Briefings in Bioinformatics, 2026, 27, bbag044
https://doi.org/10.1093/bib/bbag044
Published: 9 February 2026
Problem Solving Protocol

OXFORD

Exploring potential transcription factors and their regulatory relationships based on asymmetric covariance natural vector encoding method and machine learning algorithms

Guoqing Hu[1,‡], Mengmeng Sang[2,‡], Hao Wang[3,4,‡], Jia Ge[2], Lin Xu[2], Stephen S.-T. Yau[4,5,*]

[1]Hetao Institute of Mathematics and Interdisciplinary Sciences (HIMIS), Shenzhen 518000, Guangdong, P.R. China
[2]Department of Immunology, School of Medicine, Nantong University, Nantong 226001, Jiangsu, P.R. China
[3]Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, P.R. China
[4]Beijing Institute of Mathematical Sciences and Applications (BIMSA), Beijing 101408, P.R. China
[5]Department of Mathematical Sciences, Tsinghua University, Beijing 100084, P.R. China

*Corresponding author. Beijing Institute of Mathematical Sciences and Applications (BIMSA), Beijing 101408, P.R. China. E-mail: yau@uic.edu
‡Guoqing Hu, Mengmeng Sang, and Hao Wang contributed equally to this work.

Abstract

Transcription factors (TFs) orchestrate cellular programs by activating or repressing gene expression in response to diverse stimuli. Although advances in experimental and computational biology have expanded our understanding of TFs, existing prediction methods still struggle to accurately capture TF–target regulatory relationships and determine their directionality (activation versus inhibition). Here, we propose ACNVE-K, an integrative framework combining k-mer decomposition with asymmetric covariance natural vector encoding to convert amino acid sequences into multidimensional feature vectors. Using Leveraging eXtreme Gradient Boosting (XGBoost), Gradient Boosting (GB), and Random Forest (RF) algorithms, we constructed five predictive models for TF identification, target gene inference, and regulatory direction classification. Benchmarking analyses demonstrated that XGBoost achieved the highest predictive performance across human and mouse genomes, particularly with updated genome annotations. The 5-mer configuration provided an optimal balance between feature richness and computational efficiency. Collectively, ACNVE-K offers a robust and interpretable framework for decoding transcriptional regulation, facilitating advances in precision medicine, regulatory genomics, and machine-learning–based gene network reconstruction.

Keywords *k*-mer encoding, asymmetric covariance natural vector, machine learning, transcription factor–gene prediction, regulatory directionality

Introduction

Transcription factors (TFs) are DNA-binding proteins that act as master regulators of gene expression by orchestrating transcriptional initiation through sequence-specific recognition of regulatory DNA elements [1, 2]. These molecular switches operate via direct DNA interaction through specialized DNA-binding domains that target promoter or enhancer regions [3], followed by dynamic protein–protein interactions that modulate transcriptional activation or repression [4]. Their regulatory influence extends beyond DNA recognition, as TFs recruit chromatin-modifying complexes—including histone acetyltransferases and methyltransferases—that remodel nucleosome architecture and control chromatin accessibility [5]. The functional versatility of TFs arises from combinatorial regulation: they form context-dependent complexes with other TFs, coactivators, and corepressors to achieve precise spatiotemporal control of gene expression [6, 7]. This process is tightly integrated with cellular signaling networks, where post-translational modifications such as phosphorylation translate extracellular cues (e.g. hormones and growth factors) into transcriptional responses [8, 9]. Furthermore, feedback and feedforward loops involving TFs and their target genes create dynamic regulatory circuits that maintain cellular homeostasis while enabling adaptive responses to environmental or developmental changes [10]. Collectively, these multi-layered mechanisms establish the transcriptional and epigenetic landscape that governs cell identity, differentiation, and phenotypic plasticity.

Approaches to identify transcription factors have combined experimental and computational strategies, each with inherent strengths and limitations. Experimental techniques such as chromatin

Received: October 14, 2025. **Revised:** December 15, 2025. **Accepted:** January 7, 2026

immunoprecipitation (ChiP) and electrophoretic mobility shift assays provide direct evidence of TF–DNA binding but suffer from limitations in antibody specificity, experimental scalability, and the inability to capture transient or cell type–specific interactions [11]. In contrast, computational methods leverage DNA sequence features to predict TF binding sites using position weight matrices or motif models [12]. Databases such as JASPAR [13], TRANSFAC [14], ENCODE [15], Cistrome DB [14], and TRRUST [16] have significantly advanced our understanding of TF–target relationships, yet they remain constrained to known motifs and experimentally validated interactions, limiting discovery of novel or species-specific TFs.

Recent advances in machine learning (ML) and deep learning (DL) have revolutionized TF prediction. DeepSEA integrates genomic context to predict TF binding and chromatin effects [17]; DeepBind applies convolutional neural networks (CNNs) to capture local sequence features [18]; TF-MoDISco interprets patterns learned by DL models to infer TF binding logic [19]; and DeepTFactor combines CNN-based protein sequence analysis with explainable AI for TF discovery [20]. Despite these advancements, current models primarily extrapolate from existing data, lacking the capacity to infer previously unknown TF–target regulatory relationships or the directionality (activation versus repression) of such interactions.

To overcome these limitations, we developed ACNVE-K (Asymmetric Covariance Natural Vector Encoding with *k*-mer integration), an innovative computational framework that expands upon the natural vector method proposed by Deng et al. [21] and its symmetric covariance extension by Sun et al. [22]. ACNVE-K integrates *k*-mer decomposition with asymmetric covariance encoding to transform nucleotide and protein sequences into multidimensional numerical representations, capturing both compositional and directional characteristics of genomic interactions. Using advanced ML algorithms—including GB, RF, and XGBoost—we established predictive models for TF identification, target gene inference, and regulatory direction classification across human and mouse genomes. This framework enables researchers to predict whether genes of interest possess TF or target gene potential and to uncover possible regulatory relationships between them, including activation and inhibition patterns. Collectively, our approach provides a scalable, interpretable, and high-performance platform for deciphering transcriptional regulation and reconstructing gene regulatory networks in complex biological systems.

Materials and methods

Data source

Regulatory relationships between TFs and their target genes in *Homo sapiens* and *Mus musculus* were retrieved from the CollecTRI database [23]. The human dataset of CollecTRI database included 1186 TFs, 6692 target genes, and 43 178 regulatory interactions, comprising 37 360 activating and 5818 inhibitory relationships. The mouse dataset of CollecTRI database contained 1072 TFs, 6058 target genes, and 38 665 regulatory interactions, including 33 237 activating and 5428 inhibitory relationships. These well-curated datasets provided a reliable benchmark for model construction, optimization, and validation in this study. In addition, we used the human data from the TRRUST (v2) database (https://www.grnpedia.org/trrust/) [16] to independently validate our modeling framework; the human TRRUST (v2) dataset includes 795 TFs, 2492 target genes, and 9396 regulatory interactions, comprising 3149 activating and 1922 repressing relationships.

Asymmetric covariance natural vector encoding method

In this study, we applied the ACNVE method to model genomic sequences. Unlike traditional symmetric covariance approaches, ACNVE explicitly captures the directional dependencies between nucleotides, providing a more accurate representation of biological sequence order. For instance, the trinucleotides *AGT* and *TGA* encode distinct proteins despite comprising the same bases in different orders [24], a difference neglected by symmetric models.

Definition of basic statistics

Let the nucleotide sequence be denoted as $S = s_1 s_2 \cdots s_n$, where $s_i \in \mathcal{L} = \{A, C, G, T\}$. For each nucleotide $m \in \mathcal{L}$, we calculate the occurrence frequency n_m and its mean positional index:

$$\mu_m = \frac{1}{n_m} \sum_{j=1}^{n} j \cdot \mathbb{I}\left(s_j = m\right) \tag{1}$$

where $\mathbb{I}\left(s_j = m\right)$ is an indicator function that equals 1 if the *j-th* nucleotide is *m*, and 0 otherwise.

Construction of asymmetric covariance features

Higher-order directional covariances were introduced to encode fragment-level dependencies.

Second-order case (k = 2):

For a nucleotide pair ($m_1 m_2$), the directional indicator function is defined as:

$$\hat{\omega}_{m_1 m_2}\left(s_i s_{i+1}\right) = \begin{cases} 1, & if\ m_1 m_2 = s_i s_{i+1} \\ 0, & otherwise \end{cases} \tag{2}$$

The second-order asymmetric covariance feature is then given by:

$$\alpha - Cov\left(m_1 m_2\right) = \frac{1}{n_{m_1} n_{m_2}} \sum_{i=1}^{n-1} \left(\left[i - \mu_{m_1}\right]\left[i - \mu_{m_2}\right]\right) \cdot \hat{\omega}_{m_1 m_2}\left(s_i s_{i+1}\right) \tag{3}$$

Third-order case (k = 3):

For a nucleotide triplet ($m_1 m_2 m_3$):

$$\hat{\omega}_{m_1 m_2 m_3}\left(s_i s_{i+1} s_{i+2}\right) = \begin{cases} 1, & if\ m_1 m_2 m_3 = s_i s_{i+1} s_{i+2} \\ 0, & otherwise \end{cases} \tag{4}$$

The third-order asymmetric covariance feature is defined as:

$$\alpha - Cov\left(m_1 m_2 m_3\right) = \frac{1}{\prod_{j=1}^{3} n_{m_j}} \sum_{i=1}^{n-2} \left(\prod_{j=1}^{3} \left(i - \mu_{m_j}\right)\right) \cdot \hat{\omega}_{m_1 m_2 m_3}\left(s_i s_{i+1} s_{i+2}\right) \tag{5}$$

These second- and third-order covariance features form the foundational units of our method, capturing directional structural patterns from nucleotide sequences.

Building upon the third-order asymmetric covariance features, we generalize the formulation to model ordered nucleotide fragments of arbitrary order $k \geq 2$. For any nucleotide combination of length k,

Table 1 Indicator function values for 2-mers in sequence ACGGTAGTCA.

Sequence	A	C	G	G	T	A	G	T	C	A
Position	1	2	3	4	5	6	7	8	9	10
$I_{AC}(s_i, s_{i+1})$	1	1	0	0	0	0	0	0	0	0
$I_{AG}(s_i, s_{i+1})$	0	0	0	0	0	1	1	0	0	0
$I_{CA}(s_i, s_{i+1})$	0	0	0	0	0	0	0	0	1	1
$I_{CG}(s_i, s_{i+1})$	0	1	1	0	0	0	0	0	0	0
$I_{GG}(s_i, s_{i+1})$	0	0	1	1	0	0	0	0	0	0
$I_{GT}(s_i, s_{i+1})$	0	0	0	1	1	0	1	1	0	0
$I_{TA}(s_i, s_{i+1})$	0	0	0	0	1	1	0	0	0	0
$I_{TC}(s_i, s_{i+1})$	0	0	0	0	0	0	0	1	1	0

denoted as $m_1 m_2 \ldots m_k$, the directional indicator function is defined as:

$$\hat{\omega}_{m_1 \ldots m_k}\left(s_i \ldots s_{i+k-1}\right) = \begin{cases} 1, & if\ m_1 \ldots m_k = s_i \ldots s_{i+k-1} \\ 0, & otherwise \end{cases} \tag{6}$$

The corresponding k-order directional covariance feature is given by:

$$\alpha - Cov\left(m_1 \ldots m_k\right) = \frac{1}{\prod_{j=1}^{k} n_{m_j}} \sum_{i=1}^{n-k+1} \left(\prod_{j=1}^{k} \left(i - \mu_{m_j}\right) \right) \cdot \hat{\omega}_{m_1 \ldots m_k}\left(s_i \ldots s_{i+k-1}\right) \tag{7}$$

The resulting ACNVE feature vector integrates base frequencies, positional indices, and directional covariance features from multiple orders, providing a compact yet information-rich representation of nucleotide spatial structure.

Feature dimensionality of ACNVE

Let $B = |\mathcal{L}|$denote the alphabet size and k_{max} the maximum fragment order. The total feature dimensionality is :

$$Dim_{total} = 2B + \sum_{k=2}^{k_{max}} B^k \tag{8}$$

For DNA (B = 4) with k_{max}= 3:

$$Dim_{total} = 2 \times 4 + 4^2 + 4^3 = 88$$

This structured feature expansion enables flexible control of representational power. As k increases, dimensionality scales approximately with $4^k + 4^{k-1}$, capturing richer spatial and dependency patterns. For example, 3-mer, 4-mer, and 5-mer encodings yield feature vectors of 88, 344, and 1368 dimensions, respectively (Supplementary Tables S1–S3).

Illustrative example

For the sequence *ACGGTAGTCA (Table 1)*:

$$n_A = 3,\ \ n_C = 2,\ \ n_G = 3,\ \ n_T = 2,$$

$$\mu_A = 5.67,\ \ \mu_C = 5.5,\ \ \mu_G = 4.67,\ \ \mu_T = 6.5.$$

After calculating directional covariances via Equations (2–3), the resulting 24-dimensional feature vector: (3, 2, 3, 2, 5.67, 5.5, 4.67, 6.5, 0, 0.119, 1.381, 0, 0, 0.107, 0, 0.299, 1.272, 0.551, 0, 0, 0.034, 0, 0.625, 0) encodes both compositional and spatial dependencies, enabling downstream machine learning models to capture biological sequence logic.

Machine learning model frameworks

Five predictive model frameworks were developed by integrating ACNVE-K features with ensemble learning algorithms to identify potential TFs, target genes, and regulatory relationships (activation or inhibition).

(1) **Model 1 –** Identification of TFs.

This model was designed to predict whether a given gene sequence functions as a TF. Known TFs obtained from the TRRUST database served as positive samples, while genes without any regulatory annotations were used as negative samples. To ensure robust performance, the dataset was divided into 90% training, 9% validation, and 1% test sets, with 3-fold cross-validation applied. Three machine learning algorithms—GB [27], RF [28], and XGBoost [29]—were employed for model training. A comprehensive grid search was conducted for parameter optimization: GB was tuned with *learning_rate* ∈ {0.1, 0.15}, *max_depth* ∈ {5, 10}, and *n_estimators* ∈ {50, 100, 150, 200}, yielding 16 models; RF with *max_depth* ∈ {20, 50, 100}, *min_samples_leaf* ∈ {2, 4, 6, 8}, *min_samples_split* ∈ {10, 20}, and *n_estimators* ranging from 1000 to 2000, producing 168 models; and XGBoost with *learning_rate* ∈ {0.05, 0.1, 0.15}, *max_depth* ∈ {5, 20, 40, 60}, and *n_estimators* ∈ {250, 350, 500, 550}, generating 48 models.

(2) **Model 2 framework –** Identification of Target Genes.

The second model aimed to determine whether a gene exhibits characteristics of being regulated by transcription factors. Positive samples consisted of experimentally verified target genes from the TRRUST database, while negative samples were genes lacking evidence of TF regulation. The data partitioning strategy, algorithmic approaches, and parameter tuning procedures were identical to those employed in Model 1, ensuring methodological consistency and comparability across models.

(3) **Model 3 framework –** Prediction of TF–target regulatory relationships.

Model 3 was constructed to identify potential regulatory relationships between TFs and their target genes. Positive samples consisted of known TF–target gene pairs curated from the TRRUST database. Negative samples were generated by randomly pairing TFs (with prediction $P < .1$ from Model 1) and target genes (with prediction $P < .1$ from Model 2). Training, validation, and testing procedures were consistent with Model 1, ensuring a rigorous and standardized modeling framework.

(4) **Model 4 framework –** Prediction of activation relationships.

This model was designed to specifically predict transcriptional activation relationships between TFs and target genes. Positive samples comprised TF–target pairs with experimentally verified activating interactions in the TRRUST database, while negative samples were created following the same strategy as in Model 3. The Gradient Boosting, Random Forest, and XGBoost algorithms were again employed, using the identical parameter settings as in Model 1.

(5) Model 5 framework – Prediction of inhibitory relationships.

The final model was established to predict inhibitory regulatory relationships between TFs and their targets. Positive samples were inhibitory TF–target pairs derived from the TRRUST database, whereas negative samples were generated in accordance with the random pairing strategy used in Model 3. All machine learning algorithms and hyperparameter configurations were maintained as in previous models to guarantee consistency across the analytical pipeline.

Collectively, these five machine learning frameworks constitute an integrated and systematic platform capable of identifying TFs, recognizing their potential target genes, and elucidating the directional nature of their regulatory interactions—whether activating or inhibitory—thereby offering a comprehensive computational approach for decoding transcriptional regulatory networks. Furthermore, we applied SHAP analysis to interpret the important gene features and feature groups identified by GB, RF and XGBoost within each machine-learning framework.

Results

Overview of the analytical framework

To identify potential TFs, target genes, and their regulatory, activating, and inhibitory interactions, we established a systematic analytical pipeline (Fig. 1). First, genomic sequences were numerically encoded using the ACNVE-K method. Subsequently, three machine learning algorithms—GB, RF, and XGBoost—were applied to construct five predictive models corresponding to distinct regulatory tasks.

3Representation of gene sequences in multidimensional feature spaces

Comprehensive gene annotations in GTF format and genomic sequences in FASTA format for *H. sapiens* and *M. musculus* were obtained from the GENCODE database (https://www.gencodegenes.org/) [30]. Human genome releases H31, H43, and H47, and mouse releases M24, M30, and M36 were used.

Five representative genes—*SMAD3*, *SMAD4*, *STAT5A*, *MYC*, and *FERD3L*—were selected to illustrate the multi-order (2-mer to 5-mer) features derived from the ACNVE-K method (Fig. 2; Supplementary Table S4). The human genome sequence lengths were 131 411, 56 518, 24 397, 7518, and 640 bp, respectively, while the mouse sequences measured 111 228, 64 772, 25 819, 4984, and 886 bp.

For *SMAD3* (human H47 release), dimensions 1–4 of the natural vector corresponded to the nucleotide counts (A = 31,643; C = 30,426; G = 33,090; T = 36,252; Fig. 2A), and dimensions 5–8 represented the mean positional indices (65 675.9; 66 459.9; 65 586.4; 65 208.7). Dimensions 9–24 captured second-order asymmetric covariances between nucleotide pairs (ranging 38 539–198 046), dimensions 25–88 represented third-order features (2600–23 268), dimensions 89–344 represented fourth-order (720–16 512), and dimensions 345–2368 represented fifth-order covariances (39–13 771). Similar distributional patterns were observed across both human and mouse orthologs (Figs 2A–F), confirming that ACNVE accurately captures gene-specific structural characteristics and directional dependencies across species.

Predicting potential TFs by model 1 framework

We first evaluated how k-mer dimensionality affects TF prediction using human (H47) and mouse (M36) genomes. Across training datasets, 5-mer representations consistently achieved higher AUC (Area Under the ROC Curve) values of ROC curves than lower-order configurations, with XGBoost outperforming GB and RF (Fig. 3A and B). Optimal models were selected and further validated across all sets (Figs 3C–F), demonstrating excellent predictive reliability.

When comparing genome releases (H31, H43, H47 for human; M24, M30, M36 for mouse), the latest versions yielded superior model performance (Figs 3G–L). Using the final 5-mer XGBoost models (available at https://github.com/sangmm12/TF/model1), we predicted TF probabilities for 77 307 human (H47) and 77 981 mouse (M36) genes. Most genes lacked TF characteristics (Fig. 3M and N); however, 11 127 human and 10 077 mouse genes had $P > .9$, and 3300 human and 3813 mouse genes exceeded 0.99 (Supplementary Tables S5–S6). Notably, >99% of known TFs were correctly predicted with $P > .95$ (except *ZFPM2*, *NRG1* in human; and *Sox5*, *Npas3*, *Camta1*, *Hdac9*, *Zbtb20*, *Rora* in mouse; Fig. 3O and P; Supplementary Tables S7–S8). Most predicted TFs were protein-coding genes (Supplementary Fig. S1). Moreover, a comparative analysis identified 60 TFs that were uniquely present in humans but not in mice. Evaluation with the mouse prediction model revealed that most of the corresponding genes had TF $P >$ .8 (Supplementary Table S9). In contrast, only two TFs were exclusively detected in mice, among which one displayed a probability exceeding 0.8 in humans based on the human prediction model (Supplementary Table S10).

Predicting potential target genes by model 2 framework

We next built Model 2 to identify potential target genes. Consistent with Model 1, the 5-mer XGBoost configuration achieved the best performance across human and mouse genomes (Figs 4A–F). The latest genome versions again showed enhanced predictive power (Supplementary Fig. S2).

Using the optimal models (https://github.com/sangmm12/TF/model2), we found that among 77,307 human genes, 16,085 had target-like $P > .9$, 7498 > .99, and 122 > 0.999 (Fig. 4G); for mouse genes, 16 268 > 0.9, 9382 > 0.99, and 1677 > 0.999 (Fig. 4H).

Known targets were predicted with >0.95 accuracy (except *CTNNA3* and *CTAG1A* in human; Figs 4I and J; Supplementary Tables S11–S12), confirming that the 5-mer XGBoost model generalizes effectively.

Predicting potential TF–target regulatory relationships by model 3 framework

Model 3 integrated the predictions from Models 1 and 2 to explore potential regulatory interactions. The 5-mer feature configuration again provided superior discrimination, with XGBoost achieving the highest ROC scores across datasets (Figs 5A–F). The latest genome releases improved model accuracy (Supplementary Fig. S3). Among known TF–target pairs, ~84.9% (human) and ~89.2% (mouse) showed predicted $P >$ 99% (Supplementary Tables S13–S14). The finalized

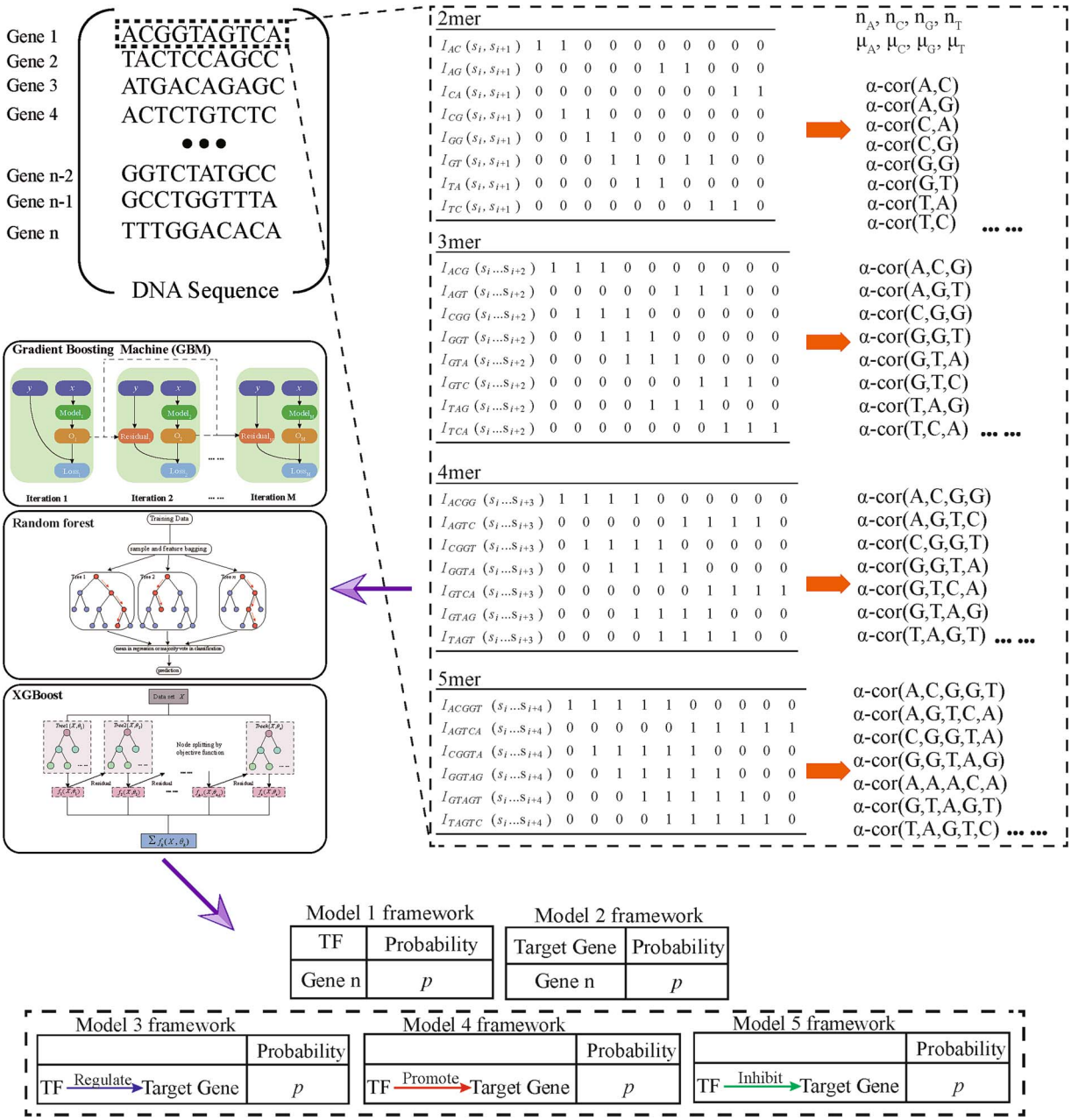

Figure 1 Overview and analytical framework of the study.

models (https://github.com/sangmm12/TF/model3) enable the prediction of novel gene pairs in which one functions as a TF and the other as its regulatory target.

Predicting potential activating relationships by model 4 framework

Model 4 focused on distinguishing transcriptional activation events. The 5-mer gene features again yielded the highest predictive performance (Figs 6A–F), particularly in the latest genome releases (Supplementary Fig. S4). XGBoost consistently outperformed GB and RF in both species. The final models (https://github.com/sangmm12/TF/model4) achieved exceptional accuracy: 99.1% of known human and 98.5% of mouse activation relationships had $P > 99\%$ (Figs 6G–H; Supplementary Tables S15–S16). This framework successfully captures potential positive regulatory links where one gene acts as a TF and the other as its activated target.

Predicting potential inhibitory relationships by model 5 framework

Finally, Model 5 was designed to identify inhibitory TF–target interactions. Across all datasets, the 5-mer representation provided the highest accuracy (Figs 7A–F), with newer genome versions improving predictive stability (Supplementary Fig. S5). In the human genome, 85.9% of known inhibitory pairs had $P > 99\%$ (Fig. 7G; Supplementary Table S17). In mouse, 51.5% had $P > 99\%$, 85.2% > 95%, and 94.6% > 90% (Fig. 7H; Supplementary Table S18).

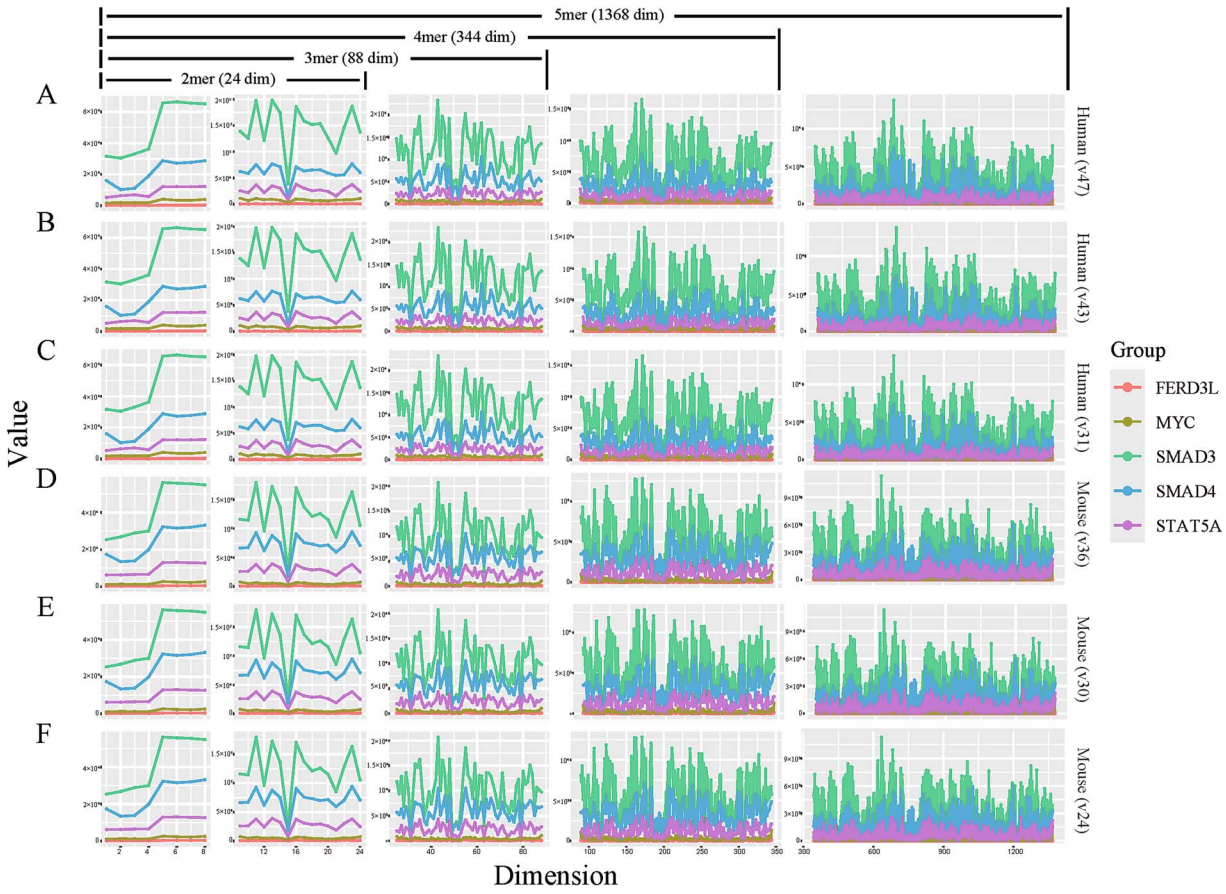

Figure 2 Genetic feature representations (2-mer, 3-mer, 4-mer, and 5-mer dimensions) of FERD3L, MYC, SMAD3, SMAD4, and STAT5A calculated using the ACNVE-K method. Results are shown for (A–C) human genome versions release_H31, release_H43, and release_H47, and (D–F) mouse genome versions release_M24, release_M30, and release_M36 from the GENCODE database.

The final models (https://github.com/sangmm12/TF/model5) demonstrate robust performance in capturing negative regulatory dynamics across species.

Cross-database validation using the TRRUST resource

To assess whether our CollecTRI-based models can generalize beyond their training resource, we performed an external validation using the human TRRUST database by restricting the analysis to TFs, target genes, and TF–target interactions that are present in TRRUST but absent from CollecTRI. Within this non-overlapping TRRUST subset, we identified 125 unique TFs, 53 target genes, and 520 TF–target regulatory relationships, including 119 activating and 185 inhibitory interactions (Supplementary Tables S19–S23). Applying the five human machine-learning frameworks to this independent benchmark, 104 of the 125 TFs (83.2%) and 37 of the 53 target genes (69.8%) were assigned predicted $P > .5$. At the interaction level, 471 of the 520 TF–target pairs (90.6%), 73 of the 119 activating edges (61.3%), and 92 of the 185 inhibitory edges (49.7%) received predicted probabilities above 0.5 (Supplementary Tables S19–S23, Supplementary Fig. S6).

Cross-species validation of TF–target regulatory predictions

Furthermore, for activating and inhibitory regulatory relationships that are known in humans but not in mice, we applied the optimal mouse Models 5 and 6 to predict their counterparts in mice. Interestingly, these models also identified similar relationships in mice, such as *Rela* activating *Tnfrsf10b* with a predicted $P = .9999$ and *Nr5a1* inhibiting *Cyp11b2* with a $P = .8571$ (Supplementary Table 24). Conversely, the optimal human Models 5 and 6 successfully predicted activating and inhibitory regulatory relationships that are unknown in humans but known in mice, highlighting the cross-species predictive capability of our framework (Supplementary Table 24).

SHAP-based analysis of ACNVE-K gene feature importance

To identify which gene features are most informative across the five model frameworks, we used SHAP to evaluate the importance of all 1368 gene-feature dimensions based on RF, GB, and XGBoost algorithms. In Model 1, we first calculated feature importance with GB, RF, and XGBoost and examined the top 20 features for each algorithm; CGGCG, CGCCG, and CGCAG consistently ranked among the most important features across all three (Supplementary Fig. S7). In Model 2, CG, CGA, TAAA, CGGCG, CGCCG, and CGCGG emerged as key features (Supplementary Fig. S8). In Model 3, the target-gene features AA, AATA, TAAA, AATAA, CGGCG, and ATAAA showed the highest importance (Supplementary Fig. S9), whereas in Model 4 the TF features GATCA, GCTTG, and TAACG were particularly prominent (Supplementary Fig. S10). In contrast, for Model 5 no gene feature

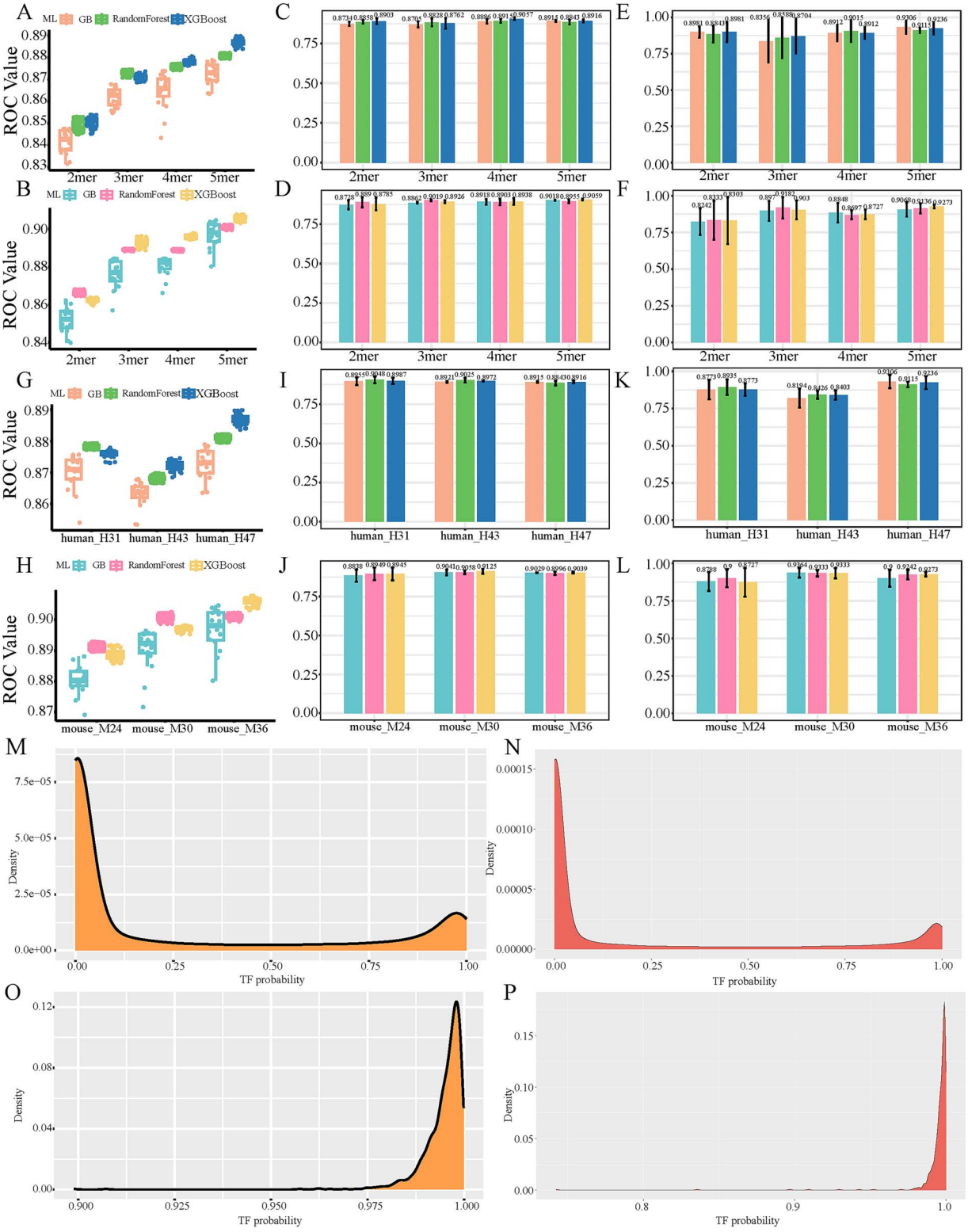

Figure 3 Prediction of potential TFs using the model 1 framework. (A, B) Receiver operating characteristic (ROC) curves of the training sets evaluated using three machine learning algorithms (GB, RF, and XGBoost) across different k-mer dimensions in (A) the human genome (release_H47) and (B) the mouse genome (release_M36). (C–F) ROC curves of the (C, D) validation and (E, F) testing sets based on the optimal GB, RF, and XGBoost models across different k-mer dimensions in (C, E) the human genome (release_H47) and (D, F) the mouse genome (release_M36). (G, H) ROC curves of the training sets across 5-mer dimensions in three genome versions of (G) human and (H) mouse. (I–L) ROC curves of the (I, J) validation and (K, L) testing sets based on the optimal GB, RF, and XGBoost models across 5-mer dimensions in three genome versions of (I, K) human and (J, L) mouse. (M, N) Density distribution of predicted TFs in (M) human and (N) mouse genomes. (O, P) Prediction probabilities of known TFs in (O) human and (P) mouse genomes.

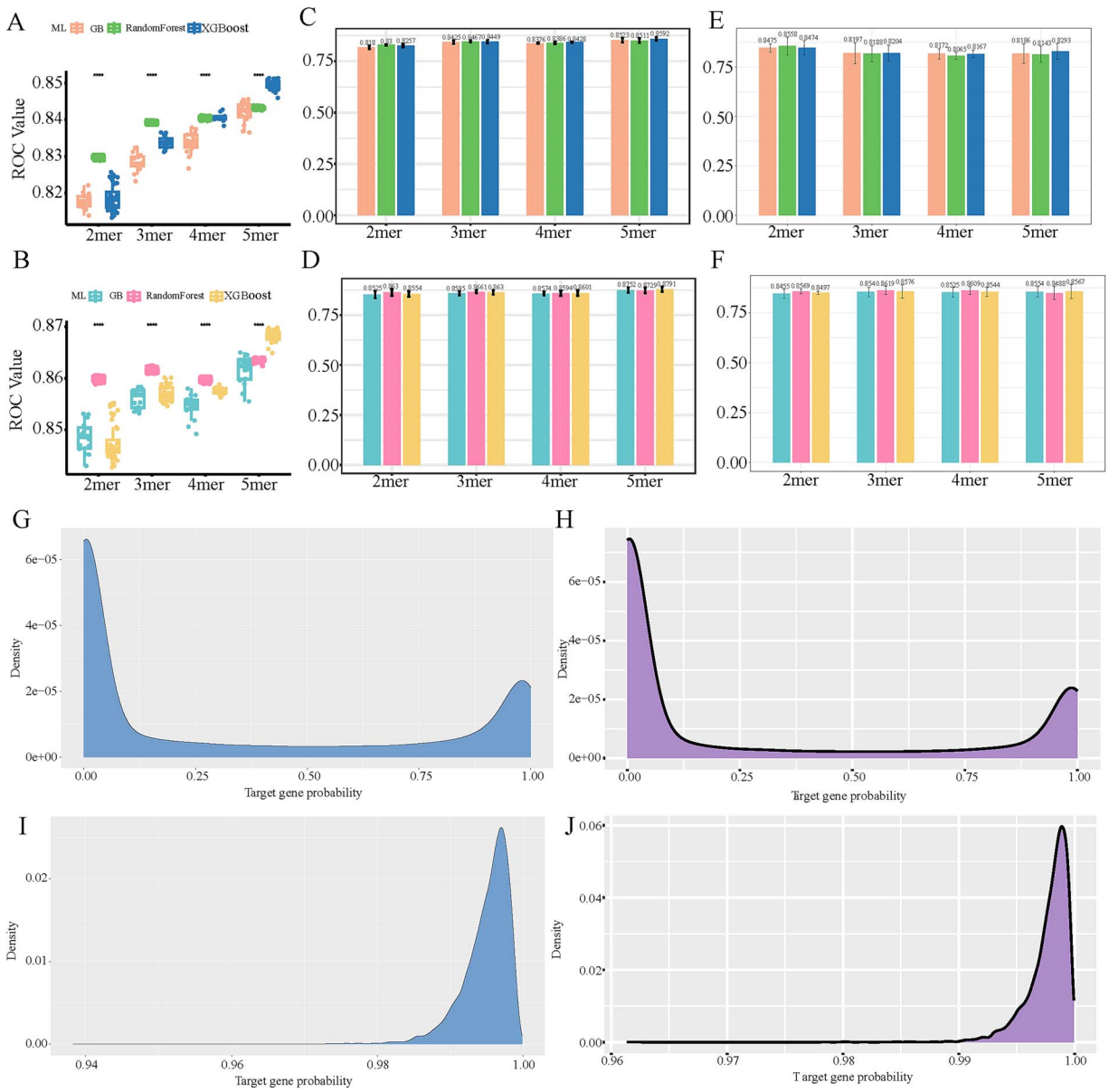

Figure 4 Prediction of potential target genes using the model 2 framework. (A, B) ROC curves of the training sets evaluated using three machine learning algorithms (GB, RF, and XGBoost) across different k-mer dimensions in (A) the human genome (release_H47) and (B) the mouse genome (release_M36). (C–F) ROC curves of the (C, D) validation and (E, F) testing sets based on the optimal GB, RF, and XGBoost models across different k-mer dimensions in (C, E) the human genome (release_H47) and (D, F) the mouse genome (release_M36). (G, H) Density distribution of predicted target genes in (G) human and (H) mouse genomes. (I, J) Prediction probabilities of known target genes in (I) human and (J) mouse genomes.

consistently stood out among the top 20 features across the three algorithms (Supplementary Fig. S11).

Furthermore, we decomposed the 1368-dimensional gene features into five groups according to the order of the directional covariance: NV1 (V1–V8), base counts and first positional moments (traditional natural vector components); NV2 (V9–V24), 2-mer asymmetric covariance features; NV3 (V25–V88), 3-mer asymmetric covariance features; NV4 (V89–V344), 4-mer asymmetric covariance features; and NV5 (V345–V1368), 5-mer asymmetric covariance features. Using GB, RF, and XGBoost, we then assessed the grouped importance of NV1–NV5 in the Model 1 and Model 2 frameworks. In both models, all three algorithms consistently identified NV5 as the dominant feature group, with its importance exceeding the combined contribution of NV1–NV4 (Supplementary Fig. S12 and S13).

Discussion

The prediction of TFs and their regulatory relationships represents a critical link between genomic information and phenotypic function. Traditional TF prediction strategies have primarily depended on curated databases [25, 26], which, despite their utility, exhibit several limitations —including insufficient coverage of noncoding regions [27], high false-positive rates in ChIP-seq experiments [28], and limited generalizability of AI-based approaches [29]. To address these

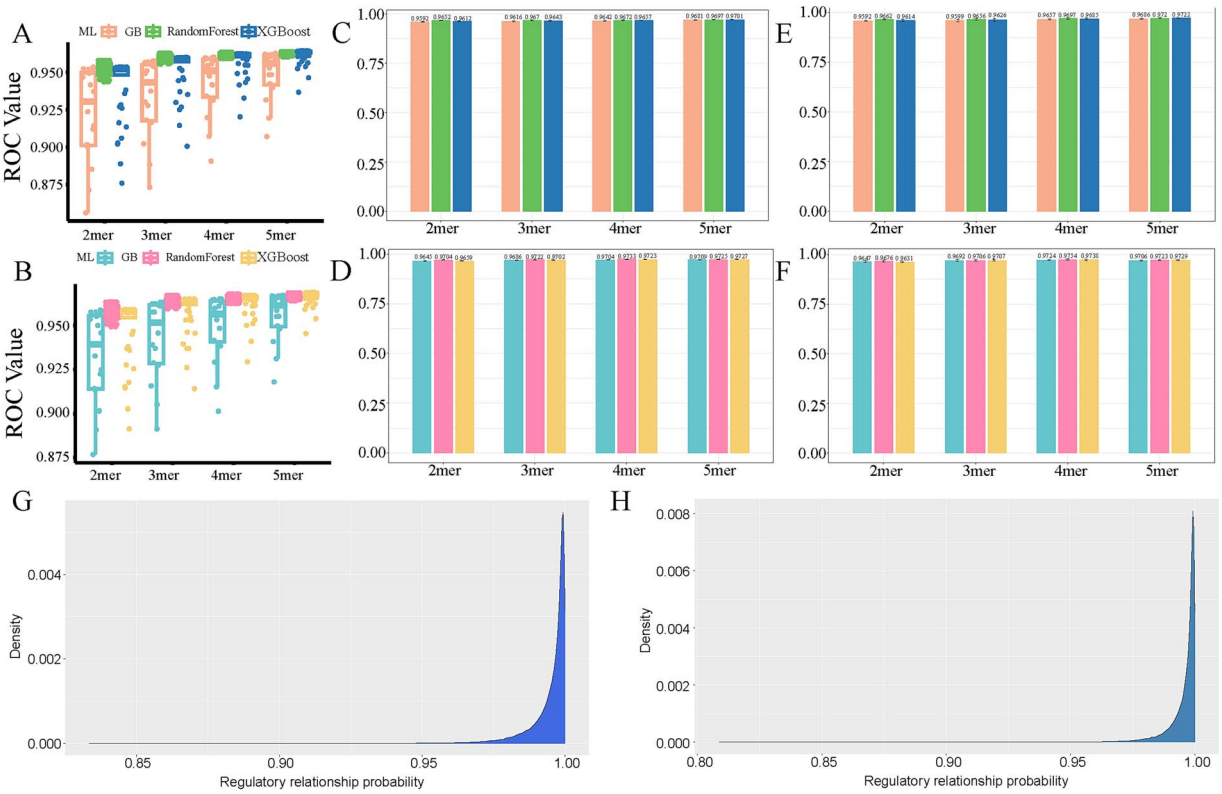

Figure 5 Prediction of potential regulatory relationships between TFs and target genes. (A, B) ROC curves of the training sets evaluated using three machine learning algorithms (GB, RF, and XGBoost) across different k-mer dimensions in (A) the human genome (release_H47) and (B) the mouse genome (release_M36). (C–F) ROC curves of the (C, D) validation and (E, F) testing sets based on the optimal GB, RF, and XGBoost models across different k-mer dimensions in (C, E) the human genome (release_H47) and (D, F) the mouse genome (release_M36). (G, H) Predicted probabilities of known regulatory relationships in (G) human and (H) mouse genomes.

issues, we developed the ACNVE-K framework, which integrates the ACNVE method with *k-mer decomposition* to achieve multidimensional digital representation of genomic sequences. These representations were subsequently incorporated into machine learning models to predict TFs, target genes, and their potential regulatory, activating, and inhibitory relationships. The predicted associations were then reconstructed into systematic regulatory networks, thereby bridging sequence-level information with functional regulation.

The natural vector framework was originally proposed to numerically characterize genomic sequences by encoding the composition and distribution of nucleotides into a 12-dimensional feature vector [21, 22]. While effective in capturing single-base distribution features, this approach lacked the ability to reflect interactions among nucleotide positions. To address this, Sun *et al.* introduced *symmetric covariance natural vectors*, which incorporated pairwise nucleotide covariances and expanded representation to 18 dimensions [21, 22]. However, symmetric covariance captures only the spatial relationships between nucleotides and neglects directionality. In biological sequences, the order of adjacent bases (e.g. *AC* versus *CA*) is non-interchangeable and carries functional implications, as directional changes in DNA can alter downstream RNA and protein sequences [24]. Our study introduced asymmetric covariance natural vectors, embedding directionality into the covariance structure and thus effectively capturing orientation-dependent nucleotide interactions. This enhancement expanded the feature representation to 24 or higher dimensions, significantly improving the expressive capacity for biological sequence modeling.

The *k*-mer method has long been a cornerstone of sequence analysis, constructing features by counting all possible subsequences of a given length *k*. However, traditional *k*-mer encoding lacks directional and positional sensitivity and often introduces redundancy. By combining the ACNVE with *k*-mer encoding, our study achieved a more comprehensive representation that integrates both positional and directional information. Specifically, genomic sequences were partitioned into continuous fragments of lengths 2–5 (and potentially longer), and asymmetric covariance was applied to each. Empirical results demonstrated that 5-mer data provided the highest predictive accuracy, while increasing *k* generally improved performance. Nonetheless, higher-order *k*-mers also exponentially increased feature dimensionality (24 → 88 → 344 → 1368), introducing sparsity and computational inefficiency. Future research will explore optimization strategies—such as dimensionality reduction, sparse matrix regularization, and kernel-based methods—to balance model complexity, accuracy, and efficiency while enabling scalable exploration of higher-order *k*-mers (e.g. 6–8-mers).

Recent advances in sequencing technologies have substantially improved the completeness and accuracy of reference genomes, particularly for *H. sapiens* and *M. musculus* [30]. Older human genome assemblies (e.g. GRCh37/hg19) contained unresolved gaps and ambiguities in telomeric [31, 32], centromeres [33, 34], and highly repetitive regions [34] as well as polymorphic loci such as the HLA cluster [35]. These omissions limited accurate gene annotation and hindered structural variant and regulatory analyses. The advent of long-read sequencing technologies (PacBio, Oxford Nanopore) enabled the

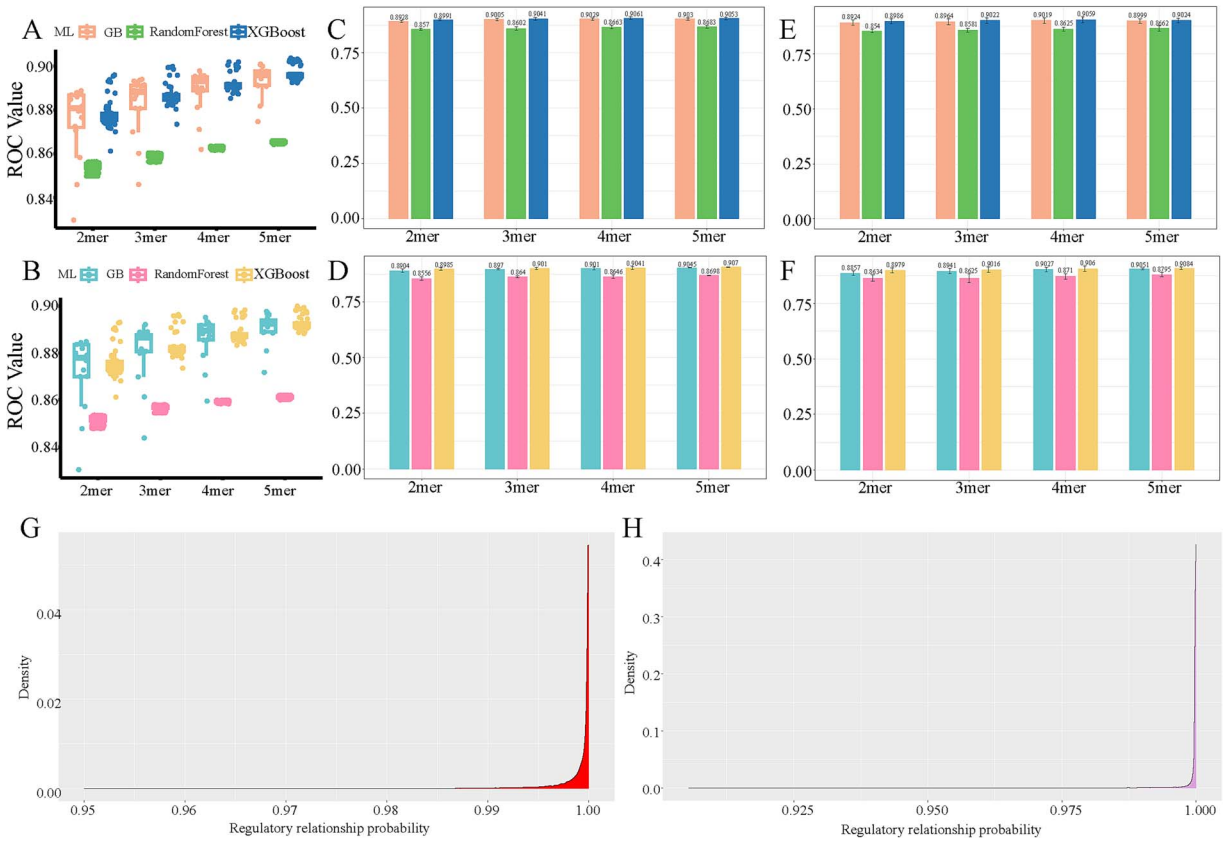

Figure 6 Prediction of potential positive regulatory relationships between TFs and target genes. (A, B) ROC curves of the training sets evaluated using three machine learning algorithms (GB, RF, and XGBoost) across different k-mer dimensions in (A) the human genome (release_H47) and (B) the mouse genome (release_M36). (C–F) ROC curves of the (C, D) validation and (E, F) testing sets based on the optimal GB, RF, and XGBoost models across different k-mer dimensions in (C, E) the human genome (release_H47) and (D, F) the mouse genome (release_M36). (G, H) Predicted probabilities of known positive regulatory relationships in (G) human and (H) mouse genomes.

assembly of GRCh38 and subsequent releases, filling ~150 Mb of previously unresolved sequences and correcting prior annotation errors [32, 36, 37]. Our findings indicate that models trained on the most recent genome versions achieved notably superior predictive performance compared with older assemblies, underscoring the critical role of high-quality genomic references in computational genomics.

To rigorously assess the advantage of ACNVE-K over existing approaches, we benchmarked our Model 3 framework against several representative methods, including GatConv [38], GraphConv [39], GraphSAGE [40], Metapath2vec [41] and HGETGI [42], on the TRRUST regulatory database. Across these comparisons, ACNVE-K consistently achieved the best performance, with an average AUC of 0.9694 on the validation set and 0.9648 on the test set, substantially exceeding the strongest baseline model (AUC = 0.8519; Supplementary Table S25). These results indicate that the directional asymmetric covariance representation captures regulatory signals more effectively than conventional graph-convolutional and embedding-based strategies, providing quantitative support for the superiority of ACNVE-K in TF–target relationship prediction.

Furthermore, our modeling framework is fundamentally different from TF predicted tools such as TFpredict [43], DeepTFactor [20], DeepSEA [17], DeepBind [18] and TF-MoDISco [44], which mainly focus on protein-sequence–based TF identification or local binding-site/motif prediction. In contrast, ACNVE-K uses gene sequences to infer TF identity, target-gene likelihood, TF–target transcriptional relationships and their regulatory direction. Thus, ACNVE-K should be viewed as complementary to these methods, providing genome-level inference of potential TFs, target genes and their interactions.

From our human genome Model 1, *RGL2* was identified as a potential TF (P = .7248; Supplementary Table S5), corresponding to *Rgl2*, a known TF in the mouse genome (Supplementary Table S6). In Model 2, *MFRP* was identified as a potential target gene in humans (P = .9965; Supplementary Table S5), consistent with *Mfrp* as a known mouse target (Supplementary Table S6). Consistent with our sequence-based predictions, existing evidence suggests that RGL2 can act in a transcriptional regulatory capacity, whereas MFRP behaves as a regulated target gene. In plants, RGL2 has been characterized as a key gibberellin-responsive transcriptional repressor, forming promoter-bound complexes (e.g. at the GATA12 and ABI5 loci) that control hormone-responsive gene expression [45, 46]. Although these mechanistic data were obtained in *Arabidopsis* rather than mammalian systems, they demonstrate that RGL2-family proteins can be incorporated into transcriptional regulatory complexes; accordingly, we regard human RGL2 as a candidate TF-like regulator whose role remains to be experimentally tested. For MFRP, Hayward *et al.* showed that Mfrp and Ctrp5 are transcribed as a bicistronic unit in human and mouse retina, and subsequent work has identified MFRP as a direct microRNA target [47] and as a downstream gene of multiple TFs in curated TF–target resources (TRRUST [16],GeneCards [48], GTRD [49], MotifMap [50], JASPAR [51], including BPTF, EGR1, YY1, PITX2

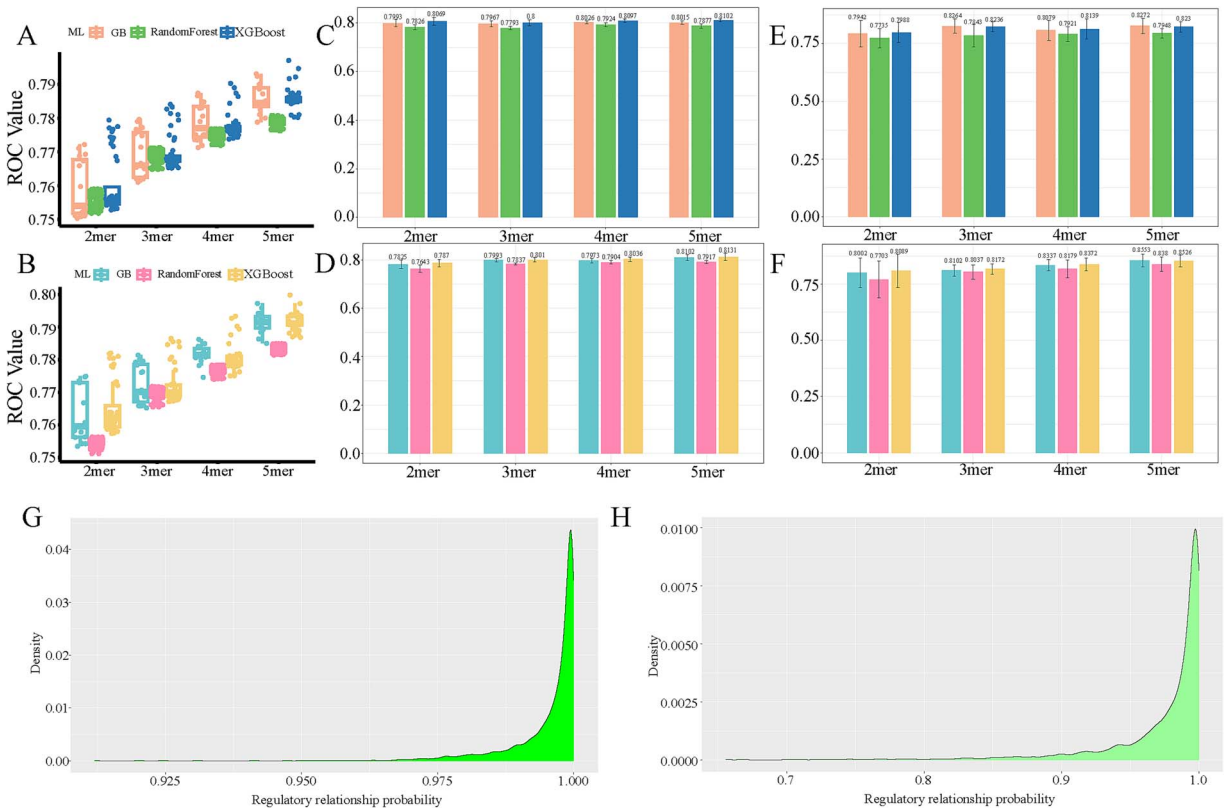

Figure 7 Prediction of potential negative regulatory relationships between TFs and target genes. (A, B) ROC curves of the training sets evaluated using three machine learning algorithms (GB, RF, and XGBoost) across different k-mer dimensions in (A) the human genome (release_H47) and (B) the mouse genome (release_M36). (C–F) ROC curves of the (C, D) validation and (E, F) testing sets based on the optimal GB, RF, and XGBoost models across different k-mer dimensions in (C, E) the human genome (release_H47) and (D, F) the mouse genome (release_M36). (G, H) Predicted probabilities of known negative regulatory relationships in (G) human and (H) mouse genomes.

and SOX10. Taken together, these convergent lines of evidence lend biological plausibility to our assignment of *RGL2* as a TF-like regulator and *MFRP* as a regulated target gene within the ACNVE-K framework.

Additionally, high-probability TF-like candidates such as *CBX4*, *CHD3*, and *KDM6B* exhibited strong regulatory potential. *CBX4*, a Polycomb family protein, participates in chromatin remodeling and transcriptional repression [52]; *CHD3* functions as a core subunit of chromatin-remodeling complexes [53]; and *KDM6B* (*JMJD3*), an H3K27 demethylase, plays key roles in epigenetic regulation, differentiation, and development [54]. Although not classical TFs, these *TF-like* genes mediate regulatory control through chromatin remodeling, RNA processing, and signaling cascades, reflecting their broader importance in gene regulation networks.

Moreover, the predictive models successfully captured cross-species regulatory conservation. For example, while *PITX1* is known to suppress *IFNA1* in humans, our model predicted an activating relationship—consistent with the experimentally verified activation of *Ifna11* by *Pitx1* in mice. Similarly, our human-specific predictions (*MEF2C* → *OPN1MW* inhibition; *RELA* → *TNFRSF10B* suppression) corresponded to verified interactions in the mouse dataset, demonstrating the model's ability to uncover conserved regulatory logic between species.

Notably, several of the top-ranked k-mers in Model 1 (CGGCG, CGCCG, and CGCAG) are GC-rich and contain CpG dinucleotides, closely resembling short GC-box–like elements. GC-box motifs (e.g. GGGCGG) are well-established binding sites for Sp1/Sp3 and related GC-preferring transcription factors in CpG-island promoters [55–57] and are extensively catalogued in motif databases such as JASPAR [51]. More broadly, transcription factor binding preferences are largely encoded by short DNA sequence motifs, and large-scale high-throughput assays have systematically characterized many GC-rich motifs for human transcription factors [58, 59]. This motif-level correspondence provides a mechanistic link between the most important sequence features identified by ACNVE-K and specific TF-binding architectures, further supporting the biological relevance of our feature-importance analysis.

Conclusion

In this study, we established five machine-learning-based predictive frameworks founded on the ACNVE-K encoding scheme to identify TFs, target genes, and their regulatory interactions (activating and inhibitory). These models collectively offer a systematic approach to deciphering gene regulatory dynamics at the sequence level. By integrating advanced feature encoding with robust machine learning strategies, this work provides a computational foundation for guiding experimental validation and accelerating the discovery of functional regulatory mechanisms in complex biological systems.

Limitation and future directions

Gene regulation is inherently context-dependent, exhibiting tissue, developmental, disease, and cell-type specificity. The current

predictions are sequence-based and thus do not yet incorporate such biological contexts. Future work will integrate multi-omics and condition-specific regulatory datasets to generate more context-aware models of gene regulation. Furthermore, the current models provide high-confidence predictions that can guide experimental validation, reducing trial-and-error costs in functional genomics. However, gene regulation often involves synergistic and dynamic interactions that static models cannot fully capture. To address this, we aim to construct dynamic gene regulatory networks that model temporal and multi-factor interactions, thereby improving interpretability in complex diseases. Finally, to enhance accessibility, we plan to develop an interactive visualization platform to enable researchers to explore predicted TF–target relationships intuitively and leverage our models for downstream discovery.

In addition, we plan to compare the ACNVE-K framework with advanced bioinformatics methods incorporating feature extraction and bidirectional attention mechanisms, such as FT_ANPD [60] and Biformer networks [61], and to integrate their advantageous concepts into future extensions of our model.

Key Points

- We integrated the asymmetric covariance natural vector method with the k-mer strategy in a synergistic framework, enabling multi-dimensional numerical representation of gene sequences.
- Using the machine learning algorithms, we developed a predictive model to decipher TF-target gene regulatory relationships, including both activation and repression mechanisms, thereby capturing the complexity of dual regulatory dynamics in gene networks.

Author contributions

Stephen S.-T. Yau, G.H., and M.S. conceived and designed the study. G.H., M.S., H.W., J.G., and L.X. analyzed the data, coding the model and drafted the initial manuscript. Stephen S.-T. Yau made the revision for the initial manuscript. The manuscript was reviewed and approved by all authors.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work is supported by the National Natural Science Foundation of China (No. 12171275), Tsinghua University Education Foundation, the Beijing Natural Science Foundation (Grant No. IS25032, IS25081), the Nantong Science and Technology Project (MS2024053, JC2024064), and the grant from Clinical Medicine Special Research of Nantong University (2024JY061).

Data availability

The datasets analyzed during the current study are available in the TRRUST database (https://www.grnpedia.org/trrust/).

Code availability

The underlying code for this study is available and can be accessed via this link https://github.com/sangmm12/TF/code. The model 1 frameworks for human and mouse genome were in https://github.com/sangmm12/TF/model1. The model 2 frameworks for human and mouse genome were in https://github.com/sangmm12/TF/model2. The model 3 frameworks for human and mouse genome were in https://github.com/sangmm12/TF/model3. The model 4 frameworks for human and mouse genome were in https://github.com/sangmm12/TF/model4. The model 5 frameworks for human and mouse genome were in https://github.com/sangmm12/TF/model5. Other data will be provided upon reasonable request.

Ethics approval

Not applicable.

References

1. Vaquerizas JM, Kummerfeld SK, Teichmann SA *et al.* A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* 2009;**10**:252–63. https://doi.org/10.1038/nrg2538.
2. Oksuz O, Henninger JE, Warneford-Thomson R *et al.* Transcription factors interact with RNA to regulate genes. *Mol Cell* 2023;**83**:e2413. https://doi.org/10.1016/j.molcel.2023.06.012.
3. Boija A, Klein IA, Sabari BR *et al.* Transcription factors activate genes through the phase-separation capacity of their activation domains. *Cell* 2018;**175**:e1816. https://doi.org/10.1016/j.cell.2018.10.042.
4. Trojanowski J, Frank L, Rademacher A *et al.* Transcription activation is enhanced by multivalent interactions independent of phase separation. *Mol Cell* 2022;**82**:e1810. https://doi.org/10.1016/j.molcel.2022.04.017.
5. Fry CJ, Peterson CL. Chromatin remodeling enzymes: who's on first? *Curr Biol* 2001;**11**:R185–97. https://doi.org/10.1016/s0960-9822(01)00090-2.
6. Goos H, Kinnunen M, Salokas K *et al.* Human transcription factor protein interaction networks. *Nat Commun* 2022;**13**:766. https://doi.org/10.1038/s41467-022-28341-5.
7. Balaji S, Babu MM, Iyer LM *et al.* Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J Mol Biol* 2006;**360**:213–27. https://doi.org/10.1016/j.jmb.2006.04.029.
8. McDowell IC, Barrera A, D'Ippolito AM *et al.* Glucocorticoid receptor recruits to enhancers and drives activation by motif-directed binding. *Genome Res* 2018;**28**:1272–84. https://doi.org/10.1101/gr.233346.117.
9. Pendurthi UR, Williams JT, Rao LV. Acidic and basic fibroblast growth factors suppress transcriptional activation of tissue factor and other inflammatory genes in endothelial cells. *Arterioscler Thromb Vasc Biol* 1997;**17**:940–6. https://doi.org/10.1161/01.atv.17.5.940.
10. Li Y, Liang C, Easterbrook S *et al.* Investigating the functional implications of reinforcing feedback loops in transcriptional regulatory networks. *Mol BioSyst* 2014;**10**:3238–48. https://doi.org/10.1039/c4mb00526k.

11. Berenson A, Lane R, Soto-Ugaldi LF *et al.* Paired yeast one-hybrid assays to detect DNA-binding cooperativity and antagonism across transcription factors. *Nat Commun* 2023;**14**:6570. https://doi.org/10.1038/s41467-023-42445-6.
12. Weighill D, Ben GM, Glass K *et al.* Predicting genotype-specific gene regulatory networks. *Genome Res* 2022;**32**:524–33. https://doi.org/10.1101/gr.275107.120.
13. Rauluseviciute I, Riudavets-Puig R, Blanc-Mathieu R *et al.* JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2024;**52**:D174–82. https://doi.org/10.1093/nar/gkad1059.
14. Taing L, Dandawate A, L'Yi S *et al.* Cistrome data browser: integrated search, analysis and visualization of chromatin data. *Nucleic Acids Res* 2024;**52**:D61–6. https://doi.org/10.1093/nar/gkad1069.
15. Kheradpour P, Kellis M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* 2014;**42**:2976–87. https://doi.org/10.1093/nar/gkt1249.
16. Han H, Cho JW, Lee S *et al.* TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res* 2018;**46**:D380–6. https://doi.org/10.1093/nar/gkx1013.
17. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 2015;**12**:931–4. https://doi.org/10.1038/nmeth.3547.
18. Alipanahi B, Delong A, Weirauch MT *et al.* Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;**33**:831–8. https://doi.org/10.1038/nbt.3300.
19. Eraslan G, Avsec Z, Gagneur J *et al.* Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* 2019;**20**:389–403. https://doi.org/10.1038/s41576-019-0122-6.
20. Kim GB, Gao Y, Palsson BO *et al.* DeepTFactor: a deep learning-based tool for the prediction of transcription factors. *Proc Natl Acad Sci USA* 2021;118:e2021171118. https://doi.org/10.1073/pnas.2021171118.
21. Deng M, Yu C, Liang Q *et al.* A novel method of characterizing genetic sequences: genome space with biological distance and applications. *PLoS One* 2011;**6**:e17293. https://doi.org/10.1371/journal.pone.0017293.
22. Sun N, Pei S, He L *et al.* Geometric construction of viral genome space and its applications. *Comput Struct Biotechnol J* 2021;**19**:4226–34. https://doi.org/10.1016/j.csbj.2021.07.028.
23. Muller-Dott S, Tsirvouli E, Vazquez M *et al.* Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Res* 2023;**51**:10934–49. https://doi.org/10.1093/nar/gkad841.
24. Rits S, Olsen BR, Volloch V. Protein-encoding RNA to RNA information transfer in mammalian cells: RNA-dependent mRNA amplification. Identification of chimeric RNA intermediates and putative RNA end products. Ann Integr. *Mol Med* 2019;**1**:23–47.
25. Zhao A, Zhou S, Yang X *et al.* Transcription factor networks and novel immune biomarkers reveal key prognostic and therapeutic insights in ovarian cancer. *Discov Oncol* 2025;**16**:309. https://doi.org/10.1007/s12672-025-01788-w.
26. Wanniarachchi DV, Viswakula S, Wickramasuriya AM. The evaluation of transcription factor binding site prediction tools in human and Arabidopsis genomes. *BMC Bioinf* 2024;**25**:371. https://doi.org/10.1186/s12859-024-05995-0.
27. Barenboim M, Manke T. ChroMoS: an integrated web tool for SNP classification, prioritization and functional interpretation. *Bioinformatics* 2013; **29**:2197–8. https://doi.org/10.1093/bioinformatics/btt356.
28. Karczewski KJ, Snyder M, Altman RB *et al.* Coherent functional modules improve transcription factor target identification, cooperativity prediction, and disease association. *PLoS Genet* 2014;**10**:e1004122. https://doi.org/10.1371/journal.pgen.1004122.
29. Zhang Y, Wang Z, Zeng Y *et al.* High-resolution transcription factor binding sites prediction improved performance and interpretability by deep learning method. *Brief Bioinform* 2021;22:bbab273. https://doi.org/10.1093/bib/bbab273.
30. Dongare DB, Nishad SS, Mastoli SY *et al.* High-throughput sequencing: a breakthrough in molecular diagnosis for precision medicine. *Funct Integr Genomics* 2025;**25**:22. https://doi.org/10.1007/s10142-025-01529-w.
31. Miga KH, Koren S, Rhie A *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 2020;**585**:79–84. https://doi.org/10.1038/s41586-020-2547-7.
32. Jarvis ED, Formenti G, Rhie A *et al.* Semi-automated assembly of high-quality diploid human reference genomes. *Nature* 2022;**611**:519–31. https://doi.org/10.1038/s41586-022-05325-5.
33. Sullivan LL, Sullivan BA. Genomic and functional variation of human centromeres. *Exp Cell Res* 2020;**389**:111896. https://doi.org/10.1016/j.yexcr.2020.111896.
34. Heard E, Johnson AD, Korbel JO *et al.* The X chromosome from telomere to telomere: key achievements and future opportunities. *Fac Rev* 2021;**10**:63. https://doi.org/10.12703/r-01-000001.
35. Wang S, Wang M, Chen L *et al.* SpecHLA enables full-resolution HLA typing from sequencing data. *Cell Rep Methods* 2023;**3**:100589. https://doi.org/10.1016/j.crmeth.2023.100589.
36. Schuy J, Saether KB, Lisfeld J *et al.* A combination of long- and short-read genomics reveals frequent p-arm breakpoints within chromosome 21 complex genomic rearrangements. *Genet Med Open* 2024;**2**:101863. https://doi.org/10.1016/j.gimo.2024.101863.
37. Hsieh P, Dang V, Vollger MR *et al.* Evidence for opposing selective forces operating on human-specific duplicated TCAF genes in Neanderthals and humans. *Nat Commun* 2021;**12**:5118. https://doi.org/10.1038/s41467-021-25435-4.
38. Veličković P., Cucurull G., Casanova A. *et al.* J. a. p. a. Graph attention networks. 2017. https://doi.org/10.48550/arXiv.1710.10903
39. Kipf T. J. a. p. a. Semi-supervised classification with graph convolutional networks. 2016. https://doi.org/10.48550/arXiv.1609.02907
40. Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. *Presented in Part at the Proceedings of the 31st International Conference on Neural Information Processing Systems*. California, USA: Long Beach, 2017:1024–34.
41. Dong Y, Chawla NV, Swami A. metapath2vec: Scalable representation learning for heterogeneous networks. *Presented in Part at the Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. NS, Canada: Halifax, 2017:135—44.
42. Huang YA, Pan GQ, Wang J *et al.* Heterogeneous graph embedding model for predicting interactions between TF and target gene. *Bioinformatics* 2022;**38**:2554–60. https://doi.org/10.1093/bioinformatics/btac148.
43. Eichner J, Topf F, Drager A *et al.* TFpredict and SABINE: sequence-based prediction of structural and functional characteristics of transcription factors. *PLoS One* 2013;**8**:e82238. https://doi.org/10.1371/journal.pone.0082238.
44. Shrikumar A., Tian K., Avsec Ž. *et al.* J. a. p. a. Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5. 6.5. 2018. https://doi.org/10.48550/arXiv.1811.00416

45. Ravindran P, Verma V, Stamm P *et al.* A novel RGL2-DOF6 complex contributes to primary seed dormancy in Arabidopsis thaliana by regulating a GATA transcription factor. *Mol Plant* 2017;**10**:1307–20. https://doi.org/10.1016/j.molp.2017.09.004.
46. Liu X, Hu P, Huang M *et al.* The NF-YC-RGL2 module integrates GA and ABA signalling to regulate seed germination in Arabidopsis. *Nat Commun* 2016;**7**:12768. https://doi.org/10.1038/ncomms12768.
47. Tian X, Zheng Q, Xie J *et al.* Improved gene therapy for MFRP deficiency-mediated retinal degeneration by knocking down endogenous bicistronic Mfrp and Ctrp5 transcript. *Mol Ther Nucleic Acids* 2023;**32**:843–56. https://doi.org/10.1016/j.omtn.2023.05.001.
48. Safran M., Dalah I., Alexander J. *et al.* GeneCards version 3: the human gene integrator. *Database (Oxford)* 2010;**2010**:baq020. https://doi.org/10.1093/database/baq020**.**
49. Yevshin I, Sharipov R, Kolmykov S *et al.* GTRD: a database on gene transcription regulation-2019 update. *Nucleic Acids Res* 2019;**47**:D100–5. https://doi.org/10.1093/nar/gky1128.
50. Daily K, Patel VR, Rigor P *et al.* MotifMap: integrative genome-wide maps of regulatory motif sites for model species. *BMC Bioinformatics* 2011;**12**:495. https://doi.org/10.1186/1471-2105-12-495.
51. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2022;**50**:D165–73. https://doi.org/10.1093/nar/gkab1113.
52. Hu Q, Su L, Zhao W *et al.* CBX4 regulation of senescence and associated diseases: molecular pathways and mechanisms. *Pharmacol Res* 2025;**215**:107705. https://doi.org/10.1016/j.phrs.2025.107705.
53. Smith R, Sellou H, Chapuis C *et al.* CHD3 and CHD4 recruitment and chromatin remodeling activity at DNA breaks is promoted by early poly(ADP-ribose)-dependent chromatin relaxation. *Nucleic Acids Res* 2018;**46**:6087–98. https://doi.org/10.1093/nar/gky334.
54. Cao Z, Shi X, Tian F *et al.* KDM6B is an androgen regulated gene and plays oncogenic roles by demethylating H3K27me3 at cyclin D1 promoter in prostate cancer. *Cell Death Dis* 2021;**12**:2. https://doi.org/10.1038/s41419-020-03354-4.
55. Kuwahara J, Yonezawa A, Futamura M *et al.* Binding of transcription factor Sp1 to GC box DNA revealed by footprinting analysis: different contact of three zinc fingers and sequence recognition mode. *Biochemistry* 1993;**32**:5994–6001. https://doi.org/10.1021/bi00074a010.
56. Shimada J, Suzuki Y, Kim SJ *et al.* Transactivation via RAR/RXR-Sp1 interaction: characterization of binding between Sp1 and GC box motif. *Mol Endocrinol* 2001;**15**:1677–92. https://doi.org/10.1210/mend.15.10.0707.
57. Kishikawa S, Murata T, Kimura H *et al.* Regulation of transcription of the Dnmt1 gene by Sp1 and Sp3 zinc finger proteins. *Eur J Biochem* 2002;**269**:2961–70. https://doi.org/10.1046/j.1432-1033.2002.02972.x.
58. Jolma A, Yan J, Whitington T *et al.* DNA-binding specificities of human transcription factors. *Cell* 2013;**152**:327–39. https://doi.org/10.1016/j.cell.2012.12.009.
59. Weirauch MT, Yang A, Albu M *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 2014;**158**:1431–43. https://doi.org/10.1016/j.cell.2014.08.009.
60. Norouzi R, Norouzi R, Abbasi K *et al.* DFT_ANPD: a dual-feature two-sided attention network for anticancer natural products detection. *Comput Biol Med* 2025;**194**:110442. https://doi.org/10.1016/j.compbiomed.2025.110442.
61. Kianfar A, Razzaghi P, Asgari Z. Integrating convolutional layers and biformer network with forward-forward and backpropagation training. *Sci Rep* 2025;**15**:7230. https://doi.org/10.1038/s41598-025-92218-y.

Briefings in Bioinformatics, 2026, **27**, bbag044
https://doi.org/10.1093/bib/bbag044
Problem Solving Protocol