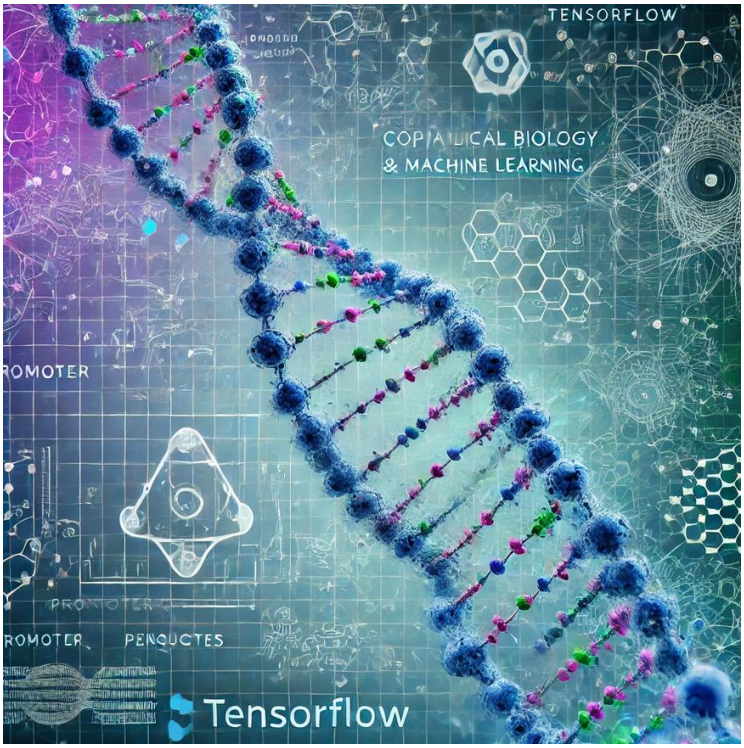
Computational prediction of promotors using the machine learning technique

Using TensorFlow to train a Convolutional Neural Network (CNN) model for predicting transcription factor binding sites of PhoP



Master research project

Msc Biology

Under supervision of Prof. Dr. Anne de Jong

Josep Bellés Roca (s4052900)

October 19, 2024

1. Abstract

This study explores the use of Convolutional Neural Networks (CNNs) to predict PhoP transcription factor binding sites across diverse bacterial species. By leveraging machine learning techniques, this approach aims to overcome experimental limitations in identifying DNA patterns where transcription factors interact.

We developed a predictive model for identifying protein sequences by constructing an initial training dataset of 169 sequences, which was refined through iterative improvements. Early results indicated high precision but low recall, highlighting missed true positives. To improve sensitivity, we expanded the positive sequence dataset, balanced the ratio of positive to negative sequences, and optimized model parameters. Through further hyperparameter tuning and the use of sequence padding to focus on relevant motifs, the model achieved a final precision of 0.95, recall of 0.85, and an F1 score of 0.9, demonstrating enhanced robustness and generalizability.

Overall, this study highlights the potential of CNN models in uncovering PhoP binding motifs and underscores the importance of dataset diversity and careful hyperparameter tuning in improving model accuracy and generalization.

INDEX

2. Introduction

2.1 PhoP transcription factor role

PhoP is a critical transcription factor that plays a pivotal role in bacterial pathogenicity and adaptation to host environments. As part of the two-component regulatory system PhoP/PhoQ, it responds to various environmental stimuli, particularly those encountered within host organisms. Upon activation, PhoP regulates the expression of numerous genes involved in virulence, stress responses, and metabolic pathways (Groisman, E. A., 2011).

The PhoP protein binds to its DNA targets in vivo only when phosphorylated (Shin and Groisman, 2005; Shin et al., 2006). This phosphorylation is regulated by the PhoQ protein (Groisman and Mouslim, 2006), which activates PhoP under conditions such as low extracellular Mg2+ levels (Garcia Vescovi et al., 1996), the presence of antimicrobial peptides (Bader et al., 2005), or acidic pH (Prost et al., 2007). Once phosphorylated, PhoP (PhoP-P) binds to a hexanucleotide direct repeat, separated by five nucleotides, known as the PhoP box, located in its target promoters (Kato et al., 1999) (Figure 1).

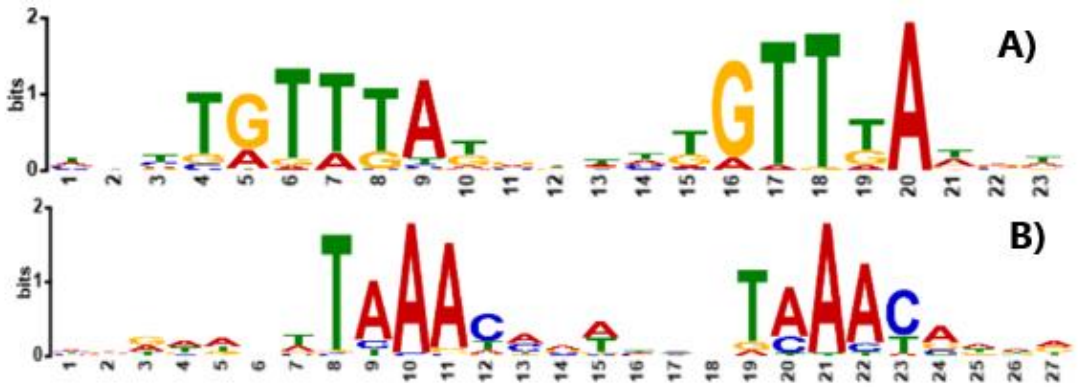


Figure 1. The consensus PhoP binding motif, also known as the PhoP box. A) PhoP box in direct orientation on the forward strand. The consensus PhoP box sequence varies slightly depending on the species. The known sequences for E. coli are KGTTANNNNNKGTTA and for S. enterica KGTTTANNNNNKGTTTA. Here, K refers to T or G, and W refers to A or T. B) PhoP box sequence in reverse orientation on the complementary strand (TAAACANNNNNTAAACA). Direct orientation refers to the half PhoP box 5' (T/G)GTTTA 3' pointing towards the -10 hexamer sequence, while reverse orientation indicates the half PhoP box sequence pointing away from the predicted -10 hexamer sequence. These sequences were extracted using MEME software from the intergenic regions of genes directly regulated by PhoP.

The importance of PhoP in pathogenicity has been demonstrated across various bacterial species. In Mycobacterium tuberculosis, it is essential for virulence (Ryndak et al., 2008), while in Yersinia pestis, it aids in immune evasion and survival in mammalian cells (Vadyvaloo et al., 2015). In avian pathogenic Escherichia coli, PhoP is crucial for biofilm formation (Gao et al., 2015), and in Pseudomonas aeruginosa, it regulates genes responsible for antibiotic resistance (Yang et al., 2021). These studies underscore the versatility and importance of the PhoP transcription factor in regulating bacterial

responses to environmental challenges and its role in the pathogenicity of various bacterial species. Further research is essential to deepen our understanding of the pathogenic mechanisms influenced by PhoP.

2.2 Challenges and Advances in Studying DNA-Protein Interactions

Identifying the binding sites of regulatory proteins on DNA is crucial for understanding microbial processes. Consequently, studying DNA-protein interactions is a key area of research. Traditionally, the study of transcription factors has relied on experimental methods that are often laborious and impractical for large-scale analyses. Techniques such as gel shift assays, DNase footprinting, and ChIP-seq have commonly been used to investigate transcription factor-DNA interactions both in vitro and in vivo (Tognon et al., 2023). However, these methods are limited to controlled laboratory settings and may not be feasible for comprehensive studies (Nakato & Sakata, 2021).

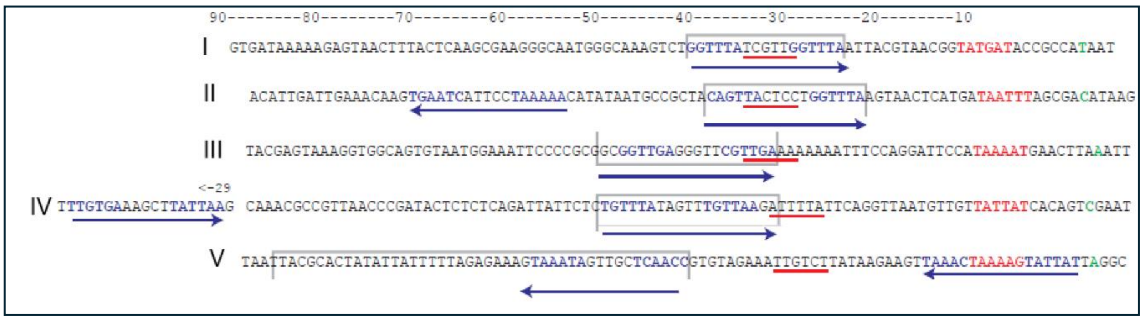


Figure 2. An example of a PhoP-activated promoter DNA sequence for each architecture and its cis-acting promoter features obtained using DNase footprinting by Zwir et al. (2012). The predicted -35 hexamer is underlined in red, and the predicted -10 hexamer is also indicated in red. PhoP binding motifs are enclosed within boxes. The green nucleotide marks the Transcription Start Site (TSS). The arrow indicates the orientation of the PhoP box. I, II, III, IV, and V refer to the different specific promoter architectures identified in Zwir et al. (2012).

Different studies have demonstrated that PhoP can cause varied gene expression depending on its cis-acting promoter features (promoter architectures), which include the number of PhoP binding sites, their orientation, location relative to RNA polymerase binding sites, and the sequence itself (Figure 2). Moreover, it has been shown that certain promoter architectures appear to be species-specific (Zwir et al., 2012). This variability suggests that PhoP employs different mechanisms to initiate gene transcription (Barnard et al., 2004; Shi et al., 2004). Zwir et al. (2012) showed that PhoP employs five distinct promoter architectures to regulate the transcription of its activated genes in Salmonella. These architectures consist of specific, rather than random, combinations of cis-acting regulatory elements (Figure 2). This complexity renders the study of PhoP-mediated actions a challenging task.

While DNA sequencing has become increasingly cost-effective and rapid, there remains a significant challenge in developing a robust system for annotating the pathogenicity-

conferring regulon—comprising promoters and transcription factor binding sites—using genomic information. This gap highlights the need for more efficient and scalable approaches to predict and study transcription factors (Cao et al., 2023).

2.3 Leveraging machine learning and overcoming experimental limitations

To overcome the limitations of traditional experimental techniques, research has increasingly shifted towards machine learning approaches for identifying DNA-binding proteins and predicting DNA-protein binding sites using ChIP-seq data (Elnitski et al., 2006). Recent advances in machine learning, particularly the use of Convolutional Neural Networks (CNNs), have revolutionized this field. CNNs are powerful models designed to recognize patterns in complex data, making them ideal for identifying motifs within promoter sequences. A CNN operates by passing data through multiple layers, each processing the input with activation functions that introduce nonlinearity, enabling the model to capture intricate patterns that traditional linear models might miss (Shujaat et al., 2021).

However, training a CNN model presents challenges such as overfitting and underfitting, which directly affect the model's ability to generalize well to new data. Underfitting occurs when the model is too simplistic and fails to learn the underlying patterns in the data, leading to poor performance. Conversely, overfitting happens when the model becomes too specialized to the training data, performing well on known data but poorly on unseen data. This can result in the model memorizing irrelevant details, leading to errors such as false predictions (Gavrilov et al., 2018).

To achieve optimal performance, it is often necessary to experiment with different parameters, such as learning rate, batch size, and the number of epochs. By carefully tuning the model architecture and employing strategies to prevent overfitting, CNNs can be trained to accurately predict transcription factor binding sites and unravel key biological processes. This advancement in computational methods offers a powerful tool for the large-scale analysis of transcription factors, including those like PhoP that play a crucial role in bacterial pathogenicity (Hu et al., 2019; J. Chen and L. Deng, 2020).

Given the importance of PhoP in bacterial virulence and its limited study across various species, it presents an intriguing target for training a neural network to predict its binding sites. To date, no published studies have comprehensively uncovered the PhoP regulon across bacterial species. The aim of this study is to create a robust dataset of binding sites, which can then be used to train a CNN model to accurately identify these sites in pathogenic bacteria. Data preparation and model training were conducted in Python using packages such as TensorFlow and Pandas.

3. Material and methods

3.1 Dataset generation

Positive and negative datasets were created to train and validate the neural network used for predicting PhoP transcription factor binding sites. Positive dataset was generated from external sources using CollecTF database, since it was the unique that contained available information of a list of known PhoP binding sites for different bacteria, as well as, already published scientific literature. Moreover, positive dataset was also complemented using two different approaches: motif matching approach and protein matching approach. Conversely, negative dataset was generated randomly. All genome and protein files used in this study were downloaded from GenBank. All scripts developed and utilized throughout the project are available in the supplementary data.

3.1.1 Motif matching

For the motif matching analysis, the intergenic regions upstream of genes encompassing *Salmonella enterica* PhoP regulon (Table 1) were extracted, filtered, and saved in FASTA format using the script '**generating_motifs_v8.py'**. All the upcoming scripts can be found in 'scripts' folder in Habrok software. The script was run with default parameters, which include specifying the bacterial genome in FASTA and GFF formats via the ***-genome*** option and providing a CSV file containing a list of regulon gene names under the 'Name' column through the ***-genes*** option. The resulting FASTA file containing these intergenic regions was then analyzed for the presence of the known PhoP motif using the MEME webtool (version 5.5.5), with the settings adjusted to search for up to 10 motifs, zero or one occurrence per sequence (zoops) and minimum length of 15 and maximum of 30 nucleotides.

1	*hemL*	25	*pipB2*	49	*ssaC*	73	*sseA*
2	*hilA*	26	*pipD*	50	*ssaD*	74	*sseB*
3	*lpxO*	27	*pmrD*	51	*ssaE*	75	*sseC*
4	*mgtA*	28	*pmrF*	52	*ssaF*	76	*sseD*
5	*mgtB*	29	*prgH*	53	*ssaG*	77	*sseE*
6	*mgtC*	30	*prgJ*	54	*ssaH*	78	*sseF*
7	*mig-14*	31	*prgL*	55	*ssaI*	79	*ssrA*
8	*nagB*	32	*prop*	56	*ssaJ*	80	*ssrB*
9	*nmpC*	33	*rstA*	57	*ssaK*	81	*steC*
10	*ompX*	34	*sifA*	58	*ssaL*	82	*tuaA*
11	*orgA*	35	*sifB*	59	*ssaM*	83	*udg*
12	*orgB*	36	*slyB*	60	*ssaN*	84	*Ugd*
13	*orgC*	37	*sopD2*	61	*ssaO*	85	*ugtL*
14	*pagC*	38	*spiA*	62	*ssaP*	86	*virK*
15	*pagD*	39	*spiB*	63	*ssaQ*	87	*yaiB*
16	*pagK*	40	*spiC*	64	*ssaR*	88	*ybiF*
17	*pagL*	41	*spiD*	65	*ssaS*	89	*ybjX*

18	*pagP*	42	*spiE*	66	*ssaT*	90	*ybjY*
19	*pcgL*	43	*spiF*	67	*ssaU*	91	*ycfQ*
20	*pdgL*	44	*spiG*	68	*ssaV*	92	*yfbE*
21	*pgtE*	45	*spiR*	69	*ssaW*	93	*yfbG*
22	*phoN*	46	*spvB*	70	*ssaX*	94	*yobG*
23	*phoP*	47	*ssaA*	71	*ssaY*	95	*yrbL*
24	*phoQ*	48	*ssaB*	72	*ssaZ*		

Table 1. The list of potential genes regulated by *Salmonella enterica* PhoP regulon according to databases and literature.

3.1.2 Protein matching

The process of identifying potential PhoP binding site in the 28 pathogenic bacteria (Table 2) using ortholog proteins consists of several steps: Protein sequence extraction of PhoP regulon, ortholog proteins identification, pathogenic bacteria intergenic region extraction, motif discovery and motif validation.

The first step consisted of retrieving the protein sequences of the *S. enterica* PhoP regulon using the script '**regulon_proteins.py'**. The script was run with default parameters: ***-faa*** for the protein sequences in FASTA format, ***-gff*** for the genome annotation file in GFF format, and ***-genes*** for a CSV file with a Name column containing gene names of *S. enterica* that codify proteins being regulated by PhoP regulon.

Then, these protein sequences were then bidirectionally BLASTed against the UniProt database using the shell script **01.DIAMOND_regulon_UniProt.sh**, yielding a list of proteins with greater than 99% match accuracy. The orthologs of these proteins were identified in their respective bacteria using the script '**compareUniProtIDs.py'**. This script was run with default parameters: ***-query*** for a query.g2d.diamond.tab file generated by the FACoP.V2 web tool ((http://facop.molgenrug.nl) (version May 2023) for each bacteria and ***-regulon*** for the RegulonID file containing all UniProt identifiers of the orthologs discovered through BLASTp. As a result, a list of PhoP ortholog genes for each analyzed bacterium was generated.

Bacteria (Gram stain +/-)	**Strain**	**Pathogenicity**
Actinomyces Johnsonii (+)	CCUG3487	Non-pathogenic
Bacillus anthracis (+)	Ames	Anthrax
Bordetella pertussis (-)	Tohama I	Whooping cough
Borreliella burgdorferi (-)	B31	Lyme disease
Campylobacter jejuni (-)	NCTC 11168-Kf1	Gastroenteritis
Chlamydia pneumoniae (-)	TW-183	Respiratory infections, inc. pneumonia
Clostridioides difficile (+)	630	Pseudomembranous colitis
Clostridium botulinum (+)	A str. ATCC 3502	Botulism
Corynebacterium diphtheriae (+)	NCTC 13129	Diphtheria

Enterococcus faecalis (+)	V583	Urinary tract infections, endocarditis
Escherichia coli (-)	K-12 substr. MG1655	Urinary tract infections, gastroenteritis
Haemophilus influenzae (-)	Rd KW20	Urinary tract infections, meningitis
Helicobacter pylori (-)	26695	Peptic ulcers, gastric cancer
Legionella pneumophila (-)	Philadelphia-1	Legionnaires' disease
Listeria monocytogenes (+)	10403S	Listeriosis
Mycobacterium leprae (+)	Kyoto.2	Leprosy
Mycobacterium tuberculosis (+)	H37Rv	Tuberculosis
Neisseria gonorrhoeae (-)	FA1090	Gonorrhea
Niallia circulans (+)	PK3_109	Non-pathogenic
Paenibacillus xylanexedens (+)	PAMC22703	Non-pathogenic
Pseudomonas aeruginosa (-)	PAO1	Causes opportunistic infections
Salmonella enterica (-)	Typhimurium LT2	Salmonellosis (food poisoning)
Shigella sonnei (-)	2015C-3566	Shigellosis (dysentery)
Staphylococcus aureus (+)	NCTC 8325	Skin infections, pneumonia
Streptococcus pneumoniae (+)	TIGR4	Pneumonia, meningitis
Streptococcus pyogenes (+)	MGAS5005	Strep throat, rheumatic fever
Vibrio cholerae (-)	01 Biovar ElTor str. N16961	Cholera
Yersinia pestis (-)	CO92	Plague

Table 2. The list of bacteria from which it was attempted to uncover the PhoP regulon to expand the list of PhoP binding sites. Additional information includes if the bacteria is gram-positive (+) or gram-negative (-), which strain was used and what its pathogenicity is.

The intergenic regions upstream of these genes were then extracted, filtered, and saved in FASTA format using the **'regulon_intergenic.py**' script, with default parameters set for ***-genome***, the bacterial genome in FASTA format and GFF format, and ***-regulon*** which is the output file that contains the orthologs of the PhoP regulon obtained from compareUniProtIDs.py.

To discover potential PhoP motifs within these intergenic regions, the FASTA files were input into MEME (version 5.5.5) with the following settings: number of motifs set to 10 and 'Select the site distribution' as 'any number of repetition (anr)', select the option to check the reverse complement of the input sequences and minimum motif width of 15 nucleotides and maximum of 22. The motifs that according to MEME information sources resembled known PhoP binding sites were then validated using Tomtom, a tool within the MEME-suite web tool (version 5.5.5).

3.2 Input modelling

The motif-containing sequences come from 3 different sources: motif and protein matching, CollecTF and scientific literature. To make the motifs generated by protein matching via MEME webtool (version 5.5.5) compatible with the model, the motifs needed

to be converted into the browser extensible data (BED) format. This was done using the script ***'motifs_to_bed_v6'*** (default parameters: *'-meme'* the motifs downloaded from MEME in FASTA format and *'-inter'* the general features format of the intergenic regions created using '**intergenic_gff.py'**).

The motifs obtained from CollecTF were processed by padding the sequence with random nucleotides added consecutively on each side when the sequence was shorter than 26 nucleotides. For longer sequences, an equal number of nucleotides were removed from each side, ensuring that the core binding motif remained intact and was not affected during this process up to reach the 26 nucleotides. The range of nucleotides that needed to be added or removed on each side never exceeded four.

From scientific literature it was collected the motifs itself plus the surrounding nucleotides that formed the sequence. In case, the sequence was not 26 nucleotides long it was proceeded like previously mentioned for CollecTF sequences (Zhang et al., 2013, Gao et al., 2008, Zhang et al., 2013, Harari et al., 2009, Mcphee et al., 2006, Aguirre et al., 2006, Eguchi et al., 2004, Perez, J. C. & Groisman, E. A., 2009).

The motif-containing sequences from protein matching and non-motif-containing sequences were extracted from the BED files using the '**ProPr_makeCNNmodel_prepareSeqlist.py**' script. The script was run with default parameters, including *-lof* for the text file containing the path to the composite FASTA file of all 28 bacteria and the corresponding BED file. It also used *-FragLen* 26 to specify the length of the random DNA fragments for the non-motif sequence, and *-ratio* 80 to generate non-motif sequences in a 80:1 ratio relative to motif sequences. This resulted in both the training and background sequences being 26 bp long. Then, the positive sequences from CollectTF and literature were manually concatenated to the generated training sequence.

3.3 The CNN model architecture

The CNN model was developed using the script '**ProPr_makeCNNmodel_from_SeqList.py'**, which builds the model by utilizing a directory containing both motif-containing and non-motif-containing sequences. This model consisted of 10 layers. The first layer, an input layer, defined the shape of the dataset, which was based on the dimensions of the image generated through one-hot encoding of the input sequences. The second layer, a Conv1D layer, acted as a pattern recognition mechanism, detecting features by sliding across the encoded sequences. The third and fourth layers are batch normalization layer and a max-pooling layer, respectively, which were used to enhance the features detected by the Conv1D layer.

To prevent overfitting, the fifth was a dropout layer with a rate of 0.2 was incorporated. The sixth layer, a biLSTM layer, analysed the spatial relationships within the sequence, interpreting the information before and after a specific motif. This was followed by the seventh one, a flatten layer to reshape the multi-dimensional output into a flat vector. The eighth layer consisted of another dropout layer with a 0.5 dropout rate for further regularization. The final two layers were dense layers, designed to store the extracted

features and assign probabilities for classification, resulting in a binary output that distinguished between motif and non-motif sequences.

3.4 Model training

The model was trained using the same script employed for its generation (**'ProPr_makeCNNmodel_fromSeqlist.py'**). This script aggregated all motif and non-motif sequences into a comprehensive dataset, which was subsequently partitioned into 90% training data and 10% of the data for the test set to evaluate the model after it is fully trained. The validation set was adjusted to 20% of the training dataset. The dataset was reshuffled at the start of each epoch, where an epoch represents one complete iteration over the entire dataset. Training was terminated after 20 epochs. Additionally, the dataset was subdivided into smaller units, referred to as batches, with a default batch size of 128 sequences utilized during training. A learning rate of 0.000001 was applied during the training process.

3.5 Model evaluation

The model's performance throughout the 20 training epochs is represented via a model-loss plot. For each batch of 128 sequences, the model determines the proportion of incorrect predictions, termed as loss. The average loss is computed for all batches within each epoch, and these values are plotted on the y-axis for both the training and validation datasets, with the x-axis representing the number of epochs. An ideal loss curve exhibits a consistent decline, suggesting effective learning, while deviations from this pattern could indicate overfitting.

At the conclusion of training, a confusion matrix is generated to assess the model's predictions against the true labels, providing essential evaluation metrics:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{(equation 1)}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{(equation 2)}$$

$$\text{F1 score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{(equation 3)}$$

The precision assesses the accuracy of the model's positive predictions (equation 1). The recall measures the model's ability to correctly identify actual positive cases (equation 2). The F1 represents a balanced measure of precision and recall (equation 3). Each of these metrics ranges between 0 and 1, where values closer to 1 signify optimal model performance.

4. Results

This study aimed to characterize the PhoP regulon across multiple bacterial species to establish a robust training set of PhoP transcription factor binding sites. Two complementary approaches, protein matching and motif matching, were employed to expand the list of known PhoP binding motifs sequences publicly accessible via papers or databases. A comprehensive initial regulon was essential for both approaches, and the PhoP regulon of *Salmonella enterica* was chosen as the reference. Orthologous PhoP regulons were first identified through protein matching, followed by the extraction of intergenic regions upstream of the corresponding genes. These regions were subsequently scanned for motifs resembling the consensus PhoP motif of *S. enterica*. The resulting dataset consisting of 325 positive sequences was formatted and used to train a convolutional neural network (CNN) model. The model's performance was evaluated using key metrics, including precision, accuracy, and F1 score.

4.1 PhoP binding motif successfully identified in *Salmonella enterica* using motif matching

A manually curated list of 95 genes (table 1) belonging to the *Salmonella enterica* regulon was compiled using scientific literature and online databases such as CollecTF, RegPrecise and Prodoric. Analysis of the intergenic regions with the motif discovery tool MEME identified 4 different motif sites with a total of 33 binding motif sequences (Figure 3A), which closely matched the consensus PhoP motif for *S. enterica* (Figure 3B).

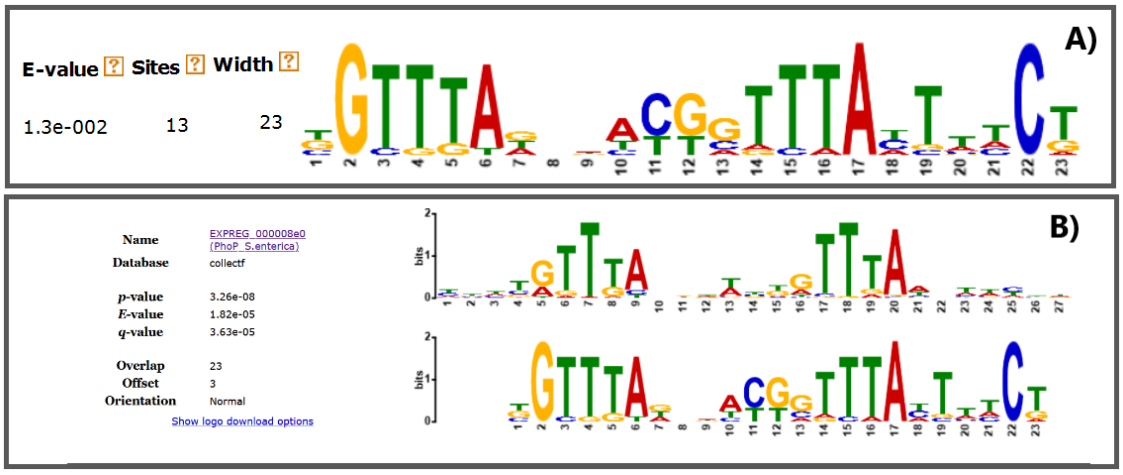


Figure 3. One of the identified motifs in *Salmonella enterica* regulon intergenic regions using MEME and its comparison to the known PhoP binding motif using TOMTOM. (A) The sequence logo of this specific motif identified by MEME, which was detected 13 times in the upstream regions of PhoP-regulated genes. The consensus motif spans 23 bases, with the size of each nucleotide indicating its relative frequency at each position. **(B)** A comparison between the MEME-derived consensus motif and the established PhoP consensus motif for *Salmonella enterica*, obtained from the CollecTF database.

Using the motif comparison tool Tomtom, the previously identified motif (Figure 3A) was queried against all known motif databases. To reduce the likelihood of false positives in the motif alignment analysis, a significance threshold of p-value < 10e-20 was applied. Tomtom identified the PhoP motif from *Salmonella enterica* as the most statistically significant match (Figure 3B), with a p-value of 3.26e-08. Furthermore, the motif also showed significant matches to the CcpA motifs of *Yersinia pestis* (p-value: 2.91e-06) and *Escherichia coli* (p-value: 1.19e-4).

4.2 Orthologous genes identified across 28 bacterial species through protein homology analysis

BLAST analysis of the PhoP protein sequences from the *Salmonella enterica* regulon against the UniProt database yielded 415 orthologous proteins with greater than 99% sequence identity. Cross-referencing these UniProt identifiers with those from the respective bacterial genomes revealed a total of 397 orthologous genes distributed across 28 bacterial species (Table 3).

The protein homology-based approach identified 82 novel PhoP binding sites across 13 out of 28 bacterial species (Table 3), representing a 31% increase relative to the 263 experimentally validated PhoP motifs found among scientific literature and databases. This expansion raised the total number of motif-containing sequences available for model training to 345. In all cases, the protein matching approach resulted in fewer motifs than those reported in CollecTF and obtained from literature.

Bacteria (Gram stain +/-)	By motif matching and MEME	Ortholog genes protein match (>99% match)	By protein matching and MEME	PhoP sites from CollecTF	PhoP sites from literature
Actinomyces Johnsonii (+)	-	9	0	0	0
Bacillus anthracis (+)	-	26	2	0	0
Niallia circulans (+)	-	16	2	0	0
Bordetella perthussis (-)	-	12	5	0	0
Borrelia burgdorferi (-)	-	1	0	0	0
Campylobacter jejuni (-)	-	7	0	0	0
Chlamydia pneumoniae (-)	-	2	0	0	0
Clostridioides difficile (+)	-	38	4	0	0
Clostridium botulinum (+)	-	24	6	0	0
Corynebacterium diphtheriae (+)	-	8	0	0	0
Enterococcus faecalis (+)	-	14	0	0	0
Escherichia coli (-)	-	22	11	43 (2)	17
Haemophilus influenzae (-)	-	3	0	0	0
Helicobacter pylori (-)	-	5	0	0	0
Legionella pneumophila (-)	-	8	0	0	0
Listeria monocytogenes (+)	-	10	0	0	0
Mycobacterium leprae (+)	-	6	0	0	0
Mycobacterium tuberculosis (+)	-	15	5	0	0
Neisseria gonorrhoeae (-)	-	1	0	0	0
Paenibacillus xylanexedens (+)	-	26	4	0	0
Pseudomona aeruginosa (-)	-	27	0	0	13
Salmonella spp. (-)	33	32	6	42	54 (2)
Staphylococcus aureus (+)	-	4	0	0	0
Streptococcus pneumoniae (+)	-	5	0	0	0
Streptococcus pyogenes (+)	-	10	6	0	0
Shigella sonnei (-)	-	22	5	0	0
Vibrio cholerae (-)	-	16	17	0	0
Yersinia pestis (-)	-	28	9	46 (3)	48
Total:	33	397	82	131	132

Table 3. This section summarizes the results of efforts to elucidate the PhoP regulon across multiple bacterial species, with the goal of expanding the dataset of PhoP binding sites. The first column provides the bacterial species and their respective Gram stain classifications. The second column shows the number of sequences obtained by motif matching approach and MEME. – means to non-having been tested. The third column details the number of orthologous genes with >99% sequence identity, identified via DIAMOND BLAST of the *Salmonella enterica* Phop regulon against the UniProt database. The fourth column lists the PhoP binding sites discovered in each species by protein matching approach and MEME. The fifth column shows the number of experimentally validated PhoP binding sites for each species, as documented in the CollecTF database. The last column refers to the number of sequencies for each species obtained from scientific publications. The values between ( ) means the number of sources (subspecies) that contribute to the total number (number of his left).

In data gathered from scientific literature, Pseudomonas aeruginosa was an exception—none of its 13 documented binding sites were identified via motif analysis of intergenic regions or found in databases. In contrast, the other bacteria (Escherichia coli, Salmonella enterica, and Yersinia pestis), which have PhoP sites recorded in both literature and databases, showed the highest concordance between PhoP sites identified through motif analysis and those discovered via protein matching approaches.

Any of all the sequences of the 3 previously mentioned bacteria identified by protein matching was found in the external positive sequences sources, either scientific research or databases. This implies for the fact that this approach accomplishes its task to uncover potential PhoP binding sites to extend the positive training dataset for future predictions.

4.3 Inferred PhoP Binding Motifs and Hexamer Sequences in *S. enterica*

According to Zwir et al. (2012), the natural PhoP-activated promoters utilize a limited number of combinations (promoter architectural landscape) of cis-acting features which include number of PhoP binding sites, orientation and location with respect to the sites bound by RNA polymerase. With the objective to elucidate interesting patterns which provide biological information such as the position of the binding motif related to cis-acting promoter features, it was paid attention in MEME motifs search to the motifs discovered representation to check if interesting patterns were discovered.

As can be seen in the figure 4., MEME represents the position on the intergenic region and strand of the different discovered motifs in *Salmonella enterica* PhoP regulated intergenic regions with the next search settings: 'Any number of repetition (anr)', Minimum length: 15 and Maximum length: 23. The orange-colored sequence, which appears six times, contains a subsequence (TATAAG at positions 8-13) resembling the -10 hexamer consensus sequence (TATAAT). This sequence frequently occurs near the red-colored PhoP binding motif, which appears 13 times. On three occasions, the orange sequence is particularly close to the PhoP motif. This proximity is evident in two of the five observed promoter architectures (Figure 2), demonstrating the recurring pattern of spatial relationship between the motifs.

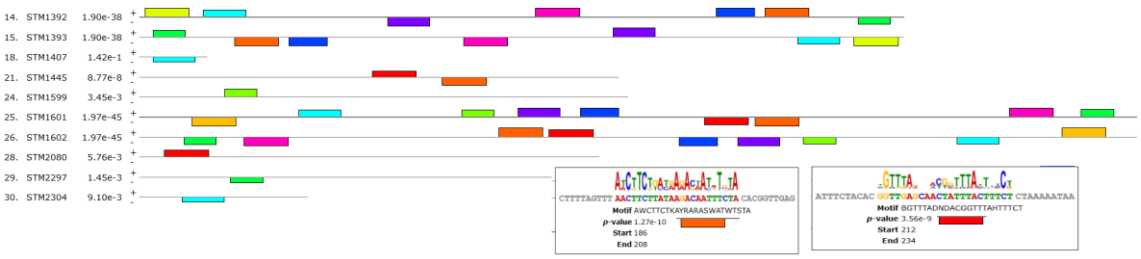


Figure 4. Representation by MEME of different discovered motifs locations. The orange-coloured motif (AWCTTCTKAYRAARASWATWTSTA) contains a similar (TATAAG) -10 hexamer sequence (TATAAT). The red-coloured motif (BGTTTADNDACGGTTTAMTTTCT) contains a statistically significant similar PhoP binding motif sequence (GGTTGA-NNNNN-TATTTA). The single number of the first column represents the number ID of the sequence of the input file. The second column shows the gene identifier. The third column is the combined match p-value. The + or – refers to the forward or reverse strand of the DNA where the motif is located. The rectangles refer to a motif discovered which are differently coloured to differentiate between them. The height of a block gives an indication of the significance of the site as taller blocks are more significant

MEME software was used with the objective to find interesting biological information such as the position of the binding motif related to cis-acting promoter features in the same discovered motif, with the next search settings: Minimum 6 – Maximum 51, and the input the same file that the one used for protein matching the Salmonella enterica, the 42 intergenic region filtered sequences (.fnn). TomTom with an e-value of >-20 was used to align the sequences with different databases and find similarities with the motifs found by MEME. In this case, two different discovered motifs, found 6 and 11 times, respectively, exhibit similarity with PhoP binding motif, as well as presence of -35 or -10 hexamer sequence experimentally validated by Zuir et al. (2012) (Table 4).

Discovered motif	Sequence	Gene	Hexamer type	Hexamer sequence	Position	Promoter architecture
TYTTCYTAATKM TMACRYCATCR WKTA	**TcTTcc**TAATG**atAACA**CCAT**cGaTTa**	STM1393	-35	*CGATT*	21	III, V
			-35	*CCATC*	17	I, II
	TcTTccTAATG**atAACA***CCAT* **cGaTTa**	STM1392	-35	*CGATT*	21	III, V
			-35	*CCATC*	17	I, II
	TTT**TcTTaA**TT**ctcACA**TCAT**caTgTA**	STM1602	NA	NA	NA	II, IV, V
	TTT**TcTTaA**TT**ctcACA**TCAT**caTgTA**	STM1601	NA	NA	NA	II, IV, V
	TcTTcATAGTG**at*AACg*** *T* CACCCTGGA	STM0383	-10	*TAACGT*	12	V
	TTCCCT**TcAcCA***TAACGT* CAT*CGATT* A	STM2782	-10	*CGATT*	21	II, V
			-10	*TAACGT*	12	II, V
KGTWDARVAAC GNTTTAMTDYC KSGMTA	**TGTTTA**GATAC***gGTTT*** **A**CTTTC**TGgTTA**	STM0941	-35	*GGTTT*	11	I, II
	TGTTTAGATAC***gGTTT*** **A**CTTTC**TGgTTA**	STM0940	-35	*GGTTT*	11	I, II
	TGTaacAGAA***cGTTTc*** CATATCGCGCTA	STM0833	-35	*CGTTT*	10	I, II
	TGTaacAGAA***cGTTTc*** CATATCGCGCTA	STM0832	-35	*CGTTT*	10	I, II
	GT**gAAAgc**GAC***gGTTTA*** ATGCCGCGCTA	STM2782	-35	*GGTTT*	11	I, II
	TGTTTAAGCCC***gGTTT*** **A**ATACCGGGCTT	STM2080	-35	*GGTTT*	11	I, II
	gGTTgAGCAAC***TaTTT*** **A**CTTTC**TcTaaA**	STM1602	-35	*TATTT*	11	I, II
	GGTTGAGCAAC***TaTTT*** **A**CTTTC**TcTaaA**	STM1601	-35	*TATTT*	11	I, II
	C**GTATCA**CGACGCGTTGCTGT*C* ***TGgcTa***	STM2781	-35	*CTGGCT*	21	I, II
	GTC**TGagaA**A***cGTTT*** **c**GCTGC**gcgACA**	STM0383	-35	*CGTTT*	10	II, V
	TGTTTAAACAC**gc*TTTATTTc*** CTCCGCC	STM3764	-35	*CTTTA*	12	I, II
			-35	*TATTT*	15	I, II

Table 4. *Salmonella enterica* intergenic discovered motifs by MEME in a search for motifs between 6 and 51 nucleotides length that exhibits a statistically significant alignment with PhoP binding motif by TOMTOM. The first column is the motif pattern followed by all its respective motif sequences (Y = C or T, K = G or T, W = T or A, R = A or G, S = C or G, D = G, T or A, V = G, C or A, M = G, A or C). The second column is the sequence found in a intergenic region. The nucleotides highlighted in black letters are the parts of the sequence that resemble PhoP box (at least 50%). The nucleotides which are in lowercases are because they differ from the consensus PhoP box (Figure 1). The nucleotides underlined and in cursive are

the ones belonging to an experimentally demonstrated by Zwir et al. (2012) -35 or -10 hexamer sequence. It can be possible that one nucleotide is in black letters, underlined and cursive. It means it belongs to a hypothetical hexamer sequence also form part of the PhoP box. The third column is the gene identifier. The fourth column is the hexamer type either -35 or -10. The fifth column is the experimentally validated hexamer sequence that MEME have found in the sequence motif. The sixth column indicates in which position of the sequence the hexamer sequence starts. The seventh column indicates which promoter architecture according to Zwir et al. (2012) this sequence shares more similarities on according to location from hexamer, orientation…

The results showed that in the first motif there was a similar number of -35 and -10 hexamer sequences, while in the second motif there were just -35 hexamer sequences. Overall, in the first motif, the -10 hexamer sequences in STM0383 seemed to overlap with a reverse complemented orientation PhoP box, whereas the -35 hexamer sequences coincide the first nucleotides with a forward orientated PhoP box. Similarly, in the second motif all the -35 hexamer sequences mostly overlapped a forward oriented PhoP box.

In the second motif, just the STM0383 did not overlap neither coincide with what apparently seemed to be a reverse complement oriented PhoP box, similarly like STM2782 in the first motif. Just two sequences of the first motif (STM1602 and STM 1601) did not contain any hexamer sequence. The STM1393 and STM1392 sequences showed a -35 hexamer sequence in the position of the random nucleotides that separate the two repetitions of the binding motif, in that case what seemed to be one forward oriented and one reverse complement oriented PhoP box.

4.4 Developing the model for an enhancement performance

To obtain a CNN model with an excellent performance as an exploratory tool to predict the locations of potential PhoP transcription factor binding site in pathogenic bacteria, the model must be appropriately trained adapting its architecture and parameters to the objective. In an exploratory tool which must predict minimising the error is so important to train the model with the correct training and background data. Through the project it was needed to retrain the model few times changing its training conditions to improve its performance. In the table 5 can be seen the evolution of the model and how the changes in different parameters affected the metrics value to assess it.

			Precision	Recall	F1 score
Model 1	Positive sequences number	169	0.97	0.38	0.55
	Ratio	1:100			
	Epochs	10			
	Batch size	64			
	Learning rate	0.00001			
	Validation set	0.25			
Model 2	Positive sequences number	325	0.74	0.79	0.76
	Ratio	1:80			
	Epochs	10			
	Batch size	64			
	Learning rate	0.00001			
	Validation set	0.25			
Model 3	Positive sequences number	325	0.95	0.85	0.9
	Ratio	1:80			
	Epochs	20			
	Batch size	128			
	Learning rate	0.000001			
	Validation set	0.2			

Table 5. Metrics results after having tested the model in three different conditions along some corrections were made to improve the model performance. The first column is the model identifier. They are in chronological tested order. The second column are the different parameters that were punctually changed to correct the model. Third column shows the values of these parameters. The grey coloured values indicates that for this model they were changed compared to the previous model. Fourth, fifth and sixth columns are the evaluation metrics precision, recall and F1 score, respectively. Positive sequence number is the amount of containing motif sequences that composed the training dataset. Ratio is the rate of containing motif sequence in comparison with non-motif containing sequences. Epochs are the number of times that a training dataset is used. Batch size is number of training samples processed before the model's internal parameters (weights) are updated. Learning rate is a hyperparameter that controls how much the model's weights are adjusted during training in response to the computed error. Validation test is a separate dataset during training to evaluate the model's performance after each training epoch

As it can be noticeable the first run model did not get a positive outcome as it is shown with the F1 score = 0.55. Due to this poor result, model 2 was run increasing the positive sequence number up to 325 and decreasing the ratio to 1:80 to balance the training and background dataset. Even though, in general the model performance was enhanced (F1 score = 0.76), it was to cost of the precision which decreased from 0.97 to 0.74. With no possibility to add more positive sequences and thus, improve the data balancing, there were some model parameters changed. Number of epochs and batch size were doubled, learning rate was one order decreased and validation set was 0.05 decreased. In this context, the model achieved prominent results reaching almost its highest ever performed precision and increasing the recall almost a 50% up to get 0.85.

4.5 Excellent sequence classification on training data

For model training, the 345 binding site motifs were extended to 26 bases. Non-motif-containing sequences were generated at an 80:1 ratio, yielding a total of 41.520 background sequences. These two sets were combined into a single dataset. During each epoch, the dataset was randomly partitioned into an 80% training set (33.492 sequences) and a 20% validation set (8.373 sequences). The model loss plot depicts the training and validation loss across 20 epochs. Initially, the training loss starts high at around 0.07 and sharply declines over the first few epochs, with a constant decrease after 5 epochs. The validation loss, on the other hand, remains consistently low throughout the training process, starting at around 0.01 and fluctuating slightly. After the first epoch, the validation loss move around 0.005, indicating a good fit of the model to the validation data. Notably, the validation loss remains lower than the training loss for the majority of the training process, which suggests the model is not overfitting and is generalizing well to unseen data. However, slight fluctuations in the validation loss after epoch 8 could indicate minor overfitting, but the overall trend shows both losses decreasing, with the model demonstrating effective learning and stable performance.

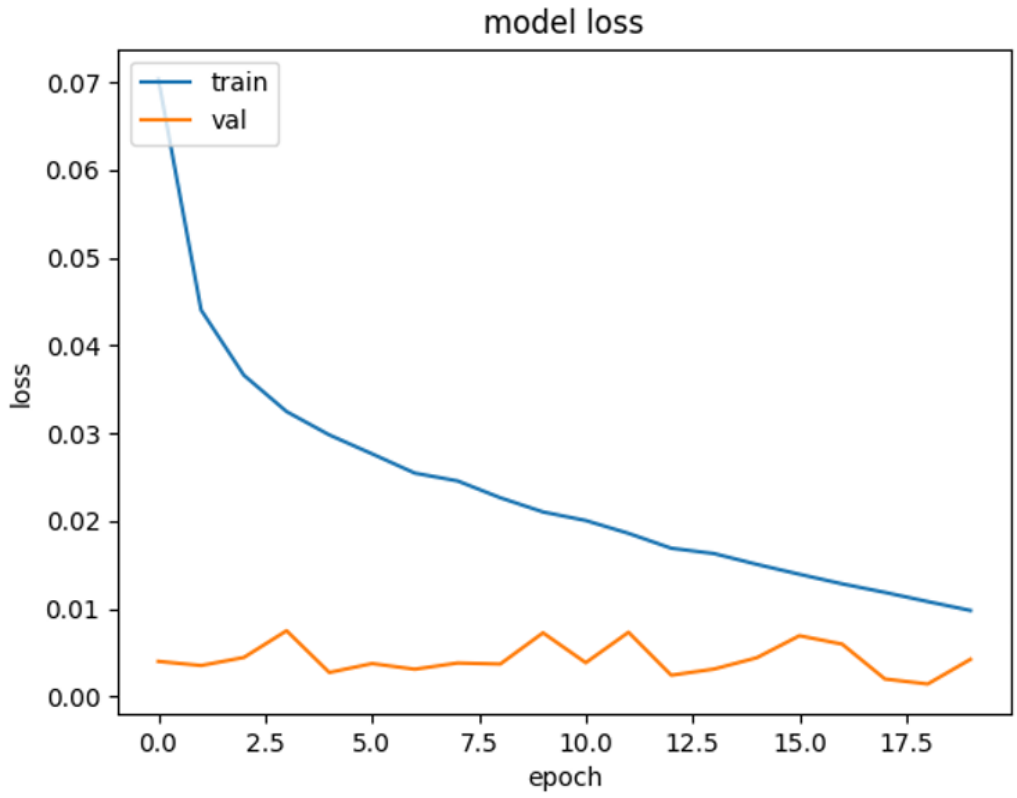


Figure 5. The model loss plot for the training dataset (blue) and validation dataset (orange). The y-axis represents the average loss per data batch, with 'loss' referring to the measure of how accurately the model's predictions align with the true values in each batch. The x-axis displays the number of epochs.

The confusion matrix illustrates the model's ability to distinguish between motif-containing and non-motif-containing sequences effectively. The matrix shows 37,355

true negatives (TN) and 249 true positives (TP), with only 13 false positives (FP) and 43 false negatives (FN). Given the dataset's 80:1 ratio of negative to positive sequences, this high number of true negatives is expected and highlights the CNN's strong capability in recognizing non-motif-containing sequences. Of the 345 motif-containing sequences, the model correctly identified 85%, reflected in the recall score of 0.85. The model's precision was 0.95, indicating that 95% of the sequences predicted as positive were correct, with a low number of false positives (13). The overall F1 score of 0.9 reflects a good balance between precision and recall. These results demonstrate the CNN's robust performance as a sequence classification model, especially in distinguishing between positive and negative sequences in the dataset.

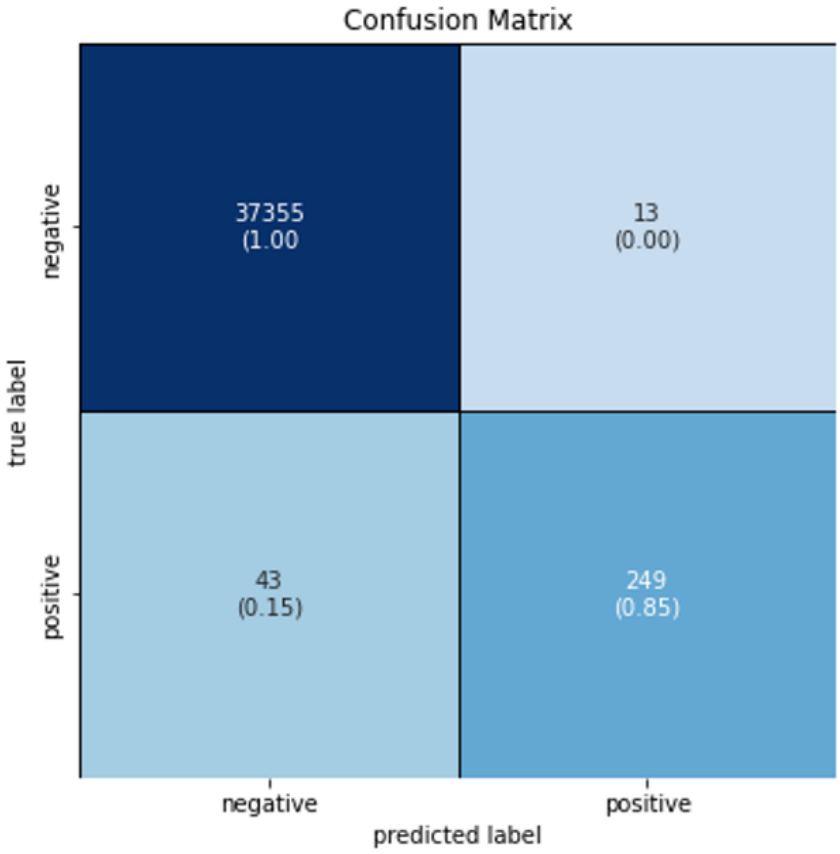


Figure 6. Confusion matrix illustrating the model's performance during the 20th and final epoch of training on a dataset comprising 37.660 sequences. The true labels indicate whether a sequence contains a binding site (positive) or is a randomly generated sequence (negative). The matrix presents the model's predictions as either positive or negative. A false positive refers to a random sequence that the model incorrectly classified as containing a binding motif.

5. Discussion

The goal of this study was to develop a robust dataset and train a convolutional neural network to recognize and predict potential PhoP binding sites in pathogenic bacteria. The dataset was partially constructed using experimentally identified PhoP binding sites from databases and literature reviews and was further supplemented by employing motif and protein matching techniques to identify potential PhoP DNA binding sites across various pathogenic bacteria.

5.1 Comparative Study of PhoP Regulation: Implications for Gram-Negative and Gram-Positive Bacteria

Even though PhoP is considered a well-studied transcription factor, its prevalence in Gram-negative bacteria has led to substantial research being concentrated primarily on a few species, such as *Salmonella enterica*, *Escherichia coli*, and *Yersinia pestis*. In contrast, less extensive research has been conducted on other bacteria, including *Pseudomonas aeruginosa* and *Mycobacterium tuberculosis* (Zwir et al., 2005). As shown in Table 3, sequences from external sources such as databases or published studies were obtained for only 4 out of 28 bacterial species. This observation is further supported by the results from TOMTOM, where motif matches consistently aligned with PhoP sequences from Gram-negative bacteria.

These findings highlight the potential and importance of developing a robust training dataset to improve the model's predictive capabilities for less-studied bacterial species. The limited focus on certain bacteria may be attributed to the more robust PhoP-PhoQ regulatory systems found in specific Gram-negative species. As illustrated in Table 3, the bacteria that have been most extensively studied in relation to PhoP also exhibit the highest number of motifs identified through the protein matching approach. This suggests a correlation between the strength of the regulatory system and the level of research attention received.

While PhoP homologs are present in some Gram-positive bacteria, they are less common and often fulfill different roles compared to their counterparts in Gram-negative bacteria. For instance, in Gram-positive bacteria like *Bacillus subtilis*, PhoP is primarily involved in phosphate regulation rather than in virulence or antimicrobial peptide resistance (Hullet, F. M., 1996). This variation in prevalence and function is largely due to the structural differences in the cell envelopes of Gram-negative and Gram-positive bacteria. Gram-negative bacteria possess an outer membrane containing lipopolysaccharides (LPS), which PhoP helps modify to confer antimicrobial resistance. In contrast, Gram-positive bacteria lack this outer membrane and instead utilize different regulatory systems, such as CodY, to control virulence factors (Stenz et al., 2011).

Additionally, other transcription factors mediating virulence in Gram-positive bacteria, such as CsrS, CovR, or GraR, do not share similar binding motifs with PhoP. This was evidenced by the absence of motif alignments with PhoP binding sites in TOMTOM analyses. Moreover, these regulatory systems differ significantly from PhoP-PhoQ in

their responses to environmental stimuli, underscoring the distinct regulatory mechanisms between Gram-negative and Gram-positive bacteria.

5.2 Exploring PhoP Binding Motifs: Generalization and Performance Across Diverse Bacterial Species

An important part of this research focused on uncovering the PhoP regulon across target bacteria using motif and protein matching techniques. The scarcity of research on PhoP binding motifs in a wide range of bacterial species, combined with the selection of 28 evenly represented Gram-positive and Gram-negative bacteria, limited the number of PhoP sites discovered using the protein matching approach. However, the 115 sequences identified through MEME motif analysis accounted for 30.42% of the complete training dataset, representing a significant portion of the overall data.

In addition to the sites discovered in the most studied species, such as *Salmonella enterica*, *Escherichia coli*, and *Yersinia pestis*, binding sites were identified in 13 out of the 28 species, covering 46% of the total bacterial species analyzed. Despite the higher prevalence of PhoP in Gram-negative bacteria, more potential PhoP binding sites were discovered in Gram-positive species, with 7 Gram-positive versus 6 Gram-negative bacterial species exhibiting these potential binding sites. However, it is noteworthy that two-thirds of the 115 sequences identified were found in Gram-negative species, highlighting a contrast between species diversity and sequence abundance.

This alternation and complementarity led to a training dataset derived from a diverse range of bacterial species. Such diversity enhances the model's ability to generalize, making it better equipped to handle unseen data from new bacterial species. This is evidenced by the low validation loss compared to the training loss, indicating that the model performs well on unseen validation data and supports the conclusion that the model is generalizing effectively. Since the loss curves do not diverge significantly, there is no substantial overfitting, and the diversity in the training data is contributing positively to the model's performance on new data.

The diverse dataset appears to have a beneficial influence on the model's performance with unseen data, as shown by the convergence of the training and validation loss curves. However, the slight fluctuations observed in the validation loss suggest that further fine-tuning may be advantageous, particularly if specific bacterial species exhibit distinct patterns that challenge the model's generalization capabilities. While a diverse training dataset improves generalization, it can also introduce challenges. Different bacterial species may possess unique sequence characteristics, and the model might struggle to fully capture patterns specific to each species if these are too distinct (Gosh et al., 2024). Nonetheless, based on the confusion matrix results (Precision: 0.95, Recall: 0.85, F1 Score: 0.9), the model demonstrates high overall performance, indicating that it has successfully balanced learning from various bacterial species without compromising accuracy.

5.3 Challenges in Validating Newly Identified PhoP Binding Sites Across Diverse Bacterial Species

The limited availability of experimental data and predicted research on PhoP binding sites in many bacterial species makes it challenging to draw direct comparisons between newly identified binding sites and those previously reported in scientific literature. This scarcity of information presents a significant obstacle when attempting to validate or contextualize the newly discovered PhoP binding sites across various bacterial species.

For instance, only 3 out of 28 bacterial species showed sequences in the different databases used. Similarly, a literature review using search terms like “PhoP binding sites” combined with specific bacterial species in databases such as PubMed, PLOS, or Google Scholar revealed a lack of scientific screening—either experimental or in silico—for 60% of the 28 bacteria used to uncover the PhoP regulon. This demonstrates a research focus on studying PhoP binding sites in specific bacteria, while highlighting the absence of targeted research in other organisms where functional studies have been conducted (Yang et al., 2021; Gao et al., 2015). Overall, the protein matching method identified more motifs than those retrieved from CollecTF only in species where no PhoP sites were collected from CollecTF.

For example, regarding sequences obtained from databases, in *Salmonella enterica* (the reference species for protein matching), none of the 33 sequences obtained through motif matching or the 6 obtained through protein matching coincided with the sequences provided by CollecTF. Similarly, in *Yersinia pestis* CO92, the strain used to uncover the PhoP operon, none of the 9 discovered sequences matched those obtained from CollecTF. Ensuring non-redundant sequences in the CNN training dataset significantly contributes to the model's strong generalization capabilities. By avoiding repetitive examples, the model is encouraged to learn diverse patterns, making it less prone to overfitting. This diversity allows the model to perform well on unseen data, as it focuses on recognizing general features rather than memorizing specific instances (Zeng et al., 2016). Consequently, the absence of redundant sequences likely explains the robust performance and reduced overfitting observed in the model's results.

From the 39 total sequences of *Salmonella enterica* obtained through either motif or protein matching, 19 unique genes were identified (*sseL*, *sipC*, *ttrB*, *ssaJ*, *yohH*, *iroB*, *iroC*, *eutQ*, *eutP*, *stdA*, *spiC*, *ybhC*, *fhuA*, *stfC*, *stmA*, *sopD*, *phoP*, *phoQ*, *yohI*). Only 4 out of these 19 genes were found in the scientific publications used to identify positive PhoP binding motif sequences. This implies that the combined approaches of motif and protein matching are effective tools for extending the training database and identifying potential PhoP binding sites in different genomic regulatory sequences. However, the fact that scientific literature has reported PhoP binding sites not discovered by MEME indicates a high false negative rate for the MEME analysis.

5.4 Exploring PhoP Box Architectures in *Salmonella enterica*: Potential Insights from MEME and Promoter Structures

According to Zwir et al. (2012), the PhoP protein utilizes a variety of specific promoter architectures to activate gene transcription in *S. enterica* serovar Typhimurium, which are critical to the regulatory functions carried out by the PhoP protein. All these architectures share at least one PhoP box located upstream of the predicted -10 hexamer. By using the experimentally validated -35 and -10 hexamer sequences from this study as reference points, potential PhoP boxes were identified in the intergenic regions of *Salmonella enterica* genes regulated by PhoP. These discovered PhoP boxes, identified through MEME analysis, appear to align with the previously demonstrated promoter architectures.

Comparing the gene identifiers with their corresponding gene names can help determine whether the genes detected by MEME represent potential new PhoP binding sites in promoter sequences that have not been experimentally validated. This approach is promising, as research suggests that the PhoP protein can interact with RNA polymerase (RNAP) in various ways depending on the specific promoter architecture, thereby regulating gene expression. Additionally, expression microarray analysis comparing wild-type and PhoP mutant strains (Harari et al., 2010) indicates that promoter architecture correlates closely with promoter strength. Therefore, examining the promoter architecture of the model-predicted PhoP binding sites, particularly in *Salmonella enterica*, could reveal significant biological insights.

Further analysis is required to determine whether PhoP binding sites can be functional when they share only 50% similarity with the consensus binding sequence, as observed in most of the 23 experimentally validated sequences reported by Zwir et al. (2012). Alternatively, it is possible that what was previously considered a single consensus sequence actually exhibits greater variability than previously recognized. It may also be the case that PhoP has a high affinity for TA-rich areas, with certain critical nucleotides playing a key role in its binding specificity.

5.5 Optimizing CNN Model Performance for PhoP Binding Site Detection: Dataset Expansion and Hyperparameter Tuning

Initially, the model was trained using a dataset of 168 sequences, composed solely of protein matching and CollecTF sequences. This model achieved a precision of 0.97 but had a recall of just 0.38. The low recall indicated that the model missed many true positives, demonstrating its inability to detect a significant number of positive cases, which resulted in a high false-negative rate. This conservative behavior in making positive predictions suggested that the model was learning too narrowly, focusing only on specific patterns present in the limited dataset. The imbalance in the dataset, with significantly more negative sequences than positive ones, likely contributed to this issue. The model did not have sufficient examples to learn the diversity of patterns associated with positive cases. To address this, the next step was to increase the number of positive

sequences to balance the training dataset and reduce the negative-to-positive ratio to 80:1. This adjustment aimed to enhance the model's ability to learn diverse, generalizable features that would help it detect a broader variety of true positives (Zeng et al., 2016).

The increase in positive sequences, achieved by incorporating experimentally validated sequences from various scientific studies, led to the development of a new model. This updated model achieved a precision of 0.74, a recall of 0.79, and an F1 score of 0.76. The model's ability to capture more true positives indicated improved sensitivity to positive cases, as it learned to recognize a wider range of patterns. Although the precision decreased slightly, this was expected. As the model became more sensitive in detecting positive cases, it occasionally misclassified negative examples as positive, leading to an increase in false positives. Overall, the model became less reliant on a narrow set of patterns, enhancing its ability to accurately identify a broader spectrum of positive sequences.

To further improve the model's performance, hyperparameter tuning was conducted, as adding more positive sequences was not feasible. The following parameters were adjusted: the learning rate was reduced from 0.00001 to 0.000001, the batch size was increased from 64 to 128, the number of epochs was increased from 10 to 20, and the validation set size was reduced from 25% to 20%. These adjustments resulted in a model with a precision of 0.95, a recall of 0.85, and an F1 score of 0.9, demonstrating a significant positive impact on performance.

Lowering the learning rate contributed to the increase in recall by allowing the model to update its weights more gradually, enabling it to capture subtle patterns and achieve better convergence. This adjustment helped prevent the model from overshooting optimal weights during training, as evidenced by the smoother decrease in training loss (Shujaat et al., 2021). Increasing the batch size improved both precision and recall by allowing the model to process more data per training step. This made the learning process more stable and enhanced the model's generalization ability by reducing noise introduced by smaller batch sizes. Extending the training from 10 to 20 epochs provided the model with more opportunities to learn from the data, contributing to overall improvements in precision and recall. The steady decline in training and validation loss indicated that the model did not overfit during the additional epochs. Lastly, reducing the test size from 25% to 20% increased the amount of training data available, which typically improves the model's generalization by exposing it to more examples during training. This change likely contributed to the improved recall by allowing the model to learn from a more comprehensive set of positive examples.

The F1 score of 0.9 indicates that the model achieved an excellent balance between precision and recall, demonstrating its effectiveness at classifying motif-containing and non-motif-containing sequences.

Despite these excellent results, there is still potential to improve the model's performance using advanced strategies. For example, applying class weighting could

adjust the loss function to penalize the model more for misclassifying positive examples, potentially boosting recall without significantly reducing precision. Enhancing feature extraction with attention mechanisms could help capture subtle but important patterns in the data. Optimizing the model architecture by adding more convolutional layers or filters could allow the model to capture more complex patterns, or experimenting with a BiCNN model might offer additional improvements. However, no advanced strategies are currently needed to prevent overfitting, as the model already demonstrates robust generalization capabilities.

5.6 Optimizing Nucleotide Length for Training a CNN Model: Balancing Information and Overfitting Risks

To maintain the model's strong performance, it was essential to determine the optimal nucleotide length for the positive training dataset and ensure a consistent input size for the model. Since the sequences originated from three different sources (CollecTF, protein matching, and scientific literature), there were limitations with CollecTF and literature-derived sequences, as their original nucleotide contexts were unknown. Due to the inability to create a BED file to retrieve the original nucleotides flanking our motifs of interest, the nucleotide length for the training dataset was selected based on a length closer to the shortest sequences (18 nucleotides) rather than the longest (32 nucleotides). This choice was made to preserve focus on the binding motif patterns of interest.

The length of training sequences can significantly influence model performance and behavior. While extending the nucleotide length might provide more contextual information—potentially enabling the detection of features like the -35 and -10 hexamers and the identification of promoter architectures as shown by Zwir et al. (2012)—using longer sequences could introduce noise. Given that our motif is relatively short (17 nucleotides), incorporating longer sequences might include irrelevant information, increasing the risk of overfitting as the model could learn spurious patterns that do not generalize well to unseen data (Shujaat et al., 2021).

Considering the model’s performance and the primary goal of enhancing prediction accuracy rather than identifying specific promoter architectures, a shorter sequence length of 26 nucleotides was chosen. This decision strikes a balance between providing enough contextual information for accurate predictions while minimizing the risk of overfitting.

5.7 Impact of Padding on Sequence Length and Generalization in CNN Models

Padding was applied to adjust sequences that were not 26 nucleotides long, ensuring a fixed input size for the CNN model. However, no constant padding technique, such as using the same base or zero padding, was employed. Introducing arbitrary padding nucleotides that do not exist

in the real sequences can create “noise” in the input, potentially causing the model to learn irrelevant features. This issue becomes particularly pronounced if certain nucleotides are disproportionately represented in the padding. To mitigate this potential noise, random padding was chosen.

This random padding was especially useful in our dataset, since it can also act as a regularization technique by forcing the model to focus on the core sequence and not be overly dependent on the exact positional features and because our sequences varied in length but shared similar motifs in different positions. Since our padding was carried out randomly it might reduce the positional sensitivity of certain motifs, learning to be less dependant on the exact position of the motif and more focused on detecting patterns regardless of where they are in the sequence, but it might slightly reduce the model’s ability to capture all biologically relevant features like real flanking regions if real nucleotides from actual biological sequences were used.

Overall, padding contributed to the model’s strong generalization ability. By encouraging the model to focus on the patterns of interest rather than specific nucleotide positions, padding helped prevent overfitting. This led to improved metrics, such as recall and F1 score, as the model became more robust in detecting patterns in varied positions and contexts.

6. Conclusion

This study demonstrates the effectiveness of Convolutional Neural Networks (CNNs) in predicting PhoP transcription factor binding sites across a broad range of bacterial species. By leveraging machine learning techniques, the research successfully addressed experimental limitations in identifying DNA-protein interactions. The compilation of a comprehensive dataset, including positive and negative sequences from motif matching, protein homology analysis, databases, and scientific literature, was crucial to the model's development.

The CNN model’s performance was significantly enhanced through hyperparameter optimization and the careful handling of dataset imbalances. The incorporation of a diverse training dataset improved the model's precision, recall, and generalization capabilities, while managing the input sequence length helped mitigate overfitting. Additionally, the use of random padding allowed the model to detect motifs in various positions, further contributing to its robustness.

These findings provide valuable insights into the regulatory role of PhoP in pathogenic bacteria and underscore the potential for continued model refinement. Expanding the binding site data and exploring advanced machine learning techniques could further enhance prediction accuracy, offering promising avenues for future research in bacterial gene regulation.

7. Reference

1- Aguirre, A., Cabeza, M. L., Spinelli, S. V., McClelland, M., García Véscovi, E., & Soncini, F. C. (2006). PhoP-induced genes within Salmonella pathogenicity island 1. *Journal of Bacteriology, 188*(19), 6889–6898. https://doi.org/10.1128/JB.00804-06

2- Bader, M.W., Sanowar, S., Daley, M.E., Schneider, A.R., Cho, U., Xu, W., *et al*. (2005) Recognition of antimicrobial peptides by a bacterial sensor kinase. *Cell* **122:** 461–472.

3- Barnard, A., Wolfe, A., and Busby, S. (2004) Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes. *Curr Opin Microbiol* **7:** 102–108.

4- Cao, X., Zhang, L., Islam, M. K., Zhao, M., He, C., Zhang, K., Liu, S., Sha, Q., & Wei, H. (2023). TGPred: efficient methods for predicting target genes of a transcription factor by integrating statistics, machine learning and optimization. NAR Genomics and Bioinformatics, 5(3), lqad083.

5- Eguchi, Y., Okada, T., Minagawa, S., Oshima, T., Mori, H., Yamamoto, K., Ishihama, A., & Utsumi, R. (2004). Signal transduction cascade between EvgA/EvgS and PhoP/PhoQ two-component systems of *Escherichia coli*. *Journal of Bacteriology, 186*(10), 3006–3014. https://doi.org/10.1128/JB.186.10.3006-3014.2004

6- Elnitski L, Jin VX, Farnham PJ, Jones SJ. Locating mammalian transcription factor bind-ing sites: a survey of computational and experimental techniques. Genome Res. 2006 Dec;16(12):1455-64. doi: 10.1101/gr.4140006. Epub 2006 Oct 19. PMID: 17053094.

7- Gao, H., Zhou, D., Li, Y., Han, Y., Song, Y., Zhai, J., & Yang, R. (2008). Identification and characterization of PhoP regulon members in *Yersinia pestis* biovar Microtus. *BMC Genomics, 9*, 143. https://doi.org/10.1186/1471-2164-9-143

8- Gao, Q., Ye, Z., Wang, X., Mu, X., Gao, S., & Liu, X. (2015). RstA is required for the virulence of an avian pathogenic Escherichia coli O2 strain E058. Infection, Genetics and Evolution, 29, 180-188.

9- Garcia Vescovi, E., Soncini, F.C., and Groisman, E.A. (1996) Mg2+ as an extracellular signal: environmental regulation of *Salmonella* virulence. *Cell* **84:** 165–174.

10- Gavrilov, A. D., Jordache, A., Vasdani, M., & Deng, J. (2018). Preventing Model Overfitting and Underfitting in Convolutional Neural Networks. *International Journal of Software Science and Computational Intelligence (IJSSCI)*, 10(4), 19-28. https://doi.org/10.4018/IJSSCI.2018100102

11- Ghosh, N., Santoni, D., Saha, I., & Felici, G. (2024). Predicting Transcription Factor Binding Sites with Deep Learning. *International Journal of Molecular Sciences, 25*(9), 4990. https://doi.org/10.3390/ijms25094990

12- Groisman, E. A. (2001). The pleiotropic two-component regulatory system PhoP-PhoQ. *Journal of Bacteriology, 183*(6), 1835-1842.

13- Groisman, E.A., and Mouslim, C. (2006) Sensing by bacterial regulatory systems in host and non-host environments. *Nat Rev Microbiol* **4:** 705–709.
14- Harari, O., del Val, C., Romero-Zaliz, R., Shin, D., Huang, H., Groisman, E. A., & Zwir, I. (2009). Identifying promoter features of co-regulated genes with similar network motifs. *BMC Bioinformatics, 10*(Suppl 4), S1. https://doi.org/10.1186/1471-2105-10-S4-S1
15- Harari, O., Park, S.Y., Huang, H., Groisman, E.A., and Zwir, I. (2010) Defining the plasticity of transcription factor binding sites by deconstructing DNA consensus sequences: the PhoP-binding sites among gamma/enterobacteria. *PLoS Comput Biol* **6:** e1000862.
16- Hu S, Ma R, Wang H. An improved deep learning method for predicting DNA-binding proteins based on contextual features in amino acid sequences. PLoS One. 2019 Nov 14;14(11):e0225317. doi: 10.1371/journal.pone.0225317. PMID: 31725778; PMCID: PMC6855455.
17- Hulett, F. M. (1996). The Pho regulon and phosphate metabolism in *Bacillus subtilis*. *Molecular Microbiology*, 19(5), 933–939. https://doi.org/10.1046/j.1365-2958.1996.430949.x
18- J. Chen and L. Deng, "DeepARC: An Attention-based Hybrid Model for Predicting Transcription Factor Binding Sites from Positional Embedded DNA Sequence," 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea (South), 2020, pp. 180-185, doi: 10.1109/BIBM49941.2020.9313249.
19- Kato, A., Tanabe, H., and Utsumi, R. (1999) Molecular characterization of the PhoP–PhoQ two-component system in *Escherichia coli* K-12: identification of extracellular Mg2+-responsive promoters. *J Bacteriol* **181:** 5516–5520.
20- McPhee, J. B., Bains, M., Winsor, G., Lewenza, S., Kwasnicka, A., Brazas, M. D., Brinkman, F. S. L., & Hancock, R. E. W. (2006). Contribution of the PhoP-PhoQ and PmrA-PmrB two-component regulatory systems to Mg^{2+}-induced gene regulation in *Pseudomonas aeruginosa*. *Journal of Bacteriology*, 188(11), 3995–4006. https://doi.org/10.1128/JB.00053-06
21- Nakato, R., & Sakata, T. (2021). Methods for ChIP-seq analysis: A practical workflow and advanced applications. In *Methods* (Vol. 187, pp. 44–53). Academic Press Inc. https://doi.org/10.1016/j.ymeth.2020.03.005
22- Perez, J. C., & Groisman, E. A. (2009). Transcription factor function and promoter architecture govern the evolution of bacterial regulons. *Proceedings of the National Academy of Sciences*, 106(11), 4319–4324. https://doi.org/10.1073/pnas.0810343106
23- Prost, L.R., Daley, M.E., Le Sage, V., Bader, M.W., Le Moual, H., Klevit, R.E., and Miller, S.I. (2007) Activation of the bacterial sensor kinase PhoQ by acidic pH. *Mol Cell* **26:** 165–174.
24- Ryndak, M., Wang, S., & Smith, I. (2008). PhoP, a key player in Mycobacterium tuberculosis virulence. Trends in Microbiology, 16(11), 528-534.

25- Shi, Y., Latifi, T., Cromie, M.J., and Groisman, E.A. (2004) Transcriptional control of the antimicrobial peptide resistance *ugtL* gene by the *Salmonella* PhoP and SlyA regulatory proteins. *J Biol Chem* **279:** 38618–38625.
26- Shin, D., and Groisman, E.A. (2005) Signal-dependent binding of the response regulators PhoP and PmrA to their target promoters *in vivo*. *J Biol Chem* **280:** 4089–4094
27- Shin, D., Lee, E.J., Huang, H., and Groisman, E.A. (2006) A positive feedback loop promotes transcription surge that jump-starts Salmonella virulence circuit. Science 314:1607–1609.
28- Shujaat, M., Wahab, A., & Tayara, H. (2021). pcPromoter-CNN: A CNN-Based Prediction and Classification of Promoters. International Journal of Molecular Sciences, 22(1), 296.
29- Stenz, L., Francois, P., Whiteson, K., Wolz, C., Linder, P., & Schrenzel, J. (2011). The CodY pleiotropic repressor controls virulence in gram-positive pathogens. FEMS Immunology & Medical Microbiology, 62(2), 123-139. https://doi.org/10.1111/j.1574-695X.2011.00812.x
30- Tognon, M., Giugno, R., & Pinello, L. (2023). A survey on algorithms to characterize transcription factor binding sites. *Briefings in Bioinformatics*, *24*(3). https://doi.org/10.1093/bib/bbad156
31- Vadyvaloo, V., Viall, A. K., Jarrett, C. O., Hinz, A. K., Sturdevant, D. E., & Hinnebusch, B. J. (2015). Role of the PhoP-PhoQ gene regulatory system in adaptation of Yersinia pestis to environmental stress in the flea digestive tract. Microbiology, 161(6), 1198-1210
32- Yang, B., Liu, C., Pan, X., Fu, W., Fan, Z., Jin, Y., Bai, F., Cheng, Z., & Wu, W. (2021). Identification of Novel PhoP-PhoQ Regulated Genes That Contribute to Polymyxin B Tolerance in Pseudomonas aeruginosa. Microorganisms, 9(2), 344.
33- Zeng, H., Edwards, M. D., Liu, G., & Gifford, D. K. (2016). Convolutional neural network architectures for predicting DNA–protein binding. *Bioinformatics, 32*(12), i121–i127. https://doi.org/10.1093/bioinformatics/btw255
34- Zhang, Y., Wang, L., Fang, N., Qu, S., Tan, Y., Guo, Z., Qiu, J., Zhou, D., & Yang, R. (2013). Reciprocal regulation of pH 6 antigen gene loci by PhoP and RovA in *Yersinia pestis* biovar Microtus. *Future Microbiology, 8*(2), 271–280. https://doi.org/10.2217/FMB.12.146
35- Zhang, Y., Wang, L., Han, Y., Yan, Y., Tan, Y., Zhou, L., Cui, Y., Du, Z., Wang, X., Bi, Y., Yang, H., Song, Y., Zhang, P., Zhou, D., & Yang, R. (2013). Autoregulation of PhoP/PhoQ and positive regulation of the cyclic AMP receptor protein-cyclic AMP complex by PhoP in Yersinia pestis. Journal of Bacteriology, 195(5), 1022-1030. https://doi.org/10.1128/JB.01530-12
36- Zwir, I., Latifi, T., Perez, J. C., Huang, H., & Groisman, E. A. (2012). The promoter architectural landscape of the Salmonella PhoP regulon. *Molecular Microbiology*, 84(3), 463–485. https://doi.org/10.1111/j.1365-2958.2012.08036.x
37- Zwir, I., Shin, D., Kato, A., & Groisman, E. A. (2005). Dissecting the PhoP regulatory network of *Escherichia coli* and *Salmonella enterica*. *Proceedings of*

the National Academy of Sciences of the United States of America, 102(8), 2862–2867. https://doi.org/10.1073/pnas.0409436102