International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 10 | October - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

AI Algorithms for Precision Motif Refinement Enhancement

Prince Joseph
Government Polytechnic College
Pala Kottayam

Abstract:The identification of motifs is a crucial stage in the process of exploring the function of genes in many different systems. The identification of transcription factor binding sites (TFBSs), is an crucial step in the process of comprehending the regulatory mechanisms that regulate gene expression. Motif discovery is an essential step in this process. However, existing motif discovery methods face challenges such as computational complexity, sensitivity to parameters, and assumptions about motif characteristics. As a consequence, we introduced the Modified Freeze Firefly method for motif discovery based on Gibbs sampling, as well as the Harmony Search with Logistic regression for predicting. This research is distinctly focused on harnessing the power of Artificial Intelligence (AI) algorithms tailored to refine and optimize precision in motif identification. The methodology begins with the acquisition of data from the JASPAR database, followed by meticulous preprocessing. Motif discovery is then executed using the Modified Freeze Firefly with Gibbs Sampling (MFF-GS). Subsequently, the identified motifs undergo a precision enhancement and refinement process through the innovative Harmony Search with Logistic Regression (HS-LR). Finally, the performance metrics are rigorously evaluated, encompassing Running time vs. Maximum Motif length, running time vs. Motif length, running time vs. motif count (d), Prediction rate, and Mean Square Error. This comprehensive approach signifies a cutting-edge integration of AI techniques, promising to elevate the accuracy and effectiveness of motif refinement in the context of gene regulation studies.

Index words: **Motifs, Transcription Factor Binding Sites (TFBSs), JASPAR database, Modified Freeze Firefly with Gibbs Sampling (MFF-GS), Harmony Search with Logistic regression (HS-LR)**

I. INTRODUCTION

A main function of motif is to control gene expression at both the transcriptional and posttranscriptional stages. Motifs in DNA and RNA serve crucial roles in a broad range of biological functions, involving but not limited to splicing alternatives, transcription, and translation.Determining patterns in DNA sequences is one of the most challenging tasks in both computer science and molecular biology. Understanding the expression of genes requires the identification of regulatory motifs. The idea that every gene contains the instructions needed to make a protein is fundamental to the study of gene expression.The same protein's binding sites are often conservative, brief sequences known as motifs. Protein binding sites were initially identified using conservative sequencing. Numerous molecular mining algorithms arise when researchers get a better understanding of molecular research [1,2]. Several known protein factors bind to initiate the expression process. They attach themselves to promoter and enhancer sequences as transcription factors. The first step is transcription, which entails making an RNA "copy" of a certain DNA sequence. The second step of the process, called translation, involves reading and interpreting this RNA sequence to produce a protein. The combined effect of these two processes is gene expression. Many regulating transcription factors (TFs), also known as transcription factor binding sites (TFBS), attach to certain DNA regions to control the expression of genes. In the last ten years, a notable method for understanding transcription regulatory networks has emerged: the study of DNA sequence data for computational identification of TFBS. Since sequence motifs are short (about 6–12 bp) and intergenic regions are quite extensive and very variable, finding sequence motifs may be difficult. Sequence motifs have a set size, are regularly repeated, and are preserved. Understanding the processes controlling gene expression is aided by the identification of TF-BSs, which is made possible by these patterns 3. Plant, structured, gapped, sequence, network, and motif categories are available for motifs [3-5]. The majority of early mobile mining techniques fall into two categories: enumeration techniques and probabilistic techniques.Transcription factor binding sites are intimately associated with DNA sequence specificity. By examining DNA sequence specificity, one may build a more comprehensive regulatory model of biological systems and gain insight into the transcription process of DNA. Furthermore, by examining the specificity of DNA sequences and the relationships between various illnesses, investigators have the ability to identify and explain disease variations. Finding diseases or illness patterns in the early phases of a disease has significant consequences for the medical industry and may also be accomplished via sequence-level research. Utilizing computational techniques to investigate the specificity of DNA sequences is becoming more and more crucial with the advent of technologies such as chip SEQ (chromatin immunoprotection sequencing) [6-8]. Utilizing biochemical experimental approaches to investigate the specificity of DNA sequence requires a significant investment of time, money, and Labor. These days, the sample size for data about biology. A depiction of motifs is the position weight matrix (PWM), whose entries show how often each of the four bases occurs at a specific location. The motif level is often used to characterize DNA sequence specificity. It is easy to comprehend how to characterize motif levels, which makes genome-scale binding site scanning faster. This description suggests that motif discovery is the primary method used in current investigations to identify DNA sequence specificity. Numerous motif mining methods have been presented as of recently. These techniques can mine the motif efficiently because of its length. By using deep learning to tackle a few

sequencing problems, this study explores the specificity of DNA sequence[9,10].Weak motif finding is a major problem in computational biology. It is difficult to fix since the actual theme and its altered versions have so many inconsistencies that the real ones might be hidden by misleading signals. Furthermore, because regulatory parts are usually brief and variable, it is difficult to find and identify them using computer algorithms. The issue in trying to tackle the theme-finding problem is to identify conserved and overrepresented motifs from the set of sequences of DNA that are predicted to develop into transcription factor binding sites. A protein known as a transcription factor regulates the expression of genes by controlling the initiation of the transcription process, which uses DNA as a template to create mRNA. The common sequence is referred to as a motif. A transcription factor's binding site "pattern". Identifying themes will aid in understanding illness vulnerability and developing therapeutic interventions[11].This study focuses on using the capability of Artificial Intelligence (AI) algorithms designed to enhance and maximize accuracy in motif recognition.

A. *Motivations and objectives*

The majority of the existing works are in motif discovery and prediction but this results in an increase in complexity, lack of attention to motifs, and lower prediction accuracy of the fragment. Despite focusing on these issues, current research does not yet offer an appropriate solution for precision motif refinement and enhancement. We are motivated by several issues addressed in the previous studies, specifically the following,

- **Inefficient Motif Discovery Algorithms:** The existing methods face challenges in capturing rich motifs, increasing complexity without the explosion of parameters.
- **Reduce the precision of prediction:**Some existing methods face challenges in predicting the accuracy of the fragment, and when the data is very large, it impacts the prediction and subsequent motif mining. It is necessary to enhance the prediction model by looking at other sources of sequenced-derived data.

The main objective of this research using AI algorithms aimed to refine and optimize precision in motif identification.

- Improve the efficiency of motif discovery by employing a modified sampling approach followed by a motif-finding technique, with a focus on reducing complexity.
- Enhance prediction accuracy through a precision enhancement and refinement process, utilizing innovative methods. The aim is to achieve more accurate predictions in relevant applications.

B. *Research Contributions*

The main goal of this research using AI algorithms is to refine and optimize precision in motif identification. The following are some of the research's contributions:

- We collect the JASPAR database and the database is normalized and scaled appropriately.
- For reducing complexity, we propose motif discovery using Modified Gibbs sampling followed by motif finding using the Freeze Firefly algorithms. The novelty is called Modified Freeze Firefly with Gibbs Sampling (MFF-GS).
- To increase the prediction accuracy using the precision enhancement and refinement process through the innovative Harmony Search with Logistic Regression (HS-LR).

C. *Research organizations*

The remainder of the paper is arranged as follows: Section II is a literature assessment that identifies and addresses research gaps. Section III provides a comprehensive summary of the work to be done, including necessary pseudocode and representations. The comparative assessment and research summary are included in Section IV. Section V goes into extensive detail on the predicted work's conclude.

II. LITERATURE SURVEY

According to the author of [12], the Machine Learning Motif Extractor (ML-MotEx) employs Shapley augmented explanations SHAP values to discover model features that are essential for fit quality following an ML algorithm has been trained on several fits. They apply the technique to four distinct chemical systems, such as clusters and disordered nanomaterials. ML-MotEx allows for a kind of modeling in which explainable machine learning is used to give a significant value for each feature in a model based on fit quality. The author of [13] describes the various encoding schemes and machine learning combinations that may be used to anticipate distinct structural/functional themes. The use of protein computational models to encode proteins together with their physicochemical characteristics and developmental information is especially exciting. The most current predictors established for transmembrane classifications, phosphorylation sites, sorting signals, and lipidation can be thoroughly investigated in order to evaluate contemporary facilities, with an emphasis on how effectively protein language models function for various types of positions. This shows that to fully use the potent machine learning techniques now accessible, additional experimental data are required. Particle display (PD) is used by the author of [14] to partition a library of aptamers in accordance with affinity. ML algorithms are then trained on this data to estimate affinities in silico. The approach discovered high-affinity aptamers of DNA using empirical selections at an 11-fold faster rate than random perturbation and created new, high-affinity aptamers at a faster rate than PD alone. Truncated aptamers were made easier to manufacture since they are 70% shorter and have a greater binding affinity (1.5 nM) than the best candidate discovered in investigations.Authors in [15] focused on considerably bigger RNA with lengths up to 3000 nucleotides to expand these findings. They also discover that huge natural and random structures are extremely similar when compared to typical structures taken from the spaces of all conceivable RNA structures, by looking at both abstract forms and structural

motif frequencies. Another finding from the motif frequency research is that machine learning algorithms may effectively categorize natural and random RNA with high accuracy by using the frequencies of various motifs, particularly for longer RNA.The data-efficient author of [16] estimates complex adsorption material binding motifs and related adsorbed energy levels at transition metals (TMs), including their alloys, based on tailored data. The Gaussian Process Regression using the Wasserstein Weisfeiler-Lehman graph kernel. The model demonstrates high prediction accuracy not only for the strain constituent TMs but also for an alloy that is derived from these TMs. Furthermore, an out-of-domain TM may be predicted with minimum integration of new training data. They offer a tool for estimating ensemble uncertainty since they expect the model will be useful in active learning tactics.

The authors of [17] provided the DeepSEED, an AI-assisted system that effectively creates artificial promoters by fusing deep learning methods with expert knowledge. DeepSEED has been shown effective in enhancing the characteristics of constitutive, IPTG-inducible, and doxycycline (Dox)-inducible promoters in Escherichia coli and mammalian cells. Moreover, the findings demonstrate that DeepSEED effectively extracts implicit data from flanking sequences, including DNA shape traits and k-mer frequencies, which are essential for identifying promoter characteristics.

The author of [18] highlighted recent advancements in DNN model interpretation, with an emphasis on its uses in epigenomics and genomics. First, they show the most advanced DNN interpretation techniques in common machine learning domains. Then, they go over the DNN interpretation techniques utilized in modern genomics and epigenomics research, with an emphasis on data- and computationally-intensive subjects including sequence motif identification, gene expression, chromatin interactions, genetic variants, non-coding RNAsand genetic variants. Also provided are the biological conclusions that resulted from various interpretive methods. The author of [19] proposed a faster, enhanced version of the program for Bayesian Markov models, called BaMMmotif2. They tested it on a large number of HTSELEX and ChIP-seq datasets using state-of-the-art molecular discovery methods. In testing across platforms and cell lines, BaMMmotif2 models demonstrated similar improvements over the next best tool without exhibiting any signs of overtraining. These results demonstrate that most TF binding models are significantly improved by dependencies beyond the first level. The author of [20] suggested a method that quantifies the allelic difference of projected epigenetic signals to further assess the functional consequences of noncoding variations on an individual basis. They show that the suggested method can: determine canonical motifs referred to control the transcription of Alzheimer's disease(AD) causal genes; improve the partitioning element of hereditary factor assessment; and rank possible causative variants in a GWAS risk locus. It may also predict quantitative genome-wide epigenetic modification communication in key genomic areas of Alzheimer's disease (AD)-related genes. They do this by following the technique proposed for the cohort of the Religious Orders Study/Memory and Aging Project (ROSMAP) that is investigating AD.The authors of [21] present Explainable Neural Networks (ExplaiNN), a transparent, fully interpretable sequence-based deep learning model for genomic issues that takes inspiration from NAMs. They test ExplaiNN on a variety of tasks and show that it outperforms the most current models, offering both local and global interpretation that is faster and easier than with more complicated approaches. They then demonstrate that the patterns produced by ExplaiNN's convolutional filters match those discovered by de novo techniques on the same data. The authors of [22] recommend using secondary structure fingerprints, which may be separated into two categories: Free energy fingerprints, which depend on a particular repertoire of tiny structural motifs, as well as higher-level representations provided by RNA-As-Graphs (RAG), are two examples. The fingerprints consider the distinctions between local and global structural matching. Additionally, they used K-mers to assess the deep learning architecture. The author of [23] represents the structural motif components of viral genomes as well as various techniques for predicting and characterizing RNA structures. We provide an overview of several research about the genomes of viruses, with a focus on severe acute respiratory syndrome coronavirus (SARS-CoV-2) and influenza A virus (IAV), based on current literature. Here, we highlight how the structure-function link and, therefore, the identification of novel antiviral therapies might be facilitated by a deeper comprehension of the architecture of viral genomes. Authors in [24] determine the regions enclosed by transcription factors (TFs), which are now essential in molecular and cellular biology because of their major function in regulating gene expression.Deep learning (DL)-based techniques have been presented more often in recent years with outstanding prediction performance for identifying TFBSs. Yet, these approaches fall short in precisely identifying motifs and TFBSs and instead primarily concentrate on predicting the sequence specificity of TF-DNA binding, which is an analogous challenge to a binary classification problem at the sequence level. Authors in [24] develops a "Fully Convolutional Network with Global Average Pooling (FCNA)" that can detect motifs and locate TFBSs in detail, making it a nucleotide-level binary classification problem. Since FCNA cannot properly pinpoint TFBSs, FNCA must anticipate some false-positive samples. The trials utilized a high threshold value to eliminate false-positive samples, however this also deleted some true-positive samples. FCNA depends heavily on nucleotide-level labels since it predicts motifs using strongly-supervised label information. Thus, future studies should suggest more thorough solutions to the two challenges. For 5-fold model validation, k-fold cross-validation may not be effective if the dataset has a temporal structure where sample order matters. Folds may leak information due to temporal interdependence. Five-fold cross-validation involves training and testing the model five times, which may be computationally demanding for complicated models or huge datasets.Authors in [25] proposed the Convolutional Auto Encoder and Convolutional Neural Network (CAE-CNN), a unique architecture that combines a convolutional autoencoder with a convolutional neural network. Specifically, using the picture reconstruction

as input, we use a convolutional autoencoder to uncover important characteristics in DNA nucleotides from the positive data. Thus, the learned characteristics will be used by the convolutional neural network in the training phase. To further better capture the properties of DNA nucleotides, we further use a highway link layer and a gated unit. Interpreting the learnt representations might be difficult, and the overall architecture may grow complicated. It could be more difficult to comprehend the model's internal operations and the importance of taught aspects. TFBS activity may show temporal dynamics and rely on the situation. Cross-validation may not be able to capture temporal fluctuations, thus it's important to think about techniques like temporal validation or how to handle time-series data in a cross-validated dataset. This study [26] uses one-dimensional SMILES as the inputs of ligand and binding location residue for protein in order to computational effectively predict unknown ligand-target interactions. Utilizing inputs including motif-rich binding site sequences of peptides and one-dimensional SMILES for medicines, researchers first create a deep learning CNN model. They believe that integrating structural protein information into drug-target interaction prediction models with big datasets would be a beneficial research technique, offering improved interpretability, high throughput, and wide application. Protein structures and chemical compounds are represented by high-dimensional data. Handling these high-dimensional inputs requires substantial computational resources. As the size of the dataset grows, scalability issues may arise, impacting the efficiency of the model. The goal of the authors in [27] was to determine if motifs in the Ceratocystidaceae family could be found using in silico methods, and if they were, whether they correlated with transcription factors that were already known. In order to find patterns, this research focused from the BUSCO dataset. The analytic tools MEME and Tomtom were used to identify conserved motifs at the family level. The findings demonstrate that the Ceratocystidaceae as well as unrelated species might be identified using similar in silico techniques by looking for known regulatory motifs. This work supports further attempts to find motifs using in silico studies. The proposed methodologies were mainly focused on establishing appropriate discovery algorithms and accurate prediction models using AI algorithms.

III. PROBLEM STATEMENT

The author of [28] represents the capacity of a DNA sequence to bind certain proteins known as DNA sequence specificity. These proteins are essential for the control of genes via processes including transcription and alternative splicing. To discover pathogenic variations and develop the biological system's regulatory model, obtaining DNA sequence specificity is crucial. DNA segments that bind to certain proteins have sequence patterns known as motifs. Currently, a few motif mining methods that work well with a specific motif length have been presented. Prediction model construction using the CNN. Regarding the motif level description, this work develops an AI-based technique to forecast the motif's length.

- The results show that the average prediction accuracy of the fragment decreases with the motif's complementation ability, or its capacity to fill in missing data.

The author of [29] proposed a novel attention-based deep convolutional neural network (CNN) model called DeepVISP is created to effectively predict oncogenic virus integration sites (VISs) in the human genome. By autonomously learning useful characteristics and crucial genomic sites just from the DNA sequences, DeepVISP delivers excellent accuracy and robust performance for all three viruses using the carefully selected benchmark integration data. Furthermore, cis-regulatory factors that may be implicated in carcinogenesis and viral integration may be decoded by DeepVISP. There are many lines of evidence in the literature that support these conclusions. The informative motifs clustering study shows that the representative k-mers in clusters may aid in the virus's identification of the host genes. Using DeepVISP, an approachable web server is created to anticipate potential oncogenic VISs in the human genome.

- In identifying and evaluating oncogenic viral integration, the study emphasizes the value of deep learning, in particular convolutional neural networks. Nevertheless, the low number of VISs in certain viruses has an impact on the effectiveness of categorization and the ensuing motif mining.

The authors of [30] investigated the DeepPPF framework's ability to discover rich motifs for functional categories using the fewest sequences of proteins. The findings show that deep learning needs a rich motif identification procedure in order to improve protein family modeling performance. Finally, in order to find the best network architecture for hierarchical level modeling and prediction, transfer the data in the lower hierarchical functioning domain to two target functional levels. The results we have obtained imply that the transfer learning technique might be employed to increase performance.

- Using the DeepPPF to increase the model's prediction capability by examining new sources of sequence-derived information. Quantitative biophysical characteristics, for example, can be considered.
- There is also a requirement to capture more rich motifs for the model by raising its complexity without exploding its parameters.

Research solutions:

The proposed research addresses several challenges in existing motif discovery algorithms, highlighting issues in capturing rich motifs and the impact of large datasets on prediction accuracy. To overcome these challenges, a novel Modified Freeze Firefly Algorithm with Gibbs Sampling (MFF-GS) is introduced, integrating key elements from the Firefly Algorithm, Gibbs sampling, and a freezing mechanism. The algorithm aims to improve motif identification by combining local and global search strategies. Additionally, the research incorporates an AI-based refinement approach using the Harmony Search

algorithm with Logistic Regression, providing an enhanced predictive model construction process. These innovations collectively contribute to more effective motif discovery and improved prediction accuracy in complex biological sequence analysis scenarios.

IV. PROPOSED METHODS

The suggested approaches were primarily concerned with developing suitable discovery algorithms and developing accurate prediction models utilizing AI algorithms. Fig.1 represents the overall architecture of this research. The several processes involved in the proposed work are discussed in this section briefly into three major processes namely,

- ❖ Sample data collection and Data preprocessing
- ❖ Modified Freeze Firefly algorithm for motif identification based on Gibbs sampling (MFF-GB)
- ❖ AI-based refinement enhancement for predicting model construction.

A. Sample Data Collection and preprocessing

The PWM of human transcription factor binding sites was taken from the Jaspar database in this study, and the matching chip SEQ data set was retrieved from the encoded database. JASPAR is an open-access library of curated, non-redundant TF binding profiles for TFs from different species in six taxonomic categories recorded as position frequency matrices (PFMs) and TF flexible models (TFFMs). The below-mentioned link is a database link-https://jaspar.genereg.net/matrix/MA0003.4/Ensure that the PWMs are normalized and scaled appropriately. Inconsistent or missing data should be checked. Verify the sequences and motifs' integrity.

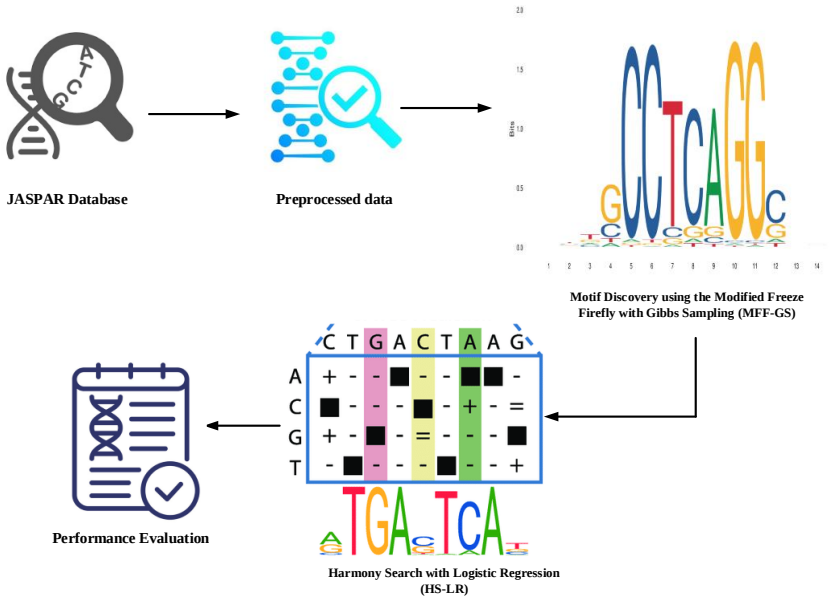

Fig. 1: Overall architecture of this research

B. Gibbs sampling-based modified Freeze Firefly method for motif identification

To construct a modified Freeze Firefly Algorithm for motif discovery based on Gibbs sampling with both local and global search, the essential components of the Firefly Algorithm, Gibbs sampling, and the freezing mechanism must be incorporated. The Gibbs sampling method follows these steps: it is a random algorithm that is a specific instance of the Markov motif:

*Step i:*With a predetermined sequence length, randomly choose a subsequence fragment from every sequence;

Step ii: Use these chosen subsequences to construct PWM;

Step iii: Choose a random subset of the initial sequences;

Step iv: Use the sliding-window approach to score all conceivable sequence motifs using the PWM; each sliding's length unit represents an amino acid or basic;

Step v: The motif that has the highest probability is determined and is thought to be the new motif;

Step vi: The PWM is updated by substituting the maximum-likelihood motif for the original sequence's chosen subsequence;

Stepvii: Continue out the iterative computation until the position weight matrix probability and likelihood score result is constant.

A scoring formula is established for the grading for every sequence string (length l) as the motif assessment criterion in order to assess potential motif effectiveness for motif identification. The similarity among the pattern string and the PWM is described by the scoring function. The candidate pattern strings are closer to motifs when they get a higher score. The following equation defines the scoring function:

$$P = \sum_{j=1}^{l} f_i, jw_i, j \quad (1)$$

In order to guarantee that the algorithm approaches the global optimum, the paper proposes a motif identification firefly method based on Gibbs sampling that combines ideas

from Gibbs sampling and PWM. The algorithm's core principle is as follows: initially, each sequence's starting point is randomly chosen in order to construct a l-tuple, where l is the motif's length. The algorithm's starting population and position weight matrix are made up of these l-tuples. Next, the PWM (sliding window with the sliding length unit of one base) is used to score each sequence's potential mode strings. The next step is to derive the final motif from the Gibbs sampling. TFBSs in the remaining set of sequences (seg2) are then discovered using the MFFF algorithm. The Firefly method was utilized and altered in order to address the problem of planted motif search. There are too many attractions since the whole attraction model makes all the fireflies in the FA flock to the brightest firefly. Even if many repetitions are required, the output will be imperfect, and the convergence rate will be sluggish. As a consequence, the freezing approach should be used with the FA. The Local freeze retains the best positions, whereas the Global freeze, which combines all of the local freeze outcomes, modifies the random directions. That would suggest the accuracy of the MFF functions.

- Regardless of gender, each firefly will be drawn by the other fireflies.
- The brighter the firefly, the more appealing it is. A less brilliant firefly will gravitate toward a brighter one in this case.
- The goal function is connected with the brightness of the firefly.

The FA is mostly focused on the attraction and movement of fireflies.

a. Attractiveness Function

The generalized version of the monotonically reducing attractiveness function (μ) is as follows:

$$\mu = \alpha_0 * c^{\gamma r^2} \quad (2)$$

$$r_{ji} = \left|\left|\chi_j - \chi_i\right|\right| = \sqrt{\Sigma_{\mathrm{K}}^{\mathfrak{D}}(\chi_{j,\mathrm{K}} - \chi_{i,\mathrm{K}})^2} \quad (3)$$

The Euclidean distance (δ) among the two fireflies, j and i, is calculated and represented as r_{ji}, wherein $\mathfrak{D}$ is the size of an optimization issue. There exists a light coefficient of absorption and a firefly brightness of α_0. The attraction value and the light absorbance coefficient control how quickly a firefly converges.

b. Movement of files

The movement of the less bright firefly j towards the brighter firefly i is shown by:

$$\chi_j = \chi_j \alpha_0 c^{\gamma r^2}\left(\chi_i - \chi_j\right) + \beta\ (r\ and - 0.5) \quad (4)$$

β is the random parameter that determines the random movement of the firefly. R is an operating system that produces random numbers that are usually between 0 and 1. Every firefly in FA is attracted to another and approaches to establish a dependence. The population is made up of one motif segment from each PMS sequence. Based on their Hamming Distance (HD), one population (firefly) is drawn to another and travels closer for reliance.The number of sites ϕ such that $s_1[\phi] \neq s_2[\phi]$gives the Hamming Distance (HD (s_1, s_2)) between two identical strings, s_1 and s_2 . The New Potential Motif location $N\vartheta\ jk$ is produced as a result of the higher HD value moving towards the lower one, as shown by Eq. (9). To be more precise, the distance away from the hamming determines how each population location moves toward the other population.

$$\delta = \sqrt{\Sigma_{j=1}^{\varepsilon}\Sigma_{i=1}^{\varepsilon}\Sigma_{k=1}^{t}(\rho_{rn}[j][k] - \rho_{rn}[i][k])^2} \quad (5)$$

$$\mu = \alpha_0 * c^{(-\gamma * \delta * \delta)^2} * (\rho_{rn}[j][k] - \rho_{rn}[i][k]) \quad (6)$$

$$\theta = \beta\ (r\ and - 0.5) \quad (7)$$

$$\Sigma_{i=1}^{\varepsilon}\Sigma_{k=1}^{t}(\rho_{rn}[j][k] = \rho_{rn}[j][k] + \mu + \theta) \quad (8)$$

$$\vartheta_{jk} = \begin{cases} N\vartheta\ jk, \\ \vartheta\ jk, \end{cases} \quad (9)$$

For theδ, use eq. (5).po refers to the population. Two factors, μ and θ randomness (eqs. 6 and 7) respectively, fix the new direction of po_2. First, use equation (5) to get the δ between ρ_{rn} of po_1 and po_2 . Then, use δ to compute the μ and determine the θ. Now, update the existing ρ_{rn} for these two values (μ, θ). The New Potential Motif position ($N\vartheta\ jk$) is the name given to the newly created position (eq. 9). Every sequence location in po_2 results in the creation of this new position.Initially, all ρ_{rn} of po_2 are kept and there is no movement if po_2has a lower HD (lesser bright) than po_1. In summary, $N\vartheta\ jk$ is created when all of the locations of the $j^{th}po$ migrate towards the $j^{th}po$ due to a greater HD of the $j^{th}po$ than the $j^{th}po$. Other than that, nothing has changed, and every place in the $i^{th}po$ is in its original position. This is seen in Eq. (9). The values and parameters utilized in this method are listed in Table 1.

Table 1
Parameters for MFFA

Parameters	Optimal values
γ	0.5
α_0	1.0
β	0.2
ε	30
$\mathbb{I}_\tau$	1
$\mathbb{U}_\tau$	1000
$Best_{accuracy}$	0

❖ Local freeze

The default Firefly Algorithm produced just 15% of matches, took longer, and had no discernible outcomes. The present work uses the FFF algorithm to overcome these limitations, and the suggested approach incorporates two freezing techniques, namely “Local Freeze (LF) and Global Freeze (GF)". LF was able to get an accuracy of up to 40% in a minimal time. Each sequence generates random sites where the altered motifs are inserted; these spots are referred to as

implanted positions. The issue with Firefly PMS is that, although it is less brilliant, it would happily fit in some of the implanted spots. There could be less of a hammering distance between them, however, therefore, as such a shift takes place, certain advantageous spots in fewer fireflies can be lost. We are maintaining that advantageous position because of this, and we term it FREEZING.

$$N\vartheta\, jk = \begin{cases} \vartheta\, jk, if\ (\mathfrak{D}_j = \mathfrak{D}_{\vartheta\, jk}) \\ \vartheta\, jk \pm \sigma, \sigma > 0, otherwise \end{cases} (10)$$

❖ Global Freeze

With just 40% success in LF, the focus is shifting to GF to improve precision. LF is executed and all phases are collected.Until every location is locked or as precise as possible, the LF is replicated 'n' times. Because the firefly's attraction and unpredictable nature fluctuate, the freezing sites at each moment are decided by the random places established in the beginning stage. Consequently, if motif searching is conducted in a certain way and the site is deemed unacceptable, it may provide some poor results. Despite many repetitions, the software fails to provide the desired results. By freezing in all random directions or achieving the highest level of motif detection accuracy, this GF is utilized to enhance performance. To ensure that the implanted location is ideal, many LF tests are run, and further refinement is done. The best solution is obtained by combining "n" times module-wise freezing (η) to create the GF. The operation ends if "freeze count (ψ)" equals the number of sequences; else, the module-wise freeze count, which is shown in Eq. 11, is increased with the freeze count.

$$\psi = \begin{cases} \psi + \eta, if\,(n! = \psi) \\ \psi, otherwise \end{cases} \quad (11)$$

Until all locations match the implanted positions, this procedure is repeated again. The loop ends and our goal is accomplished if the ψ equals t. All of the Gibbs sampling is done continually using the aforementioned procedure. In the end, we determined the optimal Gibbs sample while maintaining all levels of precision, and we found the final motif (TF) and the places that correlate to it (TFBs).

Input: $S = \{s_1, s_2 \ldots . s_t\}$ with length n, length "l", mutation "d"
BEGIN
While (true)
for$\phi = 1$ to count
Create a random position for each S
$$for\ iteration = \mathbb{I}_\tau\ to\ \mathbb{U}_\tau$$
For $j = 0\ to\ \varepsilon$
For $i = 0\ to\ \varepsilon$
$$\delta = 0$$
For $k = 0\ to\ t$
If $(\rho_{\mathcal{H}}[j] > \rho_{\mathcal{H}}[i]$

End if
$$\delta = math.sqrt(\delta)$$
End for k
For k=0 to t
$$\mu = \alpha_0 * c^{(-\gamma*\delta*\delta)^2} * (\rho_{rn}[j][k] - \rho_{rn}[i][k])$$
$$\theta = \beta\ (r\ and - 0.5)$$
Temp 1= $(\rho_{rn}[j][k] + \mu + \theta$
Freeze= Local freeze (j, k, l)
If (freeze==0)
$$\rho_{rn}[j][k] = (int)temp\ 1;$$
Else
Don't alter that position. Freeze it
end if
end for k
Finding a HD for the new locations.
end for i
end for j
end for iteration
callGF
$$ð_j = 0$$
For k=0 to t
if $(\aleph[k]! = -1)$
$$ð_j \longleftarrow ð_j + 1$$
end if
end for k
accuracy=$ð_j * t/100$
if $(Best_{accuracy} < accuracy)$
$$Best_{accuracy} = accuracy$$
$$Best_{motif} = Gibbs\ sampling\ [\phi]$$
For $h = 0\ to\ t$
$$Best_{position}[\phi][h] = \aleph[h]$$
end for h
end if
end for ϕ
End while
END

C. AI-based refinement enhancement for predicting model construction

For AI-based refinement enhancement using the Harmony Search with Logistic regression. The motifs identified using the MFF-GB are likely to represent crucial patterns in the sequences. These motifs can be transformed into numerical features, which may include properties such as motif length, mutation count, position weight matrix (PWM) scores, and other relevant characteristics.

a. Harmony Search Algorithm (HSA)

The Harmony Search algorithm is employed to further refine the features extracted from the identified motifs. HS optimizes these features by iteratively adjusting their values to enhance their contribution to the overall model. A relatively novel meta-heuristic algorithm, Harmony Search (HS) finds a pleasant harmony by emulating the improvisation process of musicians. The approach has several benefits in comparison to conventional optimization methods: (i) it is a straightforward meta-heuristic algorithm that eliminates the need for initial setting of decision variables; (ii) it employs stochastic random searches, which eliminates the need for derivative information; and (iii) it possesses a limited number of parameters that allow for fine-tuning. This is readily apparent in the literature as discrete and continuous optimization problem applications. HS is predicated on the notion that improvisation is similar to the optimization method that engineers use to solve problems, with

artists experimenting with different combinations of known (memorized) frequencies. Thus, to find the global optimum, a feasible solution is referred to as a "harmony," and each option variable is associated with a note that produces a value.

The following stages make up the HS algorithm:

i. Initiate the problem and algorithm

The definition of the optimization issue is:

$Min\ (or\ max)\ f(\mathfrak{X})$ (12)

$\vartheta\beta_i \leq \mathfrak{X}_i \leq \cup\ \beta_i$ (13)

An objective function is denoted by $f(\mathfrak{X})$,, where $\mathfrak{X}$ is a potential solution composed of the choice variables $\mathfrak{X}_i$.For every decision variable, the lower and higher limits are represented by $\vartheta\beta_i$ and $\cup\ \beta_i$, respectively.

ii. Initialize the Harmony Memory (HM)

HM represents a memory region solution vector, akin to the genetic pool in a Genetic Algorithm (GA). The initial HS is formed through a uniform distribution of ranges, populating a matrix with an equivalent number of solution vectors generated randomly, mirroring the size of HMS..

$\mathfrak{X}_i^l = \vartheta\beta_i + G\ and\ \mathfrak{X}\ (\vartheta\beta_i - \cup\ \beta_i)$(14)

iii. Create a new Harmony

Improvisation is the process of creating a new harmonic. Three principles are used to create this new harmony vector, $\mathfrak{X}' = ,\mathfrak{X}'_1, \mathfrak{X}'_2, \ldots \mathfrak{X}'_N$. Random selection, pitch modification, and memory consideration.

iv. Update the HM

The new harmony is integrated while the current worst harmony is eliminated if its objective function value is greater than that of the most unfavourable harmony vector previously documented in the HM and no harmony vector of the same kind is present in the HM.

v. Check the Stopping criterion

When the allotted number of improvisations is achieved, the algorithm terminates. If not, repeat steps three and four.

b. Logistic Regression

With the refined features obtained from the HS algorithm, a logistic regression model is constructed. Logistic regression is chosen for its suitability in binary classification problems, which is common in motif prediction tasks.A tailored regression model known as logistic regression describes and illustrates the relationship between a linear mixture of explanatory variables and a categorical response variable. It may include categorical and/or continuous variables. Multinomial, ordinal, and binomial logistic regressions are the three different types of logistic regressions. When there are just two potential values for the dependent variable (usually "0" and "1"), binomial logistic regression models are applied. Multinomial logistic regression models, on the other hand, are applied to explanatory variables with three or more possible outcomes (three or more). Ordinal logistic regression is specifically designed for situations in which the dependent variable consists of ordered outcomes. This study employed logistic regression to analyze the dependent variable, which was categorized as "1" for greater stability and "0" for less stability.

The binary logistic regression classifier is a commonly used technique in regression modeling that is utilized for modeling dichotomous dependent variables, such as the degree of stability of the index. Groups are typically encoded as (zero) "0" and (one) "1" to facilitate the interpretation of the results. Thus, items are classified in a binary manner using basic logistic regression.The following provides the logistic regression specification:

$\pi(\mathfrak{X}) = P(\mathbb{Y} = 1) = \frac{1}{1+\exp\{-(\wp_o + \sum_{j=1}^{3} \wp_j X_j)\}} = [1 +$

$\exp\{-\{-(\wp_o + \sum_{j=1}^{3} \wp_j X_j)\}]^{-1}$(15)

$P(\mathbb{Y} = 0) = 1 - P(\mathbb{Y} = 1)$ (16)

$P(\mathbb{Y} = 0) = \frac{\exp\{-(\wp_o + \sum_{j=1}^{3} \wp_j X_j)\}}{1+\exp\{-(\wp_o + \sum_{j=1}^{3} \wp_j X_j)\}}$(17)

Where

$y_i = \begin{cases} 1\ for\ more\ stable \\ 0\ for\ less\ stable \end{cases}$ (18)

The coefficients of the 3 independent variables, or the maximum likelihood estimable parameters of the linear model, are represented by the symbol $\wp_1, \wp_2 \ldots \wp_3$, which stands for the coefficient of the constant term. The variables that are independent are represented by the notation $\mathfrak{X}_1, \mathfrak{X}_2 \ldots \mathfrak{X}_3$.

For logit regression of this probability, the logistic model has a linear shape:

$Logit[\pi(\mathfrak{X})] = \log\left(\frac{\pi(\mathfrak{X})}{1-\pi(\mathfrak{X})}\right) = \pi_o + \sum_{j=1}^{3} \wp_j X_j$ (19)

Where the odds $= \frac{\pi(\mathfrak{X})}{1-\pi(\mathfrak{X})}$, where 1- $\pi(\mathfrak{X})$ represents the likelihood of failure and $\pi(\mathfrak{X})$ represents the probability of success.

c. Harmony Search with Logistic Regression

The Harmony Search algorithm is employed to optimize and refine the features extracted from the motifs. HS iteratively adjusts the values of these features to enhance their contribution to the overall predictive model. HS emulates the process of musical improvisation, treating each candidate solution (harmony) as a set of decision variables. The Harmony Memory stores promising solutions, and the algorithm improvises new solutions based on memory consideration, pitch adjustment, and random selection.A logistic regression model is built using the enhanced characteristics that come from HS. Because it works well for binary classification tasks, logistic regression may be used to anticipate outcomes like stable or less stable circumstances.Logistic regression learns a decision boundary that separates the two classes (more stable and less stable) in the feature space.Post-training, the Harmony Search algorithm is again employed to fine-tune the parameters of the logistic regression model. HS optimizes these parameters to enhance the predictive performance of the logistic regression model.

Given a set of new input features (potentially representing motifs in unseen sequences), the hybrid model utilizes the trained and refined logistic regression model for prediction.The logistic regression equation is used to calculate the log odds (logit) of the positive outcome (more stable condition). This logit is then transformed into a probability.These models are then used to scan other sequences to predict the presence of similar motifs.

IV. EXPERIMENTAL ANALYSIS

This section demonstrates the experimental examination of the motif discovery and prediction that has been suggested for performance assessment. The outcomes demonstrate the great efficiency of the proposed. This part is divided into 2 subsections, including a comparison, and a summary of the research.

A. *Comparative analysis*

Finally, the performance metrics are rigorously evaluated, encompassing Running time vs. Maximum Motif length, running time vs. Motif length, running time vs. motif count (d), Prediction rate, and Mean Square Error using the existing methods such as Straglr [31], STREME [32], Counting Motif Algorithm [33], G- Protein Coupled Receptors (GPCRs) [34], Fully Convolutional Network with Global Average Pooling (FCNA), Convolutional Autoencoder And Convolutional Neural Network (CAE-CNN) [25], self-attention graph network- drug–target affinity (SAG-DTA) [35].

a. Running time vs. Maximum Motif length

The running time increases with the maximum motif length, as shown in Fig. 2 and Table 2. In Straglr, at a motif length of 30, the running time is highest among the existing algorithms. The running time also increases with the motif length, but it seems to increase at a slower rate compared to Straglr. The proposed algorithm outperforms Streme across all motif lengths. The running time for the proposed algorithm is consistently lower than both Straglr and Streme. It demonstrates better efficiency, especially as the motif length increases.

TABLE 2

Numerical Outcomes of Maximum Motif Length

Maximum Motif length	Running time (seconds/motifs)		
	Straglr	STREME	Proposed
10	100	0	0
15	150	100	80
20	200	110	90
25	350	150	100
30	600	190	130

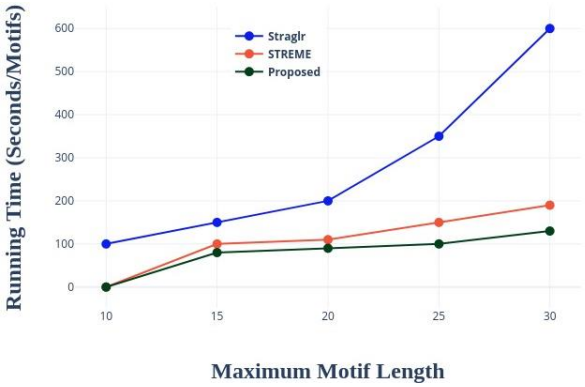

Fig. 2: Running time vs. Maximum Motif length

b. Running time vs. Motif length

Counting motif algorithm, at a motif length of 2, the running time is 23 seconds. As the motif length increases, the running time also increases, reaching 58 seconds at a motif length of 10. GPCRs, at a motif length of 2, the running time is 21 seconds. The running time increases with motif length, reaching 56 seconds at a motif length of 10. Finally, in the proposed, motif length of 2, the running time is 19 seconds. The proposed algorithm consistently outperforms both existing methods, reaching 38 seconds at a motif length of 10. The proposed algorithm consistently outperforms both existing algorithms, Counting Motif and GPCRs, in terms of running time for motif processing. As the motif length increases, the efficiency gains of the proposed algorithm become more pronounced. The running time of the proposed algorithm is consistently lower than the existing methods, making it a more efficient choice, especially for motifs with longer lengths represented in Fig. 3 and Table 3.

TABLE 3

Numerical Outcomes of Motif length (l)

Motif length (l)	Running time (s)		
	Counting Motif Algorithms	GPCRs	Proposed
10	23	21	19
15	45	41	31
20	49	40	34
25	56	50	36
30	58	56	38

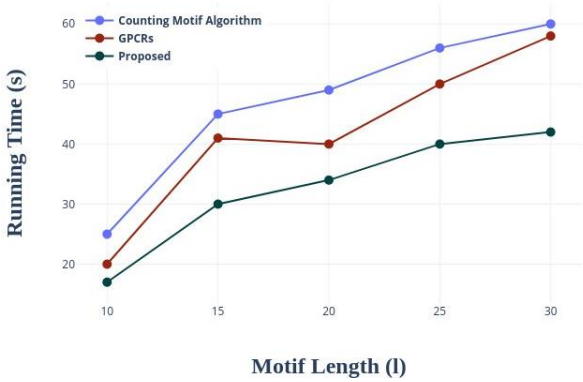

Fig. 3: Running time vs. Motif length (l)

c. Running time vs mutation count (d)

The running time increases as the mutation count (d) increases as shown in Fig. 4 and Table 4. At a mutation count of 2, the running time is 23 seconds. At a mutation count of 4, the running time increases to 45 seconds. This upward trend continues, reaching 59 seconds at a mutation count of 10. At a mutation count of 2, the running time is 21 seconds. At a mutation count of 4, the running time increases to 41 seconds. The running time further increases, reaching 56 seconds at a mutation count of 10. At a mutation count of 2, the running time is 19 seconds. At a mutation count of 4, the running time is 31 seconds. The running time increases gradually, reaching 40 seconds at a mutation count of 10. The proposed algorithm consistently outperforms both existing algorithms, Counting Motif and GPCRs, in terms of running time for motif processing. As the mutation count increases, the efficiency

gains of the proposed algorithm become more pronounced. The running time of the proposed algorithm is consistently lower than the existing methods, making it a more efficient choice, especially for motifs with a higher number of mutations.

TABLE 4

Numerical Outcomes of Mutation Count (d)

Mutation count (d)	Running time (s)		
	Counting Motif Algorithms	GPCRs	Proposed
10	23	21	19
15	45	41	31
20	49	40	34
25	56	50	38
30	59	56	40

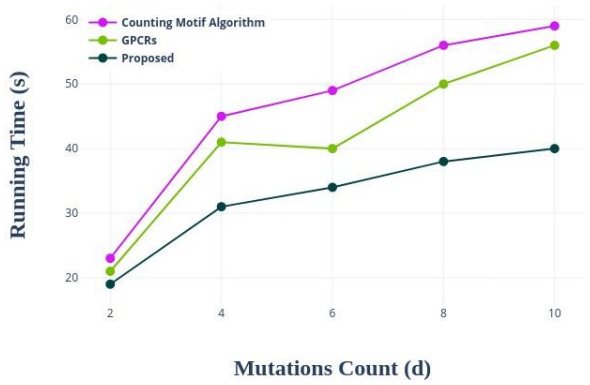

Fig. 4: Running time vs motif count (d)

d. Prediction rate

In FCNA, at 10 epochs, the prediction rate is 12%. As the number of epochs increases, the prediction rate improves, reaching 85% at 50 epochs. In CAE-CNN, at 10 epochs, the prediction rate is 19%. The prediction rate increases with the number of epochs, reaching 89% at 50 epochs. In the suggested algorithm at 10 epochs, the prediction rate is 21%.

The suggested algorithm consistently outperforms both existing methods, reaching a prediction rate of 97% at 50 epochs. The proposed algorithm consistently outperforms both existing algorithms, FCNA and CAE-CNN, in terms of prediction rate after training for a specific number of epochs.

As the number of epochs increases, the suggested algorithm demonstrates superior learning and achieves a higher prediction rate compared to the existing methods. Fig. 5 and Table 5 suggest that the suggested algorithm is more effective in capturing patterns and making accurate predictions as it undergoes more training epochs.

TABLE 5

Numerical Outcomes of Prediction Rate

Number of epochs	Prediction rate (%)		
	FCNA	CAE-CNN	Proposed
10	12	19	21
20	25	47	49
30	45	51	59
40	65	69	79
50	85	89	97

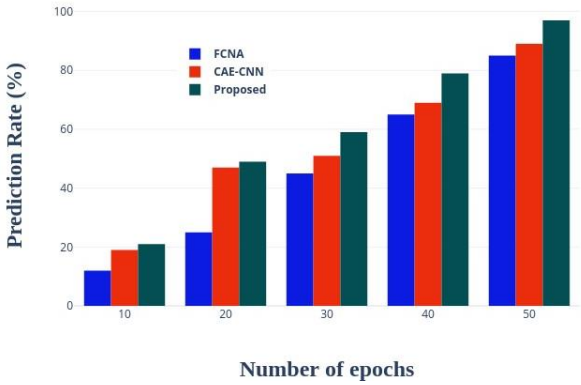

Fig. 5: Prediction rate

e. Mean Square Error

In CNN, at 10 epochs, the MSE is 98%. As the number of epochs increases, the MSE decreases, reaching 50% at 50 epochs. SAG- DTA at 10 epochs, the MSE is 95%. The MSE decreases with the number of epochs, reaching 42% at 50 epochs. At 10 epochs, the MSE is 89%. The suggested methods consistently outperform both existing methods, achieving a lower MSE of 25% at 50 epochs. The suggested methods consistently outperforms both existing algorithms, CNN and SAG-DTA, in terms of mean square error after training for a specific number of epochs. As the number of epochs increases, the proposed algorithm demonstrates superior learning and achieves a lower mean square error compared to the existing methods. Fig. 6 and Table 6 suggest that the suggested algorithm is more effective in minimizing errors and improving accuracy as it undergoes more training epochs.

TABLE 6

Numerical Outcomes of Mean Square Error

Number of epochs	MSE (%)		
	CNN	SAG-DTA	Proposed
10	12	19	21
20	25	47	49
30	45	51	59
40	65	69	79
50	85	89	97

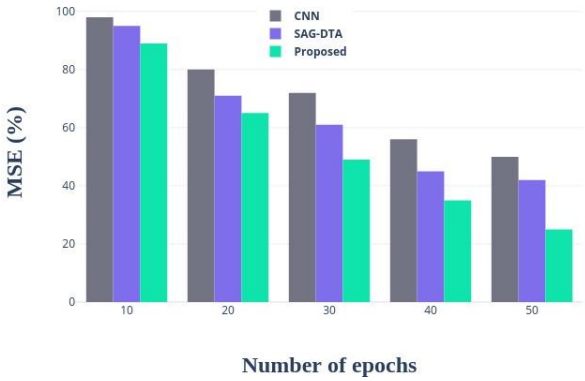

Fig. 6: Mean Square Error

B. Research summary

The research combines motif discovery and prediction through a hybrid approach. It starts with data collection from Jaspar and ENCODE databases, ensuring normalization and integrity. Motif identification utilizes a Modified Freeze Firefly Algorithm with Gibbs sampling, enhancing accuracy through freezing mechanisms. The Harmony Search algorithm refines motifs, optimizing parameters, and fine-tuning a logistic regression model. The final hybrid model integrates motif features, demonstrating effectiveness in predicting stable and less stable conditions in biological sequences.

V. CONCLUSION

Finally, the study's hybrid technique, which combines motif identification and prediction, demonstrates a complete framework for studying biological sequences. The Modified Freeze Firefly Algorithm is used with Gibbs sampling for motif discovery, followed by refining using Harmony Search and logistic regression, to provide a strong prediction model. The effective implementation of these tools demonstrates their utility in comprehending complicated biological patterns. The study not only enhances motif identification but also underlines the need of combining varied techniques to increase prediction accuracy in biological stability. Overall, the hybrid model provides a viable path for researchers to investigate complex interactions in genomic data.

Reference

1. He, Y., Shen, Z., Zhang, Q., Wang, S., & Huang, D. S. (2021). A survey on deep learning in DNA/RNA motif mining. *Briefings in Bioinformatics*, *22*(4), bbaa229.
2. Zhang, Q., Wang, S., Chen, Z., He, Y., Liu, Q., & Huang, D. S. (2021). Locating transcription factor binding sites by a fully convolutional neural network. *Briefings in Bioinformatics*, *22*(5), bbaa435.
3. Machlab, D., Burger, L., Soneson, C., Rijli, F. M., Schübeler, D., & Stadler, M. B. (2022). monaLisa: an R/Bioconductor package for identifying regulatory motifs. *Bioinformatics*, *38*(9), 2624-2625.
4. Park, M., Singh, S., Khan, S. R., Abrar, M. A., Grisanti, F., Rahman, M. S., & Samee, M. A. H. (2022). Multinomial convolutions for joint modeling of regulatory motifs and sequence activity readouts. *Genes*, *13*(9), 1614.
5. de Martin, X., Sodaei, R., &Santpere, G. (2021). Mechanisms of binding specificity among bHLH transcription factors. *International Journal of Molecular Sciences*, *22*(17), 9150.
6. Vogelgsang, L., Nisar, A., Scharf, S. A., Rommerskirchen, A., Belick, D., Dilthey, A., & Henrich, B. (2023). Characterisation of Type II DNA Methyltransferases of Metamycoplasmahominis. *Microorganisms*, *11*(6), 1591.
7. Peterson, J. M., O'Leary, C. A., Coppenbarger, E. C., Tompkins, V. S., & Moss, W. N. (2023). Discovery of RNA secondary structural motifs using sequence-ordered thermodynamic stability and comparative sequence analysis. *MethodsX*, *11*, 102275.
8. Zang, X., Zhao, X., & Tang, B. (2023). Hierarchical molecular graph self-supervised learning for property prediction. Communications Chemistry, 6(1), 34.
9. Jia, Z., Lin, Y., Wang, J., Ning, X., He, Y., Zhou, R., ... & Li-wei, H. L. (2021). Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 29, 1977-1986.
10. Han, K., Shen, L. C., Zhu, Y. H., Xu, J., Song, J., & Yu, D. J. (2022). MAResNet: predicting transcription factor binding sites by combining multi-scale bottom-up and top-down attention and residual network. Briefings in Bioinformatics, 23(1), bbab445.
11. Cao, F., Zhang, Y., Cai, Y., Animesh, S., Zhang, Y., Akincilar, S. C., ... & Fullwood, M. J. (2021). Chromatin interaction neural network (ChINN): a machine learning-based method for predicting chromatin interactions from DNA sequences. *Genome biology*, *22*, 1-25.
12. Anker, A. S., Kjær, E. T., Juelsholt, M., Christiansen, T. L., Skjærvø, S. L., Jørgensen, M. R. V., ... & Jensen, K. M. (2022). Extracting structural motifs from pair distribution function data of nanostructures using explainable machine learning. *npj Computational Materials*, *8*(1), 213.
13. Savojardo, C., Martelli, P. L., &Casadio, R. (2023). Finding functional motifs in protein sequences with deep learning and natural language models. *Current Opinion in Structural Biology*, *81*, 102641.
14. Bashir, A., Yang, Q., Wang, J., Hoyer, S., Chou, W., McLean, C., ... & Ferguson, B. S. (2021). Machine learning guided aptamer refinement and discovery. *Nature Communications*, *12*(1), 2366.
15. Wu, X., Guo, Y., Xue, J., Dong, Y., Sun, Y., Wang, B., ... & Liu, Y. (2023). Abnormal and Changing Information Interaction in Adults with Attention-Deficit/Hyperactivity Disorder Based on Network Motifs. *Brain Sciences*, *13*(9), 1331.
16. Xu, W., Reuter, K., & Andersen, M. (2022).

Predicting binding motifs of complex adsorbates using machine learning with a physics-inspired graph representation. *Nature Computational Science*, *2*(7), 443-450.
17. Zhang, P., Wang, H., Xu, H., Wei, L., Liu, L., Hu, Z., & Wang, X. (2023). Deep flanking sequence engineering for efficient promoter design using DeepSEED. *Nature Communications*, *14*(1), 6309.
18. Talukder, A., Barham, C., Li, X., & Hu, H. (2021). Interpretation of deep learning in genomics and epigenomics. *Briefings in Bioinformatics*, *22*(3), bbaa177.
19. Ge, W., Meier, M., Roth, C., &Söding, J. (2021). Bayesian Markov models improve the prediction of binding motifs beyond first order. *NAR Genomics and Bioinformatics*, *3*(2), lqab026.
20. Wang, Y., & Chen, L. (2022). DeepPerVar: a multi-modal deep learning framework for functional interpretation of genetic variants in personal genome. *Bioinformatics*, *38*(24), 5340-5351.
21. Novakovsky, G., Fornes, O., Saraswat, M., Mostafavi, S., & Wasserman, W. W. (2023). ExplaiNN: interpretable and transparent neural networks for genomics. *Genome Biology*, *24*(1), 1-24.
22. Sutanto, K., & Turcotte, M. (2021). Assessing global-local secondary structure fingerprints to classify RNA sequences with deep learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.
23. Nalewaj, M., & Szabat, M. (2022). Examples of Structural Motifs in Viral Genomes and Approaches for RNA Structure Characterization. *International Journal of Molecular Sciences*, *23*(24), 15917.
24. Zhang, Q., Wang, S., Chen, Z., He, Y., Liu, Q., & Huang, D. S. (2021). Locating transcription factor binding sites by fully convolutional neural network. *Briefings in bioinformatics*, *22*(5), bbaa435.
25. Zhang, Y., Qiao, S., Zeng, Y., Gao, D., Han, N., & Zhou, J. (2021). CAE-CNN: Predicting transcription factor binding site with convolutional autoencoder and convolutional neural network. *Expert Systems with Applications*, *183*, 115404.
26. D'Souza, S., Prema, K. V., Balaji, S., & Shah, R. (2023). Deep Learning-Based Modeling of Drug–Target Interaction Prediction Incorporating Binding Site Information of Proteins. *Interdisciplinary Sciences: Computational Life Sciences*, *15*(2), 306-315.
27. Maseko, N. N., Steenkamp, E. T., Wingfield, B. D., & Wilken, P. M. (2023). An in Silico Approach to Identifying TF Binding Sites: Analysis of the Regulatory Regions of BUSCO Genes from Fungal Species in the Ceratocystidaceae Family. *Genes*, *14*(4), 848.
28. Zhai, X., &Tuerxun, A. (2022). DNA Sequence Specificity Prediction Algorithm Based on Artificial Intelligence. *Mathematical Problems in Engineering*, *2022*.
29. Xu, H., Jia, P., & Zhao, Z. (2021). DeepVISP: deep learning for virus site integration prediction and motif discovery. *Advanced Science*, *8*(9), 2004958.
30. Yusuf, S. M., Zhang, F., Zeng, M., & Li, M. (2021). DeepPPF: A deep learning framework for predicting protein family. *Neurocomputing*, *428*, 19-29.
31. Chiu, R., Rajan-Babu, I. S., Friedman, J. M., & Birol, I. (2021). Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome Biology*, *22*(1), 224.
32. Bailey, T. L. (2021). STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, *37*(18), 2834-2840.
33. Ren, Y., Sarkar, A., Veltri, P., Ay, A., Dobra, A., & Kahveci, T. (2021). Pattern discovery in multilayer networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(2), 741-752.
34. Bongers, B. J., Gorostiola González, M., Wang, X., van Vlijmen, H. W. T., Jespers, W., Gutiérrez-de-Terán, H., ... & van Westen, G. J. P. (2022). Pan-cancer functional analysis of somatic mutations in G protein-coupled receptors. *Scientific Reports*, *12*(1), 21534.
35. Zhang, S., Jiang, M., Wang, S., Wang, X., Wei, Z., & Li, Z. (2021). SAG-DTA: prediction of drug–target affinity using self-attention graph network. *International Journal of Molecular Sciences*, *22*(16), 8993