

Theses of the PhD thesis

**Transcription Factor Binding Site Detector
Neural Networks trained with Various DNA
Representations**

Gergely Pap

Supervisors:

László Tóth, PhD, Associate Professor

Hegedűs Zoltán, PhD, Senior Research Fellow

Doctoral School of Computer Science

University of Szeged

Department of Computer Algorithms and Artificial Intelligence

2023

1 Introduction

Oftentimes in life sciences an abundance of data is present where the processing of said data requires manual labour. Extracting meaningful patterns costs a lot of time and energy which must be covered by the professionals of the given field. As such, automatisation of the information extraction pipeline is an area of active research. Approaches from artificial intelligence, machine learning and deep learning are all applied to datasets in order to automatically solve a task, be it classification or regression, the two most widely spread tasks regarding life sciences. Recently deep learning has made enormous steps in the research areas of computer vision and natural language processing, where the architectures and methods considered to be the most advanced are being utilized. However many of these complex and innovative techniques can be applied to other fields, such as bioinformatics and computational biology. Most deep learning methods thrive when data is abundantly available. Modern analytical tools - for example Next Generation Sequencing - produce huge amounts of raw information. Therefore the processing of nucleotide sequences with deep learning models is an effective pairing of the two rapidly developing techniques.

1.1 Transcription Factors in Biology

One of the most important processes in a cell's biology is an event called transcription. In molecular biology, the central dogma describes transcription as the DNA's conversion to RNA. This complex pipeline has profound effects on the cell's life as it controls and regulates the expression of genes. Understanding the mechanics and rules behind transcription is a long-standing challenge in biology. Transcription Factors (TFs) are proteins that regulate transcription, either by inhibiting or by promoting the process. Thus revealing information about the workings of TFs can be key to gaining new insights into the regularization of gene expression.

1.2 Problem statement and motivation.

In this thesis I have explored and analysed the pairing of deep learning with a DNA-protein binding classification task. As detecting proteins that bind DNA at specific locations is expensive experimentally, approaches that rely on data-driven methods are practical and sought after. Deep Neural Networks (DNNs) and especially Convolutional Neural Networks (CNNs) have made an impact in this field by proving to be excellent detectors of Transcription Factor Binding Sites (TFBSs) [9]. Once trained, inference on unknown entities is generally quick and accurate for most datasets. On the other hand, despite these encouraging results, there are many issues with deep learning on DNA sequences. Firstly, the choice of the learner's architecture plays an important role in the observed performance. Secondly, the representation of DNA as nucleotide-based sequences is the traditional way for network training. This is motivated by the fact that sequencers produce their output in nucleotide format which can be one-hot encoded as input for learning models. However other types of DNA representation could prove to be advantageous for training [14]. Experiments run on different sequence representations show that the various aspects in which they differ from a nucleotide based one enable the learner to exploit patterns not directly present in the aforementioned traditional format, therefore the discovery of train-

ing settings with appropriate network structures promises new and competitive prediction models.

1.3 Overview of the main goals

The PhD thesis presents three key experiments all concerning the biological classification task of transcription factor binding site detection using Deep Learning (DL). The first chapter presents model training with the Functional Group DNA Representation (FGDR), the second chapter shows two deep learning approaches using the Physicochemical and Conformational Descriptors (PCDs) representation of DNA and the third and final chapter sheds light on the problem of robustness and robust network training with regards to TFBS classification.

2 Classification using a functional group-based data representation

Most machine learning applications for TFBS classification use a nucleotide based representation of DNA sequences. Depending on the type of the learner, these sequences are often encoded into vectors using the one-hot encoding scheme. An other popular approach is to use k-mers as additional features extending the single nucleotide resolution data. Other experiments with embedding the k-mer sequences also yielded models with good performance. However, all of these methods rely on nucleotide data, which might have its disadvantages that are not really known since research employing different kinds of representations is scarce in this field. The usage of a different DNA input format can be beneficial as most machine learning approaches are heavily influenced by the type and format of their input data. So it is possible that a new representation might enable the learners to explore new relations.

In Chapter 2 of the PhD thesis, experiments with the FGDR representation are conducted. Several methods for different input formatting are presented as FGDR data was originally intended as a visualization technique [8]. I found that the properties of the input space have a profound impact on the learning trajectories of the models. Therefore I conducted a systematic search to find the set of options controlling the FGDR representation's preprocessing for optimal model training. After fixing the input details I continued to optimize the architecture of the CNNs. Finally in an ensemble setting, where one learner is using nucleotides and the other FGDR values, I showed an increase in performance surpassing standalone approaches.

2.1 Input format for model training

Originally the FGDR values are calculated based on a nucleotide sequence. Its numerical values range from 0 to 8 (or 9) and it is a $7 \times L$ matrix where L is the sequence length and the 7 rows represent the topological positions of the different chemical functional groups.

I presented experiments with several possible FGDR preprocessing and formatting steps, during which I varied the extent of the numerical values' range (between 8 or 9 - depending on an electrochemical choice). Furthermore I trained networks with an input matrix

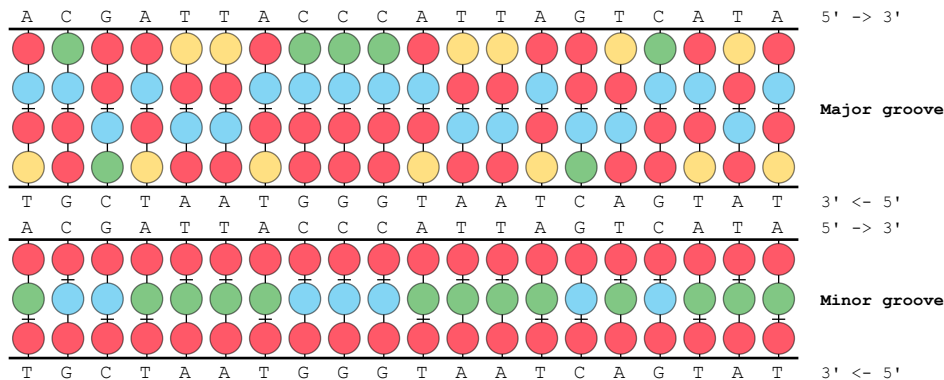


Figure 1: *FGDR Space - Illustration taken from `brc. drv. hu` [8]*

resembling the default FGDR values, but a one-hot encoded format resulted in better performances.

2.2 Optimal architecture

After observing CNNs trained with FGDR input classify the sequences with reasonable accuracy, I started to search for a set of hyper-parameters that were more suited to the representation at hand. A thorough and extensive exploration of the search space included the number of layers (both convolutional and dense), choice of the optimizer and its learning rate and regularization values (e.g., dropout probabilities and the strength of weight decay).

Table 1: *AUC scores*

Method	DeepBind	Zeng	Nuc.	FGDR	Ens.
AUC	0.863472	0.904524	0.9142	0.9145	0.9171

2.3 Ensemble learning

To further improve the performance of the learners, I trained models using both input formats (FGDR and nucleotide) in an ensemble setting (results shown in Table 1). Averaging the models output probabilities yielded further correct classifications, surpassing contemporary approaches, such as the work of Zeng et al. or DeepBind [9, 17]. The findings of this thesis point have been published in a conference proceedings.

3 Training models with physico-chemical features

Building upon the previously presented idea (in Section 2), and using an other representation format from DRV [8], we experimented with TFBS classifier training with physico-chemical features. The new Physico-Chemical Descriptor (PCD) based DNA representation is suitable for learning models to detect binding sites. In Table 2 a sample of PCD values

Table 2: *PCD sample*

	Positions	1	2	3	4	5	6	7	8	9	10
	Nucleotides	C	C	A	G	C	C	C	C	T	G
Physico-chemical Descriptors	Aida BA transition	2.26	5.1	0.79	8.28	2.26	2.26	2.26	0.79	5.1	2.26
	A-philicity	0.19	1.04	0.33	0.73	0.19	0.19	0.19	0.33	1.04	0.19
	Base stacking	-8.26	-6.57	-6.78	-14.59	-8.26	-8.26	-8.26	-6.78	-6.57	-8.26
	B-DNA twist	35.3	37.7	30.6	38.4	35.3	35.3	35.3	30.6	37.7	35.3
	Bending stiffness	130	60	60	85	130	130	130	60	60	130
	Breslauer dG	3.1	1.9	1.6	3.1	3.1	3.1	3.1	1.6	1.9	3.1

based on a nucleotide sequence is shown. We theorize that showing the network the values describing the physical and chemical state of the nucleotide pairs or trios could be beneficial as it might contain extra information about the binding sites compared to the nucleotide input format.

In this thesis group, I present networks trained on PCDs for TFBS detection. Moreover, as the format of the PCD representation is largely different from the typical input arrangements that are employed for nucleotide based CNNs, I experimented with a novel network architecture to better capture the binding patterns of the sequences. In the literature of CNNs for TFBS classification one usually presents the nucleotides as the channels of the input (quite similarly as how one inputs an RGB image regarding its colour channels to the first layer of a CNN). While the convolutional kernels proved to be able to learn the properties of the binding sites, I show that a different setup of architecture is more advantageous. The main idea is that PCDs often contain information with respect to each other (i.d., along the colour channels or depth axis) and the traditional CNN setup does not make use of these relations during feature extraction. (As the convolutions only happen in 1D separately from the different PCD channels.) The improved network is outfitted with a depthwise separable convolutional layer to better capture the relevant patterns of the PCD representation [11]. We show that the network with Depthwise Separable Convolution using PCD input (denoted as DSC (PCD)) manages to surpass other methods with statistical relevance on several TFBS detection tasks.

Algorithm 1 Creating feature subsets based on their correlation values in case of pcp-2

- 1: INPUT: PCD features for dimers from all pcp-2: (X) set
 - 2: $\text{corr_mat} \leftarrow \text{pearson_corr}(X)$
 - 3: **do**:
 - 4: $x_A, x_B \sim$ pick two PCD feature randomly
 - 5: **if** $\text{corr}[x_A, x_B] > \rho$ **then**
 - 6: remove one randomly
 - 7: add other to pcd_subset_ρ
 - 8: **while**: $\text{pearson_corr} < \rho$ for all element pairs in corr_mat
 - 9: **repeat** from step 3 for $\rho = [0.9, 0.7, 0.5]$
-

Table 3: TFBS classification accuracy with PCDs and nucleotides

TFs	Sp1	Mafk	Cjun	Cmyc	Max	Mxi1
PCD	0.6957	0.9258	0.8320	0.7265	0.7387	0.6946
Nuc.	0.7289	0.9238	0.8462	0.7593	0.7662	0.7290

3.1 Networks trained with physico-chemical features

In the first part of Chapter 3 of the PhD thesis, TFBS detector networks trained with PCDs using a relatively small parameter count are discussed. Moreover subsets of distinct PCD features are shown to have a different impact on model performance. For feature selection we developed a method for reducing the size of the input space regarding PCD subsets (see Algorithm 1). The network architecture remained similar to the CNN presented by Zeng et al. [17]. The number of trainable weights were kept low to be as close as possible to the referenced work in this task for better comparisons regarding the representational choice and classification performance. Results are in Table 3. A model structure optimized for PCDs is presented in the following subsection.

3.2 Depthwise separable convolutions for PCDs

In the second part of Chapter 3 I presented a novel structural approach for training networks with PCDs. The new model structure makes use of depthwise separable convolutions in order to learn additional relationships from the PCD features. The described architecture outperforms other methods on several datasets. The observed performance in various metrics are presented in Table 4.

Table 4: Performance of the different models and representations on the 50 datasets

Model type	ACC	AUC	AUPR
CNN-ARCH (nuc)	0.8492	0.8491	0.8045
TBiNet (nuc)	0.7849	0.7847	0.7413
DSC (nuc)	0.8515	0.8515	0.8062
CNN (PCD)	0.8243	0.8242	0.7767
DSC (PCD)	0.8555	0.8554	0.8123

4 Translational Robustness of Nucleotide Sequence Classifiers

In this thesis point I have experimented with the nucleotide representation to train neural networks. Recent works [13, 15, 16] suggest that suitable DL models trained on a nucleotide representation are interpretable. That is, researchers are able to explain to some degree the DL models' decision making process or classification mechanism. While interpretability is a desired property of machine learning approaches and several claims hold true regarding modern architectures trained for TFBS detection tasks, there are probably many unknowns still about the models' inner behaviour. To further support the chance

Table 5: Three shifting methods for evaluation and training. S means HaibH1hesSp1Pcr1x.

TF	train	discovery			occupancy		
		evaluation strat.			evaluation strat.		
		No Shift	Rnd	Worst	No Shift	Rnd	Worst
S-95	No Shift	0.7466	0.7463	0.6834	0.7542	0.7544	0.7017
	Rnd	0.7468	0.7456	0.6894	0.7505	0.7507	0.6891
	Worst	0.7607	0.7620	0.7212	0.7563	0.7571	0.7265
S-101	No Shift	0.7578	n/a.	n/a.	0.7537	n/a.	n/a.

of misinterpretation issues, in the field of robust neural networks and adversarial training recent advancements highlighted several pressing concerns in connection to interpretability. To our knowledge, the robustness of DNA-protein binding classifier models were not examined previously, so here we present measurements to determine the translational robustness of two different networks. After observing significant performance losses, we implement a strategy to increase the learners’ accuracy when faced with adversarial attacks.

4.1 Creating adversarial examples

To our knowledge, no adversarial attacks were launched against TFBS classifier models. Although the input matrix of a nucleotide sequence can resemble a binary image, modifying the content is not as straightforward as in the case of computer vision or image classification tasks. While small perturbations to an image of a cat might be unrecognisable for humans and can be misleading for networks, the same cannot be said about DNA sequences. It is hard to say that a modified sequence will still be a binding one. Moreover the values are discrete binaries and not floats thus changing a nucleotide base can result in the destruction of the original ground truth labelling, because the new sequence might not be bound in vivo by the protein in question. Considering these issues we came up with shifting strategies during which the original sequence (and the binding site) remains unchanged in terms of the nucleotide content. Cropping from the flanking regions, randomly cropping from both ends, and cropping based on the loss of all possible crop positions are examined.

4.2 Augmenting training

We observed significant performance drops when the adversarial entities were used for inference. Incorporating these shifting strategies as an augmentation method and training the networks while some of the sequences in a mini-batch undergo transformations resulted in models that are more robust (results shown in Table 5).

4.3 Extracting motifs from the learners

In order to measure and observe the differences regarding humanly interpretable features between the robustly trained and the unmodified networks, I extracted PWMs from the

Table 6: *E-values for the extracted motifs*

Kernels	Training	<i>Mafk</i> occupancy	<i>Znf</i> occupancy	<i>Sp1</i> occupancy
All	Non-robust	0.281168	0.331671	0.442
	Robust	0.252381	0.302872	0.425485
5 best	Non-robust	3.18E-06	0.002226	0.002219
	Robust	1.56E-06	0.000315	0.001415

neurons of the first convolutional layer. For each convolutional weight matrix, I scanned over the test sequences containing binding sites and noted the positions of highest activation. Using the nucleotide frequencies gathered from these positions I constructed PPMs and I visualised them as sequence logos. I determined the binding motifs present in the test sequences by using MEME [10]. Then using Tomtom [12] I compared the matrices that were extracted from the neurons with the matrices discovered by MEME.

In Table 6 a comparison for the motifs is shown. The E-values are calculated by Tomtom from the MEME Suite. Lower E-values mean better matches. The occupancy task’s positive test entities were used to establish the reference matrices. For the three TFs (*Mafk*, *Sp1* and *Znf143*) the calculated E-values were averaged over all kernels. In addition, the five kernels with the lowest E-values were selected for comparisons. The networks were the same ones from before and their performances are reported above. The non-robust networks were all middle crop variants with 90 length inputs. The *Mafk* and *Znf143* networks were worst-crop variants, but in the case of *Sp1* the random crop one was used. We can see in Table 6 that in all cases augmented training resulted in better motifs.

5 Contributions of the thesis

In the **first thesis group**, my contributions are related to transcription factor binding site detection with a new input description, FGDR. Instead of using the traditional nucleotide format, I trained neural networks with a DNA representation based on functional groups. I managed to outperform concurrent methods and show that an ensemble technique yields even further improvements. Detailed discussion can be found in Chapter 2.

- I / 1. I experimented with finding a suitable FGDR input format for learning convolutional models and I presented a way for convolutional neural network training on the modified representation containing functional group information from DNA.
- I / 2. I introduced an optimal network architecture, and showed that the learned models perform the detection task with high proficiency.
- I / 3. I showed that an ensemble training scenario (using the nucleotide and the FGDR models together) is beneficial and yields additional accuracy gains.

In the **second thesis group**, my contributions are training TFBS classifiers using the PCD representation, showing that selecting a good subset of PCDs (thus reducing computational

costs) is enough for competitive performance. In addition I presented a novel architecture using depthwise separable convolutions for training classifiers with good performance scores measured in various accuracy metrics on PCD datasets. The detailed discussion can be found in Chapter 3.

- II / 1. I showed that training networks with the PCD representation can produce competitive classification results.
- II / 2. I proposed a feature selection approach to reduce the number of input features and speed up training while preserving most of the classificational performance.
- II / 3. I designed a new model structure for learning CNNs with PCD input format using depthwise separable convolutions.
- II / 4. I showed that the proposed network with the PCD representation can produce accurate classifications outperforming other model-representation pairs.

In the **third thesis group**, I examined the robustness of TFBS detectors. I showed that shifting the input sequences can result in a significant drop of performance. I proposed an augmentation method for training more robust classifiers. In Chapter 4 of the dissertation an in-depth discussion of this topic can be found.

- III / 1. I showed that DNA-protein binding detectors are easily misled by cropping the input test sequences.
- III / 2. I proposed three shifting strategies to evaluate the vulnerabilities of the TFBS models to adversarial examples. I showed that both a smaller model trained on shorter sequences and a more advanced model with a larger dataset fail to correctly classify a significant number of cropped entities.
- III / 3. I designed an augmentation method for training, and proved that introducing shifted examples during fitting can improve robustness and classification accuracy.

Table 7 summarizes the relation between the thesis points and the corresponding publications.

Table 7: *Correspondence between the thesis points and my publications*

Publication	Thesis point									
	I/1	I/2	I/3	II/1	II/2	II/3	II/4	III/1	III/2	III/3
[1]	•	•	•							
[2]				•	•					
[3]						•	•			
[4]								•	•	•

The author's publications on the subjects of the thesis

Full papers in conference proceedings

- [1] **G. Pap**, Z. Györgypál, K. Ádám, L. Tóth. and Z. Hegedűs. Transcription factor binding site detection using convolutional neural networks with a functional group-based data representation. In *Journal of Physics: Conference Series, Volume 1824, The 2020 International Conference on Artificial Intelligence and Application Technologies (AIAT 2020)*, IOP Publishing, 012001, 2021.
- [2] **G. Pap**, K. Ádám, Z. Györgypál, L. Tóth. and Z. Hegedűs. Training models employing physico-chemical properties of DNA for protein binding site detection. In *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*, IEEE, 1-5, 2021.
- [3] **G. Pap**, K. Ádám, Z. Györgypál, L. Tóth. and Z. Hegedűs. Depthwise Convolutions using Physicochemical Features of DNA for Transcription Factor Binding Site Classification. In *The 6th International Conference on Advances in Artificial Intelligence (ICAAI 2022)*, ACM, 00, 2022.
- [4] **G. Pap** and I. Megyeri. Translational Robustness of Neural Networks Trained for Transcription Factor Binding Site Classification. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022)*, SciTePress, Volume 3: 39-45, 2021.

Further related publications

- [5] **G. Pap**, G. Lékó, T. Grósz. A Reconstruction-Free Projection Selection Procedure for Binary Tomography Using Convolutional Neural Networks. In *International Conference on Image Analysis and Recognition (ICIAR 2019)*, Springer, Cham, 228-236, 2019.
- [6] **G. Pap**, L. Tóth. A Comparison of Supervised and Semi-supervised Training Algorithms of Restricted Boltzmann Machines on Biological Data. In *2019 IEEE 19th International Symposium on Computational Intelligence and Informatics and 7th IEEE International Conference on Recent Achievements in Mechatronics, Automation, Computer Sciences and Robotics (CINTI-MACRo)*, IEEE, 000023-000028, 2019.

Journal publications

- [7] D. Varga, Sz. Szikora, T. Novák, **G. Pap**, G. Lékó, J. Mihály & M. Erdélyi Machine learning framework to segment sarcomeric structures in SMLM data. In *Scientific Reports*, Vol. 13, No. 1 p. 1582, 2023.

Other References

- [8] Krisztian Adam, Zoltan Gyorgypal, and Zoltan Hegedus. DNA Readout Viewer (DRV): visualization of specificity determining patterns of protein-binding DNA segments. *Bioinformatics*, 36(7):2286–2287, 2019.
- [9] Babak Alipanahi, Andrew DeLong, Matthew T. Weirauch, and Brendan J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.
- [10] Timothy L. Bailey, James Johnson, Charles E. Grant, and William S. Noble. The meme suite. *Nucleic acids research*, 43:W39–49, 2015.
- [11] François Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016.
- [12] Shobhit Gupta, John A. Stamatoyannopoulos, Timothy L. Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome Biology*, 8(2):R24, 2007.
- [13] Jack Lanchantin, Ritambhara Singh, Beilun Wang, and Yanjun Qi. Deep motif dashboard: Visualizing and understanding genomic sequences using deep neural networks. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 22:254–265, 2017.
- [14] Wenxiu Ma, Lin Yang, Remo Rohs, and William Stafford Noble. DNA sequence+shape kernel enables alignment-free modeling of transcription factor binding. *Bioinformatics*, 33:3003–3010, 2/14.
- [15] Sungjoon Park, Yookyung Koh, Hwisang Jeon, Hyunjae Kim, Yoonsun Yeo, and Jae-woo Kang. Enhancing the interpretability of transcription factor binding site prediction using attention mechanism. *Scientific Reports*, 10(1):13413, 2020.
- [16] Dailun Wang, Qinhu Zhang, Chang-An Yuan, Xiao Qin, Zhi-Kai Huang, and Li Shang. Motif discovery via convolutional networks with k-mer embedding. In De-Shuang Huang, Kang-Hyun Jo, and Zhi-Kai Huang, editors, *Intelligent Computing Theories and Application*, pages 374–382, Cham, 2019. Springer International Publishing.
- [17] Haoyang Zeng, Matthew Edwards, Ge Liu, and David Gifford. Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics*, 32:i121–i127, 2016.

6 Összefoglalás

Az értekezés mély tanulási módszereket mutat be orvosbiológiai adatokon. Az elsődleges feladat a fehérje kötő DNS szekvenciák detektálása neurális hálózatokkal. A bemutatott megközelítések közös vonása a mély tanuló modellek vizsgálata nukleotid vagy egyéb adatábrázolási megközelítés esetében.

A munka három fő témakörből áll. Az első fejezetben a funkciós csoportokat ábrázoló, a másodikban a fiziko-kémiai, míg a 3. fejezetben a nukleotid alapú megközelítések vizsgálata olvasható.

Az osztályozás funkciós csoportokra épülő reprezentációval című fejezetben a neuronális hálózatokat nem a hagyományos nukleotid alapú szekvenciákkal tanítom, hanem egy új vizualizációs módszer adatábrázolási megközelítésével. A szekvenciákat dinukleotidokból számolható értékekkel jellemezzük, a funkciós csoportok elektrokémiai viselkedéséből alakítunk ki bemeneti jellemzőket. Mivel nem triviális ennek a formának a felhasználása konvolúciós rétegek tanításához, első lépésként a különböző formai elrendezésekkel illetve előfeldolgozási lépésekkel foglalkoztam. Azután bemutattam egy olyan modell architektúrát, amely kiemelkedő teljesítményt ér el transzkripció faktor kötőhely detekciós feladatok esetében. Végül elkészítettem egy együttes (ensemble) modellt, ahol a nukleotidokra és a funkciós csoportokra épülő hálók becsléseit átlagolva a kimeneteknél még további fejlődést értem el.

A modellek tanítása fiziko-kémiai jellemzőkkel című fejezetben szintén egy, a nukleotidoktól eltérő adatábrázolási módszer segítségével tanítottam osztályozókat. Az új reprezentáció a DNS szál különböző fizikai és kémiai tulajdonságait írja le folytonos értékekkel. A fejezet első felében bemutattam, hogy ezen a bemeneti fajtán is taníthatóak modellek, amelyek teljesítménye az ismertebb megoldásokhoz hasonlóan teljesít. Továbbá egy jellemzőválogatásos módszer segítségével csökkenthető a bemenő jellemzők száma, így csak kis osztályozási hibanövekedés mellett gyorsabbak és olcsóbbak a tanítások. A fejezet második részében egy olyan megközelítést mutatok be, amely lehetővé teszi a mély tanulónak, hogy új összefüggéseket vegyen észre a fiziko-kémiai reprezentációban. A módszer lényege az, hogy a hálózat architektúrájában mélységi szétválasztható konvolúciós réteget használok, amely az eddig közvetlenül nem tanulható mélységi dimenzió mentén is tanulhatóvá tette az összefüggéseket. Így több, azonos feladatra publikált és ismert modell teljesítményét sikerült számos adathalmazon felülmúlnom.

A nukleotid szekvenciákra épülő osztályozók translációs robusztussága című fejezetben mesterséges intelligencián alapuló DNS-fehérje kötő detektorok robusztusságát és az ellenük felhasználható ellenséges példák előállításának lehetőségeit vizsgáltam. A feltevés az volt, hogy túlságosan érzékenyek ezek a modellek egyéb tényezőkre, amelyek a valós címkét (azaz a szekvencia biológiai funkcióját) nem befolyásolják. Továbbgondolva, ha arrébb toljuk a szekvenciákat úgy, hogy a kötőhely (tehát a meghatározó jellemző) érintetlen marad, akkor azt várnánk, hogy a modellek ettől függetlenül felismerik és helyesen döntenek. Azonban azt tapasztaltuk, hogy egy pár nukleotidos hosszanti eltolás is elegendő ahhoz, hogy félrevezessük a modelleket. Kidolgoztam három különböző eltolási stratégiát, amelyek alkalmazásakor a kiértékelt hálózatok pontosságbeli romlást szenvednek el. Ezen felül megadtam egy augmentációs tanítási módszert, amely segítségével a robusztus pontosság növelhető, így a hálózatok kevésbé vagy egyáltalán nem lesznek érzékenyek a vágásokra \eltolásokra.

Társszerzői nyilatkozat

Kijelentem, hogy ismerem **Pap Gergely** PhD fokozatra pályázó „**Transcription Factor Binding Site Detector Neural Networks trained with Various DNA Representations**” című disszertációját. A disszertációban szereplő és a

1. Gergely Pap, Krisztián Ádám, Zoltán Györgypál, László Tóth and Zoltán Hegedűs: Transcription factor binding site detection using convolutional neural networks with a functional group-based data representation; Journal of Physics: Conference Series, Vol. 1824, No. 1 IOP Publishing p. 012001
2. Gergely Pap, Krisztián Ádám, Zoltán Györgypál, László Tóth and Zoltán Hegedűs: Training models employing physico-chemical properties of DNA for protein binding site detection; 2021 International Conference on Applied Artificial Intelligence (ICAPAI), 2021, pp. 1-5,
3. Gergely Pap, Krisztián Ádám, Zoltán Györgypál, László Tóth and Zoltán Hegedűs: Depthwise Convolutions using Physicochemical Features of DNA for Transcription Factor Binding Site Classification; 2022 The 6th International Conference on Advances in Artificial Intelligence (ICAAI 2022)

cikkekben publikált közös eredményekre vonatkozóan kijelentem, hogy a következő eredményekhez való hozzájárulásunk oszthatatlan:

- Az adatrepresentációval kapcsolatos kísérletek megtervezése mély tanuláshoz [1; 2; 3]

A következő eredményekre vonatkozóan kijelentem, hogy a pályázó hozzájárulása volt a meghatározó:

- A DNS-fehérje kötőhelyek detekciójával kapcsolatos szakirodalom feldolgozása [1; 2; 3]
- A neurális hálózatok fejlesztése, tanítása és kiértékelése [1; 2; 3]
- Az eredmények értelmezése, összehasonlítása, statisztikák készítése [1; 2; 3]

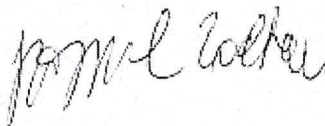
és ezeket az eredményeket nem használtam fel és a jövőben sem használom fel tudományos fokozat megszerzéséhez.

A következő eredményekben a társszerzők hozzájárulása volt a meghatározó:

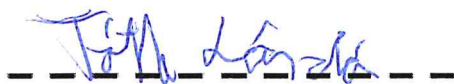
Az adatok (nukleotid szekvenciák) formátumának átalakítása [1; 2; 3]



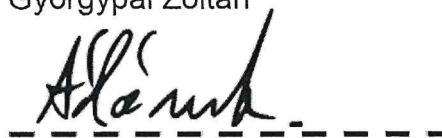
Hegedűs Zoltán



Györgypál Zoltán



Tóth László



Ádám Krisztián

Társszerzői nyilatkozat

Kijelentem, hogy ismerem **Pap Gergely** PhD fokozatra pályázó „**Transcription Factor Binding Site Detector Neural Networks trained with Various DNA Representations**” című disszertációját. A disszertációban szereplő

1. Pap G. és Megyeri I. (2022). Translational Robustness of Neural Networks Trained for Transcription Factor Binding Site Classification. In Proceedings of the 14th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART, ISBN 978-989-758-547-0, pages 39-45. DOI: 10.5220/0010769100003116,

cikkben publikált közös eredményekre vonatkozóan kijelentem, hogy a következő eredményekhez való hozzájárulásunk oszthatatlan:

- Az támadásokkal kapcsolatos kísérletek implementásása, a hálózatok kiértékelése [1]

A következő eredményekre vonatkozóan kijelentem, hogy a pályázó hozzájárulása volt a meghatározó:

- A DNS-fehérje kötőhelyek detekciójával összefüggő támadások megtervezése [1]
- Az adatbázisok előkészítése, a neurális hálózatok tanítása [1]
- A robusztusságot növelő augmentációs tanítási módszer elkészítése [1]

és ezeket az eredményeket nem használtam fel és a jövőben sem használom fel tudományos fokozat megszerzéséhez.

A következő eredményekben a társszerző hozzájárulása volt a meghatározó:

- Az ellenséges példákat és a robusztus tanítást illetően az eszköztár kidolgozása [1]

Megyeri István

Szeged, 2023. február 20.

