A structure-guided approach to non-coding variant evaluation for transcription factor binding using AlphaFold 3

Lukas Gerasimavicius[1], Simon C. Biddie[1,2]* & Joseph A. Marsh[1]*

[1]MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom
[2]NHS Lothian, Edinburgh, United Kingdom
*joseph.marsh@ed.ac.uk; simon.biddie@ed.ac.uk

Abstract

Non-coding single-nucleotide variants (SNVs) that alter transcription factor (TF) binding can profoundly affect gene expression and cellular function. Although sequence-based computational methods can excel at predicting TF binding sites, they often exhibit TF-specific performance biases and rely heavily on high-quality training data. Here, we propose a structure-guided approach using AlphaFold 3 (AF3) to model TF-DNA complexes for downstream biophysical evaluation with FoldX. We benchmark this method on a SNP-SELEX dataset in which reference and alternative DNA alleles are tested for differential TF binding *in vitro*. For each allele pair, we generate AF3 models and assess binding differences using FoldX-derived binding energy changes and AF3's interface predicted template modelling (ipTM) score. Across three TFs, SPIB, SF-1 (steroidogenic factor 1, encoded by NR5A1), and PAX5, differences in ipTM (ΔipTM) often align with SNP-SELEX results, slightly surpassing energy-based metrics, although performance varies by TF, with notably poorer performance for PAX5. We further apply this approach to clinically relevant non-coding SNVs from published studies, accurately recapitulating 5 of 6 reported effects. These findings highlight the promise of highly interpretable structural modelling for non-coding variant interpretation, which could even encompass complex homomeric or heteromeric TF assemblies. AF3 provides robust TF-DNA complex structures that can capture subtle differences in binding induced by single-nucleotide changes. Augmented by biophysical binding evaluation, such an integrative approach can elucidate how non-coding variants affect TF binding and regulatory function.

Introduction

Transcription factors (TFs) are regulatory proteins that recognize and bind short DNA sequence motifs in *cis*-regulatory elements, modulating downstream gene expression, and thus controlling diverse cellular processes including development and responses to environmental cues[1]. As such, single nucleotide variants (SNVs) affecting TF-DNA binding have been implicated in human disease[2–5]. Missense variants, which are mutations that change the amino acid identity within a protein, can affect TF function directly by either entirely destabilizing the protein or more subtly affecting its binding preference[6]. For example, missense variants in PAX6, an essential TF in eye development, can affect its DNA-binding specificity and affinity, leading to worse-than-null eye malformations[7,8]. However, expression of genes is also affected by mutations in the non-coding genome at TF bind sites, disrupting existing enhancer and promoter motifs, or enabling new TF binding sites, by affecting chromatin accessibility or generating functional TF binding motifs[9–11].

While disease variant databases like ClinVar[12] are mostly populated with variants primarily associated with simple Mendelian diseases, non-coding variation is significantly more difficult to causally associate and characterize. In contrast, most SNVs associated with common disease and

complex phenotypes through genome-wide association studies (GWAS) occur in non-coding regions, with many variants contributing small effect sizes[13,14]. More than 90% of such associated GWAS SNVs are found in the non-coding genome, with 70% of these being estimated to overlap TF binding motifs[15,16]. However, clinical examples of regulatory variants directly leading to detectable disease through interplay with TFs are scarce. Despite efforts, the direct functional consequences of most disease-associated non-coding SNVs from GWAS remain unelucidated.

To tackle this issue, numerous high-throughput experimental approaches have been developed to examine TF-DNA binding and the effects of SNVs on these interactions. *In vivo* experiments such as chromatin immunoprecipitation sequencing (ChIP-seq) can identify TF binding sites, in context with epigenetic and cell type-specific factors like chromatin accessibility, which govern available TF binding sites in cellular contexts[17]. A ChIP-seq approach also enables the comparison of TF allelic binding preference, facilitating the evaluation of SNV effects[18]. However, TF binding sites identified by ChIP-seq can span hundreds of nucleobases[16], while TFs are known to bind more specific 6-20 base motifs[19]. Binding sites can be identified at a higher resolution through *in vitro* methods such as protein binding microarrays[20] or systematic evolution of ligands by exponential enrichment (SELEX) experiments[21,22]. An iteration on the latter, termed SNP-SELEX, has recently been utilized to compare TF-specific binding events between reference and SNV DNA sequences, constrained to a 40 base resolution[18]. All of these approaches have contributed to the identification of TF binding motifs in database resources such as JASPAR[23], ENCODE[24], ADASTRA[18] and ANANASTRA[25].

Despite the success of these methods, experimental approaches are often time-consuming and expensive to perform at scale, which has led to a massive effort in developing computational methodologies to predict sequences bound by TFs and the effects of non-coding SNVs. Position weight matrices (PWMs) are among the simplest methods for evaluating TF binding[26]. They utilise experimental data from the aforementioned databases to generate TF-specific motif binding logos, which represent the nucleotide probabilities at each position within a sequence that is likely to be bound by a given TF. However, far more advanced strategies have also been developed, such as gapped k-mer approaches like deltaSVM[27], which can capture nucleotide dependencies and have been recognised as top-performing methods in benchmarks[28]. Deep learning has also been applied to TF binding site prediction, with methods like DeepSEA[29] being shown to more accurately predict *in vivo* binding by incorporating complex features like chromatin accessibility[16]. Although these approaches differ in their training data and the scope of their feature engineering, they generally all rely on sequence information.

While sequence-based approaches have achieved significant success, they often overlook the structural complexity of TF-DNA interactions, which can provide deeper mechanistic insights into binding specificity and the effects of genetic variants. Previous studies have demonstrated that the effects of missense mutations in TFs can be accurately assessed using structure-based stability predictors, particularly when a TF-DNA complex structure is available[30]. FoldX, a computational stability prediction tool, evaluates effects of missense mutations, but additionally has the functionality to mutate DNA and RNA sequences in a complex structure, allowing assessment of non-coding SNVs effects on TF binding[31–33]. However, the major hurdle up until recently has been a lack of viable TF-DNA structures in existing databases, and the limitations of current structure modelling tools. With the recent release of AlphaFold 3[34] (AF3) and similar methods[35,36], we are now able to predict biomolecular complex structures, integrating proteins, DNA, RNA and ligands within a single model. This paves the way for leveraging downstream structure-based methods to benchmark the accuracy of TF-DNA binding predictions against extensive TF binding datasets.

In this work we present an approach for evaluating the impact of non-coding SNVs on TF binding by utilizing structural TF-DNA models predicted using AF3. We utilise the abundant available SNP-SELEX data as a ground truth for assessing TF allelic binding preferences. This dataset provides a simplified binding scenario, free from confounding factors such as DNA methylation and chromatin accessibility, making it well-suited for directly evaluating the utility of structure-based biomolecular stability predictors like FoldX, as well as the confidence metrics provided by AF3. While we demonstrate that several FoldX prediction approaches can recapitulate the SNP-SELEX assay results to a high degree, we find that the AF3 predicted interface confidence metric, ipTM, performs better than the downstream stability predictors in identifying differentially bound alleles and the preferred variant. However, we demonstrate that the performance strongly varies on a per-TF basis, indicating situation-specific adjustments might be required, such as limiting the scope of a model only to specific binding domains or considering cooperative or homomeric TF modelling. This work represents a proof-of-concept into the utility of AF3 and other downstream structural tools for non-coding variant evaluation. We evaluate and discuss the possible advantages of a structural approach for assessing non-coding variant impact on TF binding.

Results

Figure 1 summarizes our framework, which utilizes preferential binding scores (PBS) from the SNP-SELEX experiment and the AlphaFold Server to derive and test structure-based allele preference metrics. Due to computational resource constraints, we limited our investigation to a small number of viable TFs and representative allele pairs, covering both differentially bound variants and non-hits with insignificant PBS p-values. As the SNP-SELEX experiment allowed formation of homomeric complexes, to ensure the highest possible concordance between the assay values and predictive metrics, we attempted to identify TFs more likely to only bind DNA as monomers through a survey of Uniprot. We identified SPIB, SF-1 (steroidogenic factor 1, encoded by NR5A1) and PAX5 as candidates with a high number of significant differentially bound alleles and little indication of homomeric binding. For each SNV, both the reference and the alternative allele TF-DNA complex structures were generated using the AF3 server with full-length canonical TF sequences. This allowed us to evaluate the impact of any alteration in the predicted binding site, totaling 4,554 structures for the entire dataset, after excluding pairs with any failed AF3 runs. Finally, the structures were processed through the FoldX stability evaluation platform, which can assess the impact of both amino acid and nucleotide changes on the stability of biomolecular complexes using structural information.

Using the FoldX BuildModel and AnalyseComplex functions, we derived two energetic impact scores. The first approach involved calculating a 'symmetric' ΔΔG value ($\Delta\Delta G_{sym}$) as the mean of both the forward (reference allele to alternative allele on the reference structure) and inverse reverse (negative value of the alternative allele to reference allele on the alternative structure) mutations. This is important because the ΔΔG from the reference allele structure will not necessarily be the inverse of the ΔΔG calculated from the alternative allele structure, especially if there is a shift in binding site between the two predicted models. The second approach, ΔInterface, directly evaluates the difference in interaction energy between the protein and the DNA chains for the two structures. Finally, we also made use of the AF3 interface predicted template modelling (ipTM) score, a confidence measure for the predicted multimeric complex interactions that is provided with the output structures, to derive ΔipTM.

We considered all variants that showed significant allele-specific binding to SPIB, SF-1 and PAX5, splitting them into two groups based on whether they showed preferred binding to either the

reference or the alternative alleles. We also considered a control set of variants that did not show significant allele-specific binding (PBS p-value >0.01). For all three structural metrics, the control variants should ideally tend to have values close to zero, while the differentially bound allele pairs should have large magnitudes, with positive Δ values indicating TF preference for the reference allele, and negative values corresponding with TF bias towards the alternative allele (full dataset accessible through a link in the **Data availability** section).

Figure 2 demonstrates that all three metrics show strong agreement with the experimental assay, with highly significant differences between the reference, alternative and control groups. For all three, the alternative allele preference groups show the lowest values and the reference preference groups show the highest values, while the control groups have values close to zero, as expected. However, the methods do show performance variability, with $\Delta\Delta G_{sym}$ demonstrating an overall shift towards more positive Δ values for all three allele preference bins, corresponding to more stable interactions in the reference models, and a disparity between the mean and median values, indicating strong outliers. ΔInterface on the other hand demonstrates a considerably more symmetrical distribution, centered around the zero value for the control non-hits. The observed differences between the two metrics could be explained by the ΔInterface representing purely the energetics of the interface, which are unaffected by the TF proximity to the variant location. Whereas, if one of the models in a pair is predicted to bind away from the variant, the energetic term from that model in $\Delta\Delta G_{sym}$ would not produce a significant contribution. Surprisingly, however, we discovered that the best performance is achieved by utilizing the AF3 ipTM value difference between the reference and alternative SNV complex models, with the smallest p-values between bins. ΔipTM scores also lead to the tightest distribution of non-hit variants, and differentially bound allele scores being more strongly shifted towards their respective directions. Nonetheless, while it appears that the two FoldX approaches have worse performance, they are being both driven and constrained by the underlying AF3 binding site prediction. The ability of the FoldX approaches to maintain the overall correct directionality of the predicted TF preference is reassuring, as any noise introduced by additional structure relaxation and sidechain adjustment does not seem to negate the discriminatory power of the binding site choice.

The success of the ΔipTM score reflects the ability of AF3 to accurately predict TF binding sites. **Supplementary Figure 1** illustrates model pairs for three SNV cases from the SNP-SELEX assay (TF in the reference allele model in blue, and red in the alternative allele model), with the associated PBS and ipTM scores, showing a variant with diminished TF binding (reference allele preference), a variant that does not impact TF binding (non-hit) and a variant that is more strongly bound by the TF (alternative allele preference). The DNA variants occur at the center of the DNA duplex (yellow), and the result for the non-hit variant is immediately apparent from the AF3 models, which demonstrate that in either case the most likely binding site for the TF does not occur near the variant location, with the mutation having no effect on binding. Respectively, in the case of differential binding, at least one of the TF models appears to show binding at the SNV location, with a TF being shifted away from the allele in the case of reference variant preference, and being shifted towards the SNV when a mutation creates a stronger motif to bind.

As we found the performance of the ΔipTM metric to be promising when utilizing the pooled TF variant dataset, we tested whether AF3 can predict the binding sites for all TFs equally well. **Figure 3** reveals that while ΔipTM performs particularly well for SPIB in discriminating between non-hits and differentially bound variants, with the preference magnitude distributions remaining highly symmetrical, some underperformance can be seen for the other two TFs. While all bin comparisons for SF-1 allele pairs remain statistically significant, differentially bound variants show smaller ΔipTM

magnitudes, particularly for the allele pairs where the SNV is preferred. In the case of PAX5, the distinction between the preference groups is even smaller, with alternative allele preference pairs no longer showing a statistically significant difference from control non-hits. Interestingly, while the FoldX metrics display the same performance bias order, TF-specific analysis reveals that ΔInterface underperforms on both SF-1 and PAX5, with none of the bins for the latter showing statistically significant differences, with the control variant pairs being predicted to have the most positive values (**Supplementary Figure 2**). To evaluate bias more quantitatively, we merged the two differentially bound allele bins into one, with adjusted directionality, and carried out a receiver operating characteristic (ROC) analysis with ΔipTM scores. This revealed that the impact of DNA variants on TF binding was most accurately predicted for SPIB, with an area under the curve (AUC) of 0.757, with SF-1 not far behind at 0.737, while discrimination for PAX5 was poor, with an AUC of 0.576. As AF3 is a trained model, predictions for some TFs may not be as accurate due to fewer available TF-DNA complexes in the PDB, or lower quality binding motif data in JASPAR, which was used to positively reinforce the model. Considering the particularly strong underperformance of the ΔInterface metric, the multi-domain DNA binding nature of PAX family proteins could also likely influence the structure-based scores by introducing additional opportunities (more distinct interfaces) for prediction errors. Regardless, the observed performance heterogeneity suggests that considerations should be taken when utilizing AF3 to predict TF-DNA complex models, such as ensuring the functionally relevant multimeric state is being modelled as well as pre-screening TFs that perform well on benchmarks.

Next, we used the metrics to explore several SNVs from literature linked to clinical phenotypes, in a qualitative manner. Plaisancié *et al.* have recently characterized a regulatory rs745674596 G>A variant responsible for reducing the expression of the FOXE3 transcription factor, leading to ocular anomalies in a patient[37]. The transcription factor USF2A was found to preferentially bind the reference allele, with reduced binding to the alternative allele sequence in a pull-down assay. We applied our approach to model the preferential binding of the USF2A-USF1 TF complex, as Uniprot indicated the heteromeric state is responsible for most DNA binding events. **Figure 4** demonstrates that AF3 predicts distinct TF binding sites for the reference and alternative allele sequences. Surprisingly, the ΔipTM of 0.03 does not indicate differential binding, although the downstream FoldX predictions show a reference allele preference for the TF complex. This result demonstrates that while ΔipTM performs the best generally, a consensus of the three prediction metrics may need to be considered. Despite the FoldX approach being successful, both of the predicted models have low overall confidence, with an ipTM of 0.36 for the reference and 0.33 for the alternative allele complexes. This might result from a lack of positive reinforcement examples from JASPAR motifs or a lack of existing PDB complexes, as only one structure containing USF2 exists on the PDB (8IA3), where the bound DNA is an E-box sequence, similar to the site predicted to be bound in the alternative allele model. Interestingly, despite the lack of structural training data and the low confidence score, the reference allele model shows consistent binding at the allele location upon repeat AF3 runs, which likely drives the accurate downstream FoldX performance.

We also explored two variants, characterized by electrophoretic mobility shift assays, which were shown to affect the binding of the NKX2-5 homeodomain in opposite directions[38]. Variants affecting the function of NKX2-5 have been implicated in congenital heart disorders. **Figure 5** shows the predicted effects of rs61216514 G>A and rs7350789 G>A, which were respectively demonstrated to reduce and increase the binding preference of NKX2-5 through experimental assays. The predicted models clearly indicate a reference allele preference in the case of rs61216514 G>A, and increased binding for the alternative rs7350789 G>A variant, with the ΔipTM values recapitulating assay results. Despite this, some variability can be seen with the FoldX metrics in terms of reduced

magnitude or even the wrong direction, as in the case of $\Delta\Delta G_{sym}$ for rs7350789 G>A. However, a consensus approach would accurately predict these disease-associated variants in line with the experimental data.

A rare non-coding chr11:31685945 G>T variant has been shown to affect an autoregulatory PAX6 enhancer, lowering the binding affinity of PAX6 and disrupting the feedback loop crucial for correct eye development[39]. However, our structure-guided approach failed to conclusively reproduce the experimentally validated effect of the variant, with the FoldX metrics either not showing a strong enough signal or indicating alternative allele preference, and only ΔipTM indicating the variant reduces binding preference, despite both models essentially having the same binding site (**Figure 6**). This example indicates a pattern of less accurate allelic preference prediction for PAX family TFs through structure-derived metrics, concerning both clinical and *in vivo* assay data. A possible explanation is that the JASPAR motifs which were used to optimize AF3 for DNA binding might not actually be representative enough of all cases, as DeltaSVM PAX5 motif models trained on HT-SELEX[40] data failed to break 0.75 AUC in the SNP-SELEX experiment. Another possibility is that AF3 might be biased by the PDB templates, as the resulting models are considerably similar to the PAX6 PDB structure containing a DNA-bound PAX6 paired domain (**Supplementary Figure 3**). Upon complex structure alignment, the T-rich nucleotide sequence from the PAX6 structure overlaps the model DNA, with a T nucleotide being positioned exactly at the variant location, with T also being the alternative allele in the models, which may have inflated the ipTM value. Finally, ΔipTM and ΔInterface are not solely influenced by the interactions of protein and DNA at the variant location, but by all interfaces made, which may introduce additional noise for PAX TFs and similar proteins which have multiple DNA-interacting domains.

Diverse physiological processes are controlled by the complex interplay of regulatory elements, and non-coding variation in these elements often serves as the basis for cancer. However, while some variants may increase susceptibility to cancer, others may have a protective effect by downregulating the expression of oncogenes by abrogating TF binding. Increased expression of LMO1, a transcriptional co-factor, has been implicated in neuroblastomas, where LMO1 acts as an oncogene. However, the rs2168101 C>A variant (forward strand) in an LMO1 enhancer, bound by GATA3, has been identified to diminish the binding of the TF through ChIP-seq experiments[41]. **Figure 7** demonstrates that the structure-guided approach could reproduce the experimental results, with all three metrics being in agreement. Furthermore, we tested the rs339331 T>C oncoprotective variant in an RFX6 enhancer, a gene which, if upregulated, is associated with tumour progression and metastasis in prostate cancers[42]. **Figure 8** reveals that while both models were predicted to be bound by HOXB13 at the same site of the variant location, two of the three metrics correctly predict that the alternative allele is not preferentially bound, functioning as a disease-protective variant.

We considered the utility of a consensus approach following the observations from the previous examples. We examined whether a FoldX ΔΔG score derived from the model structure with the higher ipTM out of the allele pair could serve as a more accurate method. We combined both the information about the AF3 confidence of binding site as well as a biophysical evaluation of the variant itself to derive an ipTM-guided $\Delta\Delta G_{max(ipTM)}$ metric, which relies on the 'correct' higher confidence AF3 model structure to predict a FoldX ΔΔG value for the non-coding mutation. $\Delta\Delta G_{max(ipTM)}$ does not contain directionality, as both allele preference groups display a complex destabilizing effect due to the nature of the metric in all cases (**Supplementary Figure 4a**), thus subsequent analyses were carried out comparing non-hits against pooled differentially bound alleles. The two-class comparison reveals that while $\Delta\Delta G_{max(ipTM)}$ is also discriminatory, it does not suffer from TF-specific performance heterogeneity to the same extent as $\Delta\Delta G_{sym}$ (**Supplementary**

Figure 4b), displaying a performance similar to ΔipTM on PAX5, likely as a result of utilizing both ipTM and FoldX in this approach. Interestingly, if we derive a reverse metric, where ΔΔG is calculated only using the lower ipTM models out of the pair, **Supplementary Figure 4c** shows the scores are unable to differentiate between hit and non-hit pairs across the TFs, demonstrating the importance of using the correct AF3 model.

The clinical variants explored previously, as well as the examples of differentially bound alleles from **Supplementary Figure 1**, in most cases display a strong shift in the TF binding site based on the AF3 predictions. This raises the question of whether the favourable performance of ΔipTM and FoldX metrics in distinguishing differentially bound alleles relies solely on whether AF3 predicts distinct binding sites for the variant pairs. To explore this, we derived a distance metric for each pair of alleles, which describes the spatial shift of a representative DNA-binding residue in a given TF between the closest nucleotide in the reference against the same nucleotide in the alternative allele sequence (see **Methods**). To explore the role of predicted binding site differences we classified all the allele pairs into 'same' and 'shifted' binding sites based on a comparison of distances between representative TF and DNA points in reference and alternative sequence models for each pair (see **Methods**). **Supplementary Figure 5, a** demonstrates that binding sites in differentially bound allele pairs tend to be shifted, with a relative 13% increase in cases compared to control non-hits. However, while the difference is statistically significant, in both, AF3 predicts more than half of the allele pairs to be bound by the TF at the same site. We further explored whether the shifted and same binding site groups showed any performance differences for the various structure-based differential binding metrics. It was found that the performance of ΔipTM and ΔInterface was not significantly impacted by shifted sites, with very similar effect sizes between the groups, with only $\Delta\Delta G_{sym,}$ seeing a 0.08 increase in effect size between control and differentially bound pairs when sites are shifted (**Supplementary Figure 5, b**). Lastly, we explored whether the binding site distance shift on its own could be used as a predictive metric to identify allele preference. **Supplementary Figure 6** demonstrates that the distance metric itself can have significant discriminatory power, where differentially bound alleles displaying large shifts in the predicted TF binding site location. However, it seems to show worse TF-specific performance than the previous metrics, with both non-hit and differentially bound allele pairs being predicted to have drastic binding site shifts in PAX5. This result indicates that favourable utilization of ΔipTM, the best metric identified in this work, is not simply due to instances where shifted binding sites are predicted. The neural network architecture and training approaches of AF3 result in a method that is, to a high degree, able to successfully capture TF-DNA interactions and the impact of variants.

Finally, a ROC analysis was carried out to compare the performance of all derived metrics in separating differentially bound allele pairs from pairs that show no binding preference for TFs, after adjusting their scores for two-class comparison (**Figure 9**). The hybrid $\Delta\Delta G_{max(ipTM)}$ approach demonstrates increased performance compared to $\Delta\Delta G_{sym}$ in all scenarios, and better performance than ΔInterface except in the case of SPIB. However, the binding site shift distance generally shows lacklustre performance that is distinct from ΔipTM, indicating the latter is predictive even without apparent TF binding site shifts between the models. Finally, ΔipTM dominates over other metrics across all three TFs, illustrating the top performance score from the specific AF3 models. AF3 influences all downstream derivative metrics by its choice of binding sites in each model, with ipTM representing the confidence at a given binding site.

Discussion

While numerous sequence-based TF binding predictors exist, ranging from the simplest PWMs[16,43], unable to account for nucleotide dependency, to multi-feature deep learning[44] and gapped k-mer models trained on large amounts of experimental data[27], structure-based approaches have not seen much development. It is not surprising, as, until recently, acquiring the relevant protein-DNA structural information was difficult, with sparse structural databases and limited modelling approaches. However, with advances in the latter we can now predict not only protein models, but also their complexes with other biomolecules, such as RNA, DNA and various ligands[35,36,45]. The AlphaFold 3 server, specifically, enables users to access this functionality without requiring in-house high-performance compute systems[34]. Straightforward generation of TF-DNA complex models now allows us to more readily utilize downstream structure-based tools to evaluate biomolecular binding and predict genetic variant effects.

The *in vitro* SNP-SELEX experiment represents a very simplified DNA binding scenario for TFs, which does not account for *in vivo* factors such as chromatin structure, methylation or histone modification. However, data from this type of experimental setup provides the most suitable platform to test structure-based TF binding and stability predictors, which directly or indirectly approximate biophysical interactions on a limited scale. In this work we aimed to explore the utility of structural approaches that utilize AF3 server models downstream, such as FoldX, but demonstrate that AF3 itself provides sufficient information in the form of the ipTM metric to reproduce SNP-SELEX assay results to a large extent. However, this study represents only a proof-of-concept, being limited to a selection of allele pairs from three TFs, due to computational resource constraints. With the recent full release of the AF3 model weights, the utility of structural approaches to TF-DNA binding could be explored on a larger scale. In this work we did not test or provide comparisons to the numerous existing TF binding site prediction tools, such as PWMs and more complex sequence-based predictors, extensively tested in previous benchmarking studies[18,28,46,47]. AF3 performance is currently outclassed by top performing methods that can achieve near perfect AUCs on certain datasets. Nevertheless, in comparison, structural approaches to predict TF binding, and the structural effects of non-coding variants, have not been as extensively evaluated, nor trained on specific allelic datasets.

In this work we also aimed to showcase other possible advantages that are afforded by a structural approach. We demonstrated that AF3 models provide an easily interpretable representation of the effects (or lack, thereof) that non-coding variants can have on TF binding. Complex models may not drastically change between the reference and the alternative allele DNA sequences when the variant is not in proximity to the predicted TF binding site, or is judged not to change the binding site, but clear binding site shifts can be observed for differentially bound SNP-SELEX variants when existing motifs are disrupted, or new ones emerge. In fact, we demonstrated differentially bound allele pairs are significantly more predicted to be bound by TFs at distinct shifted sites than control pairs. However, when differentially bound alleles do appear to be bound at the same location, the ΔipTM and FoldX metrics are still able to distinguish allele preference. Further exploration will be required to characterize AF3's behaviour in terms of site shifts when predicting TF-DNA complexes containing palindromic sequences or overall low-ipTM scenarios, where high uncertainty could lead to artificial binding site shifts. Another possible advantage of AF3 models is the simulation of homomeric or even heteromeric TF complex binding to DNA, which is not currently possible with most other known tools. While AF3 performance is likely driven by the positive reinforcement using sequences generated from known JASPAR motif PWMs, the FoldX stability predictions derived using the models do not significantly deviate in the magnitude or direction of scores, suggesting that AF3 has learned

some underlying biophysical representation of TF and DNA binding patterns on top of sequence motif recognition.

Computational resource constraints have only allowed us to explore structural metric performance for three TFs, however, this was sufficient to observe performance heterogeneity, with differential PAX5 binding being difficult to predict for all tested metrics. Surprisingly, PAX5 was the one TF annotated as binding under monomeric conditions, which we speculated may allow higher correspondence between the SNP-SELEX assay scores and AF3 models. However, it was also one of the TFs that were tested in SELEX experiments using a reduced protein sequence, covering just the paired domain, while here we modelled the full-length TF sequences. Interestingly, despite both the HT-SELEX and SNP-SELEX experiments using just the PAX5 paired domain, deltaSVM models derived on HT-SELEX[40] data also failed to achieve sufficient performance (<0.75 AUC) when predicting SNP-SELEX values, indicating a deeper issue. Variation in the experimental setup, such as the limitations in the diversity of the underlying DNA library, could potentially result in a misleading motif binding profile. An even more considerable disconnect could arise between performance of methods trained on *in vitro* SELEX data and *in vivo* ChIP-seq[28] motifs. We also demonstrate that our approach did not conclusively reproduce the effects of a variant on PAX6 binding, showing only a marginal ΔipTM result in the right effect direction. We speculate that most failures being associated with PAX proteins might not be a coincidence, and that overreliance on existing PDB templates, like 6PAX[48], or the separate sub-domains and the multi-interface nature of PAX TF binding might be negatively affecting the binding site choice and the resulting ipTM values.

A more extensive benchmark would need to be carried out, including significantly more TFs and variants, to gauge the utility of AF3 in identifying the effects of non-coding variants on TF binding. The utility of AF3 for *in vivo* binding prediction is likely limited, without accounting for chromatin accessibility and other complex features. However, it serves as a useful source of structural TF-DNA models with biologically relevant binding conformations, that enables use of downstream analysis tools such as biophysical stability predictors like FoldX. AF3 may prove to be invaluable in a non-coding variant evaluation pipeline which incorporates DNA sequences for methylation, chromatin accessibility, cell-specificity and other features, predicting binding events in conjunction with functional data. While the trained nature of AF3 means it might currently not be able to accurately evaluate binding for all possible TFs and their DNA complexes, with respect to choosing the correct binding site, it can nonetheless be forced to generate structural TF-DNA models when given a motif-length DNA sequence. For instance, a longer DNA sequence may be split up in a sliding window fashion, to generate structures that are then analyzed using FoldX, potentially identifying the most likely binding site through a measure of stability, even though AF3 might not have chosen it due to limitations in training examples. With the several clinical variant examples we have explored in this work, we have shown that ΔipTM does not always provide an accurate depiction of the binding preference, and that a consensus score incorporating FoldX methods can provide a more robust approach. Overall, we have demonstrated AF3 as a beneficial tool in non-coding variant evaluation and make a case for a more in-depth exploration of a structural approach to characterize the effects of non-coding variants in *cis*-regulatory element on TF binding.

Methods

Original batch SNP-SELEX data for SPIB, SF-1 and PAX5 was retrieved from GVATdb[40]. The 40bp oligos with a variant at the 21st position on the forward strand were extracted using the genomic coordinates provided using the 'BSgenome'[49] R package in the hg19 reference genome. Due to computational resource constraints, namely the AlphaFold 3 Server quota limitations, we had to

prioritize the number of transcription factors and variants to explore. The specific targets were chosen based on the high amount of significantly differentially bound alleles detected in the assay, as well as through a manual survey of transcription factor Uniprot entries to identify the ones that were more likely to bind DNA as monomers. This was done to simplify the structural modelling procedure and to maximize the likelihood the structures would be more reflective of the assay interactions. Only PAX5 was explicitly stated to bind DNA as a monomer. Finally, the variants themselves were prioritized, utilizing all significant hits found in the SNP-SELEX study (preferential binding score p-value <0.01), while also retrieving 400 of the least significant non-hits for each TF as controls.

For each transcription factor and SNP entry, both the reference and alternative allele interactions were predicted using the AlphaFold 3 Server using a random seed. Full-length canonical Uniprot entry human transcription factor sequences were used. All transcription factor sequences were submitted as monomers while the forward strand of the 40bp oligo was used to derive a reverse complement strand to model a DNA duplex. CIF files for the top predicted complex models were converted to PDBs using BeEM[50] for downstream analyses. FoldX[33] was then used on reference and alternative allele structures to evaluate the energetic impact of DNA mutations. Before proceeding, all complex structures were run through the RepairPDB procedure. In the first instance, the impact of the DNA variants was assessed using the BuildModel procedure, by generating and scoring both the reference-to-alternative (forward) and the alternative-to-reference (reverse) mutations, mutating the respective starting model structures. The scores were then used to calculate the "symmetric" ΔΔG ($\Delta\Delta G_{sym}$), which ought to better represent the impact due the likely shifts in binding sites between the reference and alternative AF3 models. In the second instance, we used the AnalyseComplex procedure on the reference and alternative model structures to derive interface strength scores and derived a ΔInterface metric. Finally, we also extracted the ipTM scores for each model structure from the AF3 json files and derived the difference score for all paired alleles. $\Delta\Delta G_{max(ipTM)}$ was derived simply as the ΔΔG score of the model with the higher ipTM in an allele pair, with the reference model score representing a reference-to-alternative mutation, and the alternative score being derived on the alternative model with an alternative-to-reference BuildModel mutation.

For the distance analyses, models with the highest ipTM scores for each TF were investigated to identify a representative residue position in close proximity to the variant nucleotides, choosing positively charged residues that are predicted to enter the DNA groove. As such, residues 229, 39 and 67 were chosen as representative for SPIB, SF-1 and PAX5, respectively. Using the chosen positions, atomic resolution distances were calculated to identify the closest DNA nucleotide in each reference model in a pair, allowing measurements between any atoms of the residue and nucleotides excluding hydrogen. For the alternative model, the same residue and nucleotide were evaluated for any change in distance, and a binding site was deemed shifted if the alternative model distance exceeded that of the distance in the reference by more than 5Å.

Statistical group analysis first involved binning the SNPs (the reference and alternative allele pairs) into three groups, depending on the SNP-SELEX preferential binding scores (PBS) for the pairs, as well as the PBS p-values. As for SNP-SELEX, differentially bound variants were those with a PBS p-value <0.01, where reference allele preference was indicated by positive PBS values and alternative allele preference occurred under negative PBS values. Control samples of non-differentially bound allele pairs were chosen by taking pairs with the least significant p-values for each TF. For two-class non-hit *vs* differentially bound pair analysis, the PBS, structure-derived stability metric and ipTM values were multiplied by -1 for the alternative allele preference group, and merged into the

reference allele preference group. Two-sample comparisons involved a two-sided Mann-Whitney *U* test, while pairwise statistical comparisons between groups were carried out using a Dunn's test implementation in the R 'ggstatsplot'[51] package, with the *p*-values for comparisons involving more than two groups being adjusted through Holm's multiple comparison correction[52,53]. Class proportion comparison was performed using the chi-squared test, with p-values derived in two-way and one-way table scenarios. ROC analyses were carried out using the 'pROC'[54] package, where the case-control direction was implicitly set for each predictor and the AUC was approximated using DeLong's method.

Data availability

The data generated in this study have been deposited in the OSF database at https://doi.org/10.17605/OSF.IO/UZKG5.

Acknowledgement

This project was supported by funding from the Medical Research Council (MRC) Human Genetics Unit core grant (MC_UU_00035/9).

References

1. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
2. French, J. D. & Edwards, S. L. The Role of Noncoding Variants in Heritable Disease. *Trends in Genetics* **36**, 880–891 (2020).
3. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Human Molecular Genetics* **24**, R102–R110 (2015).
4. Lee, T. I. & Young, R. A. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **152**, 1237–1251 (2013).
5. Iñiguez-Muñoz, S. *et al.* Hidden secrets of the cancer genome: unlocking the impact of non-coding mutations in gene regulatory elements. *Cellular and Molecular Life Sciences* **81**, 274 (2024).
6. Barrera, L. A. *et al.* Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science* **351**, 1450–1454 (2016).
7. Williamson, K. A. *et al.* Recurrent heterozygous PAX6 missense variants cause severe bilateral microphthalmia via predictable effects on DNA–protein interaction. *Genetics in Medicine* **22**, 598–609 (2020).
8. McDonnell, A. F. *et al.* Deep mutational scanning quantifies DNA binding and predicts clinical outcomes of PAX6 variants. *Molecular Systems Biology* **20**, 825–844 (2024).
9. Yang, J. & Adli, M. Mapping and Making Sense of Noncoding Mutations in the Genome. *Cancer Research* **79**, 4309–4314 (2019).
10. Jin, Y., Jiang, J., Wang, R. & Qin, Z. S. Systematic Evaluation of DNA Sequence Variations on in vivo Transcription Factor Binding Affinity. *Frontiers in Genetics* **12**, (2021).
11. Deplancke, B., Alpern, D. & Gardeux, V. The Genetics of Transcription Factor DNA Binding Variation. *Cell* **166**, 538–554 (2016).
12. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research* **42**, D980–D985 (2014).
13. Spielmann, M. & Mundlos, S. Looking beyond the genes: the role of non-coding variants in human disease. *Human Molecular Genetics* **25**, R157–R165 (2016).

14. Gloss, B. S. & Dinger, M. E. Realizing the significance of noncoding functionality in clinical genomics. *Experimental & Molecular Medicine* **50**, 1–8 (2018).
15. Yao, Q. *et al.* Motif-Raptor: a cell type-specific and transcription factor centric approach for post-GWAS prioritization of causal regulators. *Bioinformatics* **37**, 2103–2111 (2021).
16. Wang, Z. *et al.* Towards a better understanding of TF-DNA binding prediction from genomic features. *Computers in Biology and Medicine* **149**, 105993 (2022).
17. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics* **43**, 264–268 (2011).
18. Abramov, S. *et al.* Landscape of allele-specific transcription factor binding in the human genome. *Nature Communications* **12**, 2751 (2021).
19. Srivastava, D. & Mahony, S. Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1863**, 194443 (2020).
20. Berger, M. F. & Bulyk, M. L. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature Protocols* **4**, 393–411 (2009).
21. Pantier, R., Chhatbar, K., Alston, G., Lee, H. Y. & Bird, A. High-throughput sequencing SELEX for the determination of DNA-binding protein specificities in vitro. *STAR Protocols* **3**, 101490 (2022).
22. Riley, T. R. *et al.* SELEX-seq: A Method for Characterizing the Complete Repertoire of Binding Site Preferences for Transcription Factor Complexes. in *Hox Genes: Methods and Protocols* (eds. Graba, Y. & Rezsohazy, R.) 255–278 (Springer New York, New York, NY, 2014). doi:10.1007/978-1-4939-1242-1_16.
23. Rauluseviciute, I. *et al.* JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **52**, D174–D182 (2024).
24. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
25. Boytsov, A. *et al.* ANANASTRA: annotation and enrichment analysis of allele-specific transcription factor binding at SNPs. *Nucleic Acids Research* **50**, W51–W56 (2022).
26. Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23 (2000).
27. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nature Genetics* **47**, 955–961 (2015).
28. Han, D. *et al.* Comparative analysis of models in predicting the effects of SNPs on TF-DNA binding using large-scale in vitro and in vivo data. *Briefings in Bioinformatics* **25**, bbae110 (2024).
29. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods* **12**, 931–934 (2015).
30. Gerasimavicius, L., Livesey, B. J. & Marsh, J. A. Correspondence between functional scores from deep mutational scans and predicted effects on protein stability. *Protein Science* **32**, e4688 (2023).
31. Delgado, J. *et al.* FoldX Force Field revisited, an improved version. *Bioinformatics* btaf064 (2025) doi:10.1093/bioinformatics/btaf064.
32. Blanco, J. D., Radusky, L., Climente-González, H. & Serrano, L. FoldX accurate structural protein–DNA binding prediction using PADA1 (Protein Assisted DNA Assembly 1). *Nucleic Acids Research* **46**, 3852–3863 (2018).

33. Delgado, J., Radusky, L. G., Cianferoni, D. & Serrano, L. FoldX 5.0: Working with RNA, small molecules and a new graphical interface. *Bioinformatics* **35**, 4168–4169 (2019).
34. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
35. Krishna, R. *et al.* Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, eadl2528.
36. Baek, M. *et al.* Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nature Methods* **21**, 117–121 (2024).
37. Plaisancié, J. *et al.* Insights into the FOXE3 Transcriptional Network and Disease Mechanisms from the Investigation of a Regulatory Variant Driving Complex Microphthalmia. *bioRxiv* 2025.01.13.632782 (2025) doi:10.1101/2025.01.13.632782.
38. Peña-Martínez, E. G. *et al.* Disease-associated non-coding variants alter NKX2-5 DNA-binding affinity. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1866**, 194906 (2023).
39. Bhatia, S. *et al.* Disruption of Autoregulatory Feedback by a Mutation in a Remote, Ultraconserved PAX6 Enhancer Causes Aniridia. *The American Journal of Human Genetics* **93**, 1126–1134 (2013).
40. Yan, J. *et al.* Systematic analysis of binding of transcription factors to noncoding variants. *Nature* **591**, 147–151 (2021).
41. Oldridge, D. A. *et al.* Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism. *Nature* **528**, 418–421 (2015).
42. Huang, Q. *et al.* A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. *Nature Genetics* **46**, 126–135 (2014).
43. Ambrosini, G. *et al.* Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. *Genome Biology* **21**, 114 (2020).
44. Kathail, P., Bajwa, A. & Ioannidis, N. M. Leveraging genomic deep learning models for non-coding variant effect prediction. (2024).
45. Esmaeeli, R., Bauzá, A. & Perez, A. Structural predictions of protein–DNA binding: MELD-DNA. *Nucleic Acids Research* **51**, 1625–1636 (2023).
46. Boytsov, A., Abramov, S., Makeev, V. & Kulakovskiy, I. Positional weight matrices have sufficient prediction power for analysis of noncoding variants [version 3; peer review: 2 approved]. *F1000Research* **11**, (2022).
47. Wagih, O., Merico, D., Delong, A. & Frey, B. J. Allele-specific transcription factor binding as a benchmark for assessing variant impact predictors. *bioRxiv* 253427 (2018) doi:10.1101/253427.
48. Xu, H. E. *et al.* Crystal structure of the human Pax6 paired domain–DNA complex reveals specific roles for the linker region and carboxy-terminal subdomain in DNA binding. *Genes Dev.* **13**, 1263–1275 (1999).
49. BSgenome. *Bioconductor* http://bioconductor.org/packages/BSgenome/.
50. Zhang, C. BeEM: fast and faithful conversion of mmCIF format structure files to PDB format. *BMC Bioinformatics* **24**, 260 (2023).
51. Patil, I. Visualizations with statistical details: The 'ggstatsplot' approach. *Journal of Open Source Software* **6**, 3167 (2021).
52. Dunn, O. J. Multiple Comparisons Using Rank Sums. *null* **6**, 241–252 (1964).
53. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* **6**, 65–70 (1979).

54. Turck, N. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **8**, 12–77 (2011).

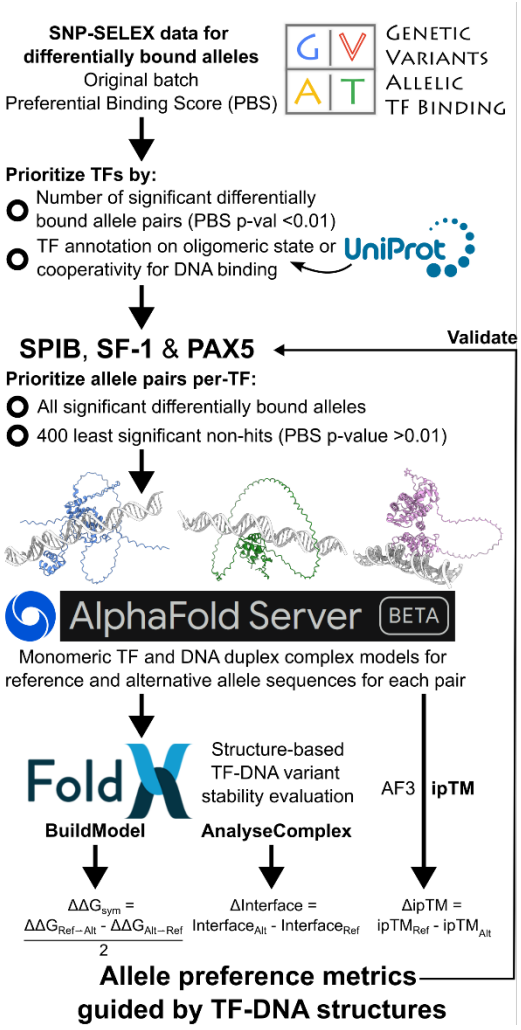


Figure 1. Schema for the derivation and testing of structure-guided transcription factor (TF) allele preference metrics using SNP-SELEX data. Three TFs were prioritized based on the number of identified significant SNP-SELEX hits (significant Preferential Binding Scores - PBS) and a lack of strong annotation for multimeric DNA binding. Allele pairs were further prioritized, modelling structures for all differentially bound allele pairs as well as 400 least significant PBS non-hits, per-TF. Using the TF-DNA complex structure model pairs, derived for both the reference and the alternative DNA sequences at a locus, TF preference was assessed through two separate complex stability prediction approaches based on FoldX, as well as using the AF3 model confidence metric ipTM. The metrics were compared against the experimentally derived PBS scores.

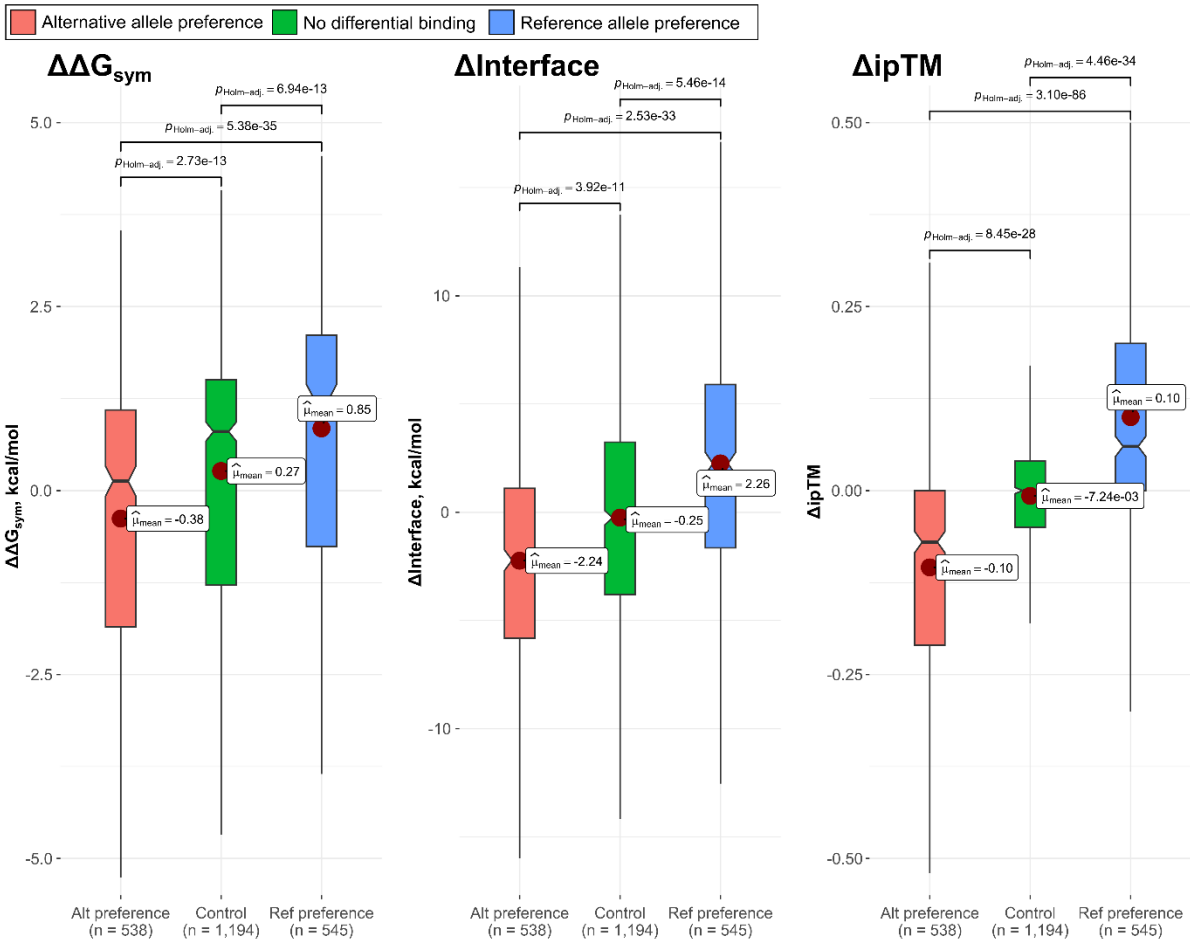


Figure 2. Structure-derived TF-DNA stability metrics recapitulate SNP-SELEX results, with the AlphaFold 3 ipTM model quality score showing greater accuracy. Boxes denote data within 25th and 75th percentiles, and contain median (middle line) and mean (red dot) value notations. Whiskers extend from the box to furthest values within 1.5x the inter-quartile range. Significance values are derived using two-sided Holm-corrected Dunn's tests. Sample size (n) indicates the number of allele pairs.

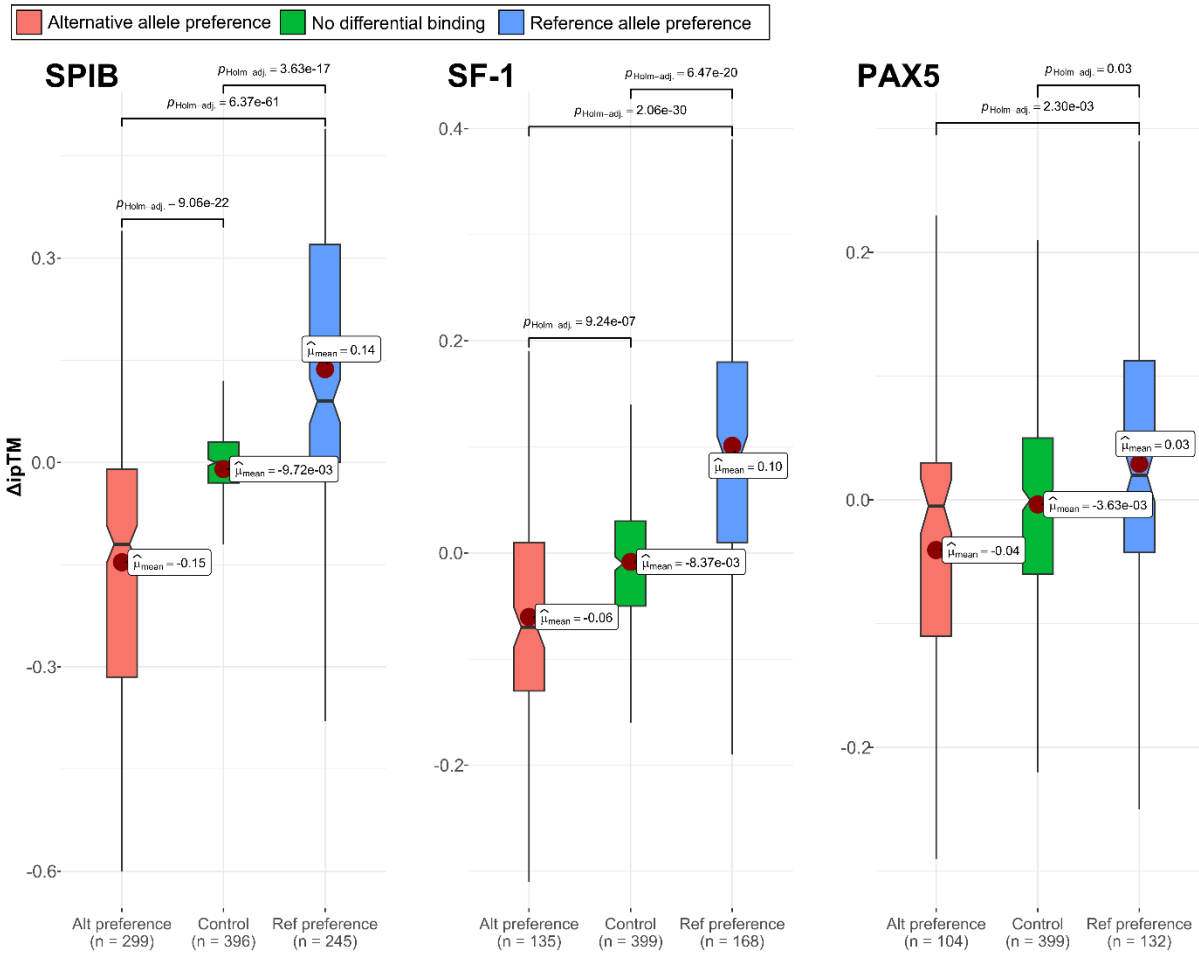


Figure 3. Differential binding prediction performance using ipTM is heterogeneous across different TFs. Boxes denote data within 25th and 75th percentiles, and contain median (middle line) and mean (red dot) value notations. Whiskers extend from the box to furthest values within 1.5x the inter-quartile range. Significance values are derived using two-sided Holm-corrected Dunn's tests. Sample size (n) indicates the number of allele pairs.

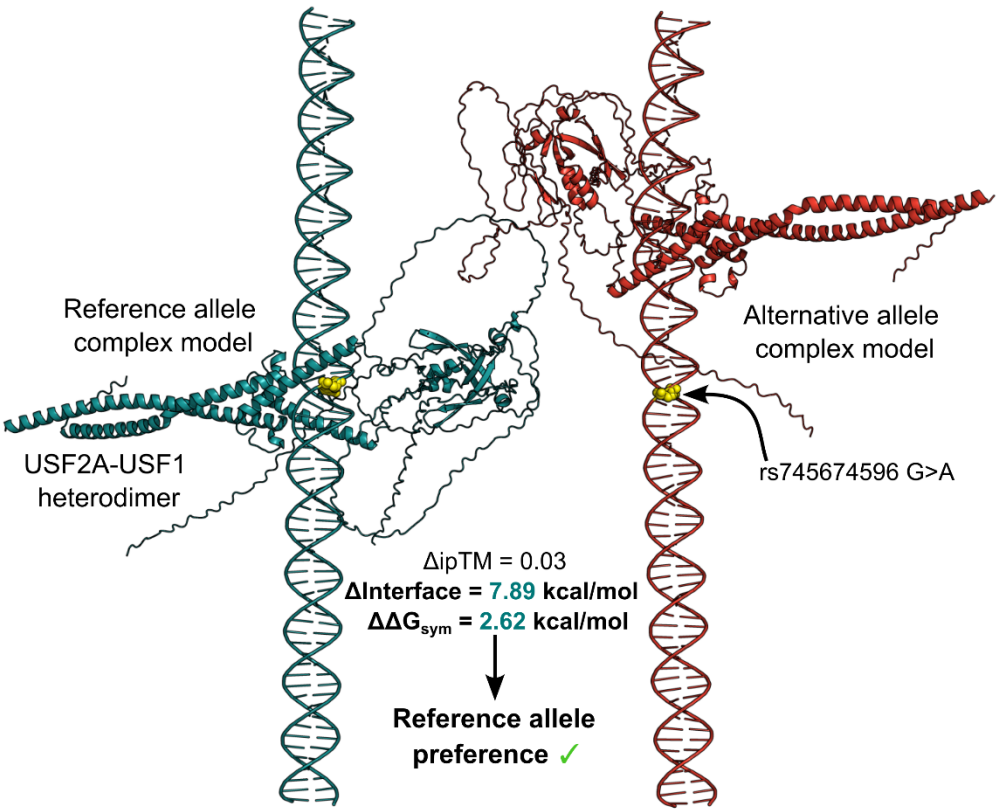


Figure 4. Downstream FoldX protocols can recapitulate assay data for clinical variants where AF3 ipTM fails to inform of differential binding. USF2A-USF1 complexes were modelled with 81bp sequences (variant positions with 40bp flanking sequences) used in Plaisancié *et al.*, 2025, exploring the *in vivo* and *in vitro* effects of the regulatory rs745674596 G>A variant[37]. The blue model (left) represents the TF binding site for the reference DNA sequence, while the red structure (right) shows the predicted binding site for the alternative allele. The yellow orbs represent the SNV location on the forward strand.

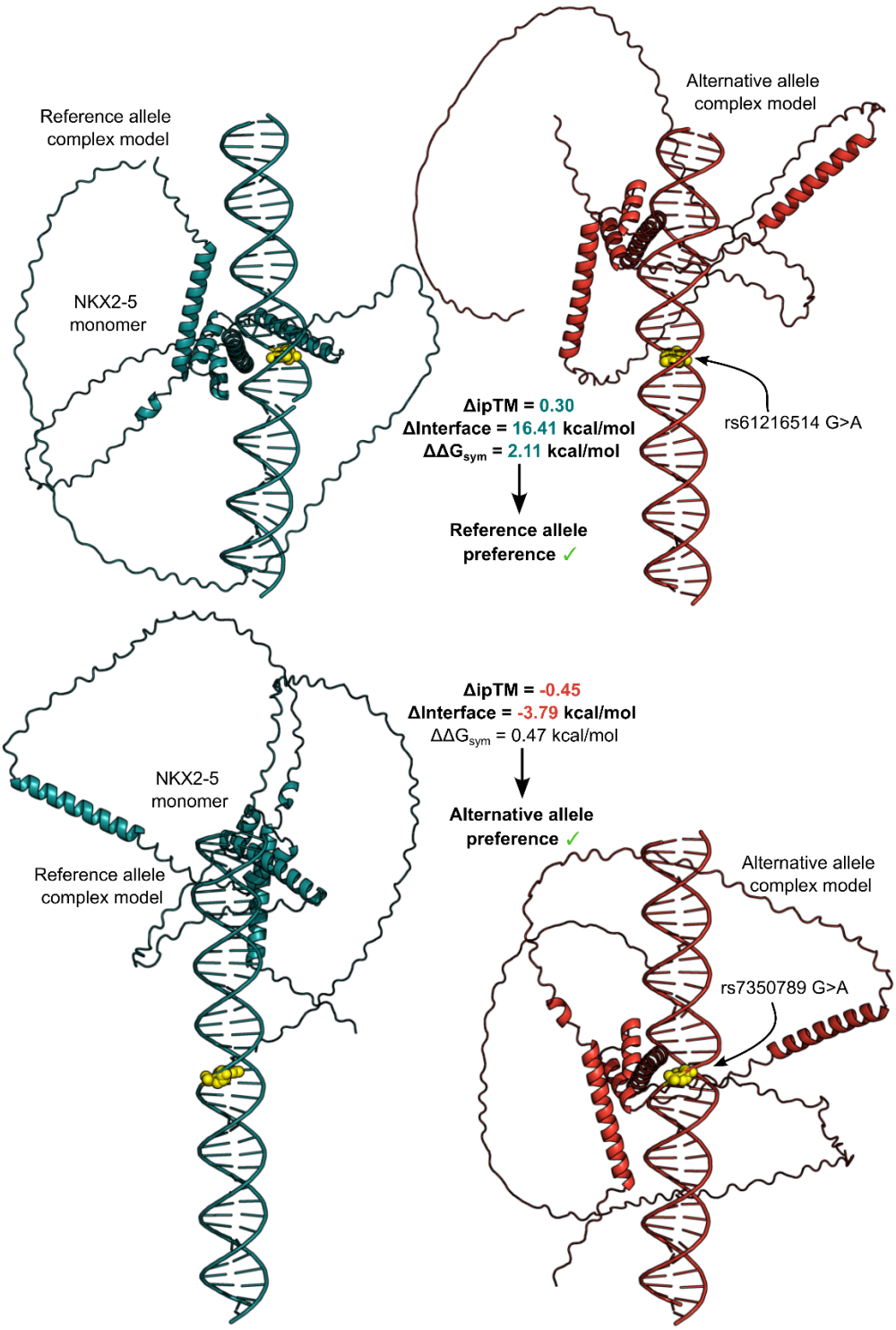


Figure 5. Structure-guided allele preference metrics show heterogenous performance, but score consensus recapitulates effects in line with experimental and clinical phenotype results. To more closely match the context of the electrophoretic mobility shift assay with purified NKX2-5 homeodomains outlined in Peña-Martínez, *et al.* 2023, monomeric NKX2-5 was modelled with 41bp long DNA sequences centred on the rs61216514 G>A and rs7350789 G>A variants[38]. In each case, blue models (left) represent the TF binding pose and site for the reference DNA sequence, while the red structures (right) show the predicted binding site for the alternative allele. The yellow orbs represent the SNV nucleotide on the forward strand.

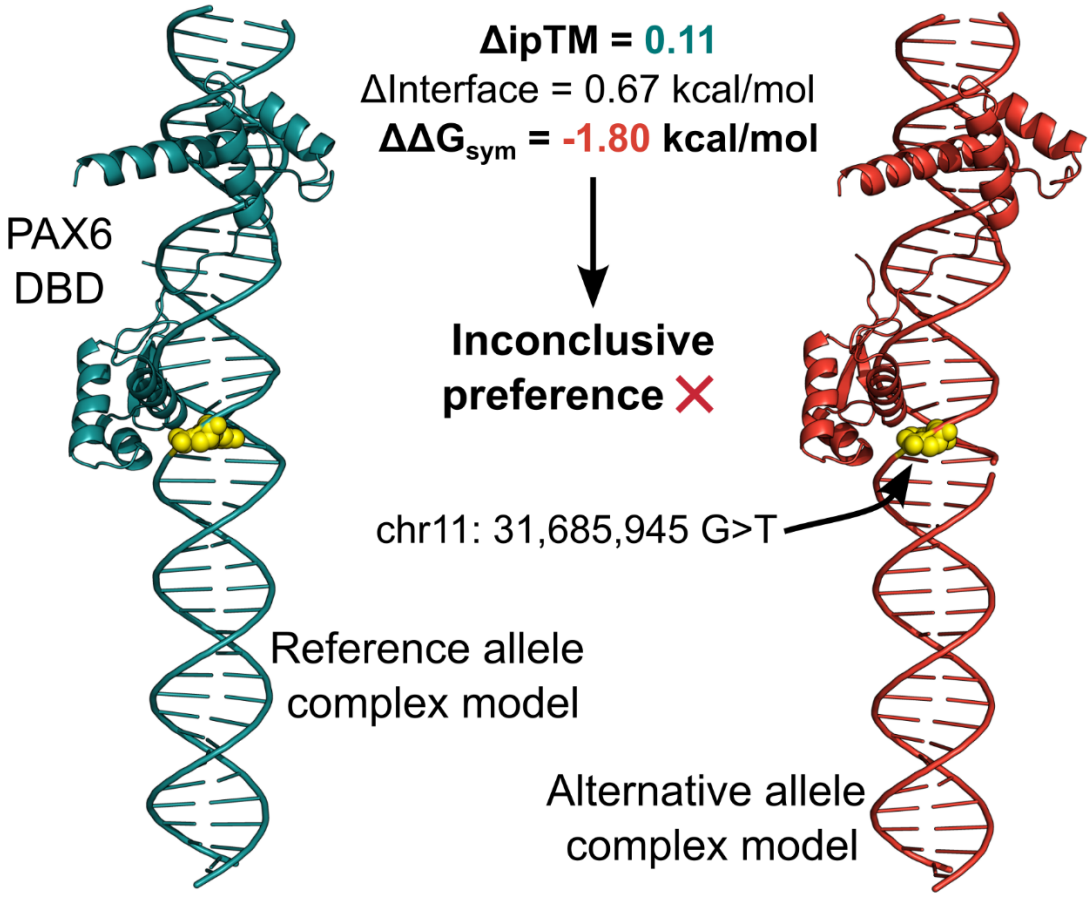


Figure 6. The structure-guided approach fails to conclusively identify the differential binding direction for a regulatory PAX6 enhancer variant. The G>T variant located at chr11:31685945 was found to disrupt autoregulatory PAX6 binding at an enhancer and show reference allele preference in experiments carried out by Bhatia *et al*.[39]. The blue model (left) represents the TF binding site for the reference DNA sequence, while the red structure (right) shows the predicted binding site for the alternative allele. The yellow orbs represent the SNV nucleotide on the forward strand.

Figure 7. Structure-guided metrics correctly predict the effects of an oncoprotective variant. LMO1, a transcription regulator, has been shown to act as a neuroblastoma oncogene. Oldridge *et al.*,2015 showed that am experimentally identified super-enhancer variant rs2168101 C>A decreased total LMO1 expression in neuroblastomas by abrogating GATA3 binding[41]. The blue model (left) represents the TF binding site for the reference DNA sequence, while the red structure (right) shows the predicted binding site for the alternative allele. The yellow orbs represent the SNV nucleotide on the forward strand.

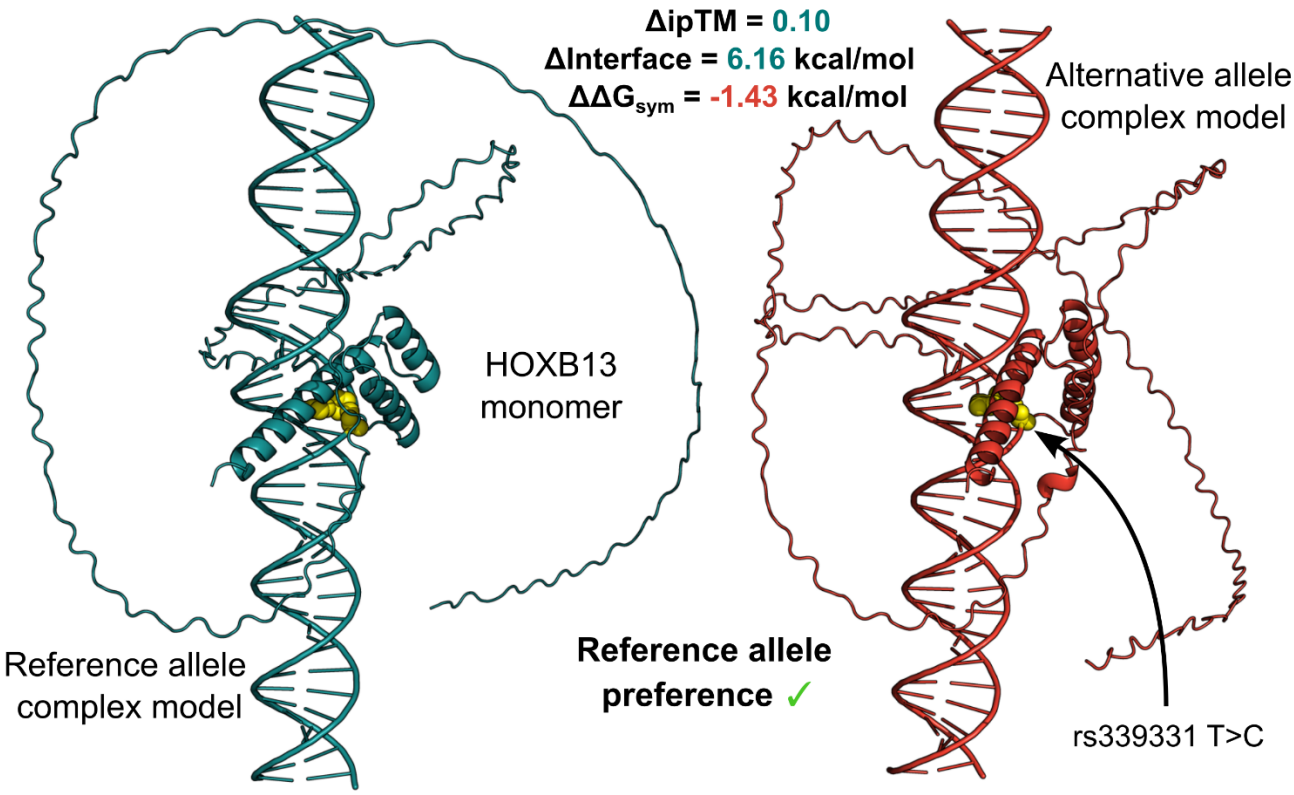


Figure 8. A consensus approach of structural metrics can identify differential allele binding even without obvious binding site shifts. Huang *et al*,.2014 demonstrated that the rs339331 T>C variant in an RFX6 enhancer reduces the binding affinity of HOXB13, imparting an oncoprotective effect that diminishes tumour progression and metastasis in prostate cancers[42]. The structural complex models were predicted using the AF3 server. The blue model (left) represents the TF binding site for the reference DNA sequence, while the red structure (right) shows the predicted binding site for the alternative allele. The yellow orbs represent the SNV nucleotide on the forward strand.

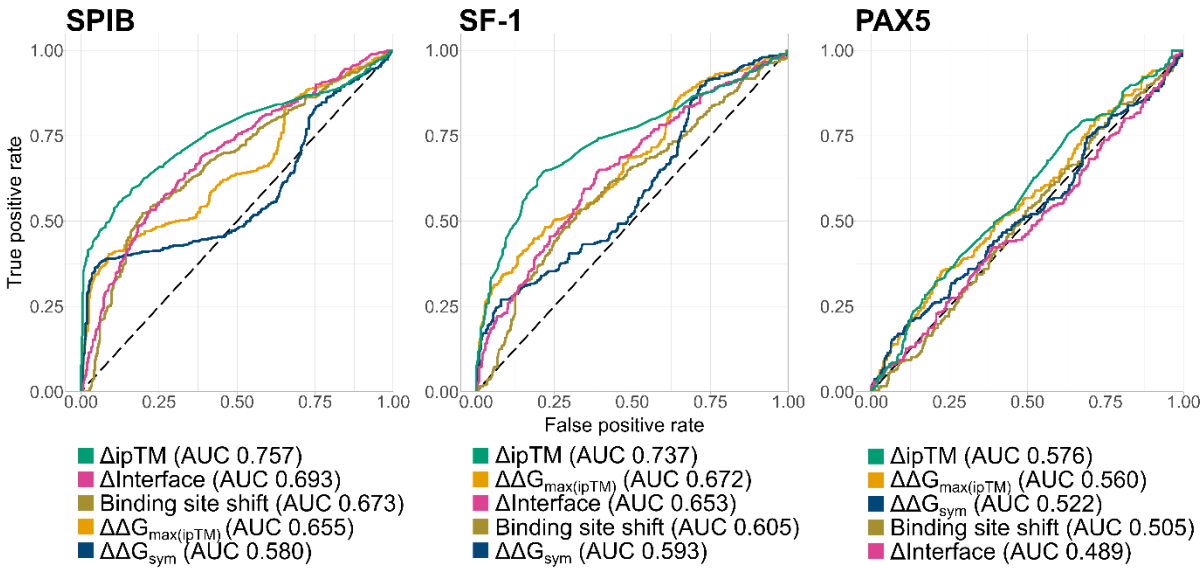
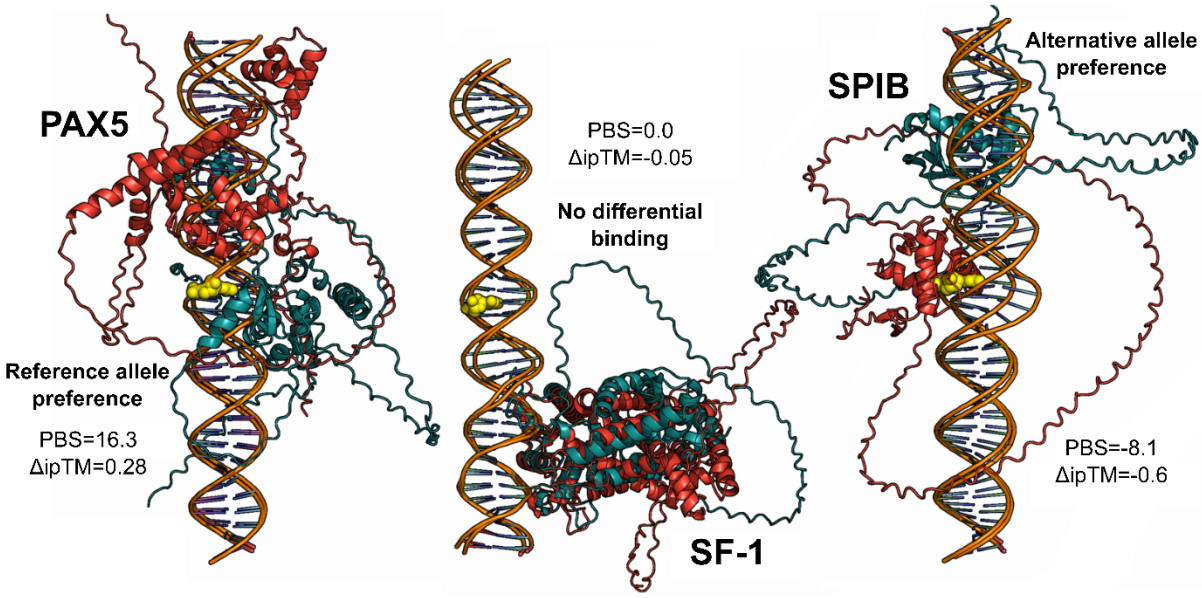
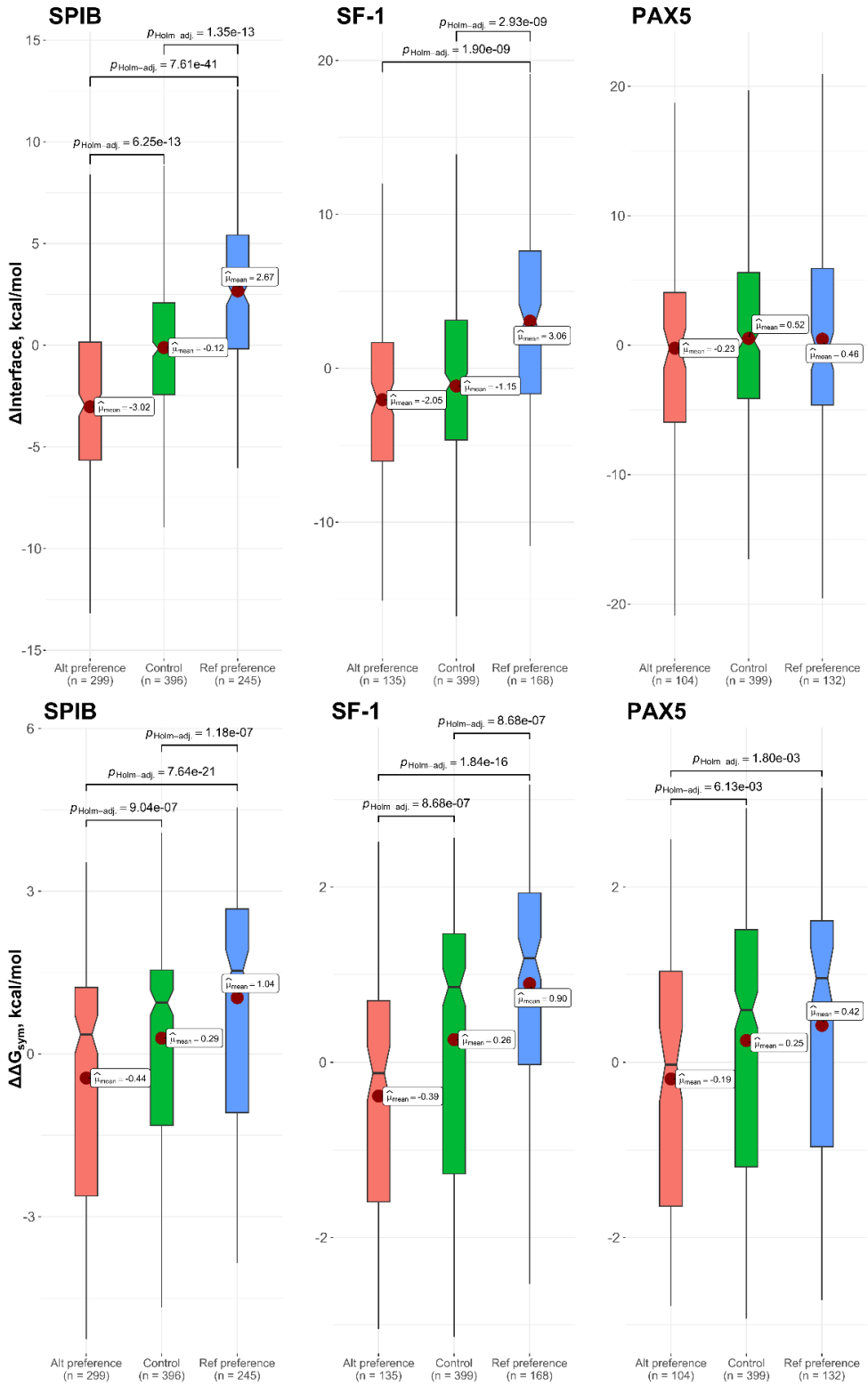


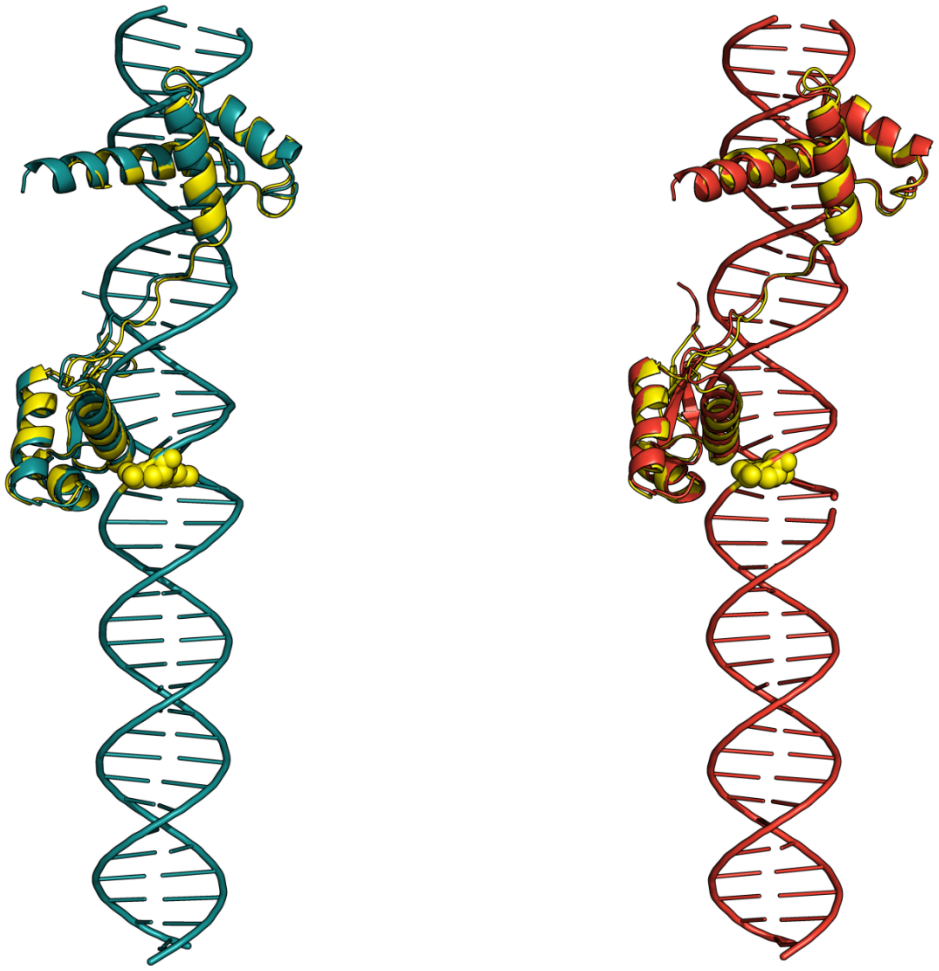
Figure 9. Performance comparison of structure-derived TF-DNA binding metrics. Variants were categorized into non-hit controls and differentially bound alleles, adjusting the scores of ΔipTM, $\Delta\Delta G_{sym}$ and ΔInterface metrics for alternative allele preference pairs to match the direction of reference allele preference. ROC AUCs were calculated using DeLong's method.



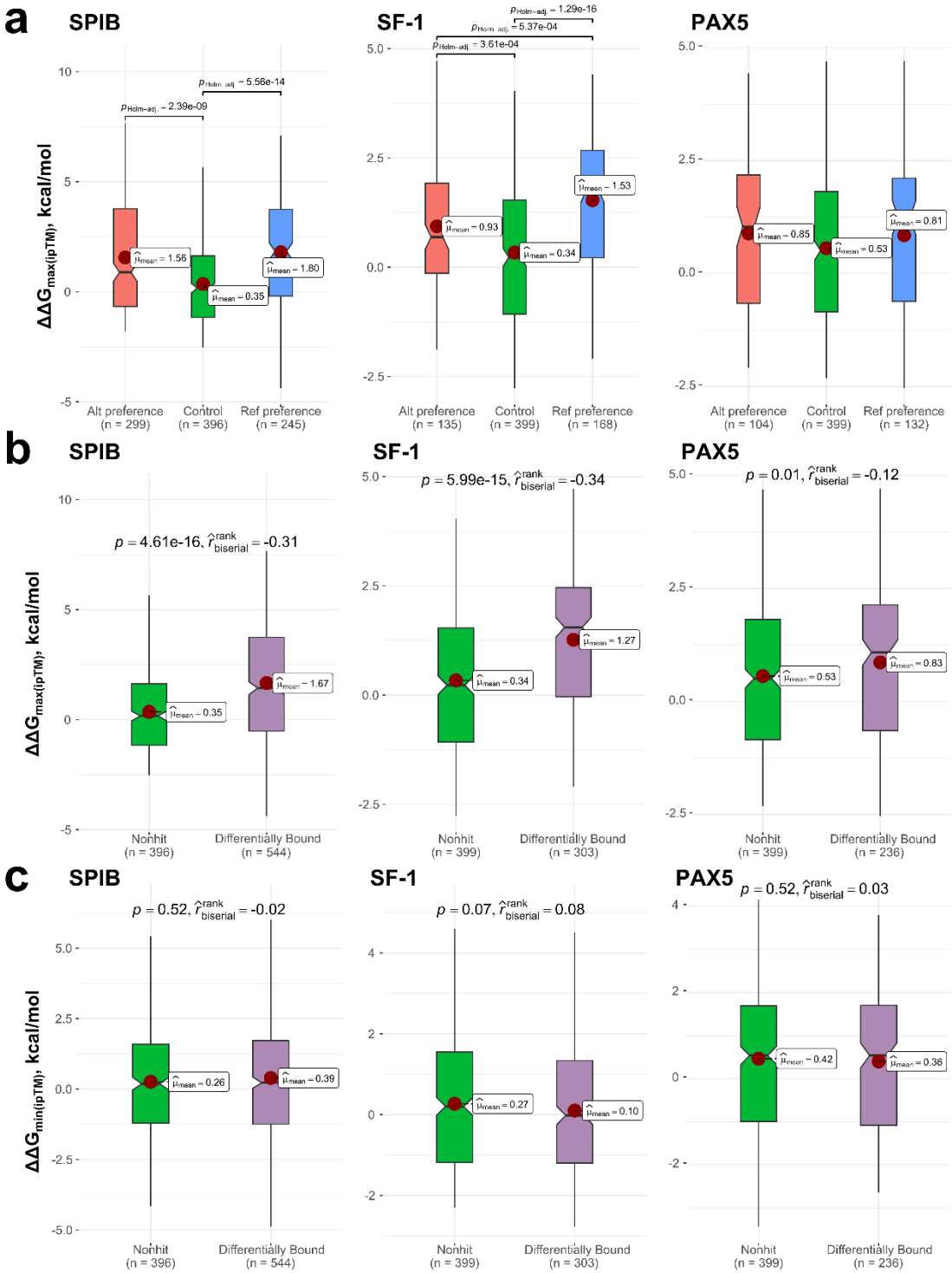
Supplementary Figure 1. AlphaFold 3 provides highly interpretable predictions for assessing the impact of non-coding SNVs on transcription factor binding. The structural complex models were predicted using the AF3 server. In each case, blue TF chains represent the binding pose and site of the reference DNA sequence, while the red structures show the predicted binding site for the alternative allele. The yellow orbs represent the SNV location on the forward strand. PBS values are the preferential binding scores from the original batch SNP-SELEX experiments, while ΔipTM is the difference in the AF3 interface modelling score between the reference and alternative allele complex structures (in the range of -1 to 1). In both cases, positive values indicate reference allele preference, negative values indicate binding bias for the alternative allele, while values close to zero indicate no differential binding.



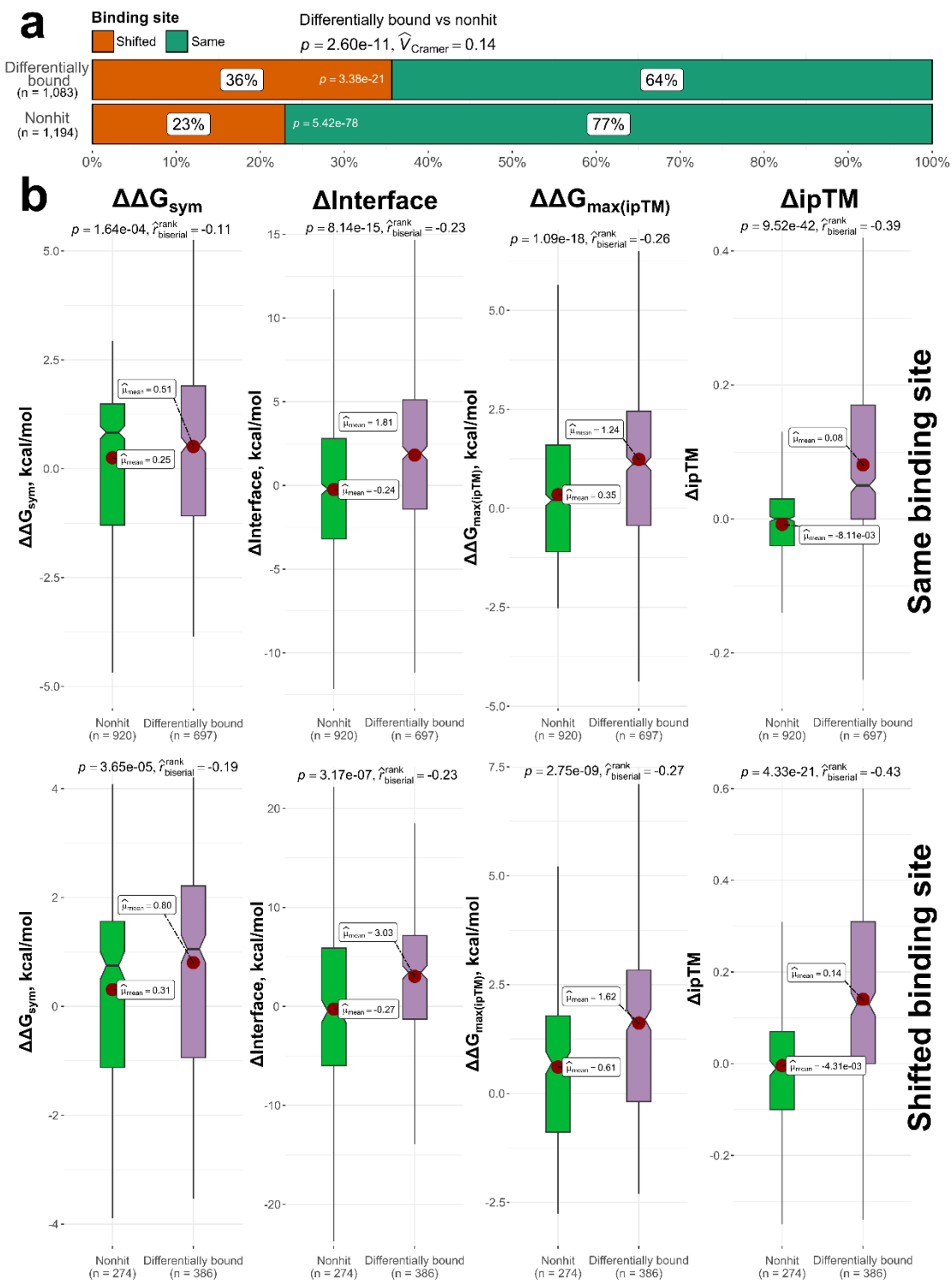
Supplementary Figure 2. Heterogeneity in differential binding prediction performance across TFs in FoldX-derived metrics. Boxes denote data within 25th and 75th percentiles, and contain median (middle line) and mean (red dot) value notations. Whiskers extend from the box to furthest values within 1.5x the inter-quartile range. Significance values are derived using two-sided Holm-corrected Dunn's tests. Sample size (n) indicates the number of allele pairs.



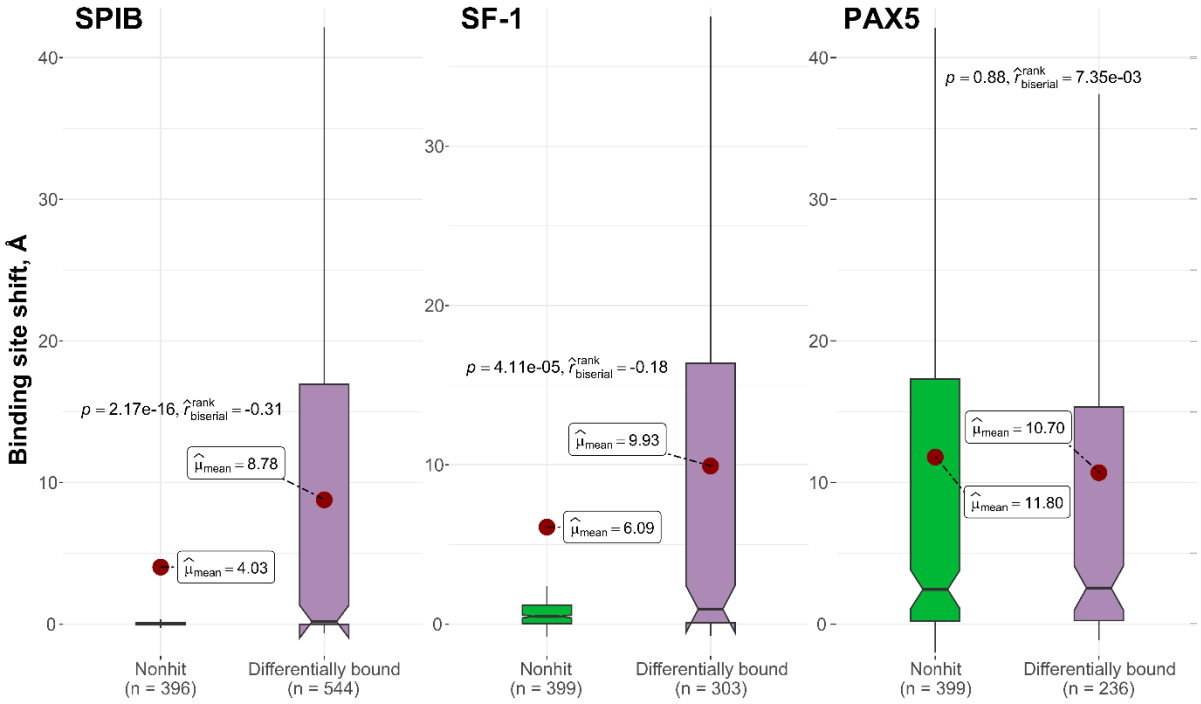
Supplementary Figure 3. PAX6 paired domain models predicted by AF3 are considerably influenced by PDB templates. The teal (left) and red (right) models represent the reference and alternative sequence PAX6-DNA models for the chr11:31685945 G>T variant, respectively. Yellow indicates an aligned PAX6 paired domain-DNA complex taken from the PDB (PDB ID 6PAX).



Supplementary Figure 4. ΔΔG calculated for only the top ipTM model in the allele pair is predictive of differential binding. Boxes denote data within 25th and 75th percentiles, and contain median (middle line) and mean (red dot) value notations. Whiskers extend from the box to the furthest values within 1.5x the inter-quartile range. Significance values are derived using the two-sided Holm-corrected Dunn's test (**a**) or two-sided Mann-Whitney *U* tests (**b-c**). Sample size (n) indicates the number of allele pairs.



Supplementary Figure 5. Differentially bound allele pairs more often show shifted binding sites, but the performance of ΔipTM and most other metrics is not notably affected by binding site shifts. a – comparison of shifted TF binding sites in allele pair models between differentially and non-differentially bound pairs; **b** – performance of all other structural differential binding metrics according to site shift class. Bar plot p-values were derived in two-way and one-way table scenarios using the chi-squared test (**a**). Boxes denote data within 25th and 75th percentiles (**b**) and contain median (middle line) and mean (red dot) value notations. Whiskers extend from the box to the furthest values within 1.5x the inter-quartile range. Significance values are derived using two-sided Mann-Whitney *U* test. Sample size (n) indicates the number of allele pairs.



Supplementary Figure 6. Distance of binding site shifts between reference and alternative sequence models shows heterogenous predictive power for differential binding. Boxes denote data within 25th and 75th percentiles, and contain median (middle line) and mean (red dot) value notations. Whiskers extend from the box to the furthest values within 1.5x the inter-quartile range. Significance values are derived using two-sided Mann-Whitney *U* test. Sample size (n) indicates the number of allele pairs.