



UNIVERSITÀ
DI PAVIA

Dipartimento di Biologia e Biotecnologie
“Lazzaro Spallanzani”

Laurea Magistralis in Molecular Biology and Genetics

Computational prediction of Transcription factor (TF) isoforms-
DNA interactions with RoseTTAFold2NA, AlphaFold 3, and
Boltz-2

Predizione computazionale delle interazioni tra isoforme di fattori
di trascrizione (TF) e DNA con RoseTTAFold2NA, AlphaFold 3 e
Boltz-2

Supervisor:

Prof. Federico Forneris

Co-supervisor:

Dr. Khalique Newaz

Experimental thesis by

Sara Soleimani Jevinani

Sommario

Lo splicing alternativo dei geni dei fattori di trascrizione genera isoforme proteiche che possono differire in modo sostanziale dalle controparti di riferimento per specificità di legame al DNA e attività trascrizionale, tuttavia la base strutturale di questa divergenza funzionale all'interfaccia proteina-DNA rimane ancora poco caratterizzata. Presentiamo un confronto strutturale sistematico di 693 isoforme derivate da 246 geni di fattori di trascrizione, utilizzando tre framework di deep learning complementari, AlphaFold 3, RoseTTAFold2NA e Boltz-2, per modellare e confrontare le strutture tridimensionali predette dei complessi isoforma-DNA di riferimento e alternativi. Dimostriamo che le metriche di confidenza strutturale di tutte e tre le piattaforme sono sensibili alle differenze di sequenza specifiche delle isoforme localizzate al di fuori del dominio di legame al DNA, e che le isoforme alternative che conservano un dominio di legame al DNA completamente intatto mostrano in modo consistente un errore posizionale inter-dominio predetto inferiore rispetto alle controparti di riferimento in tutte e tre le piattaforme. Questo schema di stabilizzazione strutturale è coerente con un modello autoinibitori in cui le regioni intrinsecamente disordinate fiancheggianti nelle isoforme di riferimento vincolano fisicamente il dominio di legame al DNA e ne limitano la libertà conformazionale, e la loro rimozione attraverso lo splicing alternativo libera questo vincolo. L'analisi di otto isoforme selezionate attraverso il filtraggio sperimentale di dati discordanti di interazione proteina-DNA mostra che i segnali di stabilizzazione strutturale sono più consistenti e interpretabili quando tutte e tre le piattaforme concordano, con DLX4-2 che mostra le evidenze più forti supportate da dati sperimentali, e PRRX1-3 che rappresenta il candidato computazionale novel più coerente. L'analisi della curva ROC su 79 isoforme uniche dimostra che nessuna metrica strutturale individuale predice in modo affidabile gli esiti di legame sperimentali quando i modelli vengono applicati senza addestramento su dati di legame, sebbene un modello di regressione logistica combinato che integra le metriche di tutte e tre le piattaforme raggiunga un'area sotto la curva di 0.730. Questi risultati forniscono un framework computazionale multi-piattaforma validato tramite benchmarking per caratterizzare le differenze strutturali tra le isoforme dei fattori di trascrizione e stabiliscono una base di riferimento per futuri approcci che incorporano dati di legame sperimentali nell'addestramento dei modelli.

Abstract

Alternative splicing of transcription factor genes generates protein isoforms that differ substantially from their reference counterparts in DNA-binding specificity and transcriptional output, yet the structural basis for this functional divergence at the protein-DNA interface remains poorly characterized. We present a systematic structural comparison of 693 isoforms derived from 246 transcription factor genes, using three complementary deep learning frameworks, AlphaFold 3, RoseTTAFold2NA, and Boltz-2, to model and compare the predicted three-dimensional structures of reference and alternative isoform-DNA complexes. We show that structural confidence metrics from all three platforms are sensitive to isoform-specific sequence differences located outside the DNA-binding domain, and that alternative isoforms retaining a fully intact DNA-binding domain consistently show lower predicted inter-domain positional error than their reference counterparts across all three platforms. This pattern of structural stabilization is consistent with an autoinhibitory model in which flanking intrinsically disordered regions in reference isoforms physically constrain the DNA-binding domain and limit its conformational freedom, and their removal through alternative splicing releases this constraint. Analysis of eight isoforms selected through experimental filtering of discordant protein-DNA interaction data shows that structural stabilization signals are most consistent and interpretable when all three platforms agree, with DLX4-2 showing the strongest cross-platform evidence supported by experimental data, and PRRX1-3 representing the most coherent novel computational candidate. Receiver operating characteristic analysis across 79 unique isoforms demonstrates that no individual structural metric reliably predicts experimental binding outcomes when the models are applied without training on binding data, though a combined logistic regression model integrating metrics from all three platforms achieves an area under the curve of 0.730. These results provide a benchmarked multi-platform computational framework for characterizing structural differences between transcription factor isoforms and establish a performance baseline for future approaches that incorporate experimental binding data into model training.

Table of Contents

1 INTRODUCTION.....	9
1.1 DNA and chromosomes.....	9
1.2 Transcription.....	10
1.3 Transcriptional Regulation.....	12
1.4 Transcription Factors.....	13
1.4.1 General transcription factors (GTFs).....	14
1.4.2 Sequence-specific transcription factors.....	15
1.4.3 The Regulatory Landscape and Chromatin Accessibility.....	17
1.4.4 Molecular Principles and Modular Domain Architecture.....	18
1.4.5 Structural Plasticity: The Continuum of Order and Disorder in TFs.....	19
1.4.6 Network Complexity and Isoform Diversity.....	20
1.4.7 Impact on Human Health and Disease.....	20
1.5 Recognition of Transcription Factor Binding Sites.....	21
1.6 Motif Databases.....	22
1.7 Alternative Splicing.....	23
1.8 Transcription Factor Isoforms.....	26
1.8.1 Two major classes of alternative isoforms: Negative regulators and rewirers.....	27
1.8.2 TF- DNA interaction.....	28
1.8.3 DNA Binding Is Influenced by Sequences Both Inside and Outside the DBD.....	29
1.8.4 The Influence of Flanking and Distal Regions on DNA Binding.....	29
1.8.5 The Autoinhibitory Model: Evidence from DLX4 and PKNX1.....	30
1.8.6 Developmental Implications: The TBX5 Model.....	30
1.9 Computational prediction of protein-DNA interactions.....	31
1.9.1 RF2NA: Generalized Prediction of Protein-Nucleic Acid Systems.....	31
1.9.2 AF3: A Unified Framework for All-Atom Structural Prediction.....	32
1.9.3 Boltz-2: Assessing Interaction Stability and Binding Affinity.....	32
1.10 Aims of the research.....	33
2 Material and Methods.....	35
2.1 Data Collection and Processing.....	35
2.2 Mathematical Framework for Information Content.....	35
2.2.1 Handling Multiple Motifs and Selection Logic.....	36
2.3 Statistical Analysis.....	37
2.4 Structural Modeling and Predictive Framework.....	37
2.4.1 Architectural Framework of RF2NA.....	38
2.4.2 AF3: A Unified All-Atom Architecture.....	39
2.4.3 Boltz-2: Integrated Structure and Affinity Prediction.....	40
2.4.4 Model Quality and Validation Metrics.....	41
2.5 Receiver operating characteristic (ROC) Curve Analysis.....	42
2.5.1 Dataset Preparation and Deduplication.....	42
2.5.2 Logistic Regression Model.....	43
2.5.3 Correlation Analysis.....	43
3 Results.....	45

3.1 Performance comparison of RF2NA, AF3, and Boltz-2 on protein-DNA complex predictions for reference TFs.....	45
3.1.1 Assessment of Model Reliability and Success Thresholds.....	46
3.1.2 AF3 Prediction Performance Across TF Families.....	49
3.2 Performance comparison of RF2NA, AF3, and Boltz-2 on protein-DNA complex predictions for alternative TFs not perturbing DBDs.....	51
3.2.1 Paired Comparative Structural Analysis of Alternative TF Isoforms Without DBD Loss, Ordered by RF2NA Prediction.....	52
3.3 Structural impact of DBD loss in alternative TF isoforms.....	61
3.4 Evaluating the Predictive Accuracy of AI Structural Models for Protein–DNA Interaction...	63
4 Discussion.....	69
5 Conclusion.....	79
6 References.....	80
Acknowledgment.....	89

Figures

Figure 1.1 The organization of DNA within the chromatin structure.....	11
Figure 1.2 Overview of eukaryotic transcription.....	12
Figure 1.3 Eukaryotic transcriptional regulation.....	14
Figure 1.4 Schematic of DNA-binding and effector domains in TFs from different families.....	19
Figure 1.5 Schematic of general TF architecture illustrating IDR sequence, structure, and functional features.....	21
Figure 1.6 Schematic of alternative splicing.....	25
Figure 1.7 Types of commonly observed splicing events.....	27
Figure 2.1 Isoform classification flowchart.....	38
Figure 2.2 Architectural Overview of RFNA.....	40
Figure 2.3 AF3 Inference Architecture.....	41
Figure 2.4 Boltz-2 Model Architecture.....	42
Figure 3.1 Comparative evaluation of structural prediction methods for reference TF–DNA complexes.....	47
Figure 3.2 Distribution of AF3 ipTM scores for reference and alternative isoforms. The red dashed line indicates the success threshold of ipTM greater than or equal to 0.6.....	48
Figure 3.3 RF2NA mean PAE score distribution by isoform type. The red dashed line indicates the success threshold of PAE below 10.....	50
Figure 3.4 AF3 prediction performance across all TF families represented in the dataset.....	52
Figure 3.5 Paired comparison of structural confidence metrics between reference and alternative TF–DNA complexes lacking DBD loss across prediction frameworks.....	53
Figure 3.6 Paired comparative analysis of structural metrics alternative TF Isoforms without DBD loss ordered by RF2NA Prediction (A-H panels).....	59
Figure 3.7 Structural modeling performance of AF3, Boltz-2, and RF2NA on TF isoforms with DNA-binding DBD loss.....	64
Figure 3.8 Receiver operating characteristic (ROC) analysis of AI structural confidence metrics for predicting protein–DNA interactions across 79 unique isoforms.....	67
Figure 3.9 Correlation analysis of AI structural confidence metrics against experimental binding outcomes for 79 unique isoforms.....	69

Tables

Table 1.1 General transcription factors and their functions.....	16
Table 2.1 Confidence metrics used to evaluate structural predictions.....	43

List of Abbreviations

Transcription factor	TF
Deoxyribonucleic acid	DNA
RNA Polymerase II	Pol II
TATA-binding protein	TBP
Transcription factor IIE	TFIIE
Transcription factor IIH	TFIIH
Transcription Factor II D	TFIID
DNA Binding Domain	DBD
pre-initiation complex	PIC
Alternative Splicing	AS
C-terminal domain	CTD
general transcription factors	GTFs
transcription factor binding sites	TFBS
cis-regulatory elements	CREs
position weight matrices	PWMs
MANE	Matched Annotation from NCBI and EMBL-EBI
HOCOMOCO	HOmo sapiens COmprehensive MOdel COllection
RoseTTAFold2NA	RF2NA
AlphaFold3	AF3
Intrinsically Disordered Proteins	IDPs
Intrinsically Disordered Regions	IDRs
Open Reading Frames	ORFs
Benjamini–Hochberg	BH
False discovery rate	FDR
Receiver Operating Characteristic	ROC

1 INTRODUCTION

1.1 DNA and chromosomes

Deoxyribonucleic acid (DNA) is a molecule consisting of a sugar-phosphate backbone and four nitrogenous bases: adenine (A), thymine (T), cytosine (C), and guanine (G). It consists of two anti-parallel strands that are twisted to form a double helix. In eukaryotic cells, DNA is localized in the cell nucleus. In order to fit the DNA molecules, which in humans would be about 2 meters in length, are compacted in several ways [1]. First, 147 bases of DNA are wound around eight positively charged proteins called histones to form a nucleosome. The histone complex consists of two molecules each of histone H2A, H2B, H3 and H4. Another type of histone, H1, acts as a linker between adjacent nucleosomes. Nucleosomes are then compacted into two levels of fibers, and finally the fibers are further compacted into chromosomes that are visible in a light microscope after special treatment at cell division, mitosis. The mixture of DNA and proteins that builds up the chromosomes is called chromatin [2]. An overview of DNA organization is shown in Figure 1.1.

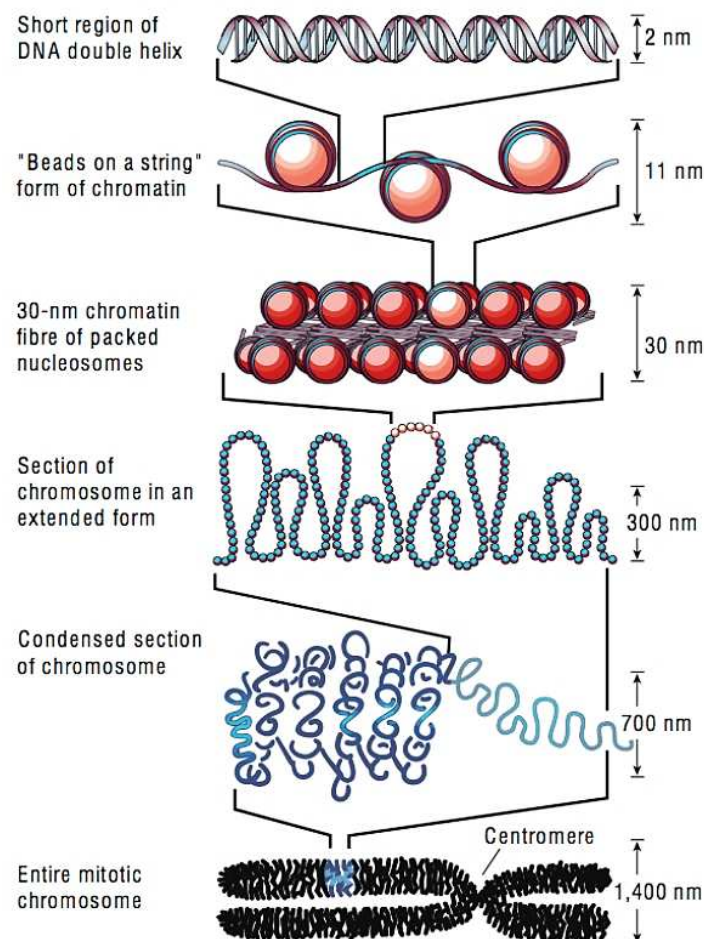


Figure 1.1 The organization of DNA within the chromatin structure. (Taken from Felsenfeld & Groudine, 2003).

1.2 Transcription

Transcription is the first and most fundamental step in gene expression, serving as the primary mechanism through which the information encoded in DNA is converted into functional RNA molecules. It refers to the synthesis of an RNA molecule based on a DNA sequence template [3]. This process is carried out by RNA polymerase (RNA Pol), a conserved enzyme found in all living organisms. In eukaryotes, RNA polymerases are subdivided into three functionally distinct classes: RNA Pol I transcribes ribosomal RNA genes, RNA Pol II synthesizes messenger RNA and small non-coding RNAs, and RNA Pol III produces transfer RNAs, 5S RNA, and a variety of other small non-coding RNAs [4]. Although all three polymerases share a conserved catalytic core, they rely on distinct sets of peripheral factors to recognize their respective gene classes and regulate transcriptional output [5].

Transcription proceeds through three sequential steps: initiation, elongation, and

termination (Figure 1.2). The process is tightly regulated, responding to metabolic requirements, developmental programs, and diverse external stimuli. This regulation is mediated at multiple levels by specific DNA sequences located in the vicinity of target genes, collectively known as cis-regulatory sequences [6]. In eukaryotes, the core promoter spans approximately 50 nucleotides upstream and downstream of the transcription start site (TSS) and contains sequence motifs that position RNA Pol II for accurate initiation [7]. Further upstream lie proximal promoter elements and, often at considerably greater genomic distances, enhancer sequences that amplify or modulate transcriptional activity [8]. The regulatory effects of these elements are exerted through trans-acting transcription factors, which bind the DNA motifs and recruit or modulate the transcriptional machinery [9, 10].

Eukaryotic Transcription

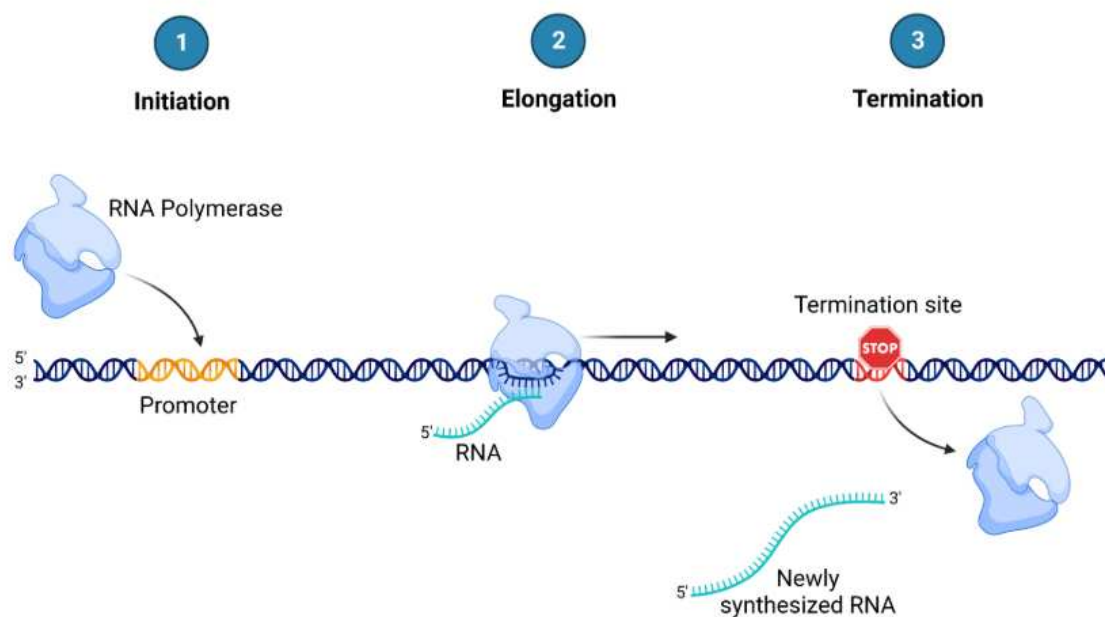


Figure 1.2 Overview of eukaryotic transcription. (Adapted from Monash University Student Academic Success).

Eukaryotic core promoters exist in multiple forms and contain diverse DNA sequence motifs that collectively determine transcription factor binding and local chromatin state. The TATA box, an AT-rich sequence element, is the best-characterized core promoter motif and serves as the principal platform for assembly of the pre-initiation complex [11]. This complex comprises a set of evolutionarily conserved general transcription factors, including TFIIA, TFIIB, TFIID, TFIIIE, TFIIF, and TFIIF. Assembly begins

when TFIID, acting through its TATA-binding protein subunit (TBP), recognizes and loads onto the promoter DNA [12, 13]. TFIIB then recognizes the TBP-DNA complex and makes direct contact with RNA Pol II, bridging the polymerase to the promoter and precisely positioning its catalytic center over the transcription start site [14].

Following this initial assembly, TFIIF associates with the TBP-TFIIB-Pol II platform, stabilizes the closed pre-initiation complex, facilitates DNA strand opening at the promoter, and helps determine the precise location of the TSS [14]. In the final step of initiation, TFIIIE recruits TFIIH and stimulates its ATP-dependent helicase and translocase activities, which unwind the DNA duplex and enable RNA Pol II to break free from the promoter and enter productive elongation. The transition from initiation to elongation is also marked by phosphorylation of the C-terminal domain of the largest RNA Pol II subunit by the TFIIH kinase subunit CDK7, which modifies serine residues at positions 5 and 7 and signals the handoff from promoter-bound initiation to processive RNA synthesis [5].

1.3 Transcriptional Regulation

Transcriptional regulation is a dynamic process essential for maintaining cellular identity and adapting gene expression to environmental signals [15]. Gene expression adjusts continuously in response to conditions such as cellular stress or developmental cues. A well-characterized example is the rapid transcriptional induction of heat-shock genes under proteotoxic stress, which drives the production of protective chaperone proteins to maintain protein homeostasis [16]. More broadly, the precise regulation of transcription allows each cell type to fulfill its specialized functional role and plays a fundamental part in connecting genetic variation to phenotypic outcomes [17]. This regulation is achieved through a complex and highly coordinated interplay between TFs, co-regulators, and cis-regulatory elements.

Transcriptional regulation is influenced by binding site context, the presence of cofactors, and the cellular environment [18]. The transcription machinery consists of proteins and complexes, including RNA polymerase, TFs, cofactors, and chromatin remodelers, which coordinate RNA synthesis from DNA. Gene expression levels change depending on the concentration of TFs, following a relationship that reflects how gene activity responds to varying TF levels [19]. Cofactors enhance the stability of TF-DNA interactions or facilitate the recruitment of additional proteins that modify chromatin structure, further supporting the transcription process. In eukaryotic cells, DNA is organized into chromatin, where it wraps around histone proteins. In its natural

state, transcription remains inactive until specific regulatory interactions take place, which means most genes are not expressed by default [20]. Transcription begins with recruitment, where TFs and proteins help RNA polymerase II bind to the promoter [21].

Transcription is also regulated by epigenetic mechanisms that alter gene expression without changing the DNA sequence. DNA methylation is a major repressive mechanism that can silence genes by blocking TF binding [22]. Additionally, histone modifications affect chromatin structure, making regions of DNA more or less accessible, while chromatin remodeling shifts nucleosomes to influence transcriptional activity. Gene expression is a complex process regulated beyond transcription as post-transcriptional mechanisms play a crucial role after RNA is synthesized from DNA [23]. Post-transcriptional regulation controls protein levels after mRNA is produced Figure 1.3 [24].

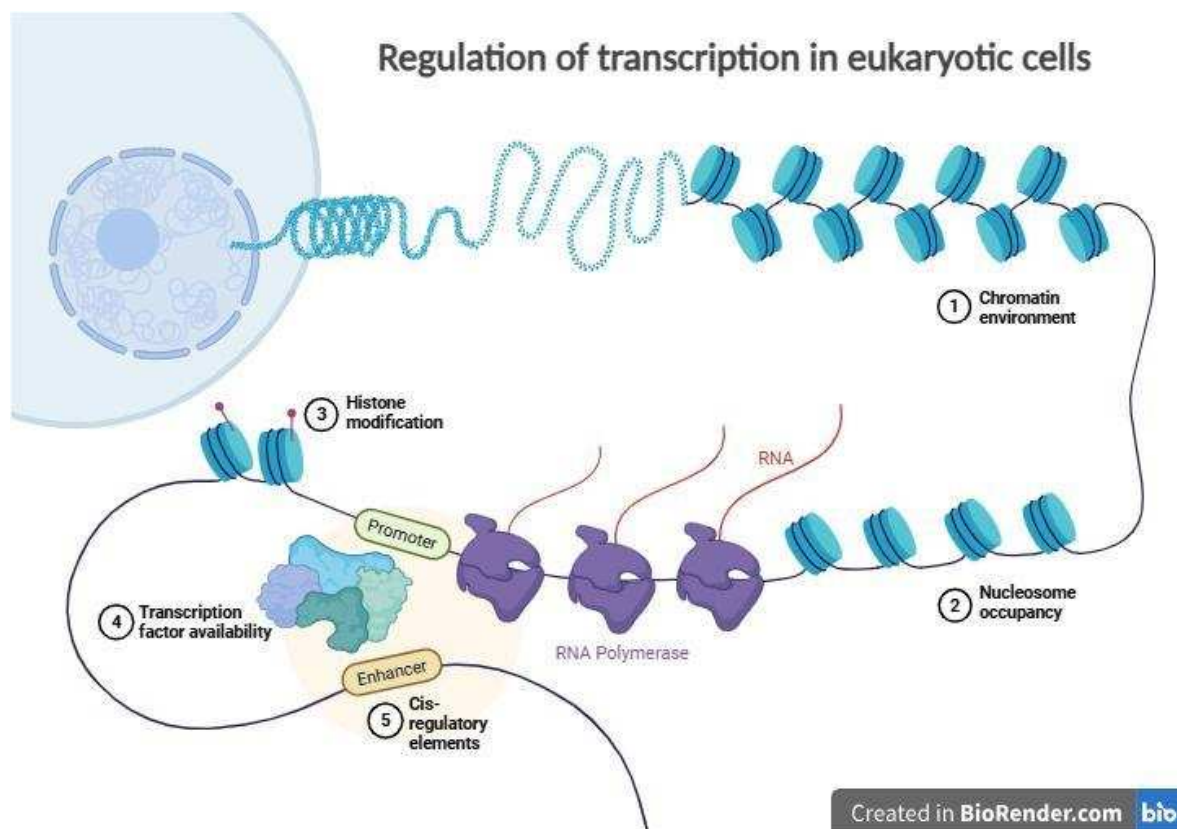


Figure 1.3 Eukaryotic transcriptional regulation. (Created with BioRender.com).

1.4 Transcription Factors

Transcription factors (TFs) are key regulators of cellular processes, governing both intrinsic programs such as development and differentiation [25], and extrinsic responses to environmental signals [26]. They are encoded by approximately 1,600 genes in the

human genome, comprising the single largest protein family in mammals. Each cell type expresses approximately 150-400 TFs, which together define and maintain the gene expression program characteristic of that cell type [27]. By directly interpreting the DNA sequence, TFs perform the first and most fundamental step in decoding genomic information. Consequently, many TFs function as "regulators" or "selector genes," exerting control over processes that specify cell identity and developmental patterning [28], as well as coordinating specific physiological pathways such as immune responses [29].

At the molecular level, TFs are proteins that bind to specific DNA sequences known as DNA binding sites or transcription factor binding sites (TFBS) and exert either activating or repressive effects on the transcriptional rate of their target genes [30]. Through this sequence-specific binding, TFs regulate the expression of particular genes and thereby influence a wide range of biological processes, including cell growth, differentiation, and adaptation to environmental signals [31].

The regulation of gene expression is mediated by the interplay between two functional classes of transcription factors: general transcription factors (GTFs), which are required for the basal expression of virtually all Pol II-transcribed genes, and sequence-specific DNA-binding transcription factors (STFs), which provide gene- and context-specific regulatory control [32].

1.4.1 General transcription factors (GTFs)

As introduced above, transcription of protein-coding genes by RNA Polymerase II proceeds through three major phases: initiation, elongation, and termination. The initiation phase requires the stepwise assembly of general transcription factors (GTFs) and Pol II at the core promoter to form the pre-initiation complex (PIC), which serves as the platform from which transcription is launched [6]. The major GTFs are summarized in Table 1.1.

GTF	Functions
TFIID	Contains two subunits, TATA-binding protein (TBP) and TBP-associated factor (TAF). TBP binds to the TATA box of the core promoter in an orientation-independent manner while TAF binds to INR and DPE and is required for promoter selection and transcriptional activation. Eukaryotes also encode a TBP-related factor (TRF) which recognises other DNA sequences in TATA-less promoters for the initiation of transcription.
TFIIA	A heterodimer which interacts with TBP and stabilises the TBP-DNA interaction. Promotes binding of TFIID to DNA. Involved in the activation of transcription by binding to activators.
TFIIB	Interacts with TFIID and RNA Pol II and is required for transcription start site selection. Binds to BRE and downstream sequences of the TATA box. A direct interacting partner of activators which may promote the recruitment of TFIIB to the promoter
TFIIF	A heterodimer containing two subunits, TFIIFβ and TFIIFα. TFIIFβ binds to either upstream or downstream of TATA box while TFIIFα binds to regions downstream of the TATA box. Binds RNA Pol II tightly and is involved in avoiding non-specific DNA binding and stabilisation of the PIC.
TFIIE	Binds to DNA sequences directly upstream of the transcription start site after the formation of the PIC. Promotes the recruitment of TFIIH and stimulates the CTD kinase and helicase activities of TFIIH.
TFIIH	Contains two subunits, a core subunit with helicase activities (XPD and XPB) and a kinase subunit (Cdk7). The helicase subunit is required in the unwinding of DNA to create an open structure while the kinase subunit phosphorylates the CTD in RNA pol II

	during the transition to elongation stage.
--	--

Table 1.1 General transcription factors and their functions

1.4.2 Sequence-specific transcription factors

While the GTFs and RNA Polymerase II account for the basal transcriptional machinery, they alone are insufficient to drive the precise, context-dependent regulation of individual genes. This role is fulfilled by sequence-specific transcription factors (STFs), which bind to cis-regulatory regions including promoters, enhancers, and silencers to activate or repress transcription in a gene- and cell-type-specific manner. STFs recognize and bind to transcription factor binding sites (TFBS), which are typically short, evolutionarily conserved DNA sequences that tolerate a certain degree of sequence degeneracy [33]. Several key properties define how STFs carry out their regulatory functions.

Modular architecture

STFs are characteristically modular proteins, composed of functionally distinct domains that can act independently or in combination. These typically include a DNA-binding module, a dimerization module, an activation or repression module, and in some cases a regulatory module. The modularity of TFs was first demonstrated in the yeast GAL4 protein, where the DNA-binding and activation functions were shown to be separable and interchangeable [34]. A wide variety of DNA-binding module families have since been identified, including the helix-turn-helix, homeodomain, zinc finger, and leucine zipper motifs [35]. Activation modules are equally diverse and can take the form of acidic or negatively charged alpha helices (Staller et al., 2018), glutamine- or proline-rich regions [36, 37], or hydrophobic beta sheets [38].

Recruitment of coactivators and corepressors

Many STFs do not act directly on the core transcriptional machinery but instead exert their effects through the recruitment of coactivators or corepressors via protein–protein interactions. Some cofactors, such as TBP-associated factors (TAFs), serve as bridging molecules that physically connect STFs to the general transcriptional machinery. Others, such as ATP-dependent chromatin remodelers and histone-modifying enzymes, alter the local chromatin landscape to either facilitate or restrict Pol II access, thereby amplifying the activation or repression signal [39].

Combinatorial control

The transcriptional output of a gene is rarely determined by a single TF acting in isolation. Instead, clusters of distinct TFBS are frequently embedded within cis-regulatory elements, creating a combinatorial code that requires the cooperative binding of multiple STFs. Such combinations of TFs often act synergistically, producing a transcriptional response that is greater than the sum of their individual contributions [40]. This combinatorial logic allows a relatively limited repertoire of TFs to generate the enormous diversity of gene expression patterns observed across cell types and developmental stages.

Post-translational modification

The activity of STFs is also dynamically regulated through post-translational modifications (PTMs), which can rapidly alter their DNA-binding affinity, transcriptional activity, stability, or subcellular localization in response to signaling events. A well-characterized example is the phosphorylation of the cyclic AMP response element-binding protein (CREB) by protein kinase A (PKA) upon cAMP stimulation, which activates CREB and drives transcription of target genes [41]. Similarly, acetylation of p53 enhances its DNA-binding affinity [42], while ubiquitylation can regulate TF turnover and transcriptional potency.

Multiprotein families

A further source of regulatory complexity is that many STFs belong to multiprotein families whose members share highly similar or identical DNA-binding motifs yet fulfill distinct biological roles. The GATA family of transcription factors provides a well-studied example. The six GATA family members (GATA1–6) are subdivided into two functionally distinct sub-families: the haematopoietic sub-family (GATA1, 2, and 3), which controls lineage commitment and specification in haematopoietic and neuronal cell lineages, and the non-haematopoietic sub-family (GATA4, 5, and 6), which governs cardiac-specific gene expression and epithelial differentiation in the digestive tract [43]. Despite their shared structural features, the distinct expression patterns and interaction partners of family members allow them to regulate non-overlapping sets of target genes at specific stages of development.

1.4.3 The Regulatory Landscape and Chromatin Accessibility

Gene transcription is precisely coordinated in response to both intracellular and extracellular signals, and this coordination depends critically on the genomic context in which TFs operate. TF binding sites are embedded within distinct classes of regulatory

regions: proximal elements located near the transcription start site (TSS), known as promoters, and distal regulatory elements called enhancers, which can act over distances of hundreds of kilobases to modulate gene expression [44]. At both promoters and enhancers, TFs work cooperatively with cofactors and other regulatory proteins to enable the recruitment and stabilization of the Pol II transcriptional machinery at the gene body [45].

A key challenge in this process is that most regulatory regions are occluded by nucleosomes under basal conditions, rendering them inaccessible to conventional TFs. A specialized class of TFs known as pioneer factors overcomes this barrier by possessing the unique ability to recognize and bind their target sequences even within compacted nucleosomal chromatin [46]. By engaging closed chromatin, pioneer factors initiate local remodeling, recruiting ATP-dependent chromatin remodelers and histone-modifying enzymes thereby increasing DNA accessibility and creating a permissive environment for the subsequent binding of additional TFs and the general transcriptional machinery.

1.4.4 Molecular Principles and Modular Domain Architecture

To understand how TFs carry out their regulatory roles, it is important to examine the molecular principles that guide their interaction with DNA and the structural features that enable sequence-specific binding. The canonical view of TFs describes them as modular proteins functioning through at least two core domains: one that contacts DNA directly and one that mediates protein–protein interactions [25].

The first and most defining component is the DNA-Binding Domain (DBD). DBDs form stable, well-conserved structural folds that recognize and bind specific DNA sequences, and multiple TFs collectively occupy enhancers and promoter-proximal regions to orchestrate gene expression [47, 48]. Because DBD sequences are evolutionarily conserved, they can be reliably detected by sequence homology and serve as the primary basis for TF classification into structural families including the C2H2 zinc finger, homeodomain, basic helix-loop-helix (bHLH), and basic leucine zipper (bZIP) families, among others [33, 49], as illustrated in Figure 1.4.

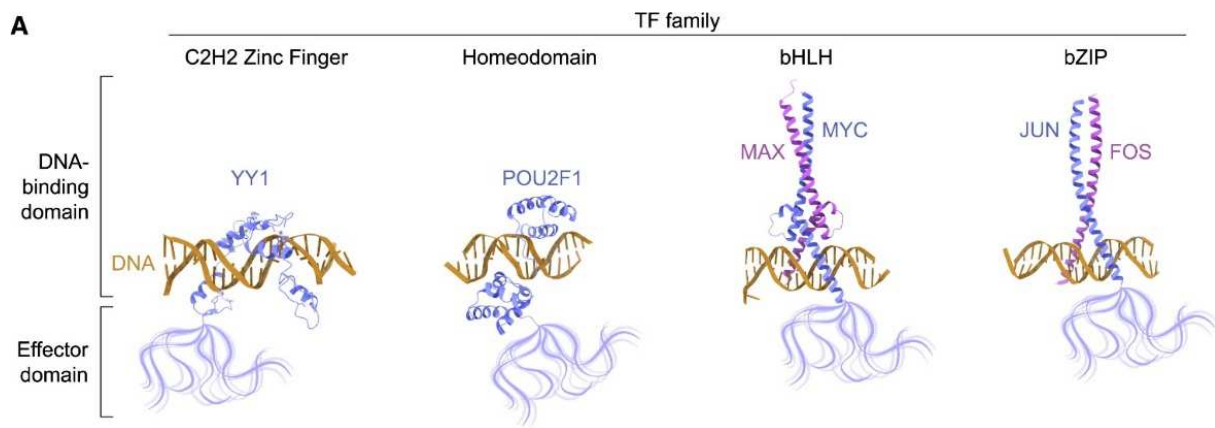


Figure 1.4 Schematic of DNA-binding and effector domains in TFs from different families. (Taken from Ozgur Oksuz et al., 2023).

The second major component is the effector domain. In contrast to DBDs, effector domains exhibit considerably less sequence conservation and adopt a wide range of transient, flexible conformations that facilitate multivalent protein–protein interactions [50, 51]. Through these interactions, effector domains recruit coactivator or corepressor complexes that contribute to gene regulation by mobilizing nucleosomes, modifying chromatin-associated proteins, shaping genome architecture, and controlling the efficiency of transcription initiation and elongation [52]. Together, the DBD and effector domain thus enable TFs to integrate sequence-specific DNA recognition with broader chromatin and transcriptional regulatory functions.

1.4.5 Structural Plasticity: The Continuum of Order and Disorder in TFs

Protein structures do not exist as static, uniform entities but rather occupy a conformational continuum, ranging from highly structured proteins with fixed tertiary folds to intrinsically disordered proteins (IDPs) that rapidly interconvert among a broad ensemble of conformational states. While approximately 60% of the human proteome consists of structured domains connected by short disordered linkers, TFs are disproportionately enriched in long intrinsically disordered regions (IDRs) (Figure 1.5), placing them toward the disordered end of this continuum [53].

These disordered segments are far from functionally inert. IDRs enable TFs to adopt distinct conformations in response to environmental cues, post-translational modifications, or binding partners, thereby expanding the regulatory repertoire of a single TF beyond what a rigid structured domain could achieve. They also facilitate dynamic, multivalent interactions with cofactors and components of the transcriptional machinery interactions increasingly recognized as central to transcriptional regulation

[53]. Yet precisely because of their structural heterogeneity, IDRs are difficult to purify and largely intractable to conventional high-resolution methods such as X-ray crystallography or cryo-EM, necessitating integrative approaches that combine ensemble techniques (NMR, SAXS), single-molecule methods (smFRET), and computational simulations to capture their dynamic behavior [54].

Figure 1.5 Schematic of general TF architecture illustrating IDR sequence, structure, and functional features. (Taken from Már et al., 2023).

TFs do not act in isolation but rather operate within highly interconnected gene regulatory networks, functioning through combinatorial interactions with multiple protein partners [55]. Traditionally, TFs have been categorized as either activators, which enhance transcription by facilitating Pol II recruitment, or repressors, which inhibit it by blocking polymerase access or recruiting silencing complexes [56]. However, this binary classification is an oversimplification. Many TFs are multifunctional, with their activating or repressive roles determined by cellular context, the availability of interacting partners, and the specific protein isoform expressed.

represents a central focus of the present thesis, as discussed in detail in subsequent sections.

1.4.7 Impact on Human Health and Disease

Gene regulatory networks (GRNs) comprising physical and functional interactions between TFs and their target regulatory elements play a critical role in orchestrating development and maintaining physiological homeostasis [58]. Disruption of these networks, through mutations that alter either TF function or the regulatory sequences they bind, underlies a broad spectrum of human diseases. Disease-associated mutations have been identified both in TF-encoding genes themselves and in non-coding regulatory sequences such as enhancers and promoters, highlighting that both the regulators and the regulated regions are frequent targets of pathogenic variation [28].

The scale of this problem is underscored by findings from genome-wide association studies (GWAS), which have revealed that approximately 90% of disease-associated genetic variants reside in the non-coding portion of the genome. Since non-coding variants are most likely to exert their effects by perturbing TF–DNA interactions at regulatory elements, a central challenge in modern genomics is to systematically map which TF interactions are disrupted by specific disease-associated variants and to understand the downstream consequences for gene regulation [58]. Addressing this challenge requires both experimental and computational approaches to characterize TF binding with high precision, a need that directly motivates the structural modeling framework developed in this thesis.

1.5 Recognition of Transcription Factor Binding Sites

As discussed above, TFs bind to specific genomic locations known as transcription factor binding sites (TFBS). These interactions are highly sequence-specific, enabling TFs to selectively regulate the transcription of their target genes [59]. The structural basis for this specificity lies in the DNA-binding domains (DBDs) of TFs, which make direct physical contact with the DNA. DBDs frequently adopt conserved structural motifs that are precisely shaped to engage specific DNA sequences. For example, the helix-turn-helix (HTH) motif inserts a recognition helix into the major groove of the DNA double helix, where it forms sequence-specific contacts with the exposed nucleotide bases [25, 60]. Similarly, the zinc finger motif relies on zinc ion coordination to stabilize the domain fold, enabling precise and stable engagement with specific DNA sequences.

Before sequence-specific binding can occur, TFs must first locate their target sites within the vast expanse of genomic DNA. This search process relies on a combination of diffusion mechanisms operating at different scales [61, 62]. In three-dimensional (3D) diffusion, TFs move freely through the nucleus, performing random collisions with DNA until they encounter a region of interest. Upon making initial contact, TFs frequently transition to one-dimensional (1D) sliding, gliding along the DNA strand and closely inspecting the local sequence. This is further supplemented by hopping and intersegment transfer, in which the TF transiently dissociates from the DNA and rebinds at a nearby site, effectively bypassing non-target sequences and accelerating the overall search [63]. Together, these complementary search strategies allow TFs to locate their binding sites with remarkable efficiency within long stretches of genomic DNA.

Once a target site is identified, binding specificity is determined by a network of molecular interactions between amino acid residues in the DBD and the nucleotide bases of the TFBS. These include hydrogen bonds, van der Waals forces, and ionic contacts, each contributing to the overall binding affinity and selectivity [64, 65]. Importantly, not all nucleotide positions within a TFBS contribute equally to binding; some positions are highly constrained while others tolerate substitutions with minimal effect on affinity. This positional variability means that TFs recognize a spectrum of related sequences with varying affinities rather than a single fixed sequence. This binding preference is commonly represented as a position weight matrix (PWM), a probabilistic model that captures the relative frequency of each nucleotide at each position within the binding site and serves as the standard framework for predicting TF occupancy across the genome [66]. While PWMs effectively describe the intrinsic sequence preferences of individual TFs, gene regulation *in vivo* involves additional layers of complexity, including DNA shape readout, chromatin accessibility, and cooperative interactions among multiple TFs aspects explored in subsequent sections.

1.6 Motif Databases

A motif and a transcription factor binding site (TFBS) are closely related but conceptually distinct terms in the study of TF-DNA interactions. A motif refers to a short, recurring DNA sequence pattern that represents the preferred binding sequence of a TF. Because TFs can recognize and bind a spectrum of similar sequences rather than a single invariant sequence, motifs summarize this binding preference in a probabilistic framework, most commonly as position weight matrices (PWMs). In

contrast, a TFBS refers to the actual genomic location such as within a promoter or enhancer where a TF physically binds to regulate gene expression. While TFBSs typically match or resemble the cognate motif, their occupancy *in vivo* is also shaped by additional factors including DNA shape, local sequence context, chromatin accessibility, and cooperative interactions with other proteins. Thus, while motifs capture the intrinsic sequence preference of a TF, TFBSs represent the functional genomic instances where this preference is realized in a living cell [55].

Two of the most widely used curated repositories of TF binding motifs are JASPAR and HOCOMOCO. JASPAR is a publicly accessible, non-redundant database of high-quality, matrix-based TFBS profiles for eukaryotic TFs, presented through a user-friendly web interface [67, 68]. Since its initial release in 2004, in which TF profiles were derived from experimentally validated binding sites identified by SELEX and related approaches, JASPAR has been updated on a biennial cycle. The most recent release at the time of writing, JASPAR 2024, represents the twentieth anniversary edition of the database and includes expanded motif collections alongside improved programmatic access through dedicated packages for R, Python, and Perl [69].

HOCOMOCO (HOMo sapiens COMprehensive MOdel COllection) serves as a complementary resource, providing high-quality PWMs primarily for human and mouse TFs derived from ChIP-seq, SELEX, and protein-binding microarray (PBM) experiments [70]. HOCOMOCO has expanded considerably since its initial release, incorporating improved quality control, confidence level annotations, and TF family classifications, and supports straightforward integration with external computational pipelines [70, 71]. Together, JASPAR and HOCOMOCO represent foundational resources for the study of TF binding dynamics and gene regulatory networks, and both are utilized in the motif selection pipeline described in the Methods section of this thesis.

1.7 Alternative Splicing

Pre-messenger RNA (pre-mRNA), the initial transcript produced from genomic DNA, is composed of protein-coding sequences called exons, intervening non-coding sequences called introns, and 5' and 3' untranslated regions (UTRs) [72]. Before the pre-mRNA can be translated into protein, it must undergo splicing, a process in which introns are excised and exons are joined together by a large ribonucleoprotein machinery known as the spliceosome. The resulting mature mRNA consists predominantly of exons and UTRs, though in certain cases introns may be retained in

the final transcript [73].

Splicing is not a rigid, deterministic process. Exons and introns can be removed and rejoined in different combinations from the same pre-mRNA, giving rise to multiple distinct transcript isoforms from a single gene, a process known as alternative splicing (AS). This process is the primary biological mechanism of interest in this thesis. As illustrated in Figure 1.6, a pre-mRNA composed of five exons can undergo different combinations of exon inclusion and exclusion to produce three distinct mature mRNA molecules, each potentially encoding a protein isoform with a unique structure and function. It is estimated that more than 95% of human genes containing multiple exons undergo AS, and that the approximately 20,000 protein-coding genes in the human genome collectively give rise to over 120,000 distinct transcripts, enormously amplifying the functional diversity of the human proteome [74].

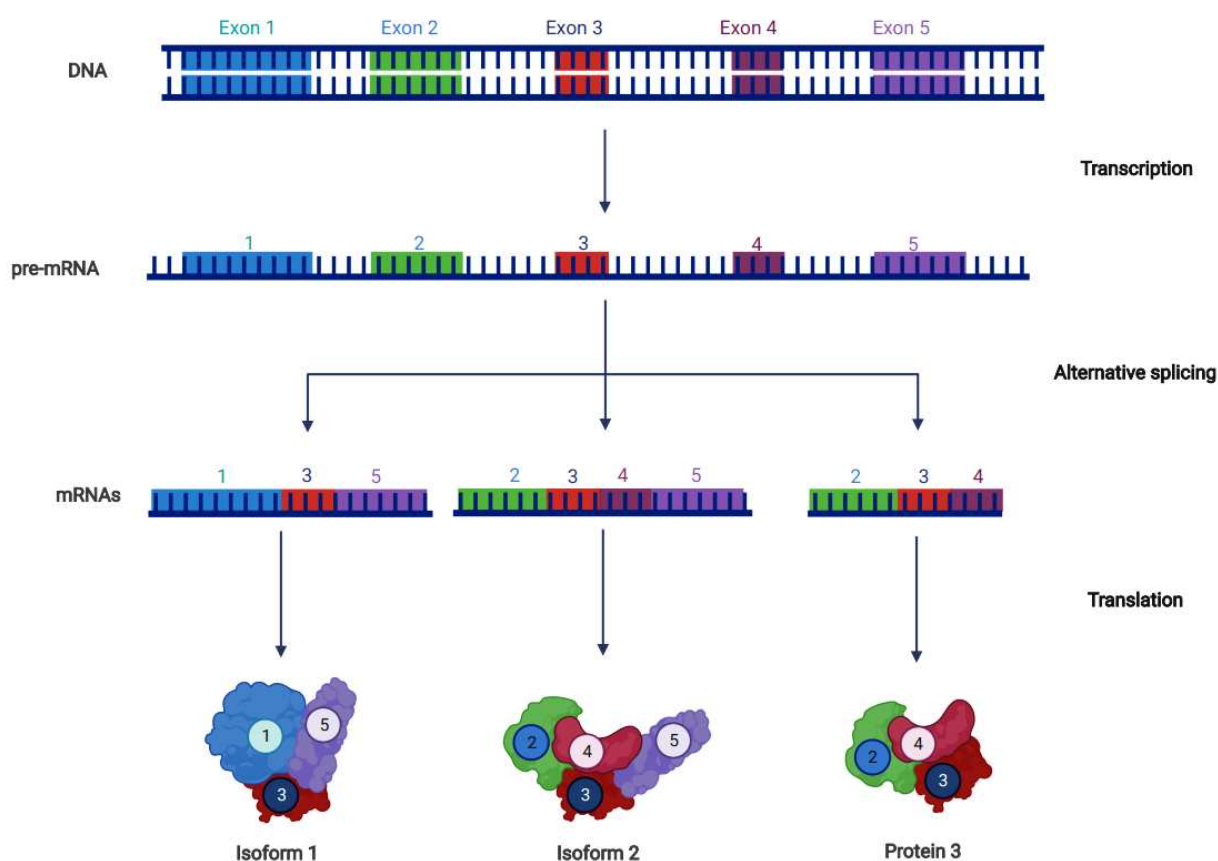


Figure 1.6 Schematic of alternative splicing. (Created with BioRender.com).

The fidelity and efficiency of AS are governed by a complex regulatory code embedded within the pre-mRNA sequence itself. Although introns do not encode

amino acids, they contain critical regulatory elements that control splicing factor binding. The most important of these is the branch point, a conserved adenosine residue located near the 3' end of the intron which base-pairs with the U2 small nuclear ribonucleoprotein (U2 snRNP) component of the spliceosome. This interaction initiates the formation of the lariat intermediate, enabling intron excision and ligation of the flanking exons [75]. Beyond the branch point, both exons and introns harbor auxiliary regulatory sequences known as splicing enhancers and silencers, which modulate the recruitment of splicing factors to promote or suppress the inclusion of adjacent exons.

Multiple mechanistic classes of AS events have been characterized, including cassette exon skipping, intron retention, and alternative selection of 3' or 5' splice sites, as well as more complex events such as mutually exclusive exon usage, where two exons are never co-included in the same transcript (Figure 1.7). In mammals, cassette exon skipping is the most prevalent and well-studied form of AS [76]. The protein variants produced from the same gene through AS are referred to as isoforms, and they frequently differ in their structure, function, interaction partners, and subcellular localization.

The functional significance of AS-generated isoforms remains an active area of investigation. While many alternatively spliced variants encode functionally distinct proteins, the extent to which all AS events produce biologically relevant products is debated [77]. Some alternatively spliced transcripts lack coding potential and are eliminated through the nonsense-mediated decay (NMD) pathway. Importantly, however, even these non-coding AS products may exert indirect regulatory effects, as AS coupled to NMD has been recognized as a conserved mechanism for controlling transcript abundance and fine-tuning gene expression levels [78]. The protein isoforms that do reach translation often exhibit substantial functional divergence from one another, a theme that is central to the present thesis and explored in depth in the following sections.

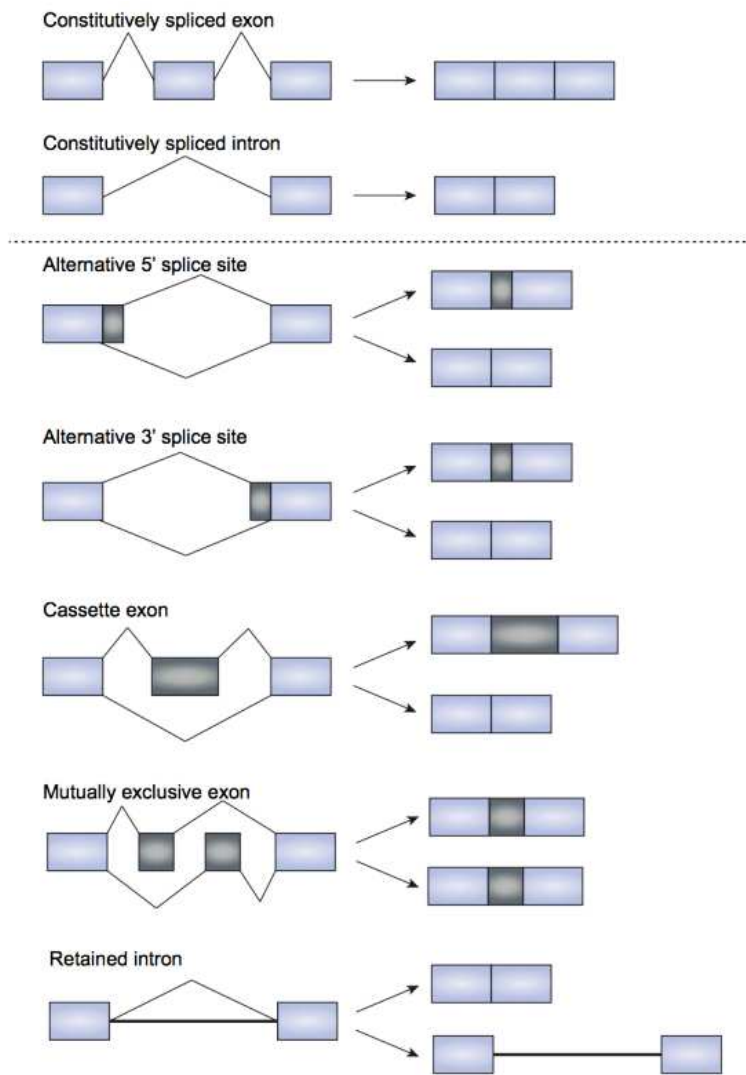


Figure 1.7 Types of commonly observed splicing events. (Taken from Blaustein et al., 2007).

1.8 Transcription Factor Isoforms

TF genes produce two main types of protein products: reference isoforms and alternative isoforms [79]. A reference isoform is defined as the canonical or representative protein product of a gene, typically selected from curated databases such as the MANE (Matched Annotation from NCBI and EMBL-EBI) transcript set, and serves as the standard against which other isoforms are compared. In contrast, alternative isoforms are additional protein variants produced from the same gene through mechanisms such as alternative splicing, alternative promoter usage, or differential exon inclusion. These alternative isoforms frequently differ in their amino acid sequence in ways that affect key functional regions, including DNA-binding domains (DBDs), effector domains, and IDRs. Through systematic comparison of

isoforms across TF genes, Lambourne et al. (2025) [79] demonstrate that the majority of alternative isoforms exhibit measurable functional divergence relative to their corresponding reference isoforms, encompassing differences in DNA-binding specificity, protein-protein interactions (PPIs), transcriptional activity, and subcellular localization. Collectively, these findings highlight that the reference isoform alone does not fully represent the functional capacity encoded by a TF gene, and that alternative isoforms contribute a significant and previously underappreciated layer of complexity to gene regulatory networks. Quantitatively, Lambourne et al. (2025) [79] demonstrated that 81% of alternative isoforms differed in protein–DNA interactions relative to their reference counterparts, underscoring the widespread functional divergence in DNA-binding behavior that motivates the structural modeling approach developed in this thesis.

The clinical relevance of TF isoform diversity is increasingly apparent. Nearly 60% of isoforms differentially expressed in tumors are alternative variants, and these frequently act as negative regulators of their cognate reference isoforms [80]. The transcription factor CREB1 provides a well-characterized example: while the reference isoform functions as an oncogene, an alternative isoform lacking key regulatory sequences may exert tumor-suppressive properties by acting as a dominant negative, and this alternative isoform is significantly downregulated in breast tumors relative to the reference [79, 81]. A complementary pattern is observed for SMAD3, a known tumor suppressor whose reference isoform is significantly downregulated in lung adenocarcinoma (LUAD). In contrast, an alternative isoform of SMAD3 that has lost part of its DBD is upregulated in the same tumors, suggesting that this variant may acquire oncogenic properties by disrupting normal SMAD3-mediated transcriptional repression [79, 82]. Together, these examples illustrate that the balance between reference and alternative TF isoforms can have profound consequences for disease progression, underscoring the importance of characterizing isoform-specific functions at the molecular level.

1.8.1 Two major classes of alternative isoforms: Negative regulators and rewirers

While the existence of alternative TF isoforms has long been recognized, their functional classification has only recently been systematically addressed. Historically, the best-characterized alternative isoforms were those acting as negative regulators variants that interfere with the activity of their cognate reference isoforms rather than executing independent regulatory programs [80, 83]. A well-studied example is

STAT3 β , which competes with the oncogenic STAT3 α isoform for DNA binding but lacks the C-terminal transactivation domain required to activate transcription. By occupying the same genomic targets as STAT3 α without driving target gene expression, STAT3 β effectively functions as a tumor suppressor through a dominant-negative mechanism [84-86]. More broadly, negative regulators are characterized by their ability to bind the same genomic targets as the reference isoform while failing to recruit key cofactors or activate transcription, thereby titrating out the reference isoform and attenuating its regulatory output.

However, the extent to which negative regulators are representative of TF isoform biology more globally remained unclear until recently [79]. Through multi-dimensional characterization of hundreds of TF isoforms encompassing DNA binding, protein–protein interactions, transcriptional activity, and subcellular localization Lambourne et al. (2025) [79] reveal that alternative isoforms fall into two primary functional categories. The first category, negative regulators, comprises isoforms that act in competition with or dependent on their reference counterparts, modulating rather than redirecting regulatory output. The second and more prevalent category, rewirers, comprises isoforms that behave in a fundamentally distinct manner from their cognate reference isoforms engaging different genomic targets, interacting with different protein partners, or driving entirely different transcriptional programs. Together, these two categories demonstrate that TF loci encode a far richer repertoire of regulatory functions than previously appreciated, with profound implications for our understanding of gene regulatory network complexity and disease [79].

1.8.2 TF- DNA interaction

TF–DNA interaction is a fundamental mechanism underlying gene regulation, in which TFs bind to specific DNA sequences through their specialized DBDs to control transcriptional activity. However, as demonstrated by Lambourne et al. (2025), these interactions are not uniform across all protein isoforms encoded by a TF gene. While reference isoforms typically represent the canonical DNA-binding behavior of a given TF, alternative isoforms can exhibit substantial variation in both their DNA-binding capacity and sequence specificity. Such variation can manifest as a complete loss of DNA binding, a gain of novel binding interactions, or a quantitative shift in binding affinity at existing target sites each with distinct consequences for downstream gene expression. Critically, TF–DNA interactions cannot be fully predicted from DBD sequence alone and are significantly shaped by isoform-specific sequence context,

underscoring the importance of characterizing alternative isoforms when modeling gene regulatory networks [79].

1.8.3 DNA Binding Is Influenced by Sequences Both Inside and Outside the DBD

While TF–DNA binding has traditionally been attributed primarily to the structured DBD, systematic analysis of protein–DNA interaction (PDI) profiles across alternative isoforms reveals that sequences both within and far outside the DBD can dictate binding outcomes [79].

Variations within the DNA-binding domain

As expected, alternative isoforms that lack a DBD entirely fail to bind DNA in enhanced yeast one-hybrid (eY1H) assays. Since all tested isoforms are fused to a nuclear localization signal (NLS) during the assay, this loss of interaction can be confidently attributed to structural deficiency rather than cytoplasmic mislocalization. Consistent with this, most isoforms retaining only a partial DBD also lose DNA-binding capability, though rare exceptions exist for example, two ZIC3 isoforms remain functional despite losing three amino acids from a zinc-finger array, suggesting that certain DBD truncations are better tolerated than others [79].

Small insertions within the DBD can be equally disruptive. A four amino acid insertion in the basic helix-loop-helix (bHLH) loop region of a HEY1 alternative isoform, for instance, results in a complete loss of PDI. This finding is consistent with structural data from the related TF MLX, in which the bHLH loop region is essential for dimer stability confirming that even minor sequence perturbations within the DBD can profoundly disrupt global TF–DNA recognition [87-89].

1.8.4 The Influence of Flanking and Distal Regions on DNA Binding

A striking finding from the systematic isoform analysis of Lambourne et al. (2025) [79] is that the majority of alternative isoforms 67% (42/63) retain a complete and unperturbed DBD. Despite this, only 19% of these isoforms shared identical DNA-binding profiles with their reference counterparts, while 21% actually gained novel binding interactions. These observations demonstrate that DBD integrity alone is insufficient to predict binding behavior. Differences in sequences immediately flanking the DBD frequently account for these shifts, highlighting the role of adjacent regions in stabilizing or orienting the domain for productive DNA contact [90, 91].

Even more striking are the effects of sequences located far from the DBD, particularly within IDRs. Analysis reveals that 65% of isoforms harboring sequence variations more than 100 amino acids away from the DBD exhibit altered DNA-binding profiles, with the majority of these changes localized to disordered segments [79]. A compelling illustration of this phenomenon is provided by CREB1, an alternative isoform of which features a 14 amino acid insertion located 165 residues from its bZIP DBD. Despite maintaining transcriptional activity confirming that the protein is properly expressed and folded this isoform loses binding across several DNA sequences. Protein-binding microarray (PBM) analysis reveals that the isoform exhibits subtly reduced DNA affinity relative to the reference [87, 92], demonstrating that distal disordered regions can fine-tune binding affinity sufficiently to drop interactions below detectable thresholds in certain assays.

1.8.5 The Autoinhibitory Model: Evidence from DLX4 and PKNOX1

To investigate how isoforms sharing identical DBDs can nonetheless produce divergent binding outcomes, Lambourne et al. (2025) [79] examined the TFs DLX4 and PKNOX1 in detail. In both cases, reference isoforms displayed highly specific DNA-binding profiles, whereas alternative isoforms lacking large N-terminal IDRs exhibited a markedly broader and less sequence-specific binding pattern. This finding supports an autoinhibitory model in which the IDR functions as a tethered inhibitor physically engaging the DBD through intramolecular interactions to selectively mask non-specific binding while permitting high-affinity, sequence-specific contacts [93-95]. Computational mean-field interaction strength calculations confirmed strong intramolecular contacts between the IDR and the homeodomain in the reference isoforms of both DLX4 and PKNOX1, interactions that were significantly attenuated in the corresponding alternative isoforms. This provides a concrete physical mechanism by which sequence changes outside the DBD can effectively rewire a TF's DNA-binding specificity without altering the domain responsible for DNA contact [79]. Lambourne et al. (2025) explicitly characterized this as a plausible and testable physical model rather than a confirmed mechanism, a description that directly motivates the computational structural validation approach undertaken in this thesis.

1.8.6 Developmental Implications: The TBX5 Model

The functional importance of IDR-driven modulation of DNA binding is further illustrated by TBX5, a critical regulator of cardiac development [96]. The reference

isoform TBX5-1 and two alternative isoforms TBX5-2 and TBX5-3 differ primarily in their N-terminal or C-terminal disordered regions, yet their DNA-binding profiles diverge considerably. TBX5-3, which carries modifications in its C-terminal region, loses approximately half of its DNA-binding targets relative to the reference. TBX5-2, which differs in its N-terminal region, retains canonical binding but exhibits significantly enhanced affinity for a subset of moderate-affinity binding motifs. Notably, ChIP-seq analysis in cardiomyocytes reveals that these moderate-affinity sites are selectively enriched *in vivo*, suggesting that TBX5-2 may fulfill a distinct and specialized regulatory role during cardiac development that is not captured by the reference isoform alone [79].

Collectively, the findings from CREB1, DLX4, PKNX1, and TBX5 converge on a unifying principle: TF IDRs are not passive structural linkers but active modulators of DNA-binding specificity and affinity. This makes it insufficient and potentially misleading to predict TF regulatory behavior from DBD sequence alone, and establishes a compelling rationale for the isoform-resolved structural modeling approach developed in this thesis.

1.9 Computational prediction of protein-DNA interactions

Recent advances in deep learning-based protein structure prediction have opened unprecedented opportunities to study biomolecular interactions at atomic resolution and at a scale previously unattainable through experimental methods alone [97]. To systematically investigate the structural basis of functional divergence between reference and alternative TF isoforms, this thesis employs a comparative structural modeling pipeline built around three state-of-the-art deep learning frameworks: AlphaFold 3 (AF3), RoseTTAFold2NA (RF2NA), and Boltz-2. Each of these tools brings distinct architectural strengths and modeling philosophies to the analysis, and by integrating their complementary outputs, this research moves beyond sequence-level analysis to provide a high-resolution, three-dimensional perspective on how alternative splicing reshapes the TF-DNA interface and the functional consequences that follow.

1.9.1 RF2NA: Generalized Prediction of Protein-Nucleic Acid Systems

RF2NA represents a significant architectural advance over previous structure prediction models by extending the RoseTTAFold framework to handle the complexities of protein-nucleic acid interactions [98]. Conventional approaches required proteins and nucleic acids to be modeled independently before being

computationally docked together, a sequential strategy that frequently failed to capture the mutual conformational influence that proteins and DNA exert on one another during binding. RF2NA overcomes this limitation through an integrated multi-track architecture that simultaneously models folding and binding from raw sequence information alone, allowing the structural consequences of protein-DNA interaction to emerge naturally from the prediction process. Trained on an augmented dataset encompassing all available RNA and DNA complex structures in the Protein Data Bank, the model learns the complex parameters necessary to capture subtle structural features, such as protein-induced DNA bending and base-specific recognition geometry. This end-to-end approach yields high-resolution structural models accompanied by integrated confidence estimates, making RF2NA a robust tool for characterizing protein–nucleic acid interactions at atomic resolution [98].

1.9.2 AF3: A Unified Framework for All-Atom Structural Prediction

AF3 was developed to provide high-accuracy structural predictions across virtually all molecular types represented in the Protein Data Bank, extending the scope of deep learning-based structure prediction well beyond proteins [99]. While earlier models were typically specialized for particular interaction types and sometimes struggled to match the accuracy of physics-based docking methods, AF3 functions as a generalist framework capable of modeling proteins, nucleic acids, small molecules, ions, and covalently modified residues within a single unified system. This versatility is achieved through a substantial architectural evolution: the computationally intensive Evoformer module of AlphaFold 2 is replaced by a more streamlined Pairformer trunk, and structure generation is delegated to a diffusion-based module that directly predicts raw atomic coordinates rather than relying on residue-specific torsion angle frames. This coordinate-based generative approach allows the network to accommodate arbitrary chemical entities including ligands, metal ions, and modified nucleotides without requiring molecule-specific handling of bonding geometry. The result is a significant advance in prediction accuracy across diverse molecular interaction types, establishing AF3 as a powerful and comprehensive tool for studying the molecular machinery of the cell [99].

1.9.3 Boltz-2: Assessing Interaction Stability and Binding Affinity

Boltz-2 is an open-source all-atom foundation model that extends the capabilities of its predecessors by integrating structural prediction with quantitative binding affinity estimation within a single unified framework [100]. While sharing the diffusion-based

architectural principles of AF3, Boltz-2 distinguishes itself through a specialized affinity module that predicts not only the three-dimensional coordinates of a biomolecular complex but also the thermodynamic strength of the interaction a quantity directly relevant to understanding the functional consequences of isoform-specific sequence changes. This integration of structure and affinity prediction is achieved through specialized training on millions of curated biochemical measurements, enabling the model to overcome the longstanding trade-off between predictive performance and computational efficiency. In the context of this thesis, Boltz-2 provides a critical functional dimension to the structural pipeline: by estimating the relative binding affinities of reference and alternative TF isoforms for their cognate DNA sequences, it enables the distinction between isoforms that structurally resemble the reference but bind with altered potency and those that engage DNA in a fundamentally different manner [101].

1.10 Aims of the research

The precise interaction between transcription factors and their genomic binding sites is fundamental to the coordination of virtually all biological processes, from embryonic development to homeostatic maintenance in adult tissues. While the existence of alternative TF isoforms generated through alternative splicing is well documented, the structural basis for their functional divergence from reference isoforms remains poorly understood. Traditional models of gene regulation have largely overlooked these alternative proteoforms, despite growing evidence that they play critical roles in both normal development and the progression of diseases such as cancer. This thesis addresses this gap by employing a comparative structural biology approach to investigate how splicing-induced sequence changes reshape the protein–DNA interface and alter the regulatory output of TF isoforms at the atomic level.

The central aim of this research is to evaluate the structural and functional differences between reference and alternative TF isoforms using a synergistic computational pipeline that integrates three state-of-the-art deep learning frameworks: AF3, RF2NA, and Boltz-2. By combining the complementary strengths of these tools, this study seeks to provide a high-resolution, three-dimensional mechanistic explanation for how alternative splicing alters TF–DNA interactions and, by extension, gene regulatory networks.

To achieve this central aim, four specific objectives were pursued:

High-Resolution Structural Comparison: Three-dimensional structural models of reference and alternative TF isoforms in complex with their DNA motif sequences were generated and systematically compared. The objective was to determine whether alternative splicing alters the binding geometry, the orientation of the DBD relative to the DNA, or the overall physical stability of the protein-DNA complex.

To ensure the reliability of the predicted structural models, standardized confidence metrics were rigorously applied across all three modeling platforms. These included the predicted local distance difference test (pLDDT) to assess local folding accuracy, the predicted aligned error (PAE) to evaluate the relative positioning of protein and DNA chains, and the interface predicted template modeling score (ipTM) to quantify confidence in the predicted protein-DNA interface.

Moving beyond static structural comparison, the binding affinity prediction capabilities of Boltz-2 and AF3 were leveraged to identify cases in which an alternative isoform adopts a structure resembling the reference but exhibits quantitatively altered binding strength changes that may fundamentally shift its biological impact without being apparent from structural superposition alone.

The influence of sequence changes located within IDRs on DNA-binding specificity and affinity was systematically examined. A central focus was to evaluate the autoinhibitory model of TF regulation, in which IDRs act as tethered intramolecular inhibitors that modulate DBD accessibility and binding selectivity in an isoform-dependent manner.

Collectively, these objectives are underpinned by the hypothesis that integrating predictions from three architecturally distinct computational modeling platforms can yield a robust consensus view of how isoform-specific sequence changes reshape protein-DNA interactions providing mechanistic insights that neither experimental nor single-model computational approaches could achieve alone.

This study was conducted at the Network and Structural Bioinformatics (NeStOme) Lab in Hamburg, Germany.

2 Material and Methods

2.1 Data Collection and Processing

In this study, TF isoform datasets and molecular interaction data were retrieved from the systematic resource established by [57]. The primary dataset comprises 693 proteins generated from 246 TF genes, which were partitioned into reference (Ref; $n = 246$) and alternative (Alt; $n = 447$) isoforms according to the classification framework defined in the source study. To evaluate the splicing variation, we characterized the Alternative isoforms based on the degree of DBD loss. Among 447 Alternative isoforms, 56.4% ($n = 252$) retained an intact DBD (0% loss), while 38.0% ($n = 170$) exhibited partial or complete domain truncation ($>0\%$ loss), with a specific subset of 83 isoforms showing total (100%) DBD removal; a small cohort of 5.6% ($n = 25$) remained unclassified due to insufficient domain annotation.

Out of the initial 246 TFs, sequence motif data were successfully retrieved for 208 TFs (84.5%) for structural modeling. These motifs were integrated from three primary repositories: JASPAR 2024 CORE non-redundant vertebrate profiles, HOCOMOCO v11 human core PWMs, and a species-specific export from CIS-BP accessed in August 2025. The remaining 38 genes were excluded due to a lack of experimentally validated Position Weight Matrices (PWMs) in the reference databases mentioned above.

2.2 Mathematical Framework for Information Content

The specificity of each motif was quantified using the Information Content (IC) metric, established by [102] and reviewed by Stormo (2013) [103]. The IC at each position i is defined as the relative entropy (Kullback-Leibler divergence) between the observed base frequencies $F(b, i)$ and the background probability $P(b)$:

$$IC(i) = 2 + \sum_b F(b, i) \log_2 F(b, i) = \sum_b F(b, i) \log_2 \frac{F(b, i)}{0.25}$$

We applied this theoretical framework using an equiprobable genomic background where $P(b) = 0.25$ for all nucleotide [15]. To determine the overall quality of a motif model of length L , the total IC (IC_{total}) for each motif was calculated by summing the IC values across all individual positions i :

$$IC_{total} = \sum_{i=1}^L IC(i) = \sum_{i=1}^L \sum_{b \in \{A, T, C, G\}} p_{b,i} \log_2 \left(\frac{p_{b,i}}{0.25} \right)$$

2.2.1 Handling Multiple Motifs and Selection Logic

Individual TFs often have multiple motifs reported across different databases due to varying experimental platforms (e.g., SELEX vs. ChIP-seq).

To ensure selection of the most biologically representative motif model, our pipeline applied a hierarchical prioritization logic based on four sequential criteria. First, a repository priority order was established, favoring JASPAR over HOCOMOCO over CIS-BP, to prioritize manually curated, high-quality, non-redundant profiles. Second, when multiple motifs for the same TF were available within a preferred repository, the motif with the highest total information content was selected, as higher IC values indicate stronger positional nucleotide preference and greater sequence specificity [104, 105]. Third, if multiple motifs had identical total information content, the motif with the greater width was selected. Finally, if both information content and width were identical, the motif with the most recent or highest-version identifier was selected as the final representative model.

As described in Lambourne et al. (2025), the original eY1H binding matrix (Table S3) represented 171 unique protein clones corresponding to 80 genes, tested against 186 DNA baits, resulting in 23,569 protein-bait combinations with a validated interaction (878 true and 22,691 false interactions). The remaining 8,237 protein-bait combinations were excluded as "empty" due to unsuccessful evaluation. Of the 23,569 protein-bait combinations, 1,489 were part of a discordant protein-bait combination pair. We define a pair of protein-bait combinations as discordant if and only if the two combinations show opposite interactions (i.e., one is true, while the other is false or NA, interaction) and the two proteins come from the same gene. In total, we found 486 such discordant protein-bait combination pairs.

DNA baits were derived from 330 human enhancer and promoter sequences selected from the VISTA Enhancer Browser, representing characterized developmental regulatory elements and systematically screened against the Open Reading Frames (ORFs) of all tested transcription factors. The bait sequences (61 to 2,090 nucleotides long) represent full-length enhancer fragments suitable for screening assays but introduce substantial sequence context beyond the core binding site. Therefore, to focus the analysis on intrinsic DNA-recognition determinants and ensure clearer structural interpretation, motif sequences representing defined binding regions were used instead. We then determined the overlap using string matching to identify 208 motifs within the 186 baits for an exact match, where 30 DNA baits showed overlap

with 56 DNA motifs. Considering these 30 DNA baits, of the 486 discordant protein-bait combination pairs, 56 remain.

From this filtered subset and based on “Table_S13” in Lambourne et al. (2025), we extracted 31 pairs (with 8 unique reference-alternative isoform pairs) with zero DBD loss for alternative isoforms and 25 pairs (with 6 unique reference-alternative isoform pairs) with different percentages in DBD loss.

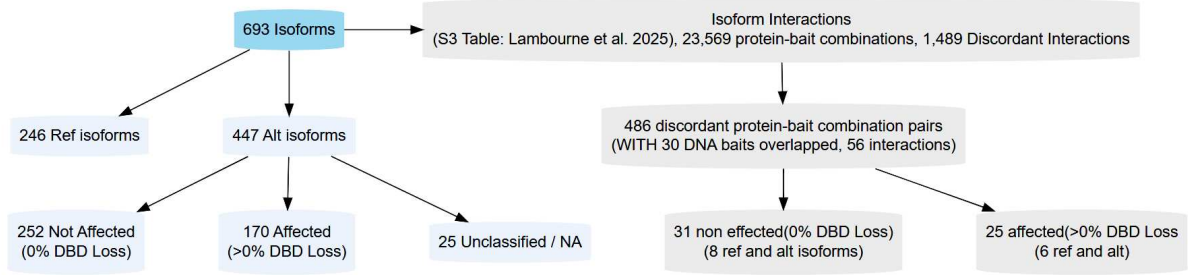


Figure 2.1 Isoform classification flowchart.

2.3 Statistical Analysis

Pairwise comparisons between model architectures were conducted using a two-sided paired Wilcoxon signed-rank test. Structural performance scores were computed for 208 curated TF–DNA motif inputs. However, due to incomplete overlap in successfully generated predictions across modeling methods, the number of paired observations varied by comparison. Specifically, AF3 and RF2NA were compared across all shared reference complexes ($n = 201$), whereas comparisons involving Boltz-2 were restricted to the subset of complexes for which Boltz-2 predictions were available ($n = 198$). Because multiple pairwise comparisons were performed across evaluation metrics, p-values were adjusted using the Benjamini–Hochberg (BH) procedure to control the false discovery rate (FDR). Statistical significance was defined at an adjusted α level of 0.05.

2.4 Structural Modeling and Predictive Framework

Structural predictions were generated for all 693 isoforms in the dataset, comprising 246 reference and 447 alternative isoforms, each paired with its corresponding DNA motif sequence as described in section 2.1. To model the three-dimensional structure of TF–DNA complexes across this full isoform library, we employed three independently developed deep learning platforms: AF3 [99], RF2NA [98] and Boltz-2 [100]. Each platform was selected because it represents a distinct computational approach to

protein–nucleic acid complex prediction, and using all three in parallel allowed us to assess whether structural findings were consistent across different modeling frameworks or specific to a single platform. Predicting protein–nucleic acid complexes is particularly challenging when close structural homologs are absent from training data (Baek et al., 2024), and a multi-platform approach provides a more robust basis for interpretation than relying on any single tool. The 56 discordant TF–DNA interaction pairs identified in section 2.2.1 were subsequently used as a specific experimental reference subset for the predictive accuracy evaluation described in sections 2.5 and 3.4.

2.4.1 Architectural Framework of RF2NA

RF2NA was developed as an extension of the RoseTTAFold framework, adapted to handle the structural prediction of protein–nucleic acid complexes within a single end-to-end network [98]. The core of the model is a three-track architecture consisting of 36 iterative processing layers followed by four specialized refinement layers, with a total of 67 million parameters. The three tracks process the biomolecular system through three parallel representations that continuously exchange information with one another throughout the prediction (Figure 2.2).

The first track operates at the sequence level and was expanded from the original 22 protein tokens to 32 tokens, with the ten additional tokens dedicated to representing DNA nucleotides, RNA nucleotides, and unknown or masked positions for both nucleic acid types. The second track operates at the level of residue pairs and was adapted from its original purpose of modeling protein–protein distances to capture geometric relationships between all molecular pairs in the complex, including amino acid–base and base–base contacts. The third track handles the three-dimensional structural representation, where proteins are described through backbone frames and chi angles while nucleotides are represented using a specialized coordinate frame centered on the phosphate group and defined by ten torsion angles spanning the backbone, ribose ring, and nucleoside geometry. These torsion angles are used to construct the complete full-atom representation of the nucleic acid chain.

The prediction process begins by generating multiple sequence alignments for both the protein and nucleic acid components. These alignments seed the three-track layers, which iteratively update all three representations in parallel. Because protein–nucleic acid complex structures are considerably less represented in the Protein Data Bank than protein-only structures, the model supplements the learned representations with

physical energy terms including Lennard-Jones potentials and hydrogen-bonding contributions in its final refinement layers. The resulting predicted frames and torsion angles are then used by an SE(3) transformer and sidechain generation module to place every atom of the protein and nucleic acid into the final three-dimensional model.

In independent validation, RF2NA achieved a mean IDDT of 0.73, with approximately 45% of high-confidence models exceeding an IDDT of 0.8 [98]. RF2NA does not output an ipTM score, and this metric was therefore not available for RF2NA predictions throughout the analyses presented in this study.

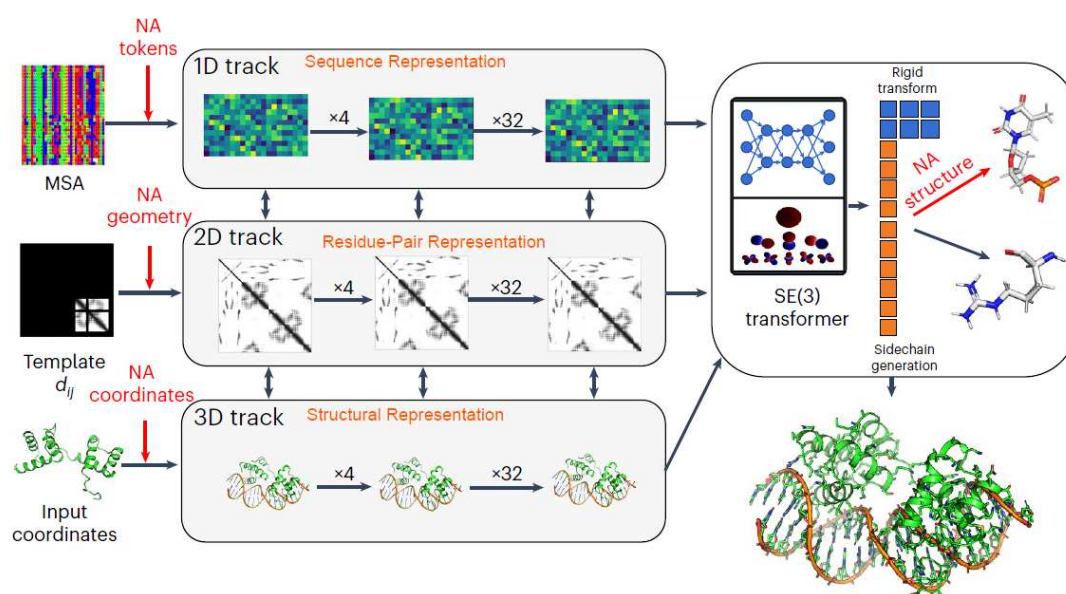


Figure 2.2 Architectural Overview of RFNA. (Taken from Baek et al., 2024)

2.4.2 AF3: A Unified All-Atom Architecture

AF3 was designed as a generalist structural prediction framework capable of modeling virtually any biomolecular system, including proteins, nucleic acids, small molecules, ions, and covalently modified residues, within a single unified pipeline [99]. This generalist scope distinguishes it from its predecessor AlphaFold 2, which was developed specifically for protein structure prediction (Figure 2.3).

The AF3 inference pipeline begins by embedding input sequences, ligand structures, and covalent bond information into initial single and pair representations through a dedicated input embedder. These representations are then refined by the Pairformer module, which serves as the central processing engine of the model and focuses on updating the pairwise representation that encodes the spatial relationships between all atoms in the system. Structural convergence is promoted through a recycling mechanism that feeds updated representations back through the Pairformer for multiple

iterations before passing them to the structure generation stage. Structure generation in AF3 is handled by a diffusion-based module that operates by beginning from a random distribution of atomic coordinates and iteratively refining them toward a physically plausible three-dimensional arrangement. This approach differs fundamentally from torsion angle-based prediction methods and allows the model to generate structures for arbitrary chemical entities without requiring specialized geometric handling for each molecular type. The reliability of the final predicted structure is assessed by a dedicated confidence module that produces pLDDT and PAE scores for each predicted complex.

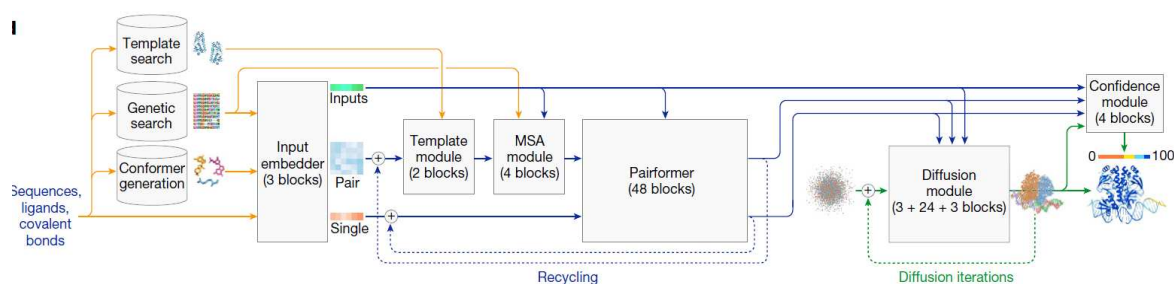


Figure 2.3 AF3 Inference Architecture. (Taken from Abramson et al., 2024).

2.4.3 Boltz-2: Integrated Structure and Affinity Prediction

Boltz-2 is an open-source all-atom structural prediction model that extends the capabilities of diffusion-based approaches by incorporating quantitative binding affinity prediction within the same computational framework [100]. While the core structural prediction mechanism of Boltz-2 shares the diffusion-based principles of AF3, the model introduces a dedicated affinity module that estimates the thermodynamic strength of the predicted biomolecular interaction alongside the three-dimensional structural coordinates (Figure 2.4).

The model processes its inputs through four sequential functional components. The trunk stage begins with an atom attention encoder that ingests the input sequences and any conditioning information, which are then passed through template, multiple sequence alignment, and Pairformer processing steps with an internal recycling mechanism to progressively refine the molecular representations. The denoising module then generates the final three-dimensional structure by applying reverse diffusion to transform an initial random distribution of atomic positions into a coherent structural model, with physics-based energy terms and conditioning signals incorporated during this process to guide the generation toward energetically plausible configurations. Once the structure is generated, the confidence module evaluates its

reliability using a Pairformer-based assessment and produces standardized confidence metrics for each predicted complex. The affinity module, which has no equivalent in AF3 or RF2NA, processes the refined molecular representations to produce a binding likelihood estimate and a quantitative affinity value. In this study, the affinity module output provided an additional functional dimension to the structural analysis by offering an estimate of how the binding strength between a TF isoform and its cognate DNA sequence differs between reference and alternative forms.

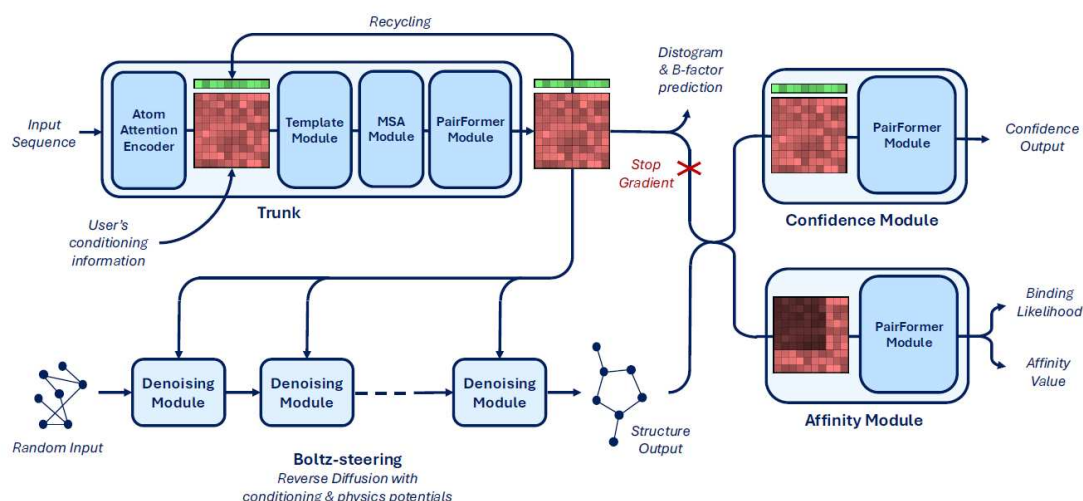


Figure 2.4 Boltz-2 Model Architecture. (Taken from Passaro et al., 2025).

2.4.4 Model Quality and Validation Metrics

Three confidence metrics were consistently applied across all three platforms to evaluate the structural quality and biological relevance of the predicted TF–DNA complexes. Before computing these metrics, the protein–DNA interface was defined geometrically by identifying all residues for which the distance between the C-alpha atom of that residue and the C-alpha atom of any residue in the opposing chain was less than or equal to 8 Å. This criterion provides a consistent and residue-type-independent definition of the interface across all complexes in the dataset.

The interface predicted template modelling score, ipTM, is a dimensionless score between 0 and 1 that quantifies the predicted accuracy of the relative positioning between the protein and DNA chains at the interface. Because it evaluates the interface specifically rather than the overall complex, ipTM is the most direct measure of predicted binding geometry among the three metrics. Following the criteria established by Helsem et al. [97], predictions were considered reliable when ipTM reached or

exceeded 0.6. This metric is produced by AF3 and Boltz-2 but is not available for RF2NA (Table 2.1).

The predicted aligned error, PAE, is reported in Ångströms and quantifies the expected positional error between any pair of residues in the predicted complex. Lower values indicate greater confidence in the overall spatial arrangement of the protein and DNA chains relative to one another. PAE is produced by all three platforms and was used as the primary reliability criterion for RF2NA, for which a mean inter-chain PAE below 10 Å was required for a prediction to be considered structurally acceptable, again following Helsem et al. [97].

The predicted local distance difference test score, pLDDT, measures per-residue confidence in local structural accuracy. AF3 and Boltz-2 report this score on a scale from 0 to 100, while RF2NA reports it on a scale from 0 to 1. To allow consistent comparison across platforms, all pLDDT values from AF3 and Boltz-2 were divided by 100 prior to analysis, bringing them onto the same 0 to 1 scale as RF2NA. Values near 1.0 reflect high confidence in a well-defined structured region, while values below 0.5 are generally associated with intrinsically disordered segments that do not adopt a stable conformation.

Metrics	Score/test description	Range	Unit	Applied and reported	Relevant platform(s)
pLDDT	Predicted local distance difference test	0-1 Very poor-very high confidence	–	Significant differences between models	AF3 Boltz-2 RF2NA
PAE	Predicted aligned error	Continuous	Å	< 10 considered successful in RF2NA	AF3 Boltz-2 RF2NA
ipTM	Predicted interface template modeling score	0-1 Poor-excellent	–	≥ 0.6 considered successful	AF3 Boltz-2

Table 2.1 Confidence metrics used to evaluate structural predictions. (Taken from Helsem et al., 2025)

2.5 Receiver operating characteristic (ROC) Curve Analysis

To assess how well the structural confidence metrics generated by each platform predict experimental protein–DNA binding outcomes, receiver operating characteristic curve analysis was performed across all three platforms and all three metrics.

2.5.1 Dataset Preparation and Deduplication

In the experimental dataset, each protein isoform was tested against multiple DNA

baits, resulting in repeated entries for the same isoform across different bait combinations. Because structural confidence scores are properties of the isoform itself rather than of any specific bait interaction, including these repeated entries would have inflated the effective sample size and biased the ROC analysis. The dataset was therefore reduced to 79 unique isoforms by retaining one entry per isoform. Each isoform was classified as a positive case if binding to at least one bait was observed, and as a negative case only if no binding was detected across any of the baits tested. This classification strategy ensures that each data point in the analysis represents a single independent biological observation.

ROC curves were computed for each confidence metric independently using the pROC package in R. For pLDDT and ipTM, where higher values are expected to indicate greater binding competency, the ROC direction was set accordingly. For PAE, where lower values reflect better structural quality and are therefore expected to be associated with binding, the direction was reversed. Area under the curve values were calculated for all available metrics across the three platforms. Because RF2NA does not produce an ipTM score, this metric was evaluated for AF3 and Boltz-2 only.

2.5.2 Logistic Regression Model

To determine whether combining structural metrics from multiple platforms improves the prediction of experimental binding outcomes, a multivariable binary logistic regression model was constructed using all available confidence scores as predictors. These included ipTM, pLDDT, and PAE from AF3 and Boltz-2, and pLDDT and PAE from RF2NA. The model was fitted in R using the *glm* function with a binomial link function, and predicted probabilities from the fitted model were used to generate a combined ROC curve. This combined curve is displayed alongside the individual metric curves in all three ROC panels to provide a consistent reference for comparison. The significance of each platform's independent contribution to the combined model was evaluated from the coefficient p-values. In addition, three platform-specific logistic regression models were fitted separately, one for each platform, using only the metrics available from that platform, to isolate the independent predictive contribution of each and reduce the collinearity introduced by overlapping metric information across platforms.

2.5.3 Correlation Analysis

To examine the linear and rank-based relationships between each structural confidence metric and experimental binding outcome, Pearson and Spearman correlation

coefficients were computed between each metric and the binary binding status variable across the 79 unique isoforms. The resulting coefficients were organized into three-by-three matrices, one for each correlation method, and visualized as heatmaps using the heatmap package in R. Diverging color scales were applied to represent both the direction and magnitude of each correlation. The Pearson and Spearman heatmaps were displayed side by side to allow direct comparison of linear and monotonic relationships across all platforms and metrics. All figures were produced using ggplot2 and saved at 300 dpi.

3 Results

3.1 Performance comparison of RF2NA, AF3, and Boltz-2 on protein-DNA complex predictions for reference TFs

To predict the three-dimensional structures of TF–DNA motif complexes, we modeled the interactions of 246 reference TFs and their corresponding DNA motifs using three complementary deep learning frameworks: AlphaFold 3 (AF3), Boltz-2, and RoseTTAFold2NA (RF2NA). We evaluated prediction reliability using the Predicted Local Distance Difference Test (pLDDT) to measure confidence in local structural accuracy, the Predicted Aligned Error (PAE) to assess confidence in the relative positioning of TF and DNA chains, and the interface predicted template modelling score (ipTM) to estimate confidence in the predicted TF–DNA binding interface. We selected these three methods because they represent complementary architectural approaches: AF3 is based on a transformer-driven multimodal architecture optimized for complex prediction, RF2NA extends the three-track architecture to protein–nucleic acid systems by integrating sequence, pairwise, and structural representations informed by evolutionary information, and Boltz-2 uses a diffusion-based generative approach designed for iterative structural refinement. By comparing these models, we aimed to assess how differences in architectural design influence TF–DNA complex prediction performance.

Across reference TF–DNA complexes, the three modeling frameworks showed distinct performance patterns (Figure 3.1). For local structural confidence (Figure 3.1A), RF2NA displayed the most consistent pLDDT distribution, largely centered between 0.6 and 0.8. AF3 showed comparable but slightly broader variability, whereas Boltz-2 exhibited a wider spread toward lower pLDDT values, indicating less uniform local backbone confidence across complexes.

In terms of global complex reliability (Figure 3.1B), Boltz-2 achieved significantly lower mean PAE values compared to AF3 and RF2NA, suggesting improved confidence in the overall protein–DNA spatial arrangement. AF3 and RF2NA showed higher and more dispersed PAE distributions. Interface-specific confidence (Figure 3.1C) revealed the clearest distinction between platforms: Boltz-2 predictions were tightly clustered near ipTM approximately 0.95, whereas AF3 scores spanned a broad range across the confidence spectrum. This indicates substantially greater consistency of interface modeling with Boltz-2.

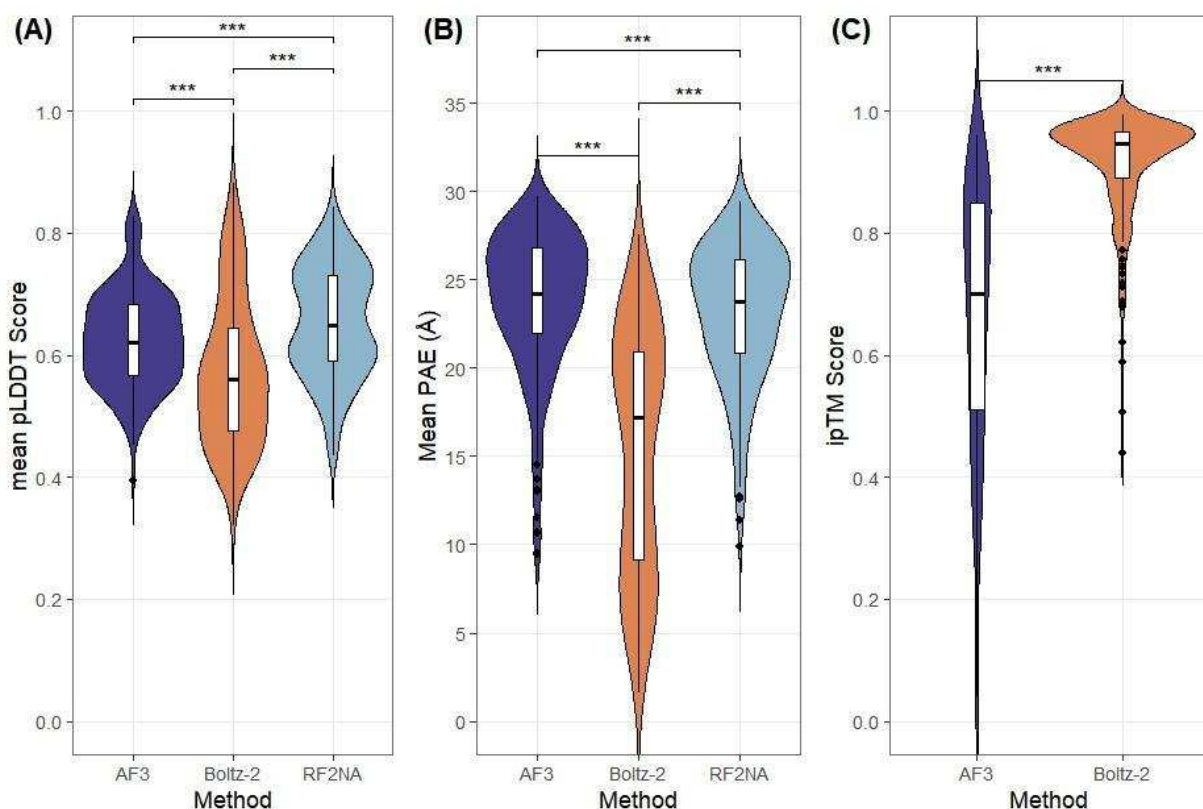


Figure 3.1 Comparative evaluation of structural prediction methods for reference TF-DNA complexes.

3.1.1 Assessment of Model Reliability and Success Thresholds

Having established the comparative performance of RF2NA, AF3, and Boltz-2 across reference TF-DNA complexes, we next applied a standardized threshold-based filtering protocol to determine which individual predictions met a minimum standard of biological reliability. This filtering was applied to AF3 and RF2NA predictions only, following the benchmarking criteria established by Helsem et al. [97]. Boltz-2 was included in all structural analyses presented in this thesis; however, no equivalent community-validated confidence threshold for Boltz-2 in the context of protein-DNA interaction prediction has been established in the literature, and threshold-based filtering was therefore not applied to Boltz-2 predictions in this section.

AF3: ipTM Greater Than or Equal to 0.6

For AlphaFold 3, model confidence was assessed using the interface predicted template modelling score (ipTM), which quantifies the predicted accuracy of the relative orientation between the protein and DNA chains on a scale from 0 to 1, where higher values indicate greater interface confidence. Following Helsem et al. [97], a prediction was classified as reliable when ipTM was greater than or equal to 0.6. Predictions

falling below this threshold were considered to lack sufficient interface confidence for downstream biological interpretation.

Applying this threshold to the full isoform library, AF3 successfully modeled 64.18% of reference isoforms and 56.16% of alternative isoforms, a difference of approximately 8 percentage points between the two groups. Reference isoforms showed a higher overall distribution of ipTM values, with a greater density of predictions clustering above the 0.6 threshold compared to alternative isoforms. This difference in reliable prediction rates between reference and alternative isoforms is consistent with the broader finding that alternative splicing frequently alters the structural context of TF–DNA complexes, though the specific structural basis for this difference would require further analysis beyond the threshold assessment alone (Figure 3.2).

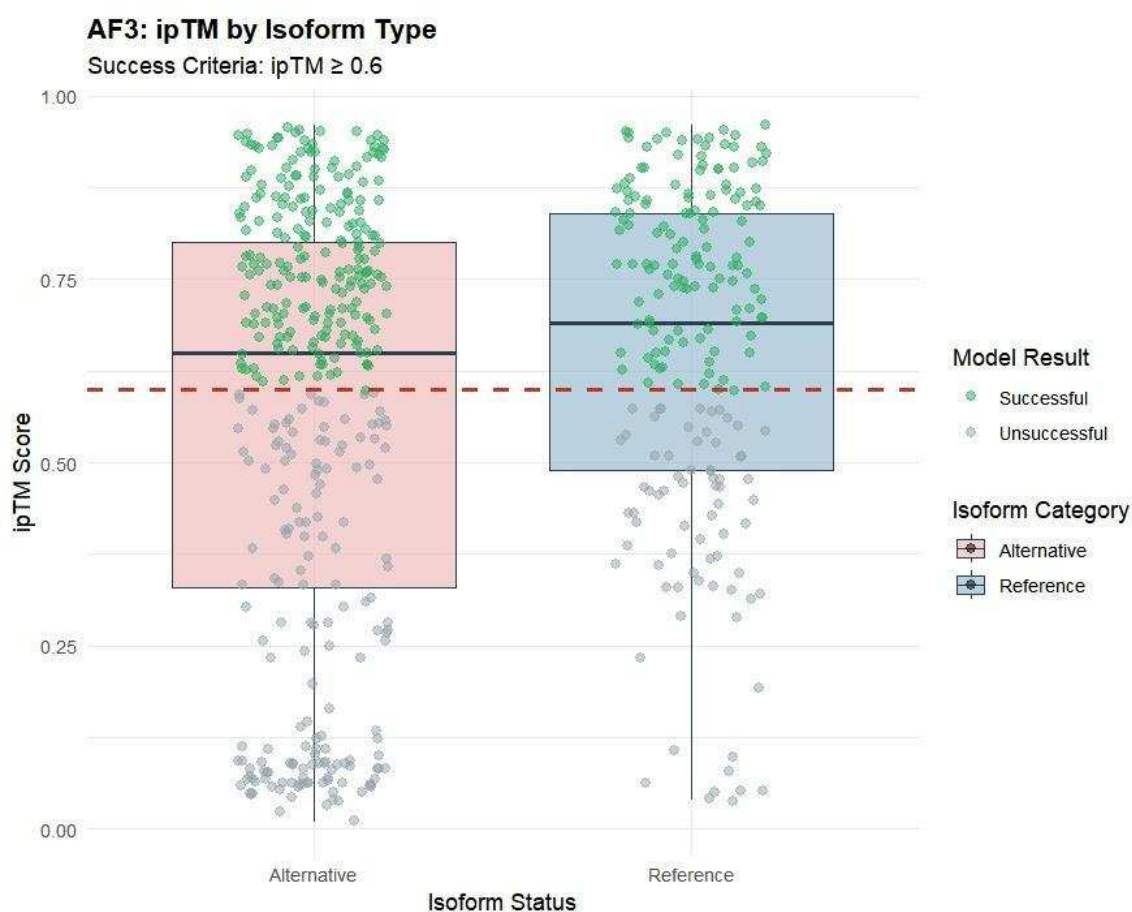


Figure 3.2 Distribution of AF3 ipTM scores for reference and alternative isoforms. The red dashed line indicates the success threshold of ipTM greater than or equal to 0.6.

RF2NA: PAE Less Than 10 Å

For RF2NA, model reliability was assessed using the mean PAE, which reflects the expected positional error in Ångströms between residue pairs across the predicted protein–DNA complex. Unlike ipTM, lower PAE values indicate greater confidence, meaning the two metrics operate in opposite directions and are not numerically comparable. Following Helsem et al. [97], a prediction was classified as reliable when mean PAE fell below 10 Å.

Applying this threshold, RF2NA predictions were largely unsuccessful across the full TF isoform library. The vast majority of PAE scores remained substantially above 10 Å for both reference and alternative isoforms, with no meaningful separation between the two groups visible in the distribution. Reliable predictions were limited to a small number of specific TF families including KLF and MAX. The proportion of isoforms meeting this threshold was too low to permit a meaningful comparison between reference and alternative groups. This pattern suggests that RF2NA, when applied to this library of human TF–DNA complexes, generally produced inter-domain positional error estimates that did not reach the level of confidence required by this threshold, indicating that this platform may face particular challenges in modeling the structural diversity of human TF–DNA interfaces (Figure 3.3).

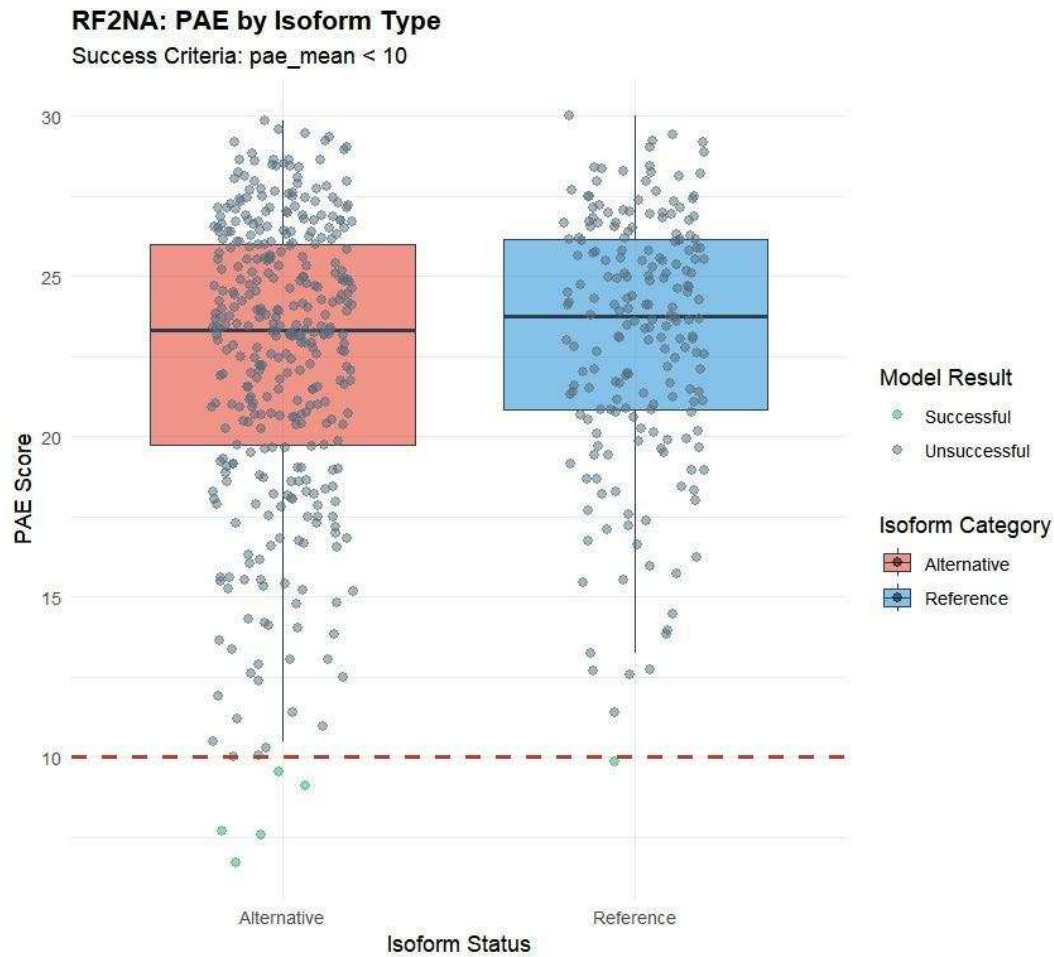


Figure 3.3 RF2NA mean PAE score distribution by isoform type. The red dashed line indicates the success threshold of PAE below 10.

3.1.2 AF3 Prediction Performance Across TF Families

To understand how modeling performance varied across the TF landscape, we examined the distribution of successful and unsuccessful AF3 predictions across all TF families represented in the dataset, using ipTM greater than or equal to 0.6 as the success criterion (Figure 3.4).

The C2H2 zinc finger family dominated the dataset numerically, accounting for 201 isoforms in total. Of these, 66 achieved reliable predictions while 135 fell below the confidence threshold, giving a failure rate of 67%. As the largest family in the dataset by a considerable margin, the C2H2 zinc finger family recorded both the highest absolute number of successful and unsuccessful predictions. The 67% failure rate also represents the highest proportional failure rate among the larger families, suggesting that the structural diversity of zinc finger array configurations presents a particular modeling challenge for AF3.

In contrast, several other well-represented families showed markedly higher proportional success rates. The Homeodomain family comprised 60 isoforms, of which 50 achieved reliable predictions and 10 did not, giving a success rate of 83%. The bHLH family achieved the highest success rate among the larger families at 95%, with 37 successful predictions out of 39 total isoforms. The HMG/Sox, bZIP, and SMAD families similarly showed high proportional success rates, with the large majority of isoforms in each group achieving reliable predictions. These families present structurally well-characterized and conserved DNA-binding interfaces that are well-represented in AF3 training data, which likely accounts for their consistently high success rates.

The Nuclear receptor family stands out as the most notable exception among the larger families. With approximately 48 isoforms in total, roughly equal numbers achieved successful and unsuccessful predictions, representing the most balanced split between the two outcomes of any well-represented family in the dataset. This suggests that nuclear receptor–DNA complexes present a particular modeling challenge for AF3, possibly reflecting the structural diversity of ligand-binding domain conformations that influence the overall complex geometry even when the DBD itself is conserved.

The IRF family showed the most unfavorable outcome of any family in the dataset, with all 7 isoforms falling below the confidence threshold, giving a 100% failure rate. Among the smaller families, p53, Rel, HSF, CSL, Paired box, T-box, Grainyhead, Homeodomain/POU, and MADS box all achieved 100% success rates, though the small number of isoforms in each of these groups limits the interpretive weight that can be placed on these observations. In contrast, several smaller families showed complete or near-complete failure: AP-2 produced 0 successful predictions from 4 total isoforms, AT hook produced 0 from 2, and TEA produced only 1 successful prediction from 4 total isoforms.

Overall, the family-level distribution reveals that AF3 performance is not uniform across TF classes. Families with structurally well-characterized and conserved DNA-binding interfaces, such as Homeodomain and bHLH, tend to achieve consistently high rates of reliable predictions. Families with greater structural diversity at the protein–DNA interface, particularly Nuclear receptor and C2H2 zinc finger, show more variable outcomes.

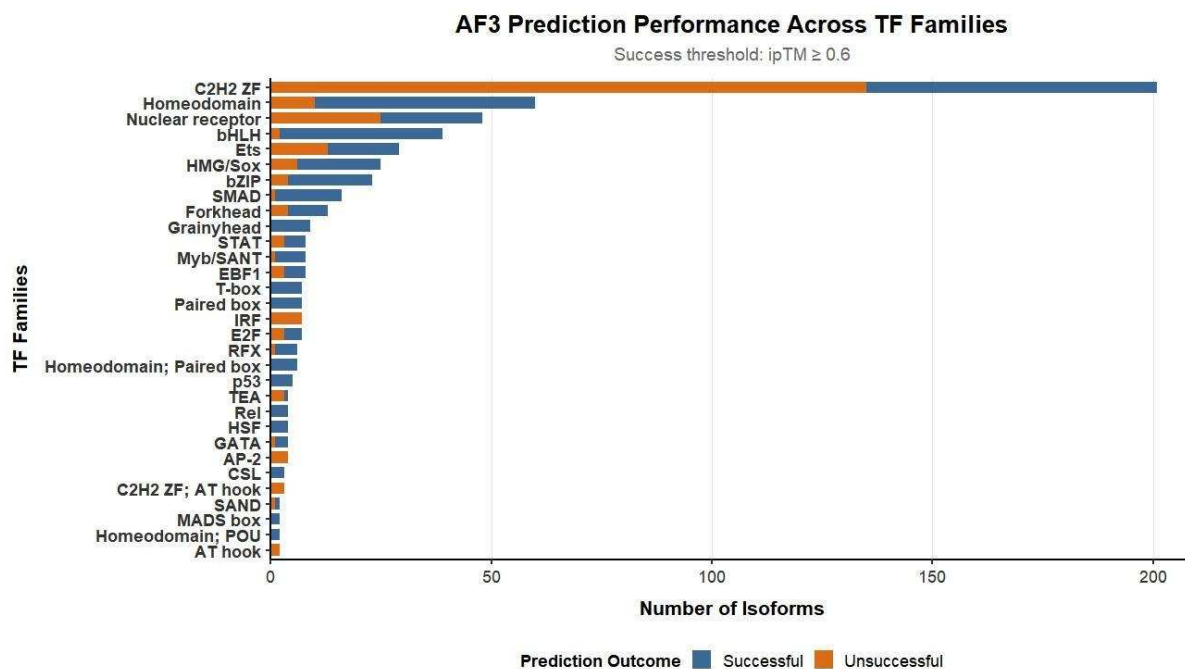


Figure 3.4 AF3 prediction performance across all TF families represented in the dataset.

3.2 Performance comparison of RF2NA, AF3, and Boltz-2 on protein-DNA complex predictions for alternative TFs not perturbing DBDs

We next assessed how well AF3, Boltz-2, and RF2NA model protein-DNA complexes of alternative isoforms that retain the DNA-binding domains (DBDs). To do this, we modeled the interactions of 252 alternative TFs with no DBD loss and the DNA motifs of their corresponding TFs (Section 2.2). Because these alternative TF isoforms retain their DBDs, it is expected that the above models predict protein-DNA complexes of such isoforms with similar accuracy as of reference TFs.

Mean pLDDT values remained remarkably stable between groups. In AF3, alternative isoforms showed nearly identical median pLDDT values to reference counterparts (0.630 vs. 0.634). Boltz-2 followed this trend with a negligible difference (0.557 alt vs. 0.559 ref). Interestingly, RF2NA predicted a slight increase in backbone confidence for alternative isoforms (0.654) relative to the reference (0.642) (Figure 3.5). This consistency suggests that the local folding of alternative isoforms is as well-defined as that of the reference TFs.

The PAE metric, which reflects the confidence in the relative orientation of domains, showed that alternative isoforms often achieve tighter structural arrangements. Boltz-2 showed the most significant improvement, with alternative isoforms reaching a median

PAE of 16.0 Å, compared to 18.0 Å for reference isoforms. AF3 and RF2NA showed minor shifts, with alternative isoforms exhibiting slightly lower (better) PAE values (AF3: 24.1 vs. 24.2 Å; RF2NA: 23.7 vs. 24.1 Å) (Figure 3.5). The lower PAE values in alternative isoforms across all methods suggest a potential for more compact or rigid domain-domain packing in these variants.

Interface confidence (ipTM) was nearly indistinguishable between the two groups. In AF3, the alternative isoforms achieved a median ipTM of 0.75, marginally higher than the reference 0.74. Boltz-2 mirrored this stability, with median scores of 0.950 for alternative versus 0.946 for reference isoforms. These high ipTM scores across both groups indicate that the interface integrity is preserved, regardless of the isoform status.

Overall, the data indicate that alternative isoforms retaining the DBD possess structural confidence levels that are broadly comparable, and in some cases superior, to their reference counterparts. While AF3 and RF2NA show minor fluctuations, Boltz-2 identifies a notable increase in structural compactness (lower PAE) for alternative isoforms.

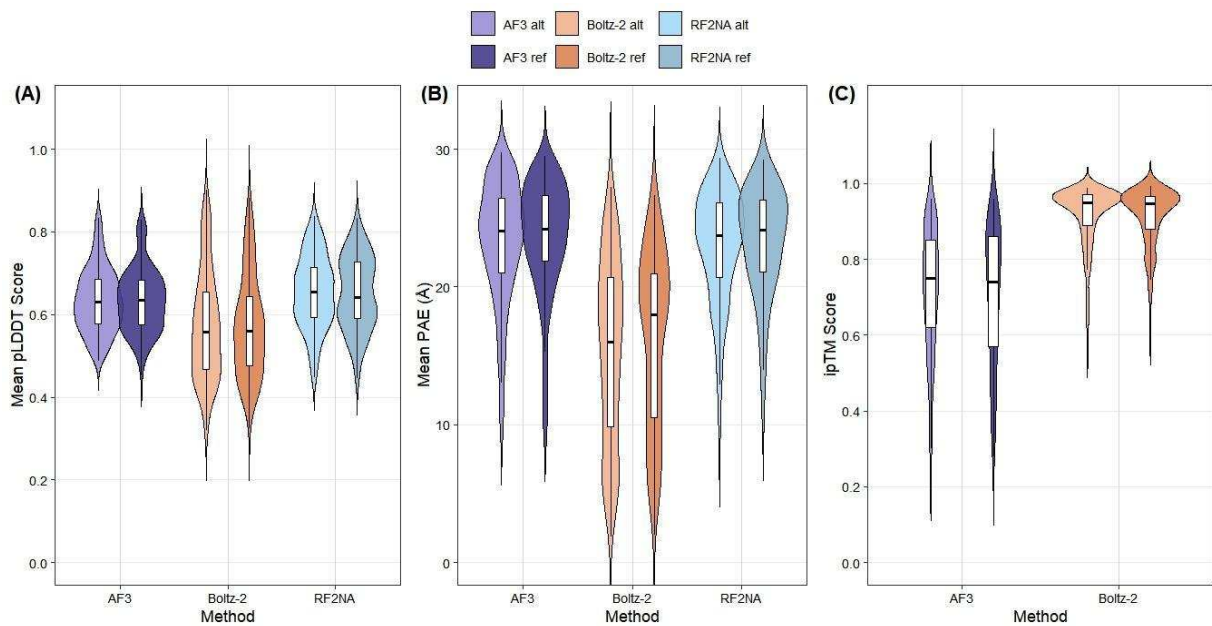


Figure 3.5 Paired comparison of structural confidence metrics between reference and alternative TF–DNA complexes lacking DBD loss across prediction frameworks.

3.2.1 Paired Comparative Structural Analysis of Alternative TF Isoforms Without DBD Loss, Ordered by RF2NA Prediction

To explore how alternative splicing affects TF–DNA complex structure when the DNA-binding domain is left completely intact, we computed paired delta scores (Δ =

Alternative – Reference) for all 252 DBD-intact alternative isoforms across three structural metrics (pLDDT, PAE, and ipTM) and three independent prediction platforms, RF2NA, AF3, and Boltz-2. All 252 alternative–reference pairs were ranked by their RF2NA Δ pLDDT score from the most negative on the left to the most positive on the right, and this same ordering was kept fixed across all eight panels (A–H), so that each isoform pair always appears in the same horizontal position regardless of which metric or platform is being examined. This makes it possible to follow any single isoform across the entire figure without losing track of where it sits in the broader landscape.

In panel A, the alternative-reference paired Δ pLDDT scores spread across a wide range in both directions and do not sit quietly around zero. Some pairs show meaningful structural loss in the alternative isoform relative to its reference, while others show meaningful structural gain, with the largest values reaching above +0.2. If the regions outside the DBD had no real structural relevance, we would expect the delta scores to be small and essentially random. This spread suggests that even when the DBD sequence is completely preserved, the regions that differ between the alternative and reference isoforms, primarily the flanking IDRs that are removed or truncated through splicing, are not structurally silent. Their presence or absence visibly shifts how confidently and how precisely the whole TF–DNA complex is predicted to fold. Alternative splicing outside the DBD is not a neutral event structurally, and the spread of values in panel A makes that clear before looking at a single individual isoform.

It is also worth noting that the distribution is not symmetric. More alternative isoforms gain structural confidence than lose it. This makes biological sense in the context of the autoinhibitory model proposed by Lambourne et al. (2025), where the IDRs present in reference isoforms physically fold back onto the DBD and constrain it. If that is true, then removing those IDRs through alternative splicing would be expected to free the DBD and allow it to fold more stably, and that is broadly what the distribution shows.

When we applied the same RF2NA-derived ranking to the AF3 and Boltz-2 pLDDT panels, panels C and F respectively, the right side of the distribution stayed predominantly blue across both platforms. Boltz-2 in particular shows the most pronounced gains, with the rightmost isoforms reaching well above +0.2. In the PAE panels, both panel B and panel D, the right end of the distribution, where the most structurally stabilized isoforms sit, is dominated by red bars, meaning the alternative

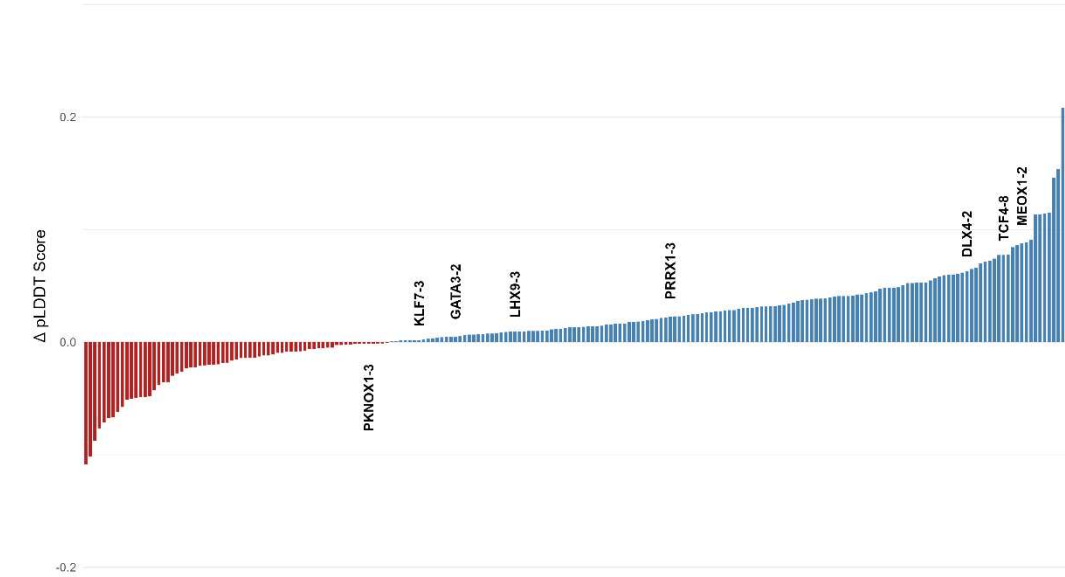
isoform shows lower PAE than the reference and therefore better inter-domain positional accuracy. The combination of gaining pLDDT while simultaneously reducing PAE at the same end of the ranked distribution is a coherent structural signal that points to globally more organized complexes in these isoforms.

In the ipTM panels, neither panel E nor panel H shows a clear directional trend, both display a roughly mixed distribution of positive and negative values throughout the ranked order. The DBD is sequence-identical between alternative and reference in all 252 cases, meaning the domain that physically contacts DNA has not changed. It would be surprising if the predicted interface quality shifted systematically in one direction, and the mixed ipTM pattern is consistent with the idea that the structural changes we observe in pLDDT and PAE are happening in the regions surrounding the interface rather than at the contact surface itself.

To identify the most biologically meaningful isoforms for structural interpretation, we filtered the Lambourne et al. (2025) experimental dataset following the approach described in the Methods. Starting from 486 discordant protein-bait pairs, where alternative and reference isoforms from the same gene showed opposite binding outcomes for the same DNA target, we narrowed the set down in two steps: first by keeping only baits where a confirmed TF binding motif was present, which brought the number down from 486 to 56 pairs, and then by keeping only those pairs where the alternative isoform had lost none of its DBD, which left us with 31 discordant pairs representing eight unique isoforms: DLX4-2, GATA3-2, KLF7-3, LHX9-3, MEOX1-2, PKNOX1-3, PRRX1-3, and TCF4-8. These eight isoforms were not chosen based on anything structural. They were identified purely through experimental evidence showing that their DNA-binding behavior differs from their reference isoform at well-defined, sequence-specific binding sites, while their DBD remains completely untouched. This combination identifies these as cases in which sequence changes outside the DBD must account for the observed functional differences, since the binding domain itself is fully intact.

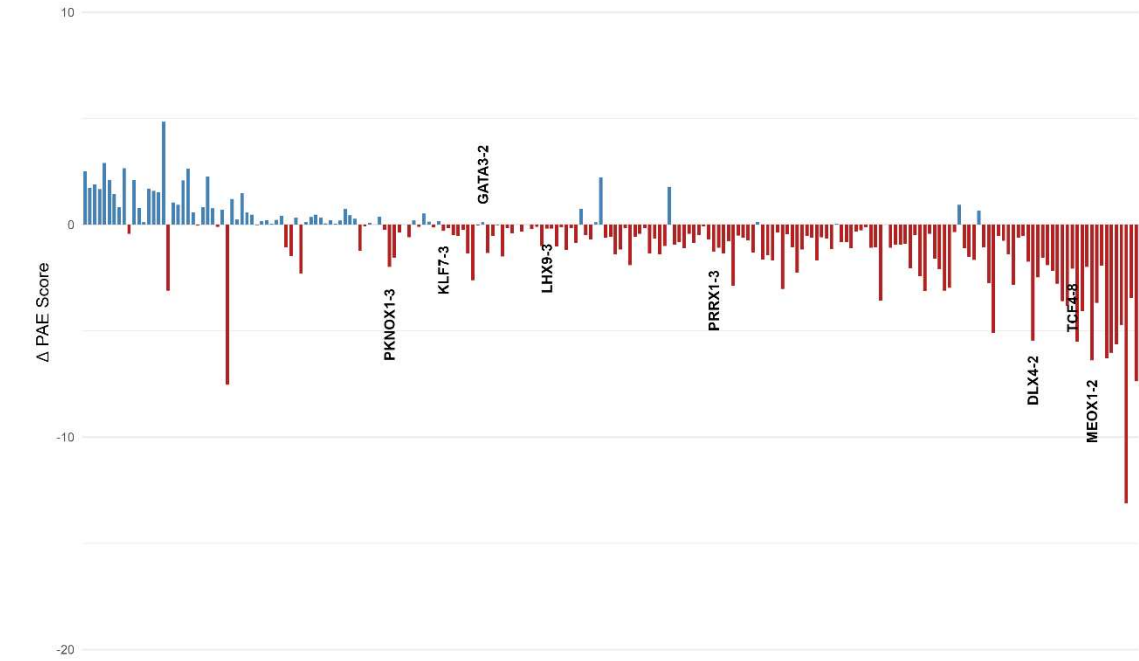
(A)

RF2NA



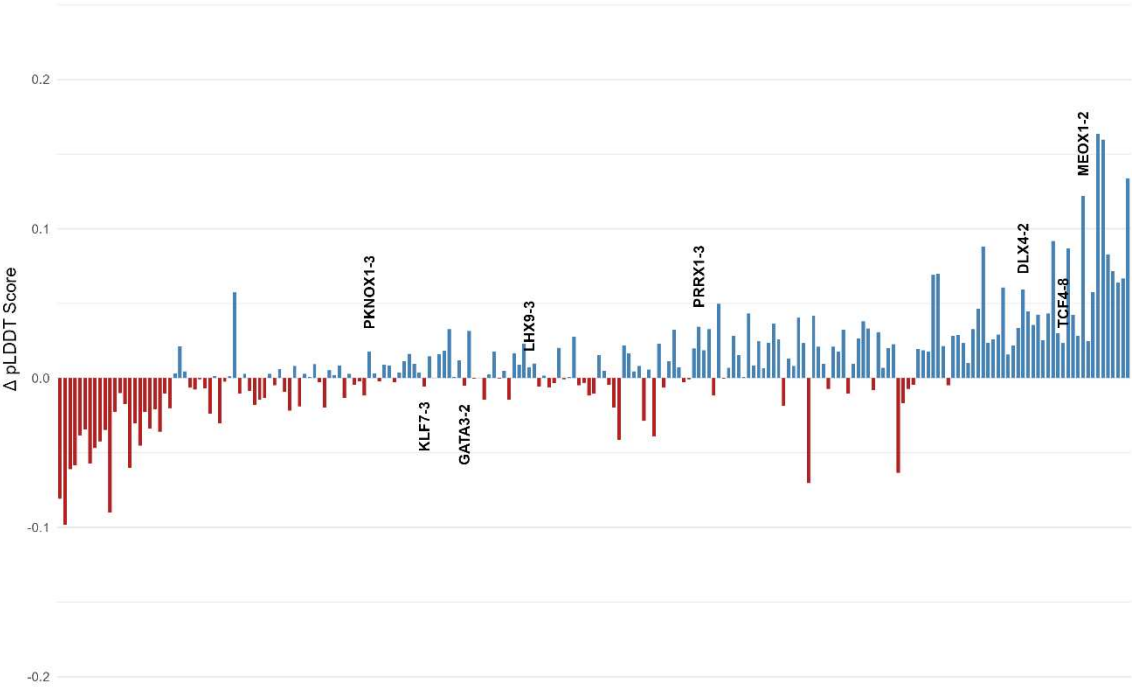
(B)

RF2NA



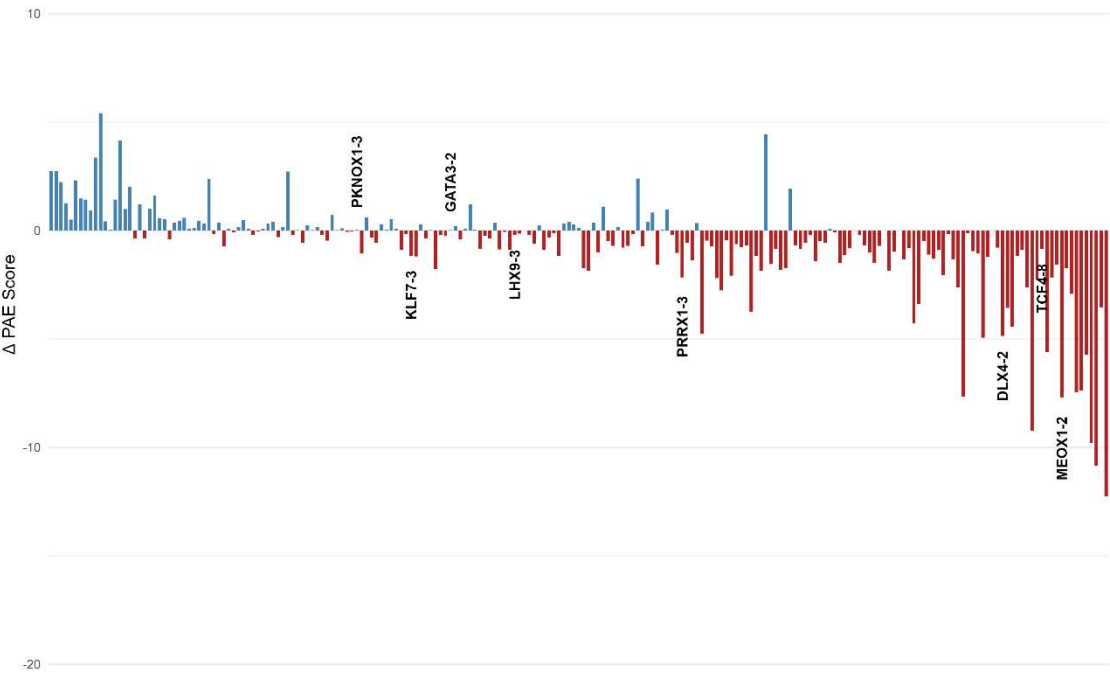
(C)

AF3



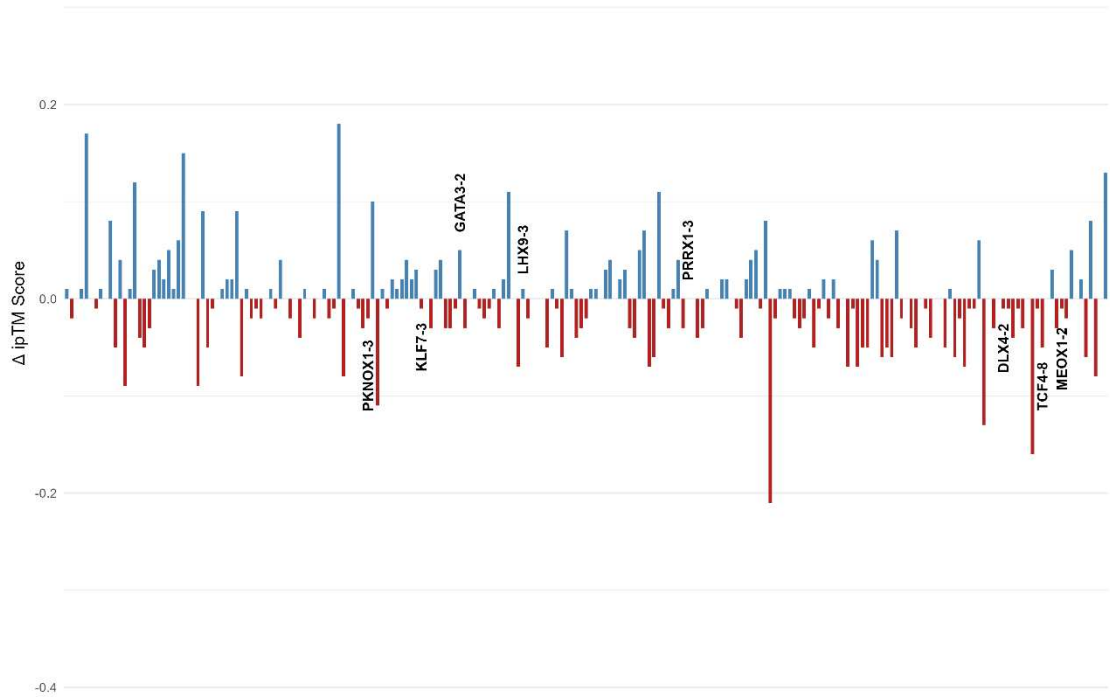
(D)

AF3



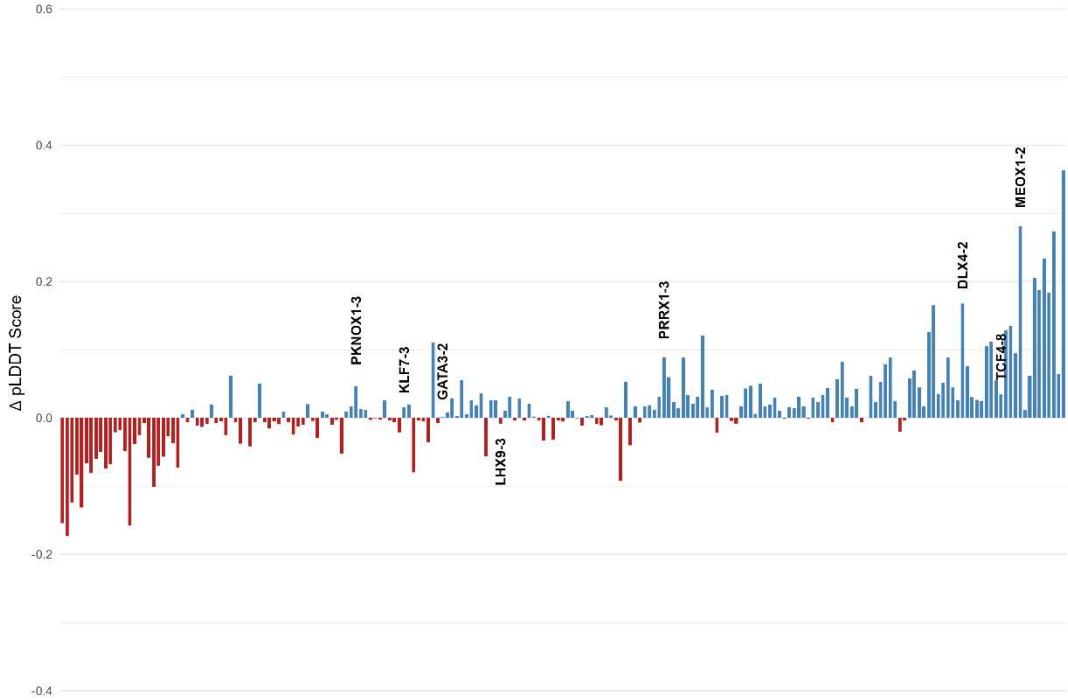
(E)

AF3



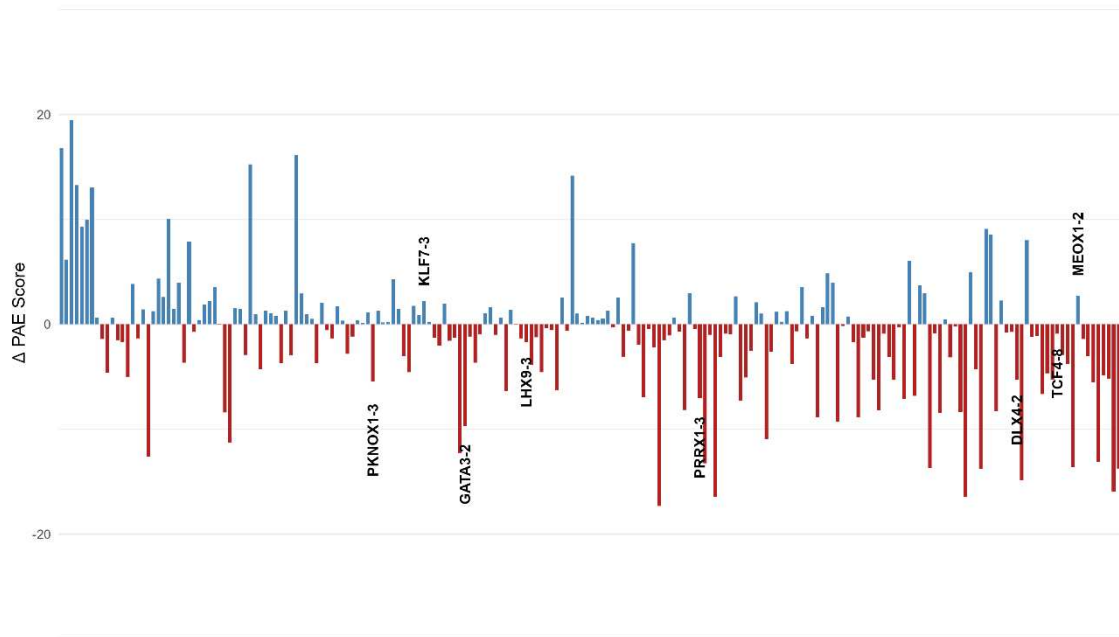
(F)

Boltz-2



(G)

Boltz-2



(H)

Boltz-2

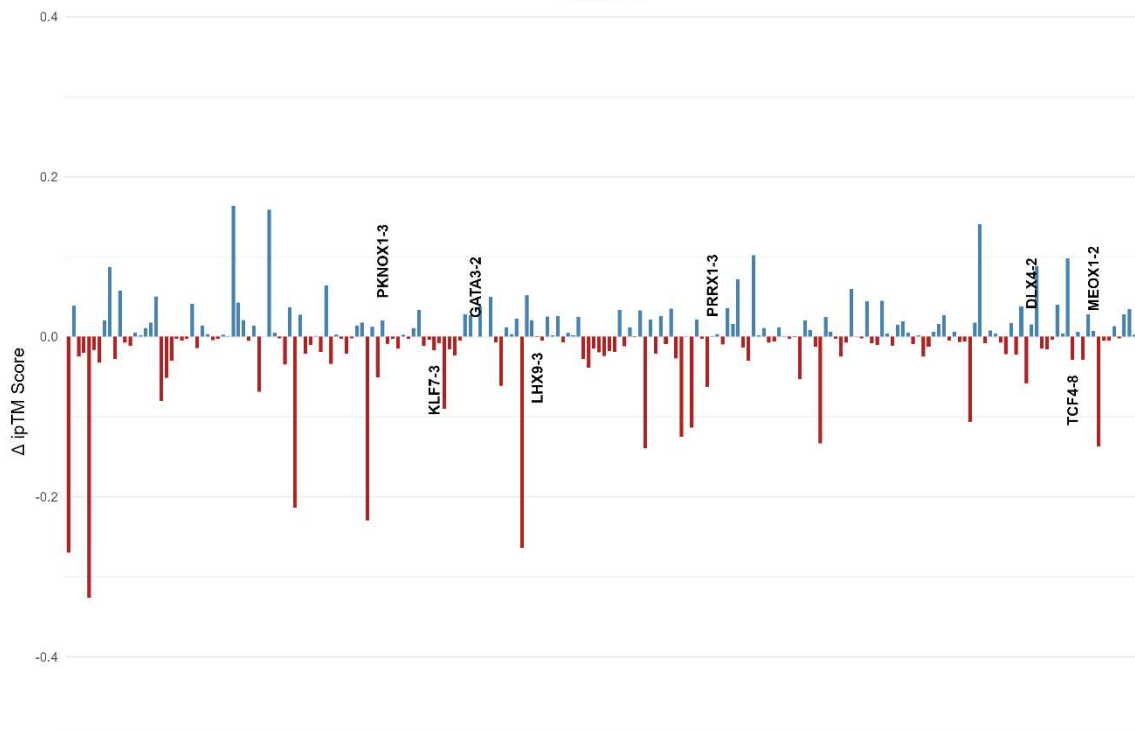


Figure 3.6 Paired comparative analysis of structural metrics alternative TF Isoforms without DBD loss ordered by RF2NA Prediction (A-H panels).

Looking at the quantitative delta scores across all eight isoforms, DLX4-2 and

PRRX1-3 stand out as the most structurally consistent, showing the same direction across every platform and every metric without a single contradictory signal. Both gained local folding confidence in RF2NA, AF3, and Boltz-2 (DLX4-2: +0.063, +0.059, +0.168; PRRX1-3: +0.023, +0.034, +0.089), both showed improved inter-domain positional accuracy across all three platforms (DLX4-2: RF2NA = -5.45 , AF3 = -4.87 , Boltz-2 = -5.27 Å; PRRX1-3: RF2NA = -1.27 , AF3 = -2.16 , Boltz-2 = -7.02 Å), and both showed near-neutral interface confidence changes with no contradictory signals anywhere in their profiles (DLX4-2: AF3 = -0.01 , Boltz-2 = $+0.015$; PRRX1-3: AF3 = 0.00 , Boltz-2 = $+0.001$). This cross-platform agreement across all three metrics makes DLX4-2 and PRRX1-3 the most computationally robust candidates in the dataset.

MEOX1-2 showed the largest pLDDT gains of any isoform across all three platforms (RF2NA = +0.087, AF3 = +0.122, Boltz-2 = +0.281), and its Δ PAE was negative in both RF2NA and AF3 (-6.39 and -7.69 Å respectively), meaning the alternative isoform is predicted to have better inter-domain positional accuracy than the reference in both platforms. Boltz-2, however, predicted a positive Δ PAE of $+2.74$ Å for this isoform, meaning the alternative isoform actually has worse inter-domain positioning than its reference according to this platform, despite folding with greater local confidence. This is an important distinction: MEOX1-2 consistently folds better locally across all three platforms, but how the domains are positioned relative to each other is where Boltz-2 disagrees with RF2NA and AF3. Its Δ ipTM is near-neutral in both AF3 (-0.01) and Boltz-2 ($+0.007$), meaning the binding interface itself remains largely unchanged between alternative and reference regardless of platform.

TCF4-8 shows a similar trajectory to DLX4-2 and PRRX1-3 in terms of pLDDT and PAE, consistently positive Δ pLDDT across all three platforms (RF2NA = +0.077, AF3 = +0.023, Boltz-2 = +0.035) and negative Δ PAE in all three (RF2NA = -2.08 , AF3 = -0.86 , Boltz-2 = -0.90). Where it clearly diverges from the others is in its Δ ipTM, which is negative in both AF3 (-0.05) and Boltz-2 (-0.029), representing the most consistently negative interface values of any isoform in the group. This tells us that while TCF4-8 gains local structural confidence and reduces inter-domain positional error, the predicted binding interface becomes slightly less precisely resolved in the alternative form. This is a distinct profile from DLX4-2 and PRRX1-3, and it may indicate that the structural reorganization in TCF4-8 extends closer to the actual DNA contact surface than in the other isoforms.

PKNOX1-3 and GATA3-2 both show genuine structural reorganization signals. PKNOX1-3 is the only isoform with a negative RF2NA Δ pLDDT (-0.001), but both AF3 and Boltz-2 predict positive pLDDT gains ($+0.018$ and $+0.047$), and its Δ PAE is negative in both RF2NA and Boltz-2 (-1.98 and -5.46), with only a near-neutral AF3 Δ PAE ($+0.03$) as the exception. Its Δ ipTM is slightly negative in AF3 (-0.02) but positive in Boltz-2 ($+0.020$), keeping the interface picture broadly neutral overall. GATA3-2 shows small and mixed pLDDT signals across platforms (RF2NA = $+0.005$, AF3 = -0.005 , Boltz-2 = $+0.001$) and near-neutral Δ PAE in RF2NA and AF3 ($+0.109$ and $+0.010$), but its Boltz-2 Δ PAE is one of the largest reductions in the entire dataset (-9.70), a striking signal suggesting that Boltz-2 may be resolving a structural feature the other two platforms do not fully capture for this isoform. Its Δ ipTM is positive in AF3 ($+0.05$) and essentially zero in Boltz-2 ($+0.0003$), consistent with a preserved or marginally improved binding interface.

LHX9-3 and KLF7-3 show the weakest and most mixed structural signals of the eight. LHX9-3 shows consistently negative Δ PAE across all three platforms (RF2NA = -0.188 , AF3 = -0.20 , Boltz-2 = -1.67 Å), meaning modest but consistent improvement in inter-domain positioning, while its pLDDT changes are essentially flat and its Δ ipTM is near-zero in both AF3 and Boltz-2. The alternative isoform is structurally very similar to its reference, with only a conservative inter-domain improvement detectable. KLF7-3 is the most inconsistent of the group. Its pLDDT is slightly positive in RF2NA and Boltz-2 but slightly negative in AF3, its Δ PAE is negative in RF2NA and AF3 meaning better inter-domain positioning, but positive in Boltz-2 ($+2.22$ Å) meaning worse positioning, and its Δ ipTM is slightly negative in both AF3 (-0.01) and Boltz-2 (-0.017). The three platforms simply do not agree on KLF7-3, making it the most difficult isoform to interpret structurally in this dataset.

Taken together, the dominant trend across all eight isoforms is one of structural refinement in the alternative form relative to the reference. Six of the eight show positive Δ pLDDT in at least two platforms, and most of these also show negative Δ PAE in at least two platforms, meaning local folding confidence and inter-domain positional accuracy improve together. The consistency of this trend across three independently built prediction frameworks, despite platform-specific differences in individual isoforms, strengthens confidence that it reflects genuine biological signal rather than modeling noise.

These structural findings are best understood in the context of the autoinhibitory model

that Lambourne et al. (2025) proposed for DLX4 and PKNOX1 specifically. According to this study, the reference isoforms DLX4-1 and PKNOX1-1 were highly specific in their DNA binding, while the alternative isoforms DLX4-2 and PKNOX1-3, which carry truncations of large N-terminal IDRs, showed considerably broader binding. To explore why, they calculated predicted interaction strengths between the isoform-specific IDRs and the DNA-binding residues in the homeodomains of both TFs, and found that the reference isoforms showed strong putative intramolecular contacts between their IDRs and their DNA-binding residues while the alternative isoforms showed comparatively weak equivalents. They interpreted this as consistent with a model where the IDR acts as a tethered inhibitor that folds back onto the DBD and restricts which DNA sequences it can reach, while being careful to note that these results should be taken qualitatively, that many additional factors could also contribute to the observed binding differences, and that this represents a plausible and testable physical model rather than a confirmed mechanism.

With that in mind, the pLDDT gains and PAE reductions we observe for DLX4-2 and PKNOX1-3 fit naturally within this framework. When we correlate these structural changes with the experimental interaction data, both isoforms show PDI rewiring or PDI gain in the wet-lab alongside structural stabilization computationally, with higher pLDDT, lower PAE, and largely preserved ipTM. This alignment suggests that removing the IDRs that were constraining the DBD allows it to settle into a more stable and precisely organized conformation, and that new structural arrangement translates into a different and broader DNA-binding profile without any loss of the canonical binding domain. For the remaining six isoforms, the same intramolecular inhibitory mechanism has not been directly tested, and we should not assume it applies equally to all of them. What we can say is that the broader observation from Lambourne et al., that IDR differences outside the DBD frequently drive changes in DNA-binding behavior even when the binding domain itself is intact, gives us a consistent biological framework for interpreting the structural stabilization we observe more widely, while acknowledging that the specific physical basis may differ between isoforms and that experimental follow-up on each one individually would be needed to establish that with confidence.

3.3 Structural impact of DBD loss in alternative TF isoforms

To investigate whether the structural differences observed between reference and alternative TF–DNA complexes are driven by partial or complete loss of the DBD, we

focused on isoform pairs in which the alternative isoform carries a measurable percentage of DBD loss, as reported in Table S13 of Lambourne et al. (2025). This allowed us to directly evaluate the structural consequences of DBD truncation on predicted complex stability and interface accuracy.

Mean pLDDT scores revealed a general decrease in local folding confidence for alternative isoforms across most platforms, with the exception of RF2NA. In AF3, DBD loss was associated with a median pLDDT drop from 0.617 in reference isoforms to 0.572 in alternative isoforms. Boltz-2 showed a similar downward trend, with median scores shifting from 0.537 in reference isoforms to 0.516 in alternative isoforms. RF2NA, however, predicted a marginal increase in backbone confidence for alternative isoforms (0.640) compared to reference isoforms (0.626), suggesting that this platform may be less sensitive to the structural consequences of DBD truncation than the other two methods.

In terms of PAE, Boltz-2 showed the most striking difference between reference and alternative isoforms. It predicted a substantial reduction in inter-domain alignment error for DBD-truncated isoforms, with a median PAE of 6.08 Å compared to 16.3 Å for reference isoforms. This suggests that in the absence of the DBD, Boltz-2 predicts a considerably more compact or rigid arrangement of the remaining domains. AF3 and RF2NA, by contrast, maintained much higher error profiles with minimal change between groups, AF3 showing almost no difference (25.1 vs. 24.7 Å) and RF2NA showing only a modest reduction (23.9 vs. 22.9 Å).

The most pronounced effect of DBD loss was observed in the ipTM scores, particularly in AF3. While reference isoforms achieved a median ipTM of 0.71, alternative isoforms with DBD loss dropped sharply to 0.27, indicating that AF3 largely fails to identify a confident protein–DNA interface when the canonical binding domain is absent. Boltz-2 also showed a decrease, but maintained a considerably higher level of interface confidence, with median scores shifting from 0.946 in reference isoforms to 0.874 in alternative isoforms.

Taken together, these results demonstrate that DBD loss significantly perturbs structural confidence metrics, though the magnitude of this effect varies considerably between platforms. AF3 appears the most sensitive, showing a near-complete collapse of predicted interface confidence when the DBD is absent. Boltz-2, by contrast, suggests that DBD-truncated isoforms may adopt more compact alternative

configurations while retaining relatively high interface scores, a pattern that likely reflects fundamental differences in how these two platforms model incomplete or truncated binding domains.

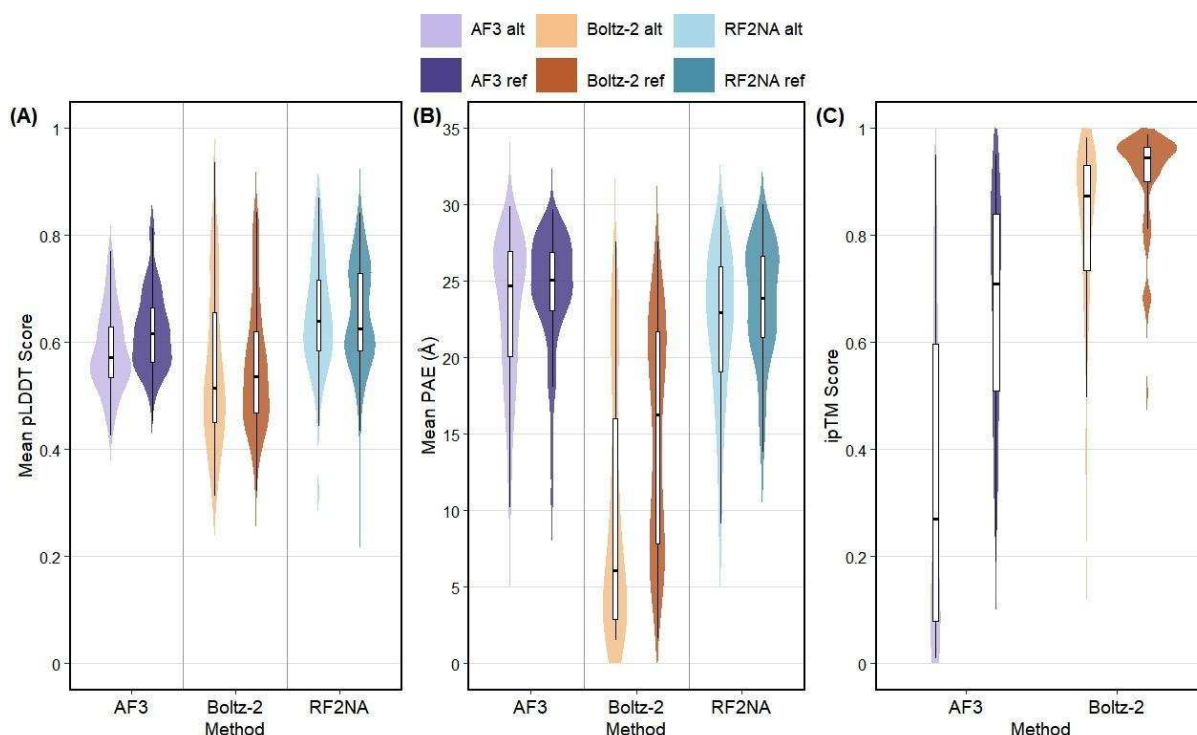


Figure 3.7 Structural modeling performance of AF3, Boltz-2, and RF2NA on TF isoforms with DNA-binding DBD loss.

3.4 Evaluating the Predictive Accuracy of AI Structural Models for Protein–DNA Interaction

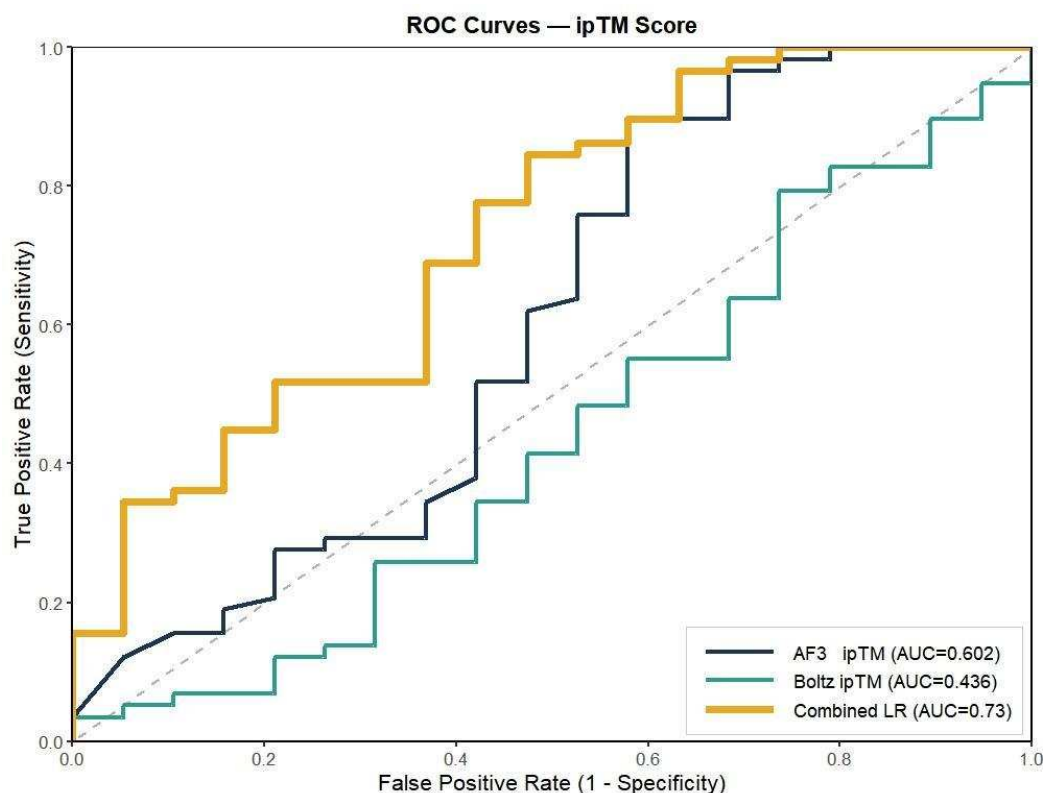
To evaluate the extent to which AI-generated structural confidence scores can predict experimental protein–DNA binding outcomes, we performed a ROC analysis across all three platforms and all three structural metrics. To ensure statistical validity, we first addressed the issue of data redundancy inherent to our experimental framework. Each protein isoform was tested against multiple DNA baits in the eY1H assay, but AI-generated structural confidence scores, pLDDT, ipTM, and PAE, are properties of the isoform structure itself and therefore remain identical regardless of which bait is paired with it in any given comparison. Retaining the raw repeated data would have artificially inflated the sample size by treating the same structural prediction as multiple independent observations. The dataset was therefore deduplicated to 79 unique isoforms. Each isoform was then assigned a binary functional classification: an isoform was designated as a positive case if it demonstrated the capacity to interact with at least one DNA bait, and as a negative case only if no interactions were detected

across all baits tested. This classification ensures that every data point in the ROC analysis represents an independent biological entity.

Evaluation of individual structural metrics revealed consistently limited discriminatory power for distinguishing interacting from non-interacting isoforms across all three platforms (Figure 3.8). All pLDDT values performed near or below the random-guess baseline of 0.50 regardless of platform (AF3 = 0.530, RF2NA = 0.507, Boltz-2 = 0.478), demonstrating that local folding confidence provides essentially no useful information for predicting protein–DNA binding competency in this dataset. PAE showed similarly weak individual discriminatory power when evaluated in the biologically expected direction, where lower PAE was treated as predictive of binding on the assumption that more precise inter-domain positioning should favor DNA interaction (AF3 = 0.514, RF2NA = 0.459, Boltz-2 = 0.303). The AUC values at or below chance for PAE confirm that this expected relationship does not hold in this dataset, a point that is further clarified by the correlation analysis discussed below. Among all individual metrics, AF3 ipTM was the strongest predictor of binding competency, achieving an AUC of 0.602, while Boltz-2 ipTM performed below chance at 0.436, indicating that Boltz-2 interface confidence scores are inversely associated with experimental binding in this dataset. RF2NA ipTM was not available for evaluation. Collectively, these results demonstrate that no individual structural confidence metric from any of the three platforms provides reliable standalone discrimination between interacting and non-interacting TF isoforms when applied without task-specific training on experimental binding data.

Combining all available structural metrics across the three platforms into a logistic regression model substantially improved predictive performance, achieving a combined AUC of 0.730 (Figure 3.8, orange line). This represents a meaningful improvement over the best individual metric (AF3 ipTM = 0.602), demonstrating that the three platforms capture partially complementary structural information that, when integrated, provides considerably greater discriminatory power than any platform alone. Each platform was built on a different computational architecture: AF3 uses a transformer-based multimodal approach, RF2NA integrates evolutionary information through a three-track network, and Boltz-2 uses a diffusion-based generative framework. Combining their outputs in a logistic regression model exploits the structural information each captures independently, which explains the substantial gain in predictive performance. The combined logistic regression curve is presented

consistently across all three ROC panels as a fixed reference to facilitate comparison regardless of which individual metric is displayed in each panel. Given the relatively modest sample size of 79 unique isoforms and the use of eight predictors in the combined model, the possibility of overfitting cannot be excluded, and the reported AUC of 0.730 should be interpreted as an upper bound on predictive performance that would benefit from validation in a larger independent dataset.



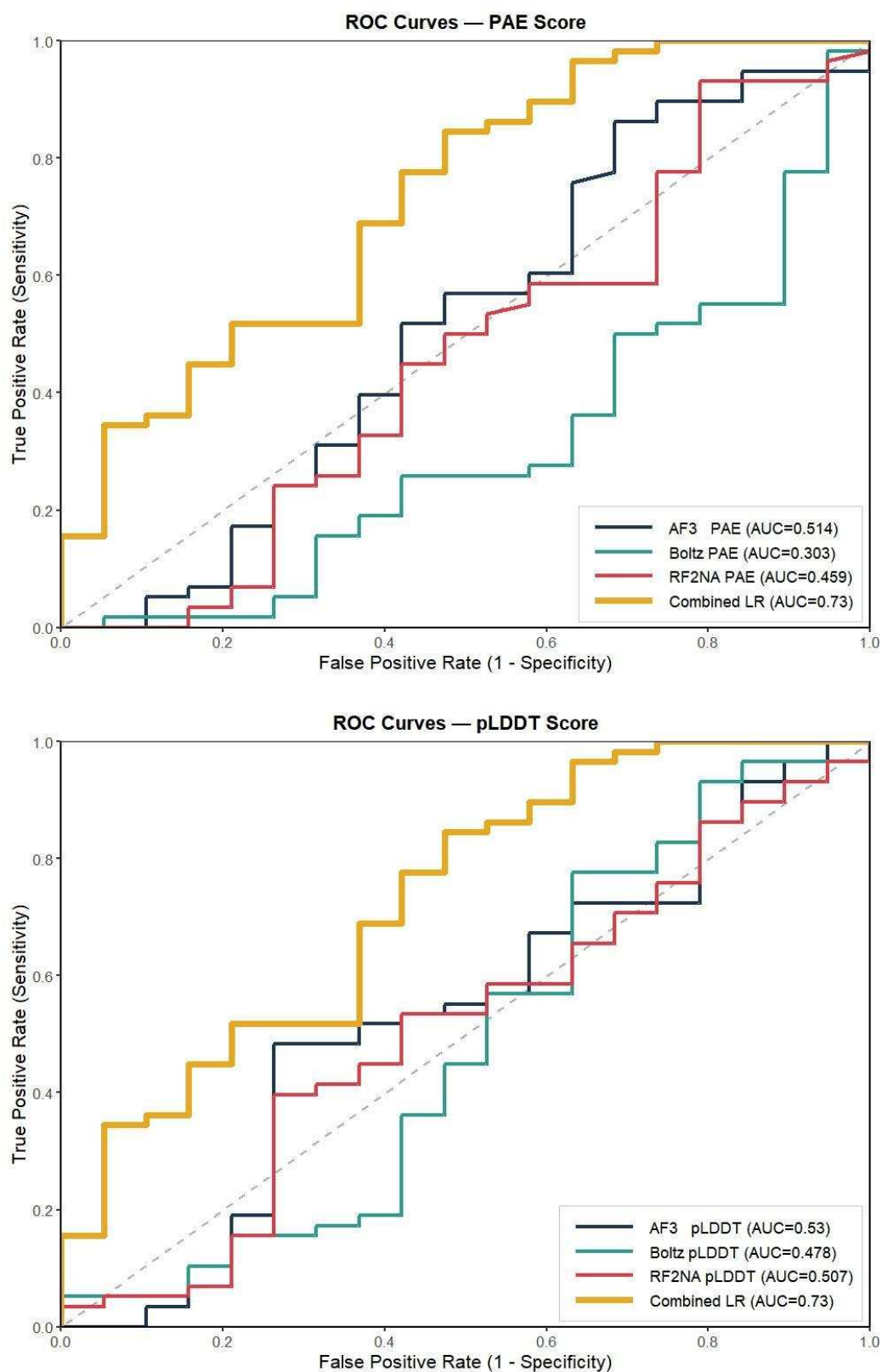


Figure 3.8 Receiver operating characteristic (ROC) analysis of AI structural confidence metrics for predicting protein–DNA interactions across 79 unique isoforms.

Correlation analysis between structural metrics and experimental binding outcomes revealed several biologically informative patterns that help explain the ROC results

(Figure 3.9). All pLDDT values showed negative correlations with experimental binding across both Pearson and Spearman coefficients (AF3: $r = -0.07$, $\rho = -0.01$; Boltz-2: $r = -0.13$, $\rho = -0.15$; RF2NA: $r = -0.18$, $\rho = -0.23$), meaning that isoforms with higher local folding confidence are slightly less likely to show experimental binding. This is the opposite of what one might intuitively expect, but it is mechanistically interpretable in the context of the structural findings discussed in section 3.2. Reference isoforms with tightly constrained, high-confidence structures frequently show restricted DNA binding due to autoinhibitory IDR contacts, while alternative isoforms with somewhat lower or more variable pLDDT show broader binding following IDR removal. In this system, high pLDDT in the reference isoform may therefore reflect the structural constraint that limits DNA binding rather than serving as a positive predictor of it, which explains why the relationship between pLDDT and binding runs in the negative direction across the dataset.

PAE showed positive correlations with experimental binding across all three platforms in both Pearson and Spearman analyses (AF3: $r = +0.15$, $\rho = +0.20$; Boltz-2: $r = +0.36$, $\rho = +0.36$; RF2NA: $r = +0.24$, $\rho = +0.28$), meaning higher inter-domain positional uncertainty is associated with greater likelihood of experimental binding. Boltz-2 PAE showed the strongest linear correlation of any individual metric in the entire dataset (Pearson $r = 0.36$). This positive direction is consistent with the ROC findings: since PAE was evaluated in the ROC analysis under the conventional assumption that lower PAE should predict binding, the resulting AUC values at or below chance confirm that this assumption does not hold here and that the actual predictive relationship runs in the opposite direction. Both analyses therefore point to the same conclusion: contrary to the biologically expected direction, isoforms with higher inter-domain positional uncertainty are more likely to show experimental binding in this system. The most plausible explanation is that isoforms with higher PAE tend to be larger or more disordered proteins with greater conformational flexibility, and such proteins may be more likely to produce detectable interactions in eY1H assays through a wider range of binding configurations. In this context, PAE is likely functioning as a proxy for protein disorder and structural flexibility rather than as a direct measure of binding precision.

Among all interface-specific metrics, AF3 ipTM showed the most biologically interpretable correlation pattern. It was the only metric to show consistent positive correlations with experimental binding across both Pearson ($r = +0.18$) and Spearman

($\rho = +0.21$) analyses, indicating that higher predicted interface confidence is modestly but consistently associated with greater likelihood of experimental binding. This is the expected direction of association for an interface metric, and the fact that AF3 ipTM is the only metric to show this expected direction clearly distinguishes it from the other metrics in this dataset. Boltz-2 ipTM showed a near-zero Pearson correlation ($r = -0.01$) and a weak positive Spearman correlation ($\rho = +0.14$), indicating that while Boltz-2 can rank isoforms with some limited consistency, its interface confidence scores do not have a meaningful linear relationship with experimental binding outcomes. The AUC of 0.436 for Boltz-2 ipTM in the ROC analysis is consistent with this finding, as performing below chance indicates that higher Boltz-2 ipTM is actually negatively associated with binding probability in this dataset, suggesting that Boltz-2 interface confidence scores may not reliably reflect the biological reality of TF–DNA binding in this context.

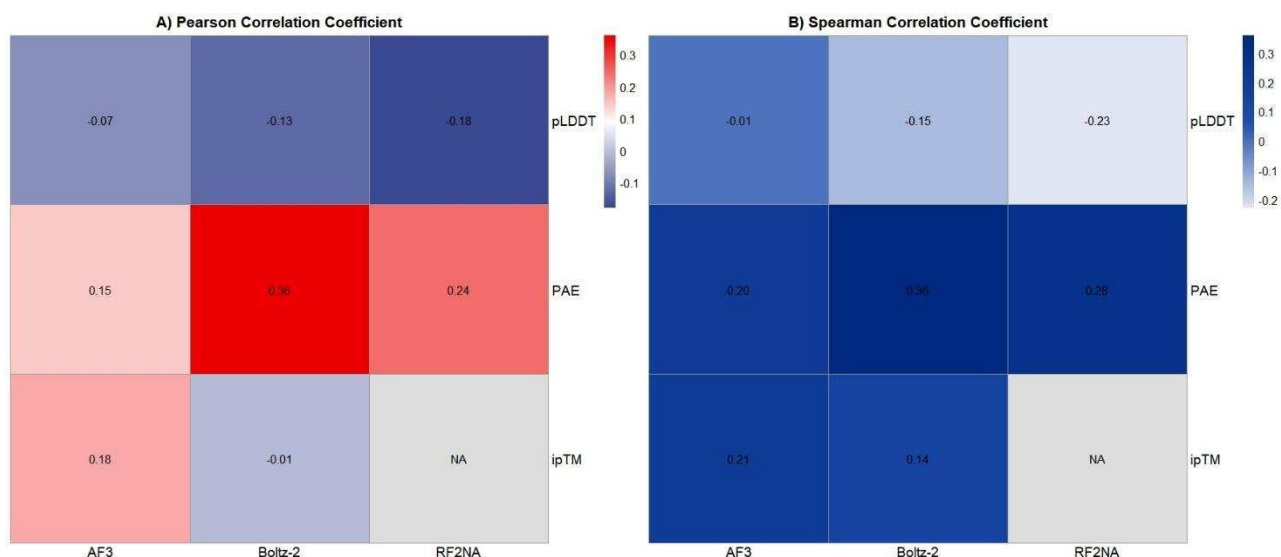


Figure 3.9 Correlation analysis of AI structural confidence metrics against experimental binding outcomes for 79 unique isoforms.

4 Discussion

The systematic characterization of transcription factor isoform diversity has long been constrained by the limitations of sequence-level analysis, which cannot capture the three-dimensional structural consequences of alternative splicing on protein–DNA recognition. This study aimed to address this limitation by applying three complementary deep learning frameworks, AF3, RF2NA, and Boltz-2, to predict and compare the three-dimensional structures of reference and alternative TF isoform–DNA complexes, using a dataset of 693 isoforms derived from 246 TF genes as characterized by Lambourne et al. (2025). The central finding is that structural confidence metrics from all three platforms are sensitive to isoform-specific sequence differences located outside the DNA-binding domain, and that the structural changes associated with IDR removal in alternative isoforms are broadly consistent with the autoinhibitory model of IDR-mediated DNA-binding regulation proposed by Lambourne et al. (2025). At the same time, no individual structural confidence metric from any platform reliably predicts experimental protein–DNA binding outcomes in a zero-shot setting, a finding that reflects the absence of task-specific training on experimental binding data and is itself biologically interpretable as discussed below.

The benchmarking of reference TF–DNA complexes revealed distinct and architecturally interpretable performance profiles for each platform. Boltz-2 demonstrated the highest and most consistent interface confidence across the full reference library, a pattern consistent with its diffusion-based architecture, which iteratively refines atomic coordinates toward energetically plausible configurations and tends to converge on high-confidence interface geometries even in the absence of close structural homologs. While this consistency is a strength in terms of prediction reliability, it also raises an important consideration for interpretation. The uniformly high Boltz-2 interface confidence values leave little dynamic range for discriminating between isoforms of differing binding competency, a limitation demonstrated directly by the ROC analysis in section 3.4.

RF2NA showed the most consistent local structural confidence across reference complexes, reflecting the evolutionary information integrated through its three-track architecture, which draws on multiple sequence alignments to constrain local structural predictions within biologically plausible bounds. However, the failure of RF2NA to achieve reliable inter-domain positional accuracy across the vast majority of complexes in this library is a significant finding that likely reflects the relatively sparse

representation of protein–nucleic acid complex structures in the Protein Data Bank compared to protein-only structures [98]. This sparsity constrains the ability of any model trained predominantly on protein structures to generalize across the full breadth of human TF–DNA interface diversity, and it suggests that RF2NA performs most reliably for TF families with well-characterized structural homologs in its training data, as evidenced by the reliable predictions obtained for specific TF families in the threshold analysis.

AF3 showed the broadest variability across all three structural confidence metrics, consistent with its generalist unified architecture, which was designed to model virtually all molecular interaction types within a single framework rather than being optimized specifically for protein–nucleic acid systems. In practice, this generalist design means that AF3 predictions are more discriminating than those of Boltz-2, producing a wider distribution that better reflects genuine differences in the structural tractability of individual TF–DNA complexes. This greater dynamic range ultimately makes AF3 the most informative platform for comparative isoform analysis, and is consistent with the finding that AF3 interface confidence is the strongest individual predictor of experimental binding outcomes. Taken together, the benchmarking results demonstrate that no single platform dominates across all evaluation criteria. The consistency of Boltz-2, the local structural accuracy of RF2NA, and the discriminative dynamic range of AF3 represent complementary rather than redundant contributions, and together they justify the multi-platform approach adopted in this study.

The family-level analysis of AF3 predictions revealed that modeling success is not uniform across TF structural classes. Families with structurally well-characterized and conserved DNA-binding interfaces, such as Homeodomain and bHLH, achieved consistently high rates of reliable predictions, while more structurally diverse families such as Nuclear receptor showed a considerably more balanced distribution of successful and unsuccessful predictions. The IRF family showed the most unfavorable proportional success rate among the better-represented families, suggesting that IRF–DNA complexes may present particular modeling challenges for AF3 that warrant further investigation. These family-level differences have practical implications for researchers applying AF3 to TF structural prediction: the platform should be applied with greater caution to structurally diverse families, and its predictions for these families should be interpreted alongside additional experimental evidence.

The analysis of DBD-intact alternative isoforms revealed that all three platforms model

these variants with structural confidence levels broadly comparable to those observed for reference isoforms, confirming that the platforms do not systematically penalize alternative isoforms simply because they are non-canonical. The stability of local folding confidence across all three platforms between reference and alternative groups confirms that preserving the DBD sequence is sufficient to maintain local structural integrity regardless of the sequence changes that occur in the flanking regions.

The more informative finding is the consistent tendency toward lower inter-domain positional error in alternative isoforms compared to their reference counterparts, observed across all three platforms though most pronounced in Boltz-2. A reduction in predicted inter-domain positional error in the alternative form, despite the DBD sequence being identical to that of the reference, suggests that these changes reflect the structural consequences of IDR removal in the regions flanking the DBD rather than changes within the binding domain itself. This observation is consistent with the autoinhibitory model proposed by Lambourne et al. (2025), in which IDRs present in reference isoforms physically engage the DBD through intramolecular contacts that restrict the domain's conformational freedom. Removing these IDRs through alternative splicing would therefore be expected to release the DBD from intramolecular constraint and allow it to adopt a more precisely organized conformation in the context of the DNA complex, and the inter-domain positional error reductions we observe across all three platforms are consistent with this prediction. The asymmetric distribution of paired local folding confidence differences, where more alternative isoforms gain structural confidence than lose it when compared to their reference counterparts, further supports this interpretation at the population level and would not be expected if IDR removal were a structurally neutral event.

An important distinction must be drawn between the utility of Boltz-2 inter-domain positional error as a within-gene comparative metric and its utility as an across-isoform predictor of experimental binding outcomes. While Boltz-2 shows the greatest sensitivity to structural changes associated with IDR removal in the paired analysis, it performs substantially below chance as a standalone predictor of experimental binding in the ROC analysis, with a positive rather than negative correlation with binding outcomes. This apparent contradiction is resolved by recognizing that these two analyses address fundamentally different biological questions. The paired delta analysis evaluates relative structural change within matched isoform pairs from the same gene, where Boltz-2 inter-domain positional error is sensitive to the compactness changes

associated with IDR removal. The ROC analysis evaluates absolute values as predictors of binding across structurally heterogeneous isoforms from diverse TF families, where higher inter-domain positional error is associated with larger and more disordered proteins that are more likely to produce detectable interactions in eY1H assays regardless of their structural precision. Boltz-2 inter-domain positional error is therefore most informative when used in a paired comparative framework where the baseline structural context is controlled, and least informative when used as a standalone predictor across structurally heterogeneous isoforms from different TF families.

The eight isoforms identified through the three-stage experimental filter, where discordant eY1H binding outcomes, motif-bait overlap, and zero DBD loss converged on the same set of cases, provided an opportunity to connect structural modeling results directly to experimental evidence of functional divergence. The dominant trend across these isoforms was one of structural refinement in the alternative form relative to the reference, with the majority showing gains in local folding confidence and reductions in inter-domain positional error in at least two of the three platforms. The consistency of this trend across three independently built prediction frameworks strengthens confidence that it reflects genuine biological signal rather than modeling noise, and it is most naturally interpreted within the autoinhibitory framework proposed by Lambourne et al. (2025) [79].

DLX4-2 and PRRX1-3 emerged as the most computationally robust candidates in the dataset, showing concordant structural stabilization across all three platforms and all three metrics without a single contradictory signal. For DLX4-2 specifically, this computational evidence aligns directly with the experimental and computational findings of Lambourne et al. (2025), who demonstrated that the reference isoform carries strong putative intramolecular contacts between its N-terminal IDR and its homeodomain DNA-binding residues, and that the alternative isoform, which lacks this IDR, shows considerably broader and less sequence-specific DNA binding. The structural stabilization we observe computationally is therefore consistent with IDR removal releasing the homeodomain from intramolecular constraint and allowing it to engage DNA more broadly and with greater conformational precision. The same reasoning applies directly to PKNOX1-3, for which Lambourne et al. (2025) [79] reported analogous experimental findings. It is important to note that while the autoinhibitory model applies directly and with experimental support to DLX4-2 and

PKNOX1-3, the same mechanism has not been directly tested for the remaining six isoforms in this dataset. The broader observation from Lambourne et al. (2025) that IDR differences outside the DBD frequently drive changes in DNA-binding behavior even when the binding domain is intact provides a consistent biological framework for interpreting the structural stabilization we observe more widely, while acknowledging that the specific physical basis may differ between isoforms and that experimental follow-up on each individual case would be required to establish the mechanism with confidence.

TCF4-8 presents a distinct structural profile that warrants specific consideration: despite gaining local folding confidence and reducing inter-domain positional error, its predicted binding interface becomes slightly less precisely resolved in the alternative form across both platforms that report this metric. This suggests that the structural reorganization in TCF4-8 may extend closer to the actual DNA contact surface than in the other isoforms, potentially reflecting a more substantive remodeling of DBD positioning relative to DNA rather than simply a relaxation of flanking IDR constraint. GATA3-2 produced the most pronounced platform-specific signal in the dataset, with Boltz-2 predicting a substantially larger inter-domain positional improvement than either RF2NA or AF3. Whether this represents a genuine structural feature that Boltz-2 is uniquely resolving or a platform-specific artifact of its generative architecture cannot be determined from structural predictions alone, and this isoform represents a priority for future experimental structural investigation. KLF7-3 was the most internally inconsistent isoform across all three platforms and the most difficult to interpret structurally in this dataset, making it the highest priority for experimental structural validation.

The analysis of alternative isoforms with measurable DBD loss revealed a clear and interpretable pattern: the structural consequences of domain truncation are most sensitively captured by AF3 interface confidence, which showed a near-complete collapse when the canonical binding domain is absent. This collapse is mechanistically expected, as the DBD provides the primary structural scaffold for protein–DNA contact, and its removal eliminates the structural basis for a confident interface prediction. From a practical standpoint, this finding suggests that AF3 interface confidence could serve as a useful computational indicator for identifying alternative isoforms that have lost functional DNA-binding capacity, complementing sequence-based annotation of DBD integrity with a structural readout that is sensitive

to the consequences of domain loss at the interface level. The clinical relevance of DBD loss is illustrated by the finding of Lambourne et al. (2025) that an alternative isoform of the tumor suppressor SMAD3, which carries partial DBD loss and loses interaction with its known binding partner FOXH1, is significantly upregulated in lung adenocarcinoma tumors while the reference isoform is downregulated. This example demonstrates that structural disruption of the DNA-binding interface through alternative splicing can have direct oncogenic consequences, and suggests that the AF3 interface confidence collapse we observe for DBD-truncated isoforms may serve as a structural marker for identifying alternative isoforms with potentially oncogenic properties.

The behavior of Boltz-2 in the context of DBD loss presents an important interpretive challenge. Despite the removal of the canonical binding domain, Boltz-2 maintained relatively high interface confidence and showed a substantial reduction in inter-domain positional error, suggesting that the truncated isoforms adopt apparently compact and well-organized configurations complex with DNA. A more mechanistically plausible explanation is that Boltz-2, with its diffusion-based generative architecture, continues to converge on structurally plausible configurations even when the canonical protein–DNA interface is absent. This interpretation reinforces the earlier conclusion that Boltz-2 confidence scores should be interpreted with caution when the structural context departs substantially from that of a canonical TF–DNA complex, and it highlights the importance of using multiple platforms and multiple metrics when evaluating predictions for structurally atypical isoforms.

The ROC and correlation analyses demonstrated that none of the individual structural confidence metrics from any of the three platforms provides reliable prediction of experimental protein–DNA binding outcomes when applied without task-specific training. The primary reason is that none of the three platforms were trained on experimental TF–DNA binding data from the target system, which fundamentally limits their ability to discriminate between binding-competent and non-binding isoforms in a zero-shot setting.

The pattern of deviations from expectation in the correlation analysis is itself biologically informative. The negative local folding confidence correlations with experimental binding across all three platforms are interpretable within the autoinhibitory framework: reference isoforms with tightly constrained, high-confidence structures are precisely those whose DBD access is most restricted by intramolecular

IDR contacts, making them the least likely to show broad experimental binding in the eY1H assay. High structural confidence in the reference isoform therefore reflects autoinhibitory constraint rather than binding competency in this dataset. The positive inter-domain positional error correlations similarly reflect the role of protein disorder as a determinant of binding promiscuity in the eY1H assay, where larger and more conformationally flexible proteins with higher PAE values are more likely to produce detectable interactions across a range of binding configurations. AF3 interface confidence is the only metric that shows the expected direction of association with experimental binding in both Pearson and Spearman analyses, indicating that interface-specific confidence captures something genuine about the likelihood of protein–DNA interaction that global structural metrics do not. The improvement in predictive performance achieved by the combined multi-platform logistic regression model, reaching an AUC of 0.730 compared to the best individual metric of 0.602, demonstrates that the three platforms capture partially complementary structural information whose integration provides greater discriminatory power than any single metric alone. Given the sample size of 79 unique isoforms and eight predictors, this combined AUC should be treated as an upper bound on predictive performance rather than a reliable estimate of generalization to new datasets.

Several limitations of this study should be acknowledged. All three prediction platforms operate in a zero-shot setting with respect to TF–DNA binding, meaning that none of them were trained on experimental binding data from the target system. Task-specific fine-tuning on experimentally validated TF–DNA binding datasets would likely substantially improve predictive performance, though such datasets of sufficient scale and diversity do not currently exist for the breadth of TF families represented in this study.

Structural predictions are inherently static snapshots that capture a single low-energy conformation of the complex rather than the dynamic ensemble of states that a protein occupies in solution. IDRs are by definition conformationally heterogeneous, and their regulatory effects on DBD accessibility are likely mediated through transient intramolecular contacts rather than a single fixed conformation. The structural stabilization we observe in alternative isoforms upon IDR removal should therefore be interpreted as reflecting a thermodynamic tendency rather than a discrete conformational switch.

The autoinhibitory model has been directly tested only for DLX4 and PKNOX1 by

Lambourne et al. (2025) [79], and we apply it as a broader interpretive framework for the remaining priority isoforms and for the full population of DBD-intact alternative isoforms. While the population-level structural trends are consistent with this model, the specific intramolecular contacts and energetic contributions responsible for the observed structural changes have not been characterized for most isoforms, and experimental validation of the mechanism in each individual case would be required before drawing firm mechanistic conclusions.

The eY1H assay used to generate the experimental binding data has known limitations, including sensitivity to protein expression levels, potential for false positives from non-specific binding to the long bait sequences used in the assay, and false negatives arising from isoforms that bind only weakly or transiently. A specific and experimentally documented example of this sensitivity limitation is provided by Lambourne et al. (2025), who demonstrated that an alternative isoform of CREB1 retains transcriptional activity confirming proper expression and folding, yet shows complete loss of eY1H binding despite maintaining an intact bZIP DBD. Protein-binding microarray experiments revealed that this isoform has subtly lower DNA affinity than the reference, with the difference being sufficient to drop binding signals below the eY1H detection threshold. This demonstrates that eY1H-based negative classifications do not necessarily indicate an inability to bind DNA but may reflect affinity reductions below assay sensitivity, a limitation that would reduce the apparent predictive accuracy of structural confidence metrics for reasons independent of the structural predictions themselves.

The combined logistic regression model was trained and evaluated on the same dataset of 79 unique isoforms using eight predictors, introducing a meaningful risk of overfitting. Validation in an independent dataset of comparable size and experimental quality would be needed to establish the true predictive utility of the combined model. Finally, no experimental structural validation of any of the predicted complexes was performed in this study, and the degree to which the predicted structures accurately represent the true three-dimensional geometry of the TF–DNA complexes in solution remains to be established.

Most immediately, the eight isoforms identified through the three-stage experimental filter represent strong candidates for experimental structural characterization by cryo-electron microscopy or NMR spectroscopy. Structural determination of the DLX4-2 and PKNX1-3 complexes in particular would provide the first direct

atomic-resolution test of the autoinhibitory model proposed by Lambourne et al. (2025), and would establish whether the structural stabilization predicted computationally corresponds to a genuine conformational change in the experimental structure.

Molecular dynamics simulations of the reference and alternative isoform complexes would complement the static structural predictions by capturing the dynamic behavior of IDRs and their transient intramolecular contacts with the DBD. This approach would be particularly valuable for isoforms such as KLF7-3 and GATA3-2, where inter-platform disagreements reflect genuine structural ambiguity that static prediction cannot resolve, and for establishing the physical basis of the structural stabilization observed in DLX4-2 and PKNX1-3 at a mechanistic resolution that static structural prediction alone cannot achieve.

The TBX5 isoforms described by Lambourne et al. (2025) illustrate an important limitation of the current approach. An alternative isoform differing only in its N-terminal disordered region retains binary binding status but shows significantly enhanced affinity for a subset of moderate-affinity binding sequences enriched at physiological TBX5 binding sites in cardiomyocytes. This demonstrates that IDR-mediated effects on DNA binding can manifest as quantitative affinity shifts rather than binary changes in binding status, a level of resolution that the current structural confidence metrics cannot capture. Extending the structural pipeline to incorporate quantitative affinity prediction using the Boltz-2 affinity module, and validating against protein-binding microarray data for isoforms such as TBX5-2, would represent a direct step toward capturing this level of functional resolution.

From a computational methodology perspective, the most impactful future development would be task-specific fine-tuning of one or more platforms on experimentally validated TF–DNA binding datasets. The zero-shot performance ceiling established in this study defines the upper bound of what is currently achievable without task-specific training, and fine-tuning on binding data would enable these platforms to move from structural plausibility assessment toward genuine binding competency prediction. As large-scale experimental resources such as the TFIso collection from Lambourne et al. (2025) continue to grow, this approach becomes increasingly feasible.

More broadly, the computational pipeline developed in this study could be extended to

the full set of alternative isoforms in the Lambourne et al. (2025) [79] dataset and applied to disease-associated TF isoforms identified from tumor transcriptome data. Given that nearly 60% of isoforms differentially expressed in tumors are alternative variants (Lambourne et al., 2025), a structural understanding of how these variants differ from their reference isoforms at the protein–DNA interface could provide mechanistic insight into their oncogenic or tumor-suppressive properties and potentially identify new targets at the level of TF isoform-specific interactions.

5 Conclusion

This study demonstrates that deep learning-based structural prediction frameworks can detect isoform-specific structural differences that are consistent with experimental evidence of functional divergence in TF–DNA binding, even when the DNA-binding domain itself is completely preserved between isoforms. The structural stabilization observed in DBD-intact alternative isoforms across three independent platforms, combined with the near-complete collapse of AF3 interface confidence in DBD-truncated isoforms, confirms that these tools are sensitive to the structural consequences of alternative splicing at both the local and interface level. The family-level analysis of AF3 predictions further revealed that modeling success varies considerably across TF structural classes, with well-characterized families achieving near-complete reliability and more structurally diverse families showing greater variability, a finding with practical implications for the application of these tools in future structural studies of TF biology.

While no individual structural metric currently provides reliable standalone prediction of experimental binding outcomes in a zero-shot setting, the integrated multi-platform logistic regression model achieves a practically meaningful level of predictive performance, establishing a baseline that future task-specific fine-tuning approaches can build upon. The counterintuitive correlation patterns observed between structural metrics and experimental binding outcomes are themselves biologically informative rather than representing failures of the metrics. They reflect the autoinhibitory structural constraints that characterize reference isoforms in this dataset and the role of protein disorder as a determinant of binding promiscuity in the eY1H assay, providing mechanistic insight that extends beyond what a purely predictive approach would reveal.

Collectively, these findings advance our understanding of how alternative splicing reshapes TF–DNA interfaces at the structural level and benchmarked multi-platform computational framework. As experimental datasets characterizing isoform-specific binding continue to grow and as structural prediction tools are refined through task-specific training, the approach developed here is well positioned to contribute to a mechanistic understanding of how the human TF isoform landscape encodes regulatory complexity beyond what the reference proteome alone can capture.

6 References

1. Felsenfeld, G. and M. Groudine, *Controlling the double helix*. Nature, 2003. **421**(6921): p. 448-53.
2. Talbert, P.B. and S. Henikoff, *Histone variants at a glance*. J Cell Sci, 2021. **134**(6).
3. Venters, B.J. and B.F. Pugh, *How eukaryotic genes are transcribed*. Crit Rev Biochem Mol Biol, 2009. **44**(2-3): p. 117-41.
4. Casamassimi, A. and A. Ciccodicola, *Transcriptional Regulation: Molecules, Involved Mechanisms, and Misregulation*. International Journal of Molecular Sciences, 2019. **20**(6): p. 1281.
5. Vannini, A. and P. Cramer, *Conservation between the RNA polymerase I, II, and III transcription initiation machineries*. Mol Cell, 2012. **45**(4): p. 439-46.
6. Malik, S. and R.G. Roeder, *Regulation of the RNA polymerase II pre-initiation complex by its associated coactivators*. Nat Rev Genet, 2023. **24**(11): p. 767-782.
7. Zhang, Z., et al., *Rapid dynamics of general transcription factor TFIIB binding during preinitiation complex assembly revealed by single-molecule analysis*. Genes Dev, 2016. **30**(18): p. 2106-2118.
8. Chen, H.T., L. Warfield, and S. Hahn, *The positions of TFIIF and TFIIE in the RNA polymerase II transcription preinitiation complex*. Nat Struct Mol Biol, 2007. **14**(8): p. 696-703.
9. Fishburn, J., et al., *Double-stranded DNA translocase activity of transcription factor TFIIF and the mechanism of RNA polymerase II open complex formation*. Proc Natl Acad Sci U S A, 2015. **112**(13): p. 3961-6.
10. Giordano, G., et al., *Uncoupling the TFIIF Core and Kinase Modules Leads To Misregulated RNA Polymerase II CTD Serine 5 Phosphorylation*. bioRxiv, 2024.
11. Luse, D.S., *Promoter clearance by RNA polymerase II*. Biochim Biophys Acta, 2013. **1829**(1): p. 63-8.
12. Hsin, J.P. and J.L. Manley, *The RNA polymerase II CTD coordinates transcription and RNA processing*. Genes Dev, 2012. **26**(19): p. 2119-37.
13. Guo, J. and D.H. Price, *RNA Polymerase II Transcription Elongation Control*. Chemical Reviews, 2013. **113**(11): p. 8583-8603.
14. Mischo, H.E. and N.J. Proudfoot, *Disengaging polymerase: terminating RNA polymerase II transcription in budding yeast*. Biochim Biophys Acta, 2013. **1829**(1): p. 174-85.
15. Casamassimi, A. and A. Ciccodicola, *Transcriptional Regulation: Molecules, Involved Mechanisms, and Misregulation*. Int J Mol Sci, 2019. **20**(6).
16. Morimoto, R.I. and E.A.A. Nollen. *CHAPTER 310 – The Heat-Shock Response: Sensing the Stress of Misfolded Proteins*. 2003.
17. Sperling, S., *Transcriptional regulation at a glance*. BMC Bioinformatics, 2007. **8 Suppl 6**(Suppl 6): p. S2.
18. Fry, C.J. and P.J. Farnham, *Context-dependent transcriptional regulation*. J Biol Chem, 1999. **274**(42): p. 29583-6.
19. Tkačik, G., C.G. Callan, and W. Bialek, *Information flow and optimization in transcriptional regulation*. Proceedings of the National Academy of Sciences, 2008. **105**(34): p. 12265-12270.
20. Wray, G.A., et al., *The evolution of transcriptional regulation in eukaryotes*. Mol Biol Evol, 2003. **20**(9): p. 1377-419.
21. Farnung, L. and S.M. Vos, *Assembly of RNA polymerase II transcription initiation complexes*. Curr Opin Struct Biol, 2022. **73**: p. 102335.
22. Kaluscha, S., et al., *Evidence that direct inhibition of transcription factor binding is the prevailing mode of gene and repeat repression by DNA methylation*. Nat Genet, 2022. **54**(12): p. 1895-1906.

23. Furlan, M., S. de Pretis, and M. Pelizzola, *Dynamics of transcriptional and post-transcriptional regulation*. Brief Bioinform, 2021. **22**(4).
24. Franks, A., E. Airoidi, and N. Slavov, *Post-transcriptional regulation across human tissues*. PLoS Comput Biol, 2017. **13**(5): p. e1005535.
25. Spitz, F. and E.E. Furlong, *Transcription factors: from enhancer binding to developmental control*. Nat Rev Genet, 2012. **13**(9): p. 613-26.
26. Lambert, S.A., et al., *The Human Transcription Factors*. Cell, 2018. **172**(4): p. 650-665.
27. Oksuz, O., et al., *Transcription factors interact with RNA to regulate genes*. Molecular Cell, 2023. **83**(14): p. 2449-2463.e13.
28. Lee, T.I. and R.A. Young, *Transcriptional regulation and its misregulation in disease*. Cell, 2013. **152**(6): p. 1237-51.
29. Singh, H., A.A. Khan, and A.R. Dinner, *Gene regulatory networks in the immune system*. Trends Immunol, 2014. **35**(5): p. 211-8.
30. Kim, S. and J. Wysocka, *Deciphering the multi-scale, quantitative cis-regulatory code*. Mol Cell, 2023. **83**(3): p. 373-392.
31. Cosma, M.P., *Ordered recruitment: gene-specific mechanism of transcription activation*. Mol Cell, 2002. **10**(2): p. 227-36.
32. Zheng, J., et al., *Epistatic relationships reveal the functional organization of yeast transcription factors*. Mol Syst Biol, 2010. **6**: p. 420.
33. Oksuz, O., et al., *Transcription factors interact with RNA to regulate genes*. Mol Cell, 2023. **83**(14): p. 2449-2463.e13.
34. Brent, R. and M. Ptashne, *A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor*. Cell, 1985. **43**(3, Part 2): p. 729-736.
35. Pabo, C.O. and R.T. Sauer, *Transcription factors: structural families and principles of DNA recognition*. Annu Rev Biochem, 1992. **61**: p. 1053-95.
36. Mermod, N., et al., *The proline-rich transcriptional activator of CTF/NF-I is distinct from the replication and DNA binding domain*. Cell, 1989. **58**(4): p. 741-753.
37. Courey, A.J. and R. Tjian, *Analysis of Sp1 in vivo reveals multiple transcriptional domains, including a novel glutamine-rich activation motif*. Cell, 1988. **55**(5): p. 887-98.
38. Van Hoy, M., et al., *The acidic activation domains of the GCN4 and GAL4 proteins are not α helical but form β sheets*. Cell, 1993. **72**(4): p. 587-594.
39. Jenster, G., *Coactivators and corepressors as mediators of nuclear receptor function: an update*. Mol Cell Endocrinol, 1998. **143**(1-2): p. 1-7.
40. Bhattacharjee, S., et al., *Combinatorial control of gene expression*. Biomed Res Int, 2013. **2013**: p. 407263.
41. Gonzalez, G.A. and M.R. Montminy, *Cyclic AMP stimulates somatostatin gene transcription by phosphorylation of CREB at serine 133*. Cell, 1989. **59**(4): p. 675-80.
42. Ray-Jones, H. and M. Spivakov, *Transcriptional enhancers and their communication with gene promoters*. Cell Mol Life Sci, 2021. **78**(19-20): p. 6453-6485.
43. Badia, I.M.P., et al., *Gene regulatory network inference in the era of single-cell multi-omics*. Nat Rev Genet, 2023. **24**(11): p. 739-754.
44. Zaret, K.S. and J.S. Carroll, *Pioneer transcription factors: establishing competence for gene expression*. Genes Dev, 2011. **25**(21): p. 2227-41.
45. Avsec, Ž., et al., *Base-resolution models of transcription-factor binding reveal soft motif syntax*. Nat Genet, 2021. **53**(3): p. 354-366.
46. Panne, D., T. Maniatis, and S.C. Harrison, *An atomic model of the interferon-beta enhanceosome*. Cell, 2007. **129**(6): p. 1111-23.
47. Levitsky, V.G., et al., *Asymmetry of Motif Conservation Within Their Homotypic Pairs Distinguishes DNA-Binding Domains of Target Transcription Factors in ChIP-Seq Data*. Int J Mol Sci, 2025. **26**(1).
48. Soto, L.F., et al., *Compendium of human transcription factor effector domains*. Molecular Cell, 2022. **82**(3): p. 514-526.

49. Vaquerizas, J.M., et al., *A census of human transcription factors: function, expression and evolution*. Nat Rev Genet, 2009. **10**(4): p. 252-63.
50. Henninger, J.E. and R.A. Young, *An RNA-centric view of transcription and genome organization*. Mol Cell, 2024. **84**(19): p. 3627-3643.
51. Már, M., K. Nitsenko, and P.O. Heidarrsson, *Multifunctional Intrinsically Disordered Regions in Transcription Factors*. Chemistry – A European Journal, 2023. **29**(21): p. e202203369.
52. Jeyaraj, G., et al., *High-resolution protein modeling through Cryo-EM and AI: current trends and future perspectives - a review*. Front Mol Biosci, 2025. **12**: p. 1688455.
53. Inukai, S., K.H. Kock, and M.L. Bulyk, *Transcription factor-DNA binding: beyond binding site motifs*. Curr Opin Genet Dev, 2017. **43**: p. 110-119.
54. Martinez-Corral, R., et al., *Emergence of activation or repression in transcriptional control under a fixed molecular context*. bioRxiv, 2024.
55. Lambourne, L., et al., *Widespread variation in molecular interactions and regulatory properties among transcription factor isoforms*. bioRxiv, 2024.
56. Fuxman Bass, J.I., et al., *Human gene-centered transcription factor networks for enhancers and disease variants*. Cell, 2015. **161**(3): p. 661-673.
57. Awdeh, A., M. Turcotte, and T.J. Perkins, *Identifying transcription factors with cell-type specific DNA binding signatures*. BMC Genomics, 2024. **25**(1): p. 957.
58. Schnepf, M., et al., *Transcription Factor Binding Affinities and DNA Shape Readout*. iScience, 2020. **23**(11): p. 101694.
59. de Jonge, W.J., et al., *Following the tracks: How transcription factor binding dynamics control transcription*. Biophys J, 2022. **121**(9): p. 1583-1592.
60. Bauer, M., et al., *Real sequence effects on the search dynamics of transcription factors on DNA*. Scientific Reports, 2015. **5**(1): p. 10072.
61. Suter, D.M., *Transcription Factors and DNA Play Hide and Seek*. Trends Cell Biol, 2020. **30**(6): p. 491-500.
62. Luscombe, N.M., R.A. Laskowski, and J.M. Thornton, *Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level*. Nucleic Acids Res, 2001. **29**(13): p. 2860-74.
63. Yesudhas, D., et al., *Proteins Recognizing DNA: Structural Uniqueness and Versatility of DNA-Binding Domains in Stem Cell Transcription Factors*. Genes, 2017. **8**(8): p. 192.
64. Krieger, G., et al., *Evolution of transcription factor binding through sequence variations and turnover of binding sites*. Genome Res, 2022. **32**(6): p. 1099-1111.
65. Sandelin, A., et al., *JASPAR: an open-access database for eukaryotic transcription factor binding profiles*. Nucleic Acids Res, 2004. **32**(Database issue): p. D91-4.
66. Ambrosini, G., et al., *Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study*. Genome Biology, 2020. **21**(1): p. 114.
67. Rauluseviciute, I., et al., *JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles*. Nucleic Acids Research, 2024. **52**(D1): p. D174-D182.
68. Vorontsov, I.E., et al., *HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors*. Nucleic Acids Res, 2024. **52**(D1): p. D154-d163.
69. Kulakovskiy, I.V., et al., *HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis*. Nucleic Acids Res, 2018. **46**(D1): p. D252-d259.
70. Yeo, G., et al., *Variation in alternative splicing across human tissues*. Genome Biology, 2004. **5**(10): p. R74.
71. Ellis, J.D., et al., *Tissue-specific alternative splicing remodels protein-protein interaction networks*. Mol Cell, 2012. **46**(6): p. 884-92.
72. Nilsen, T.W. and B.R. Graveley, *Expansion of the eukaryotic proteome by alternative splicing*. Nature, 2010. **463**(7280): p. 457-63.
73. Braunschweig, U., et al., *Dynamic integration of splicing within gene regulatory pathways*.

Cell, 2013. **152**(6): p. 1252-69.

74. Blake, D. and K.W. Lynch, *The three as: Alternative splicing, alternative polyadenylation and their impact on apoptosis in immune function*. Immunol Rev, 2021. **304**(1): p. 30-50.
75. Lambourne, L., et al., *Widespread variation in molecular interactions and regulatory properties among transcription factor isoforms*. Molecular Cell, 2025. **85**(7): p. 1445-1466.e13.
76. Belluti, S., G. Rigillo, and C. Imbriano, *Transcription Factors in Cancer: When Alternative Splicing Determines Opposite Cell Fates*. Cells, 2020. **9**(3).
77. Sapio, L., et al., *Targeting CREB in Cancer Therapy: A Key Candidate or One of Many? An Update*. Cancers (Basel), 2020. **12**(11).
78. Aragón, E., et al., *Structural basis for distinct roles of SMAD2 and SMAD3 in FOXH1 pioneer-directed TGF- β signaling*. Genes Dev, 2019. **33**(21-22): p. 1506-1524.
79. López, A.J., *Developmental role of transcription factor isoforms generated by alternative splicing*. Dev Biol, 1995. **172**(2): p. 396-411.
80. Zammarchi, F., et al., *Antitumorigenic potential of STAT3 alternative splicing modulation*. Proc Natl Acad Sci U S A, 2011. **108**(43): p. 17779-84.
81. Zhang, H.X., et al., *STAT3beta, a distinct isoform from STAT3*. Int J Biochem Cell Biol, 2019. **110**: p. 130-139.
82. Schaefer, T.S., et al., *Functional differences between Stat3 α and Stat3 β* . Molecular and cellular biology, 1997. **17**(9): p. 5307-5316.
83. Ma, L., et al., *A critical role for the loop region of the basic helix-loop-helix/leucine zipper protein Mlx in DNA binding and glucose-regulated transcription*. Nucleic Acids Res, 2007. **35**(1): p. 35-44.
84. Englert, C., et al., *Truncated WT1 mutants alter the subnuclear localization of the wild-type protein*. Proc Natl Acad Sci U S A, 1995. **92**(26): p. 11960-4.
85. Larsson, S.H., et al., *Subnuclear localization of WT1 in splicing or transcription factor domains is regulated by alternative splicing*. Cell, 1995. **81**(3): p. 391-401.
86. Siggers, T., et al., *Diversification of transcription factor paralogs via noncanonical modularity in C2H2 zinc finger DNA binding*. Mol Cell, 2014. **55**(4): p. 640-8.
87. Eddy, S.R., *Accelerated Profile HMM Searches*. PLoS Comput Biol, 2011. **7**(10): p. e1002195.
88. Berger, M.F., et al., *Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities*. Nat Biotechnol, 2006. **24**(11): p. 1429-35.
89. He, F., et al., *Interaction between p53 N terminus and core domain regulates specific and nonspecific DNA binding*. Proc Natl Acad Sci U S A, 2019. **116**(18): p. 8859-8868.
90. Krois, A.S., H.J. Dyson, and P.E. Wright, *Long-range regulation of p53 DNA binding by its intrinsically disordered N-terminal transactivation domain*. Proc Natl Acad Sci U S A, 2018. **115**(48): p. E11302-e11310.
91. Bjarnason, S., et al., *DNA binding redistributes activation domain ensemble and accessibility in pioneer factor Sox2*. Biophysical Journal, 2024. **123**(3): p. 73a.
92. Steimle, J.D. and I.P. Moskowitz, *TBX5: A Key Regulator of Heart Development*. Curr Top Dev Biol, 2017. **122**: p. 195-221.
93. Helsem, S.A., et al., *Two different and robustly modeled DNA binding modes of Competence Protein ComP - systematic modeling with AlphaFold 3, RoseTTAFold2NA, Chai-1 and re-docking in HADDOCK*. PLOS ONE, 2025. **20**(5): p. e0315160.
94. Baek, M., et al., *Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA*. Nature Methods, 2024. **21**(1): p. 117-121.
95. Abramson, J., et al., *Accurate structure prediction of biomolecular interactions with AlphaFold 3*. Nature, 2024. **630**(8016): p. 493-500.
96. Passaro, S., et al., *Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction*. bioRxiv, 2025: p. 2025.06.14.659707.
97. Schneider, T.D., et al., *Information content of binding sites on nucleotide sequences*. Journal

- of Molecular Biology, 1986. **188**(3): p. 415-431.
98. Stormo, G.D., *Modeling the specificity of protein-DNA interactions*. Quantitative Biology, 2013. **1**(2): p. 115-130.
 99. Schneider, T.D. and R.M. Stephens, *Sequence logos: a new way to display consensus sequences*. Nucleic Acids Res, 1990. **18**(20): p. 6097-100.
-
1. Felsenfeld, G. and M. Groudine, *Controlling the double helix*. Nature, 2003. **421**(6921): p. 448-53.
 2. Talbert, P.B. and S. Henikoff, *Histone variants at a glance*. J Cell Sci, 2021. **134**(6).
 3. Venters, B.J. and B.F. Pugh, *How eukaryotic genes are transcribed*. Crit Rev Biochem Mol Biol, 2009. **44**(2-3): p. 117-41.
 4. Casamassimi, A. and A. Ciccodicola, *Transcriptional Regulation: Molecules, Involved Mechanisms, and Misregulation*. International Journal of Molecular Sciences, 2019. **20**(6): p. 1281.
 5. Vannini, A. and P. Cramer, *Conservation between the RNA polymerase I, II, and III transcription initiation machineries*. Mol Cell, 2012. **45**(4): p. 439-46.
 6. Malik, S. and R.G. Roeder, *Regulation of the RNA polymerase II pre-initiation complex by its associated coactivators*. Nat Rev Genet, 2023. **24**(11): p. 767-782.
 7. Zhang, Z., et al., *Rapid dynamics of general transcription factor TFIIB binding during preinitiation complex assembly revealed by single-molecule analysis*. Genes Dev, 2016. **30**(18): p. 2106-2118.
 8. Chen, H.T., L. Warfield, and S. Hahn, *The positions of TFIIF and TFIIE in the RNA polymerase II transcription preinitiation complex*. Nat Struct Mol Biol, 2007. **14**(8): p. 696-703.
 9. Fishburn, J., et al., *Double-stranded DNA translocase activity of transcription factor TFIIH and the mechanism of RNA polymerase II open complex formation*. Proc Natl Acad Sci U S A, 2015. **112**(13): p. 3961-6.
 10. Giordano, G., et al., *Uncoupling the TFIIH Core and Kinase Modules Leads To Misregulated RNA Polymerase II CTD Serine 5 Phosphorylation*. bioRxiv, 2024.
 11. Luse, D.S., *Promoter clearance by RNA polymerase II*. Biochim Biophys Acta, 2013. **1829**(1): p. 63-8.
 12. Hsin, J.P. and J.L. Manley, *The RNA polymerase II CTD coordinates transcription and RNA processing*. Genes Dev, 2012. **26**(19): p. 2119-37.
 13. Guo, J. and D.H. Price, *RNA Polymerase II Transcription Elongation Control*. Chemical Reviews, 2013. **113**(11): p. 8583-8603.
 14. Mischo, H.E. and N.J. Proudfoot, *Disengaging polymerase: terminating RNA polymerase II transcription in budding yeast*. Biochim Biophys Acta, 2013. **1829**(1): p. 174-85.
 15. Casamassimi, A. and A. Ciccodicola, *Transcriptional Regulation: Molecules, Involved Mechanisms, and Misregulation*. Int J Mol Sci, 2019. **20**(6).
 16. Morimoto, R.I. and E.A.A. Nollen. *CHAPTER 310 – The Heat-Shock Response: Sensing the Stress of Misfolded Proteins*. 2003.
 17. Sperling, S., *Transcriptional regulation at a glance*. BMC Bioinformatics, 2007. **8 Suppl 6**(Suppl 6): p. S2.
 18. Fry, C.J. and P.J. Farnham, *Context-dependent transcriptional regulation*. J Biol Chem, 1999. **274**(42): p. 29583-6.
 19. Tkačik, G., C.G. Callan, and W. Bialek, *Information flow and optimization in transcriptional regulation*. Proceedings of the National Academy of Sciences, 2008. **105**(34): p. 12265-12270.
 20. Wray, G.A., et al., *The evolution of transcriptional regulation in eukaryotes*. Mol Biol Evol, 2003. **20**(9): p. 1377-419.

21. Farnung, L. and S.M. Vos, *Assembly of RNA polymerase II transcription initiation complexes*. Curr Opin Struct Biol, 2022. **73**: p. 102335.
22. Kaluscha, S., et al., *Evidence that direct inhibition of transcription factor binding is the prevailing mode of gene and repeat repression by DNA methylation*. Nat Genet, 2022. **54**(12): p. 1895-1906.
23. Furlan, M., S. de Pretis, and M. Pelizzola, *Dynamics of transcriptional and post-transcriptional regulation*. Brief Bioinform, 2021. **22**(4).
24. Franks, A., E. Airolidi, and N. Slavov, *Post-transcriptional regulation across human tissues*. PLoS Comput Biol, 2017. **13**(5): p. e1005535.
25. Spitz, F. and E.E. Furlong, *Transcription factors: from enhancer binding to developmental control*. Nat Rev Genet, 2012. **13**(9): p. 613-26.
26. Lambert, S.A., et al., *The Human Transcription Factors*. Cell, 2018. **172**(4): p. 650-665.
27. Oksuz, O., et al., *Transcription factors interact with RNA to regulate genes*. Molecular Cell, 2023. **83**(14): p. 2449-2463.e13.
28. Lee, T.I. and R.A. Young, *Transcriptional regulation and its misregulation in disease*. Cell, 2013. **152**(6): p. 1237-51.
29. Singh, H., A.A. Khan, and A.R. Dinner, *Gene regulatory networks in the immune system*. Trends Immunol, 2014. **35**(5): p. 211-8.
30. Kim, S. and J. Wysocka, *Deciphering the multi-scale, quantitative cis-regulatory code*. Mol Cell, 2023. **83**(3): p. 373-392.
31. Cosma, M.P., *Ordered recruitment: gene-specific mechanism of transcription activation*. Mol Cell, 2002. **10**(2): p. 227-36.
32. Zheng, J., et al., *Epistatic relationships reveal the functional organization of yeast transcription factors*. Mol Syst Biol, 2010. **6**: p. 420.
33. Oksuz, O., et al., *Transcription factors interact with RNA to regulate genes*. Mol Cell, 2023. **83**(14): p. 2449-2463.e13.
34. Brent, R. and M. Ptashne, *A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor*. Cell, 1985. **43**(3, Part 2): p. 729-736.
35. Pabo, C.O. and R.T. Sauer, *Transcription factors: structural families and principles of DNA recognition*. Annu Rev Biochem, 1992. **61**: p. 1053-95.
36. Mermod, N., et al., *The proline-rich transcriptional activator of CTF/NF-I is distinct from the replication and DNA binding domain*. Cell, 1989. **58**(4): p. 741-753.
37. Courey, A.J. and R. Tjian, *Analysis of Sp1 in vivo reveals multiple transcriptional domains, including a novel glutamine-rich activation motif*. Cell, 1988. **55**(5): p. 887-98.
38. Van Hoy, M., et al., *The acidic activation domains of the GCN4 and GAL4 proteins are not α helical but form β sheets*. Cell, 1993. **72**(4): p. 587-594.
39. Jenster, G., *Coactivators and corepressors as mediators of nuclear receptor function: an update*. Mol Cell Endocrinol, 1998. **143**(1-2): p. 1-7.
40. Bhattacharjee, S., et al., *Combinatorial control of gene expression*. Biomed Res Int, 2013. **2013**: p. 407263.
41. Gonzalez, G.A. and M.R. Montminy, *Cyclic AMP stimulates somatostatin gene transcription by phosphorylation of CREB at serine 133*. Cell, 1989. **59**(4): p. 675-80.
42. Gu, W. and R.G. Roeder, *Activation of p53 sequence-specific DNA binding by acetylation of the p53 C-terminal domain*. Cell, 1997. **90**(4): p. 595-606.
43. Molkenkint, J.D., *The zinc finger-containing transcription factors GATA-4, -5, and -6. Ubiquitously expressed regulators of tissue-specific gene expression*. J Biol Chem, 2000. **275**(50): p. 38949-52.
44. Ray-Jones, H. and M. Spivakov, *Transcriptional enhancers and their communication with gene promoters*. Cell Mol Life Sci, 2021. **78**(19-20): p. 6453-6485.
45. Badia, I.M.P., et al., *Gene regulatory network inference in the era of single-cell multi-omics*. Nat Rev Genet, 2023. **24**(11): p. 739-754.
46. Zaret, K.S. and J.S. Carroll, *Pioneer transcription factors: establishing competence for gene*

- expression*. Genes Dev, 2011. **25**(21): p. 2227-41.
47. Avsec, Ž., et al., *Base-resolution models of transcription-factor binding reveal soft motif syntax*. Nat Genet, 2021. **53**(3): p. 354-366.
 48. Panne, D., T. Maniatis, and S.C. Harrison, *An atomic model of the interferon-beta enhanceosome*. Cell, 2007. **129**(6): p. 1111-23.
 49. Levitsky, V.G., et al., *Asymmetry of Motif Conservation Within Their Homotypic Pairs Distinguishes DNA-Binding Domains of Target Transcription Factors in ChIP-Seq Data*. Int J Mol Sci, 2025. **26**(1).
 50. Soto, L.F., et al., *Compendium of human transcription factor effector domains*. Molecular Cell, 2022. **82**(3): p. 514-526.
 51. Vaquerizas, J.M., et al., *A census of human transcription factors: function, expression and evolution*. Nat Rev Genet, 2009. **10**(4): p. 252-63.
 52. Henninger, J.E. and R.A. Young, *An RNA-centric view of transcription and genome organization*. Mol Cell, 2024. **84**(19): p. 3627-3643.
 53. Már, M., K. Nitsenko, and P.O. Heidarrsson, *Multifunctional Intrinsically Disordered Regions in Transcription Factors*. Chemistry – A European Journal, 2023. **29**(21): p. e202203369.
 54. Jeyaraj, G., et al., *High-resolution protein modeling through Cryo-EM and AI: current trends and future perspectives - a review*. Front Mol Biosci, 2025. **12**: p. 1688455.
 55. Inukai, S., K.H. Kock, and M.L. Bulyk, *Transcription factor-DNA binding: beyond binding site motifs*. Curr Opin Genet Dev, 2017. **43**: p. 110-119.
 56. Martinez-Corral, R., et al., *Emergence of activation or repression in transcriptional control under a fixed molecular context*. bioRxiv, 2024.
 57. Lambourne, L., et al., *Widespread variation in molecular interactions and regulatory properties among transcription factor isoforms*. bioRxiv, 2024.
 58. Fuxman Bass, J.I., et al., *Human gene-centered transcription factor networks for enhancers and disease variants*. Cell, 2015. **161**(3): p. 661-673.
 59. Awdeh, A., M. Turcotte, and T.J. Perkins, *Identifying transcription factors with cell-type specific DNA binding signatures*. BMC Genomics, 2024. **25**(1): p. 957.
 60. Schnepf, M., et al., *Transcription Factor Binding Affinities and DNA Shape Readout*. iScience, 2020. **23**(11): p. 101694.
 61. de Jonge, W.J., et al., *Following the tracks: How transcription factor binding dynamics control transcription*. Biophys J, 2022. **121**(9): p. 1583-1592.
 62. Bauer, M., et al., *Real sequence effects on the search dynamics of transcription factors on DNA*. Scientific Reports, 2015. **5**(1): p. 10072.
 63. Suter, D.M., *Transcription Factors and DNA Play Hide and Seek*. Trends Cell Biol, 2020. **30**(6): p. 491-500.
 64. Luscombe, N.M., R.A. Laskowski, and J.M. Thornton, *Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level*. Nucleic Acids Res, 2001. **29**(13): p. 2860-74.
 65. Yesudhas, D., et al., *Proteins Recognizing DNA: Structural Uniqueness and Versatility of DNA-Binding Domains in Stem Cell Transcription Factors*. Genes, 2017. **8**(8): p. 192.
 66. Krieger, G., et al., *Evolution of transcription factor binding through sequence variations and turnover of binding sites*. Genome Res, 2022. **32**(6): p. 1099-1111.
 67. Sandelin, A., et al., *JASPAR: an open-access database for eukaryotic transcription factor binding profiles*. Nucleic Acids Res, 2004. **32**(Database issue): p. D91-4.
 68. Ambrosini, G., et al., *Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study*. Genome Biology, 2020. **21**(1): p. 114.
 69. Rauluseviciute, I., et al., *JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles*. Nucleic Acids Research, 2024. **52**(D1): p. D174-D182.
 70. Vorontsov, I.E., et al., *HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors*. Nucleic Acids Res, 2024. **52**(D1): p. D154-d163.

71. Kulakovskiy, I.V., et al., *HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis*. Nucleic Acids Res, 2018. **46**(D1): p. D252-d259.
72. Yeo, G., et al., *Variation in alternative splicing across human tissues*. Genome Biology, 2004. **5**(10): p. R74.
73. Ellis, J.D., et al., *Tissue-specific alternative splicing remodels protein-protein interaction networks*. Mol Cell, 2012. **46**(6): p. 884-92.
74. Nilsen, T.W. and B.R. Graveley, *Expansion of the eukaryotic proteome by alternative splicing*. Nature, 2010. **463**(7280): p. 457-63.
75. Braunschweig, U., et al., *Dynamic integration of splicing within gene regulatory pathways*. Cell, 2013. **152**(6): p. 1252-69.
76. Blake, D. and K.W. Lynch, *The three as: Alternative splicing, alternative polyadenylation and their impact on apoptosis in immune function*. Immunol Rev, 2021. **304**(1): p. 30-50.
77. Tress, M.L., F. Abascal, and A. Valencia, *Alternative Splicing May Not Be the Key to Proteome Complexity*. Trends Biochem Sci, 2017. **42**(2): p. 98-110.
78. Green, R.E., et al., *Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes*. Bioinformatics, 2003. **19 Suppl 1**: p. i118-21.
79. Lambourne, L., et al., *Widespread variation in molecular interactions and regulatory properties among transcription factor isoforms*. Molecular Cell, 2025. **85**(7): p. 1445-1466.e13.
80. Belluti, S., G. Rigillo, and C. Imbriano, *Transcription Factors in Cancer: When Alternative Splicing Determines Opposite Cell Fates*. Cells, 2020. **9**(3).
81. Sapio, L., et al., *Targeting CREB in Cancer Therapy: A Key Candidate or One of Many? An Update*. Cancers (Basel), 2020. **12**(11).
82. Aragón, E., et al., *Structural basis for distinct roles of SMAD2 and SMAD3 in FOXH1 pioneer-directed TGF- β signaling*. Genes Dev, 2019. **33**(21-22): p. 1506-1524.
83. López, A.J., *Developmental role of transcription factor isoforms generated by alternative splicing*. Dev Biol, 1995. **172**(2): p. 396-411.
84. Zammarchi, F., et al., *Antitumorigenic potential of STAT3 alternative splicing modulation*. Proc Natl Acad Sci U S A, 2011. **108**(43): p. 17779-84.
85. Zhang, H.X., et al., *STAT3beta, a distinct isoform from STAT3*. Int J Biochem Cell Biol, 2019. **110**: p. 130-139.
86. Schaefer, T.S., et al., *Functional differences between Stat3 α and Stat3 β* . Molecular and cellular biology, 1997. **17**(9): p. 5307-5316.
87. Ma, L., et al., *A critical role for the loop region of the basic helix-loop-helix/leucine zipper protein Mlx in DNA binding and glucose-regulated transcription*. Nucleic Acids Res, 2007. **35**(1): p. 35-44.
88. Englert, C., et al., *Truncated WT1 mutants alter the subnuclear localization of the wild-type protein*. Proc Natl Acad Sci U S A, 1995. **92**(26): p. 11960-4.
89. Larsson, S.H., et al., *Subnuclear localization of WT1 in splicing or transcription factor domains is regulated by alternative splicing*. Cell, 1995. **81**(3): p. 391-401.
90. Siggers, T., et al., *Diversification of transcription factor paralogs via noncanonical modularity in C2H2 zinc finger DNA binding*. Mol Cell, 2014. **55**(4): p. 640-8.
91. Eddy, S.R., *Accelerated Profile HMM Searches*. PLoS Comput Biol, 2011. **7**(10): p. e1002195.
92. Berger, M.F., et al., *Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities*. Nat Biotechnol, 2006. **24**(11): p. 1429-35.
93. He, F., et al., *Interaction between p53 N terminus and core domain regulates specific and nonspecific DNA binding*. Proc Natl Acad Sci U S A, 2019. **116**(18): p. 8859-8868.
94. Krois, A.S., H.J. Dyson, and P.E. Wright, *Long-range regulation of p53 DNA binding by its intrinsically disordered N-terminal transactivation domain*. Proc Natl Acad Sci U S A, 2018.

- 115**(48): p. E11302-e11310.
95. Bjarnason, S., et al., *DNA binding redistributes activation domain ensemble and accessibility in pioneer factor Sox2*. Biophysical Journal, 2024. **123**(3): p. 73a.
 96. Steimle, J.D. and I.P. Moskowitz, *TBX5: A Key Regulator of Heart Development*. Curr Top Dev Biol, 2017. **122**: p. 195-221.
 97. Helsem, S.A., et al., *Two different and robustly modeled DNA binding modes of Competence Protein ComP - systematic modeling with AlphaFold 3, RoseTTAFold2NA, Chai-1 and re-docking in HADDOCK*. PLOS ONE, 2025. **20**(5): p. e0315160.
 98. Baek, M., et al., *Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA*. Nature Methods, 2024. **21**(1): p. 117-121.
 99. Abramson, J., et al., *Accurate structure prediction of biomolecular interactions with AlphaFold 3*. Nature, 2024. **630**(8016): p. 493-500.
 100. Passaro, S., et al., *Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction*. bioRxiv, 2025.
 101. Passaro, S., et al., *Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction*. bioRxiv, 2025: p. 2025.06.14.659707.
 102. Schneider, T.D., et al., *Information content of binding sites on nucleotide sequences*. Journal of Molecular Biology, 1986. **188**(3): p. 415-431.
 103. Stormo, G.D., *Modeling the specificity of protein-DNA interactions*. Quant Biol, 2013. **1**(2): p. 115-130.
 104. Stormo, G.D., *Modeling the specificity of protein-DNA interactions*. Quantitative Biology, 2013. **1**(2): p. 115-130.
 105. Schneider, T.D. and R.M. Stephens, *Sequence logos: a new way to display consensus sequences*. Nucleic Acids Res, 1990. **18**(20): p. 6097-100.

Acknowledgment

None of this would have come together without **Dr. Khalique Newaz**. From the earliest stages of this project to its final form, his guidance shaped the way I think about the problems at the heart of this work. I am genuinely grateful, not just for the scientific direction, but for the patience and generosity he showed throughout.

To my **mom** and **Armin**, I am not sure I have the right words. They made it possible for me to leave, to stay, and to keep going when things got hard. That kind of support is not something you can repay. I hope this is, at least, a small proof that it was worth it.

And to **Shidokht** and **Rimon**: thank you for the dinners, the conversations, all moments that had nothing to do with work. These past three years would have been a much lonelier place without you in them.

Note on AI use: This thesis was checked for grammar and language using Claude. Claude was used strictly for paraphrasing and refining the writing. All scientific content, analyses, interpretations, and figures are entirely my own.