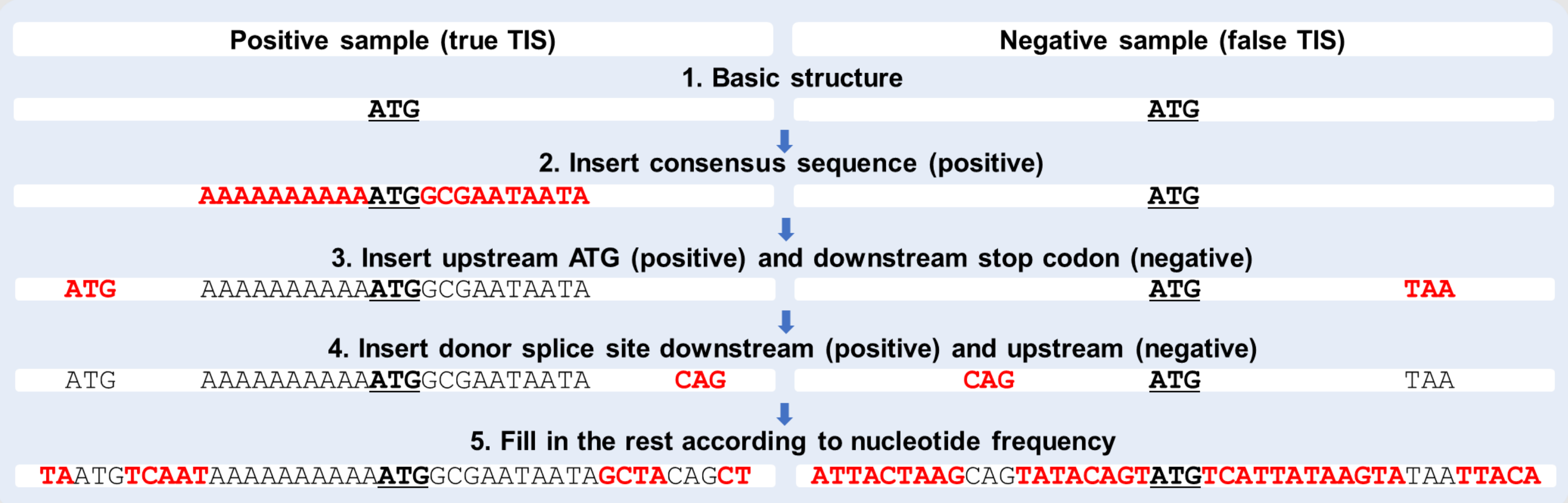
TRANSLATION INITIATION SITE PREDICTION USING DEEP LEARNING AND SYNTHETIC DATASETS

Introduction and Motivation

Building a prediction model for translation initiation sites (TISs) and determining their important features may aid in uncovering new translation mechanisms and give emphasis to already existing ones. However, interpretation is difficult, as many machine learning models are black box in nature. Therefore, to better understand the relevant features, we investigate the use of synthetic data in the context of TIS prediction for *A. thaliana* and, through transfer learning, for *H. sapiens*.

Data Generation and Model Training

Synthetic Dataset → Synthetic Black-Box Model (SBBM)



Positive sample (true TIS) | Negative sample (false TIS)

1. Basic structure
ATG | ATG

2. Insert consensus sequence (positive)
AAAAAAAAAAATGGCGAATAATA | ATG

3. Insert upstream ATG (positive) and downstream stop codon (negative)
ATG AAAAAAAAAAATGGCGAATAATA | ATG TAA

4. Insert donor splice site downstream (positive) and upstream (negative)
ATG AAAAAAAAAAATGGCGAATAATA CAG | CAG ATG TAA

5. Fill in the rest according to nucleotide frequency
TAATGTCAATAAAAAAAAAAATGGCGAATAATAGCTACAGCT | ATTACTAAGCAGTATACAGTATGTCATTATAAGTATAATTACA

Real Dataset (Magana-Mora et al., 2012) **→ Real Black-Box Model (RBBM)**

Combined Dataset → Combined Black-Box Model (CBBM)

Generated by combining the synthetic and real datasets in a 1:1 ratio.

Results

Accuracy and F1 score of SBBM, RBBM, and CBBM have been recorded.

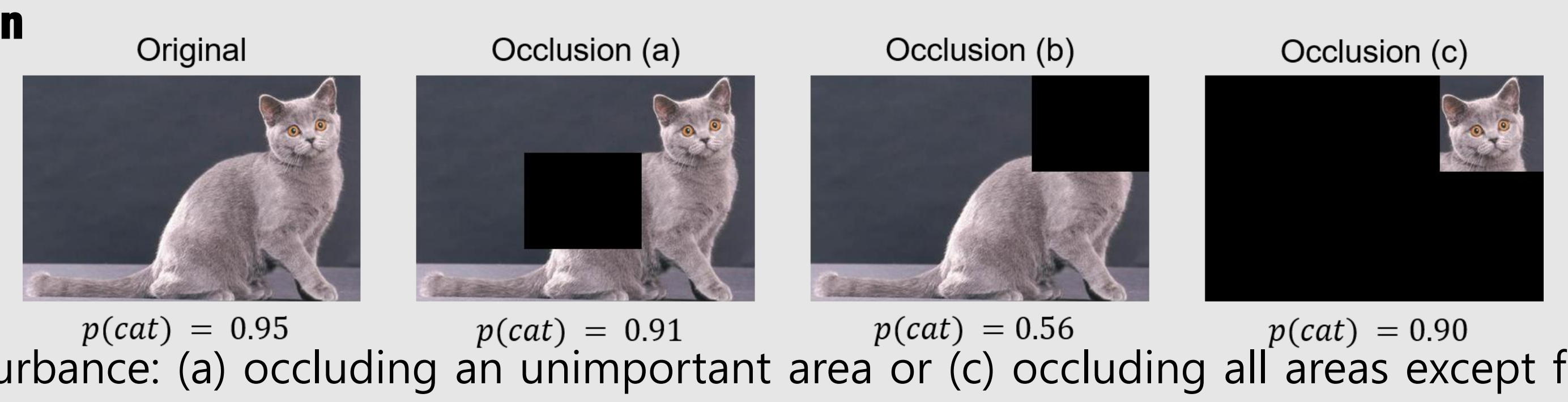
- CBBM learns features from both the real and synthetic datasets.
- The features of SBBM may be overlapping with those of RBBM.

Evaluation of all models (in %)

Models	Accuracy		F1 score	
	Mean	SD	Mean	SD
SBBM	85.53	0.099	85.42	0.122
RBBM	90.74	0.120	90.64	0.160
CBBM	90.68	0.126	90.68	0.163

Feature Analysis

Occlusion



Perturbance: (a) occluding an unimportant area or (c) occluding all areas except for an important area keeps prediction stable while (b) occluding an important area influences prediction greatly.

Occlude only the feature of interest

The lower the value, the more influential the feature.

Evaluation of SBBM with missing features.

Missing feature	F1 score (%)	
	Mean	SD
Consensus sequence	83.45	0.099
Upstream ATG	85.84	0.212
Downstream stop codon	84.42	0.273
Donor splice site	85.78	0.104

Occlude all featues but the feature of interest

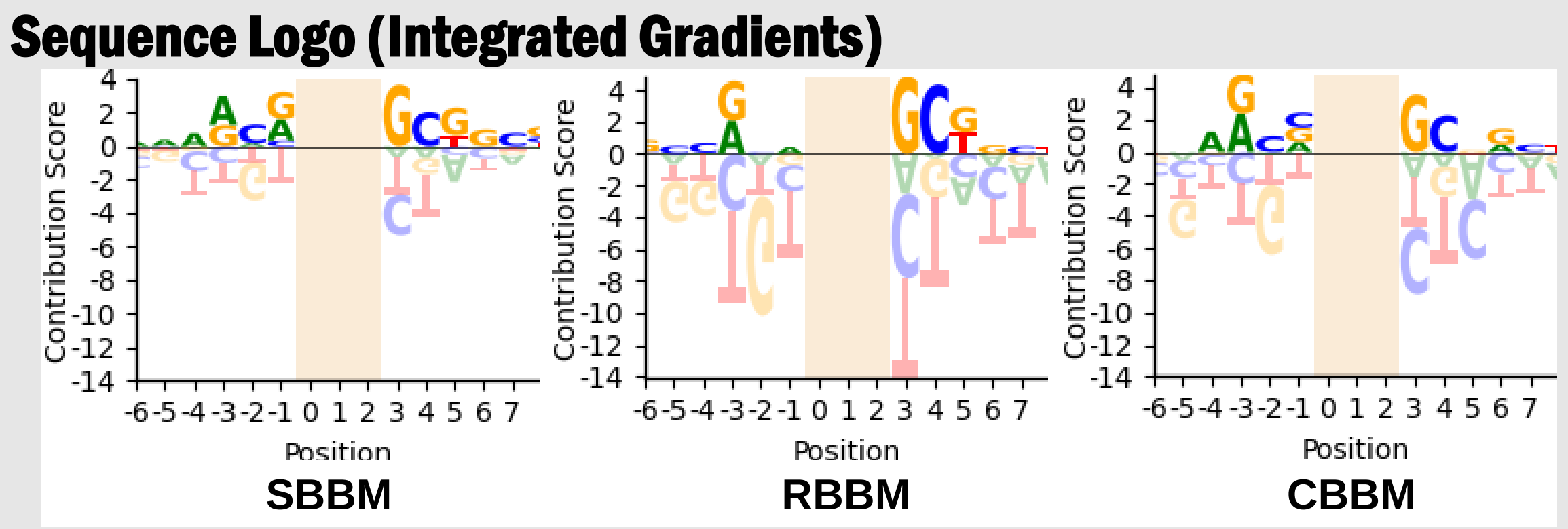
The higher the value, the more influential the feature.

Evaluation of SBBM with single features.

Single feature	F1 score (%)	
	Mean	SD
Nucleotide frequency	80.89	0.369
Consensus sequence	84.75	0.252
Upstream ATG	79.22	0.705
Downstream stop codon	83.46	0.199
Donor splice site	81.92	0.205
Codon usage	62.90	0.840

- Nucleotide frequency and consensus sequence are the most positively influencing features.
- Downstream stop codon has a slight positive influence on TIS prediction.
- Upstream ATG and donor splice site are neutrally influencing or slightly negatively influencing features.
- Codon usage is a negatively influencing feature.

Sequence Logo (Integrated Gradients)



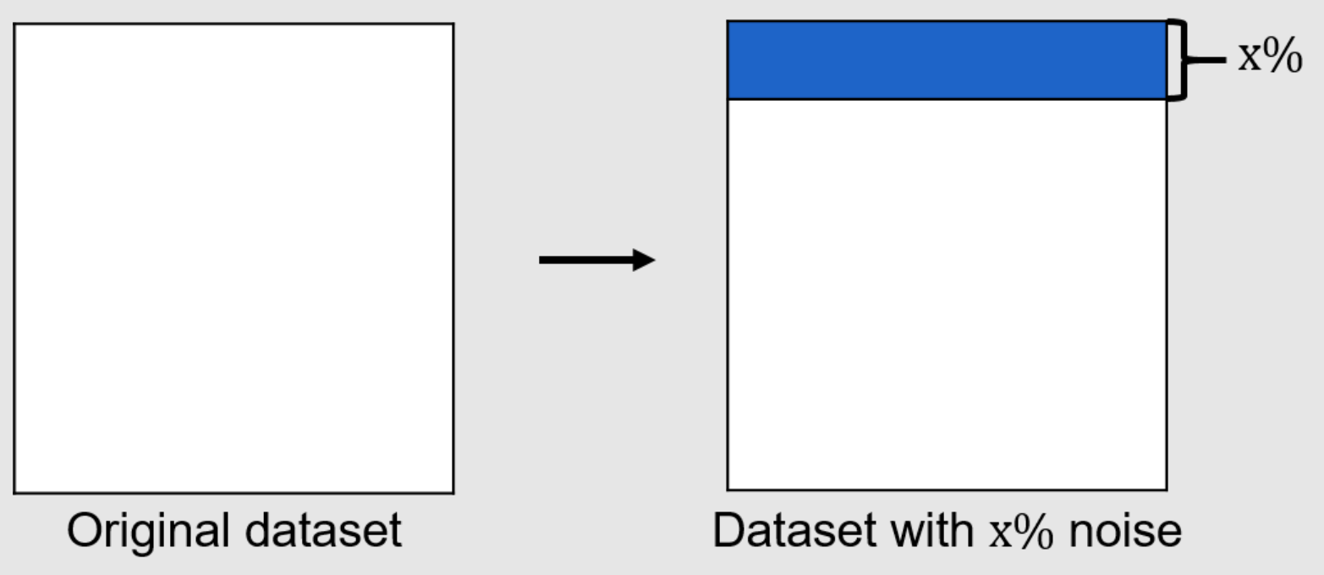
In the sequence logo, only the consensus sequence (positions -5 to +7) are shown, with positions 0 to 2 denoting the TIS

Important consensus nucleotides of SBBM are similar to those of RBBM, albeit in different magnitudes.

Noise Analysis

Equalized Loss of Accuracy (ELA)

$$\mathrm{ELA}_{x\%} = \frac{100 - \mathrm{A}_{x\%}}{\mathrm{A}_{0\%}}$$ (Sàez et al., 2016)



Class Noise (Misclassification)

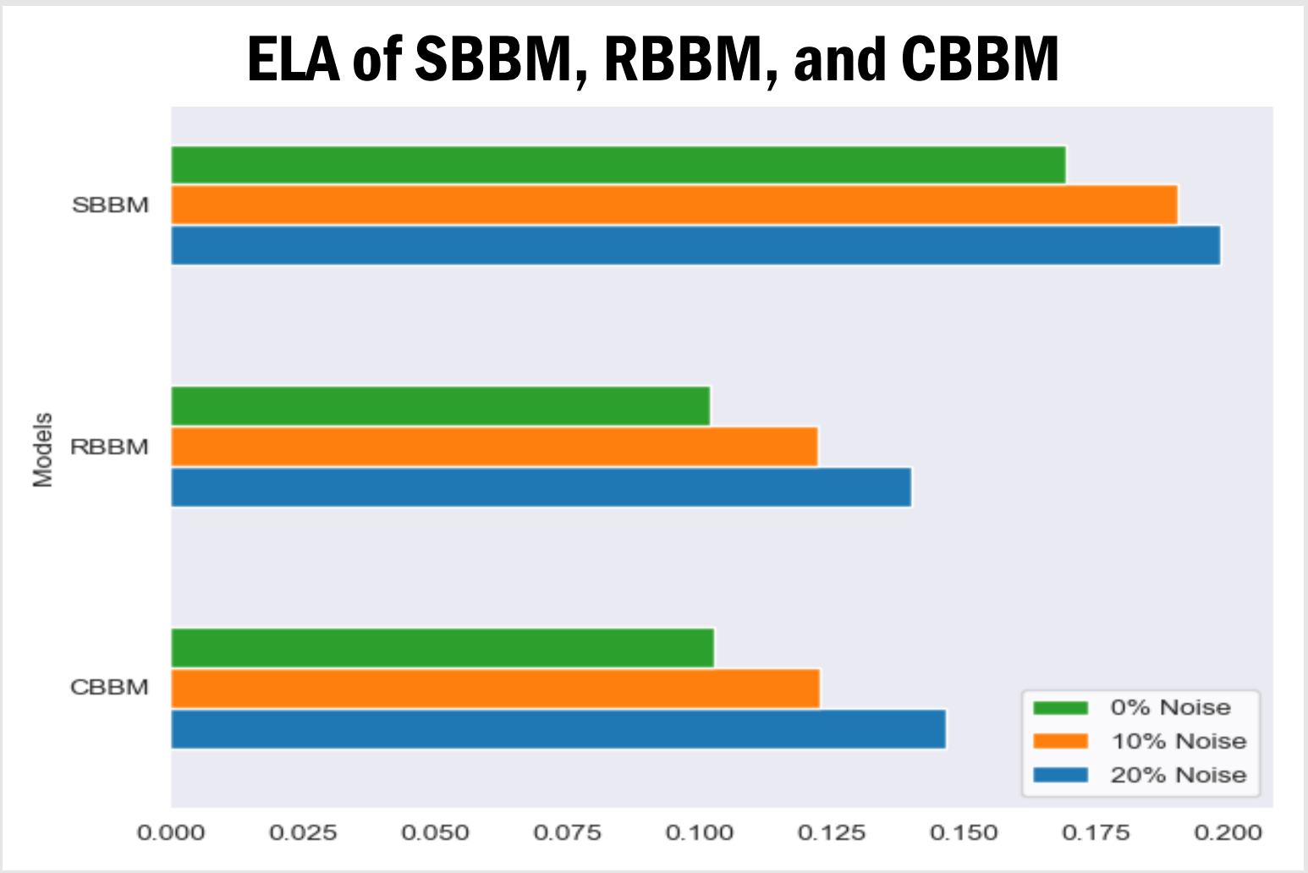
Switch positive and negative datasets.

Attribute Noise (Incorrect Features)

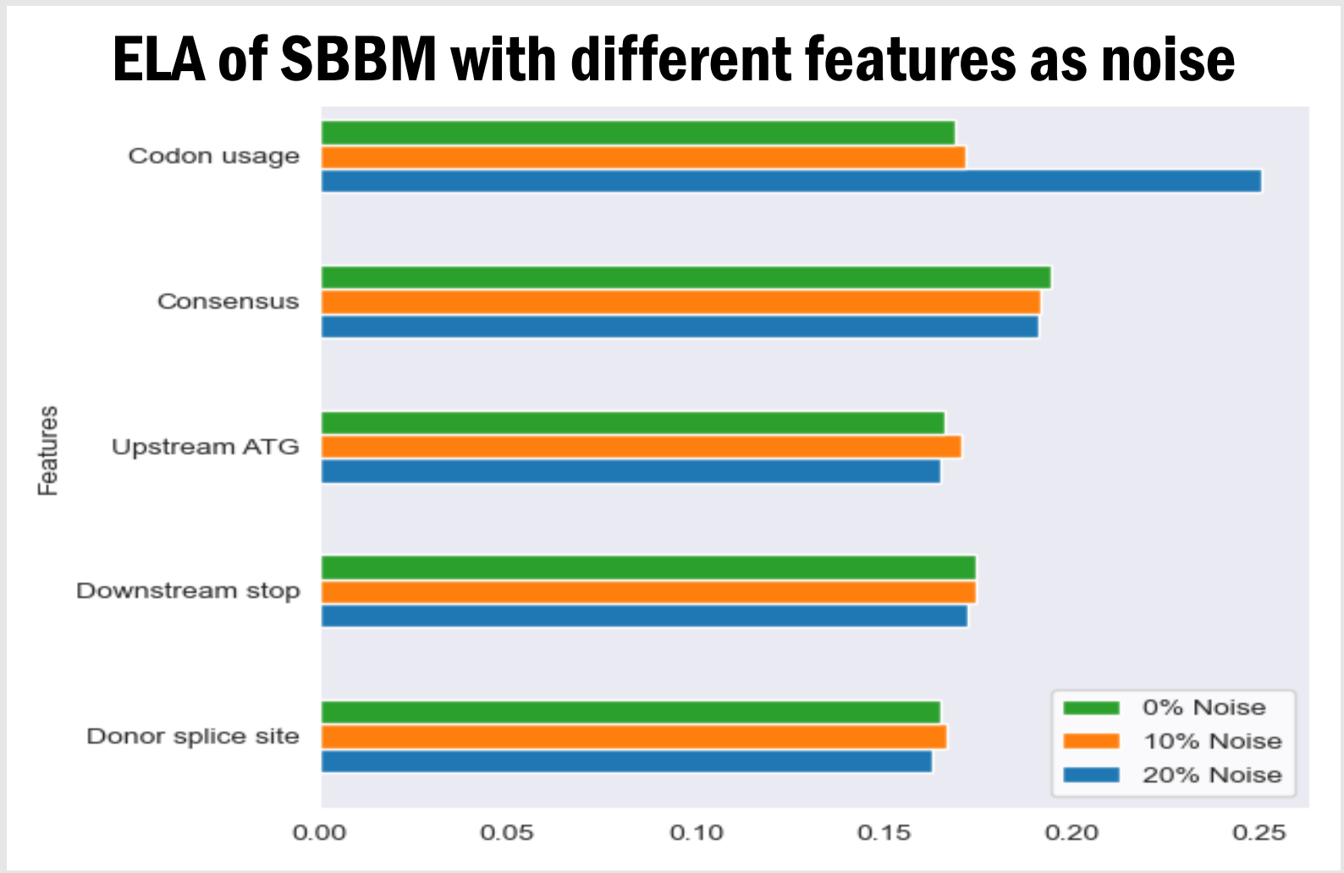
Synthetic features are considered as 'incorrect.'

Results

ELA of SBBM, RBBM, and CBBM



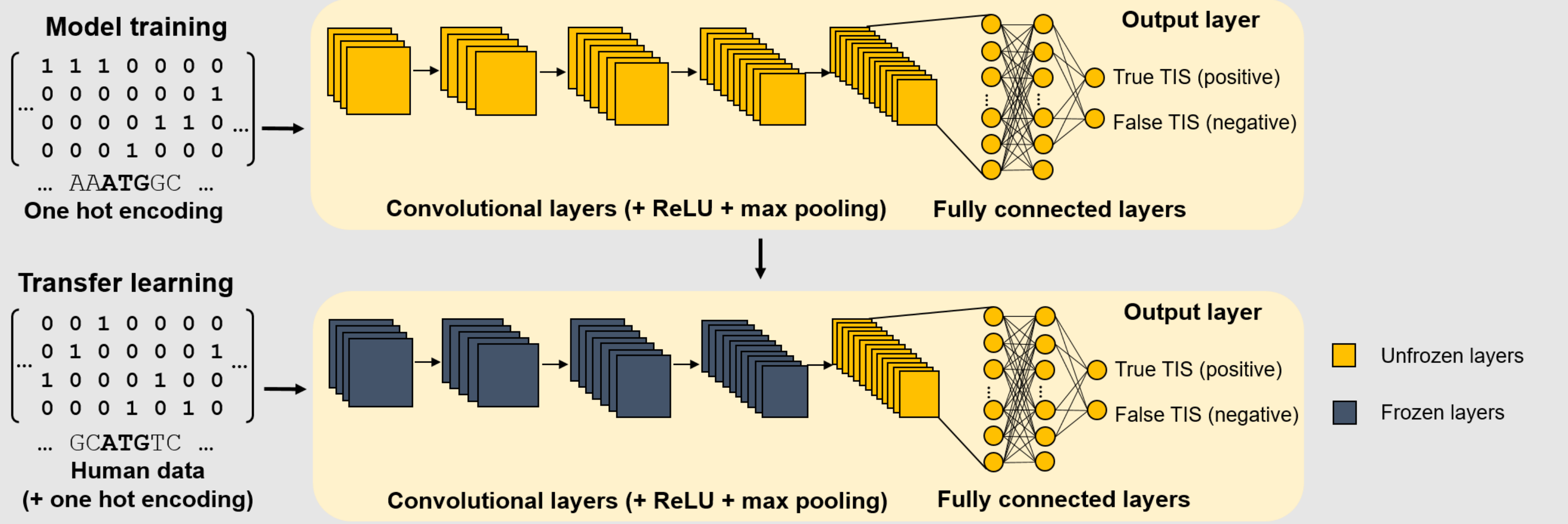
ELA of SBBM with different features as noise



The lower the ELA value, the better the model deals with noise.

- CBBM behaves similarly to RBBM and does not show a high spike like SBBM.
- Consensus sequence is the most contributing feature.

Transfer Learning



The previously trained models (SBBM, RBBM, CBBM) were used as pre-trained models, with some of their layers frozen, to train on human data (Chen et al., 2014), whose small size causes overfitting.

Results

Evaluation of models trained with human data (in %)

Pre-trained models	Accuracy		F1 score	
	Mean	SD	Mean	SD
None	78.40	1.744	81.37	1.180
SBBM	78.80	2.384	81.28	1.513
RBBM	85.40	0.860	86.16	0.613
CBBM	83.75	0.316	84.90	0.184

- CBBM and RBBM increased prediction effectiveness and reduced overfitting
- RBBM gives better results than CBBM.
- CBBM has the potential of being made into a general model for TIS prediction, making it possible to cover other data-insufficient species.

References

Chen, W., Feng, P.-M., Deng, E.-Z., Lin, H., & Chou, K.-C. (2014). iTIS-PseTNC: A sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Analytical Biochemistry*, 462, 76–83. doi: 10.1016/j.ab.2014.06.022.

Magana-Mora, A., Ashoor, H., Jankovic, B. R., Kamau, A., Awara, K., Chowdhary, R., Archer, J. A., & Bajic, V. B. (2012). Dragon TIS Spotter: an Arabidopsis-derived predictor of translation initiation sites in plants. *Bioinformatics*, 29(1), 117–118. doi: 10.1093/bioinformatics/bts638.

Sàez, J. A., Luengo, J., and Herrera, F. (2016). Evaluating the classifier behavior with noisy data considering performance and robustness: The Equalized Loss of Accuracy measure. *Neurocomputing*, 176:26–35. doi:10.1016/j.neucom.2014.11.086.

Conclusions and Future Work

- SBBM learns similar features as RBBM, albeit in different magnitudes.
- CBBM has a similar effectiveness as RBBM, with its usage reducing overfitting when training on small data.
- Consensus sequence and nucleotide frequency have the most significant (positive) influence on TIS prediction.
- Codon usage has a negative influence on prediciton effectiveness, which could be further investigated.
- In future research, a more sophisticated and generalized synthetic dataset could be generated, to be used effectively for data-insufficient use cases.
- Furthermore, new data-efficient strategies may be unlocked for data-hungry models in genomics.

Contact

Yunseol.Park@ghent.ac.kr

Universiteit Gent
@ugent; @ugentkorea
Ghent University; Yunseol Park

GHENT UNIVERSITY
GLOBAL CAMPUS

Center for
Biosystems & Biotech Data Science