Layman Summary

The application of machine learning approaches has become increasingly popular in the field of DNA analysis. Specifically the use of machine learning to detect parts of the DNA which are responsible for the regulation of protein levels. Since proteins are created from the DNA in a process called transcription (followed by translation), we refer to this phenomenon as transcriptional regulation. This process is responsible for instructing cells to perform different functions at different times and different contexts. For example, the cells that make up muscles having the ability to expand and contract, and the cells in the brain having the ability to receive and send electrical signals. In a sense, the complex networks created by regulation of transcription carry the instructions for creating and sustaining a living organisms on a cellular level. Throughout the lifetime of an organism, changes in the DNA code occur. Caused by specific exposures such as cigarette smoke, UV-light etc., and by the passing of time. It has been shown that ~80% of these DNA changes associated with diseases, occur in regions responsible for transcriptional regulation. This underlines the importance of understanding these regions, referred to as regulatory elements.

In recent years, the primary way of elucidating regulatory elements is through a combination of natural language processing and machine learning. Natural language processing refers to finding patterns within textual data, such as the DNA. These patterns can then be used to predict the location and functionality of specific regulatory elements. Since the genome of an average person contains about 6.270.000.000 bases, and the interactions between regulatory elements varies based on the sequence context, cell type and developmental stage, machine learning approaches outperform human analyses. Particularly machine learning approaches that take advantage of using many layers in their architecture. These extra layers allow such tools to transform the input more and in novel ways, thus leveraging more complex patterns within the DNA to make predictions about regulatory elements.

However, considering that these regulatory features of the DNA represent such an immensely promising field of study, based on the amount of disease mutations that target them. It is crucial to think about the risks related to the application of these tools. Most notably, the lack of interpretability of machine learning tools which apply many layers. Since every layer is free to transform the output of the previous layer in ambiguous ways, it becomes more complicated to explain how predictions were made. And should the perfect prediction be the end goal? Or the perfect understanding?

Through reviewing historical challenges, the methods used to surmount them and the hurdles that still remain, we aim to empower the reader to be aware of the implications of the current approaches. We highlight the methods by which the interpretability of many-layered machine learning tools can be enhanced, alongside the value of testing hypotheses provided by such tools. In doing so, we reach a positive but weary conclusion, namely that the current direction of the field is more promising than any before it, but that it cannot be walked blindly.

Generative AI statement

The author acknowledges limited application of Generative AI in the creation of this review. The application of these tools can be split into two categories; Namely, the identification of relevant sources, and the alteration of sentence/paragraph structure.

The identification of relevant sources exclusively involved supplying a generative AI with a topic and requesting possibly relevant sources. These sources where subsequently manually reviewed and used to increase the understanding of the topic. The alteration of sentence/paragraph structure was limited to supplying a generative AI with information and text created by the author and requesting a clearer or more concise structure.

All concepts, ideas, and opinions were created by the author; who expressly discourages the muddled understanding of a topic through over application of generative AI. Which is one of the primary arguments within the review itself. The author approximates that 15-25% of the text in the final product was subject to some degree of structural alteration. Herein no specific sections are exempt.

Influence of Machine Learning Advancements on the Detection of DNA Regulatory Elements

Wietse Niederländer
Utrecht University
27-12-2023

Abstract

This review explores the intersection of Natural Language Processing (NLP), machine learning, and genomics, with a focus on the understanding and interpretation of DNA regulatory elements. Tracing the evolution of NLP from its linguistic roots to its current state as 'neural NLP', powered by probabilistic methods and machine learning. It highlights the role of DNA regulatory elements as the 'switches' of the genome and discusses the challenges and breakthroughs in elucidating these elements. The review critically examines the shift towards machine learning NLP based methods in genomic research, determining if this approach may lead to incomplete or incorrect understanding. By delving into historical challenges, recent advancements, and ongoing hurdles, this review aims to provide a comprehensive overview of the state-of-the-art machine learning based approaches in the study of DNA regulatory elements and their potential impact on disease treatment and prevention. In doing so we observed risks regarding the interpretability of computational tools released after 2020, due to the deep-learning trend in the field of machine learning.

Introduction

Natural language processing (NLP) is a field of research aimed at reading, deciphering and assigning meaning to textual data. While it originated in the field of linguistics, it has since become an interdisciplinary subfield of both linguistics and computer science. Herein the computer science aspect can be split into the rule-based and the probabilistic. Where the probabilistic methods have seen a significant increase in popularity due to advancements in the field of machine-learning. The unison between machine learning and NLP is now referred to as 'neural NLP'. These approaches have had increasing success in deciphering more complex texts, such as the genome [1]. For in the simplest sense, the genome is but a 'text' structured in such a way, so that it can facilitate facets required to sustain life.

In order to perform this function of sustaining life, the regulation of the transcription of the genome has shown to be of critical importance. As the interaction between gene transcripts, and their protein products form complex gene regulatory networks (GRNs). These GRNs control gene expression cascades responsible for development of multicellular constructs, such as us, and the cellular interactions that sustain them. They are spatially and temporally adaptive, making sure distinct cell-types fulfil their required function when and where they are supposed to [2]. Due to their fundamental nature, many disease phenotypes can be traced back to the mis-regulation of GRNs, from the loss of DNA repair functionality in cancer [3] to the mis regulation of cerebral layer formation in autism spectrum disorder [4]. GWAS studies have shown that ~ 80% of disease-causing variants occur not in coding but in regulatory regions of the genome [5].

GRNs rely on the interplay between gene activation and repression, whereby the transcription of one gene can influence that of one or more other genes. This regulation of transcription is facilitated by DNA regulatory elements. These elements, which include enhancers, promoters, silencers, insulators, and transcription factor binding sites, function as the 'switches' of the genome, increasing or decreasing the level of gene transcription. They play a pivotal role in determining when, where, and how much of a gene is expressed [6]. The mode of regulation is different between these distinct elements. For instance, enhancers elevate the expression of a gene compared to its level of transcription when the enhancer is absent, while silencers have the inverse effect [7]. The effect of each type of element will be further discussed later.

Elucidation of the exact mechanisms of the interplay between DNA regulatory elements, has historically represented a significant challenge for researchers. Recent technological advancements together with our in-depth understanding of evolutionary principles, applied to artificial neural networks, have enabled the creation of machine learning tools capable of detecting patterns in vast and complex data. These tools have since shown to out-perform experimental approaches, which were historically the predominant way of investigating DNA regulatory elements [8].

Improving our understanding of DNA regulatory elements holds the potential to revolutionize the way we approach disease treatment and prevention. As through understanding genomic linguistics, we might be enabled to edit genomic information to restore cellular function or even create entirely new functionality if so desired [9]. When we consider the immense potential that the understanding of DNA regulatory elements holds, we should ask ourselves if the current shift in our approach of elucidating these regulatory elements, is indeed the optimal one. Or if the current shift towards machine learning NLP based methods has potential pitfalls, leading to an incomplete or even incorrect understanding.

This review seeks to answer that question by delving into the historical challenges encountered in this field, the breakthroughs that surmounted them, and the hurdles that continue to impede state-of-the-art machine learning based approaches.

Background

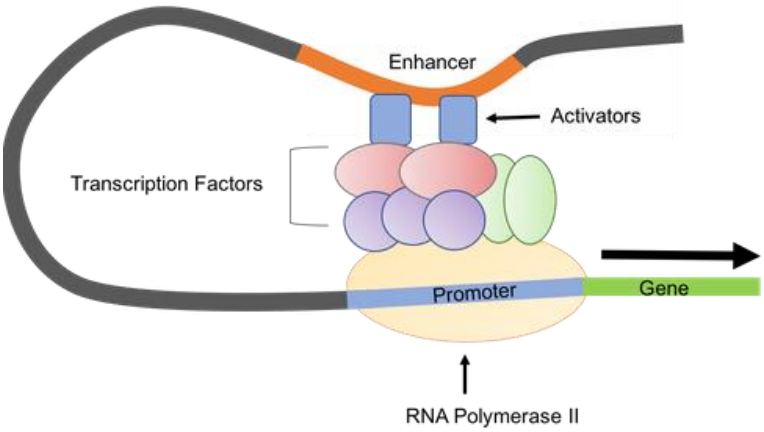
DNA Regulatory Elements

DNA regulatory elements are integral components of the genome that control gene expression. These elements, which include enhancers, promoters, silencers, insulators, repressors, boundary elements and transcription factor binding sites, function as the 'switches' of the genome, increasing or decreasing gene expression. They play a pivotal role in determining when, where, and how much of a gene is expressed, thereby influencing a wide range of biological processes and phenotypic traits.

Regulatory elements were split into two categories, namely cis and trans elements. Here cis-regulatory elements (CREs) are those that have an allele-specific effect, regulating specifically targeted alleles. While trans-regulatory elements (TREs) are regulatory elements trat encode transcription factors, RNA-binding proteins, and non-coding RNAs, which affect regulation of multiple alleles simultaneously. Depending on the TRE product, they can influence regulation at level of the DNA, RNA or protein, ranging from a specific target within the same domain to any number of targets across the cellular neighbourhood [10]. Furthermore, regulatory elements are split in another two, related but clearly distinct, classifications. Namely, proximal, or distal. Proximal regulatory elements are those located near the gene target, generally within a few hundred base pairs of the transcription start site. While distal elements are those located far from the target sequence. These distal elements can be separated from their target genes by millions of base pairs [11]. In the three-dimensional space of the nucleus however, distal elements are often localized in close physical proximity to the gene-proximal regulatory sequences through the formation of chromatin loops [12]. TREs are often automatically classified as distal and CREs as proximal, as the majority of CREs are in close proximity to their gene targets due to their allele-specific effects. While TREs can inherently act at greater distances. However, both proximal TREs and distal CREs have been described, indicating the importance of the distinction between these classifications.

As the scope of this review is limited to DNA regulatory element, the primary focus will be on CREs. These regions of non-coding DNA are split into several sub-classifications, acting together through collaboration of hundreds of transcription factors. Herein, promotor and enhancer elements are those facilitating transcription initiation, these elements share features such as divergent transcription and the modes of transcription factor binding. [13]. Where initiation of transcription through one or the other has been described to influence transcript stability [14]. Conventionally, promoters are considered as the primary site for the assembly of the transcriptional machinery [15]. Promoters are associated with the initiation of RNA synthesis, whereas enhancers stimulate specific promoter activity, providing further specificity in gene regulation through targeting limited core promoter elements, see figure 1 [16,17,18, 19,20].

Fig 1. Representation of the interaction between enhancer and promotor elements. *The proximal promotor facilitates the assembly of the transcriptional machinery. The proximal or distal enhancer stimulates promotor activation*[1].



[1] http://www.cs.ucf.edu/~xiaoman/ET/ 17-12-2023

Besides the initiation and promotion of transcription, DNA regulatory elements responsible for the downregulation and obstruction of transcription Silencers, which, when bound by proteins known as repressors, downregulate the expression of a target gene. They function in a manner opposite to that of enhancers, repressing gene expression by recruiting proteins that disrupt or inactivate the formation of polymerase II transcription complexes at otherwise accessible promoters, see figure 2 [21]. Silencers have been identified to act both distal- and proximally at multiple genes and at the level of chromosomal domains, indicating their broad regulatory impact [22,23]. Similarly, insulators are regulatory elements which can impede the undesired transcription of genes. They serve as boundaries between different regions of the genome. Playing a crucial role in maintaining the integrity of the genome by preventing the interaction between enhancers, silencers and promoters that are not meant to interact. This ensures that genes are correctly regulated and prevents inappropriate gene expression or the silencing thereof, see figure 2 [24]. Furthermore, insulators contribute to the establishment of functionally independent regions within genomes and confer autonomy to each domain.

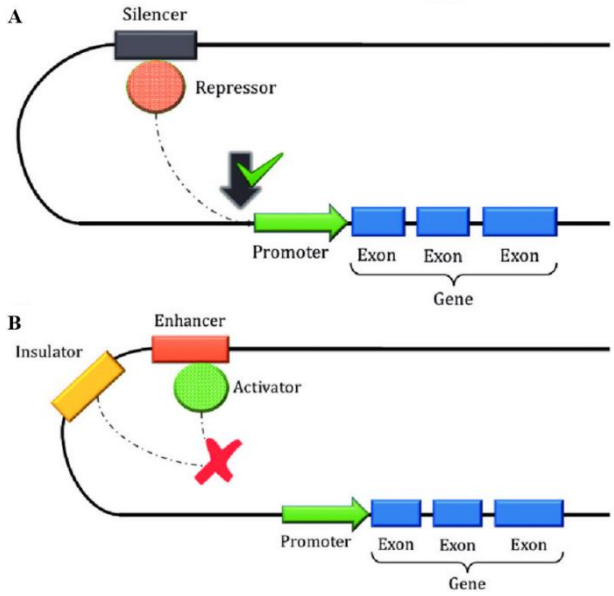


Fig 2. Schematic representation of silencer and insulator functionality. A) *Shows the mode of repressing gene transcription by a silencer element, impeding the formation of the transcriptional machinery at the promotor.* ***B)*** *The insulator element is shown to obstruct the activation stimulation by the enhancer, thus downregulating transcription*[2].

Another important player in the study of regulatory elements is chromatin structure. Chromatin is a complex of DNA and protein found in eukaryotic cells [25]. Its primary function is to package long DNA molecules into more compact, denser structures. By determining which parts of the genome are 'packaged', it defines which regions are aren't accessible, defined as chromatin accessibility. Chromatin accessibility, or the physical access to chromatinized DNA, is a widely studied characteristic of the eukaryotic genome. As active regulatory DNA elements are generally 'accessible', the genome-wide profiling of chromatin accessibility can be used to identify candidate regulatory genomic regions in a tissue or cell type [26].

The precise identification and understanding of regulatory elements are crucial in comprehending the intricate landscape of gene regulation. Kern et al. emphasized that gene regulatory elements are central drivers of phenotypic variation and are critical for understanding the genetics of complex traits [27]. Furthermore, Zhou et al. highlighted the unmasking of thousands of functional cis-regulatory elements integral to transcriptional regulation, such as enhancers and promoters, within the noncoding genome, underscoring their significance in genetic and epigenetic alterations associated with cancer [28]. Heidenreich and Kumar also stressed the impact of genetic alterations within regulatory elements, emphasizing the potential identification of extended therapeutic targets through an understanding of such modifications [29].

[2] https://academic.oup.com/bib/article/20/5/1639/5035219?login=false 27-12-2023

Experimental/biochemical approaches

Due to the limitations in sequencing technologies and the lack of computational assays, DNA regulatory elements relied heavily on experimental approaches until the end of the 20th century. These approaches can be generalized to the induction or observation of a specific perturbation followed by the detection of gene expression changes. These types of analyses were mostly focused on a single or small number of transcription factors, to be able to determine causality. Due the complexity of the regulatory landscape, the detection of regulatory elements was an especially challenging task. [30]. While historical techniques were unable to take into account elements acting from greater distances and the more complicated influence of the regulatory context, they provided valuable insights. Primarily the detection of promotor regions, whose gene relative locus is more conserved. shedding light on the ways gene expression could be regulated. Thus, building the foundation within the field [31].

High throughput assays

With the advent of high-throughput sequencing platforms, the world of DNA-regulatory element detection changed rapidly, enabling comprehensive and systematic analyses of the genome. Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) emerged as a powerful technique for mapping protein-DNA interactions, allowing the genome-wide identification of transcription factor binding sites and histone modifications [32]. Additionally, Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) has enabled the sensitive and rapid profiling of open chromatin regions, providing insights into the interplay between genomic locations of open chromatin, DNA binding proteins, individual nucleosomes, and higher-order compaction at regulatory regions with nucleotide resolution. Gaulton et al identified a map of open chromatin in human pancreatic islets, revealing the presence of cell-selective regulatory domains associated with single genes [33]. Rijnkels et al described interactions between the extracellular matrix and chromatin conformation, suggesting the role of the extracellular matrix as an epigenetic regulator [34]. In summary, high-throughput methods have significantly advanced our understanding of DNA regulatory elements, offering a comprehensive view of the dynamic chromatin landscape and the regulatory networks that govern gene expression.

Computational approaches

Computational approaches, including advanced algorithms and machine learning techniques, have revolutionized the management of the vast data generated by high-throughput techniques. These methods, such as deep learning, have significantly improved the predictive power of DNA-binding proteins by training on larger, higher-quality annotated genomic sequences [35]. Furthermore, integrative methods have been developed to prioritize alterations in regulatory elements and facilitate the systematic analysis of gene regulatory networks. The popularity of computational rendering of high-level gene regulatory networks has led to the development of computational network inference approaches, particularly in the context of high-quality genomes and transcriptomes [36].

Machine learning, particularly deep learning, has been instrumental in predicting the location and function of regulatory elements in large annotated genomic sequences. By training models on a large dataset of genomic sequences with known locations of regulatory elements, these approaches can predict the locations of regulatory elements in new, unannotated genomic sequences by identifying patterns learned during training. This approach allows for the analysis of vast amounts of genomic data and can uncover complex, non-linear relationships between genomic features and regulatory elements. The development of computational strategies has significantly enhanced our ability to manage the vast amounts of data generated by high-throughput approaches, providing valuable insights into the regulatory landscape of the genome and the functional roles of DNA regulatory elements. These advancements have not only improved our understanding of gene regulatory networks but have also paved the way for the identification of key pathways and genes in various biological processes, such as obesity and cancer [37,38].

Historical challenges and breakthroughs in investigating DNA Regulatory Elements

Identification of gene targets

Determining the gene targets of regulatory elements presents significant challenges due to the complex nature of regulatory interactions and the cryptic properties these elements possess. Franco-Zorrilla et al have shown that transcription factors can bind to hundreds or thousands of DNA loci, with only a fraction of targets responding transcriptionally, making it difficult to distinguish relevant binding sites [39]. Additionally, Snetkova & Skok emphasized the variability in the distance separating enhancers and their target promoters, which poses a challenge in determining which elements engage in the regulation of a particular gene [40]. The cell-type specificity and difficulties in characterizing the regulatory targets of these elements further limit the ability to identify causal genetic variants [41]. This phenomenon, referred to as polysemy, describes how a single regulatory element can affect different genes based on its context. Such as the cell type, cellular neighbourhood, or developmental stage. Consequently, the amount of data required to study each element increases, having to consider many contexts to determine gene targets, making its detection and characterization more challenging. Despite these challenges, understanding the polysemy of regulatory elements is crucial for a comprehensive understanding of gene regulation and is a key focus in the field of regulatory element detection [42].

The creation of specialized corpora annotated with regulatory DNA elements has provided valuable resources for training NLP models to disambiguate polysemous terms in the context of DNA regulatory element detection. These corpora have been instrumental in training NLP algorithms to accurately identify and interpret regulatory element-related terms within their specific biological contexts, thereby overcoming the challenges posed by polysemy [43]. Furthermore, the detection of distal regulatory elements and their targets has seen considerable progress. Herein computational tools have been developed with the goal to consider larger sequences and use positional encoding to allow these systems to consider interactions regardless of distance. More on this in the section on ML in NLP.

Elucidation of regulatory interactions

Determining interactions between regulatory elements presents significant challenges due to the complex and dynamic nature of these interactions. Regulatory elements engage in looping interactions that have been implicated in gene regulation, but the precise mechanisms and dynamics of these interactions remain elusive [44]. The surrounding sequence context of interacting regulatory elements is also known to affect local transcriptional output as well as the enhancer and promoter activity of individual elements, adding to the complexity of understanding their interactions. Three-dimensional interaction between regulatory elements is a fundamental process in gene regulation, and questions regarding the structural interaction between regulatory elements can be addressed by analysing interactions between these elements in individual cells [45]. Furthermore, Valuchova et al applied a rapid method for detecting protein-nucleic acid interactions by protein-induced fluorescence enhancement and showed that protein/nucleic acid interactions are further affected by neighbouring proteins, small molecules, and the physical properties of the environment, such as temperature or pH, highlighting the complexity of these interactions [46].

Recent advancements in NLP & ML have played a crucial role in overcoming historical challenges in the field of DNA regulatory element research. By developing innovative NLP techniques and integrating computational approaches, researchers have made considerable progress in accurately interpreting and disambiguating the diverse meanings of regulatory element-related terms, thereby enhancing the precision and reliability of DNA regulatory element detection in the face of polysemy and 3-dimensional/distal interactions. Lee & Rhie highlight the molecular and computational approaches to map regulatory elements in 3D chromatin structure, emphasizing the significance of computational strategies in understanding the complex regulatory landscape [47]. Additionally, Doane & Elemento discuss technical advances and current challenges for the mapping of regulatory elements at the genome-wide scale, underscoring the role of computational methods in uncovering these elements via reconstructing regulatory networks from large genomic datasets [48].

ML in NLP & DNA Regulatory Element detection

Feedforward Neural Network

A feedforward neural network (FNN) is the first of two types of artificial neural networks discussed in this review. They are defined by the one-directional flow of information through the system. Starting at the input layer, followed by a layer of hidden nodes, and finally to the output nodes. There are no cycles or loops in the network. The working of a feedforward neural network involves two phases: the feedforward phase and the backpropagation phase. In the feedforward phase, the input data is fed into the network, and it propagates forward through the network. Once a prediction is made, backpropagation is applied to minimize the error (difference between the predicted output and actual output). This minimizing of the error is achieved through the adjustment of the weights of the hidden neuron layer(s) , see figure 3 [49].

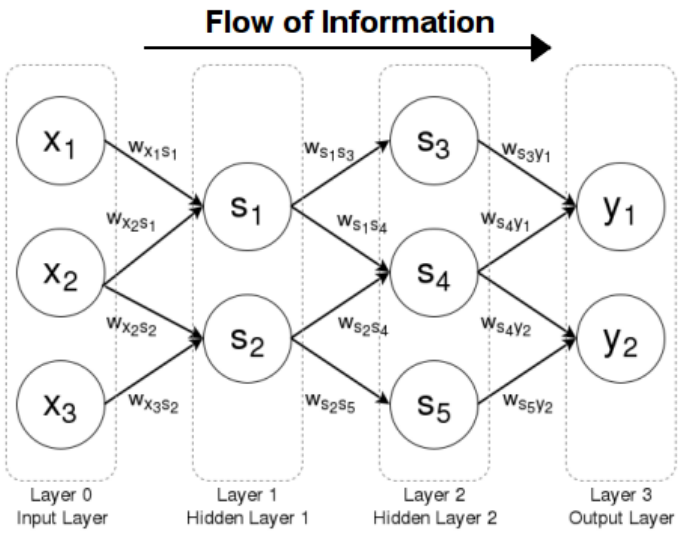


Fig 3. Schematic representation of simple feedforward neural architecture*. X defines the system inputs, W defines the weights applied to the input of the prior layer, determined through backpropagation to produce output y with minimalized error*[3].

Convolutional Neural Network

Convolutional neural networks are a type of feedforward network, primarily applied to image recognition. It improves upon the standard feedforward architecture through introduction of filters to the hidden layers. These filters represent patterns for the network to recognize. Neurons within the network will judge the similarity of their input to given patterns. Subsequent neuron layers of the network take the most relevant output of the previous layer to create more complicated filters. By increasing the number of hidden neuron layers, convolutional models are able to recognize more complicated patterns [50,51]. This topic will be further discussed under the term coined for the application of multiple/many hidden neuron layers, 'deep learning'.

Besides image recognition, convolutional networks were among the first with the ability to process temporal data such as speech and text, thus having created the first artificial network capable of natural language processing. Collobert & Weston defined a general convolutional network architecture and described its application to various natural language processing (NLP) tasks, including part-of-speech

[3] *(CSE 415 –(c) S.Tanimoto, 2002) 27-12-2023*

tagging, chunking, named-entity recognition, learning a language model, and semantic role-labeling. This demonstrates the early utilization of convolutional networks in processing textual data [52]. Word embedding, the act by which words are translated into vectors, significantly decreased the computational power and time required to train these networks. On top of this, word embedding allows text with similar meanings to be represented by similar vectors. These advancements have significantly enhanced the efficacy of CNN [53].

Standard CNN have significant limitations in the context of identifying DNA regulatory elements using NLP. This is due to the filter size used for the convolution, which is only able to consider the information in the immediate vicinity of the input [54]. While CNNs have shown state-of-the-art predictions for transcription factor binding and DNA accessibility, their application in identifying regulatory elements is hindered by the inability to capture the long-range interactions and variable sizes of these elements [55,56].

Transformer

The transformer architecture facilitated a significant increase in the efficacy of ML models in the context of DNA regulatory element detection. It is a type of feedforward neural network architecture, now including positional encoders, which complement word vectors by generating a vector to represent distance between input sequences [57]. Subsequently the word vectors with this positional information are passed to an attention layer and a feedforward layer, which together make up an encoder block. The attention layer enables the model to emphasize distinct parts of the input sequence regardless of distance, thereby enabling the model to consider dependencies between sequences regardless of their positions in the input sequence [58].

Through the application of encoder and decoder blocks, this type of neural network is enabled to translate information types between layers, such as positional information of regulatory sequences to the association between regulatory elements. The encoder consists of 2 parts: the self-attention mechanism and the position-wise feed-forward network. The self-attention mechanism allows different positions in a single sequence to interact to compute a representation of that sequence with contextual information. Thus, allowing this architecture to weigh the importance of parts of a sequence, such as regulatory elements. The output of the self-attention mechanism is fed to the position-wise feedforward layer. In which each position of the input sequence has its own path to this feedforward layer, allowing them to be considered individually. The combination of these parts of the encoder allows for understanding the context and relationships in the data, see figure 4 [59].

The decoder in the transformer architecture comprises two attention layers and a feedforward layer. The secondary attention layer incorporates the attention weights from the encoder and its preceding attention layer. This design empowers the network to perform sequence-to-sequence translation tasks, such as converting a DNA sequence into a probability vector representing sequences linked to promoter regions [60].

Recurrent Neural Networks

As opposed to feedforward architectures, Recurrent neural networks include feedback loops, allowing them to consider previous inputs when processing current ones, see figure 4 [61,62]. Which makes the network behave more dynamically, making them better suitable for modeling time series data and sequential information [63]. The performance of these networks exceeded that of the (multilayer) feedforward architectures in the context of sequence analysis. However, due to the sequential nature of the network, it has been shown that these networks do not handle long-term dependencies very well. This limitation is particularly significant in tasks such as natural language processing, speech recognition, and time series analysis, where understanding and retaining long-range dependencies is crucial [64].

Variants of the traditional Recurrent neural networks such as 'Long Short-Term Memory' and 'Gated Recurrent Units' have been designed to circumvent the vanishing gradient problem. These variants have demonstrated superior performance in capturing and retaining valuable information over extended sequences, improving their suitability for tasks involving sequential data analysis [65].

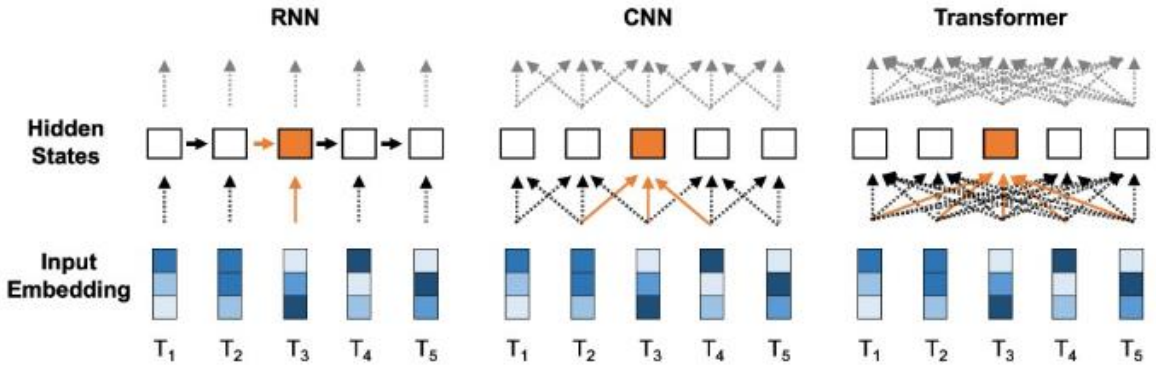


Fig 4. Overview of differences between Recurrent, Convolutional and Transformer Neural Networks[4].

[4] https://academic.oup.com/bioinformatics/article/37/15/2112/6128680 27-12-2023

Deep learning

Deep learning is a branch of machine learning that has revolutionized many aspects of research, industry, The "depth" of these models comes from the multitude of hidden neuron layers that allow them to learn hierarchical representations, with lower-level features (like lines and edges in image processing) being combined and abstracted in higher layers to detect more complex concepts (like shapes or objects). This depth enables deep learning models to perform tasks that require an understanding of complex patterns and relationships in data, such as image recognition, natural language processing. The inclusion of many hidden neuron layers enables deep learning architecture to process more complex input data, not requiring feature extraction, such as featurization using a database of known motifs. Without requiring pre-processing of input data, the system is enabled to detect relations without requiring prior knowledge, see figure 5 [66].

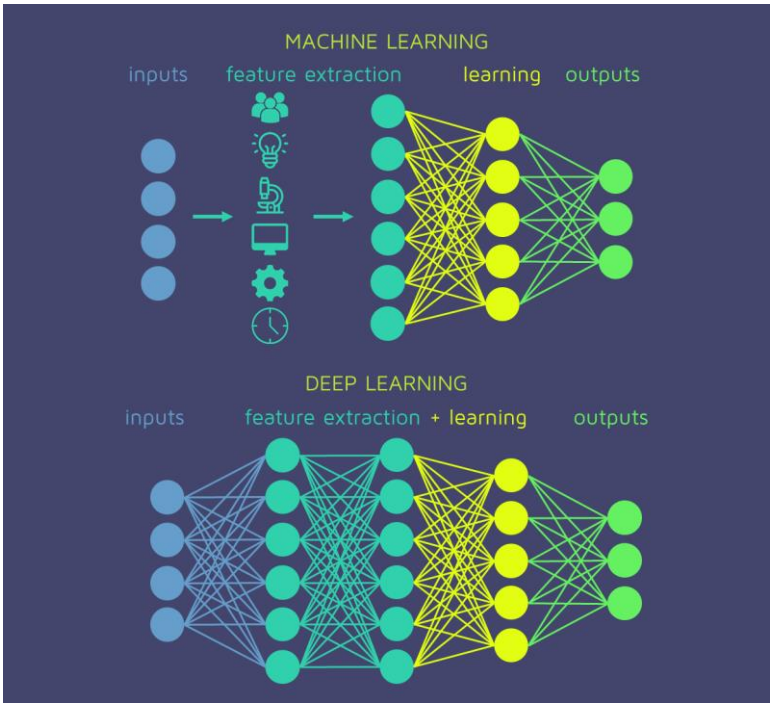


Fig 5. Schematic representation of the difference between 'standard' machine learning and Deep learning. *A primary difference between classical machine learning algorithms is that classical architectures often require structured data, while deep learning models can work with unstructured data such as images, audio and text* [5].

The complexity of genomic regulation, as evidenced by the intricate structures observed in gene regulatory networks (GRNs), underscores the importance of unbiased detection of all possible relationships within input sequences to enhance the efficacy of deep learning models GRNs constructed from gene expression data reflect the interactions of regulatory elements in biological systems, revealing the inner complex mechanisms that drive adaptability to the environment and the growth and development of organisms [67]. However, deep learning architectures come with several drawbacks. For instance, they are more computationally intensive and require large amounts of data and time to train. Furthermore, deep learning models lack the interpretability of classical machine learning models, meaning that it is easier to see why a model made a certain prediction. This drawback is often overlooked when these models are applied and generate accurate and functional predictions. These predictions are widely applied without consideration of the underlying reasoning, for instance. Can we truly consider a field to be understood when much of the knowledge is obtained using indescribable reasoning? Molnar et al emphasized the need for tools for model-agnostic interpretability methods to improve the adoption of machine learning, indicating that interpretability is a recognized challenge in the field of machine learning [68]. Similarly, Chen & Deng highlighted the

[5] https://medium.com/@rishithakurr/a-practical-use-case-of-machine-learning-in-amazon-b0f2249ccf7c 27-12-2023

lack of interpretability in deep learning models, despite their superior performance in predictions in many fields, suggesting that this drawback is a significant concern, particularly in practical applications [69].

As of writing this review, Deep CNN have shown to outperform other architectures in predicting gene expression and the identification of DNA regulatory elements, in both human and mouse genomes [70]. O'Donovan et al demonstrated that CNNs significantly outperformed classical machine learning techniques in predicting the time series of gene expression in primary human hepatocytes given a measured time series of gene expression from primary rat hepatocytes following exposure to a previously unseen compound [70]. Additionally, Trabelsi et al found that deeper and more complex architectures, particularly hybrid CNN/RNN architectures, outperformed other methods in terms of accuracy for predicting DNA/RNA sequence binding specificities [71]. Furthermore, Yuan & Bar-Joseph showed that their method, CNN for co-expression, improved upon prior methods in tasks ranging from predicting transcription factor targets to identifying disease-related genes to causality inference [72]. However, it is important to note that not all studies support the claim. Elbashir et al highlighted that the availability of gene expression data influences prediction accuracy of deep-learning models. Where a small number of samples and a relatively large number of dimensions, may not be suitable for deep CNN architectures [73].

Recent advancements

The field of NLP and ML has been rapidly evolving and shows no signs of slowing down. In this period of rapid advancements, we aim to elucidate the ways ML based tools for the detection of DNA regulatory elements have transformed recent history. Based on the evolution of these tools we attempt to chart the path in which the field is expected to develop, which problems are likely to be overcome and which hurdles might remain.

Starting with 'DNABert', a machine learning based tool developed in 2021. It was developed to satisfy three properties put forward by its inventors, which are considered required for the ideal computational method to detect regulatory elements; namely, (i) The ability to globally take all the contextual information into account to distinguish polysemous CREs; (ii) develop generic understanding transferable to various tasks; (iii) generalize well when labelled data is limited [74].

Their approach to satisfy these requirements saw them adapt a bi-directional encoder transformer architecture. Pre-training was performed on unlabelled 'human genome data obtained through direct non-overlap splitting and random sampling'. DNA-sequences with a max length of 512bp were tokenized using a k-mer representation. This is done to enrich the contextual information the model can consider. The model was then fine-tuned for the prediction of promotors, transcription factor binding sites and splice sites. Furthermore, this tool includes direct visualization of relevancy of nucleotides to regulatory elements and semantic relationship within sequences [75]. While this visual representation increases the interpretability of the results, this model still contains a large 'black box' as the mechanisms by which predictions are made are left ambiguous due to the many hidden layers included in the network.

In 2021 'DeepRegFinder' was released. As the name implies this is another deep learning architecture. It aims to further improve the efficacy of DNA regulatory elements by focusing on the detection of both promotors and enhancers. While most promotor elements are considered to have been identified prior, the same cannot be said for enhancer elements, whose mode of regulation has obstructed their identification. Namely, the potential distance from the transcription start site, which can be up to 1Mb [76]. Deep learning architectures such as DeepRegFinder', are not reliant on labelled data and perform their own feature extraction, outperforming prior approaches in detection of enhancer regions. Herein DeepRegFinder has been applied to identify 7,796 putative enhancer sites [76]. To avoid false positive

predictions, this model applied mean average precision instead of the AUC score, which have been described to be more meaningful interpretable [77]. The input flexibility, extensive pre-training, and ability to use different network architectures make this model very accessible. In 2022 the predictive abilities across diverse biological contexts were further enhanced by the release of 'Sei'. The model encompasses an estimated 1,000 nonhistone DNA-binding proteins, 77 histone marks, and chromatin accessibility across more than 1,300 cell lines and tissues. This comprehensive model aims to associate specific sequences or variants to regulatory activities, predicting 21,907 distinct chromatin profiles. The predictive abilities of Sei are further reinforced by its integration of tissue-specific expression, expression quantitative trait loci, and evolutionary constraint data [78].

More recently, in 2023 to be exact, an updated version of the DNABert model was released, the aptly named 'DNABert-2' represents a significant advancement over the original DNABert model. Replacing the k-mer tokenization with byte pair encoding, leveraging a compression algorithm commonly used by large language models. This approach aims to circumvent a known issue of k-mer tokenization, namely inferring semantic relationships. K-mer tokenization breaks down input text into fixed-length subsequences of characters, which may not always align with meaningful linguistic units, effectively limiting the consideration of the sequence context. Byte Pair Encoding replaces k-mer tokenization by employing a data-driven subword tokenization algorithm, which dynamically identifies and encodes frequent sequences of characters within the input sequence. This approach effectively addresses some of the limitations associated with k-mer tokenization and offers an enhanced representation of semantic relationships [79]. DNABert-2 further improves on its predecessor by getting rid of input length limitations. This is done by incorporating flash attention instead of positional embedding. In doing so, the model achieves state-of-the-art performance while simultaneously reducing computational cost 56-fold and the number of parameters 21-fold [79].

The same issues regarding k-mer tokenization were observed and tackled by a deep-transformer architecture called 'HyenaDNA', released in 2023. This model applies a single-character tokenizer instead of byte pair encoding. Effectively increasing context length to one million tokens at the single nucleotide level, or in other words a context of a million bases. Moreover, HyenaDNA has demonstrated its remarkable ability to process context lengths of up to 131,000 tokens with a shallow architecture comprising only two hidden layers, showcasing its efficiency in handling extensive genomic data without relying on deep-learning architectures [80]. However, while these results give an impressive indication of the abilities of non-deep architectures, they are still easily outcompeted by their deep counterparts. Particularly due to knowledge that context at a million bases is relevant to transcriptional regulation.

It should be noted that all the models discussed make nucleotide-level importance predictions, which enhance their interpretability. However, such predictions are akin to hypotheses proposed by the model, which require testing to confirm and convey a deeper understanding of the prediction itself. Effectively moving, not solving, the interpretability issue. Tools to evaluate these nucleotide level importance predictions at the scale of the genome, such as DeepLIFT and TF-MoDISco have been developed. They aim to cluster motifs with elevated importance predictions. Through this clustering it becomes possible to attach meaning to predictions made by deep-learning architectures at the motif level [81,82].

Conclusion

Through our exploration of regulatory element detection, the historical challenges, breakthroughs, and innovative methods used to predict them, we've gained insight into the progression of the field. We observed that computational models are actively being enhanced to cope with the inherent complexity of regulatory elements. Particularly the influence of specific cellular, sequence and developmental context which influences regulatory aspects. Along with the variety in distal and proximal regulation, where some regions of the DNA form functionally independent regions, due to insulators and chromatin accessibility. While other elements are known to interact across millions on bases, taking advantage of the 3-dimensional structure of the DNA.

With the rapid increase in computing power and the availability of extensive training data, we have seen the development of numerous tools designed to tackle these challenges. These tools leverage advanced machine learning techniques and vast genomic datasets to predict the location and functionality of regulatory elements with increasing accuracy. Deep learning models have shown great promise in this area. They are capable of handling the high-dimensional data, typical in genomics, and can model complex relationships, which are essential for understanding the polysemous nature of regulatory elements [83, 84]. However, while capable of complex data processing, deep-learning models, particularly those with many layers or neurons, are challenging to interpret. Thus, while it is expected we will accurately map the locations and functions of regulatory elements to a relatively complete degree within the next decade, we could still be met with an era of trial and error. For instance, when we have fully charted the regulatory landscape and wish to use that knowledge to bring about specific functionality in the lab, our deeper understanding of DNA regulatory elements will be required and possibly absent due to this black-box effect of deep learning models [85]. When this understanding would be absent, our primary option to achieve such a goal, would be to act according to an approach put forward by a deep architecture, following its instructions to create the desired functionality. Having no or minimal indication of the efficacy of the proposed solution until it is attempted, in other words, an era of trial and error [86].

However, this view can, and hopefully will be, be considered short-sighted, as efforts are being made to develop tools with increased interpretability and separate tools to untangle the black-box aspects of deep-learning architectures. For instance, TF-MoDISco, a tool aimed at clustering motifs based on nucleotide level importance predictions. Allowing for functional annotation of motif clusters. These hypothesized functional clusters can be tested experimentally or computationally [87]. However, it should be noted that polysemy will necessitate significant data variability, as motif clusters are context dependent [88].

Regardless of these concerns, deep learning architectures have demonstrated their power in the identification of regulatory elements, particularly in situations with abundant ChIP-seq, ATAC-seq, DNase-seq, gene expression, and methylation data [89]. We do not expect non-deep architectures to compete with the efficacy of these models, particularly in cases where substantial data is available. Which is the case for many of regulatory element relevant data types such as ChIP-seq, ATAC-seq, DNase-seq, gene expression and methylation data. The vast amount of data that is currently available is expected to increase, improving the accuracy of prediction in distinct contexts such as cellular, differential, and developmental states. Which are known to be critical to the functioning of regulatory elements.

In summary, deep learning architectures have proven to be invaluable tools in the study of regulatory elements, leveraging the abundance of available data to improve prediction accuracy in diverse biological contexts. However, there is a need for further research into experimental/computational confirmation of their predictions. We expect a deeper understanding of transcription regulation to have a significant positive effect on related fields. Particularly our understanding of disease mechanisms, as GWAS studies have found a significant correlation between disease phenotypes and mutations in non-coding regions. Where this understanding has the potential to improve disease treatment, prevention and even eradication.

References

1) Cho et al. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation" (2014) doi:10.3115/v1/d14-1179 publication type: other, topics: computer science, decodes Furthermore, many recent works showed that neural networks can be successfully used in a number of tasks in natural language processing (NLP)
2) Akira ISHIHAMA, Prokaryotic genome regulation: A revolutionary paradigm, Proceedings of the Japan Academy, Series B, 2012, Volume 88, Issue 9, Pages 485-508, Released on J-STAGE November 09, 2012, Online ISSN 1349-2896, Print ISSN 0386-2208, https://doi.org/10.2183/pjab.88.48
3) Zickenrott, S., Angarica, V., Upadhyaya, B. et al. Prediction of disease–gene–drug relationships following a differential network analysis. Cell Death Dis 7, e2040 (2016). https://doi.org/10.1038/cddis.2015.393
4) Kao, A., McKay, A., Singh, P. et al. Progranulin, lysosomal regulation and neurodegenerative disease. Nat Rev Neurosci 18, 325–333 (2017). https://doi.org/10.1038/nrn.2017.36
5) Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., Hansen, R. S., Neph, S., Sabo, P. J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S. R., Kaul, R., & Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science, 337(6099), 1190-1195. doi:10.1126/science.12227941
6) Weng, Q., Xing, J., Li, Z., Dong, Z., & Dong, J. (2010). Expression analysis of rus1 and construction of rus1 plant expressing vector. Frontiers of Agriculture in China, 4(1), 31-36. https://doi.org/10.1007/s11703-010-0099-6
7) Whalen, S., Truty, R., & Pollard, K. S. (2016). Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. Nature Genetics, 48(5), 488-496. https://doi.org/10.1038/ng.3539
8) Pennacchio, L. A., Bickmore, W. A., Dean, A., Nóbrega, M. A., & Bejerano, G. (2013). Enhancers: five essential questions. Nature Reviews Genetics, 14(4), 288-295. https://doi.org/10.1038/nrg3458
9) Hojo, H., & Ohba, S. (2019). Insights into Gene Regulatory Networks in Chondrocytes. International Journal of Molecular Sciences, 20(24), 6324. doi:10.3390/ijms20246324
10) Gilad, Y., Rifkin, S. A., & Pritchard, J. K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. Trends in genetics : TIG, 24(8), 408–415. https://doi.org/10.1016/j.tig.2008.06.001
11) Snetkova V, Skok JA. Enhancer talk. Epigenomics. 2018 Apr 1;10(4):483-498. doi: 10.2217/epi-2017-0157. Epub 2018 Mar 27. PMID: 29583027; PMCID: PMC5925435.
12) Khader N, Shchuka VM, Shynlova O, Mitchell JA. Transcriptional control of parturition: insights from gene regulation studies in the myometrium. Mol Hum Reprod. 2021 May 8;27(5):gaab024. doi: 10.1093/molehr/gaab024. PMID: 33823545; PMCID: PMC8126590.
13) Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., & Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature genetics, 46(12), 1311–1320. https://doi.org/10.1038/ng.3142
14) Core, L. J., Waterfall, J. J., & Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science (New York, N.Y.), 322(5909), 1845–1848. https://doi.org/10.1126/science.1162228
15) Hong, C. K. Y., & Cohen, B. A. (2022). Genomic environments scale the activities of diverse core promoters. Genome research, 32(1), 85–96. https://doi.org/10.1101/gr.276025.121
16) Mikhaylichenko, O., Bondarenko, V., Harnett, D., Schor, I. E., Males, M., Viales, R. R., & Furlong, E. E. M. (2018). The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. Genes & development, 32(1), 42–57. https://doi.org/10.1101/gad.308619.117
17) (2015). Enhancers, enhancers – from their discovery to today's universe of transcription enhancers. biological chemistry, 396(4), 311-327. https://doi.org/10.1515/hsz-2014-0303
18) (2009). The interplay between transcription factors and micrornas in genome‐scale regulatory networks. bioessays, 31(4), 435-445. https://doi.org/10.1002/bies.200800212
19) (2012). The control of histone gene expression. biochemical society transactions, 40(4), 880-885. https://doi.org/10.1042/bst20120065
20) (2012). Control of rene gene expression. pflügers archiv - european journal of physiology, 465(1), 13-21. https://doi.org/10.1007/s00424-012-1110-2
21) Qi, T., Wu, Y., Zeng, J. et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. Nat Commun 9, 2282 (2018). https://doi.org/10.1038/s41467-018-04558-1
22) Pang, B., & Snyder, M. P. (2020). Systematic identification of silencers in human cells. Nature genetics, 52(3), 254–263. https://doi.org/10.1038/s41588-020-0578-5
23) Li, Y., Cheng, T., & Gartenberg, M. R. (2001). Establishment of transcriptional silencing in the absence of dna replication. Science, 291(5504), 650-653. https://doi.org/10.1126/science.291.5504.650

24) Raab, J., Kamakaka, R. Insulators and promoters: closer than we think. Nat Rev Genet 11, 439–446 (2010). https://doi.org/10.1038/nrg2765
25) Mondal, T., Rasmussen, M., Pandey, G. K., Isaksson, A., & Kanduri, C. (2010). Characterization of the RNA content of chromatin. Genome research, 20(7), 899–907. https://doi.org/10.1101/gr.103473.109
26) Minnoye, L., Marinov, G.K., Krausgruber, T. et al. Chromatin accessibility profiling methods. Nat Rev Methods Primers 1, 10 (2021). https://doi.org/10.1038/s43586-020-00008-9
27) Kern, C., Wang, Y., Xu, X. et al. Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. Nat Commun 12, 1821 (2021). https://doi.org/10.1038/s41467-021-22100-8
28) Zhou, S., Treloar, A. E., & Lupien, M. (2016). Emergence of the Noncoding Cancer Genome: A Target of Genetic and Epigenetic Alterations. Cancer discovery, 6(11), 1215–1229. https://doi.org/10.1158/2159-8290.CD-16-0745
29) Heidenreich, B. and Kumar, R. (2017), Altered TERT promoter and other genomic regulatory elements: occurrence and impact. Int. J. Cancer, 141: 867-876. https://doi.org/10.1002/ijc.30735
30) Elemento, O., Slonim, N., & Tavazoie, S. (2007). A universal framework for regulatory element discovery across all genomes and data types. Molecular Cell, 28(2), 337-350. https://doi.org/10.1016/j.molcel.2007.09.027
31) Identification of Essential cis-Acting Regulatory Elements for Transcription of the Rat DAN Gene; TOSHINORI OZAKI and SHIGERU SAKIYAMA, DNA and Cell Biology 1997 16:6, 779-786
32) Ji, H., Jiang, H., Ma, W. and Wong, W.H. (2011), Using CisGenome to Analyze ChIP-chip and ChIP-seq Data. Current Protocols in Bioinformatics, 33: 2.13.1-2.13.45. https://doi.org/10.1002/0471250953.bi0213s33
33) Gaulton, K., Nammo, T., Pasquali, L. et al. A map of open chromatin in human pancreatic islets. Nat Genet 42, 255–259 (2010). https://doi.org/10.1038/ng.530
34) Rijnkels, M., Kabotyanski, E., Montazer-Torbati, M.B. et al. The Epigenetic Landscape of Mammary Gland Development and Functional Differentiation. J Mammary Gland Biol Neoplasia 15, 85–100 (2010). https://doi.org/10.1007/s10911-010-9170-4
35) Ahmed, S.F., Alam, M.S.B., Hassan, M. et al. Deep learning modelling techniques: current progress, applications, advantages, and challenges. Artif Intell Rev 56, 13521–13617 (2023). https://doi.org/10.1007/s10462-023-10466-8
36) Zarayeneh, N., Ko, E., Oh, J. H., Suh, S., Liu, C., Gao, J., Kim, D., & Kang, M. (2019). Integration of multi-omics data for integrative gene regulatory network inference. International Journal of Data Mining and Bioinformatics, 20(24), 6324. doi:10.3390/ijms20246324
37) Joshi, H., Vastrad, B., Joshi, N., Vastrad, C., Tengli, A., & Kottursheti, I. (2021). Identification of Key Pathways and Genes in Obesity Using Bioinformatics Analysis and Molecular Docking Studies. Frontiers in Endocrinology, 12. doi:10.3389/fendo.2021.628907
38) Jung HC, Kim SH, Lee JH, Kim JH, Han SW. Gene Regulatory Network Analysis for Triple-Negative Breast Neoplasms by Using Gene Expression Data. J Breast Cancer. 2017 Sep;20(3):240-245. https://doi.org/10.4048/jbc.2017.20.3.240
39) Franco-Zorrilla, J. M., López-Vidriero, I., Carrasco, J. L., Godoy, M., Vera, P., & Solano, R. (2014). DNA-binding specificities of plant transcription factors and their potential to define target genes. Proceedings of the National Academy of Sciences of the United States of America, 111(6), 2367-2372. doi:10.1073/pnas.1316278111
40) Snetkova, V., & Skok, J. A. (2018). Enhancer talk. Epigenomics, 10(4). doi:10.2217/epi-2017-0157
41) Song, M., Yang, X., Ren, X. et al. Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. Nat Genet 51, 1252–1262 (2019). https://doi.org/10.1038/s41588-019-0472-1
42) (2021). Co-regulated genes and gene clusters. genes, 12(6), 907. https://doi.org/10.3390/genes12060907
43) Jin, X., Zhang, D., Zhu, H., Xiao, W., Li, S., Wei, X., Arnold, A., & Ren, X. (2022). Lifelong Pretraining: Continually Adapting Language Models to Emerging Corpora 2021
44) Sanyal, A., Lajoie, B., Jain, G. et al. The long-range interaction landscape of gene promoters. Nature 489, 109–113 (2012). https://doi.org/10.1038/nature11279
45) Oudelaar, A.M., Davies, J.O.J., Hanssen, L.L.P. et al. Single-allele chromatin interactions identify regulatory hubs in dynamic compartmentalized domains. Nat Genet 50, 1744–1751 (2018). https://doi.org/10.1038/s41588-018-0253-2
46) Valuchova, S., Fulnecek, J., Petrov, A. et al. A rapid method for detecting protein-nucleic acid interactions by protein induced fluorescence enhancement. Sci Rep 6, 39653 (2016). https://doi.org/10.1038/srep39653
47) Lee, B.H., Rhie, S.K. Molecular and computational approaches to map regulatory elements in 3D chromatin structure. Epigenetics & Chromatin 14, 14 (2021). https://doi.org/10.1186/s13072-021-00390-y
48) Doane, A.S. and Elemento, O. (2017), Regulatory elements in molecular networks. WIREs Syst Biol Med, 9: e1374. https://doi.org/10.1002/wsbm.1374
49) Baldi, P., & Vershynin, R. (2019). The capacity of feedforward neural networks. Neural networks : the official journal of the International Neural Network Society, 116, 288–311. https://doi.org/10.1016/j.neunet.2019.04.009
50) Vaz JM, Balaji S. Convolutional neural networks (CNNs): concepts and applications in pharmacogenomics. Mol Divers. 2021 Aug;25(3):1569-1584. doi: 10.1007/s11030-021-10225-3. Epub 2021 May 24. PMID: 34031788; PMCID: PMC8342355.
51) Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. Insights into imaging, 9(4), 611–629. https://doi.org/10.1007/s13244-018-0639-9

52) Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: deep neural networks with multitask learning. Proceedings of the 25th International Conference on Machine Learning
53) Asudani DS, Nagwani NK, Singh P. Impact of word embedding models on text analytics in deep learning environment: a review. Artif Intell Rev. 2023 Feb 22:1-81. doi: 10.1007/s10462-023-10419-1. Epub ahead of print. PMID: 36844886; PMCID: PMC9944441.
54) Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., & Mirny, L. A. (2020). Predicting 3D genome folding from DNA sequence with Akita. Nature chemical biology, 16(2), 198-204. doi:10.1038/s41592-020-0958-x
55) SENIES: DNA Shape Enhanced Two-layer Deep Learning Predictor for the Identification of Enhancers and Their Strength" (2021) doi:10.1101/2021.05.14.444093
56) Wang, J., Long, Q., Li, Y., Nguyen, T., Basith, S., & Zhou, H. (2022). Genome-wide identification and characterization of DNA enhancers with a stacked multivariate fusion framework. Plos computational biology. https://doi.org/10.1371/journal.pcbi.1010779
57) Ravanmehr, V., Blau, H., Cappelletti, L., Fontana, T., Carmody, L., Coleman, B., George, J., Reese, J., Joachimiak, M., Bocci, G., Hansen, P., Bult, C., Rueter, J., Casiraghi, E., Valentini, G., Mungall, C., Oprea, T. I., & Robinson, P. N. (2021). Supervised learning with word embeddings derived from PubMed captures latent knowledge about protein kinases and cancer. NAR genomics and bioinformatics, 3(4), lqab113. https://doi.org/10.1093/nargab/lqab113
58) Transformers: State-of-the-Art Natural Language Processing; (https://aclanthology.org/2020.emnlp-demos.6) (Wolf et al., EMNLP 2020)
59) Asudani DS, Nagwani NK, Singh P. Impact of word embedding models on text analytics in deep learning environment: a review. Artif Intell Rev. 2023 Feb 22:1-81. doi: 10.1007/s10462-023-10419-1.
60) Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 6000–6010). Curran Associates Inc.
61) Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2016). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. arXiv preprint arXiv:1502.03044.
62) Connor, J. T., Martin, R. D., & Atlas, L. E. (1994). Recurrent neural networks and robust time series prediction. IEEE transactions on neural networks, 5(2), 240–254. https://doi.org/10.1109/72.279188
63) John JN, Sid E, Zhu Q. Recurrent Neural Networks to Automatically Identify Rare Disease Epidemiologic Studies from PubMed. AMIA Jt Summits Transl Sci Proc. 2021 May 17;2021:325-334. PMID: 34457147; PMCID: PMC8378621.
64) Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. Neural computation, 31(7), 1235–1270. https://doi.org/10.1162/neco_a_01199
65) Schmidhuber J. (2015). Deep learning in neural networks: an overview. Neural networks : the official journal of the International Neural Network Society, 61, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003
66) Xing, L., Guo, M., Liu, X. et al. An improved Bayesian network method for reconstructing gene regulatory network based on candidate auto selection. BMC Genomics 18 (Suppl 9), 844 (2017). https://doi.org/10.1186/s12864-017-4228-y
67) Molnar et al., (2018). iml: An R package for Interpretable Machine Learning . Journal of Open Source Software, 3(26), 786, https://doi.org/10.21105/joss.00786
68) Chen, H., Deng, W. Interpretable patent recommendation with knowledge graph and deep learning. Sci Rep 13, 2586 (2023). https://doi.org/10.1038/s41598-023-28766-y
69) Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., & Troyanskaya, O. G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nature genetics, 50(8), 1171–1179. https://doi.org/10.1038/s41588-018-0160-6
70) O'Donovan SD, Driessens K, Lopatta D, Wimmenauer F, Lukas A, et al. (2020) Use of deep learning methods to translate drug-induced gene expression changes from rat to human primary hepatocytes. PLOS ONE 15(8): e0236392. https://doi.org/10.1371/journal.pone.0236392
71) Ameni Trabelsi, Mohamed Chaabane, Asa Ben-Hur, Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities, Bioinformatics, Volume 35, Issue 14, July 2019, Pages i269–i277, https://doi.org/10.1093/bioinformatics/btz339
72) Yuan, Y., & Bar-Joseph, Z. (2019). Deep learning for inferring gene relationships from single-cell expression data. Proceedings of the National Academy of Sciences, 116(52), 26543-26549. doi:10.1073/pnas.1911536116
73) M. K. Elbashir, M. Ezz, M. Mohammed and S. S. Saloum, "Lightweight Convolutional Neural Network for Breast Cancer Classification Using RNA-Seq Gene Expression Data," in IEEE Access, vol. 7, pp. 185338-185348, 2019, doi: 10.1109/ACCESS.2019.2960722.
74) Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics, 37(15), 2112–2120. https://doi.org/10.1093/bioinformatics/btab083
75) Naranjo, S., Voesenek, K., de la Calle-Mustienes, E., Robert-Moreno, A., Kokotas, H., Grigoriadou, M., Economides, J., Van Camp, G., Hilgert, N., Moreno, F., Alsina, B., Petersen, M. B., Kremer, H., & Gómez-Skarmeta, J. L. (2010). Multiple enhancers located in a 1-Mb region upstream of POU3F4 promote expression

during inner ear development and may be required for hearing. Human genetics, 128(4), 411–419. https://doi.org/10.1007/s00439-010-0864-x
76) Neiro, J., Sridhar, D., Dattani, A., & Aboobaker, A. (2022). Identification of putative enhancer-like elements predicts regulatory networks active in planarian adult stem cells. eLife, 11, e79675. https://doi.org/10.7554/eLife.79675
77) Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432
78) Chen, K. M., Wong, A. K., Troyanskaya, O. G., & Zhou, J. (2022). A sequence-based global map of regulatory activity for deciphering human genetics. Nature genetics, 54(7), 940–949. https://doi.org/10.1038/s41588-022-01102-2
79) Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics (Oxford, England), 37(15), 2112–2120. https://doi.org/10.1093/bioinformatics/btab083
80) Nguyen, E., Poli, M., Faizi, M., Thomas, A., Birch-Sykes, C., Wornow, M., Patel, A., Rabideau, C., Massaroli, S., Bengio, Y., Ermon, S., Baccus, S. A., & Ré, C. (2023). HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. ArXiv, arXiv:2306.15794v2.
81) Smith, G. D., Ching, W. H., Cornejo-Páramo, P., & Wong, E. S. (2023). Decoding enhancer complexity with machine learning and high-throughput discovery. Genome biology, 24(1), 116. https://doi.org/10.1186/s13059-023-02955-4
82) Minnoye, L., Taskiran, I. I., Mauduit, D., Fazio, M., Van Aerschot, L., Hulselmans, G., Christiaens, V., Makhzami, S., Seltenhammer, M., Karras, P., Primot, A., Cadieu, E., van Rooijen, E., Marine, J. C., Egidy, G., Ghanem, G. E., Zon, L., Wouters, J., & Aerts, S. (2020). Cross-species analysis of enhancer logic using deep learning. Genome research, 30(12), 1815–1834. https://doi.org/10.1101/gr.260844.120
83) Alipanahi, B., Delong, A., Weirauch, M. et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 33, 831–838 (2015). https://doi.org/10.1038/nbt.3300
84) Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome research, 26(7), 990–999. https://doi.org/10.1101/gr.200535.115
85) Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. Briefings in bioinformatics, 18(5), 851–869. https://doi.org/10.1093/bib/bbw068
86) Lanchantin, J., Singh, R., & Wang, B. (2016). Deep Motif Dashboard: Visualizing and Understanding Genomic Sequences Using Deep Neural Networks. In Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (pp. 25-32). doi:10.1142/9789813207813_0025
87) Shrikumar, A. (2018). Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5. arXiv preprint arXiv:1811.00416.
88) Izabella Krystkowiak, Norman E. Davey, SLiMSearch: a framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions, Nucleic Acids Research, Volume 45, Issue W1, 3 July 2017, Pages W464–W469, https://doi.org/10.1093/nar/gkx238
89) Chen, K.M., Wong, A.K., Troyanskaya, O.G. et al. A sequence-based global map of regulatory activity for deciphering human genetics. Nat Genet 54, 940–949 (2022). https://doi.org/10.1038/s41588-022-01102-2