nature portfolio

Peer Review File

TransBind allows precise detection of DNA-binding proteins and residues using language models and deep learning

Corresponding Author: Dr Md Shamsuzzoha Bayzid

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

A version of this paper was originally rejected for publication by Communications Biology, however that decision was reconsidered after appeal by the authors.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
It is my pleasure to express my gratitude to the author for his or her hard work on this study. In spite of this, I would like to point out below the points that need to be clarified and improved in order to make the article clearer and more convincing in the future.

1. The proposed method is expected to be superior to previous studies utilizing PSSM and multiple alignments in terms of performance and prediction time. However, in the results section, the study only provided execution time for independent data sets and could not compare it with previous research methods which based on alignment. It would be helpful if the author provided this comparison instead, or at least gave a table showing how long each alignment-based method took to execute. By doing so, we will be able to emphasize the point of the proposed method.

2. Based on Table 5, the proposed model provides an improvement in sensitivity while maintaining the same level of accuracy as the previous method. That result, however, also indicates a different balance between sensitivity and specificity. It is important to provide balanced results for the two units above, so that previous methods can be compared in an even manner.

3. I suggest re-designing figures 1 and 2 in a more professional manner, even though they are clear. Make sure the Local Extraction module and the Classification module are clearly defined.

4. Aims to contribute to the research community by providing an innovative method for predicting DNA-binding proteins and residues. It would be best if the team published the tool as a web application along with some examples of input.

Reviewer #2

(Remarks to the Author)
This manuscript introduces a sequence-based method to predict DNA-binding residues and DNA-binding proteins, which falls in the general approaches of deep learning. Specifically, the proposed method adopts a widely-used pre-trained protein language model ProtTrans to extract sequence features, which are subsequently fed into a CNN module to learn the binding patterns. It is worth noting that similar strategies have already been employed in the same tasks in other studies, thus diminishing the novelty of this work. In addition, the benchmark results are incomplete and unconvincing. Detailed comments regarding these concerns are provided below.

1. The proposed method is deemed simplistic and does not exhibit any significant novelty. The main contribution of this work lies in the substitution of the MSA features (PSSM and HMM) with embeddings from ProtTrans. However, ProtTrans has already been extensively employed in the predictions of binding sites, as evidenced by previous studies including [PMID: 37824738], [PMID: 38015872], [PMID: 34903827], and [https://doi.org/10.1093/bib/bbad488].

2. A substantial number of state-of-the-art methods in this field are missing, both in the benchmark studies and the Introduction section. For example, GLMSite [PMID: 37824738], GeoBind [PMID: 37070217], GraphBind [PMID: 33577689], GraphSite [PMID: 35039821], NucBind [PMID: 30169574], NCBRPred [PMID: 33454744], DRNApred [PMID: 28132027], and DNAPred [PMID: 30943723] should be included.
3. The datasets used in this work were collected from previous studies published before 2016. The authors should collect a new dataset encompassing the recently resolved proteins.
4. The authors adopted an independent test set comprising 41 proteins to benchmark the DNA-binding site predictors. However, I contend that this test set's size is relatively small, thereby undermining the persuasiveness of the obtained results. More importantly, the authors did not remove similar proteins between the test set and the training sets, nor did they eliminate redundant proteins within the test set, leading to potential bias results.
5. The absence of ablation studies in this work raises uncertainty regarding the underlying reasons for the superiority of the authors' proposed model over other competing methods. Without such analyses, the specific factors or components contributing to the observed performance remain unclear.
6. The metrics used in the benchmark are incomplete. Area under the receiver operating characteristic curve (AUC) and area under the precision-recall curve (AUPR) are two widely used indicators in this field, which should be included. Notably, AUC and AUPR are threshold-independent, thus reflecting the overall performance of a model.
7. Almost all state-of-the-art DNA-binding site predictors mentioned above offer the convenience of a web server interface, enabling seamless utilization by biologists without specialized programming expertise. Therefore, I strongly advocate for the authors to consider the development of a web server alongside the standalone program.

Reviewer #3

(Remarks to the Author)
The authors develop a tool called TransBind, which combines the features generated by protein language model with a deep learning framework to predict DNA-binding residues. In the past twenty years, extensive efforts have been devoted into this field. By contrast, the idea proposed by this work is not so compelling. Moreover, there are ambiguities in data processing, model building, and performance evaluation, so the reliability of the results needs to be further evidenced. My comments are provided as follows:

1. The authors emphasize that TransBind could be applied to orphan proteins in both the introduction and discussion sections. Please explicitly display its performance on orphan proteins and compare TransBind with other methods.

2. Among existing protein language models, ESM-based embedding usually achieves state-of-the-art performance on various protein-related classification problems. Why did the authors choose ProtT5-XL-UniRef50 to generate embedding features? The comparison between ProtTrans- and ESM-based embedding could be provided.

3. It is not necessary to introduce the framework of ProtT5-XL-UniRef50 in such a detailed way. In this work, ProtT5-XL-UniRef50 is just a tool for feature extraction.

4. The datasets used in this work are relatively old. The authors could collect the latest data deposited in PDB (e.g. from 2020 to 2024) to evaluate their algorithm.

5. The authors neglect two important evaluation metrics, namely, AUC and AUPRC, which reflect the overall performance of the method on different datasets.

6. The datasets for training, validation and testing should be clarified. The redundancy between the training and testing sets should be evaluated.

7. The authors may compare their method with recently developed SOTA methods:
GraphBind (https://doi.org/10.1093/nar/gkab044),
GraphSite (https://doi.org/10.1093/bib/bbab564),
PSTPRNA (https://doi.org/10.1093/bioinformatics/btac078),
NABind (https://doi.org/10.1371/journal.pcbi.1011428).

8. Why are different competing methods shown in Tables 2-4? The reason for the better performance of your model should be explained.

9. The significance of analyzing the prediction performance on different types of amino acids may be addressed. The reason for different performances on various amino acid types could be discussed.

10. The usefulness of the local feature extractor could be evaluated by the ablation study. Also, the impact of data augmentation on the prediction performance could be evaluated.

11. English correction should be conducted. There are many mistakes in the manuscript, such as "ppositivesamples" and "Table 3.1.2" in section 2.4, "previoslyu" in section 3.3, etc.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I fully accept the revised manuscript.

Reviewer #2

(Remarks to the Author)
1. I have considerable concern regarding the veracity and reliability of the results presented in this article. In the initial version of the manuscript, the authors claimed that they employed “synthetic data generation to tackle data imbalance”. This claim was corroborated by their provided code on GitHub, which confirmed the utilization of the ADASYN (Adaptive Synthetic Sampling) package. However, in the subsequent revised manuscript, the authors made an abrupt shift in their stated methodology, asserting that sampling techniques involving artificial data generation were eschewed due to associated risks, including overfitting and the introduction of synthetic artifacts. Instead, they declared the adoption of a weighted loss function during training. An ablation study, detailed in Table 8, was offered to substantiate the efficacy of this new approach. This inconsistency raises questions about whether a methodological change was indeed implemented between versions. More preposterously, despite the purported change in methodology, ALL the reported results remain entirely unchanged. This discrepancy necessitates clarification to ensure the integrity and reproducibility of the research.

2. The authors have committed a serious factual inaccuracy. Despite acknowledging that numerous preceding studies have utilized ProtTrans, they asserted that, in contrast to these methodologies, their proposed TransBind is a sequence-only method. Furthermore, they stated that CLAPE-DB relies on the availability of the 3D structural information. However, according to the original publication of CLAPE-DB [https://doi.org/10.1093/bib/bbad488], which explicitly mentions “CLAPE-DB did not incorporate any structure information,” it is clear that CLAPE-DB is also a sequence-based method employing ProtTrans, akin to TransBind. This discrepancy highlights a critical misrepresentation that undermines the credibility of the authors' claims, as well as the novelty of their proposed method.

3. In the initial version of the manuscript, the authors claimed that their contribution resided in the substitution of the MSA features with ProtTrans embeddings. As I previously noted, ProtTrans has already been extensively employed in this field, and the authors subsequently acknowledged this, repositioning the principal novelty of their work in the global and local feature extractors. Firstly, it should be noted that such components are not novel within this domain, and the marginal advancements introduced do not warrant publication in a journal under the Nature Research portfolio. Second, the ablation studies provided to validate the contributions of the global and local feature extractors are inadequate. For instance, the impact of removing the global context vector on performance remains unaddressed. Furthermore, there is no evaluation of how the model would perform using only ProtTrans embeddings in conjunction with a simple MLP.

4. The webserver appears to be non-functional. I input a sequence comprising 44 residues and clicked ”SUBMIT”. However, I waited for 2 hours and no result was displayed. When I submitted the example sequence provided by the author, the server returned a job id. However, when I entered this job id and clicked “FETCH OUTPUT”, the server displayed a "Processing..." message indefinitely.

Reviewer #3

(Remarks to the Author)
Although the authors have remarkably improved their manuscript, the comparison of their method and the recently proposed state-of-the-art methods is not fully addressed. Methods developed between 2020 and 2024 (as suggested by reviewers 2 and 3) are missing in the introduction and comparison sections. These methods may outperform the proposed model, but authors should objectively report the comparison results. Comprehensive comparisons are provided in the recent NABind paper, which may provide insights into this work.

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
I maintain my previous stance that the contribution of this work is incremental, as both the CNN-based local feature extractor and the sequence features based on language models are common in the field, as evidenced by studies with PMIDs 32298689 and 37824738. However, the authors have effectively addressed other issues concerning not updating tables, misclassification of one method, and the non-functioning web server.

Reviewer #3

(Remarks to the Author)
I have no more questions.

(The new/modified material in the revised manuscript is provided in blue text to make it easy to identify)

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

It is my pleasure to express my gratitude to the author for his or her hard work on this study. In spite of this, I would like to point out below the points that need to be clarified and improved in order to make the article clearer and more convincing in the future.

Response: Thank you for your encouraging comments. We have adequately addressed the points you raised. We believe you will find the revised manuscript more convincing and publishable.

1. The proposed method is expected to be superior to previous studies utilizing PSSM and multiple alignments in terms of performance and prediction time. However, in the results section, the study only provided execution time for independent data sets and could not compare it with previous research methods which based on alignment. It would be helpful if the author provided this comparison instead, or at least gave a table showing how long each alignment-based method took to execute. By doing so, we will be able to emphasize the point of the proposed method.

Response:

Thank you for your suggestion. Comparing the running time of TransBind with alignment-based methods is indeed important. We have now conducted an analysis to compare these times. Specifically, we performed multiple sequence alignment (MSA) based feature extraction (i.e., PSSM generation) for all four datasets (PDNA 224, PDNA 316, PDNA 543, and PDNA 41). We recorded the time required to generate MSA-based features for all sequences in each dataset.

In Section 5 and Table 12, we report the time required for feature generation in TransBind and the corresponding time for generating MSA-based features. Additionally, we include the inference time for TransBind.

Remarkably, even without considering the inference time for other methods, the PSSM feature generation time alone for those methods is significantly higher than for TransBind. The ratios of running times between TransBind and generating PSSM features for the four datasets (PDNA 224, PDNA 316, PDNA 543, and PDNA 41) are approximately 0.0013%, 0.0007%, 0.0005%, and 0.0097%, respectively. This indicates a drastic improvement in running time, demonstrating TransBind's efficiency and applicability in fast yet accurate prediction.

2. Based on Table 5, the proposed model provides an improvement in sensitivity while maintaining the same level of accuracy as the previous method. That result, however, also indicates a different balance between sensitivity and specificity. It is important to provide balanced results for the two units above, so that previous methods can be compared in an even manner.

Response: Thank you for your valid concern regarding the need for more balanced metrics beyond sensitivity and specificity. As we discussed in Section 3.1 (Table 1), the dataset is highly imbalanced, meaning that metrics like accuracy, sensitivity, or specificity alone cannot provide a comprehensive comparison with other methods. Although TransBind demonstrates high specificity compared to most methods, as noted by the reviewer, its sensitivity score is lower.

To address this issue, we have included the Matthews Correlation Coefficient (MCC) in Table 5. The MCC is a balanced metric that considers both sensitivity and specificity and is widely used for evaluating imbalanced datasets (Chicco et al., BioData mining, 2021). Table 5 shows that TransBind achieves the highest MCC score (0.43), representing a 36.5% improvement over the second-best model, iProDNA (0.315). Moreover, we have now included AUC and AUPR – two threshold-independent metrics. TransBind has achieved impressive AUC and AUPR across all the datasets analyzed in this study.

3. I suggest re-designing figures 1 and 2 in a more professional manner, even though they are clear. Make sure the Local Extraction module and the Classification module are clearly defined.

Response:

We appreciate your suggestion to redesign Figures 1 and 2 to enhance the professionalism of the opening diagram. In response, we have significantly updated Figure 1 to include all the major components of the TransBind architecture and its training strategies. In the revised manuscript, we have omitted the synthetic data generation process due to potential drawbacks such as overfitting and the introduction of synthetic artifacts when using sampling techniques

for class imbalance, and have introduced a weighted loss function that assigns greater weight to minority class samples. This entire pipeline is now presented in Figure 1.

Additionally, we have clearly shown the global feature extraction module (Figures 1(a) and 1(b)), the separation of residues (Figure 1(c)), the local feature extraction module (Figure 1(d)), the training strategy for the local feature extractor (Figure 1(d')), and the classification module (Figure 1(e)) in the updated diagram.

As we have excluded the synthetic data generation process, we remove figure 2 from our manuscript. We hope you find the new figure clearer and more professional.

4. Aims to contribute to the research community by providing an innovative method for predicting DNA-binding proteins and residues. It would be best if the team published the tool as a web application along with some examples of input.

Response:

Thank you for this excellent suggestion. Developing a web application for TransBind is indeed crucial to making our contributions widely accessible. We have now developed a web server and web application that allows users to submit protein sequences and identify nucleic acid-binding residues and proteins from the sequences themselves. The web server is available at https://trans-bind-web-server-frontend.vercel.app/.

Please note that while this web server is fully functional and serves its intended purpose, we are actively working on further enhancements.

Reviewer #2 (Remarks to the Author):

This manuscript introduces a sequence-based method to predict DNA-binding residues and DNA-binding proteins, which falls in the general approaches of deep learning. Specifically, the proposed method adopts a widely-used pre-trained protein language model ProtTrans to extract sequence features, which are subsequently fed into a CNN module to learn the binding patterns. It is worth noting that similar strategies have already been employed in the same tasks in other studies, thus diminishing the novelty of this work. In addition, the benchmark results are incomplete and unconvincing. Detailed comments regarding these concerns are provided below.

Response:

Thank you for your comments and suggestions. We have addressed your comments and hope you will find the revised manuscript more convincing.

1. The proposed method is deemed simplistic and does not exhibit any significant novelty. The main contribution of this work lies in the substitution of the MSA features (PSSM and HMM) with embeddings from ProtTrans. However, ProtTrans has already been extensively employed in the predictions of binding sites, as evidenced by previous studies including [PMID: 37824738], [PMID: 38015872], [PMID: 34903827], and [https://doi.org/10.1093/bib/bbad488].

Response:

We acknowledge that TransBind may not appear to be a highly sophisticated method. However, despite its simplicity, we have effectively employed various concepts that make TransBind both fast and highly accurate.We also acknowledge that the novelty of our proposed approach was not presented well in our original manuscript. We have now updated Section 2 and we hope the novelties of our proposed approach are clear in the revised manuscript. TransBind is a sequence-only model, and while the application of language models (e.g., ProtTrans-based features) is not new in this field, the existing methods do not achieve as high and balanced performance as TransBind. It is important to note that TransBind addresses both protein-level and residue-level predictions, achieving remarkable performance in both tasks.

The primary novelty of our work lies in the way we use global and local feature extractors and the separation of amino acids in the local extractor. Furthermore, unlike other methods, we have implemented an appropriate weighted loss function to train on the imbalanced dataset in this revised version.

To elaborate, TransBind first generates global features using a protein language model, allowing each amino acid to consider all its surrounding amino acids within the language model's context length. TransBind then separates the amino acids, treating each one as a single data point. Our approach effectively addresses the class-imbalance issue by treating each residue separately and assigning more weight to the less-represented class. We detail this approach in section 2.2.4 of our updated manuscript, "Training with Weighted Loss Distribution for Tackling Class Imbalance." This application of a weighted loss function not only addresses potential drawbacks such as overfitting and the introduction of synthetic artifacts when using sampling techniques for class imbalance but also improves the performance of our method. Moreover, we handle protein-level predictions of nucleic acid binding with a learned weighted pooling.

Thus, while TransBind may seem simple, it incorporates notable novelty and is effectively designed to offer a fast, highly accurate sequence-only method for both interacting residue and protein predictions. Many other methods (including those mentioned by the reviewer) rely on 3D structural information, in addition to MSA-based features. However, often the experimentally determined 3d structures are not available for newly discovered sequences. Consequently, some methods use structural prediction frameworks like AlphaFold3 to make predictions of structure first and then use that for downstream tasks, but that is a multi step and time consuming approach. In contrast, TransBind, being a sequence-only method, does not require MSA-based features or any structural information, yet it achieves better performance than those methods.

All of the four methods you mentioned use 3D structural information in addition to language model-based features and/or MSA-based evolutionary features. For example, one of the methods that you suggested, namely CLAPE-DB (https://doi.org/10.1093/bib/bbad488) is a recent method that uses pretrained language models as well as 3D structural information. Based on your and Reviewer-3's suggestions, we have now compared our method with a number of recent methods including CLAPE-DB (please see Section 3.4 and Table 7).

TransBind achieved an MCC score of 0.470 on the DNA-129 dataset, whereas that of CLAPE-DB is 0.389. TransBind's better and more balanced performance compared to other methods is primarily due to the way we separated amino acids in the local extractor and the application of

a weighted training scheme. We have discussed these results and the updated process flow of TransBind in our revised manuscript. We believe you will find the updated version of TransBind and the revised manuscript compelling.

2. A substantial number of state-of-the-art methods in this field are missing, both in the benchmark studies and the Introduction section. For example, GLMSite [PMID: 37824738], GeoBind [PMID: 37070217], GraphBind [PMID: 33577689], GraphSite [PMID: 35039821], NucBind [PMID: 30169574], NCBRPred [PMID: 33454744], DRNApred [PMID: 28132027], and DNAPred [PMID: 30943723] should be included.

Response:

Thank you for pointing this out. Based on your suggestion, we have added a new section in the updated manuscript (Section 3.4: "Comparison with Structure-Based Prediction Approaches") where we compared TransBind with most of the methods you mentioned. Following your suggestion to use a newer dataset for validation (comment 3), we have evaluated TransBind's performance on a recent dataset (DNA-129) and compared it with GraphBind (Xia et al., Nucleic Acid Research, 2021), NucBind (Su et al., Bioinformatics, 2019), DNAPred (Yan et al., Nucleic Acid Research, 2017), targetDNA (Hu et al.,IEEE/ACM transactions on computational biology and bioinformatics, 2016), and DNABind (Lui et al., PROTEINS: structure, Function, and Bioinformatics, 2013). Some other methods that do not explicitly focus on nucleic acid binding are not included in the comparison.

As shown in Table 7 of the updated manuscript, TransBind achieves the second-best performance in F1 score, MCC score, and AUC score, while outperforming the current best model, GraphBind, in precision. For the F1 score, which is an overall balanced metric for classification performance on imbalanced datasets, TransBind achieves 0.51 compared to GraphBind's 0.52, a difference of only 1.8%. Compared to DNABind, TransBind shows a performance improvement of about 16% in F1. Additionally, TransBind achieves the highest precision score, with a 12% improvement over the second-best method, GraphBind. Moreover, TransBind significantly outperformed CLAPE-DB (one of the most recently published structure dependent methods for DNA binding protein residue prediction) across all evaluation metrics.

It is important to note that GraphBind and all the other methods mentioned above use 3D structural information to predict nucleic acid binding residues. However, even though

TransBind is a sequence-only model that is independent of MSA and structural information availability, it performs very competitively with these structure-dependent methods.

3. The datasets used in this work were collected from previous studies published before 2016. The authors should collect a new dataset encompassing the recently resolved proteins.

Response: Thank you for another excellent suggestion. Based on your suggestion, we have now used two benchmark datasets of nucleic-acid-binding proteins from the BioLiP database (Yang et al., Nucleic acids research, 2012) to evaluate the performance of TransBind. These datasets, divided into training and test sets based on release dates, are available at {http://www.csbio.sjtu.edu.cn/bioinf/GraphBind/datasets.html} . The training set includes 573 DNA-binding and 495 RNA-binding protein chains, while the test set consists of 129 DNA-binding and 117 RNA-binding protein chains. Sequence and structural similarities were assessed using bl2seq and TM-align. The datasets were curated from the BioLiP release on December 5, 2018. They have been used by several methods, including GraphBind, TargetDNA, NucBind, CLAPE-DB, and DNABind. Remarkably, in addition to DNA-binding protein residue prediction, we have included an RNA-binding dataset to evaluate the robustness of TransBind. Please see Section 3.4 and Table 7. These results indicate that TransBind provides fast and reliable nucleic-acid binding predictions without relying on structural data, offering a practical solution for computational proteomics. It operates efficiently even with limited resources, meeting the growing demand to assess protein functionality as new sequences are regularly discovered.

4. The authors adopted an independent test set comprising 41 proteins to benchmark the DNA-binding site predictors. However, I contend that this test set's size is relatively small, thereby undermining the persuasiveness of the obtained results. More importantly, the authors did not remove similar proteins between the test set and the training sets, nor did they eliminate redundant proteins within the test set, leading to potential bias results.

Response:

We appreciate your feedback regarding the size of the PDNA-41 dataset. We chose this dataset because it is one of the most widely used benchmark dataset which has been analyzed by many prior studies. In response to your suggestion, we have now included two more recent independent testset: DNA-129 and RNA-117, which contain 129 and 117 non-redundant sequences, respectively, compiled and used by a recent study ((Xia et al., Nucleic Acid Research,

2021)). We have now substantially extended the evaluation study by incorporating these new datasets and comparing our results with recent methods that utilize evolutionary and structural information. We hope you will find the revised evaluation study both convincing and compelling.

Among the three benchmark test sets (PDNA-41, DNA-129, and RNA-117) we analyzed in this study, DNA-129 and RNA-117 consist of non-redundant sequences, ensuring that there is no more than 30% sequence similarity within the test sets or between the training and test data. However, neither the original study that compiled the PDNA-41 dataset nor subsequent studies have examined the sequence similarity between PDNA-41 and the corresponding training data. Therefore, we performed a CD-HIT analysis on the PDNA-41 dataset and the corresponding training datasets: PDNA-224, PDNA-316, and PDNA-543 using a 30% sequence similarity threshold. We identified 4 sequences (out of the 41 sequences) with more than 30% sequence similarity with the training sets. Consequently, we evaluated TransBind on the entire PDNA-41 dataset (to ensure a fair comparison with prior studies that reported their results on the entire set of 41 sequences) as well as on the 37 non-redundant sequences after removing these four sequences.

We have clearly discussed the redundancy information of the datasets in the revised manuscript. We reported the performance of TransBind on bothe 41 and 37 sequences (see Table 5 in the revised manuscript).

5. The absence of ablation studies in this work raises uncertainty regarding the underlying reasons for the superiority of the authors' proposed model over other competing methods. Without such analyses, the specific factors or components contributing to the observed performance remain unclear.

Response: Thank you for this nice suggestion. We have now performed an extensive ablation study to identify the critical algorithmic contributions which resulted in a superior performance (please see Sec 3.6 in the revised manuscript). We performed the ablation study in two directions: 1) ablation on local feature extraction module, and 2) ablation to assess the impact of class-weighted training scheme.

Overall, we see from the ablation study that introduction of loss function based on class weights significantly contributed to the superior performance of TransBind which allowed to effectively tackle the class imbalance issue during the training.

6. The metrics used in the benchmark are incomplete. Area under the receiver operating characteristic curve (AUC) and area under the precision-recall curve (AUPR) are two widely

used indicators in this field, which should be included. Notably, AUC and AUPR are threshold-independent, thus reflecting the overall performance of a model.

Response: Thank you for suggesting to include AUC and AUPR. These are indeed important measures of performance. We have now included AUC for the two new datasets (DNA-129 and RNA-117) that we analyzed based on your suggestion (Section 3.4, Table 7 in the revised manuscript). Notably, TransBind achieved significantly higher AUC scores than all methods using structural information and MSA-based features, with the exception of GraphBind, which slightly outperforms TransBind in terms of AUC.

Following your recommendation, we also computed AUC and AUPR values for the four datasets analyzed in our original version (PDNA-224, PDNA-316, PDNA-543, and PDNA-41). Please refer to Tables 2, 3, 4, and 5. TransBind achieved impressive AUC and AUPR values across all these four datasets. It's important to note that other methods did not report AUC values, except for EL_PSSM (Zhou et al., BMC bioinformatics, 2017) on the PDNA-224 dataset and EC_RUS (Shen et al., Molecules, 2017) for PDNA-543. Similarly, none of the earlier methods reported AUPR values, so we could not provide a comparative evaluation based on AUPR. However, as you suggested, we have computed and reported AUPR values for TransBind in the revised manuscript (see Table 6 in the revised manuscript).

7. Almost all state-of-the-art DNA-binding site predictors mentioned above offer the convenience of a web server interface, enabling seamless utilization by biologists without specialized programming expertise. Therefore, I strongly advocate for the authors to consider the development of a web server alongside the standalone program.

Response: Thank you for this excellent suggestion. Developing a web application for TransBind is indeed crucial to making our contributions widely accessible. We have now developed a web server and web application that allows users to submit protein sequences and identify nucleic acid-binding residues and proteins from the sequences themselves. The web server is available at https://trans-bind-web-server-frontend.vercel.app/.

Please note that while this web server is fully functional and serves its intended purpose, we are actively working on further enhancements with more features.

Reviewer #3 (Remarks to the Author):

The authors develop a tool called TransBind, which combines the features generated by the protein language model with a deep learning framework to predict DNA-binding residues. In the past twenty years, extensive efforts have been devoted into this field. By contrast, the idea proposed by this work is not so compelling. Moreover, there are ambiguities in data processing, model building, and performance evaluation, so the reliability of the results needs to be further evidenced. My comments are provided as follows:

Response: Thank you for your comments and suggestions. We have addressed all of your comments and we believe you will find the revised manuscript more convincing.

1. The authors emphasize that TransBind could be applied to orphan proteins in both the introduction and discussion sections. Please explicitly display its performance on orphan proteins and compare TransBind with other methods.

Response: Thank you for this excellent suggestion. We have now included a comprehensive and systematic assessment of the impact of available homologous information on the performance of various methods. Please refer to Section 3.5 in the revised manuscript. These results indicate the efficacy and superiority of TransBind for proteins with low homology information compared to the methods that rely on evolutionary features.

2. Among existing protein language models, ESM-based embedding usually achieves state-of-the-art performance on various protein-related classification problems. Why did the authors choose ProtT5-XL-UniRef50 to generate embedding features? The comparison between ProtTrans- and ESM-based embedding could be provided.

Response:

Thank you for this suggestion. We have now included a comparison between ProtTrans- and ESM-based features, as discussed in Section 3.6. We assessed the impact of different language models on the performance of TransBind using two recent benchmark datasets: DNA-129 and RNA-117. Our analysis shows that ProtTrans-based embeddings outperform ESM-based embeddings. This finding aligns with recent comparative analysis studies (Shang,

arXiv:2405.11735, 2024), which demonstrate that ProtT5-XL-UniRef50 performs better in four out of five protein downstream tasks compared to similarly sized ESM-based models.

While ESM models offer larger language models, such as the 15B model of ESM2 or ESM3 with 98B parameter (Hayes, bioRxiv, 2024), these are not feasible for inference or prediction tasks on local machines or typical standalone web servers. Given the competitive or superior performance of ProtTrans, we chose to use ProtT5-XL-UniRef50. It provides better performance than ESM-based embeddings within a similar model size and structure.

3. It is not necessary to introduce the framework of ProtT5-XL-UniRef50 in such a detailed way. In this work, ProtT5-XL-UniRef50 is just a tool for feature extraction.

Response:

Thank you. We agree with you, and have revised the manuscript accordingly. Moreover, we acknowledge that the figures and the description of our method in the original version can be improved. We have now substantially revised Section 2 and we believe this is much improved now. In our revised manuscript, we tried to focus on our contributions which were somewhat missing in our previous version. The primary novelty of our work lies in the way we use global and local feature extractors and the separation of amino acids in the local extractor. Furthermore, unlike other methods, we have implemented an appropriate weighted loss function to train on the imbalanced dataset in this revised version. We believe you will find the revised Section 2 and the updated figure more convincing.

4. The datasets used in this work are relatively old. The authors could collect the latest data deposited in PDB (e.g. from 2020 to 2024) to evaluate their algorithm.

Response: Thank you for another excellent suggestion. Based on you and Reviewer-2's suggestion, we have now used two benchmark datasets of nucleic-acid-binding proteins from the BioLiP database (Yang et al., Nucleic acids research, 2012) to evaluate the performance of TransBind (please refer to Section 3.4). These datasets, divided into training and test sets based on release dates, are available at {http://www.csbio.sjtu.edu.cn/bioinf/GraphBind/datasets.html} . The training set includes 573 DNA-binding and 495 RNA-binding protein chains, while the test set consists of 129 DNA-binding and 117 RNA-binding protein chains. Sequence and structural similarities were assessed using bl2seq and TM-align. The datasets were curated from the BioLiP release on

December 5, 2018. They have been used by several methods, including GraphBind, TargetDNA, NucBind, and DNABind. Remarkably, in addition to DNA-binding protein residue prediction, we have included an RNA-binding dataset to evaluate the robustness of TransBind.

TransBind outperforms all current methods except GraphBind across all metrics on both DNA and RNA datasets. In Table 7, TransBind shows a 12% and 10.3% increase in precision on the DNA and RNA datasets, respectively, with only a 1.1% and 0.3% decrease in MCC scores. The AUC scores for TransBind are slightly lower than GraphBind by 1.09% for DNA and 0.9% for RNA. Compared to CLAPE-DB (a recent structure-dependent method), TransBind significantly surpasses all metrics.

Notably, both GraphBind and CLAPE-DB require 3D structural information. GraphBind also uses computationally intensive graphical data augmentation to address data imbalance, which demands substantial resources. Newly discovered protein sequences without experimentally verified structural information require structure prediction frameworks like AlphaFold3 or ESM-3 before using GraphBind, adding further computational complexity – which cannot be done locally within regular GPU resources without the use of extensive cloud support.

TransBind, in contrast, provides fast and reliable nucleic-acid binding predictions without relying on structural data, offering a practical solution for computational proteomics. It operates efficiently even with limited resources, meeting the growing demand to assess protein functionality as new sequences are regularly discovered.

5. The authors neglect two important evaluation metrics, namely, AUC and AUPRC, which reflect the overall performance of the method on different datasets.

Response:

Thank you for suggesting to include AUC and AUPR. These are indeed important measures of performance. We have now included AUC for the two new datasets (DNA-129 and RNA-117) that we analyzed based on your suggestion (Section 3.4, Table 7 in the revised manuscript). Notably, TransBind achieved significantly higher AUC scores than all methods using structural information and MSA-based features, with the exception of GraphBind, which slightly outperforms TransBind in terms of AUC.

Following your recommendation, we also computed AUC and AUPR values for the four datasets analyzed in our original version (PDNA-224, PDNA-316, PDNA-543, and PDNA-41). Please refer to Tables 2, 3, 4, and 5. TransBind achieved impressive AUC and AUPR values across all these four datasets. It's important to note that other methods did not report AUC values, except for

EL_PSSM (Zhou et al., BMC bioinformatics, 2017) on the PDNA-224 dataset and EC_RUS (Shen et al., Molecules, 2017) for PDNA-543. Similarly, none of the earlier methods reported AUPR values, so we could not provide a comparative evaluation based on AUPR. However, as you suggested, we have computed and reported AUPR values for TransBind in the revised manuscript (see Table 6 in the revised manuscript).

6. The datasets for training, validation and testing should be clarified. The redundancy between the training and testing sets should be evaluated.

Response: We are sorry that we did not clearly discuss these in our original version. We have now updated the manuscript accordingly.

Among the three benchmark test sets (PDNA-41, DNA-129, and RNA-117) we analyzed in this study, DNA-129 and RNA-117 consist of non-redundant sequences, ensuring that there is no more than 30% sequence similarity within the test sets or between the training and test data. However, neither the original study that compiled the PDNA-41 dataset nor subsequent studies have examined the sequence similarity between PDNA-41 and the corresponding training data. Therefore, we performed a CD-HIT analysis on the PDNA-41 dataset and the corresponding training datasets: PDNA-224, PDNA-316, and PDNA-543 using a 30% sequence similarity threshold. We identified 4 sequences (out of the 41 sequences) with more than 30\% sequence similarity with the training sets. Consequently, we evaluated TransBind on the entire PDNA-41 dataset (to ensure a fair comparison with prior studies that reported their results on the entire set of 41 sequences) as well as on the 37 non-redundant sequences after removing these four sequences.

We have clearly discussed the redundancy information of the datasets in the revised manuscript. We reported the performance of TransBind on bothe 41 and 37 sequences (see Table 5 in the revised manuscript).

7. The authors may compare their method with recently developed SOTA methods:

GraphBind (https://doi.org/10.1093/nar/gkab044),

GraphSite (https://doi.org/10.1093/bib/bbab564),

PSTPRNA (https://doi.org/10.1093/bioinformatics/btac078),

NABind (https://doi.org/10.1371/journal.pcbi.1011428).

Response:

Thank you for this suggestion. Reviewer 2 also suggested some recent methods to include in our evaluation study. Based on your and other reviewers' suggestions, we have now compared the performance of TransBind with a number of recent approaches for protein-DNA interaction prediction including GraphBind (Xia et al., Nucleic Acid Research, 2021), NucBind(Su et al., Bioinformatics, 2019), DNAPred (Yan et al.,Nucleic Acid Research, 2017), targetDNA(Hu et al.,IEEE/ACM transactions on computational biology and bioinformatics, 2016), and DNABind (Lui et al., PROTEINS: structure, Function, and Bioinformatics, 2013). We also compared TransBind with the state-of-the-art methods for nucleic acid binding protein residue prediction CLAPE-DB (Liu et al., Briefings in Bioinformatics, 2024) . Please see Section 3.4 and Table 7 in the revised manuscript.

8. Why are different competing methods shown in Tables 2-4? The reason for the better performance of your model should be explained.

Response:

We reported the results of different competing methods on widely used benchmark datasets from prior studies. Different prior studies were evaluated on different benchmark datasets. For example, PreDNA (Li et al., Bioinformatics, 2013), PDRLGB (Qian et al., BMC Bioinformatics, 2021), and El PSSM-RT (Zhou et al., BMC Bioinformatics, 2017) reported their validation performance on PDNA-224 but did not evaluate on PDNA-543. We compiled and reported the available performance evaluation results on each dataset from the literature. Consequently, we do not have a common set of competing methods evaluated across all four datasets presented in these tables. Additionally, since these methods are not available as user-friendly standalone software, we were unable to run them ourselves.

Regarding the reasons for TransBind's superior performance, we acknowledge that the original manuscript did not adequately address this. Following your suggestion, we have now included an ablation study to evaluate the contributions of the different components of our approach.

The improvement mainly arises from our innovative use of global and local feature extractors, the separation of amino acids in the local extractor followed by the application of an appropriate weighted loss function to handle the imbalanced dataset. Thus, while TransBind may seem simple and not so compelling, it incorporates significant novelties and is effectively designed to offer a fast, highly accurate sequence-only method for both interacting residue and

protein predictions. We have discussed this in the revised manuscript. Moreover, we have substantially revised Section 2 (Approach) to better explain different components in TransBind.

9. The significance of analyzing the prediction performance on different types of amino acids may be addressed. The reason for different performances on various amino acid types could be discussed.

Response: Thank you for this suggestion. We have now substantially modified this section (Sec 3.7 in the revised manuscript) with new results/figures, highlighting the significance of this analysis and the reason for high variance in predicting capabilities across different amino acids. We believe you will find the revised section much improved and useful.

10. The usefulness of the local feature extractor could be evaluated by the ablation study. Also, the impact of data augmentation on the prediction performance could be evaluated.

Response: Thank you for this nice suggestion. We have now performed an extensive ablation study to identify the critical algorithmic contributions which resulted in a superior performance (please see Sec 3.6 in the revised manuscript). We performed the ablation study in two directions: 1) ablation on local feature extraction module, and 2) ablation to assess the impact of class-weighted training scheme.

Please note that we have replaced the application of data augmentation by an appropriately designed weighted loss function scheme. We assessed the impact of this class-weighted training scheme. Overall, we see from the ablation study that introduction of loss function based on class weights significantly contributed to the superior performance of TransBind which allowed to effectively tackle the class imbalance issue during the training.

11. English correction should be conducted. There are many mistakes in the manuscript, such as "ppositivesamples" and "Table 3.1.2" in section 2.4, "previoslyu" in section 3.3, etc.

Response: We apologize for having made such mistakes and thank you for pointing these out. We have fixed these and other similar mistakes throughout the manuscript.

(The new/modified material in the revised manuscript is provided in blue text to make it easy to identify)

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

I fully accept the revised manuscript.

Response: Thank you for fully accepting the manuscript. Your comments helped us significantly improve the manuscript.

Reviewer #2 (Remarks to the Author):

1. I have considerable concern regarding the veracity and reliability of the results presented in this article. In the initial version of the manuscript, the authors claimed that they employed "synthetic data generation to tackle data imbalance". This claim was corroborated by their provided code on GitHub, which confirmed the utilization of the ADASYN (Adaptive Synthetic Sampling) package. However, in the subsequent revised manuscript, the authors made an abrupt shift in their stated methodology, asserting that sampling techniques involving artificial data generation were eschewed due to associated risks, including overfitting and the introduction of synthetic artifacts. Instead, they declared the adoption of a weighted loss function during training. An ablation study, detailed in Table 8, was offered to substantiate the efficacy of this new approach. This inconsistency raises questions about whether a methodological change was indeed implemented between versions. More preposterously, despite the purported change in methodology, ALL the reported results remain entirely unchanged. This discrepancy necessitates clarification to ensure the integrity and reproducibility of the research.

Response: Thank you for your comment, and we sincerely apologize for causing the confusion about ADASYN- and weighting-based results. In our previous revised version, we presented the results using our proposed new weighting scheme in Tables 5-11. However, we made the regrettable mistake of not updating some tables (Tables 2, 3, and 4) and a few entries in Table 5 (TransBind-41 Sequences) and Table 8, which still reflected the old ADASYN-based results. You rightly pointed this out and we sincerely apologize for this mistake. This unfortunate oversight made it appear as though we had not conducted the new analysis, which in reality, we did. Your comment, "More preposterously, despite the purported change in methodology, ALL the reported results remain entirely unchanged," reflects this, which we fully acknowledge was

caused by our oversight. We have now updated these Tables in the revised manuscript. The code for our weighting-based scheme is also available at our github repository (https://github.com/TokiTahmid64/TransBind/tree/main). We believe it will not be fair if this correctable mistake from our side overshadows the substantial improvements we have made in the manuscript (including the weighting scheme to handle data imbalance).

As for shifting from artificial data generation to a new weighting-based scheme: we initially used ADASYN, a synthetic oversampling technique to address extreme data imbalance present in the datasets in this domain. While artificial data generation techniques like ADASYN, SMOTE are being used to address data imbalance, they have also been criticized for introducing artificial data due to potential drawbacks such as overfitting and the introduction of synthetic artifacts when using sampling techniques. Therefore, we have been trying to propose a new technique for combating this challenge without having to generate artificial data. Finally, in the revised manuscript, we omitted the synthetic data generation process, and have introduced a weighted loss function, designed to handle data imbalance without generating artificial data.

In addition to updating Tables 2, 3, and 4 in the manuscript, we have also submitted a separate file with only these tables (updated-tables.pdf) for your review. To make it easy to compare and review, in this updated-tables.pdf file, we have included both the new weighting scheme-based (TransBind-weighted) and ADASYN-based (TransBind-augmented) results side by side. These results suggest that our weighting-based approach is competitive with or slightly better than the ADASYN-based data augmentation technique. Thus, we were able to achieve the desired performance without having to generate artificial data. We believe it is a notable methodological improvement. This idea can be used in other related problems in this domain to tackle class imbalance.

We sincerely apologize once again for the oversight in not updating the tables with the new results from our proposed weighting scheme. It is important to emphasize that the new weighting scheme achieves competitive, and in some cases, slightly better performance than the ADASYN-based approach. Therefore, there is no reason to intentionally withhold these updates, and the omission was purely an unfortunate mistake on our part. We hope that with these corrections, any concerns regarding the veracity and reliability of the results are fully addressed. We are also willing to provide any additional data, code, or clarifications to further ensure the transparency and integrity of our study.

2. The authors have committed a serious factual inaccuracy. Despite acknowledging that numerous preceding studies have utilized ProtTrans, they asserted that, in contrast to these methodologies, their proposed TransBind is a sequence-only method. Furthermore, they stated that CLAPE-DB relies on the availability of the 3D structural information. However, according to the original publication of CLAPE-DB [https://doi.org/10.1093/bib/bbad488], which explicitly mentions “CLAPE-DB did not incorporate any structure information,” it is clear that CLAPE-DB is also a sequence-based method employing ProtTrans, akin to TransBind. This discrepancy highlights a critical misrepresentation that undermines the credibility of the authors' claims, as well as the novelty of their proposed method.

Response:

We apologize for this mistake of incorrectly labeling CLAPE-DB as a structure-based method. We compared our method in our revised version with methods that utilize structural information. However, CLAPE-DB, like our method (TranBind), is purely sequence-based. You have rightly pointed out this factual inaccuracy, and we acknowledge this mistake. However, we emphasize that this was an unintentional error. Moreover, our method significantly outperformed CLAPE-DB across all evaluation metrics. So there is no reason to intentionally mislabel CLAPE-DB – it was just a silly mistake/misstatement which can be easily corrected. We have corrected our statement on CLAPE-DB in the updated manuscript. It is to be noted that, we call TransBind a sequence-only model, because it considers only the language model based features and does not require any other features such as evolutionary features (e.g., PSSM, HMM, etc.) and structural information.

It is disheartening for us to see concerns raised about the integrity of our study when the error was a simple misstatement rather than any intentional misrepresentation to mislead. Given the rigor and honesty with which we revised the manuscript, along with the six months of effort spent addressing all concerns from the first round of review, we are deeply disheartened by the subsequent rejection based on these mistakes. Unfortunately, this rejection was mostly due to two silly and correctable, yet damaging, errors on our part, which created the impression of a credibility issue.

3. In the initial version of the manuscript, the authors claimed that their contribution resided in the substitution of the MSA features with ProtTrans embeddings. As I previously noted, ProtTrans has already been extensively employed in this field, and the authors subsequently acknowledged this, repositioning the principal novelty of their work in the global and local feature extractors. Firstly, it should be noted that such components are not novel within this domain, and the marginal advancements introduced do not warrant publication in a journal under the Nature Research portfolio. Second, the ablation studies provided to validate the contributions of the global and local feature extractors are inadequate. For instance, the impact of removing the global context vector on performance remains unaddressed. Furthermore, there is no evaluation of how the model would perform using only ProtTrans embeddings in conjunction with a simple MLP.

Reponse:

While our method may not be highly sophisticated, we believe that our method brings notable contributions to the field. The integration of local and global feature extraction, combined with our proposed weighting scheme to address data imbalance, enables our method to outperform even a number of structure-based methods such as DNABind, DNAPred, and COACH-D. Additionally, TransBind achieves the best performance among sequence-only models, highlighting its practical value and innovation.

Our main objective was to develop a method that delivers fast and reasonably accurate predictions without relying on MSA-based evolutionary features or structural data. Please note that we included a comprehensive and systematic assessment of the impact of available homologous information on the performance of various methods in our last revised version (Section 3.5). These results indicate the efficacy and superiority of TransBind for proteins with low homology information (e.g., orphan proteins) compared to the methods that rely on evolutionary features. Thus, while our method may appear simplistic, it provides an efficient and effective sequence-only solution in the field.

We conducted ablation studies in three key areas to evaluate the impact of:

i) different configurations of the local feature extractors,

ii) the use of a weighted loss function to address class imbalance, and

iii) various global features (comparing ProtTrans vs. ESM).

You have now suggested exploring the effect of removing the global feature vector. However, since our approach is a sequence-only method that relies solely on ProtTrans-based global features, removing them would not be applicable within the context of our proposed method. So, we are sorry that, as we understand it, your comment cannot be addressed.

Additionally, you recommended evaluating the performance of ProtTrans features using a simple MLP. Thank you for this nice suggestion. Following your suggestion, we have now extended our ablation study to include this analysis (please refer to Table 8 in our revised manuscript). As seen in the results, a simple MLP with ProtTrans features (without any local extractor) does not yield satisfactory results, highlighting the effectiveness of using both local and global feature extractors in TransBind, which achieves remarkably better performance with ProtTrans features than a simple MLP with ProtTrans features.

Finally, despite the notable performance of TransBind, fixing the unintentional mistakes (as discussed earlier) and tremendous efforts and improvements over our prior version to address all the reviewers' comments, if you still feel this work is not suitable for a Nature portfolio journal, we will consider it an unfortunate outcome for our team.

4. The webserver appears to be non-functional. I input a sequence comprising 44 residues and clicked ”SUBMIT”. However, I waited for 2 hours and no result was displayed. When I submitted the example sequence provided by the author, the server returned a job id. However, when I entered this job id and clicked “FETCH OUTPUT”, the server displayed a "Processing..." message indefinitely.

Response:

We sincerely apologize for the unresponsiveness of the web server. This is likely due to a temporary power or internet outage. In addition, the web server is under active development for further enhancements, so it could also be due to the ongoing development. But we can confirm that the webserver is fully operational. To prevent future disruptions like those you experienced before, we have now hosted the server to a cloud facility, which can be accessed at https://trans-bind-web-server-frontend.vercel.app/. It typically takes around 2 minutes to receive a response from the web server on the first attempt, as it needs to establish a database connection. However, once the connection is established, subsequent responses are much quicker and no longer require this extra time for database connections.

Reviewer #3 (Remarks to the Author):

Although the authors have remarkably improved their manuscript, the comparison of their method and the recently proposed state-of-the-art methods is not fully addressed. Methods developed between 2020 and 2024 (as suggested by reviewers 2 and 3) are missing in the introduction and comparison sections. These methods may outperform the proposed model, but authors should objectively report the comparison results. Comprehensive comparisons are provided in the recent NABind paper, which may provide insights into this work.

Response:

Thank you for your encouraging comment on the improvements we have made. We hope we were able to address your comments/suggestions you made on our earlier draft.

We thank you for your suggestion to extend the evaluation study by including NAbind. In response to your earlier comments during the first round of review, we made efforts to enhance our evaluation study, although we acknowledge that it may not have been comprehensive enough. We acknowledge the importance of your comment on comparing our method to NAbind, a recent method which has been evaluated against some of the leading methods in the field, such as GraphBind, NucBind, and DNABind. Please note that we had already compared our methods with these leading methods (GraphBind, NucBind, DNABin), in our last revised version. Following your latest suggestions, we have now compared our results with NABind as well (please see Table 7) in our revised manuscript. Our evaluation study has now included 2 new datasets (DNA-129 and RNA-117) which are widely used in recent literature and a number of recent methods that are published after 2021 including GraphBind, NABind, CLAPE-DB etc. We hope these updates meet your expectations and provide a more comprehensive assessment of our method.

NABind achieves the highest scores across all metrics, with the exception of precision on the RNA 117 dataset, where TransBind delivers the best performance. Overall, TransBind remains highly competitive with both NABind and GraphBind, with only marginal differences in performance. It is important to highlight that NABind leverages a combination of language-based features, MSA-derived evolutionary features (PSSM, HMM), and structural

information. In contrast, TransBind is a sequence-only method that relies solely on ProtTrans-based features. Despite this, TransBind outperforms several structure-based methods, including DNABind, DNAPred, NucBind, and COACH-D, demonstrating its effectiveness even without the use of evolutionary or structural data.

It is important to note that structure-based methods require known PDB structure data, and analyzing 3D structural data is computationally more expensive than sequence-based analysis. Moreover, 3D structural information is often lacking for newly discovered proteins. As a result, methods like GraphSite depend on inferring 3D structures using predictive tools such as AlphaFold, a process that is both time-consuming and resource-intensive. Additionally, we like to note here that, based on one of your important comments, we included a comprehensive and systematic assessment of the impact of available homologous information on the performance of various methods in our last revised version (Section 3.5). These results indicate the efficacy and superiority of TransBind for proteins with low homology information (e.g., orphan proteins) compared to the methods that rely on evolutionary features. Thus, TransBind provides fast and reliable nucleic-acid binding predictions without relying on structural data and MSA-based evolutionary features, offering a practical solution for computational proteomics.

Therefore, while TransBind (a sequence-only method) does not outperform NABind, which utilizes both MSA-based features and structural information, it remains highly competitive and represents a significant contribution to the field of sequence-only methods.

(The new/modified material in the revised manuscript is provided in blue text to make it easy to identify)

Reviewer #2 (Remarks to the Author):

I maintain my previous stance that the contribution of this work is incremental, as both the CNN-based local feature extractor and the sequence features based on language models are common in the field, as evidenced by studies with PMIDs 32298689 and 37824738. However, the authors have effectively addressed other issues concerning not updating tables, misclassification of one method, and the non-functioning web server.

Response: Thank you for your comments. We are pleased to hear that we successfully addressed your concerns regarding the manuscript's tables, comparisons with other methods, and the web server functionality. Your valuable feedback has significantly improved the quality of our manuscript, and we sincerely appreciate your insights and suggestions.

We respectfully disagree with the assertion that TransBind's contribution is incremental. While it is true that convolutional neural networks (CNNs) and protein language models (LMs) are being used in the field, their use does not inherently negate the novelty of our work. LMs, especially the ProtBERT, have been extensively used in various fields of computational proteomics. CNNs are also widely used in computational biology. However, TransBind combines the strengths of these distinct architectures in an innovative manner, which is not incremental or not directly built upon existing work, to address critical challenges in DNA-binding residue prediction.

Specifically, TransBind introduces an innovative way, that sets it apart from existing work, to integrate local-global feature extraction by leveraging language models (LMs) to capture global context while incorporating a carefully designed local feature analyzer (with separation of amino acids) based on a stacked inception network. Unlike conventional CNN-based methods, our stacked inception network is fine-tuned through ablation studies to extract rich, residue-level features. Furthermore, we tackle the challenge of class imbalance with an innovative weighted loss function, a first in the field, ensuring balanced and robust performance across all classes.

We now discuss the two studies you mentioned (PMIDs 32298689 and 37824738) in detail below.

PMID 32298689

Predicting protein-peptide binding sites with a deep convolutional neural network (Visual)

Main similarity: Both Visual and TransBind employ local feature extraction (albeit there are significant differences in the local feature extraction module).

Key differences:

1. **Local Feature Extraction Window:** Visual relies on a fixed local sliding window of size 7, which captures features from three residues on either side of the central residue. TransBind, in contrast, uses a window size of 1, allowing finer residue-level granularity. Subsequently, this design enables TransBind to implement weighted class training effectively, improving its ability to handle data imbalance which results in a greater overall MCC score.

2. **Local Feature Extraction Network:** Visual employs a simple CNN for feature extraction. TransBind incorporates a three-layer stacked inception network, which has been validated through extensive ablation studies to optimize performance , and are reported in Table 8, where we vary the number of stacked inception layers as well as compare performance with a simple MLP based prediction head.

3. **Global Features:** TransBind uses protein LMs, such as ProtBERT, which leverage self-attention mechanisms to capture long-range dependencies and encode global context across the sequence. Visual lacks this capability, focusing only on local sequential features without considering distant residue interactions.

4. **Balanced MCC Score Through Weighted Training:** A significant limitation of Visual is its inability to address data imbalance. Visual considers entire protein sequences as inputs, making it challenging to target underrepresented classes (e.g., DNA-binding residues). As a result, Visual achieves a low MCC score of 0.17 on the TS125 dataset, despite higher overall accuracy. TransBind, by treating individual residues as independent data points and applying a weighted loss function, significantly improves MCC scores, demonstrating its balanced performance across classes.

Thus, the only similarity is that both have the local feature extractor module, but TransBind's module is significantly different and improved than that of Visual. Moreover, unlike Visual, we coupled the global feature extractor with a global feature extractor using LM as well as a weighted loss function to address class imbalance. Additionally, there are other notable differences as we have discussed above.

PMID 37824738

Accurately identifying nucleic-acid-binding sites through geometric graph learning on language model predicted structures (GLMSite)

Main similarity: Both GLMSite and TransBind utilize protein LMs to generate global embeddings.

Key differences:

1. **Input modality for the network:** GLMSite requires both sequential data and 3D structural information, generated using ESMFold, whereas TransBind is a sequence-only method. It demands significant computational resources, especially when 3D structures are unavailable. TransBind, in contrast, operates solely on sequence data, making it more computationally efficient.

2. **Method architecture:** GLMSite employs a graph neural network to integrate 3D structural features with sequential embeddings. TransBind, on the other hand, relies on a stacked inception network to process sequence-based features. This design choice makes TransBind simple yet effective, avoiding the need for additional computational steps like structure prediction.

3. **Handling class imbalance:** GLMSite indirectly addresses class imbalance by leveraging structural biases—specifically, the tendency of DNA-binding residues to reside on protein surfaces. Several studies, such as Gainza et al., Nature Methods (2020) and Tubiana et al., Nature Methods (2022), have demonstrated that insights into the surface topology of a protein provide significant information about its binding residues. This is primarily because binding pockets are predominantly located on the protein's surface (Gainza et al., 2020). Consequently, structural data helps mitigate the issue of data imbalance by naturally deprioritizing residues buried within the protein folds when identifying nucleic acid-binding residues. However, these methods are computationally very demanding and heavily depend on the availability of reasonably accurate structural information. Thus, TransBind, in contrast, handles imbalance purely through a learning-based approach, applying a weighted loss function directly to residues without relying on structural biases. This strategy ensures balanced performance across classes while maintaining computational efficiency.

In light of this discussion, we believe that our methodological advancement is neither incremental nor directly built upon existing work. While our approach employs CNNs and LMs, these components are innovatively utilized. The primary novelty lies in the integration of global and local feature extractors,

specifically the separation of amino acids in the local extractor—a concept not previously introduced in the field.

Moreover, we have addressed the critical challenge of class imbalance by implementing a carefully designed weighted loss function, which is unique to our work. Our extensive ablation studies (some inspired by your valuable suggestions) clearly demonstrate that simply employing LMs or conventional CNNs does not yield superior performance. Instead, it is the innovative way of combining LMs, CNNs with a stacked inception network on individual amino acids, and the weighted loss function in our proposed architecture that provides the remarkable performance of our model, outperforming even several structure-based methods such as DNABind, DNAPred, and COACH-D.

Based on your comment and suggestion, we have now precisely clarified how the proposed local feature extractor is distinct from existing methods (Please refer to Section *Local Feature Extraction Using Stacked Inception V2 Modules* in page 16 of the updated manuscript). Additionally, we have acknowledged the prior use of LMs in literature, including the specific paper you mentioned (GLMsite). We believe you will find our response and the revised manuscript convincing.

Reviewer #3 (Remarks to the Author):

I have no more questions.

Response: We are pleased that we have adequately addressed your comments. Your comments helped us significantly improve the manuscript. Thank you very much.