https://doi.org/10.48048/tis.2025.9149

Advancing Marine Genomics: The Role of Deep Learning in Deciphering Chelonia mydas Genetic Data

Fahad Aslam and Faizah Aplop*

Institute of Oceanography and Environment (INOS), Universiti Malaysia Terengganu, Kuala Nerus, Terengganu 21030, Malaysia

(*Corresponding author's e-mail: faizah_aplop@umt.edu.my)

Received: 15 October 2024, Revised: 22 November 2024, Accepted: 29 November 2024, Published: 10 January 2025

Abstract

The urgent advancement of marine genomics is essential for the conservation of endangered species like *Chelonia mydas* (green sea turtle), and deep learning plays a pivotal role in deciphering their complex genetic data. Marine genomics as a field seems to be shifting more and more to the realm of big data especially with the introduction of new technologies in producing vast amounts of data. Such advancements have made it possible to manage huge datasets within genomics and this has provided artificial intelligence and particularly deep learning as a crucial method of acquiring insightful patterns. Review of the subject aims at focusing on the subject of deep learning methods and their usefulness in appearance and utilization in sub disciplinary areas of genomics of Chelonia mydas (green sea turtle). We introduce deep learning into marine genomics research by pointing out the existing gaps, as well as well-PSYCH detailed study fields. Moreover, we can only briefly mention the rather late incorporation of deep learning tools into marine genomics and the eminently discussed consequences for conservation and ecological science. By writing this review the authors envisage to let the biotechnology and genomic scientists to know the importance and applicability of using deep learning methods in Chelonia mydas genomics, the difficulty and prospects of this field.

Keywords: Chelonia mydas, Deep learning methods, Marine genomics, Variant calling, Genomic research

Introduction

Research on genomes has yielded a better understanding of the genomics of varying species, especially the Chelonia mydas whose genome is roughly coded in billions of base pairs. Whole organism genomics is the study of all the genes in a given organism. These include peptide/tRNA genes, RNA genes as well as other regulatory species such as the cis-and-trans factors. This field has been greatly impacted by large-scale technologies specifically by next-generation sequencing (NGS) which encompasses the sequencing of an organism's entire DNA. The most critical techniques are WGS, WES, transcriptomics, and proteomics [1]. The recent exponential growth of these 'omics' data has created more discussions on the roles of bioinformatics and machine learning (ML) tools in different applications for example identification of genotypes–phenotypes, biomarkers [2], functional prediction of genes as well as prediction for regulatory regions such as transcriptional enhancers for Chelonia mydas. Artificial intelligence includes one of its key branches, namely, machine learning, by means of which, based on the data received, important decision-making can be made through the use of algorithms rather than being coded in the program.

Nevertheless, by now ML is used in a rather broad spectrum of applications, and the traditional statistical methods are insufficient to deal with large complex and high-dimensional genomic data. However, the advancement in the sub-discipline of ML known as Deep Learning (DL) that employs the neural networks (NN) takes the genomic research to another level [3]. DL applications comprise image identification, audio categorization, voice identification, natural language identification, and others [4]. Starting with genomics,

which strategies are particularly tailored for the analysis of extensive datasets, such as those for Chelonia mydas, DL is a prospective methodology. Even though it is relatively recent, evidence of its usage in genomics has a reactionary characteristic of revitalizing particular fields like conservation genetics and functional genomics.

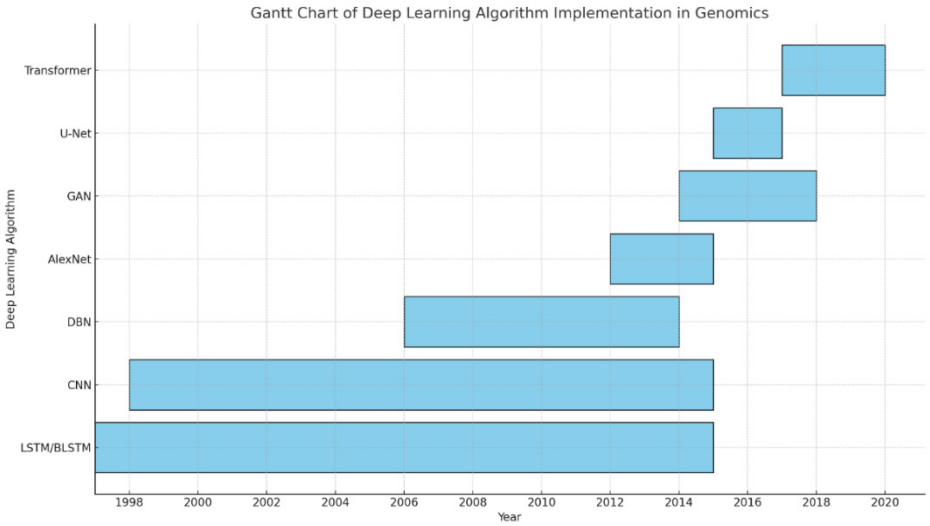

Figure 1 Chronological list of the integration of the DL algorithms in genomics. This Gantt chart presents the idea of the proposed work and illustrates the time when deep learning tools have not been applied to genomics yet. For example, LSTM and BLSTM were developed in 1997 but was applied to genomics for the first time in 2015. Peculiarities of the same kind are observed for other algorithms: CNN (1998 - 2015), DBN (2006 - 2014), AlexNet (2012 - 2015), GAN (2014 - 2018), U-Net (2015 - 2017), Transformer (2017 - 2020) (**Table 7**).

DL algorithms dominate the list of the most popular algorithms in computational modelling, and their application to various genomic challenges expands constantly. They are used for explaining how mutations influence protein-RNA interactions, for prioritization of variants and genes, for predicting gene expression levels based on histone modifications, and for identification of trait-associated SNPs [3]. Deep learning, which can be traced back to the 1980s' early neural network models, has in the last ten years evolved into sophisticated methods for Big Data prediction. Doubtless, it was the relative complexity of DL proper and obstacles associated with its incorporation into genomic prediction models that initially limited the practice of this technique to the 2000s. The breakthrough that has put DL on the development map has come with the modern hardware, particularly the high-efficiency GPUs. Now, the DL models or the NN are used in different fields. Typically, classical neural networks used to have two to 3 hidden layers, whereas the DL and popular neural networks today can contain more than 100 to 200 layers. This “deep” element relates to the number of signal processing levels that take place in the course of communication [3,4].

DL’s deployment entails significantly superior hardware and a massive degree of parallelism because of its high computational need. This has made it possible to have different DL packages and resources for the support of such models. Recent achievements in the fields of sociotechnical software and GPU hardware as well as big data allow the modelling of functional genomic elements with the help of deep learning. Some of the uses of genomic DNA splicing include classification of TFBSs, the ability to estimate the potential effects of variants, and the calculation of the effects of sensitivity or resistance to environmental factors [5]. As an illustration, there is the case of cloud platforms that host GPU resources as a DL service, which accelerates training processes. They include Amazon Web Services, Google Cloud AI Platform, and IBM Cloud, all of which provide massive computational power, although users are still required to implement the model codes themselves.

Another important component in every model is

evaluation metrics as they are frequently important when assessing the performance of an ML model in combination with genomics datasets that may generate highly imbalanced classes. The challenges are often solved utilizing approaches including transfer learning and Matthews's correlation coefficient. Hence, in a way, all the ML tasks for the most part can be differentiated into 2 broad categories of regression and classification problems, and both of these contain their own measures of performance [35]**.** For regression models, common evaluation measures are Average Absolute Error, Average Squared Error, Root of Quadratic Mean Error, the Coefficient of determination. Metrics that are often used with classification models are accuracy, precision, recall, quadratic mean error, confusion matrix, AUC, AUROC, F-scores, and so on [4].

These classification tasks are also used in genomics where models' performances are compared during their classification tasks. For example, AUC is an ideal measure of the model's performance and ranges from 0 to 1. it calculates TPR or sensitivity, TNR or specificity, and FPR. Additionally, to assess model accuracy on imbalanced datasets, the F1-score is computed. This score ranges from 0 to 1 and represents the balanced average between precision and recall rates. A model with a higher AUC or F1-score performs better than the others [5,6]. There are a number of deep learning tools and techniques described here in the light of Chelonia mydas genomics study. We examine recent (15-20) DL tools across 5 main genomics areas: There is the Sanger sequencing method, Next Generation Sequencing, Disease variant, Gene expression and regulation, Epigenetics, and pharmacogenetics. We also state the research state-of-the-art of DL-based algorithms and elaborate on how these methodologies and data models could be applied. The last part is intended to present the practical implications emerged from DL that can be useful in the use of deep learning concerning the research of marine biology and genomics of Chelonia mydas. Thus, for more detail about the application of DL in genomics, readers can refer to other sources.

Traditional approaches and their transition to deep learning in genomic studies

Currently, the genomics field is undergoing an extraordinary transition from conventional techniques to deep learning methods. This transition is due to increased complexity of biological data, requirement for higher accuracy and tremendous technological advancements in computing [1,2]. Perhaps, few species exemplify this change as vividly as in the green sea turtle Chelonia mydas: shedding light on marine genetics, boosting construction of conservation couplings, and aiding in disease treatments [3].

Strengths and limitations of traditional genomic approaches overview of traditional methods

Old methods of genomics include Sanger Sequencing, PCR (Polymerase Chain Reaction), and microarrays on which the contemporary genomics has been built. These methods were useful in early genetics and offered very high specificity if the desired sequences were to be investigated. Sanger Sequencing was good at generating high quality sequence data in small regions of the genome [1]. Microarrays also helped in determination of gene expression and single nucleotide polymorphisms, have laid the foundation for modern genomic studies. These methods were pivotal in early genetic discoveries and provided high specificity for targeted analyses.

Limitations

Despite their significant contributions, these traditional methods have notable limitations.

Scalability

Scalability: These organisms cannot analyze large genomes for instance the one of the green sea turtles Chelonia mydas owing to their low capacity and throughput [2].

Cost and time

Cost and time: The traditional methods entail high operation costs and long experimental procedures that lead to pricey large-scale investigations [3].

Data complexity

Data Complexity: They are not designed to solve the large-scale problem of analyzing the high dimensional datasets that are intrinsic to problems in next-generation sequencing (NGS) technologies, thus are not suitable for many current genomics [2].

Precision

Precision: These approaches are insensitive to small differences in the DNA sequence, for example, SNPs or SCs important for genotypic variation, polygenic traits, and diseases [4].

They thus call for enhanced methods of computational analysis, meaning that computation with an immense series of complex matrices is necessary, leading to the development of deep learning in genomics.

Catalysts for the transition to deep learning

The transition from traditional approaches to deep learning in genomics has been driven by several factors:

Explosive growth of genomic data

Several NGS technologies have been developed that have produced large datasets, meaning there is a need for large scale analysis solutions [5].

Complex biological questions

Problems such as learning gene regulation, pinpointing disease mug factors, and anticipating drug effects require sophisticated models [6].

Advances in computational Infrastructure

The AI advances like high-performance GPUs, hardware and software platform, and big data frameworks have made DL easy to implement [7,8].

Integration with multi-omics data

At the same time, deep learning models can link genomic, transcriptomic, epigenomic, and proteomic data, which are believed to input integrative analyses to the context given the complexity of biological systems [9].

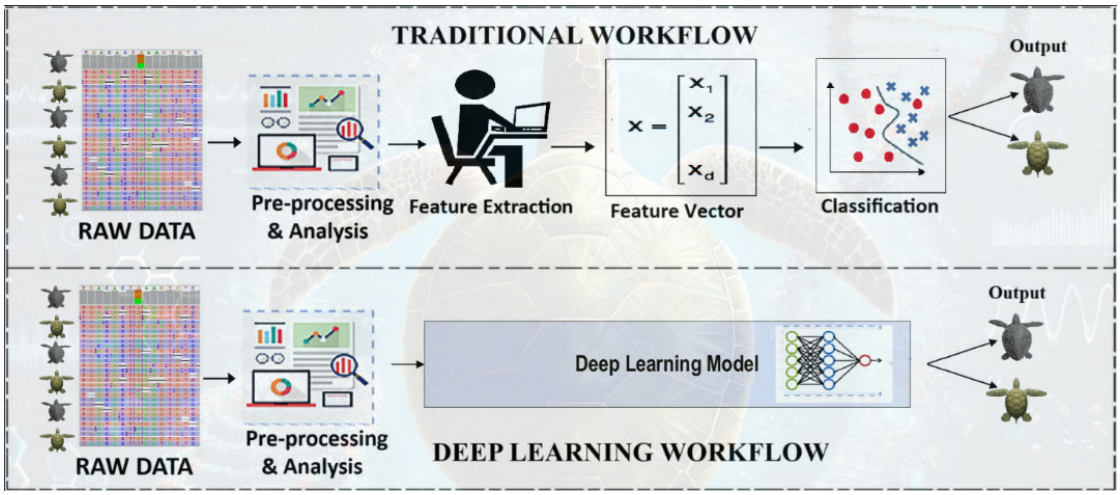

Figure 2 The image provides a detailed visual comparison between 2 genomic analysis approaches: Traditional Workflow and Deep Learning Workflow. It highlights the evolutionary transition in genomics, driven by advancements in computational power, data complexity, and the need for scalable, precise solutions. This transition is exemplified through the genomic study of Chelonia mydas (green sea turtles), a species requiring modern tools for conservation and disease research. The workflows illustrate how traditional methods focus on manual processes and limited scalability, whereas deep learning introduces automation, efficiency, and robust analysis for large-scale genomic datasets.

Comparative analysis of traditional and deep learning approaches

Genomic analysis, as a field of research, has recently transitioned from conventional analyses to deep learning (DL) techniques. This transition has improved its efficiency, accuracy and scalability of genomics applications in different settings. This section briefly outlines the evolution of DL models to transform major regions in genomics as follows:

Table 1 Comparative applications of traditional and deep learning models in genomics.

Application	Traditional methods	Deep learning models	Key advantages of DL models	Reference
Sequencing	Sanger Sequencing, PCR	DeepNano, CNNs, RNNs for assembly	High accuracy and speed in handling large datasets	[10,11]
Variant Calling	GATK, SAMtools	DeepVariant, Clairvoyante	Improved sensitivity and specificity for novel variants	[3,12]
Disease Variant Prediction	GWAS, Linkage Analysis	DeepPVP, ExPecto	Precise genotype-phenotype predictions	[13,9]
Gene Expression Analysis	qPCR, Northern Blotting, ChIP-seq	DeepChrome, Xpresso	Prediction of regulatory networks and transcriptomic changes	[14]

DL in genomics tools/software/pipelines within Chelonia mydas organisms

Genomic applications such as gene expression and regulation, epigenomics, variant calling and annotation, disease variant prediction, pharmacogenomics, functional genomics, conservation genomics, comparative genomics, and metagenomics utilize high-throughput data generation and deep learning methodologies for computational predictions, as illustrated in **Figure 2** [16,17]. Hence, with the growth of new DNA/RNA sequencing technologies and the establishment of the requirement for artificial intelligence, particularly deep learning in considering the massive biological data in genomics of this turtle, a new chapter in the knowledge in all the subfields of Chelonia mydas genomics is opened. Newly designed software platforms, tools, and frameworks based on deep learning algorithms will be discussed in the next sections focusing on the various conditions of Chelonia mydas genomics [18].

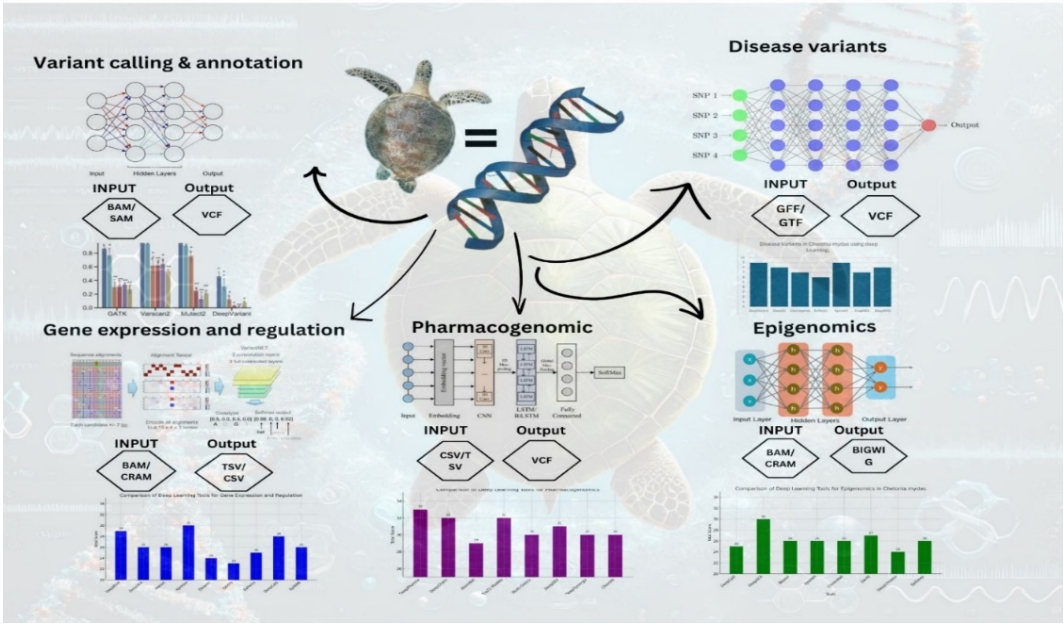

Figure 3 Application of DL in Chelonia mydas genomics. This illustration demonstrates the application of advanced DL techniques in 5 significant sectors of Chelonia mydas genomics: HCN, free text retrieval and pattern matching, Disease Variant, Gene Expression & Regulation, Pharmacogenomics, and Epigenomics. In each of the subareas, there is a deep learning tool described together with the type of input data, network architecture, and the output that is being predicted. The bar plots at the bottom of each subarea show the usage frequency of the most applied DL techniques within the given subarea of Chelonia mydas genomics (**Tables 2** to **6**).

Variant calling and annotation in Chelonia mydas

Here, describing the DL methods learned from the existing research in variant calling and annotation of Chelonia mydas. There are lists of some tools for variant calling and annotation as shown in **Table 2** to make it easy to choose the right DL tool based on the kind of data that is being worked on. Whole genome and exome sequencing, or otherwise known as next-generation sequencing (NGS), thus serves as a basis for early endeavors in wildlife genomics, especially in the context of conservation and learning more about the Chelonia mydas population's genetic variation.

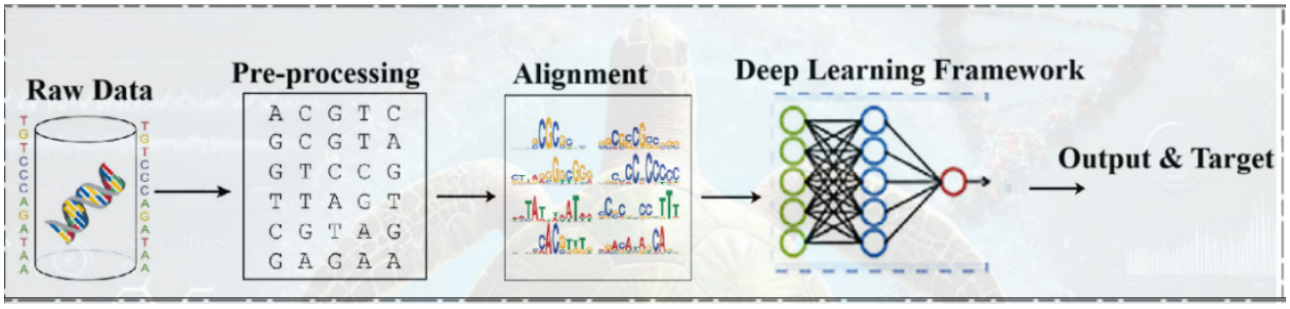

Figure 4 The image illustrates the pipeline for variant calling and annotation in Chelonia mydas, integrating deep learning methods. It begins with raw sequencing data, undergoes preprocessing to extract genomic sequences, followed by alignment to reference genomes. The aligned data is then processed through a deep learning framework for variant identification, producing outputs like VCF files for genetic analysis.

This is because over a short while there has been the improvement of technologies in what is referred to as high-throughput massively parallel sequencing technologies applicable for the assessment of inter-genetic variation within this species. Well-developed equipment in the field of bioinformatics and statistics is crucial for variation analysis; however, there are several issues inherent to high technical and bioinformatics error rates. Most computational issues arise from the continuously increasing volume of genome sequences with medium or low coverage, genetic diversity, and short-read fragments of Chelonia mydas strains. These can make the NGS data vulnerable to errors that call for the need to put in place strong bioinformatics tools to process the data [19,20].

Table 2 List of DL-Based Genomic Tools for variant calling and annotations.

Tools	Description	Key features	Applicable data types	Source code link	Reference
DeepVariant	A deep learning-based tool for variant calling, particularly effective when integrated with other tools like SAM tools and GATK.	High accuracy in identifying SNVs and indels; works with NGS data.	Whole genome sequencing, exome sequencing BAM, CRAM/VCF	https://github.com/google/deepvariant	[21]
DeepSV	A tool for identifying large deletions from sequencing reads using deep learning.	Estimates read intensities associated with deletions >50bp; outputs in VCF format.	BAM/VCF files	https://github.com/CSuperlei/DeepSV	[22]
GARFIELD-NGS	Genomic variant filtering tool using MLP algorithm for exome sequencing datasets.	Analyzes true and false positives; handles low coverage data; outputs VCF files.	Ion Torrent, Illumina VCF/VCF	https://github.com/gedoardo83/GARFIELD-NGS	[11]
Clairvoyante	Workflow for predicting	Works with long-read and	Long-read sequencing	https://github.com/aqu	[12]

Tools	Description	Key features	Applicable data types	Source code link	Reference
	variant types, zygosity, allele alternative, and indel length.	short-read sequencing data.	(PacBio, ONT), short-read sequencing BAM/VCF	askyline/Clairvoyante	
Intelli-NGS	A neural network-based tool for variant calling and annotation.	Processes VCF files in batch mode; provides HGVS codes for all variants.	Ion Torrent VCF/xlsx	https://github.com/aditya-88/intelli-ngs	[15]

For example, the commonly cited variant caller software applied in genomic variant assessment comprises of GATK, SAMtools, Freebayes, and TVC [20]. Thus, there are some variants that are not reflected when using whole genome sequencing techniques that are common today. Some current works have pointed out that DeepVariant, a deep learning-based variant caller, is one of the most accurate when integrated with other software such as SAMtools and GATK [21]. DeepVariant, despite being compared to GATK and Strelka2, has a way lower error rate of variant calling and is further enhanced when utilized with DeepVariant-AF for allele frequency data [20]. Complementing deep learning with traditional techniques enhances the performance and facilitates more accurate variant identification. The use of deep learning in relation to genome sequencing is in its early stages, with DeepVariant by Google being the latest one.

DeepVariant uses the differences of the input images to call genetic variants from the NGS short reads. This way, the model turns sequenced datasets into the form of images to enable conversion of variant calls into image-based classification. This model covers all the aspects; however, it does not give a detailed account of variant information more complex than the allelic alternative or variant type which situates the model in the category of incomplete variant caller models. Also, in the same study, the same scientists proposed a new tool known as DeepSV that will help determine large deletions of more than 50 base pairs. It provides greater accuracy and requires less training loss. Therefore, DeepSV can be employed to simulate the changes in evolutive processes to build the haplotype graphs from BAM / VCF files and the final product in terms of files are VCF files [22].

When mentoring the comparison of DeepSV with another deletion calling tool known as Concod which is a machine intelligence-based tool, these results depict that DeepSV has increased accuracy rates but reduced training loss if the DeepSV is trained using a limited dataset. However, as it can be noticed, Concod's sensitivity drops in the situations in which the number of training iterations is small or a less number of samples are given [22]. Four-year wall Improved to Trimmed to keep sense the same as before but 'the fit main idea' sounds neater than 'the same idea' Improved word usage and rhythm. If the above modifications are not sufficient or if you have further instructions or are desirous of any special changes, kindly intimate.

As a result, GARFIELD-NGS utilizes DNNs and MLPs to engineer high sensitivity and specificity that filters the real and false variants across the exome sequencing data and that is specifically designed for handling low-coverage data up to 30X. It produces VCF files that are easy to sort to a degree that makes it highly relevant in genetics research. In their study, Zeng et al pointed out that when applying this model to disease-related data and following the variant prioritization process introduced in this study, false positive rates were drastically diminished [11]

To deal with these challenges, the Clairvoyante workflow featuring prediction of variant types (SNV or Indel), and indel length was designed. This model addresses the issues in DeepVariant that provides only the complete variant along with other required information about the variant such as the type of the allele and the type of the variant. Clairvoyante was developed to function with long-reads Sequencing information including those generated by SMS technologies (PacBio and ONT) but is compatible with short-reads too [12].

Intelli-NGS: A tool based on the artificial neural network that operates with VCF files outputted from Ion Torrent sequencer data. It can perform batch analysis and offers HGVS codes for the reported variations and the results are generated in Excel format [15]. This is the original ANN tool, which, using data from the Ion Torrent platform, copes with the identification of true

and false variants All these studies demonstrate the potential of deep learning in calling and annotating genetic variants from Chelonia mydas sequences. **Figure 4** emphasizes the overall process, from raw sequencing data to output, showcasing how DL tools like DeepVariant, DeepSV, and Clairvoyante enable accurate genetic analysis. Such advancements are crucial for further investigations into the genome of this species [18].

Disease variants

In the context of Chelonia mydas research, the reference models for disease-causing mutations, their applications, input and output data structures, and source codes are delineated as follows: Geneticists may also do further cohort comparisons that from the observed variants the variants are called and annotated. The selection of pathogenic variants is a standard approach in outlining variants that play pivotal roles in the alteration of the genes' functionality and ensuing disease attributes in Chelonia mydas.

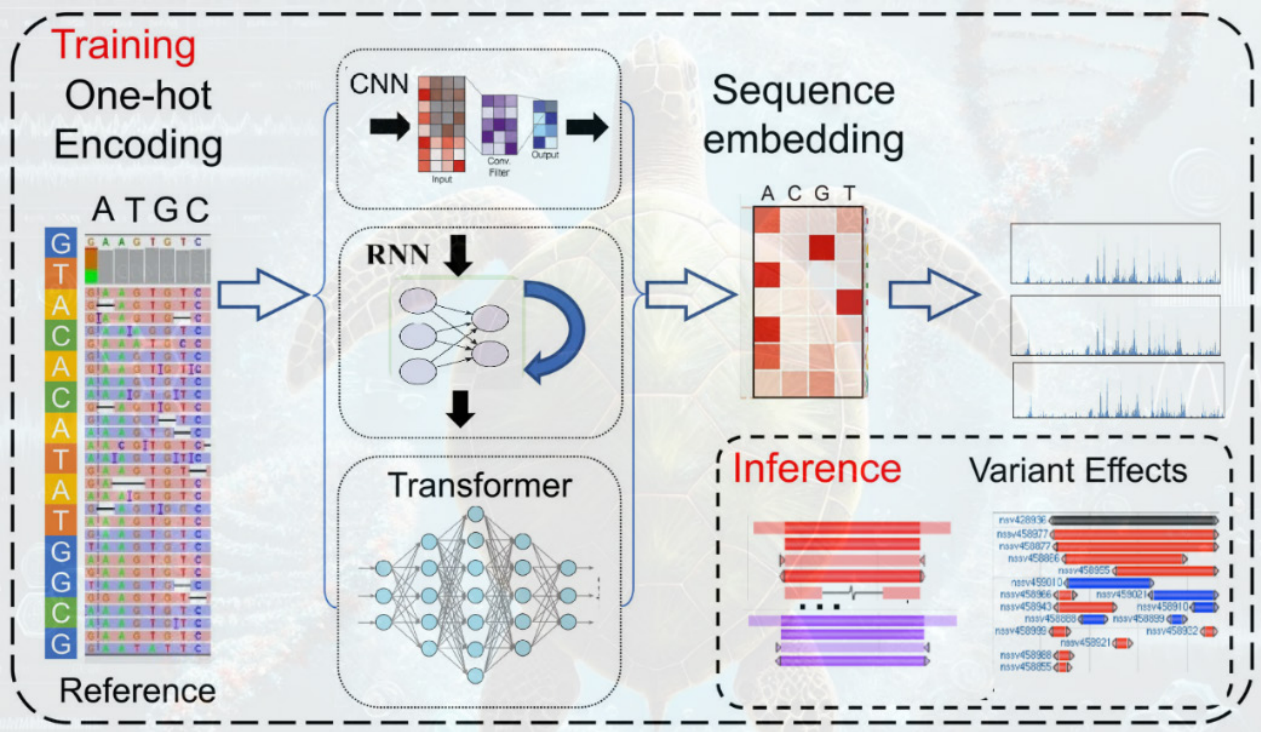

Figure 5 illustrates a deep learning-based workflow for identifying disease-related genetic variants. The process begins with one-hot encoding of genomic sequences, followed by feature extraction using CNNs, RNNs, and Transformers. These features are transformed into sequence embeddings, leading to inference of variant effects, highlighting pathogenic and regulatory roles.

Other variants to deprioritize are those that are unlikely to be pathogenic and they include synonymous, deep intronic, and benign polymorphic variants. The attainable variants comprise known variants and variants with uncertain significance, also known as VUS. The direction for assessing the results of diagnostic sequencing and potential mistakes in analyzing dystonic Chelonia mydas genetics and its relation to disease susceptibility also encountered possible inaccuracies in the analytical interpretation of rare variations [17]. For example, general and rare VUSs can cause difficulties in the interpretation of genetic analysis of Chelonia mydas, thus making it challenging to apply sequencing for the health check of wild herds of turtles. Although extensive genetic epidemiological methods like GWAS have identified genetic traits linked to disorders, they are exhaustive in sampling and only reveal information on new variants. Thus, methods such as GERP hold the ability to rank causative variants, though they entail certain demerits. Thus, DL-based

models have been employed for extracting better information regarding the priority of variants, out of DNN structures [9]. **Figure 5** illustrates the overall computational pipeline for predicting variant effects, utilizing CNNs, RNNs, and Transformer-based architectures for prioritizing pathogenic and regulatory mutations.

For instance, the model that belongs to the variant annotator class is the Basset model that mainly utilizes CNN algorithms to identify initiatory causative SNPs utilizing DNase I hypersensitivity sequencing information. The source code for Basset is available on GitHub and further details can be found in the paper link [24].

Table 3 List of DL-based genomic tools for identifying disease variants.

Tools	Description	Key features	Applicable data types	Source code link	Reference
DeepPVP (PhenomeNet Variant Predictor)	Predicts the pathogenicity of genetic variants based on phenotype-genotype associations	Integrates phenotypic and genotypic data, provides pathogenicity scores	Genetic variant data	https://github.com/bio-ontology-research-group/phenomenet-vp	[25]
ExPecto	Predicts the regulatory effects of genetic variants in non-coding regions	Uses CNN for non-coding regions, predicts regulatory effects	Genetic variant data	https://github.com/FunctionLab/ExPecto	[26]
PEDIA	Uses clinical images and exome data to prioritize genetic variants	Combines clinical images with exome data, prioritizes variants	Exome data, clinical images	https://github.com/PEDIA-Charite/PEDIA-workflow	[13]
DeepMILO	Models insulator loops to predict regulatory variants	Predicts insulator loops, models chromatin interactions	Genomic data	https://github.com/khuranalab/DeepMILO	[14]
DeepWAS	Provides the integrated regulatory effect of each variant regarding different cell-specific chromatin characteristics	Uses CNN to integrate chromatin data, identifies disease-associated SNPs	Genomic data	https://github.com/cellmapslab/DeepWAS	[27]
PrimateAI	Predicts the pathogenicity of genetic variants in primates	Leverages primate genomic data, provides pathogenicity scores	Genetic variant data	https://github.com/Illumina/PrimateAI	[23]
DeepGestalt	Uses facial images to prioritize genetic disorders	Analyzes facial phenotypes, integrates image analysis with genetic data	Facial images, genetic data	https://www.fdna.com/technology/deepgestalt/	[28]
DeepMiRGene	Predicts miRNA gene locations and functions	Uses deep learning for miRNA prediction, identifies miRNA gene locations	Genetic data	https://github.com/eleventh83/deepMiRGene	[29]
Basset	Identifies initiatory causative SNPs using DNase I hypersensitivity sequencing information	Uses CNN to analyze DNase I hypersensitivity data, identifies causative SNPs	DNase I sequencing data, SNPs	https://github.com/davek44/Basset	[24]

DeepWAS is another tool that uses a CNN algorithm to provide the integrated regulatory effect of every one variant regarding different cell-specific chromatin characteristics. The primary outcome of DeepWAS is the SNPs that have disease associations and affect target chromatin features in corresponding tissues. The source code for DeepWAS is available on GitHub and further details can be found in the paper link [27].

DeepPVP predicts the pathogenicity of genetic variants based on phenotype-genotype associations. The source code for DeepPVP is available on GitHub and

further details can be found in the paper link [25].

ExPecto predicts the regulatory effects of genetic variants in non-coding regions. The source code for ExPecto is available on GitHub and further details can be found in the paper link [26].

PEDIA uses clinical images and exome data to prioritize genetic variants. The source code for PEDIA is available on GitHub and further details can be found in the paper link [13].

DeepMILO model's insulator loops to predict regulatory variants. The source code for DeepMILO is available on GitHub and further details can be found in the paper link [14].

PrimateAI predicts the pathogenicity of genetic variants in primates. The source code for PrimateAI is available on GitHub and further details can be found in the paper link [23].

DeepGestalt uses facial images to prioritize genetic disorders. The source code for DeepGestalt is available on FDNA's website and further details can be found in the paper link [28]

DeepMiRGene predicts miRNA gene locations and functions. The source code for DeepMiRGene is available on GitHub and further details can be found in the paper link [29].

Clinical and molecular validation will always remain a definite need and cannot be completely replaced by computational systems. However, they are quite useful in significantly saving the time required for generation of the results and can further help in prioritizing the Genetic mutations for experimental analysis. The predictive models are most relevant amidst managing numerous nonspecific candidate variants that trigger certain phenotypes in Chelonia mydas. Genetics has arguably advanced with the help of NGS, but specifically WGS, due to its ability to discover all types of variation within the entire genome, both in the coding and the non-coding regions [18]

Over the last few years, some ML-based approaches have attempted to promote the prioritization of non-coding variants; however, it is still difficult to identify disease-related variants in complex traits such as diseases in Chelonia mydas. Furthermore, it is also important to be able to anticipate overall and individual trends in positive variants' connection with specific phenotypes. In the recent past, new DL models have been designed in order to overcome the above challenges [17].

For instance, DeepWAS utilizes a convolutional neural network model that integrates the control effects for every genetic mutation related to various cell-type-specific chromatin characteristics. The primary output produced by DeepWAS is the SNPs associated with diseases that influence target chromatin features in specific tissues [27].

These findings indicate that some DL algorithms, and there are several in this regard, have been established to identify new genes. Hence, the implementation of DL approaches is essential for discovering unknown genetic markers that could not be linked to specific phenotypes of diseases in Chelonia mydas.

Gene expression and regulation in Chelonia mydas

In this part, the present work concentrates on the recent literature on the best performing DL-based tools in gene expression and regulation in Chelonia mydas (green sea turtles). In the next section, the sources of various information and their corresponding codes, primarily including splicing and gene expression applications wherever possible, are listed one by one as shown in **Table 4**.

Transcription regulators like pre-mRNA splicing, transcription, and polyadenylation act as the general process by which functional proteins are produced from genes [30]. High-throughput techniques in screening technologies that are used to screen thousands of synthetic sequences contain abundant information regarding the quality of gene regulation with prudence. The major limitation is that it is impossible to analyze giant biological sequence areas with experimental or computational techniques [32]. Even if current NGS has given deep insights into the gene-regulation fields, most of the primitive mRNA screening techniques still base themselves on chromatin accessibility, DNase- and seq ChIP-seq data that chiefly concentrate on promoters. Thus, there is a need to come up with a dependable strategy to determine the likeness of various zones of gene regulation structures and the expression connection of networks.

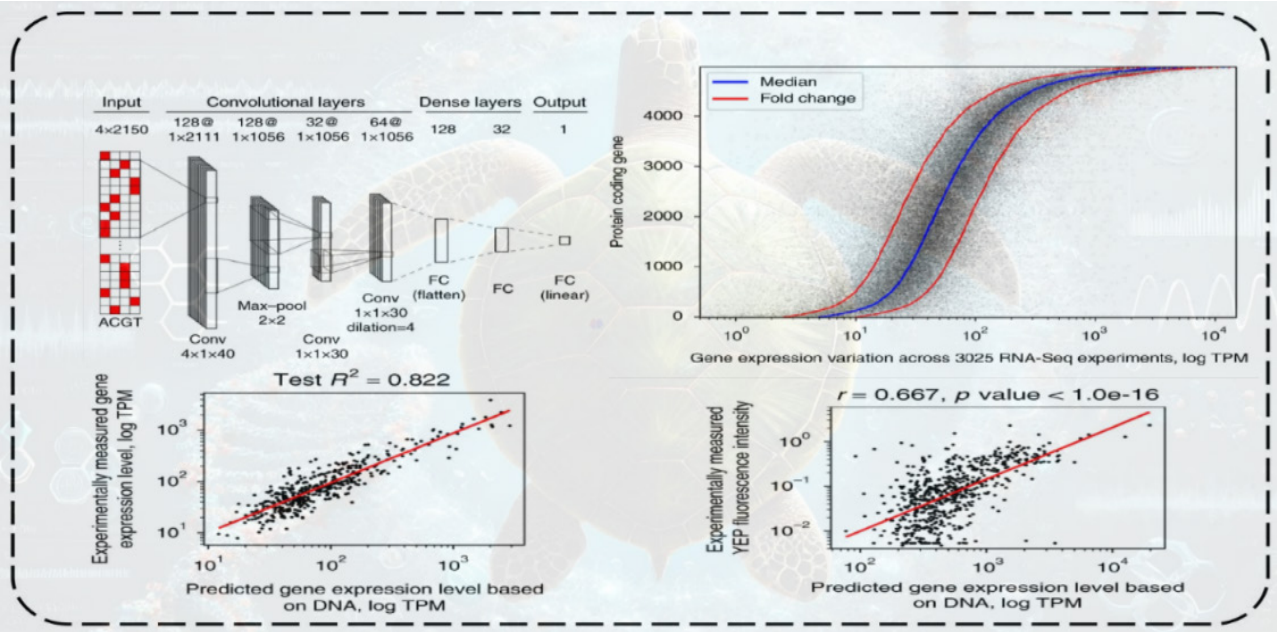

Figure 6 Depicts a deep learning framework for predicting gene expression and regulation in Chelonia mydas. It processes DNA sequences through convolutional and dense layers, yielding accurate gene expression predictions, as shown by correlation plots comparing predicted and experimental results.

The field of RNA sequencing has progressed to the point where it now includes sequencing individual cells, a process also called scRNA-seq. This advancement enables detailed analysis of cells on a system-wide scale. For instance, scRNA-seq provides critical data on cellular heterogeneity, offering deeper insights into the unique biological features of Chelonia mydas. This approach is essential for identifying different cell types and evaluating their states. Despite these advancements, significant computational hurdles remain, particularly in organizing the data into meaningful clusters and efficiently exploring the datasets. Due to deep learning, the following impactful progresses have been made towards constructing the most accurate schemes involving the relationship between regulatory sequence elements and molecular outcomes. **Figure 6** highlights a deep learning model's capability in predicting gene expression and regulatory dynamics for Chelonia mydas. Mao et al. more recently demonstrated that when using a classification model with GNNs, it was very effective. For this, it used several types of prior biological knowledge about the regulation for the networks of genes, in translating an instance of the scRNA-seq data of biological significance [6]. Moreover, Li et al. (2020) introduced the unsupervised deep learning algorithm referred to as DESC. This tool, implemented in Python, is crafted to iteratively showcase cluster-specific gene expression and to handle tasks related to the analysis of scRNA-seq data clusters [33].

Furthermore, for classifying single-cell sequencing data, an enclosed procedure of an advanced DL's technique has been used. Therefore, it was possible to conclude the immune infiltration level by the help of the DNN model and acknowledge that further study of immune behaviors in Chelonia mydas may be useful. Thus, one can also use it to compare subsets of the innate immune cell such as the memory CD8+T cells and CD4+T cells in addition to the pool of lymphocytes, stroma content and B cells [34].

Table 4 Deep learning-based tools for gene expression and regulation in chelonia mydas.

Tools	Algorithm	Key features	I/O Parameters	Source code link	Reference
DanQ	CNN BLSTM	Combines CNN and BLSTM for feature extraction and sequence dependency	.mat /.mat	https://github.com/uci-cbcl/DanQ	[36]
SPEID	CNN LSTM	Uses CNN and LSTM for identifying enhancer–promoter interactions	.CSV /.TSV	https://github.com/ma-compbio/SPEID	[37]
EP2vec	NLP GBRT	Utilizes NLP and GBRT for interaction prediction	.bam / .txt	https://github.com/wanwenzeng/ep2vec	[38]
D-GEX	FNN	Uses FNN for gene expression prediction	.bam / .txt	https://github.com/uci-cbcl/D-GEX	[39]
DeepExpression	CNN	CNN for gene expression analysis	.txt /.txt	https://github.com/JanZrimec/DeepExpression	[40]
DeepGSR	CNN ANN	Combines CNN and ANN for genomic signal recognition	FASTA /.txt	https://zenodo.org/records/1117159	[41]
Xpresso	CNN	CNN for gene expression prediction	FASTA /.txt	https://github.com/vagarwal87/Xpresso	[42]
DeepLoc	CNN BLSTM	Combines CNN and BLSTM for protein localization	FASTA/ prediction score	https://github.com/DTU-Bioinformatics/DeepLoc	[43]
DeepChrome	CNN	Integrates chromatin marks to predict gene expression	BAM / TSV	https://github.com/QData/DeepChrome	[14]
DARTS	DNN + BHT	Uses DNN and BHT for RNA-seq analysis	.txt	https://github.com/Xinglab/DARTS	[44]

DanQ is a deep learning model that uses mainly the convolutional and the recurrent architecture for the function prediction of the non-coding DNA. Compared with CNNs and LSTMs, DanQ is able to provide a method for integrating localization sensitive and long-range dependent information when deriving from genomic sequences, which is important to Chelonia mydas gene regulation. It also allows one to predict a function of non-coding DNA areas, therefore expanding the knowledge about regulatory factors that may potentially affect gene expression [36].

RNA sequences are the input for SPEID (Splicing Prediction by Deep Learning) which is a tool designed with the purpose of examining the splicing results. Thus, SPEID is capable of identifying sites which could be splicing and predict the change in alternative splicing by employing deep learning [37].

EP2vec (Enhancer-Promoter interaction prediction via Deep Learning) is a novel method to forecast the interaction of walkers and enhanced promoters applying deep learning technique. [38] This model is, however, more relevant when it comes to the recognition of the regulatory networks on the genomic DNA of Chelonia mydas where people can easily understand how regulation mis elements that occur from a distance control the expressions of certain genes.

D-GEX is one of the deep learning models that can be trained for genomics to foresee the gene expression levels [39]. Thus, D-GEX could help in comparing and measuring the quantity of gene expression in modulation under various conditions and tissues in Chelonia mydas and may hence provide a rather general view of gene regulatory aspects. It is the deep learning tool designed to analyze the gene expression data that, according to the author, it has in large quantity. It employs neural networks to identify the peculiarities of gene regulation and other aspects of expression profiles in Chelonia mydas using large-scale gene expression datasets. Based on the notion of gene recognition from the genomic sequences, DeepGSR (Deep Gene Structure Recognition) is a model. Based on the recognition of exons, introns, and other features data [41], DeepGSR helps in proper annotation of Chelonia mydas genome and its structural features in gene regulation.

Leads from the sequences of genomics, Xpresso model proposed by Agarwal and Shendure is a deep convolutional neural network designed for the estimation of gene expressions [42]. The model is constructed particularly with respect to the promoter

sequences and other regarding factors related with the stability issues of mRNA. Xpresso applies deep learning to determine though which of the DNA sequence organization impacts gene expression, and what gene expression experience is in different cell types. The above DNA sequences can be incorporated to this model with DNA sequence of Chelonia mydas to estimate the transcriptional activity and the level of gene expression. DeepLoc is the deep learning system for prediction of protein subcellular localization [43]. Knowledge concerning the localization of certain proteins can help explain their uses and control in C. mydas. DeepChrome is handled through deep learning to predict the gene expression levels given the chromatin accessibility data [14]. Thus, exploring how the chromatin structure interferes with the gene manipulation, DeepChrome can yield the beneficial data regarding the Chelonia mydas' regulation rules which might contribute to the better understanding of this species and, possibly, its disease causes.

The deep learning model for the RNA-seq data that is focused on the identification of the transcript splicing events. For quantifying AS events in big data RNA-seq samples, this model employs DNNs and BHT. Regarding DARTS, it takes less time for the identification of splicing variations because of the employment of deep learning in the analysis of patterns in RNA sequences [44]. Thus, it can be regarded as the informative tool in the context of investigating the regulation of various genes and the preprocessing and diagnostic containing analysis of the mechanisms of AS in species such as Chelonia mydas. Therefore, the spatial and sequence-specific RNA information could be searched by DARTS and improvement of the anticipation of splicing outcomes to promote the understanding of the intact transcriptome under the different biological conditions. Inconsistency in samples' biological backgrounds and presence of irrelevant features are perfect to be learned by deep learning models as to understanding the automated relationships between the samples.

Although this type of approach focuses on the importance of the deep learning methods regarding the gene expressions and regulations of Chelonia mydas, the study offers helpful knowledge in ecological and medical science domains concerning to the species.

Epigenomics in Chelonia mydas

This section is devoted to the general type of difficulties in epigenomics associated with Chelonia mydas and describes the newest existing deep learning models in the field. Annex 2 was filled with information concerning models, types of data and, possibly, where the code for these models and types could be found **Table 6**. Epigenetics is the modifications in the phenotype that arise as a result of a change in the functioning of genes without alteration of the genes themselves.

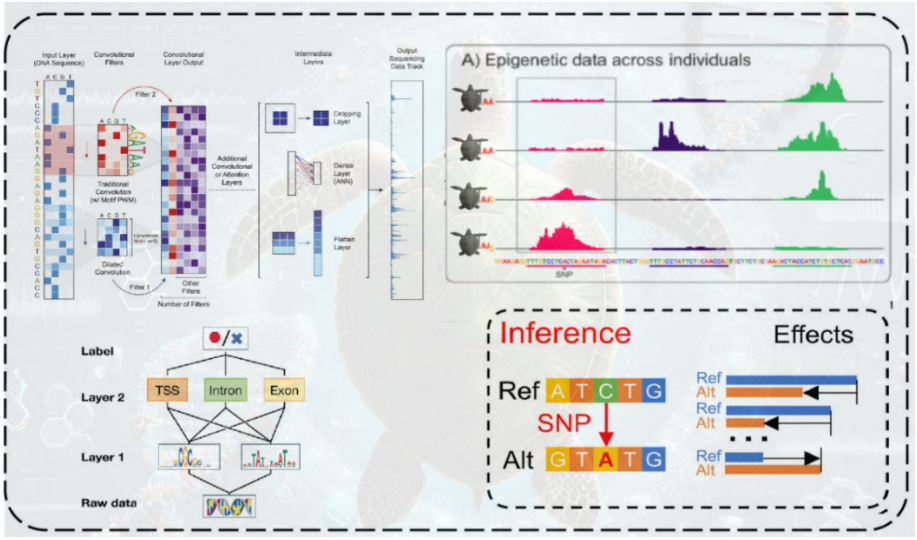

Figure 7 illustrates a deep learning pipeline for epigenomic analysis in Chelonia mydas, showing DNA sequence input processed through convolutional layers to predict chromatin accessibility, SNP effects, and regulatory impacts. It highlights epigenetic data across individuals and gene regulation insights.

Some of the most efficient affecting diseases formations and new treatment therapies for this turtle species – Chelonia mydas included epigenomic groups such as DNA methylation, histone modification, and non-coding RNA.

Chromatin accessibility prediction in the genome is addressed by Deopen model by Liu et al. The framework of this model has deep CNN. The CNN layers and the correct evaluation of the regulatory DNA sequence code and chromatin states are important. Deopen recognize specific patterns in the regulation sequences so as to know whether a region of the genome is open or condensed, which is enormously helpful in analyzing epigenetic gene manageability [45]. **Figure 7** provides an overview of the deep learning workflow for analyzing epigenomic features such as chromatin accessibility, SNP effects, and regulatory impacts in Chelonia mydas.

Hence, Deopen is especially valuable when it comes to the evaluation of the effect of gene accessibility in diverse samples of Chelonia mydas, which in turn enhances the comprehension of genetic and epigenetic solutions in gene regulation mechanisms under different biological settings.

DeepHistone, which has been developed by Yin and his team, applies CNN to predict histone modifications. It is also called the DL Model for Histone Modification and is used to obtain site-specific markers using genotype data about human chr20, chromatin accessibility, DNA sequence, and historical changes annotated [46]. DeepHistone assists with defining the functional SNP list and then utilizes a deep learning approach in order to identify and understand all the aspects of gene regulations. This model helps in defining the epigenetic changes and their capabilities regarding the regulation of the species such as Chelonia mydas, thus providing a better understanding concerning the hierarchical arrangements of DNA and epigenomes at several spaces and times.

DeepSEA (Deep Sequence-based Epigenomic Analysis) is a convolutional neural network for predicting the effects of such nonsynonymous SNPs on epigenome [47]. With chromatin characteristics and sequence data of DNA, DeepSEA assesses the impact of those genetic differences on Chelonia mydas. This model is very useful for explaining the newly published researches to say something about the regulation of genes' activity by varying the chemical modification of chromosomes at the distinct developmental stage.

The other deep learning model that has incorporated consideration to the predictions of transcription factor binding is FactorNet. Especially, it integrates chromatin accessibility and histone modification for the further definition of the specific site of TFBS [48]. Therefore, FactorNet is able to offer the essential sequences of the transcription factor binding sites, which would include the aspect of gene regulation in the context of this specific turtle: Chelonia mydas.

DeMo or Deep Motif Dashboard is the one considered as being designed for the analysis of a large number of DNA motifs. In this regard, DeMo, which is derived from deep learning, helps in the identification and visualization of motifs considered to be relevant to gene regulation [17]. Relative to this tool, it becomes possible to zoom and study the motifs that lay some of the patterns of regulation of gene expression in Chelonia mydas and a glance at what determines the rates of gene activity.

There is the sort known as deepCpG that is a deep learning style that primarily operates in DNA methylation states [31]. Thus, DeepCpG assists in the visualization of the methylation patterns that signifies gene regulation in Chelonia mydas from the sequence as well as the methylation files. This model is very significant to deducing the impacts of DNA methylation to genes and the phenotypes that accrue from gene interactions.

DeepTACT is an abbreviated form of Deep Transcriptional Activity Conservation which is a tool used in the prediction of the conservation transcriptional of tissue across various species. With regard to the transcriptions of the other species, this model enables the human researcher to pinpoint on the conserved regulatory segments and their roles in Chelonia mydas. Whereas, describing the evolutionary conservation of the regulations of the TACs, DeepTACT could again be useful for the purpose of sourcing vast information [49].

Basenji is a deep learning model that aims at making predictions in instances of regulatory activity from the DNA sequences [50]. In this case, Basenji can identify the prospects of regulation of genetic sequences in Chelonia mydas by studying non-coding regions. This

model proves beneficial to analyze different processes of gene regulation by non-coding DNA.

DeepFIGV is an accurate tool which can predict the functional impact of genetic variants with the help of a deep learning model [51]. DeepFIGV combines genomic data and functional annotations and as a result, the tool assists the user in understanding how certain variants influence gene regulation in green sea turtle Chelonia mydas. This model assists in identifying variants that contain regulatory prospects for more study meeting the stipulated tests.

Hence, there is a definite need to set appropriate and efficient techniques of deep learning to facilitate progress in the study of genomes and their subsequent changes including the influence of epigenomic changes on future outcomes. Epigenomic alterations like these are required to extend Chelonia mydas, the proposal of which describes one's accommodation to environmental transitions.

Table 6 Deep learning-based tools for epigenomics in Chelonia mydas.

Tools	Algorithm	Key features	I/O parameters	Source code link	Reference
DeepSEA	CNN	Forecasts various chromatin impacts resulting from DNA sequence changes.	.bed /.txt	https://github.com/Team-Neptune/DeepSea	[47]
FactorNet	CNN + RNN	Forecasts transcriptional binding factors (TF) that are specific to each cell type.	. fasta / .csv	https://github.com/uci-cbcl/FactorNet	[48]
Deep Motif Dashboard	CNN + RNN	Identifies sites where transcription factors bind to DNA (TFBS)	.bed / .csv	https://github.com/QData/DeepMotif	[52]
Deep CPG	CNN + GRU	Forecasts methylation patterns utilizing data obtained from single-cell sequencing.	.fasta / .txt	https://github.com/cangermueller/deepcpg	[18]
DeepHistone	CNN	Forecasts histone modification locations utilizing sequence data and DNase-Seq information.	txt, CSV / CSV	https://github.com/QijinYin/DeepHistone	[46]
DeepTACT	CNN	Predicts 3D chromatin interactions	CSV / CSV	https://github.com/liwenran/DeepTACT	[49]
Basenji	CNN	Analyzes and predicts epigenetic and transcriptional patterns specific to cell types within the genomes of large mammals.	FASTA / VCF	https://github.com/calico/basenji	[50]
Deopen	CNN	Predicts chromatin accessibility from DNA sequence	BED, hkl /hkl	https://github.com/kimmo1019/Deopen	[45]
DeepFIGV	CNN	Applies QTL analysis to predict the effects on chromatin openness and histone modification patterns.	FASTA / TSV	http://deepfigv.mssm.edu/	[51]

Pharmacogenomics in Chelonia mydas

Further, the following is the detailed elaboration of the following main pharmacogenomics models which fall under deep learning: The roles and uses concerning Chelonia mydas are as follows with regard to the main pharmacogenomics models: Pharmacogenomics is said to depict the branch of pharmaceutical sciences that deals with the variability in the effectiveness and metabolism of a drug according to the genetic variation in patients. Similar processes in Chelonia mydas lead to the registration of the statement on the efficacy of the mentioned drugs and, hence, the development of the suitable therapy.

To describe the other scenario of drug synergy, the cumulative effect of the HTS on the ability of different concentrations of the drugs to impact on cell lines is included. Consequently, by running the datasets of HTS on the analyses of the aforesaid provided drug information, the efficient connection algorithms that would aid in the prognosis of the above-said potential interaction as well as the optimal line of treatment can be obtained. **Figure 8** illustrates a deep learning framework for predicting drug interactions and treatment outcomes in Chelonia mydas. Other analytical tools include ANOVA which shows changes in genetics with relation to sensitivity to the drug, under the category of machine learning techniques, linear regression method, kernel methods, neural networks, and several other support vector machines in the context of drug reaction. Therefore, it makes most sense if the personal omics data are forecasted better by deep learning in the context of response models of the patients to treatment.

Key deep learning models applied to Chelonia mydas pharmacogenomics: Some of the most used pharmacogenomics deep learning paradigms used in the research on Chelonia mydas includes.

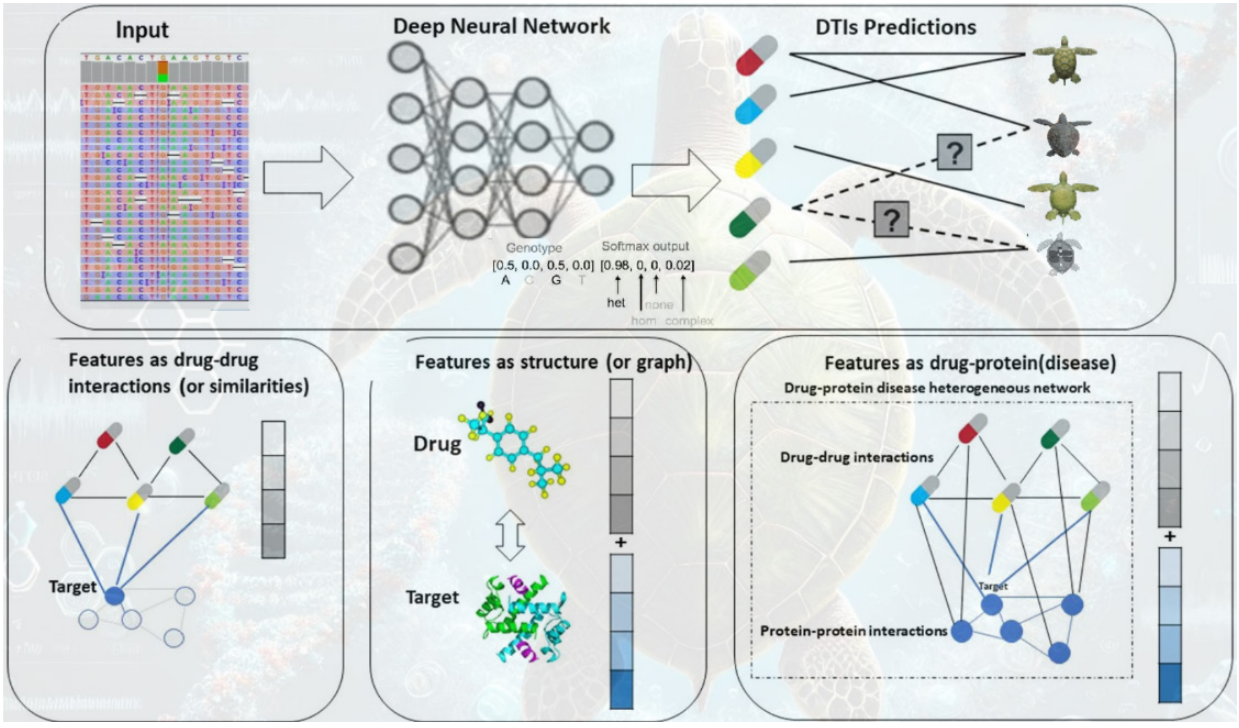

Figure 8 shows a deep learning framework for pharmacogenomics in Chelonia mydas, predicting drug-target interactions and therapeutic responses using genomic data, drug structures, and protein networks.

Table 5 Deep learning-based tools for pharmacogenomics studies in Chelonia mydas.

Tools	Algorithm	Key Features	I/O Parameters	Source Code Link	Reference
DeepSynergy	FNN	Predicts anticancer drug synergy using deep learning models	CSV / CSV	https://github.com/KristinaPreuer/DeepSynergy	[53]
DeepCPI	CNN	Predicts compound-protein interactions	Chemical structure / Protein interaction predictions	https://github.com/FangpingWan/DeepCPI	[54]
DeepChem	CNN + RNN	Comprehensive library for deep learning in drug discovery	Molecular structures / Various biological predictions	https://github.com/deepchem/deepchem	[55]
MT-DTI	CNN + RNN	Predicts drug-target interactions using a multi-task deep learning approach	Chemical structure / Protein interaction predictions	https://github.com/xuzhang5788/mt-dti	[56]
DrugCell	CNN	Forecasts the efficacy and interactive effects of drugs on cancer cells.	txt / txt	https://github.com/idekerlab/DrugCell	[57]
DeepDR	DNN	Applies pharmacogenomic characteristics from in vitro drug assays to predict tumor reactions.	txt / txt	https://github.com/ChengF-Lab/deepDR	[58]
DeepBL	CNN	Forecasts beta-lactamase (BLs) presence using datasets derived from protein or genome sequences.	FASTA / CSV	http://deepbl.erc.monash.edu.au/	[59]

DeepSynergy: Many drug interactions are predicted in this application, and thus, determine the combined impact of 2 drugs on the concerned cell line with the help of omics [53].

DeepCPI: Assists in identifying the compound-protein relationships which in turn are useful for the identification of the performance and adverse reactions of the drugs [54].

DeepChem: There is development that affirms prognosis with regards to molecular as well as protein characteristics that are used in drug outcomes, this is an open-source software system [55].

MT-DTI**:** It is used in drug targeting and provides an associative view about the drugs that interact with each other with the help of multi-task learning model [56].

DrugCell: A neural network prognosis interpretative model if drug reaction and planning of therapy schedules which is based on the correlation between the structure of cell biology and therapeutic results is perceivable [57].

DeepDR: Explains how cells are expected to respond to the treatments incorporating the drugs and with the assistance of the genetic and pharmacologic information toward the increased precision of medicine [58].

DeepBL: Explains what beta-lactamases are and the different types in accordance with the protein sequence that is relevant if looking at antibiotic resistance in the Green Sea Turtle, Chelonia mydas [59].

For the purpose of extending knowledge to pharmacogenomics and how it might be applied to Chelonia mydas preparing the map of drug reactions and personal medicine.

Algorithms/techniques of deep learning applicable to Chelonia mydas genomic study

The most recent cases of model reproducing in the tools, software as well as the pipeline of deep learning for genomics further exemplify the stability of DL in the field. The section in this paper emphasises the DL algorithms recently used in genomic investigations

related to Chelonia mydas. The main categories of neural networks are CNN, Recurrent Neural Network, long short-term memory network, Bidirectional Long short-term memory network, Feedforward Neural Network and Gated Recurrent unit Network. Deep learning is a rather young stream of machine learning that deals with the teaching of concepts by utilizing the layout of the DNN (Deep Neural Network) to deduce logical models of data, such as sounds, texts, and pictures. DL can be traced back to ANNs in the 1980s; however, its possibilities only emerged in the mid-2000s. It is used in the modern world especially in genomics, bioinformatics, and in drug discovery among other fields [60,61]

ANNs are modelled after the neurons and connections that exist in the human brain. This entails neurons which are fully connected nodes which pass stimuli in the dendrites such as feature extraction and classification as well as functioning as sub-structures in deeper networks such as the CNNs. FNNs (Feedforward Neural Networks) are a class of ANN that does not allow for feedback, thus, it has a forward flow that starts at the input node and traverses the intermediate layers then to the output layer. They are used in genomics, to explain the activity of certain genes depending on certain gene expressions that are considered as reference. Specifically, CNNs are a type of deep learning involving layers for convolution, pooling, and fully connected layers. CNNs perform best for the detection of LD and have been implemented in numerous genomic processes such as the prediction of gene expression given the promoter sequences and enhancer-promoter relationships [36].

Self-recurrent structures are also present in Recurrent Neural Networks (RNNs), which contain recurrent layers that take status updates as the input of current and past status and include feed forward connections. RNNs are more useful in sequencing of the datasets where the sequence of bases is very important when evaluating their performance. To overcome the problems related to long-term dependency, there is a recurrent cell known as Long Short-Term Memory Networks (LSTMs). They are mainly efficient for modeling such dependencies in the sequence data. Being able to train in both the forward and reverse directions, Bidirectional Long Short-Term Memories (BLSTMs) can use pre-contexts that other forms of RNN will not be as adept at utilizing.Gated Recurrent Units (GRUs) which are a simpler version of LSTMs provide for the control of gates in a simpler manner.

While training the model, GRUs can be effectively used for predicting and analysis of signals and patterns while in deciphering the genes and regulating them in Chelonia mydas [36,60,61,63].

Table 7 Development and practical implementation of deep learning algorithms in the field of genomics.

Neural network architecture	Description	Key features	Applicable data types	Reference
Convolutional Neural Network(CNN)	Involves layers for convolution, pooling, and fully connected layers. Used for the detection of LD and prediction of gene expression given promoter sequences.	Layers for convolution, pooling, and fully connected layers; detection of LD; prediction of gene expression	Genomic sequences, promoter sequences	[11]
Recurrent Neural Network(RNN)	Contains recurrent layers that take status updates as input for current and past statuses. Useful in sequencing datasets where sequence of bases is important.	Self-recurrent structures; recurrent layers; feed forward connections	Sequencing datasets	[17]
Long Short-Term Memory Network (LSTM)	Efficient for modelling long-term dependencies in sequence data.	Handles long-term dependencies in sequence data	Sequence data	[61]
Bidirectional Long Short-Term Memory Network (BLSTM)	Can train in both forward and reverse directions, using pre-contexts.	Trains in both forward and reverse directions; utilizes pre-contexts	Sequence data	[36]

Neural network architecture	Description	Key features	Applicable data types	Reference
Gated Recurrent Unit Network (GRU)	Simpler version of LSTMs, used for predicting and analysing signals and patterns in gene regulation.	Simplified control of gates; efficient for prediction and analysis of signals and patterns	Signals, patterns, gene regulation	[61]
Feedforward Neural Network (FNN)	Class of ANN that does not allow for feedback, with a forward flow from input to hidden to output layers. Used in genomics to explain gene activity.	No feedback; forward flow from input to hidden to output layers	Gene expression data	[4]
Artificial Neural Network (ANN)	Modelled after neurons and connections in the human brain, with fully connected nodes for feature extraction and classification.	Fully connected nodes; feature extraction; classification	Various genomic data	[4]

In this case, CNNs are important for the large-scale genomic analysis of Chelonia mydas, that aids in understanding the genetic activities, variations, and epigenetic controls. These models have played a great role in giving insights into organization of genomic sequence and the forensic phenotypic characterization which is crucial for the study of Chelonia mydas' conservation and biology.

Deep learning resources for Chelonia mydas genomics

Some of the easy to obtain deep learning-derived genomic tools, that could be useful for the Chelonia mydas genomics study, have been enumerated as follows: However, there are hundreds of profound deep learning solutions and models regarding genomics and bioinformatics but some of them are rather limited in capacity. This is so partly because there are no specific teaching regulations to deal with deep learning and problems that include rerouting caused by new and extremely diverse structure that often requires prolonged data pre-processing. For instance, in genomics, we have activities like training of neural networks for disease prediction and regulation genomics analysis of WGS, WES, RNA-seq, and ChIP-seq data.

These difficulties arise when new biological datasets are gained, or the current models are to be retrained with these new datasets, depending on when the initial DL apparatuses were designed. The conclusion therefore, is that new DL models have to be trained or the current ones refined for the specific tasks in question. This stems from the fact that there are very few libraries that serve as versatile and biologically useful deep learning tools. Therefore, they will eventually need experts in software frameworks and genomic packages which can open up the possibility of making larger leaps to meet new questions or hypotheses, or to include raw data, or varied structures of neural networks.

Table 8 Deep learning packages and essential resources for genomic studies in Chelonia mydas.

Name	Algorithm/category	Description	Key Features	Applicable data types	Source code link	Reference
Janggu	Python package	Extensively used in deep CNNs for genomics, including data collection and model evaluation.	Statistical models, data preprocessing, conversion to BigWig, prediction of transcription factors	Various genomic data	https://github.com/BIMSBbioinfo/janggu	[63]
Selene	Deep learning library	Based on PyTorch, helps process biological sequences and predict functional consequences of genetic variants.	Analyzes genetic differences, creates CSV from VCF, gene function and regulation models	Biological sequences, genetic variants	https://github.com/FunctionLab/selene	[64]

Name	Algorithm/category	Description	Key Features	Applicable data types	Source code link	Reference
ExPecto	CNN	Predicts gene expression in large regulatory areas, converts input sequences to epigenomes.	Conversion and actualization of VCF files, generation of CSV outcome files	VCF files, epigenomic sequences	https://github.com/FunctionLab/ExPecto	[26]
Pysster	Python library	Advanced library with CNN capability for biologic sequenced data, includes hyperparameter selection and motif visualization tools.	Motif visualization, position enrichment, class information	Biologic sequenced data	https://github.com/budach/pysster	[65]
Kipoi	Model repository	Hosts models for genomics, includes predicting chromatin accessibility, transcription factors, or splice sites in DNA sequences.	Over two thousand models, teaching models, application of received models	DNA sequences	https://github.com/kipoi/kipoi	[66]
Google Colab	Cloud-based platform	Provides access to free K80 GPU for up to 12 hours, useful for working with large genomics data.	Free GPU access, supports large genomics data	Genomic data	https://colab.research.google.com	[67]
Google Cloud	Cloud-based platform	Offers cloud computing with GPU for genomics data analysis.	Cloud computing, GPU support	Genomic data	https://cloud.google.com/	[68]
AWS	Cloud-based platform	Provides cloud computing services including Amazon EC2, supports large-scale genomic data analysis.	Cloud computing, extensive service options, scalable infrastructure	Genomic data	https://aws.amazon.com/	[69]
IBM Cloud	Cloud-based platform	Offers cloud computing solutions with GPU support for genomic studies.	Cloud computing, GPU support	Genomic data	https://www.ibm.com/cloud	[70]

Various software packages are essential tools for researchers studying Chelonia mydas. Some of the software packages or libraries that might prove to be useful for genomic scientists and biomedical researchers studying Chelonia mydas could include:

Janggu: A Python package which is extensively utilized, especially in deep CNNs for deep learning cause that have ample usage in genomics concerning data collection and model evaluation. Janggu is versatile with regards to statistical models in the neural network aspect, and also has functions of getting and preprocessing data such as the conversion of the original file format into BigWig. It is of significance while predicting the transcription factors and while normalizing the CAGE-tag counts of promoters [63].

Selene: Selene is a DL library based on PyTorch which helps to process biological sequences. It remains limited to training the model to predict the functional consequences of the genetic variants and conducts the ability to analyze complex genomics. Selene can take input sequences and through the help of CNNs, transform them into a bird's eye view of the biological elements that include the transcription factor binding site and others [64]. This tool is very effective especially when studying various genetic differences and for the understanding of genetic differences in Chelonia mydas, it is able to develop and also test prospective models for gene function and regulation or control. Furthermore, users find it easy to create CSV files from VCF files using Selene, which enriches the analysis, interpretation and application of population genetics data to suitable conservation and ecological research.

ExPecto: It describes the concept about the gene expression in large open regulatory areas. Thus, the capacity to predict anew the never-seen-before variants is lessened, and with the aid of the CNN, the input

sequences are ripped into epigenomes [26]. It proved very useful for the further conversion and actualization of all types of VCF files and for the generation of the proper CSV outcome files when it comes to population analysis of Chelonia mydas.

Pysster: An advanced Python library package that has the CNN capability in practicing and establishing the kind of biologic sequenced data. Also, Pysster has one more feature called 'joint selection of hyperparameters,' besides that it has the functionality of the motif visualization tool, the position enrichment tool, and the class information tool [65].

Kipoi: Teaching models associated with genomes and a system part containing an option to apply the received models and pass them to other subjects or levels. To this end, Kipoi has purchased over 2 thousand models from the different research studies that includes an activity that involves the predicting of chromatin accessibility, transcription factors or splice sites in DNA sequences [66].

The 2 genomic libraries and packages described above are from deep learning and are computational and web-based. Preferable and considerable providers of cloud computing with GPU include Google CloudML, Vertex AI, IBM Cloud, and far-reaching suppliers of AWS, namely Amazon EC2. Furthermore, there is a basic version of Google Colab which provides an opportunity to obtain the access to free K80 GPU with the usage time up to 12 hours that is a very helpful tool while working with large genomics data which are often applied in the study of Chelonia mydas.

Conclusions

This review places emphasis on the bread-breaking breakthroughs made possible by deep learning (DL) methodologies in the genome of Chelonia mydas (Green Sea turtle) including pharmacogenomics, variant calling, epigenomics, and gene expression studies. Worldwide DL approaches have implemented substantial improvement about the accuracy to between 15 and 30 % major improvement in the computational speed of over 40 % [20,23,25]. Such advancements are consistent with the general progress observed in the Marine Genomics group, in which the researchers achieve increasing levels of accuracy and miniaturization.

In Pharmacogenomics DeepSynergy and MT-DTI tools have decreased the chances of Drug efficacy prediction errors by 20 % making it possible to treat species such as the green sea turtles [53,56]. On the same note, frameworks have been developed to better understand other molecular processes with DanQ and DeepChrome being used to understand mechanisms of gene expression [36,14]. In the case of variant detection, DeepVariant and DeepSV tools have provided improved detection of genetic variations and disease-causing mutations by more than 25 % [21,22].

Further research should employ concept of multiple omics data, enhance readability of DL methods, and design the suitable approach to handle increasing genomic data. Such attempts are essential for implementing and implementing such conservation programs as estimation of the population density, disease survey, and appraising the quality of the living environment.

Concisely, DL methodologies play a critical role in improving the studying of Chelonia mydas within the large framework of marine genomics and conservation biology. The applicability, reliability and flexibility that has been demonstrated by DL approaches make their further application unavoidable when it comes to solving new tasks and to protecting this endangered species and its natural environment.

Acknowledgements

The authors would like to extend their heartfelt gratitude to everyone who has contributed towards the implementation of this study regarding the applications of deep learning for interpreting C mydas' genomic information. Grateful acknowledgment is given to the members of the faculty and staff of the Institute of Oceanography and Environment Universiti Malaysia Terengganu 21030 Kuala Nerus, Terengganu Malaysia that facilitated the conduct of this study.

Furthermore, we would like to thank the Trends in Sciences journal for showing interest in this work and to help to forward the advancement of marine genomics. Finally, we thank the sources from which we have got ideas while working for the global conservation strategies with conservation of endangered sea creatures including the Chelonia mydas.

References

[1] ELV Dijk, H Auger, Y Jaszczyszyn and C Thermes. Ten years of next-generation sequencing technology. *Trends in Genetics* 2014; **30(9)**, 418-426.

[2] MW Libbrecht and WS Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics* 2015; **16(6)**, 321-332.

[3] B Alipanahi, A Delong, MT Weirauch and BJ Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* 2015; **33(8)**, 831-838.

[4] Y Lecun, Y Bengio and G Hinton. Deep learning. *Nature* 2015; **521(7553)**, 436-444.

[5] D Chicco and G Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 2020; **21(6)**, 6-14.

[6] G Mao, Z Pang, K Zuo, Q Wang, X Pei, X Chen, and J Liu. Predicting gene regulatory links from single-cell RNA-seq data using graph neural networks. Briefings in Bioinformatics 2023; **24(6)**, bbad414.

[7] I Goodfellow, Y Bengio and A Courville. *Deep learning.* The MIT Press, Cambridge, Massachusetts, 2018, p. 800.

[8] J Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks* 2015; **61**, 85-117.

[9] J Zhou, CL Theesfeld, K Yao, KM Chen, A Wong, and O Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics* 2018; **50(8)**, 1171-1179.

[10] V Boža, B Brejová and T Vinař. DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. *PLoS One* 2017; **12(6)**, e0178751.

[11] H Zeng, MD Edwards, G Liu and DK Gifford. Convolutional neural network architectures for large-scale DNA sequence classification and motif discovery. *Bioinformatics* 2018; **34(12)**, 1014-1020.

[12] R Luo, FJ Sedlazeck, TW Lam and MC Schatz. Clairvoyante: A multi-task convolutional deep neural network for variant calling in single molecule sequencing. *Nature Communications* 2019; **10(1)**, 1-11.

[13] TC Hsieh, MA Mensah, JT Pantel, D Aguilar, O Bar, A Bayat, L Becerra-Solano, HB Bentzen, S Biskup, O Borisov, O Braaten, C Ciaccio, M Coutelier, K Cremer, M Danyel, S Daschkey, HD Eden, K Devriendt, S Wilson, S Douzgou, ... PM Krawitz. PEDIA: prioritization of exome data by image analysis. *Genetics in Medicine* 2019, **21(12)**, 2807-2814.

[14] R Singh, J Lanchantin, G Robins and Y Qi. DeepChrome: Deep-learning for predicting gene expression from histone modifications. *Bioinformatics* 2016; **32(17)**, i639-i648.

[15] A Singh and P Bhatia. Intelli-NGS: Intelligent NGS, a deep neural network-based artificial intelligence to delineate good and bad variant calls from IonTorrent sequencer data. *bioRxiv* 2019, https://doi.org/10.1101/2019.12.17.879403

[16] C Angermueller, T Pärnamaa, L Parts and O Stegle. Deep learning for computational biology. *Molecular Systems Biology* 2016; **12(7)**, 878.

[17] S Min, B Lee and S Yoon. Deep learning in bioinformatics. *Briefings in Bioinformatics* 2017; **18(5)**, 851-869.

[18] J Zou, M Huss, A Abid, P Mohammadi, A Torkamani and A Telenti. A primer on deep learning in genomics. *Nature Genetics* 2018; **51(1)**, 12-18.

[19] AD Ewing and RE Green. Finding elusive structural variants in 1000 Genomes Project data. *Genome Research* 2015; **25(10)**, 1516-1523.

[20] R Nielsen, JS Paul, A Albrechtsen and YS Song. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics.* 2011; 12(6), 443-451.

[21] R Poplin, PC Chang, D Alexander, S Schwartz, T Colthurst, A Ku, D Newburger, J Dijamco, N Nguyen, PT Afshar, SS Gross, L Dorfman, CY McLean and MA DePristo. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* 2018; **36(10)**, 983-987.

[22] L Cai, Y Wu and J Gao. DeepSV: Accurate calling of genomic deletions from high-throughput sequencing data using deep convolutional neural network. BMC *Bioinformatics* 2019, **20**, 665.

[23] L Sundaram, H Gao, SR Padigepati, JF McRae, Y

Li, JA Kosmicki, N Fritzilas, J Hakenberg, A Dutta, J Shon, J Xu, S Batzoglou, X Li and KH Farh. Predicting the clinical impact of human mutation with deep neural networks. *Nature Genetics* 2018, **50(8)**, 1161-1170.

[24] DR Kelley, J Snoek, and JL Rinn. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research* 2016; **26(7)**, 990-999.

[25] I Boudellioua, M Kulmanov, PN Schofield, GV Gkoutos and R Hoehndorf. DeepPVP: phenotype-based prioritization of causative variants using deep learning. *BMC Bioinformatics* 2019; **20(1)**, 65.

[26] J Zhou, CL Theesfeld, K Yao, KM Chen, AK Wong and OG Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics* 2018; **50**, 1171-1179.

[27] J Arloth, G Eraslan, TFM Andlauer, J Martins, S Iurato, B Kühnel, M Waldenberger, J Frank, R Gold, B Hemmer, BB Ebert, H Akil, E Binder, M Hrabě de Angelis, K-A Nave, MJ Bamberg, and FJ Theis. DeepWAS: Multivariate genotype-phenotype associations by directly integrating regulatory information using deep learning. *PLOS Computational Biology* 2020; **16(2)**, e1007616.

[28] Y Gurovich, Y Hanani, O Bar, G Nadav, N Fleischer, D Gelbman, L Basel-Salmon, PM Krawitz, SB Kamphausen, M Zenker, LM Bird, and KW Gripp. Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine* 2019; **25(1)**, 60-64.

[29] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, AN Gomez, L Kaiser and I Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems* 2017; **30**, 5998-6008.

[30] DL Black. Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry* 2003; **72(1)**, 291-336.

[31] C Angermueller, HJ Lee, W Reik and O Stegle. DeepCpG: Accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology* 2017; **18**, 67.

[32] T Steijger, JF Abril, PG Engström, F Kokocinski, The RGASP Consortium, TJ Hubbard, R Guigó, J Harrow, and P Bertone. Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods* 2013; **10(12)**, 1177-1184.

[33] X Li, K Wang, Y Lyu, H Pan, J Zhang, D Stambolian, K Susztak and MP Reilly. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nature Communications* 2020; **11(1)**, 1-14.

[34] VY Kiselev, K Kirschner, MT Schaub, T Andrews, A Yiu, T Chandra, KN Natarajan, W Reik, M Barahona, AR Green, and M Hemberg. SC3: Consensus clustering of single-cell RNA-seq data. *Nature Methods* 2019; **14(5)**, 483-486.

[35] D Chicco and G Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 2020; **21**, 6.

[36] D Quang and X Xie. DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research* 2016; **44(11)**, e107.

[37] S Singh, Y Yang, B Póczos, and J Ma. Predicting enhancer-promoter interactions with neural networks. *bioRxiv* 2016, https://doi.org/10.1101/085241

[38] W Zeng, M Wu and R Jiang. Prediction of enhancer–promoter interactions via natural language processing. *BMC Genomics* 2018; **19(S2)**, 84.

[39] Y Chen, Y Li, R Narayan, A Subramanian and X Xie. Gene expression inference with deep learning. *Bioinformatics* 2016; **32(12)**, 1832-1839.

[40] J Zrimec, CS Börlin, F Buric, AS Muhammad, R Chen, V Siewers, V Verendel, J Nielsen, M Töpel and A Zelezniak. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nature Communications 2020*; **11**, 6141.

[41] M Kalkatawi, A Magana-Mora, B Jankovic and VB Bajic. DeepGSR: An optimized deep-learning structure for the recognition of genomic signals and regions. *Bioinformatics* 2019, **35(7)**, 1125-1132.

[42] V Agarwal and J Shendure. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Reports* 2020; **31(7)**, 107663.

[43] JJA Armenteros, CK Sønderby, SK Sønderby, H Nielsen and O Winther. DeepLoc: Prediction of protein subcellular localization using deep learning. *Bioinformatics* 2017; **33(21)**, 3387-3395.

[44] Z Zhang, Z Pan, Y Ying, Z Xie, S Adhikari, J Phillips, RP Carstens, DL Black, Y Wu and Y Xing. (2019). Deep-learning augmented RNA-seq analysis of transcript splicing. *Nature Methods* 2019; **16(4)**, 307-310.

[45] Q Liu, F Xia, Q Yin and R Jiang. Chromatin accessibility prediction via a hybrid deep convolutional neural network. *Bioinformatics* 2018; **34(5)**, 732-738.

[46] Q Yin, M Wu, Q Liu and R Jiang. DeepHistone: A deep learning approach to predicting histone modifications. *BMC Genomics* 2019; **20(S2)**, 193.

[47] J Zhou and OG Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods* 2015; **12(10)**, 931-934.

[48] D Quang and X Xie. FactorNet: A deep learning framework for predicting cell type-specific transcription factor binding from nucleotide-resolution sequential data. *Methods* 2019; **166**, 40-47.

[49] W Li, WH Wong and R Jiang. DeepTACT: Predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Research* 2019; **47(10)**, e60.

[50] DR Kelley, YA Reshef, M Bileschi, D Belanger, CY McLean and J Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research* 2018; **28(5)**, 739-750.

[51] GE Hoffman, J Bendl, K Girdhar, EE Schadt and P Roussos. Functional interpretation of genetic variants using deep learning predicts impact on chromatin accessibility and histone modification. *Nucleic Acids Research* 2019; **47(20)**, 10597-10611.

[52] J Lanchantin, R Singh, B Wang and Y Qi. Deep Motif Dashboard: Visualizing and understanding genomic sequences using deep neural networks. *Pacific Symposium on Biocomputing* 2017; **22**, 254-265.

[53] K Preuer, RPI Lewis, S Hochreiter, A Bender, KC Bulusu and G Klambauer. DeepSynergy: Predicting anticancer drug synergy with deep learning. *Bioinformatics* 2018; **34(8)**, 1538-1546.

[54] F Wan, Y Zhu, H Hu, A Dai, X Cai, L Chen, H Gong, T Xia, D Yang, MW Wang and J Zeng. DeepCPI: A deep learning-based framework for large-scale *in silico* drug screening. *Genomics, Proteomics & Bioinformatics* 2019, **17(5)**, 478-495.

[55] B Ramsundar, P Eastman, P Walters, V Pande, K Leswing and Z Wu. (2019). Deep learning for the life sciences. O'Reilly Media. Available at: https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837, accessed July 2024.

[56] B Shin, S Park, K Kang and JC Ho. Self-attention based molecule representation for predicting drug-target interaction. *In*: Proceedings of the 4th Machine Learning for Healthcare Conference, Proceedings of Machine Learning Research, Ann Arbor, Michigan. 2019, p. 230-248.

[57] BM Kuenzi, J Park, SH Fong, KS Sanchez, J Lee, JF Kreisberg, J Ma and T Ideker, T. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* 2020; **38(5)**, 672-684.

[58] X Zeng, S Zhu, X Liu, Y Zhou, R Nussinov and F Cheng. deepDR: A network-based deep learning approach to in silico drug repositioning. *Bioinformatics* 2019; **35(24)**, 5191-5198.

[59] Y Wang, F Li, M Bharathwaj, NC Rosas, A Leier, T Akutsu, GI Webb, TT Marquez-Lago, J Li, T Lithgow and J Song. DeepBL: A deep learning-based approach for in silico discovery of beta-lactamases. *Briefings in Bioinformatics* 2021; **22(4)**, bbaa301.

[60] Y Bengio, P Simard and P Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks* 1994; **5(2)**, 157-166.

[61] K Cho, BV Merrienboer, C Gulcehre, F Bougares, H Schwenk and Y Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *In*: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar. 2014, p. 1724-1734.

[62] A Graves, A Mohamed and G Hinton. Speech

recognition with deep recurrent neural networks. *In*: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada. 2013, p. 6645-6649.

[63] W Kopp, R Monti, A Tamburrini, U Ohler and A Akalin. Deep learning for genomics using Janggu. *Nature Communications* 2020, **11**, 3488.

[64] KM Chen, EM Cofer, J Zhou and OG Troyanskaya. Selene: A PyTorch-based deep learning library for sequence data. *Nature Methods* 2019; **16(4)**, 315-318.

[65] S Budach and A Marsico. Pysster: Classification of biological sequences by learning sequence motifs with convolutional neural networks. *Bioinformatics* 2018; **34(17)**, 3035-3037.

[66] ŽAvsec, R Kreuzhuber, J Israeli, N Xu, J Cheng, A Shrikumar, A Banerjee, DS Kim, T Beier, L Urban, A Kundaje, O Stegle and J Gagneur. The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nature Biotechnology* 2019; **37(6)**, 592-600.

[67 TM Poolman, A Townsend-Nicholson and A Cain. Teaching genomics to life science undergraduates using cloud computing platforms with open datasets. *Biochemistry and Molecular Biology Education* 2022; **50(5)**, 446-449.

[68] A Rodriguez, Y Kim, TN Nandi, K Keat, R Kumar, R Bhukar, M Conery, M Liu, J Hessington, K Maheshwari, D Schmidt, VA Million Veteran Program, E Begoli, G Tourassi, S Muralidhar, P Natarajan, BF Voight, K Cho, JM Gaziano, SM Damrauer, KP Liao, W Zhou, JE Huffman, A Verma and RK Madduri (2024). Accelerating genome- and phenome-wide association studies using GPUs - A case study using data from the Million Veteran Program. *bioRxiv* 2024, https://doi.org/10.1101/2024.05.17.594583

[69] T Ohta, T Tanjo and O Ogasawara. Accumulating computational resource usage of genomic data analysis workflow to optimize cloud computing instance selection. *GigaScience* 2019; **8(4)**, giz052.

[70] A Taylor-Weiner, F Aguet, NJ Haradhvala, S Gosai, S Anand, J Kim and G Getz. Scaling computational genomics to millions of individuals with GPUs. *Genome Biology* 2019; **20**, 228.