Sequence and Structural Analysis of Six Ancient KRAB-Zinc Finger Protein DNA Binding Profiles using Machine Learning

Daniel Gallegos

College of Natural Sciences
The University of Texas at Austin

1. Introduction

Uncovering the mysteries of how living organisms function and evolve depends in large part on our ability to sequence and analyze genomic data. Advances in next generation sequencing technologies have greatly accelerated these capabilities. Whereas it took the International Human Genome Project (HGP) Consortium and Celera Genomics 13 years and significant funding to sequence the human genome, the entire human genome can now be sequenced for less than US$1000 with comparatively little effort (Gibbs, 2020). Advances in next generation sequencing have made this possible, as these technologies allow for the efficient reading of millions of DNA sequences at once. The size of sequences which can be efficiently read through NGS technologies ranges widely. At one extreme are whole genomes billions of nucleotide base pairs (bp) long; at the other extreme are comparatively tiny 35 bp sequences (Satam et al., 2023). Other next generation sequencing methods have vastly increased our understanding of the interactions between DNA, which codes for proteins and regulatory functions in all organisms, and the proteins which regulate their expression (Satam et al., 2023). Some of these proteins are transcription factors, which bind to specific locations on the genomes of organisms to promote or repress their expression. Transcription factors directly influence the downstream creation of new proteins.

Krüppel-associated box zinc finger proteins (KZFPs) are a class of transcription factors which are evolutionarily ancient and make up a significant portion of the genomes of many organisms. Imbeault et al. (2017) analyzed 15000 KZFP clusters in 191 vertebrate species and showed that they contribute to the evolution of gene regulatory networks across vertebrates. Because of their biological importance and abundance across vertebrate species, KZFPs have been routinely studied in humans, mice, and other vertebrate species.

ChIP-seq, a next generation sequencing method, enables researchers to comprehensively assess all the binding sites of KZFPs to whole animal genomes. Combining ChIP-seq and other next generation sequencing technologies with machine learning methods allows for novel analytic approaches and insight generation. This project applies machine learning methods to ChIP-seq data from six evolutionarily conserved KZFPs. Specifically, logistic regression and convolutional neural network models were used to analyze nucleotide base sequences, composition of k-mers, and physical shape features of the DNA strand at KZFP binding sites. Similarities and differences in binding patterns among the six KZFPs studied that are detectable by these models are highlighted (Figure 1).

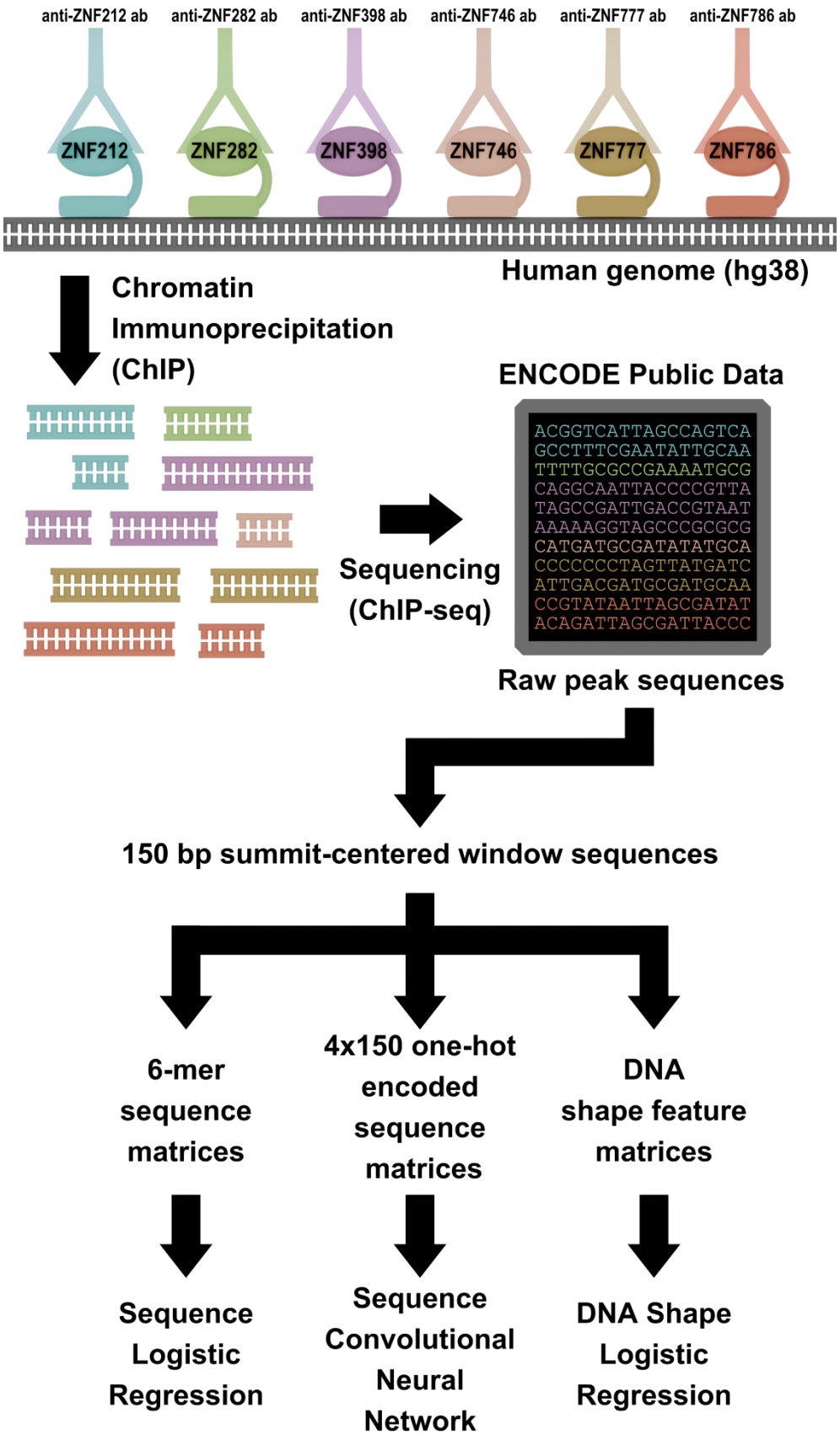
anti-ZNF212 ab
anti-ZNF282 ab
anti-ZNF398 ab
anti-ZNF746 ab
anti-ZNF777 ab
anti-ZNF786 ab
ZNF212
ZNF282
ZNF398
ZNF746
ZNF777
ZNF786
Human genome (hg38)
Chromatin Immunoprecipitation (ChIP)
ENCODE Public Data
Sequencing (ChIP-seq)
ACGGTCATTAGCCAGTCA
GCCTTTCGAATATTGCAA
TTTTGCGCCGAAAATGCG
CAGGCAATTACCCCGTTA
TAGCCGATTGACCGTAAT
AAAAAGGTAGCCCGCGCG
CATGATGCGATATATGCA
CCCCCCCTAGTTATGATC
ATTGACGATGCGATGCAA
CCGTATAATTAGCGATAT
ACAGATTAGCGATTACCC
Raw peak sequences
150 bp summit-centered window sequences
6-mer sequence matrices
4x150 one-hot encoded sequence matrices
DNA shape feature matrices
Sequence Logistic Regression
Sequence Convolutional Neural Network
DNA Shape Logistic Regression

Figure 1. Graphical abstract of project. In Chromatin Immunoprecipitation (ChIP), researchers use a targeted antibody (ab) such as an anti-KZFP ab to pull on transcription factors bound to DNA across the full human genome. To target the binding sites of ZNF212, anti-ZNF212 ab is used. Analogous antibodies are used to immunoprecipitate other KZFPs. Transcription factor binding sites (in this case, binding locations of the six KZFPs of interest) on DNA are pulled and both the transcription factor and DNA fragment are sheared from the rest of the genome. Using Chromatin Immunoprecipitation-sequencing (ChIP-seq), a next generation sequencing technique, millions of DNA fragments are sequenced and outputted as peak density and sequence data representing transcription factor binding site profiles. For the present analysis, raw peak sequence data for six ancestral KZFPs – ZNF212, ZNF282, ZNF398, ZNF746, ZNF777, and ZNF786 – were acquired from ENCODE (The Encode Project Consortium, 2012) and processed into 150 base pair summit-centered windows for machine learning analysis. These were further processed into 6-mers, 4x150 one-hot encoded matrices, and DNA shape feature matrices for sequence logistic regression and convolutional network analysis, as well as shape-only logistic regression analysis.

2. Research Background

The six KZFPs of interest are ZNF212, ZNF282, ZNF398, ZNF746, ZNF777, and ZNF786. These were chosen because Liu et al., 2014 explains that they are part of a deeply evolutionarily conserved cluster. All six are present in the genomes of human, dog, mice, opossum, with some like ZNF777 existing in chicken and lizard genomes (Liu et al., 2014). KZFPs such as these constitute the largest family of transcription factors in mammals and are characterized by rapid evolution (Bruno et al., 2025). Like other KZFPs, these six are silencers of transposable elements, which make up a significant portion of vertebrate genomes and more than half of the full human genome (Senft & Macfarlan, 2021).

Johnson et al. (2007) developed the Chromatin Immunoprecipitation-sequencing (ChIP-seq) assay to comprehensively map DNA-protein interactions across whole genomes. Their assay had sharp resolution at ±50 bp, high sensitivity and specificity and statistical significance of $p < 10^{-4}$ (Johnson et al., 2007). Guidelines for the analysis of ChIP-seq data have been iteratively developed since the development of the assay (Bailey et al., 2013). Fundamentally, ChIP-seq

involves using proteins to pull on antibodies which exclusively bind to DNA-binding proteins of interest, such as KZFPs. These DNA-protein fragments are then processed and sequenced, allowing researchers a global view of protein binding sites across the genome.

Machine learning analysis of next generation sequencing data such as ChIP-seq data has an extensive background in the literature. Alharbi et al. (2022) surveyed numerous deep learning applications in human genomics using NGS data, such as algorithms used to identify genetic variants and generate annotations; tools used to identify disease variants; algorithms used to predict gene expression regulation; as well as epigenomics and pharmacogenomics applications of deep learning. In the case of ChIP-seq specifically, shortly after the assay was developed in 2007, researchers began applying machine learning methods to ChIP-seq data. Gorkin et al. (2012) trained a supervised machine learning algorithm on melanocyte ChIP-seq data which generated a vocabulary of 6-mers with genome-wide predictive power in both mouse and human genomes. Yang et al. (2019) used deep learning to developed DEep Sequence and Shape mOtif or DESSO, which accurately predicted motifs in 690 human ENCODE ChIP-seq datasets.

Since ChIP-seq outputs sequence and protein binding peak information, it is possible to train machine learning algorithms on ChIP-seq data in multiple ways. A classical machine learning model is logistic regression, which when applied to DNA sequences can generate linear sequence motifs. Convolutional neural networks have the added advantage of being capable of generating nonlinear sequence motifs. Additionally, recent advances in DNA shape feature algorithms enable the characterization of sequences by biophysical features such as minor groove width (MGW), helical twist (HelT), propeller twist (ProT), and Roll (Zhou et al., 2013). These allow for the inference of sequence-structure relationships involved in the binding of KZFPs and other transcription factors to DNA (Figure 2).

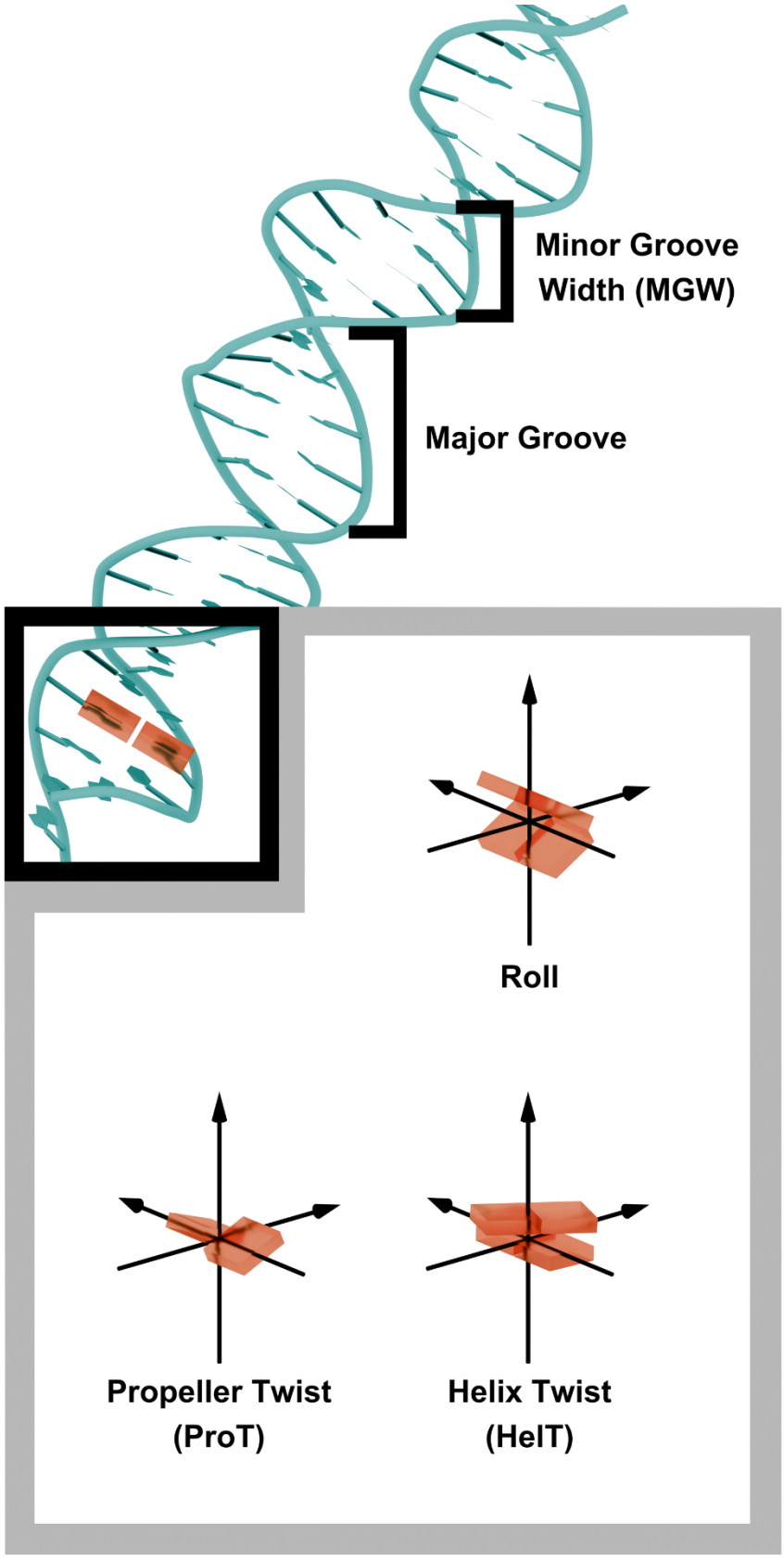
Minor Groove
Width (MGW)
Major Groove
Roll
Propeller Twist
(ProT)
Helix Twist
(HelT)

Figure 2. Biophysical Features of DNA. The four DNA shape features analyzed in the present project are pictured here. Minor Groove Width (MGW) refers to the width of the DNA double helix minor groove. The double helix major groove is also specified for additional context. For Roll, Propeller Twist (ProT), and Helix Twist, each nucleotide base is represented as an individual rectangle for simplicity. This schematization is based on the schema originally outlined by Lavery and Sklenar (1989) and further developed by Zhou et al. (2013). Roll specifies the degree of rotation between two planar sets of base pairs along an axis normal to the helix. ProT specifies the degree of rotation between one nucleotide base and its complement base. HelT refers to the degree of rotation between two planar sets of base pairs along the central helix axis. DNA model retrieved from Protein Data Bank entry 5V3M (Patel & Cheng, 2018) and processed in Jmol (Jmol Development Team, 2025).

This project combines these approaches to investigate whether nucleotide base sequence and DNA shape information can distinguish the sites of binding in ZNF212, ZNF 282, ZNF398, ZNF746, ZNF777, and ZNF786. The three machine learning models utilized are 1) a logistic regression model which analyzes 6-mers of sequence windows; 2) a convolutional neural network model which analyzes one-hot encoded sequence matrices; 3) a DNA shape model using logistic regression to analyze shape features only. Dimensionality reduction was also used to make global assessments of the ChIP-seq data. The overarching goal was to determine whether the KZFPs in this deeply evolutionarily conserved cluster have differentiable binding patterns detectable by machine learning models.

3. Materials and Methods

3.1. Data Acquisition and Processing

Data for the six ancestral KZFPs was acquired from the publicly accessible ENCODE database (The ENCODE Project Consortium, 2012; Kagda et al., 2025). The ENCODE accession numbers for all KZFP ChIP-seq datasets are included in Table 1.

Table 1. ENCODE Accession Numbers

KZFP	ENCODE Accession Number
ZNF212	ENCSR002ZGC
ZNF282	ENCSR752NDX
ZNF398	ENCSR676ZEF
ZNF746	ENCSR591MYB
ZNF777	ENCSR295BIP
ZNF786	ENCSR206BVQ

A comprehensive multi-species analysis was not possible due to the relative lack of publicly available ChIP datasets for these KZFPs. Human datasets were used exclusively.

Bed narrowPeak files were downloaded and 150 base pair (bp) windows were extracted centering on each peak. This was done by finding the genomic location of the human genome using hg38 from the UCSC Genome Browser (Perez et al., 2025). There were both biological and computational reasons for this. Biologically, the actual site of contact in KZFPs is an array of C2H2 zinc fingers which range from 3-40 fingers, with the average size being 12 fingers (Yang et al., 2017). Each finger makes direct contact with three nucleotide bases (Yang et al., 2017). This suggests that a reasonable range of contact site lengths is between (3 fingers x 3 bp) to (40 fingers x 3 bp), or 9 bp to 120 bp. A sequence window of 150 bp adequately encompasses this range.

Additionally, it is standard practice to feed bins of 100-200 bp into algorithms for machine learning analysis. Alipanahi et al. (2015) used 101 bp sequences centered at the point source for each peak to train DeepBind, a deep learning tool used to predict sequence specificities in DNA- and RNA-binding proteins. Zhou and Troyanskaya (2015) used 200 bp sequence bins to prepare whole genome data for convolutional network model analysis in order to create DeepSea. Fundamentally, machine learning models often require uniform input lengths. The machine learning models used here, namely logistic regression and convolutional neural networks, each require some form of uniform, fixed input.

Pyfaidx (Shirley et al., 2015), which draws inspiration from the widely used Samtools indexing command faidx (Li et al., 2009; Danecek et al., 2021), was used to generate 150 bp window sequences. Each individual KZFP dataset was concatenated into a final dataset. No sequences had invalid start or end locations, and zero sequences were filtered out from the final dataset. In total there were 62,617 sequences (Figure 3). GC Fraction was also calculated for each window sequence per KZFP (Figure 4), as was CpG fraction. This gives insight into possible methylation sites, since most if not all CpG islands are sites of transcription initiation and subject to repression by methylation or polycomb recruitment (Deaton & Bird, 2011). Table 2 summarizes the processed dataset.

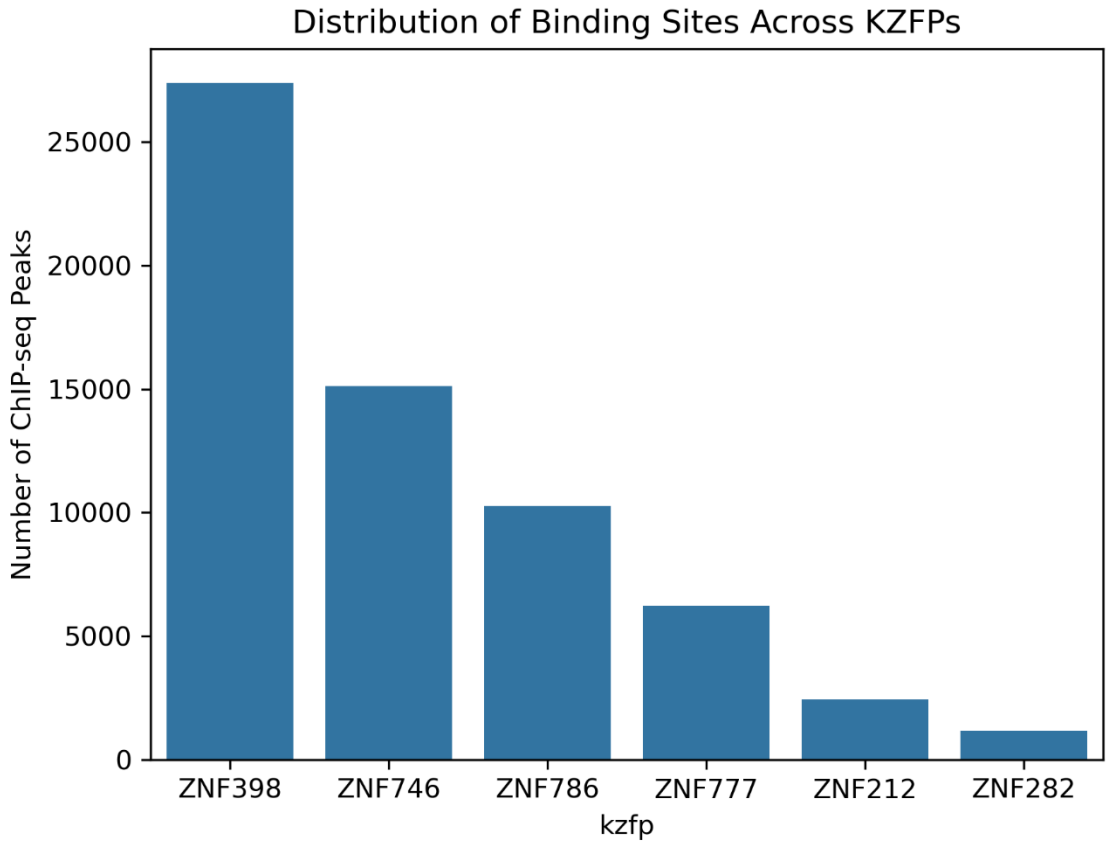

Figure 3. Number of ChIP-seq peaks for each of the six ancestral KZFPs of interest

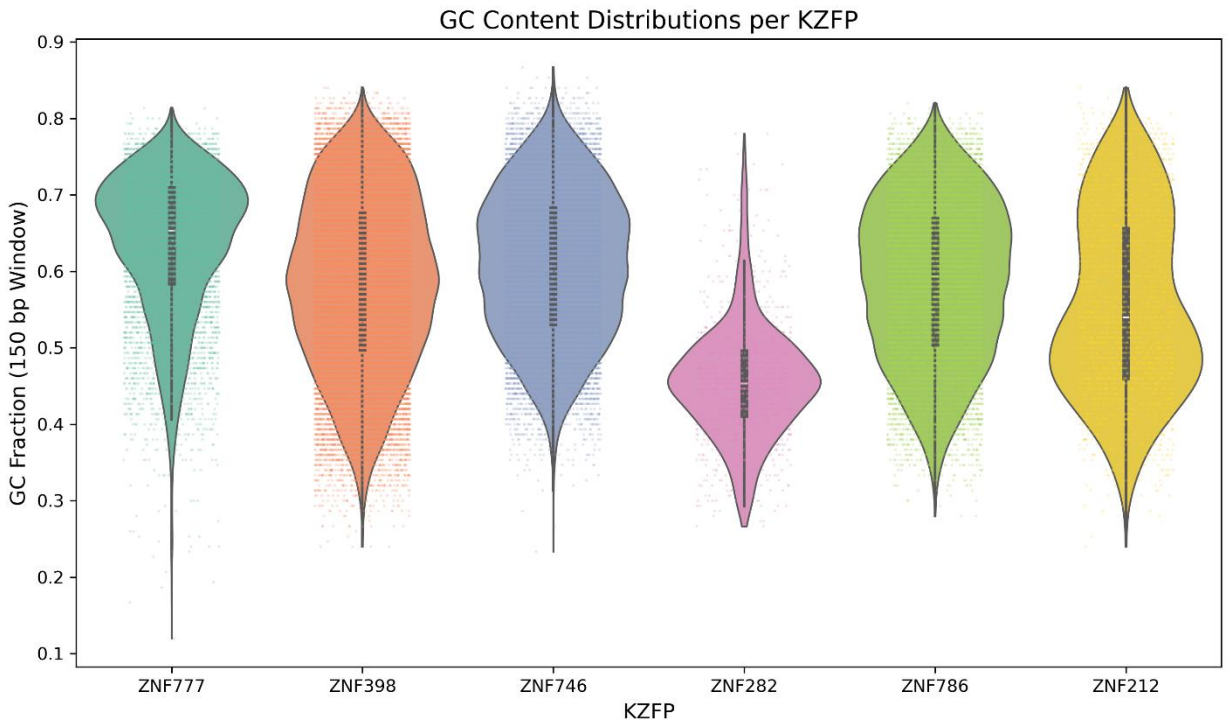

Figure 4. Fraction of guanine (G) and cytosine (C) bases across 150 bp windows per KZFP

Table 2. Dataset Summary

Field	Description	Example Record
chrom	Chromosome number on human genome (hg38)	chr12
start	Start position of given ChIP-seq signal on chromosome	12266969
end	End position of given ChIP-seq signal on chromosome	12267360
signalValue	Average read count compared to background sample; higher values indicate stronger binding	507.8467
qValue	Positive false discovery rate (Storey, 2003)	2.6561
peak	Location of signal summit relative to start position (summit = start + peak)	194
win_start	Start of 150 bp window (summit – 75)	12267088

Table 2. Dataset Summary

Field	Description	Example Record
win_end	End of 150 bp window (summit + 75)	12267238
windowSeq	Full 150 bp window sequence	GCAACGAGCCCCTTCTCCCGGTACTGCCTCCTGT ACGGCCAGGGAAGGGAGTCGGCAACCGCACGC ACAGCCTCTGCCTGAGACCCTGGGAGAGGTTCT GGGCTAGCAGAGGCGAACTGGGAGGGAAAGCC TCCTCCCGGTGGCGCATTC
gc	Proportion of G and C bases in windowSeq	0.66
cpg_density	Proportion of CpG islands in windowSeq	0.06
kzfp	Krüppel-associated box zinc finger protein	ZNF777

The 150 bp sequence windows were converted into k-mer vectors for logistic regression analysis. Each sequence window was converted to a matrix of 6-mers. 6-mers were chosen because they have been found to be more informative for machine learning analysis of ChIP-seq data than k-mers of other lengths while giving a balance of interpretability and rigor (Lee et al., 2011; Gorkin et al., 2012). For convolutional neural network (CNN) analysis, each window sequence was one-hot encoded into a 4x150 matrix, with one distinct value corresponding to each of the four nucleotide bases (A, C, G, T). In one-hot encoding of DNA sequences, each nucleotide base is represented as a binary vector with 4 elements: three 0 elements, and one identifier element, 1 (Gupta et al., 2025).

NumPy (Harris et al., 2020) and pandas (McKinney, 2010) were used for data processing. Matplotlib (Hunter, 2007) and Seaborn (Waskom, 2021) were used for data visualization. Figures utilizing 3D graphics were custom made in Blender (Blender Development Team, 2025).

3.2. Sequence Logistic Regression Model

The matrix of 6-mers for each genomic window sequence, GC content, and CpG density were combined as the features and KZFP names - ZNF212, ZNF282, ZNF398, ZNF746, ZNF777, ZNF786 - were used as labels. Using the LogisticRegression class from Scikit-learn (Pedregosa et al., 2011), the logistic regression model was trained on 80% of the dataset with 20% left for testing.

3.3. Sequence Convolutional Neural Network Model

All 62,617 window sequences were one-hot encoded as 4x150 matrices. This set of matrices was used as the feature input of the model and KZFP names were used as labels. A convolutional neural network was constructed using Tensorflow (Abadi et al., 2016) which consisted of a 1D convolutional layer with 128 filters, a kernel size of 12, and ReLU activation. Next, a max-pooling layer and a second 1D convolutional layer (64 filters, kernel size of 8, ReLU activation) followed. Additionally, a global max-pooling layer was used to convert the feature maps into fixed lengths vectors representing the maximum activation across the 150 bp sequence for each filter. These vectors were then passed through a Dense 64-unit ReLU layer with 0.3 dropout. The output layer was a 6-way softmax classifier, representing the 6 KZFP genes under analysis. The Adam optimizer was used with a $1e^{-3}$ learning rate. Categorial cross-entropy loss and accuracy were used for evaluation. Fifteen epochs were used to train the model, each with batches of 64 sequences. Ten percent of the training data was set aside for internal validation. Training history was used to confirm stable convergence across epochs.

3.4. DNA Shape Logistic Regression Model

DNAShapeR (Zhou et al., 2013; Chiu et al., 2016) was used in R to compute characteristic biophysical features of DNA, including minor groove width (MGW), propeller twist (ProT), helical twist (HelT), and Roll. Shape features were exported from R as matrices with 586 total features per sequence (147 MGW, 147 ProT, 146 Roll, 146 HelT features). Column renaming and further analysis was carried out in Python. This feature matrix was used as the feature input

of a Scikit-learn logistic regression classification model. KZFP names were used as labels for the model, which was trained on 80% of the input data.

3.5. Clustering Analysis

The 150 bp sequences were converted into 4-mers instead of 6-mers for clustering analyses as nucleotide 4-mers have been shown to perform optimally well with clustering algorithms (Yang et al., 2010). These 4-mer embeddings were used as input to a Scikit-learn principal component analysis (PCA) algorithm with two components to visualize global clustering structure. UMAP, short for Uniform Manifold Approximation and Projection for Dimension Reduction (McInnes et al., 2020), was also attempted with a reducer model of 30 neighbors, a minimum distance of 0.1, and Euclidean distance computing.

4. Results

4.1. Sequence Logistic Regression Baseline

The 6-mer sequence logistic regression model performed with an overall accuracy of 0.46, macro F1 of 0.15, and weighted F1 of 0.35. ZNF398 had the highest recall at 0.94. The next highest recall was 0.21 for ZNF746. All remaining KZFPs had a recall at or below 0.01 (Table 3).

Table 3. Classification Report for Sequence Logistic Regression

KZFP	Precision	Recall	F1-Score	Support
ZNF212	0.00	0.00	0.00	488
ZNF282	0.40	0.01	0.02	231
ZNF398	0.49	0.94	0.64	5482
ZNF746	0.33	0.21	0.26	3027
ZNF777	0.00	0.00	0.00	1244
ZNF786	0.33	0.01	0.01	2052
Accuracy	-	-	0.46	12524
Macro Avg	0.26	0.19	0.15	12524
Weighted Avg	0.35	0.46	0.35	12524

For all except 78 predictions, the logistic regression model predicted that a given record of 6-mer sequences, GC content, and CpG density belonged to either ZNF398 or ZNF746. ZNF398 arose as the most predicted KZFP, leaving all minor KZFPs misclassified (Figure 5).

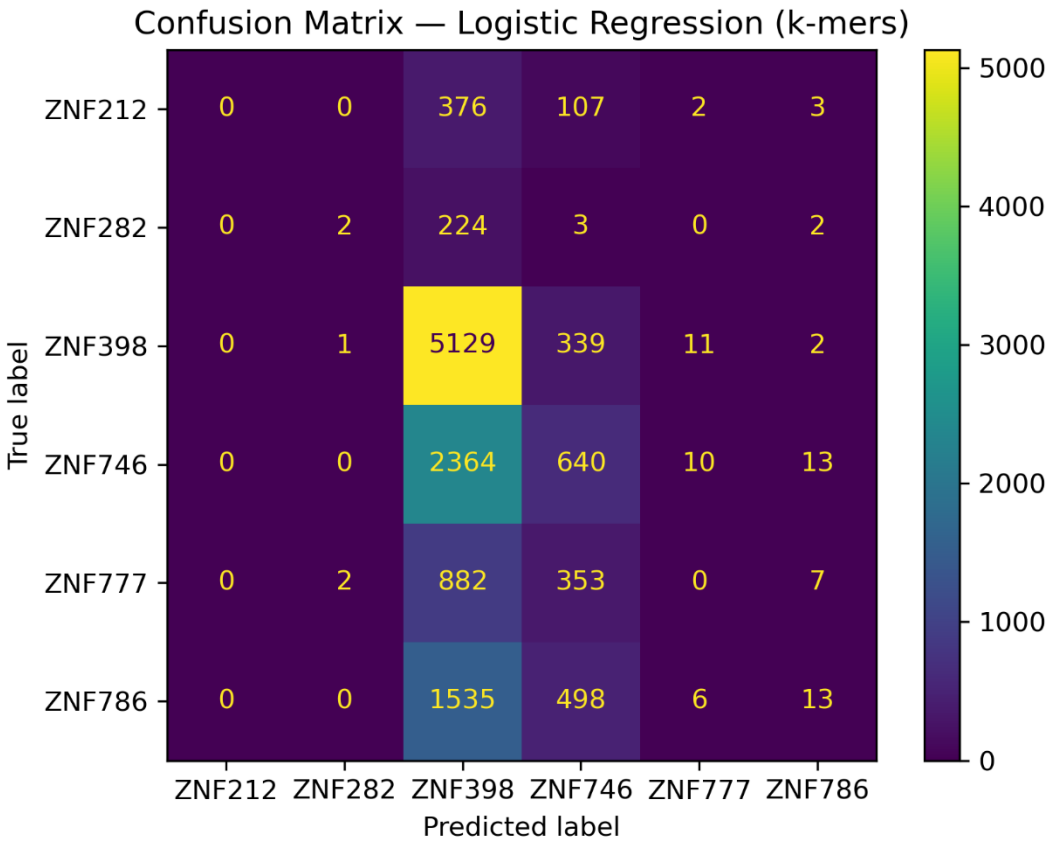

Figure 5. Confusion Matrix for Sequence Logistic Regression

4.2. Sequence Convolutional Neural Network Results

The convolutional neural network model achieved higher classification report values than the logistic regression model. It had an accuracy of 0.47, macro F1 of 0.35, and weighted F1 of 0.46 (Table 4).

Table 4. Classification Report for Sequence Convolutional Neural Network

KZFP	Precision	Recall	F1-Score	Support
ZNF212	0.45	0.09	0.15	488
ZNF282	0.57	0.42	0.48	231
ZNF398	0.64	0.68	0.66	5482
ZNF746	0.36	0.38	0.37	3027
ZNF777	0.21	0.25	0.23	1244
ZNF786	0.29	0.26	0.27	2052
Accuracy	-	-	0.47	12524
Macro Avg	0.42	0.35	0.36	12524
Weighted Avg	0.47	0.47	0.46	12524

Per-class improvements were especially notable for ZNF777 and ZNF786, which each received an order of magnitude more predictions than in the logistic regression model. Although ZNF212 and ZNF282 continued to receive few predictions, they received more than the 0-2 predictions from the logistic regression model (Figure 6).

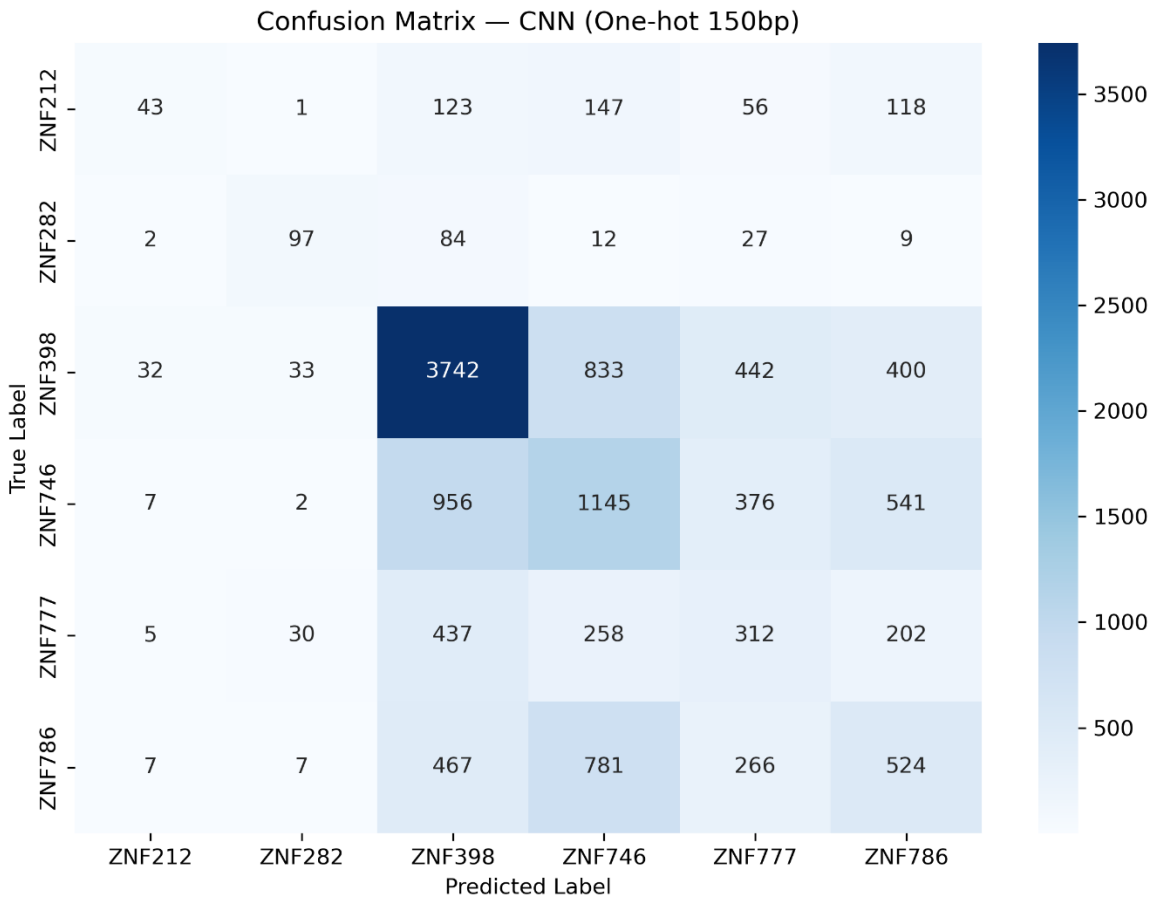

Figure 6. Confusion Matrix for Sequence Convolutional Neural Network

4.3. DNA Shape Logistic Regression Model Performance

DNA shape-only logistic regression performed similarly to sequence logistic regression. The shape-only model had an accuracy of 0.44, macro F1 of 0.20, and weighted F1 of 0.36. ZNF398 remained the dominant target of predictions while ZNF746 and ZNF282 received 0.25 and 0.16 recall respectively (Table 5).

Table 5. Classification Report for DNA Shape Logistic Regression

KZFP	Precision	Recall	F1-Score	Support
ZNF212	0.00	0.00	0.00	488
ZNF282	0.33	0.16	0.22	231
ZNF398	0.49	0.84	0.62	5482
ZNF746	0.32	0.25	0.28	3027
ZNF777	0.20	0.02	0.04	1244
ZNF786	0.21	0.04	0.07	2052
Accuracy	-	-	0.44	12524
Macro Avg	0.26	0.22	0.20	12524
Weighted Avg	0.35	0.44	0.36	12524

The six KZFPs exhibited several insights on the shape of the DNA fragments which they bind to. Based on the top ten shape features for each KZFP, ZNF212 binding sites are dominated by MGW in mid-body positions. ZNF777 binding sites have a strong dependence on Roll features. ZNF282 and ZNF786 exhibited a strong mix of features, with the former exhibiting a mix of MGW and Roll, and the latter exhibiting a mix of HelT and MGW. ZNF398 and ZNF746 had distributed MGW signals across many positions (Figure 7).

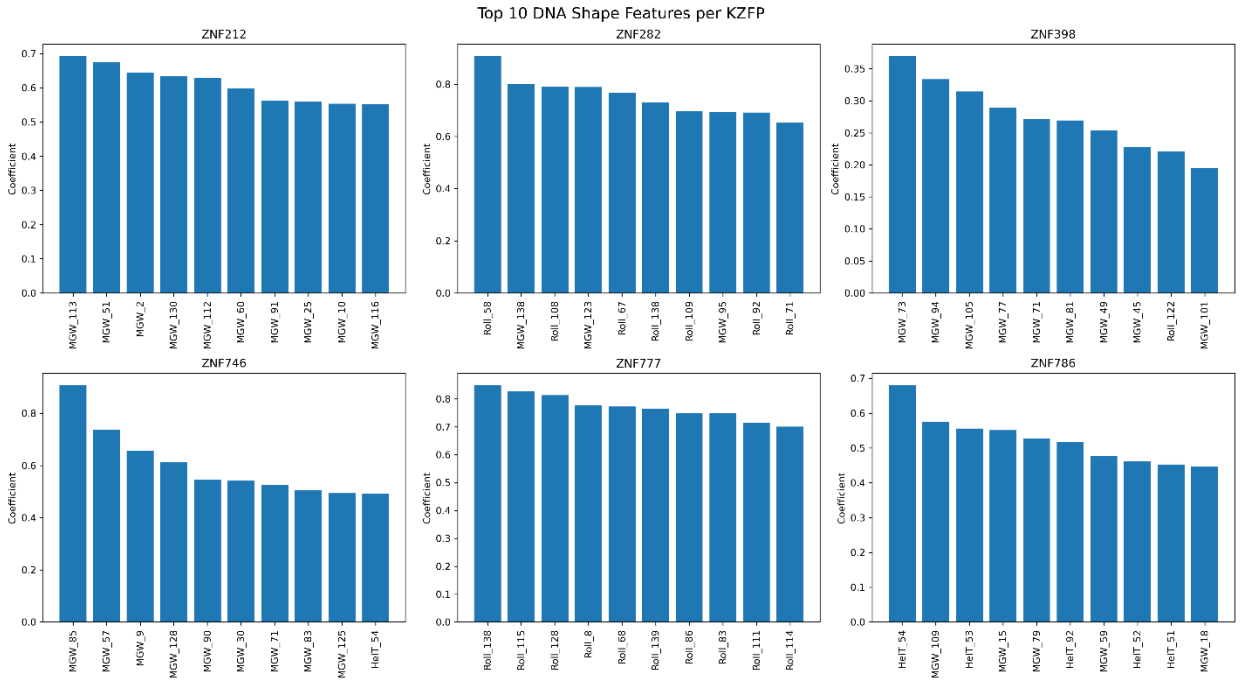

Figure 7. Top Ten DNA Shape Features per KZFP

Comparing across KZFPs reveals other patterns. ZNF282 arises as having the highest average MGW, Roll, and HelT profile and the lowest average ProT profile across the 150 bp sequence window. ZNF777 has the highest average ProT profile. The other KZFPs cluster closer together making for less distinguishing characteristics (Figure 8).

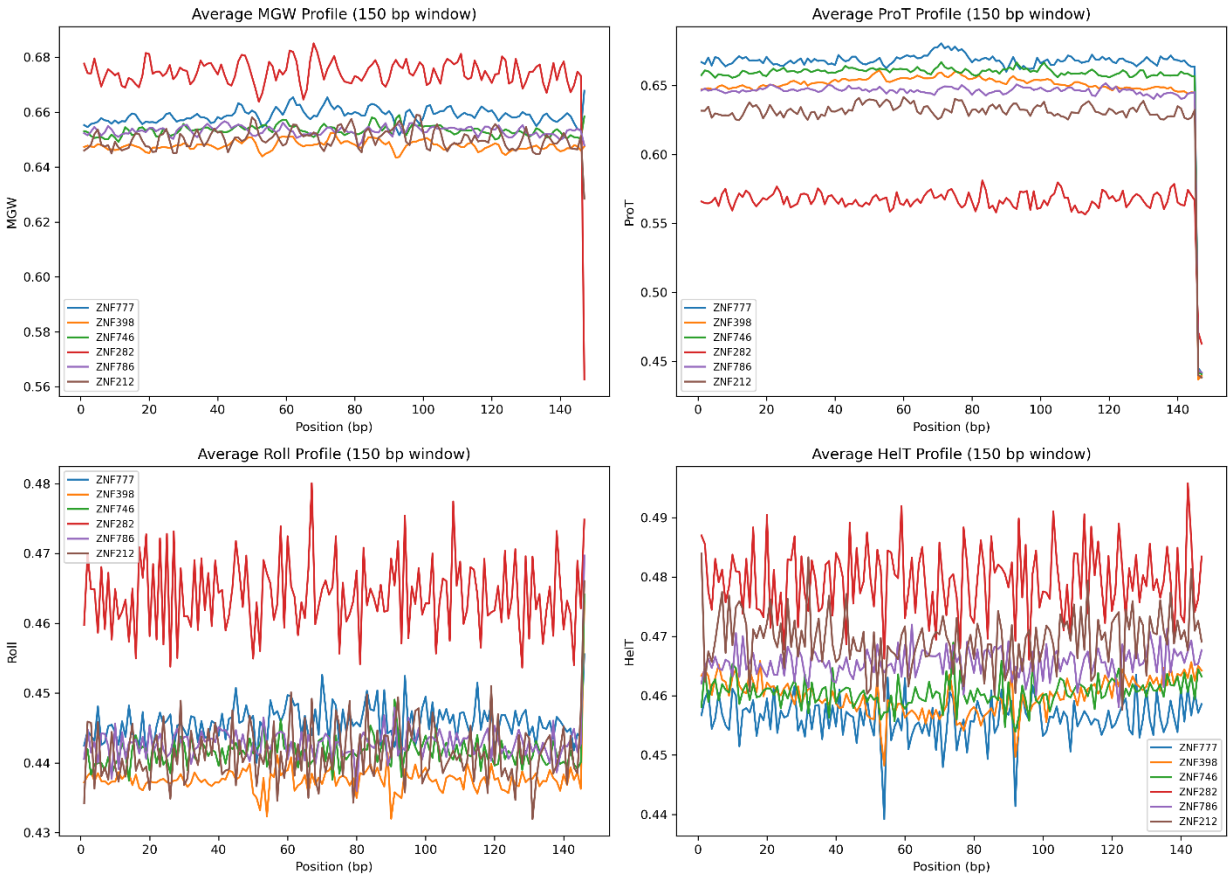

Figure 8. Average DNA Shape Profiles of each KZFP

Analysis of the cosine distance between each KZFP per DNAShape feature reveals that ZNF777 and ZNF282 are most dissimilar in terms of MGW (0.00013) and ProT (0.00022). ZNF282 and ZNF212 are most dissimilar regarding both Roll (0.0001) and HelT (0.00012; Figure 9).

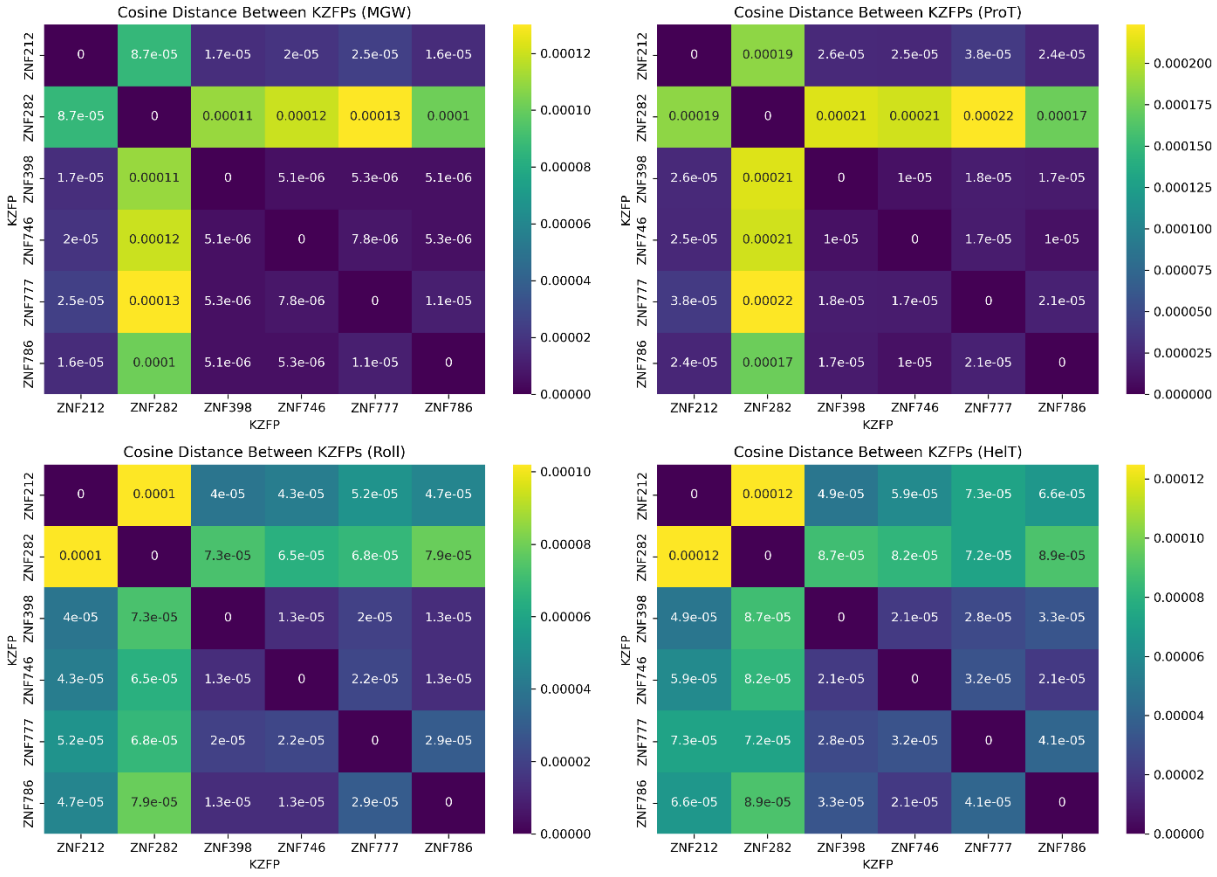

Figure 9. Cosine Distance Between KZFPs per DNA Shape Feature

4.4. Clustering on K-mers

Principal component analysis clustering revealed a partial separation of ZNF282, ZNF398, and ZNF777 along principal components. Most other KZFPs overlapped substantially (Figure 10).

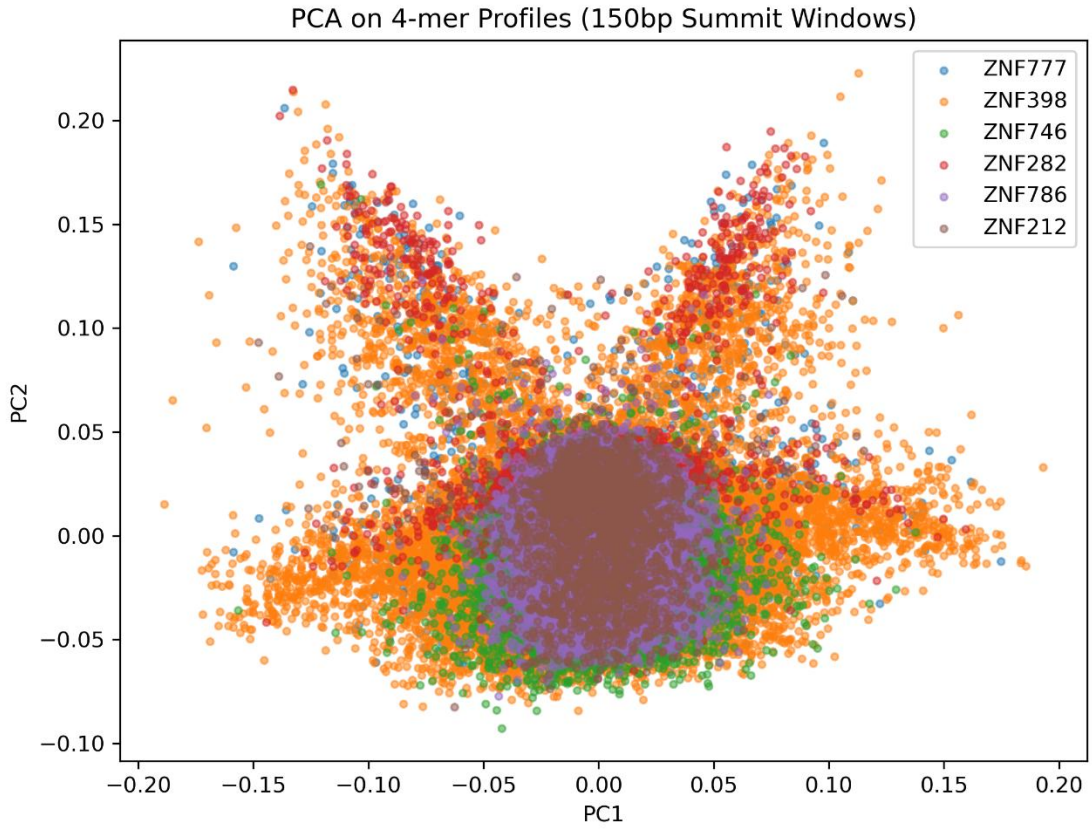

Figure 10. Principal Component Analysis (PCA) Clustering on 4-mer KZFP Profiles

UMAP on k-mers produced noisy embeddings without stable clusters (Figure 11).

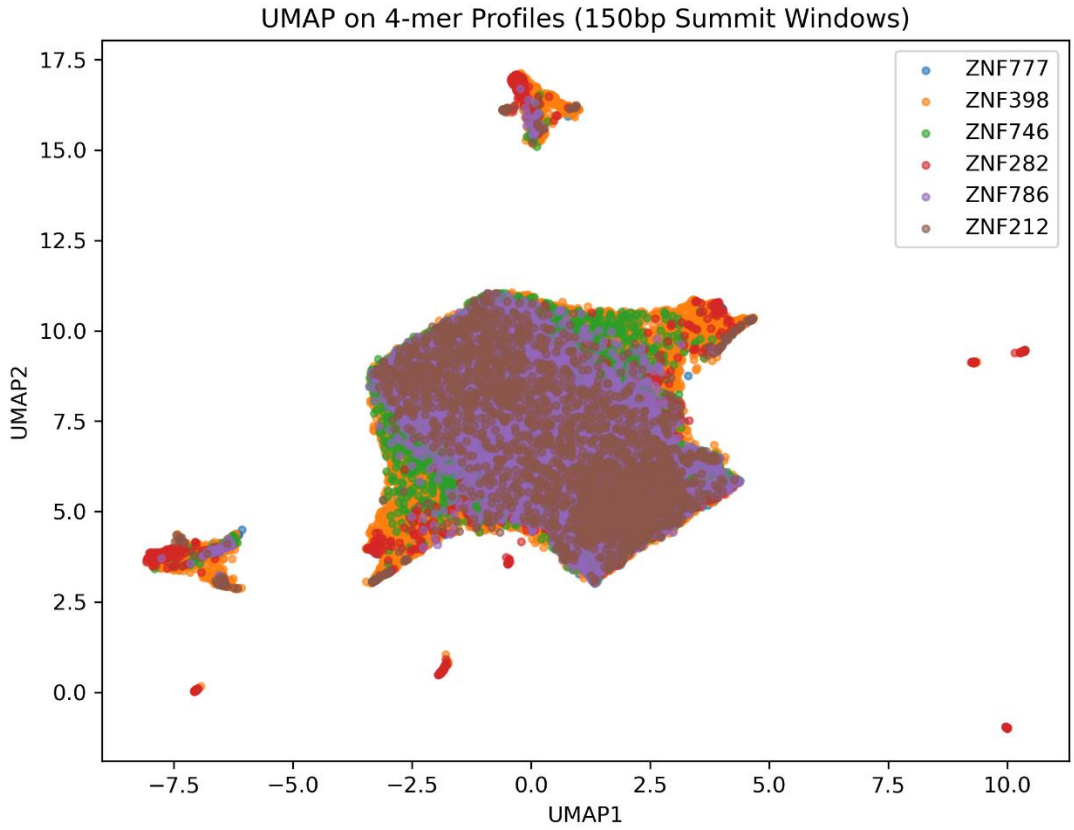

Figure 11. Uniform Manifold Approximation and Projection (UMAP) Clustering on 4-mer KZFP Profiles

5. Discussion

This project aimed to investigate whether DNA sequence and structure patterns of the binding sites of a cluster of six ancestral KZFPs can be detected and differentiated by machine learning models. Using logistic regression, a convolutional neural network, and clustering algorithms, this project has shown that some sequence and structural patterns can be differentiated within this dataset while others cannot. Four insights emerged from this project.

5.1. CNNs Outperform Logistic Regression on Imbalanced ChIP-seq Data

It is well understood that CNNs are being increasingly applied to imbalanced data in many contexts as they are especially good at dealing with such data (Dablain et al., 2024). The dataset

used in this project was imbalanced, with most records associated with ZNF398 and substantially less associated with ZNF212 and ZNF282. That logistic regression had little success at differentiating the binding sites of all KZFPs except for ZNF398 is entirely consistent with the abundance of records associated with that KZFP. However, the CNN model was more capable of differentiating the remaining KZFPs than the logistic regression model. Nevertheless, the overall model accuracy plateaued at 47%, suggesting substantial similarities indistinguishable by the present model.

5.2. DNA Shape Signatures are Differentiable and Informative

Logistic regression analysis of the DNA shape features revealed numerous insights within and across individual KZFP binding profiles. That ZNF282 arises as having the highest average MGW, Roll, and HelT profile and the lowest average ProT profile across the 150 bp sequence window may be explained by its zinc finger array. Significant ProT signal across ZNF777 binding may similarly suggest distinguishing traits about its physical composition. MGW plays a dominant role for several KZFPs while Roll comprises all ten of the top ten DNA shape features of ZNF777. These findings support the notion that KZFPs recognize shape-dependent structural motifs beyond sequence identities alone, a notion supported by recent findings (Kalsan et al., 2025).

5.3. Ancestral KZFP Sequence Signals Cluster Together

While PCA showed some separation of three KZFPs, most of the signal for all six KZFPs was centered in one noisy area. Similarly, The UMAP plot showed tight overlapping of all six KZFPs. This is partly due to the imbalance of the dataset used for analysis. It may also suggest that binding of these ancestral KZFPs arises from nuanced features that are not captured globally.

5.4. Next Generation Sequencing Data and Machine Learning Are Synergistic

The application of machine learning methods to next generation sequencing data provides significant opportunities for researchers to address questions about genome evolution and gene regulation that can only be partially answered by one domain or the other. Future work could investigate young KZFPs in contrast to ancestral KZFPs like the six analyzed here. Integrating gene annotations would provide additional context and contribute further to understanding the genetic mechanisms involved in the history of life on Earth.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Google Brain. (2016). TensorFlow: a system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)* (pp. 265-283). https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

Alharbi, W. S., & Rashid, M. (2022). A review of deep learning applications in human genomics using next-generation sequencing data. *Human Genomics, 16*(26), 1-20. https://doi.org/10.1186/s40246-022-00396-x

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology, 33*(8), 831-839. https://doi.org/10.1038/nbt.3300

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Tasli, C., & Zhang, J. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLOS Computational Biology, 9*(11), 1-8. https://doi.org/10.1371/journal.pcbi.1003326

Blender Development Team. (2025). Blender (Version 5.0.0) [Computer software]. https://www.blender.org

Bruno, M., Farhana, S. M., Mitra, A., Costello, K., Watkins-Chow, D. E., Logsdon, G. A., Gambodi, C. W., Dumont, B. L., Black, B. E., Keane, T. M., Ferguson-Smith, A. C., Dale, R. K., & Macfarlan, T. S. (2025). Young KRAB-zinc finger gene clusters are highly dynamic incubators of ERV-driven genetic heterogeneity in mice. *Nature Communications, 16*(9608), 1-16. https://doi.org/10.1038/s41467-025-64609-2

Chiu, T., Comoglio, F., Zhou, T., Yang, L., Paro, R., & Remo, R. (2016). DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics, 32*(8), 1211-1213. https://doi.org/10.1093/bioinformatics/btv735

Dablain, D., Jacobson, K. N., Bellinger, C., Roberts, M., & Chawla, N. V. (2024). Understanding CNN fragility when learning with imbalanced data. *Machine Learning, 113*, 4785-4810. https://doi.org/10.1007/s10994-023-06326-9

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience, 10*, 1-4. https://doi.org/10.1093/gigascience/giab008

Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & Development, 25,* 1010-1022. https://doi.org/10.1101/gad.2037511

Gibbs, R. A. (2020). The Human Genome Project changed everything. *Nature Reviews Genetics, 21,* 575-576. https://doi.org/10.1038/s41576-020-0275-3

Gorkin, D. U., Lee, D., Reed, X., Fletez-Brant, C., Bessling, S. L., Loftus, S. K., Beer, M. A., Pavan, W. J., & McCallion, A. S. Integration of ChIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes. (2012). *Genome Research, 22,* 2290-2301. https://doi.org/10.1101/gr.139360.112

Gupta, Y. M., Kirana, S. N., & Homchan, S. (2025). Representing DNA for machine learning algorithms: a primer on one-hot, binary, and integer encodings. *Biochemistry and Molecular Biology Education, 53,* 142-146. https://doi.org/10.1002/bmb.21870

Harris, C. R., Millman, J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E. Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Rio, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). *Nature, 585*, 357-362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: a 2D graphics environment. *Computing in science & engineering, 9*(03), 90-95. https://doi.org/10.1109/MCSE.2007.55

Imbeault, M., Helleboid, P., & Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature, 543*, 550-554. https://doi.org/10.1038/nature21683

Jmol Development Team. (2025). Jmol: an open-source Java viewer for chemical structures in 3D (Version 16.3) [Computer software]. http://www.jmol.org/

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science, 316*(5830), 1497-1502. https://doi.org/10.1126/science.1141319

Kagda, M. S., Lam, B., Litton, C., Small, C., Sloan, C. A., Spragins, E., Tanaka, F., Whaling, I., Gabdank, I., Youngworth, I., Strattan, J. S., Hilton, J., Jou, J., Au, J., Lee, J., Andreeva, K., Graham, K., Lin, K., Simison, M., Jolanki, O., … Hitz, B. C. (2025). *Nature Communications, 16*(9592), 1-11. https://doi.org/10.1038/s41467-025-64343-9

Kalsan, M., Mirza, S., Bathla, D., & Ahmad, S. (2025). Dictionary based approaches for studying intrinsic DNA shape in transcription factor recognition. *Current Opinion in Structural Biology, 95*(103166), 1-9. https://doi.org/10.1016/j.sbi.2025.103166

Lavery, R., & Sklenar, H. (1989). Defining the structure of irregular nucleic acids: conventions and principles. *Journal of Biomolecular Structure and Dynamics*, *6*(4), 655-667. https://doi.org/10.1080/07391102.1989.10507728

Lee, D., Karchin, R., & Beer, M. A. (2011). Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Research, 21,* 2167-2180. https://doi.org/10.1101/gr.121905.111

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Liu, H., Chang, L., Sun, Y., Lu, X., & Stubbs, L. (2014). Deep vertebrate roots for mammalian zinc finger transcription factor subfamilies. *Genome Biology and Evolution, 6*(3), 510-525. https://doi.org/10.1093/gbe/evu030

McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv. https://doi.org/10.48550/arXiv.1802.03426

McKinney, W. (2010). Data structures for statistical computing in Python. *scipy, 445*(1), 51-56. https://doi.org/10.25080/Majora-92bf1922-00a

Patel, A., & Cheng, X. (2018). mouseZFP568-ZnF1-11 in complex with DNA. https://doi.org/10.2210/pdb5V3M/pdb

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830. https://dl.acm.org/doi/10.5555/1953048.2078195

Perez, G., Barber, G. P., Benet-Pages, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J. N., Hinrichs, A. S., Lee, C. M., Nassar, L. R., Raney, B. J., Speir, M. L., van Baren, M. J., Vaske, C. J., Haussler, D., Kent, W. J., & Haeussler, M. (2025). The UCSC Genome Browser database: 2025 update. *Nucleic Acids Research, 53,* D1243-D1249. https://doi.org/10.1093/nar/gkae974

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A., Das, G., & Malonia, S. K. (2023). Next-generation sequencing technology: current trends and advancements. *Biology, 12*(997), 1-25. https://doi.org/10.3390/biology12070997

Senft, A. D., & Macfarlan, T. S. (2021). Transposable elements shape the evolution of mammalian development. *Nature Reviews Genetics, 22*, 691-711. https://doi.org/10.1038/s41576-021-00385-1

Shirley, M., Ma, Z., Pedersen, B., & Wheelan, S. (2015). Efficient “pythonic” access to FASTA files using pyfaidx. *PeerJ PrePrints 3*:e970v1. https://doi.org/10.7287/peerj.preprints.970v1

Storey, J. D. (2003). The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics, 31*(6), 2013-2035. https://doi.org/10.1214/aos/1074290335

The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*, 57-74. https://doi.org/10.1038/nature11247

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software, 6*(60), 3021. https://doi.org/10.21105/joss.03021

Yang, B., Peng, Y., Leung, H. C., Yiu, S., Chen, J., & Chin, F. Y. (2010). Unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers. *BMC Bioinformatics, 11*(Suppl 2), S5. https://doi.org/10.1186/1471-2105-11-S2-S5

Yang, J., Ma, A., Hoppe, A. D., Wang, C., Li, Y., Zhang, C., Wang, Y., Liu, B., & Ma, Q. (2019). Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. *Nucleic Acids Research, 47*(15), 7809-7824. https://doi.org/10.1093/nar/gkz672

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods, 12,* 931–934. https://doi.org/10.1038/nmeth.3547

Zhou, T., Yang, L., Lu, Y., Dror, I., Machado, A. C. D., Ghane, T., Felice, R. D., Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Research, 41*(W1), W56-W62. https://doi.org/10.1093/nar/gkt437