Molecular Omics

ROYAL SOCIETY OF CHEMISTRY

RESEARCH ARTICLE

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Mol. Omics*, 2025, **21**, 226

Exon–intron boundary detection made easy by physicochemical properties of DNA†

Dinesh Sharma,[a] Danish Aslam,[a] Kopal Sharma,[a] Aditya Mittal[a] and B. Jayaram*[ab]

Received 2nd December 2024,
Accepted 11th March 2025

DOI: 10.1039/d4mo00241e

rsc.li/molomics

Genome architecture in eukaryotes exhibits a high degree of complexity. Amidst the numerous intricacies, the existence of genes as non-continuous stretches composed of exons and introns has garnered significant attention and curiosity among researchers. Accurate identification of exon–intron (EI) boundaries is crucial to decipher the molecular biology governing gene expression and regulation. This includes understanding both normal and aberrant splicing, with aberrant splicing referring to the abnormal processing of pre-mRNA that leads to improper inclusion or exclusion of exons or introns. Such splicing events can result in dysfunctional or non-functional proteins, which are often associated with various diseases. The currently employed frameworks for genomic signals, which aim to identify exons and introns within a genomic segment, need to be revised primarily due to the lack of a robust consensus sequence and the limitations posed by the training on available experimental datasets. To tackle these challenges and capitalize on the understanding that DNA exhibits function-dependent local physicochemical variations, we present ChemEXIN, an innovative novel method for predicting EI boundaries. The method utilizes a deep-learning (DL) architecture alongside tri- and tetra-nucleotide-based structural and energy features. ChemEXIN outperforms existing methods with notable accuracy and precision. It achieves an accuracy of 92.5% for humans, 79.9% for mice, and 92.0% for worms, along with precision values of 92.0%, 79.6%, and 91.8% for the same organisms, respectively. These results represent a significant advancement in EI boundary annotations, with potential implications for understanding gene expression, regulation, and cellular functions.

Introduction

In the heterogenous world of genomics, eukaryotes stand apart from prokaryotes with a fascinating twist – their genetic blueprints exhibit remarkable complexity.[1] Amongst various captivating elements in eukaryotic DNA, the intriguing EI boundary regions have ignited a blazing spark of interest among researchers.

A gene in eukaryotes is a discontinuous structure composed of a protein-coding region (exon) and a non-coding stretch (intron).[2] During the process of gene expression, the introns are excised from a pre-mRNA after transcription, and the exons are joined together through splicing in various combinations to form mature mRNA products.[3] These EI boundary sites are vital for determining the encoded amino acid sequence and regulating splicing events. These boundaries hold significant medical importance, as many human genetic disorders and diseases result from irregular pre-mRNA splicing.[4] Thus, the demarcation of accurate EI architecture is crucial in eukaryotic genome annotation.

In pursuit of annotating these sites, several attempts have been made in genomics. In the early stages of exploration, researchers relied upon the consensus sequence-based approach.[5,6] Scrutinizing the sequences, character by character, and complementing the findings with the experimental data provided with the initial patterns for their identification. These signals, generally known as splice site (SS) motifs, occur in nucleotide pairs with GT and AG at the 5′ and 3′ ends of the intron, respectively.[7,8] However, at later stages, the emergence of cryptic SSs within all the genes of a particular eukaryotic species and other organisms yielded several diverse consensus stretches.[9–11] The situation is even more complex due to the prevalence of alternative splicing (AS) in eukaryotes. An individual gene can give rise to multiple mRNA isoforms through AS by selectively including or excluding different exons, creating an array of potential protein products.[12] This remarkable phenomenon adds another layer of complexity to the

[a] *Supercomputing Facility for Bioinformatics & Computational Biology (SCFBio), Kusuma School of Biological Sciences, Indian Institute of Technology (IIT) Delhi, Hauz Khas, New Delhi 110016, India. E-mail: bjayaram@chemistry.iitd.ac.in*

[b] *Department of Chemistry, Indian Institute of Technology (IIT) Delhi, Hauz Khas, New Delhi 110016, India*

† Electronic supplementary information (ESI) available. See DOI: **https://doi.org/10.1039/d4mo00241e**

identification of EI boundaries, as the traditional linear gene model no longer suffices.

Researchers have recognized the need for a more comprehensive and reliable approach. Various computational approaches, including sequence alignment-based methods, hidden Markov models (HMMs), and machine learning (ML) techniques, have long been used to annotate these evasive boundaries. For instance, some tools leverage scoring matrices holding valuable sequence pattern information from experimentally verified SSs by identifying conserved nucleotide positions and their frequencies.[13–16] Approaches like Genscan[17] and GenomeScan[18] incorporate additional information from known protein sequences to enhance their predictive power. Advanced algorithms, such as GeneWise,[19] Augustus,[20] Fgenesh,[21] GeneParser,[22] and geneid,[23] are built using dynamic programming models employing a data-driven approach to learn the sequence patterns associated with various genomic elements, including exons, introns, SSs, and other regulatory regions critical for gene structure prediction. Spliceator,[24] a recent approach to SS prediction, harnesses the power of the convolutional neural network (CNN) for its predictive capabilities. The key strength of Spliceator lies in its training process, which uses validated data from a diverse set of over 100 organisms. While these methods demonstrate substantial predictive capabilities, their effectiveness relies heavily on the availability of extensive sequence data, resulting in variable performance from species to species.[13–26]

In addition to the aforementioned methods, RNA sequence-based tools like TopHat,[27] SpliceMap,[28] MapSplice,[29] SplitSeek,[30] LEMONS,[31] and, SpliceAI[32] offer reliable predictions for organisms with or without a reference genome. Additionally, a recent tool, DeltaSplice,[33] helps predict splicing–altering mutations. Though widely used, these tools, too, fall short when it comes to annotating splice junctions at the DNA sequence level. Recent exploration of chromatin organization and nucleosome positioning approach presents a fresh perspective.[34] It has yet to achieve the desired level of sensitivity and specificity. Although valuable insights have been gathered from these studies over the years, it remains clear that novel ideas and newer models are essential for accurately identifying EI boundaries at the DNA level as RNA-Seq methods face challenges,[35,36] limiting their accessibility.

It is widely recognized that DNA within our body exhibits sequence and, more importantly, function-specific local structural and energetic variations.[37–45] These arrangements are necessary to facilitate several biological processes, such as protein interactions, gene expressions, *etc.*[46] Investigations on nucleic acid chemistry have yielded fresh insights into genome architecture, providing researchers with a new perspective on annotation. Consistent findings from studies demonstrate that similar DNA sequences often share similar biophysical properties. Interestingly, however, it is not always the case, the alternative sequences can produce DNA molecules that possess similar structures and energy properties.[47,48] This intriguing phenomenon highlights the complex relationship between DNA sequences and their resulting physicochemical properties.

Our past research has highlighted the significance of physicochemical properties in the characterization and annotation of genomic elements within DNA.[8,49–59] These findings reveal that the biophysical signatures of genomic elements are unique and conserved despite sequence variations at these sites. In line with this trend, EI boundaries also display distinctive structural and energy profiles within DNA, distinguishing them from other genomic regions.[8,58] Advancing our exploration further into EI boundaries, we present a novel approach, ChemEXIN, which utilizes structural and energy characteristics of DNA to identify these boundaries. This method capitalizes on molecular dynamic simulation (MDS) based biophysical features; encompassing the Backbone, Inter-base pair (BP) organization, Intra-BP organization, BP-axis and energetics depicted by hydrogen (H)-bond energy, stacking energy, and solvation energy of DNA to discern the precise EI junctions. ChemEXIN, a DL-based method, has undergone dedicated training and development on EI boundary junctions from the protein-coding genes in *Homo sapiens* (*H. sapiens*), *Mus musculus* (*M. musculus*), and *Caenorhabditis elegans* (*C. elegans*). It is openly accessible on GitHub (**https://github.com/rnsharma478/ChemEXIN**) and has been extensively optimized during development using comprehensive datasets involving rigorous comparisons vis a vis various classification models.

Further, comparing it against widely adopted DNA sequence-based methods such as Spliceator, Fgenesh, geneid, Genscan, and Augustus complements its versatility. While these tools have been widely used, we found that they exhibit limitations, particularly in handling larger and more complex genomic datasets. Spliceator, for example, demonstrated high misclassification rates, especially with the default reliability parameter score. This issue resulted in an over-representation of donor and acceptor SSs, leading to false positives and decreasing specificity. Similarly, Genscan showed suboptimal performance in processing large-scale datasets, highlighting a need for more refined and adaptable models. These challenges emphasize the potential for improvement in accuracy and adaptability when integrating ML and DL methods.

Our findings underscore the robustness and broad applicability of ChemEXIN in accurately predicting EI boundaries across both protein-coding and non-coding genes. This unprecedented performance not only validates the effectiveness of our approach but also positions it as a significant contributor to the progression of eukaryotic gene annotation methodologies concerned at the DNA level.

Methods

EI sequence datasets

From the human genome feature files downloaded from the GENCODE[60] database, we identified and filtered out unique Exon-Start (ES) and Exon-End (EE) positions from the exons (both internal and terminal) of all protein-coding genes (328 368 positions for both ES and EE). Using the human reference genome, we generated two positive sequence datasets around these positions for both ES and EE. Refer to Table S7 of ESI,† File 3 for genome assembly and annotation information.

The Dataset I consist of 401 nucleotides long 328 368 sequences. These sequences were generated through an extraction of 200 nucleotides located both upstream and downstream of the EE, positioned at zero. Similarly, Dataset II was created by spanning 200 nucleotides upstream and downstream of the ES. A negative control dataset consisting of 30 140 sequences, each extending 401 nucleotides, was similarly created using the coding sequences (CDS). These sequences were extracted from the middle of exons with a length greater than 1000 nucleotides.

Characterization parameters

For a comprehensive structural depiction of DNA, we have considered various aspects of its organization, including the Backbone arrangement defined by alpha, beta, gamma, delta, epsilon, zeta, chi, phase, and amplitude; Inter-BP arrangements through shift, slide, rise, tilt, roll, and twist; Intra-BP arrangements encompassing shear, stretch, stagger, buckle, propel, and opening; and the BP-axis, which takes into account *X*-displacement, *Y*-displacement, inclination, and tip.

In contrast to our previous studies, which relied on X-ray-derived di-nucleotide data,[8,56,57] our current research adopts a more comprehensive approach. We incorporate neighboring effects by analyzing the structural attributes of all distinctive tri-nucleotides to obtain parameter values for the Backbone, Intra-BP, and BP-axis and unique tetra-nucleotide steps for the Inter-BP arrangement parameters. The Nucleic Acid Database (NDB) lacks B-DNA structures encompassing all possible tri- and tetra-nucleotide steps. Therefore, akin to our recently published study,[58] we rely on atomistic MDS as the sole viable approach to obtain reliable and transferrable parameters for all the unique nucleotide steps. To obtain these structural parameters, we synthetically designed 13 oligomers and followed the exact methodology outlined in[58,61] and summarized here in Methodology S4 of ESI,† File 2. For the energy parameters, we relied upon our in-house lab software to calculate the values of H-bond energy, stacking energy, and solvation energy over all instances of tri-nucleotide steps.[52]

After computing all structural and energy parameters for each oligomer, we assessed the tri- and tetra-nucleotide steps in the 5′ to 3′ direction corresponding to each property. By averaging these occurrences, we generated comprehensive parameter value tables (present in ESI,† File 1).

EI boundary junction profiling and visualization

Using the tri- and tetra-nucleotide parameter value tables, every sequence within each dataset (328 368 ES/EE sequences and 30 140 CDS) was converted to 28 numerical profiles. To attenuate noise arising due to the flanking regions, a moving average filter of 25 base pairs, established previously in ref. 8 and 56–58, was employed over the entire length of profiles. Within this window, the values were averaged, resulting in a single value for each position. The resulting 374 and 373 long numerical profiles corresponding to the tri- and tetra-nucleotide parameters represented the parameter trend over the sequence. Thereafter, a min–max normalization was applied over these profiles, ensuring all values fell within a standardized range of zero to one.[58]

A visual representation of these profiles was achieved by creating two categories of plots for both the ES and EE parameters. In the first category, all 28 discrete properties belonging to the structural and energy class were averaged across all positions over all the sequences. This yielded a single curve for each parameter that was plotted at the ES and EE and was compared with their corresponding CDS profiles. In the second category, numerical profiles of the structural parameters within a specific class (Backbone, Inter-BP, Intra-BP, and BP-axis) were combined by doing a position-based averaging for all the normalized structural profiles (generated during the first visualization step) within a major category. The structural grouping brings forward the synergistic effects of parameters within that group and generates a single curve corresponding to the four structural classes. All these structural profiles were then plotted along with the energy profiles (the three energy parameters represent different aspects, so they were kept separate). This visualization allowed us to observe the actual trends in Backbone, Inter-BP, Intra-BP, and BP-axis organizations along with the energetic drifts.

Formulation of training datasets

For all the ES and EE sequences, the combined seven numerical profiles were processed to extract a segment of length 50, ranging from position 158 to 207 (this segment represented the exon to intron and intron to exon transitions, and the segment length was uniform throughout all profiles). These segments, during visualization, displayed a distinctive pattern to that of the CDS. These three contrasting patterns within the EI transitions and CDS profiles acted as target classes for our prediction models.

To incorporate contrasting features, the seven parameter profiles from the CDS were generated. However, this time, we employed a slightly different approach to capture the sequence characteristics and eliminated bias from a lower count of CDS (30 140). Since all the profiles of CDS had a smooth trend throughout their entire length, we extracted seven non-overlapping numerical fragments, each 50 in length. This extraction followed an organized, non-redundant approach, starting from position one and advancing in increments of 50 nucleotides (*e.g.*, from position one to position 50, position 51 to position 100, and so forth, ultimately resulting in the final fragment spanning from position 301 to position 350). Consequently, this method yielded a negative training dataset comprising 210 980 CDS corresponding to the seven final parameters.

Training pipeline

The entire approach employed for ChemEXIN is outlined as a framework in Fig. 4. Before advancing to the training phase, our analysis commenced with investigating the correlations among the final parameters. This preliminary step aimed to elucidate the relationships and potential interdependencies between the parameters, providing valuable insights into their collective behavior. Correlation analyses conducted on the 50-length segment for both ES and EE datasets yielded diverse levels of correlation among different pairs of parameters. Nonetheless, these correlations were not much pronounced, except for a few cases observed in both datasets. A feature importance analysis was performed to strengthen the conclusions further. Consequently, for the scope of this study, all seven individual parameters (four structural and three energetic) were considered, and the positive (ES/EE) and negative (CDS) datasets were integrated into a single training sequence file. To get a vigorous prediction pipeline, instead of averaging the 50 values extracted from the numerical profiles of each parameter, all 50 values corresponding to each of the seven primary categories were treated as distinct features. This method retained the full spectrum of information within each category and thus provided us with 350 derived features (50 numerical values corresponding to each category) for each sequence. Advancing towards the training process, the integrated three-dimensional (3D) dataset comprising ∼850 000 sequences was categorized into three classes: 0 for CDS, 1 for ES, and 2 for EE. These sequences were then separated into smaller datasets for extensive training-testing and evaluation. 60 000 sequences having an equal proportion of CDS, ES, and EE were chosen randomly from their respective classes and constituted the blind evaluation dataset. The remaining sequences, after randomization, which ensured unbiasedness, were subjected to a classical 80–20 split to create training-testing datasets. Various ML/DL methods were deployed over these human datasets, and the results were compared. By employing multiple models rather than relying on one, we tried to strengthen the idea that the physicochemical profiles observed at the EI boundaries play a crucial role in predictions, irrespective of the models.

Starting here, we integrated two additional eukaryotes, namely *M. musculus* and *C. elegans* (biophysical profiles at EI sites of these organisms are presented in Fig. S3 and S4 of ESI,† File 4). This approach involved a similar EI and CDS sequence extraction (information detailed in Table S7 of ESI,† File 3) and training-testing split alongside an independent extraction of a benchmarking set. This set comprised 2000 sequences each for ES and EE for each organism. Skipping the organism-specific model evaluation for mice and worm genomes, the best-performing model in humans was deployed for these organisms. Further, to maintain linearity across organisms from the evaluation dataset corresponding to *H. sapiens*, 2000 random ES and EE sequences from this set were selected as a separate benchmarking set for humans.

Moreover, given the distinct genomic structures across different kingdoms, we focused on species within the same kingdom for this study. While we profiled and visualized the EI boundaries of organisms from other kingdoms, such as *Plasmodium falciparum* (Protista), *Saccharomyces cerevisiae* (Fungi), and *Arabidopsis thaliana* (Plantae), the results (Fig. S5–S7 of ESI,† File 4) indicated that different model adaptations are required for these organisms. Therefore, incorporating these organisms into the current framework was beyond the scope of this study. However, we are actively working on addressing these challenges in the subsequent version of ChemEXIN to expand its applicability to a broader range of species.

Evaluation and comparison with the state-of-the-art

To assess our trained method, which includes models from *H. sapiens*, *M. musculus*, and *C. elegans*, we benchmarked it against five widely used gene annotation tools specific to DNA sequences. This state-of-the-art comparison against tools specific to DNA sequences was conducted using the organism-specific benchmarking datasets. Due to limitations of RNA-based biophysical characterization, annotation tools specific to other sequences (transcripts) were not benchmarked against our approach.

Additionally, an advanced comparison between the two top-performing tools and the final model was conducted using non-protein coding gene datasets. These datasets encompass sequences devoid of prior training, thus offering a rigorous evaluation of the efficacy and adaptability of our biophysical-based prediction approach.

Prediction methodology

Moving ahead with creating a novel biophysical parameters-based EI boundary prediction tool, we integrated the three benchmarked models into an easy-to-use programmed pipeline. This OS-independent pipeline, developed entirely in Python 3, is accessible as a command-line tool and is publicly available on GitHub. The exact methodology during a prediction involves validating the input sequence length and characters and then converting the input sequence into seven numerical profiles corresponding to the combined major categories. Subsequently, a transient data frame with an organization similar to our training-testing dataset is created at the backend. This data frame then employs the organism-specific models and the reliability threshold value chosen by the user in addition to the sequence input step. The predictions from the employed models pass through various filters to provide the user with the final EI boundary sites organized in an output file. The detailed prediction pipeline and all the filtering steps are available in the user manual (ESI,† File 5). To test the working of the developed pipeline and its cross-platform functionality, predictions were made by ChemEXIN on varying-length genes from the three organisms in consideration.

Results and discussion

Physicochemical profiles at the EI boundaries

Fig. 1 and Fig. S1, S2 (ESI,† File 4) depict the numerical profiles of all 28 parameters, including the nine Backbone angle parameters, six Inter-BP parameters, six Intra-BP parameters, four BP-axis parameters, and three energy parameters. These profiles were generated for ES, EE sequences, and CDS in *H. sapiens*. The results of this study indicate that the physicochemical profiles at both ES and EE sequences display unique

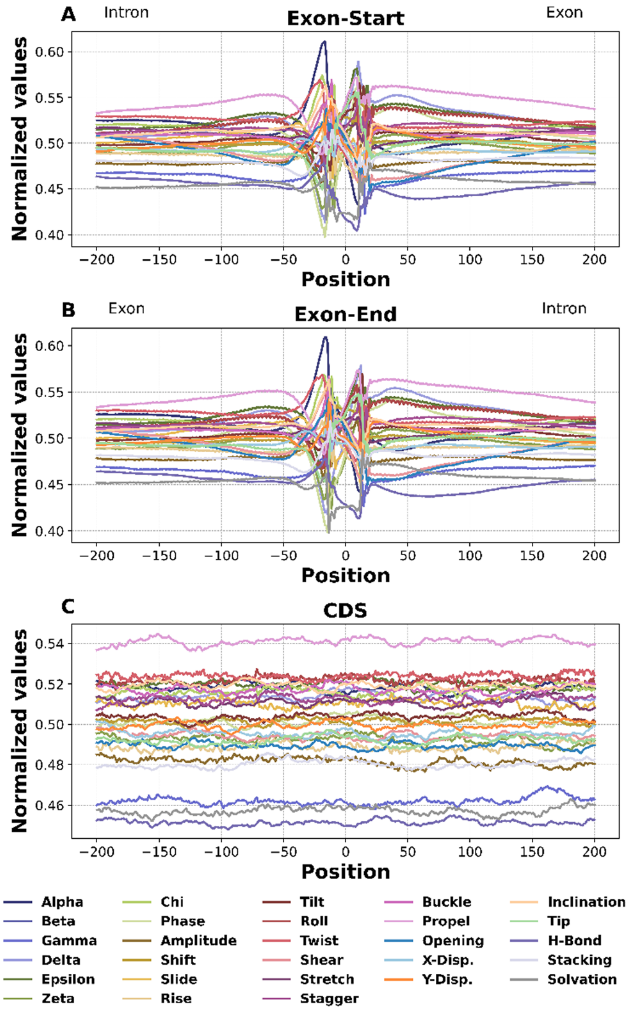

Fig. 1 Profiles of 28 normalized structural and energetic parameters for the three regions. (A) Exon-Start, (B) Exon-End, and (C) CDS regions. Each line represents a different structural or energetic parameter. These parameters show distinct shifts in their profiles at the boundary regions (Exon-Start (ES) and Exon-End (EE)), while the patterns appear stable throughout the CDS regions.

patterns, which differ significantly from those observed in CDS. The results demonstrate that while the structural and energy properties of the CDS remain relatively constant throughout the sequences, a distinct shift occurs in the biophysical profiles at the EI boundary. The structural trends observed at the EI sites for each parameter emphasize the presence of a transient thermodynamically unstable boundary. This boundary may be crucial for facilitating classical splicing events, with exons demonstrating higher stability than neighboring intronic sites.[59,62] Additionally, the energy plots at the EI junctions within these DNA sequences support the classical hypothesis that boundary elements might play a crucial role in secondary structure formation in RNA, thereby facilitating splicing.[7] The H-bond energy exhibited a rapid rise followed by a drop, implying an initial instability at the boundary position that gradually balances out as the junction site progresses. In contrast, the stacking energy reached its maximum value at the border junction, attributing to an increased flexibility in the DNA by reducing its stiffness. The observed decreased solvation energy could indicate the transiently formed stable structure at the interfaces between exons and introns.

Moving with the idea that the combined effect of smaller features brings about a concerted change, we combined the individual structural parameters belonging to the respective major categories to provide us with the actual Backbone, Inter-BP, Intra-BP, and BP-axis profiles. The synergistic visualization of these categories and the three energy parameters at the exon junctions are available in Fig. 2. These results provide us with the evident change at the boundaries for the seven structural and energy parameters. Trends, initially widespread over a region of 50–100 length (Fig. 1) within the individual parameters, are now contained uniformly within a region of ∼50 for all the categories.

The shaded region within the combined plots shows the site undergoing major structural and energy changes. Together, these individual and combined profiles offer valuable insights into the potential utility of the combined parameters for effective EI boundary identification within any given gene sequence.

Correlation analyses and feature importance

A correlation analysis was conducted to examine the interrelationships among the seven final parameters within the 50 nucleotide regions of both the ES and EE profiles in humans. The primary objective of these analyses was to assess the degree of correlation between parameters and identify any redundancy that may exist. Fig. 3 shows the correlation results. Different pairs of parameters exhibited varying degrees of correlation, ranging from moderate to high. Some parameters were dependent on each other, while others showed no correlation. Furthermore, an examination of feature importance through principal component analysis (PCA) was performed to retain the significant features without compromising on information for the downstream analysis. As summarized in Fig. 3 and Methodology S1 of ESI,† File 2, these results emphasized the significance of using all the seven parameters. The framework, as outlined in Fig. 4, was thus followed, leading to the development of the novel physicochemical property-based EI boundary prediction method, ChemEXIN.

Performance evaluation

To arrive at an optimal EI boundary prediction method, various ML/DL models were deployed during the initial development phase of ChemEXIN. The performance of these models underwent comparison using both the training-testing dataset and the evaluation dataset in humans. Model assessment and comparison were conducted using five key criteria: sensitivity, specificity, *F*1-score, precision, and accuracy. The conclusive results of the training-testing are presented in Table S1 of ESI,† File 3. These results indicate that all the models exhibit the capability to predict EI boundary sites, with accuracy levels and *F*1-scores spanning from 53% when utilizing basic models to an improved performance of ∼80% when employing DL model on

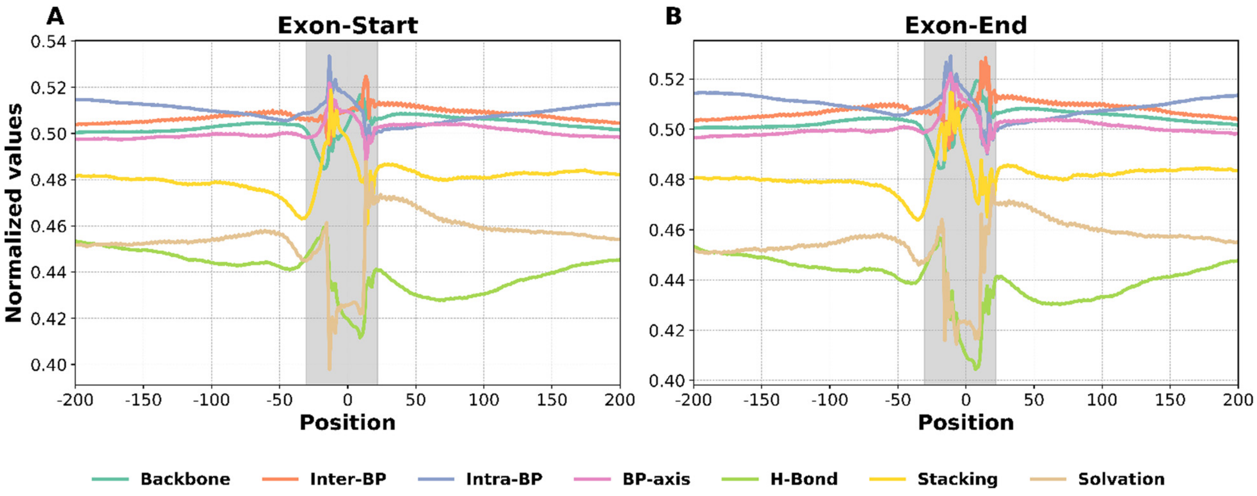

Fig. 2 Profiles of the seven normalized structural and energetic parameters at the EI junctions. (A) Exon-Start (ES), (B) Exon-End (EE). Each line represents a major structural or energetic parameter. The four structural parameters were obtained by combining the individual parameters within that category. These parameters show the actual structural and energy change at the two boundaries.

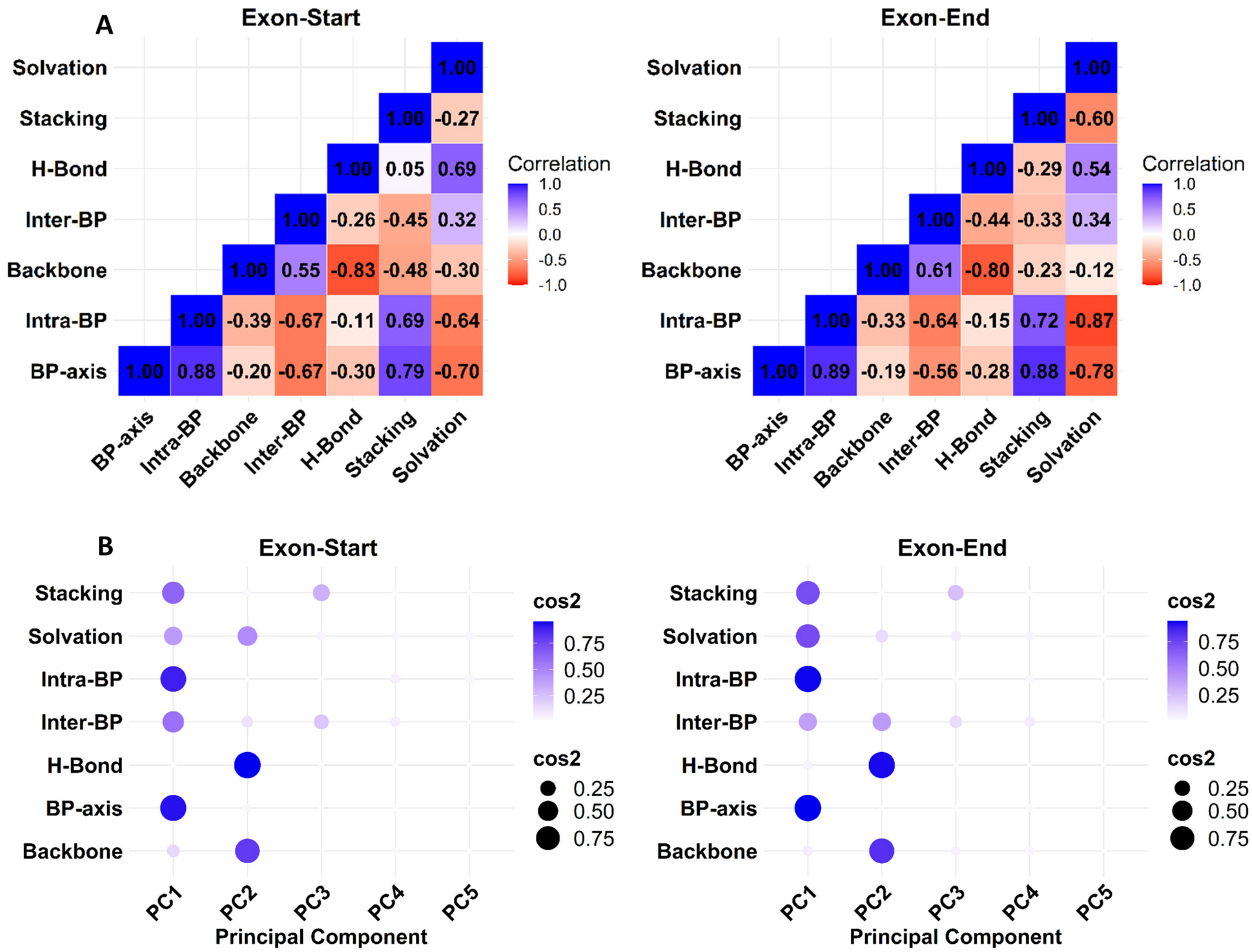

Fig. 3 Correlation and feature importance analyses. (A) Correlation matrices depicting the relationships among the seven final parameters at Exon-Start (ES) and Exon-End (EE). (B) Feature importance analysis conducted through principal component analysis (PCA), revealing significant contributions of all the seven parameters.

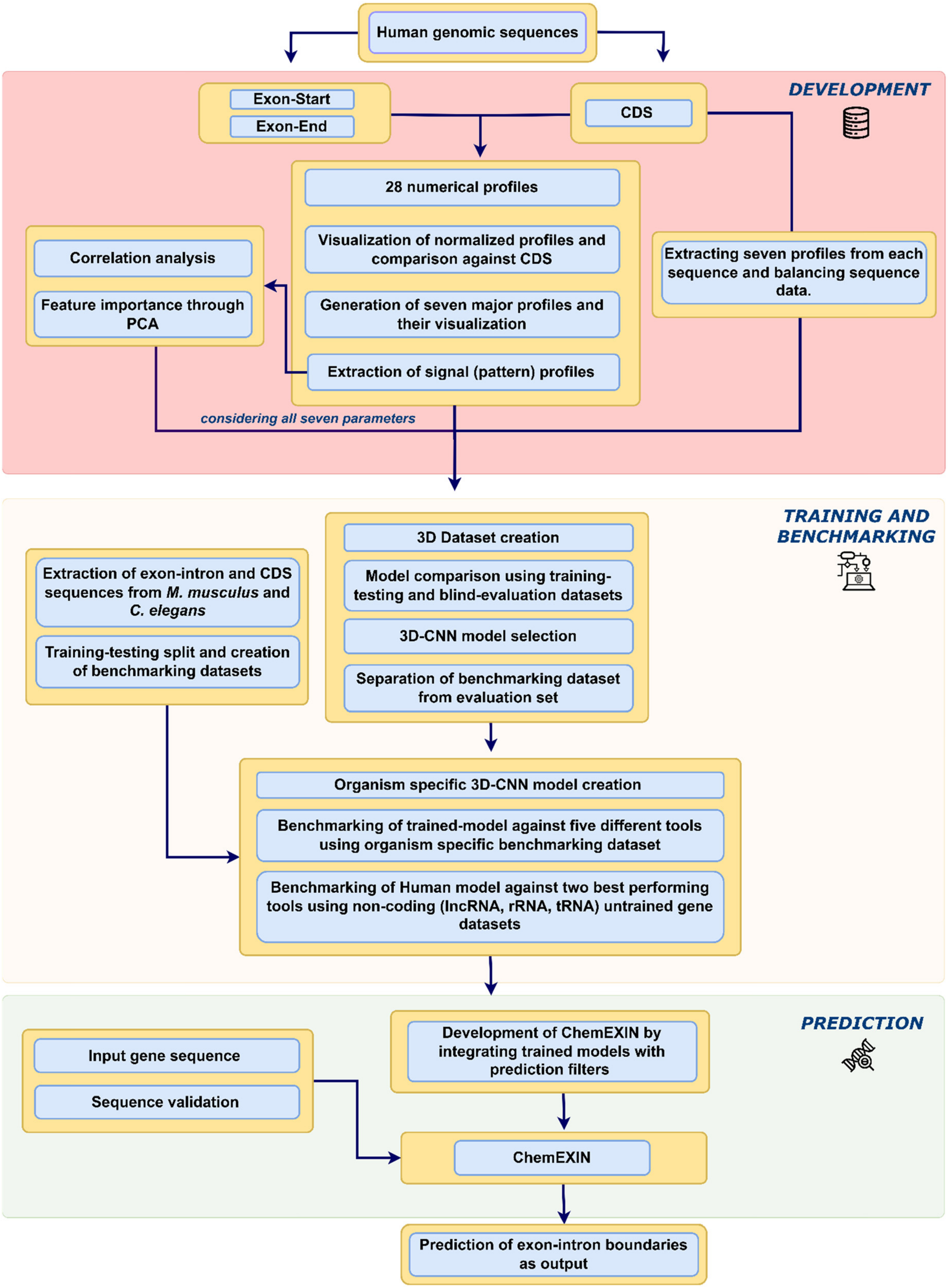
Human genomic sequences
DEVELOPMENT
Exon-Start
Exon-End
CDS
28 numerical profiles
Visualization of normalized profiles and comparison against CDS
Generation of seven major profiles and their visualization
Extraction of signal (pattern) profiles
Correlation analysis
Feature importance through PCA
Extracting seven profiles from each sequence and balancing sequence data.
considering all seven parameters
TRAINING AND BENCHMARKING
3D Dataset creation
Model comparison using training-testing and blind-evaluation datasets
3D-CNN model selection
Separation of benchmarking dataset from evaluation set
Extraction of exon-intron and CDS sequences from M. musculus and C. elegans
Training-testing split and creation of benchmarking datasets
Organism specific 3D-CNN model creation
Benchmarking of trained-model against five different tools using organism specific benchmarking dataset
Benchmarking of Human model against two best performing tools using non-coding (lncRNA, rRNA, tRNA) untrained gene datasets
PREDICTION
Development of ChemEXIN by integrating trained models with prediction filters
Input gene sequence
Sequence validation
ChemEXIN
Prediction of exon-intron boundaries as output

Fig. 4 A detailed framework highlighting the development of ChemEXIN.

the test set. On the same dataset, it is worth noticing that parameters such as specificity, which examines the model's ability to accurately detect true negatives, and sensitivity (true positive rate or recall), which evaluates the system's proficiency in predicting true positives within each category or class, demonstrated notably strong performance values.

The findings from the *H. sapiens* evaluation set comprising 60 000 held-out sequences further validate the efficacy of utilizing biophysical parameters for accurately identifying boundary regions (Table S1, ESI,† File 3). To optimize the performance, we are exploring the inclusion of physicochemical profiles at the pre-mRNA level. By capturing these biophysical features, which reflect structural and energetic variations in the RNA sequences, we anticipate that the model's predictive power could be enhanced. Additionally, we are investigating the integration of more advanced model architectures, including hybrid models that combine DL with other ML techniques. These innovations aim to further refine the model's ability to capture complex patterns, improving prediction accuracy and making it applicable to a broader range of species.

Moving further with our predictive analysis, the area under the receiver operating characteristic curve (AUROC) scores on the evaluation dataset presented in Fig. 5 show that the 3D-CNN and support vector machine (SVM) classifiers surpass other models in predictive performance across all three classes. Notably, the 3D-CNN exhibits higher area under the curve (AUC) values across all classes, signifying its efficacy in distinguishing between diverse classes. Following the comparison results of the above models and the 3D nature of our datasets, we decided to proceed with the 3D-CNN[63] trained model for subsequent analysis and its independent implementation in all three organisms under study. The architecture of the 3D-CNN model employed here is detailed in Fig. 6 and Methodology S2 of ESI,† File 2.

Comparison with the state-of-the-art tools

Five widely used DNA sequence-based gene structure organization prediction tools—Spliceator, Fgenesh, geneid, Genscan, and Augustus, were benchmarked against each of our three organism-specific trained models. The results are presented in Fig. 7 and Tables S3–S5 of ESI,† File 3, with details on the outputs available in Methodology S3 of ESI,† File 2. To ensure an unbiased comparison of our approach, we used three benchmarking datasets, each comprising 2000 randomly selected sequences from the respective organism. Most of these tools are available as web servers, which tend to crash on large input sequences and/or require input sequences in batches. Henceforth, this reasonable-sized comparison data ensured the efficient working of all the tools.

Spliceator, available as a web server,[24] employs CNN in conjunction with a user-defined reliability parameter and a sequence search window to predict the gene organization within input DNA sequences. Instead of treating individual input sequences separately, it processes them as a unified input string with a maximum length of approximately 200 500 bases. To adhere to this constraint, we divided the input sequences for ES and EE for organisms under study into two batches. We employed a default reliability parameter score of 98% and a model tailored to a 400-length search window (as our individual input sequences are 401 nucleotides long) to predict donor and acceptor sites. The output files obtained for each batch were combined into their respective categories and processed to provide a final confusion matrix. From the results, it is evident that Spliceator results are less than satisfactory for all three organisms. The observed high level of misclassification is primarily attributed to Spliceator's consensus-based approach to identifying donor and acceptor sites, resulting in an over-representation of these sites in the predictions. This over-representation tends to increase with a decrease in the reliability parameter score due to non-specific pattern matching. Moreover, there is no noticeable improvement at a 100% reliability parameter score, suggesting its high sequence specificity.

Fgenesh is available both as a web server and as a local downloadable version. Due to the requirement of several genomic feature files in processing the downloadable version, we

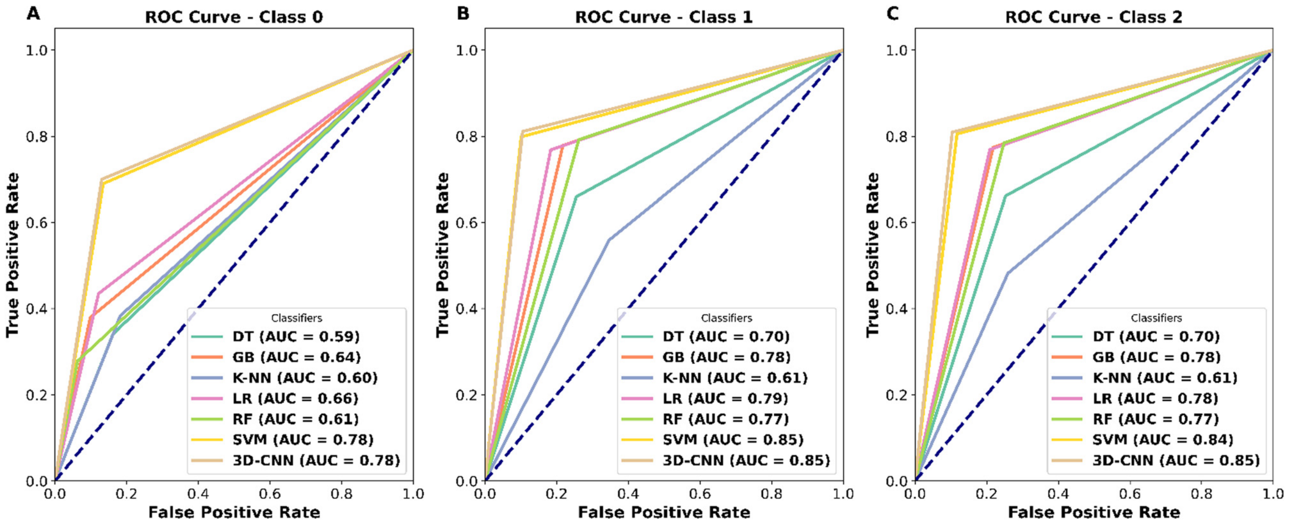

Fig. 5 Area Under the Receiver Operating Characteristic curve (AUROC) depicting AUC scores for all three classes. (A) CDS: 0, (B) Exon-Start (ES): 1, and (C) Exon-End (EE): 2 across all classifiers employed over the blind evaluation set. The classifiers used are DT: decision tree, GB: gradient boosting, K-NN: K-nearest neighbors, LR: logistic regression, RF: random forest, SVM: support vector machine, 3D-CNN: 3D convolutional neural network.

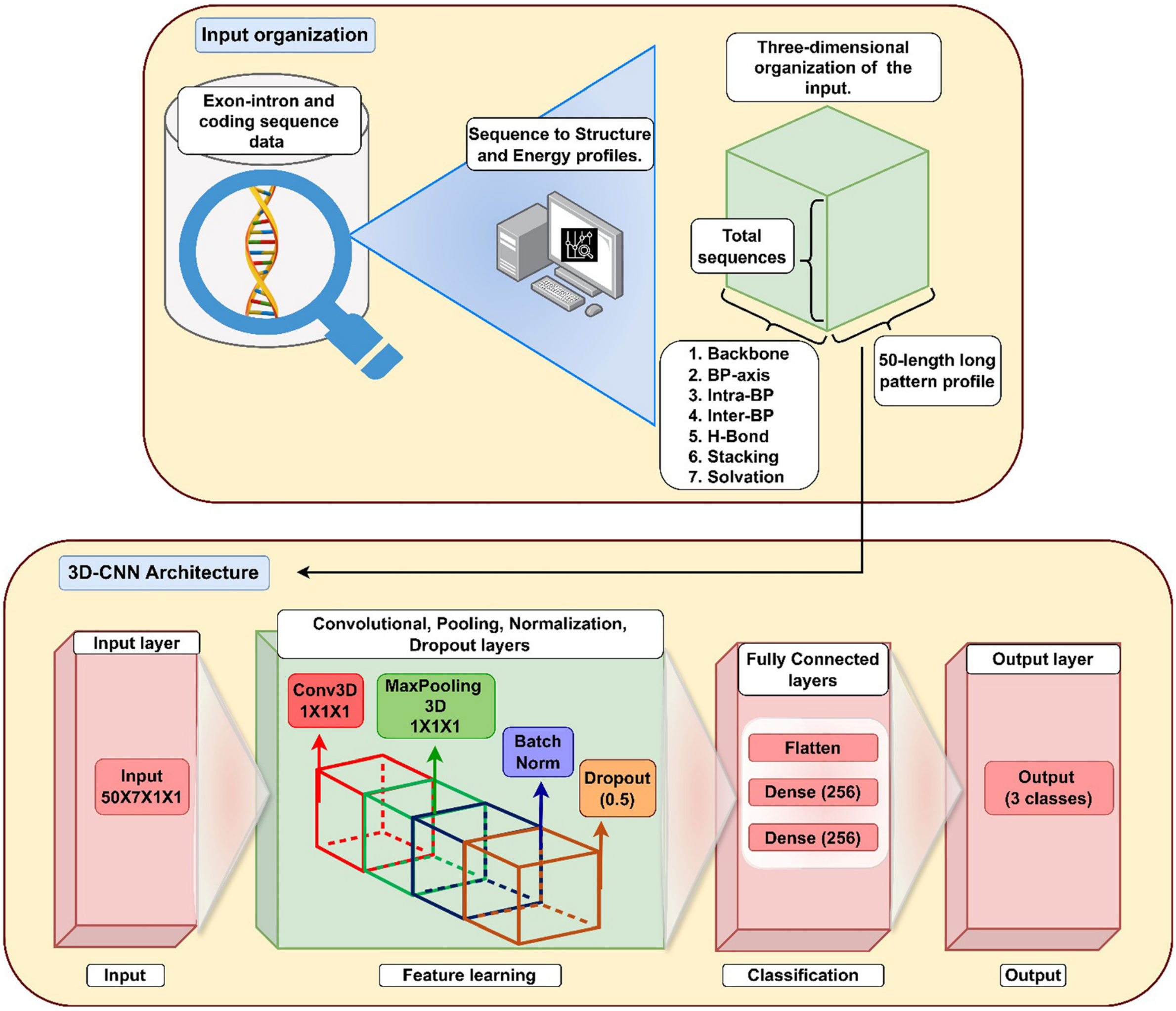

Fig. 6 3D-CNN architecture employed in ChemEXIN. The architecture was designed over the human model and replicated directly for mice and worms.

tested our sequence with the online web server.[21] The method accepts a single file as input, with each sequence represented in FASTA format. In addition to a HMM based gene prediction model trained over several eukaryotic species, though not within the scope of our research, it provides numerous user-specific advanced options. Operating it with organism-specific default parameters, the method generated output files, which were then processed into a confusion matrix. Similar to Spliceator, Fgenesh yielded comparable results for *C. elegans* and *M. musculus*. However, for humans, the precision and accuracy notably improved, reaching close to 40%. Although there is a potential for improved outcomes by utilizing targeted training with specific feature files in the downloadable version (Fgenesh++), we opted not to pursue these advanced options.

geneid, another tool in our evaluation, employs position weight arrays, scoring, and Markov models to identify gene features in DNA sequences. Although available as a web server and a GitHub repository,[23] we faced challenges with the online version, prompting us to resort to the local version downloaded from GitHub.[23] Despite processing input sequences in a manner similar to Fgenesh, the processed results more closely resemble those of Spliceator, exhibiting a high misclassification rate ranging from 80% to 90% for the organisms under consideration. The misclassification observed can be attributed to the overrepresentation of the ES and EE sites. Regardless of being trained on multiple species from all four eukaryotic kingdoms, geneid did not yield satisfactory results in our study.

Continuing our benchmarking efforts, we evaluated Genscan, a widely used tool for identifying EI structures in genomic sequences. Genscan[17] employs general probabilistic models to annotate gene features within input sequences. While Genscan can process nearly one million bases, our dataset, comprising approximately 0.8 million bases, posed a challenge to its processing capabilities. Hence, following a similar strategy employed with Spliceator, we partitioned the input sequences into two batches for both the ES and EE. Regrettably, akin to the outcomes observed with the other tools, the results were not encouraging.

Transitioning to our last tool, Augustus, our objective was to evaluate its proficiency in predicting gene structures. Beyond being accessible as a straightforward pre-trained web server and a GitHub repository, Augustus offers an improved web

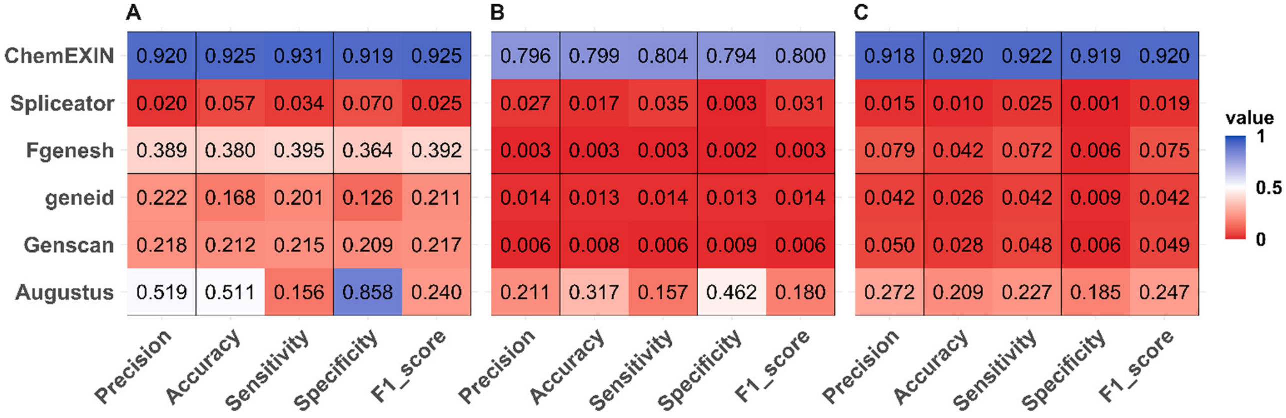

Fig. 7 Heatmaps depicting the performance of all methods across all three organisms. (A) *H. sapiens* (B) *M. musculus* (C) *C. elegans*.

server option. This server allows training sequences not listed in their database using annotation files containing information for cDNA sequences and/or hints for donor and acceptor sites (hint files). We utilized the pre-trained web server[20] and prepared a basic hint file (GFF) for input sequences (FASTA), adhering to the required format. Using a generalized HMM with an additional probabilistic model for gene structure prediction, Augustus also provides information on alternative SSs. Augustus exhibited relatively favorable performance compared to other tools, achieving a specificity of approximately 85%, notably attributed to the utilization of a hint file. In the case of humans, the misclassification rate decreased significantly to approximately 45%. However, while a similar trend was observed for other organisms, the outcomes were less favorable.

In a similar manner to the above-reported comparisons, we examined how well our models performed by looking at their predictions on the benchmarking datasets. This evaluation was essential for understanding how accurately the approach could predict EI boundaries. The analysis unequivocally demonstrates that our approach clearly outperforms other tools (Fig. 7) across all major evaluation criteria in all three organisms. The results indicate a notably low misclassification rate, ranging approximately from 0.075 to 0.20, and high precision, ranging from approximately 0.796 to 0.92. These findings indicate the reliability and accuracy of the predictions obtained through our technique. This exhaustive comparison underscores the presence of substantial sequence alternatives. However, despite these variations, the biophysical profiles at the junction sites remain largely conserved. This conservation suggests the potential utility of these profiles in facilitating precise recognition and prediction by the physicochemical property-driven 3D-CNN models.

Expanding the scope of our comparison, we further assessed the performance of the reported method alongside two top-performing tools identified in the previous benchmarking step, namely Fgenesh and Augustus. This extended comparison focused on predicting EI junctions in non-protein coding genes, including lncRNA genes; tRNA genes; and rRNA genes in humans.

Despite its widespread usage, Fgenesh failed to generate results in our comparative assessment. Unlike Augustus, while our method has not undergone specific training on the EI characteristics of these genes, the results documented in Table 1 and Fig. 8 underscore a notable performance of our framework against Augustus. As evident, Augustus shows higher specificity for EI boundary predictions in tRNA (76.8%), it achieves significantly lower sensitivity (20.0%) compared to ChemEXIN (49.6%). This indicates that Augustus is more conservative in predicting EI boundaries in tRNA, resulting in fewer false positives but missing many true positives. In contrast, ChemEXIN strikes a better balance with higher sensitivity (49.6%) and precision (52.6%), allowing it to correctly identify more true boundaries in tRNA while maintaining relatively few false positives. This leads to a higher *F*1 score for ChemEXIN (51.0%) compared to Augustus (28.3%). While differences in EI boundary predictions of tRNA are modest, ChemEXIN outperforms Augustus in other gene categories,

Table 1 Comparison of methods using untrained non-protein coding human genes (lncRNA, tRNA, and rRNA)

Method	Gene category	True positive	False positive	True negative	False negative	Sensitivity (%)	Specificity (%)	*F*1-Score (%)	Precision (%)	Accuracy (%)
ChemEXIN	lncRNA	161	178	169	171	48.5	48.7	48.0	47.5	48.6
	tRNA	122	110	146	124	49.6	57.0	51.0	52.6	53.4
	rRNA	84	113	83	94	47.2	42.3	44.8	42.6	44.7
Augustus	lncRNA	110	330	268	504	17.9	44.8	20.9	25.0	31.2
	tRNA	82	88	292	328	20.0	76.8	28.3	48.2	47.3
	rRNA	88	90	184	414	17.5	67.2	25.9	49.4	35.1

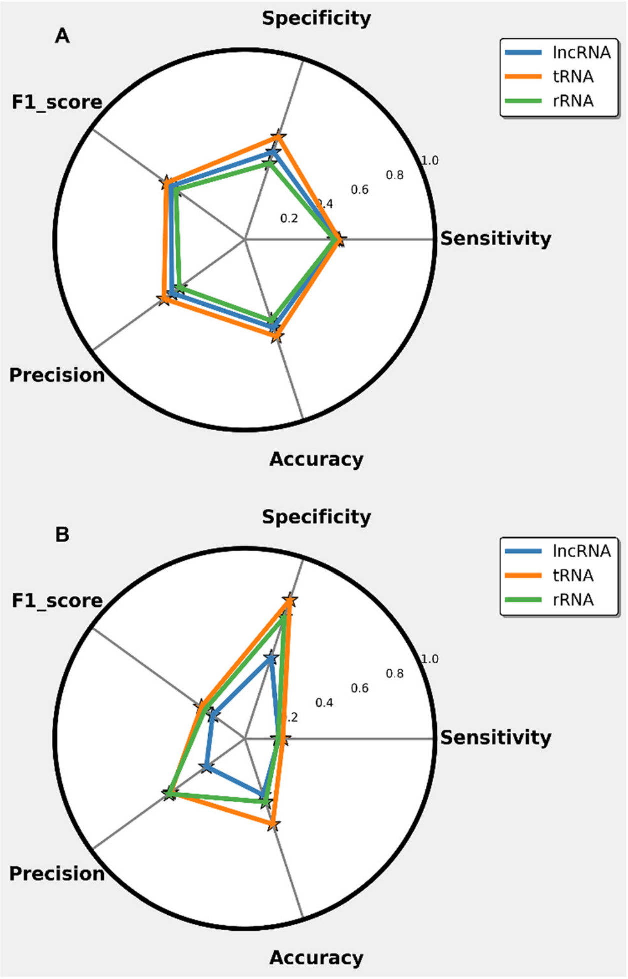

Fig. 8 Performance evaluation of ChemEXIN against Augustus on non-protein coding gene datasets. (A) ChemEXIN, (B) Augustus.

highlighting its broader applicability and robustness in EI boundary prediction indicating its adaptability and efficacy even in contexts beyond its specialized training domain. This establishes the robustness and versatility of our approach, particularly in addressing gene prediction tasks across varied genomic contexts.

Leveraging biophysical parameters and the DL method, the approach exhibited superior performance compared to existing gene annotation tools across the three organisms. Moving forward, we developed ChemEXIN, a consolidated prediction framework combining the three organism-specific pre-trained 3D-CNN models and additional prediction filters (ESI,† File 5). This approach holds significant potential for enhancing the efficiency of EI boundary annotation.

While ChemEXIN demonstrates robust performance, existing tools like Spliceator, Fgenesh, geneid, Genscan, and Augustus have room for improvement in addressing issues such as the over-representation of SS predictions. A multi-faceted approach could address the weaknesses related to the over-representation of predictions in these tools. For example, incorporating biological context through a post-prediction filtering step using experimentally validated datasets could further enhance the precision of SS identification. Another avenue to improve these tools is by increasing their adaptability. Providing organism-specific models as used by Fgenesh, geneid, Genscan, and Augustus would allow Spliceator to obtain relevant features from diverse datasets, making it more adaptable to different organisms and integrating a customization layer to its DL architecture. Furthermore, enhancing user guidance on adapting tools to new sequences—including curating hint files (already used in the case of Augustus) and customizing settings based on preliminary analyses—could improve adaptability and ensure more accurate results from these tools.

EI boundary prediction through ChemEXIN

ChemEXIN, available as an open-source tool, can be downloaded and used within a conda environment, offering an accessible platform for researchers. After the initial setup of the virtual environment through cloning, users can activate and run ChemEXIN using a Python 3 interpreter *via* a command prompt. This process involves providing essential inputs: a file containing the gene sequence of interest, the associated organism, and a threshold value that defines the probability at which prediction windows are refined. Upon receiving these inputs, ChemEXIN performs its analysis and delivers the prediction results in a comma-delimited file. For detailed instructions on setting up and using ChemEXIN, researchers can refer to the user manual (ESI,† File 5).

To assess the performance speed of ChemEXIN, we tested it on random gene sequences of varying lengths from the studied organisms, using a default probability score of 0.75. The specific outcomes of this analysis are cataloged in Table 2. Additionally, to assess ChemEXIN's compatibility across different computing environments, we executed predictions on the same gene set but on systems with various configurations. The results of this compatibility assessment are detailed in Table S6 of ESI,† File 3. Collectively, these results demonstrate that ChemEXIN is highly efficient in processing sequences of diverse lengths, a feat it accomplishes using minimal computational resources and without depending on the operating

Table 2 Speed evaluation of ChemEXIN on random genes in humans and mice

Organism	Gene	Length[a] (nt)	Predicted sites	Average time[b] (s)
H. sapiens	DMD[c]	2 220 382	47	221.77
	BDNF[d]	188 307	28	24.84
	NEU1[e]	10 881	8	8.24
M. musculus	RP1[f]	409 685	26	41.30
	CDK[g]	189 524	9	22.46
	SCAF8[h]	83 888	15	12.78

[a] Nucleotides. [b] Average processing time over three operating systems (Windows 10, Linux Ubuntu 22.04, and macOS 14) in seconds. [c] Dystrophin (muscular dystrophy, Duchenne and Becker types). [d] Brain-derived neurotrophic factor. [e] Neuraminidase-1. [f] Retinitis Pigmentosa-1. [g] Cyclin-dependent-kinase 6. [h] SR-related CTD associated factor-8.

system (OS). A detailed examination of our prediction outcomes, particularly with *H. sapiens* and *M. musculus* gene sequences, reveals that a significant number of EI boundary sites are predicted with remarkable accuracy. This strong performance can be attributed to the similarity in the biophysical profiles at the EI boundaries of humans and mice, where the structural and energetic properties of DNA at these sites closely resemble each other (Fig. 2 and Fig. S3 of ESI,† File 4). The genetic similarity between humans and mice, with approximately 85–90% of their genes being conserved,[64,65] may contribute to the model's ability to generalize effectively across these species. This shared genetic basis could be one of the factors that support the model's strong performance in both organisms. Even in instances where predictions deviate, they do so by a margin of only five to ten nucleotides from the established boundary windows, further supporting the model's ability to generalize across these species. Although the *C. elegans* model showed promising results, its predictions demonstrated relatively lower reliability compared to the other two organisms and were therefore not included in the reported results. This discrepancy is likely due to the distinct biophysical profiles at the EI boundaries in *C. elegans*, which differ from those in humans and mice (Fig. S4 of ESI,† File 4). These differences could have influenced the model's performance. Additionally, challenges such as imbalanced positive and negative datasets may have contributed to this outcome. To address these challenges, we are actively refining the model by improving the filters, extending the training process, and optimizing the model's parameters. Furthermore, we are working to expand ChemEXIN's applicability by incorporating species from different kingdoms (biophysical profiles of some of these organisms are shown in Fig. S5–S7 of ESI,† File 4), enhancing its generalizability across a wider range of eukaryotes. These ongoing improvements are expected to enhance ChemEXIN's performance in future versions.

Conclusion

This study introduces ChemEXIN, a novel tool that integrates biophysical parameters with DL to predict EI boundaries at the DNA level with enhanced accuracy across eukaryotic species. Our analysis of structural and energy profiles at the EI boundaries in multiple organisms revealed distinct physicochemical patterns essential for their recognition. By incorporating these insights, ChemEXIN outperforms existing DNA-based gene prediction tools, demonstrating superior precision in boundary identification.

ChemEXIN's integration with refinement filters further optimizes its usability, offering a user-friendly platform that efficiently processes gene sequences with minimal computational demands. Its open-source nature and adaptability across various genomic contexts position ChemEXIN as a valuable resource for the research community, advancing our understanding of gene architecture and enabling precise EI boundary annotations.

Looking ahead, we recognize the importance of increasing ChemEXIN's scope by training models on a broader range of species while also expanding its capabilities to predict AS sites. To achieve this, we plan to inspect the biophysical profiles of the organisms at the pre-mRNA level and integrate them with DNA profiles in a bottom-up approach. We speculate that this strategy will not only enhance ChemEXIN's predictive accuracy but also facilitate the identification of AS sites, further refining its performance. By broadening its applicability across diverse species and genomic elements, ChemEXIN sets a new benchmark for integrating physicochemical properties in gene prediction tasks, offering significant potential for future applications in molecular biology and genomic research.

Author contributions

The study's design was conceptualized by BJ and DS. Data collection, analysis, and development were conducted by DS, DA, and KS. DS, DA, and BJ analyzed the results and authored the manuscript. DS, DA, and KS developed and hosted the ChemEXIN on GitHub. AM made significant contributions to the conceptual framework of the investigation.

Data availability

The datasets for all the studied organisms can be downloaded from the SCFBio website (**https://www.scfbio-iitd.res.in/ChemEXIN/ChemEXIN_Datasets.tar**). ChemEXIN, available as a Python-based command line utility, is hosted at GitHub (**https://github.com/rnsharma478/ChemEXIN**).

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We express our sincere gratitude to Dr Modesto Orozco and Dr Federica Battistini from the Molecular Modelling and Bioinformatics group at the Institute of Research in Biomedicine, Barcelona, Spain, for sharing the structural MDS data utilized in this research. The authors acknowledge the Centre of Excellence Support to SCFBio, IIT Delhi, from the Department of Biotechnology (DBT), Ministry of Science and Technology, Government of India. DS thanks the Council of Scientific & Industrial Research (CSIR), Government of India, for granting him the senior research fellowship.

References

1 A. Spang, J. H. Saw, S. L. Jørgensen, K. Zaremba-Niedzwiedzka, J. Martijn, A. E. Lind, R. Van Eijk, C. Schleper, L. Guy and T. J. G. Ettema, Complex archaea that bridge the gap between prokaryotes and eukaryotes, *Nature*, 2015, **521**(7551), 173–179.

2 P. A. Sharp, Split genes and RNA splicing, *Cell*, 1994, **77**(6), 805–815.

3 M. Soller, Pre-messenger RNA processing and its regulation: a genomic perspective, *Cell. Mol. Life Sci.*, 2006, **63**, 796–819.

4 A. Anna and G. Monika, Splicing mutations in human genetic disorders: examples, detection, and confirmation, *J. Appl. Genet.*, 2018, **59**, 253–268.

5 C. Mathé, M. F. Sagot, T. Schiex and P. Rouzé, Current methods of gene prediction, their strengths and weaknesses, *Nucleic Acids Res.*, 2002, **30**(19), 4103–4117.

6 J. E. Allen, M. Pertea and S. L. Salzberg, Computational gene prediction using multiple sources of evidence, *Genome Res.*, 2004, **14**(1), 142–148.

7 J. Watson, T. Baker and S. Bell, *et al.*, *Molecular Biology of the Gene*, Cold Spring Harbor Laboratory Press, New York, 7th edn, 2013. ISBN-13: 978-0-321-76243-6.

8 A. Mishra, P. Siwach, P. Misra, S. Dhiman, A. K. Pandey, P. Srivastava and B. Jayaram, Intron exon boundary junctions in human genome have in-built unique structural and energetic signals, *Nucleic Acids Res.*, 2021, **49**(5), 2674–2683.

9 X. Roca and A. R. Krainer, Recognition of atypical 5′ splice sites by shifted base-pairing to U1 snRNA, *Nat. Struct. Mol. Biol.*, 2009, **16**(2), 176–182.

10 X. Roca, R. Sachidanandam and A. R. Krainer, Intrinsic differences between authentic and cryptic 5′ splice sites, *Nucleic Acids Res.*, 2003, **31**(21), 6321–6333.

11 G. E. Parada, R. Munita, C. A. Cerda and K. Gysling, A comprehensive survey of non-canonical splice sites in the human transcriptome, *Nucleic Acids Res.*, 2014, **42**(16), 10564–10578.

12 T. W. Nilsen and B. R. Graveley, Expansion of the eukaryotic proteome by alternative splicing, *Nature*, 2010, **463**(7280), 457–463.

13 P. Senapathy, M. B. Shapiro and N. L. Harris, *[16] Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project*, 1990.

14 S. Brunak, J. Engelbrecht and S. Knudsen, Prediction of human mRNA donor and acceptor sites from the DNA sequence, *J. Mol. Biol.*, 1991, **220**(1), 49–65.

15 G. Yeo and C. B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals, *Proceedings of the seventh annual international conference on Research in computational molecular biology*, 2003, pp. 322–331.

16 K. Sahashi, A. Masuda, T. Matsuura, J. Shinmi, Z. Zhang, Y. Takeshima and K. Ohno, In vitro and in silico analysis reveals an Efficient algorithm to predict the splicing consequences of mutations at the 5′ splice sites, *Nucleic Acids Res.*, 2007, **35**(18), 5995–6003.

17 R. Ramakrishna and R. Srinivasan, Gene identification in bacterial and organellar genomes using GeneScan, *Comput. Chem.*, 1999, **23**(2), 165–174.

18 R. F. Yeh, L. P. Lim and C. B. Burge, Computational inference of homologous gene structures in the human genome, *Genome Res.*, 2001, **11**(5), 803–816.

19 E. Birney, M. Clamp and R. Durbin, GeneWise and genomewise, *Genome Res.*, 2004, **14**(5), 988–995.

20 M. Stanke and B. Morgenstern, AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints, *Nucleic Acids Res.*, 2005, **33**(suppl_2), W465–W467.

21 V. Solovyev, P. Kosarev, I. Seledsov and D. Vorobyev, Automatic annotation of eukaryotic genes, pseudogenes and promoters, *Genome Biol.*, 2006, **7**, 1–12.

22 E. E. Snyder and G. D. Stormo, Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks, *Nucleic Acids Res.*, 1993, **21**(3), 607–613.

23 E. Blanco, G. Parra and R. Guigó, Using geneid to identify genes, *Curr. Protoc. Bioinf.*, 2007, DOI: **10.1002/0471250953.bi0403s18**.

24 N. Scalzitti, A. Kress, R. Orhand, T. Weber, L. Moulinier, A. Jeannin-Girardon and J. D. Thompson, Spliceator: multi-species splice site prediction using convolutional neural networks, *BMC Bioinf.*, 2021, **22**, 1–26.

25 G. F. Ejigu and J. Jung, Review on the computational genome annotation of sequences obtained by next-generation sequencing, *Biology*, 2020, **9**(9), 295.

26 Z. Chen, N. U. Ain, Q. Zhao and X. Zhang, From tradition to innovation: conventional and deep learning frameworks in genome annotation, *Briefings Bioinf.*, 2024, **25**(3), bbae138.

27 C. Trapnell, L. Pachter and S. L. Salzberg, TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics*, 2009, **25**(9), 1105–1111.

28 K. F. Au, H. Jiang, L. Lin, Y. Xing and W. H. Wong, Detection of splice junctions from paired-end RNA-seq data by SpliceMap, *Nucleic Acids Res.*, 2010, **38**(14), 4570–4578.

29 K. Wang, D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich and J. Liu, MapSplice: accurate mapping of RNA-seq reads for splice junction discovery, *Nucleic Acids Res.*, 2010, **38**(18), e178–e178.

30 A. Ameur, A. Wetterbom, L. Feuk and U. Gyllensten, Global and unbiased detection of splice junctions from RNA-seq data, *Genome Biol.*, 2010, **11**, 1–9.

31 L. Levin, D. Bar-Yaacov, A. Bouskila, M. Chorev, L. Carmel and D. Mishmar, LEMONS–a tool for the identification of splice junctions in transcriptomes of organisms lacking reference genomes, *PLoS One*, 2015, **10**(11), e0143329.

32 K. Jaganathan, S. K. Panagiotopoulou, J. F. McRae, S. F. Darbandi, D. Knowles, Y. I. Li and K. K. H. Farh, Predicting splicing from primary sequence with deep learning, *Cell*, 2019, **176**(3), 535–548.

33 C. Xu, S. Bao, Y. Wang, W. Li, H. Chen, Y. Shen and C. Zhang, Reference-informed prediction of alternative splicing and splicing-altering mutations from sequences, *Genome Res.*, 2024, **34**(7), 1052–1065.

34 J. A. Fincher, G. S. Tyson and J. H. Dennis, DNA-Encoded Chromatin Structural Intron Boundary Signals Identify Conserved Genes with Common Function, *Int. J. Genomics*, 2015, **2015**(1), 167578.

35 F. Geraci, I. Saha and M. Bianchini, RNA-Seq analysis: methods, applications and challenges, *Front. Genet.*, 2020, **11**, 220.

36 H. Satam, K. Joshi, U. Mangrolia, S. Waghoo, G. Zaidi, S. Rawool and S. K. Malonia, Next-generation sequencing technology: current trends and advancements, *Biology*, 2023, **12**(7), 997.

37 R. E. Dickerson and H. R. Drew, Structure of a B-DNA dodecamer: II. Influence of base sequence on helix structure, *J. Mol. Biol.*, 1981, **149**(4), 761–786.

38 K. Yanagi, G. G. Privé and R. E. Dickerson, Analysis of local helix geometry in three B-DNA decamers and eight dodecamers, *J. Mol. Biol.*, 1991, **217**(1), 201–214.

39 M. A. El Hassan and C. R. Calladine, The assessment of the geometry of dinucleotide steps in double-helical DNA; a new local calculation scheme, *J. Mol. Biol.*, 1995, **251**(5), 648–664.

40 W. K. Olson, A. A. Gorin, X. J. Lu, L. M. Hock and V. B. Zhurkin, DNA sequence-dependent deformability deduced from protein–DNA crystal complexes, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**(19), 11163–11168.

41 D. L. Beveridge, G. Barreiro, K. S. Byun, D. A. Case, T. E. Cheatham, S. B. Dixit and M. A. Young, Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps, *Biophys. J.*, 2004, **87**(6), 3799–3813.

42 S. B. Dixit, D. L. Beveridge, D. A. Case, T. E. Cheatham, E. Giudice, F. Lankas and P. Varnai, Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical structures of the 10 unique dinucleotide steps, *Biophys. J.*, 2005, **89**(6), 3721–3740.

43 R. Lavery, M. J. H. P. D. Moakher, J. H. Maddocks, D. Petkeviciute and K. Zakrzewska, Conformational analysis of nucleic acids revisited: curves+, *Nucleic Acids Res.*, 2009, **37**(17), 5917–5929.

44 R. Lavery, K. Zakrzewska, D. Beveridge, T. C. Bishop, D. A. Case, T. Cheatham III and J. Sponer, A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA, *Nucleic Acids Res.*, 2010, **38**(1), 299–313.

45 M. Pasi, J. H. Maddocks, D. Beveridge, T. C. Bishop, D. A. Case, T. Cheatham III and R. Lavery, μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA, *Nucleic Acids Res.*, 2014, **42**(19), 12272–12283.

46 R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann and B. Honig, The role of DNA shape in protein–DNA recognition, *Nature*, 2009, **461**(7268), 1248–1253.

47 K. Florquin, Y. Saeys, S. Degroeve, P. Rouze and Y. Van de Peer, Large-scale structural analysis of the core promoter in mammalian and plant genomes, *Nucleic Acids Res.*, 2005, **33**(13), 4255–4264.

48 M. M. Gromiha, J. G. Siebers, S. Selvaraj, H. Kono and A. Sarai, Intermolecular and intramolecular readout mechanisms in protein–DNA recognition, *J. Mol. Biol.*, 2004, **337**(2), 285–294.

49 S. Dutta, P. Singhal, P. Agrawal, R. Tomer, K. Kritee and B. Jayaram, A physicochemical model for analyzing DNA sequences, *J. Chem. Inf. Model.*, 2006, **46**(1), 78–85.

50 P. Singhal, B. Jayaram, S. B. Dixit and D. L. Beveridge, Prokaryotic gene finding based on physicochemical characteristics of codons calculated from molecular dynamics simulations, *Biophys. J.*, 2008, **94**(11), 4173–4183.

51 G. Khandelwal and J. Bhyravabhotla, A phenomenological model for predicting melting temperatures of DNA sequences, *PLoS One*, 2010, **5**(8), e12433.

52 G. Khandelwal and B. Jayaram, DNA–water interactions distinguish messenger RNA genes from transfer RNA genes, *J. Am. Chem. Soc.*, 2012, **134**(21), 8814–8816.

53 G. Khandelwal, J. Gupta and B. Jayaram, DNA-energetics-based analyses suggest additional genes in prokaryotes, *J. Biosci.*, 2012, **37**, 433–444.

54 G. Khandelwal, R. A. Lee, B. Jayaram and D. L. Beveridge, A statistical thermodynamic model for investigating the stability of DNA sequences from oligonucleotides to genomes, *Biophys. J.*, 2014, **106**(11), 2465–2473.

55 A. Singh, A. Mishra, A. Khosravi, G. Khandelwal and B. Jayaram, Physico-chemical fingerprinting of RNA genes, *Nucleic Acids Res.*, 2017, **45**(7), e47.

56 A. Mishra, P. Siwach, P. Misra, B. Jayaram, M. Bansal, W. K. Olson and D. L. Beveridge, Toward a universal structural and energetic model for prokaryotic promoters, *Biophys. J.*, 2018, **115**(7), 1180–1189.

57 A. Mishra, S. Dhanda, P. Siwach, S. Aggarwal and B. Jayaram, A novel method SEProm for prokaryotic promoter prediction based on DNA structure and energetics, *Bioinformatics*, 2020, **36**(8), 2375–2384.

58 D. Sharma, K. Sharma, A. Mishra, P. Siwach, A. Mittal and B. Jayaram, Molecular dynamics simulation-based trinucleotide and tetranucleotide level structural and energy characterization of the functional units of genomic DNA, *Phys. Chem. Chem. Phys.*, 2023, **25**(10), 7323–7337.

59 M. N. Nedelcheva-Veleva, M. Sarov, I. Yanakiev, E. Mihailovska, M. P. Ivanov, G. C. Panova and S. S. Stoynov, The thermodynamic patterns of eukaryotic genes suggest a mechanism for intron–exon recognition, *Nat. Commun.*, 2013, **4**(1), 2101.

60 A. Frankish, M. Diekhans, I. Jungreis, J. Lagarde, J. E. Loveland, J. M. Mudge and P. Flicek, GENCODE 2021, *Nucleic Acids Res.*, 2021, **49**(D1), D916–D923.

61 P. D. Dans, A. Balaceanu, M. Pasi, A. S. Patelli, D. Petkevičiūtė, J. Walther and M. Orozco, The static and dynamic structural heterogeneities of B-DNA: extending Calladine–Dickerson rules, *Nucleic Acids Res.*, 2019, **47**(21), 11090–11102.

62 R. I. Kraeva, D. B. Krastev, A. Roguev, A. Ivanova, M. N. Nedelcheva-Veleva and S. S. Stoynov, Stability of mRNA/DNA and DNA/DNA duplexes affects mRNA transcription, *PLoS One*, 2007, **2**(3), e290.

63 S. Ji, W. Xu, M. Yang and K. Yu, 3D convolutional neural networks for human action recognition, *IEEE Trans. Pattern Anal. Machine Intell.*, 2012, **35**(1), 221–231.

64 R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril and P. Agarwal, Initial sequencing and comparative analysis of the mouse genome: Mouse Genome Sequencing Consortium, *Nature*, 2002, 520–562.

65 K. Lindblad-Toh, M. Garber, O. Zuk, M. F. Lin, B. J. Parker, S. Washietl and M. Kellis, A high-resolution map of human evolutionary constraint using 29 mammals, *Nature*, 2011, **478**(7370), 476–482.