frontiers | Frontiers in Genetics

TYPE Review
PUBLISHED 13 August 2025
DOI 10.3389/fgene.2025.1603687

OPEN ACCESS

EDITED BY
Jared C. Roach,
Institute for Systems Biology (ISB), United States

REVIEWED BY
David N. Arnosti,
Michigan State University, United States
Faizah Aplop,
University of Malaysia Terengganu, Malaysia

*CORRESPONDENCE
Spyros Foutadakis,
foutadakiss@gmail.com
Eleni Karakike,
elkarakike@gmail.com

RECEIVED 31 March 2025
ACCEPTED 31 July 2025
PUBLISHED 13 August 2025

CITATION
Foutadakis S, Bourika V, Styliara I, Koufargyris P, Safarika A and Karakike E (2025) Machine learning tools for deciphering the regulatory logic of enhancers in health and disease. *Front. Genet.* 16:1603687. doi: 10.3389/fgene.2025.1603687

Machine learning tools for deciphering the regulatory logic of enhancers in health and disease

Spyros Foutadakis[1]*, Vasiliki Bourika[2], Ioanna Styliara[3], Panagiotis Koufargyris[1], Asimina Safarika[1] and Eleni Karakike[1]*

[1]4th Department of Internal Medicine, Medical School, National and Kapodistrian University of Athens, Athens, Greece, [2]Neonatal Unit, First Department of Pediatrics, National and Kapodistrian University of Athens, Athens, Greece, [3]Department of Obstetrics and Gynaecology, School of Medicine, University of Patras, Patras, Greece

Transcriptional enhancers are DNA regulatory elements that control the levels and spatiotemporal patterns of gene expression during development, homeostasis, and pathophysiological processes. Enhancer identification and characterization at the genome-wide scale rely on their structural characteristics, such as chromatin accessibility, binding of transcription factors and cofactors, activating histone modifications, 3D interactions with other regulatory elements, as well as functional characteristics measured by massively parallel reporter assays and sequence conservation approaches. Recently, machine learning approaches and particularly deep learning models (Enformer, BPNet, DeepSTARR, etc.) allow the prediction of enhancers, the impact of variants on their activity and the inference of transcription factor binding sites, leading, among others, to the construction of the first completely synthetic enhancers. We present the above computational tools and discuss their diverse applications towards cracking the enhancer regulatory code, which could have far-reaching ramifications for uncovering essential regulatory mechanisms and diagnosing and treating diseases. With an emphasis on sepsis, a leading cause of morbidity and mortality in hospitalized patients, we discuss computational approaches to identify sepsis-associated endotypes, circuits, and immune cell states and signatures characteristic of this condition, which could aid in developing novel therapies.

KEYWORDS

deep learning, enhancers, genomics, machine learning, sepsis, transcriptional regulation

1 Introduction

Transcriptional enhancers are DNA regulatory elements that control the levels and spatiotemporal patterns of gene expression during development, homeostasis, and pathophysiological processes (Banerji et al., 1981; Agelopoulos et al., 2021; Sur and Taipale, 2016; Rickels and Shilatifard, 2018; Furlong and Levine, 2018). Enhancers regulate transcription irrespective of orientation to the Transcription Start Site (TSS) (Banerji et al., 1981), can act over large genomic distances exceeding 1 Mbp (Lettice et al., 2003; Long et al., 2020) and can skip their closest gene in the linear DNA sequence to regulate distal genes (Kessler et al., 2023; Chen et al., 2024). There are two main models regarding the architectural organization of enhancers: the enhanceosome and the billboard model (Arnosti and Kulkarni, 2005; Jindal and Farley, 2021). The enhanceosome operates with a high degree of cooperativity between enhancer-bound proteins, with the exact

organization of transcription factor (tf) binding sites being crucial for enhancer output. The archetype enhanceosome is that of the interferon beta, induced following viral infections (Thanos and Maniatis, 1995). Nevertheless, enhanceosomes are rather rare and have been described for a limited number of cases involving mainly cytokine genes (Giese et al., 1995; Jindal and Farley, 2021). On the other hand, the billboard model posits that the binding sites are flexibly arranged and the bound proteins act as an ensemble interacting independently with their targets (Kulkarni and Arnosti, 2003; Jindal and Farley, 2021).

In the following sections, we present the main wet-lab methodologies to identify enhancers at the genome-wide scale (Catarino and Stark, 2018) and characterize their regulatory grammar (Jindal and Farley, 2021) and 3D interaction networks (Uyehara and Apostolou, 2023). We also describe booming machine learning algorithms to aid the above tasks towards a better understanding of pathophysiological conditions.

2 Enhancer identification at the genome-wide scale

Enhancer identification and characterization at the genome-wide scale rely on their structural characteristics, such as chromatin accessibility (Buenrostro et al., 2013; Marinov et al., 2023; John et al., 2013; Vierstra et al., 2020), binding of transcription factors and cofactors (Robertson et al., 2007; Lambert et al., 2018), activating histone modifications (Heintzman et al., 2007; Calo and Wysocka, 2013; Barral and Déjardin, 2023) and DNA hypomethylation (Angeloni and Bogdanovic, 2019) (Figure 1). More specifically, chromatin accessibility denotes the regulatory potential of a DNA sequence and is measured by popular assays such as DNaseI-seq (John et al., 2013; Vierstra et al., 2020) and ATAC-seq (Buenrostro et al., 2013; Marinov et al., 2023). The combinatorial presence of histone modifications on chromatin constitutes a regulatory code (Kouzarides, 2007), with the genomic distribution of individual modifications usually measured with the ChIP-seq technology (Heintzman et al., 2007; Fadri et al., 2024). Histone modifications associated with active enhancers include H3K27ac and H3K4me1, while H3K27me3 and H3K9me3 are repressive marks (Heintzman et al., 2007; Calo and Wysocka, 2013; Barral and Déjardin, 2023). Most importantly, the binding of transcription factors and coactivators such as Med1 and CBP/P300 and other members of the transcription apparatus are also hallmarks of active enhancers (Catarino and Stark, 2018). High levels of the above enhancer markers are found in long stretches of enhancer DNA termed super-enhancers or stretch-enhancers that play crucial roles in the control of cell identity and disease-related genes (Hnisz et al., 2013; Whyte et al., 2013; Parker et al., 2013). The above epigenomics methodologies are frequently paired with whole transcriptome measurements using RNA-seq (Mortazavi et al., 2008; Deshpande et al., 2023) and the integration of the complementary multiomics methodologies provides a more holistic picture of the regulatory mechanisms that govern transcriptional programs.

Until recently the above (epi)genomics methodologies were only applied in bulk cell populations, thus providing only an ensemble of averaged signals, although it is well established that there is pervasive heterogeneity and stochasticity across systems, meaning that each individual cell inevitably has a unique profile (Eling et al., 2019). In the last decade, technological advances have ushered in the era of single-cell genomics with technologies such as single-cell RNA-seq (Tang et al., 2009; Klein et al., 2015), single-cell ATAC-seq (Buenrostro et al., 2015), simultaneous measurement of different modalities (Baysoy et al., 2023) and even spatial methodologies (Moses and Pachter, 2022; Deng et al., 2022), allowing the study of individual cells (Heumos et al., 2023).

All the above structural characteristics of enhancers are just convenient proxies for enhancer identification and do not prove that a certain DNA sequence holds true enhancer potential. The classic method to measure enhancer activity is the reporter gene assay, with the genomics-based derivatives STARR-seq (Arnold et al., 2013) and MPRA (Melnikov et al., 2012) allowing the massively parallel measurement of the enhancer potential for thousands or even millions of DNA sequences (Inoue and Ahituv, 2015). A limitation of the above methodologies is their episomal nature, measuring the activity of putative enhancers outside their native chromatin environment. The random genomic integration of the tested regions with the variant MPRA technology termed lenti-MPRA (Inoue et al., 2017) alleviates this problem to a certain extent, but nevertheless it does not allow the measuring of enhancer activity at the endogenous locus. An approach to measure enhancer activity at the native chromatin environment is the detection of enhancer RNAs (eRNAs), long non-coding RNAs that are produced from transcription of enhancer sequences, usually in a bidirectional fashion (Kim et al., 2010; Sartorelli and Lauberth, 2020). Nevertheless, eRNAs are of an unstable nature, necessitating the application of nascent RNA sequencing approaches such as GRO-seq (Core et al., 2008) and PRO-seq (Mahat et al., 2016) for their detection. The ultimate test to prove the activity of an enhancer and identify its target gene(s) is the disruption or mutation of its sequence with simultaneous measurements of its effect on the expression levels and chromatin environment of nearby genes and regulatory elements, respectively. This can be achieved with CRISPR-based methodologies such as CRISPR-interference (Fulco et al., 2016) that can silence an enhancer, CRISPR-mediated saturation mutagenesis (Canver et al., 2015) and even CRISPR-mediated activation of enhancers (Heidersbach et al., 2023). The above experimental methodologies to detect enhancers are complemented by DNA sequence conservation approaches to examine enhancer evolution across species (Siepel et al., 2005; Villar et al., 2015).

Large national and international consortia like the ENCODE (ENCODE Project Consortium, 2012), the Roadmap Epigenomics (Roadmap Epigenomics Consortium, 2015), the International Human Epigenome Consortium (Stunnenberg et al., 2016) and the Genotype-Tissue Expression Project (GTEx Consortium, 2013), as well as individual labs, have produced thousands of datasets that are deposited at public repositories such as the Gene Expression Omnibus (Clough et al., 2024) and the European Nucleotide Archive (Yuan et al., 2024) and are freely available for re-analysis and as training material for machine learning applications discussed in a following section.

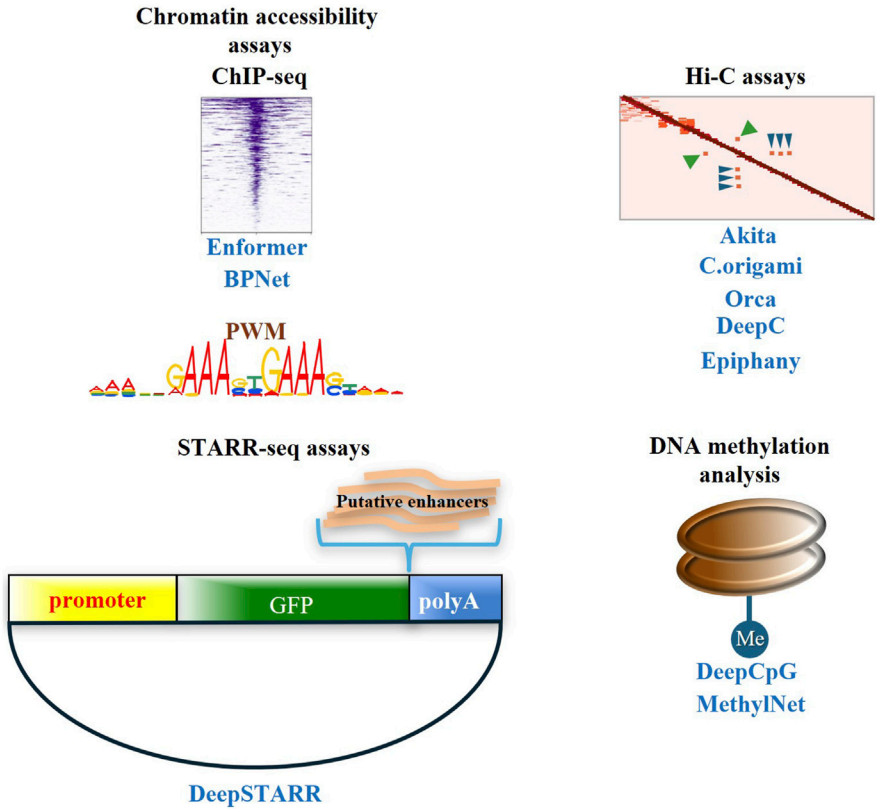

FIGURE 1
Depicted in black are the main (epi)genomics methodologies for studying enhancers, such as chromatin accessibility assays (ATAC-seq, DNaseI-seq) and ChIP-seq to identify open chromatin regions and the presence of histone modifications and tf and coactivator binding, respectively. Moreover, assays such as Hi-C are used to study the 3D organization of the genome, while massively parallel reporter assays like STARR-seq are used to examine the *bona fide* enhancer potential of putative regulatory regions. Depicted in blue are selected state-of-the-art deep learning algorithms trained with datasets produced through the above methodologies and used to identify enhancers, predict crucial motifs and assess the effect of variants on their function.

3 Three-dimensional genome organization and enhancer communication

Vertebrate genomes, apart from the linear DNA sequence, are organized in the three-dimensional space of the nucleus with important ramifications for transcription regulation, DNA replication, and genome integrity (Misteli, 2020). The main approaches to studying 3D genome organization include Hi-C-based methodologies (Rao et al., 2014) or super-resolution microscopy (Boettiger and Murphy, 2020). The higher organizational unit of the genome is the chromosome compartment, with compartment A containing mainly active DNA regions, while compartment B hosts inactive regions (Rao et al., 2014). Although older studies produced rather sparse Hi-C datasets that examined compartments at the Mbp scale, recent efforts using ultra-deep Hi-C data found that A and B compartments alternate at the kilobase scale level (Harris et al., 2023). Another chromatin organizational unit is the topologically associating domains (TADs) that usually span between a few hundred Kbp to a few Mbp (Dixon et al., 2012). TADs set the stage for the finer organizational unit, the loops, which are chromatin interactions between regulatory elements or structural loops that connect CTCF-bound sites (Rao et al., 2014; Furlong and Levine, 2018). Enhancers and promoters within a TAD have a higher probability of interacting than if these regulatory elements were on different TADs. Nevertheless, there are examples of interactions spanning TAD boundaries, especially of developmentally important genes (Hung et al., 2024).

There are two prevailing models regarding the mechanism by which an enhancer interacts with the promoter of the gene it regulates: the classic looping model and the emerging phase separation model (Popay and Dixon, 2022). It is generally accepted that spatial proximity between an enhancer and a promoter is required, but the minimum distance of this interaction and how it relates to activity is highly debated. In support of the looping model and more specifically its instructive nature, targeted tethering of a looping factor allowed the *de novo* formation of a chromatin loop and activation of transcription (Deng et al., 2012). On the other hand, it has been observed mainly through super-resolution microscopy experiments that the timing of enhancer-promoter proximity is not always correlated with activation of gene expression (Alexander et al., 2019; Benabdallah et al., 2019). The latter findings, together with evidence that depletion of CTCF or cohesin, the main organizers of loop formation, does not lead to global transcriptional changes (Rao

et al., 2017; Hsieh et al., 2022) have challenged the universality of the looping model. An alternative model, compatible with the above findings, is the phase separation model (Hnisz et al., 2017), according to which weak multivalent interactions between disordered regions in the activation domains of transcription factors and coactivators lead to condensates that separate from their surroundings, causing high local concentrations of activating factors, especially at super-enhancers. Containment of an enhancer and a promoter within the same condensate would allow gene activation without physical interaction through loop formation. Alternatively, the condensate could transiently interact with the regulatory elements and the gene locus to control gene bursting (Du et al., 2024). Clearly, more experiments are required to establish in a broader fashion the regulatory scenarios in which each of these two models for enhancer regulation applies.

4 Using machine learning to decipher the regulatory logic of enhancers

Machine learning-based approaches to discover patterns in genomics data are broadly categorized into non-neural network algorithms such as support vector machines (SVMs) and random forests (RFs) as well as the increasingly popular neural networks/deep learning approaches (Eraslan et al., 2019; Smith et al., 2023; Li et al., 2023). Neural networks utilize many layers of interconnected neurons to decipher hard-to-recognize patterns, hence the term "deep learning". There are supervised neural network approaches comprised of the most popular convolutional neural networks (CNNs), the fully connected, the recurrent and the graph convolutional, as well as unsupervised methodologies such as autoencoders and generative adversarial networks, with the latter approaches mainly applied in single-cell genomics. In supervised learning, a model is obtained that takes features as input and delivers a prediction for the target variable, which is the desired output used to train the model. A machine learning algorithm is trained using a set of features, usually genomics data such as ChIP-seq or chromatin accessibility datasets that are split into three sets: the training set used for optimizing the parameters of the model, the validation set for evaluating the performance of the model and the test set for the assessment of the best model (Eraslan et al., 2019). The parameters of the network are randomly initiated and refined in an iterative fashion using batches of the training dataset for memory efficiency and prevention of overfitting, and the process is usually parallelized using graphical processing units (GPUs). The analyst can fine-tune different hyperparameters, such as the number of layers and the batch size, using the validation set before the final evaluation of the model using the test set. On the contrary, in unsupervised learning, unlabeled data are characterized by utilizing useful properties of the data. Unsupervised methodologies include autoencoders, which embed the data into a low-dimensional space and force the network to extract the most useful features. While autoencoders have found applications using bulk sequencing data such as extraction of gene signatures from RNA-seq data (Tan et al., 2017), they are also suited for single-cell applications such as improving clustering and denoising data (Wang et al., 2021). The generative adversarial networks offer a different approach consisting of two neural networks, a discriminator and a generator trained in parallel, with the generator creating realistic data and the discriminator classifying if a sample is real or created by the generator.

While simple linear "shallow learning" models like logistic regression can take care of standard tabular data, genomic sequences pose several challenges, such as local dependencies. For example, in classifying regions as bound or unbound by a transcription factor, representing enhancers as position-weighted matrices (PWMs) or k-mers with support vector machine approaches such as gkmSVM (Ghandi et al., 2016) may miss patterns where binding depends on cooperative interaction between multiple motifs with fixed spacing. In contrast, convolutional neural networks are perfectly suited to discern such complex dependencies (Eraslan et al., 2019; Smith et al., 2023; Li et al., 2023; Toneyan et al., 2022). They are composed of multiple layers, each scanning the sequence with several filters- PWMs and quantifying similarities between the filter and the sequence, followed by a non-linear activation function and pooling. Each subsequent layer composes the output of the previous layer that can be fed to a fully connected neural network that receives all information to perform the final prediction task. Applications of deep learning tools in genomics include prediction of regulatory elements such as enhancers, tf binding and assessing the effect of DNA variants. The main models for the above tasks and their salient features are presented in Supplementary Table S1. Early seminal applications of CNNs in genomics include DeepBind (Alipanahi et al., 2015), DeepSEA (Zhou and Troyanskaya, 2015) and Basset (Kelley et al., 2016), which were trained on large-scale chromatin accessibility or transcription factor binding data and were used to prioritize variants by predicting their effect on chromatin accessibility patterns. The DanQ algorithm (Quang and Xie, 2016) uses a hybrid architecture of CNNs and bidirectional long short-term memory recurrent networks that deal effectively with long-range dependencies. Another algorithm, the improved successor of Basset, Basenji (Kelley et al., 2018), uses dilated convolution to increase the receptive field, thus accommodating inputs of 131 Kbp again to effectively take into consideration long-range dependencies. Transformers, previously used in natural language processing applications, are particularly capable of handling pairwise interdependencies in DNA sequence data. The Enformer algorithm (Avsec et al., 2021) was the first to employ a combination of CNN and Transformer architecture and achieved higher accuracy compared to previously designed tools. A recent approach, BPNet, trained with high-resolution ChIP-exo data, is a state-of-the-art algorithm and its architecture contains a 10-layer CNN, with 64 filters per layer with dilated convolution, thus achieving a receptive field of 1,034 bp for any position in the genome (Avsec et al., 2021). In general, models that use a large input sequence (receptive field), such as Basenji and Enformer, can better predict long-range enhancers. Together with BPNet, these three models are the state-of-the-art in supervised learning for tasks related to deciphering the enhancer regulatory code (Toneyan et al., 2022). Finally, it is interesting to mention that the constant maturation of deep neural network algorithms has led to a recent milestone in enhancer biology, the construction of synthetic enhancers with the aid of deep neural network models such as DeepSTARR (de Almeida et al., 2022; de Almeida et al., 2024; Taskiran et al., 2024).

There are also deep learning tools dedicated to other applications in genomics, such as predicting the 3D architecture from DNA sequence as well as predicting DNA methylation states.

Machine learning approaches such as Akita (Fudenberg et al., 2020), DeepC (Schwessinger et al., 2020), Orca (Zhou, 2022), Epiphany (Yang et al., 2023) and C. origami (Tan et al., 2023) have been used to predict the 3D genome organization from the linear sequence and/or 1D epigenomics experiments (Wang et al., 2024; Smaruj et al., 2025) (Supplementary Table S2). One major distinction between approaches is the type of input data, with some tools using only DNA sequence (Akita, DeepC, ORCA), while Epiphany inputs only epigenomics data. Sequence-based-only approaches cannot make accurate *de novo* predictions in different cell types, while epigenomics-only approaches usually require an array of different datasets to improve predictive power. Combining the above approaches, C. origami (Tan et al., 2023) is expected to achieve better predictive accuracy at the expense of possibly introducing uncertainties in model interpretation. Another discriminating characteristic of 3D predictive models includes the type of output, which is either a 2D contact map (Akita, Orca, C. origami) or predicted pixels directly from the 1D representation (DeepC, Epiphany), with the former possibly offering advantages as it makes use of the local correlation structure of contact map data (Smaruj et al., 2025).

Finally, tools such as DeepCpG (Angermueller et al., 2017) and MethylNet (Levy et al., 2020) based on CNNs have been used to predict DNA methylation states.

4.1 Model sharing and interpretation

Models are usually created using machine learning frameworks such as TensorFlow (Abadi et al., 2016), PyTorch (Adam et al., 2017) and Keras (Chollet, 2015), a user-friendly API that operates on frameworks like PyTorch. Models can be shared through repositories called model zoos available through the above frameworks as well as Kipoi (Avsec et al., 2019), a model zoo dedicated to genomics.

Although deep neural networks are often criticized as being "black boxes", they can nevertheless be interpreted with various methodologies. For example, in perturbation-based methods, the input is modified and changes in the output are inspected. In the case of DNA sequence-based models, perturbations could involve a single nucleotide substitution (Alipanahi et al., 2015; Zhou and Troyanskaya, 2015). Nevertheless, the above approaches are computationally expensive, in contrast to attribution-based approaches such as saliency maps (Simonyan et al., 2013). The latter attribute a model's intermediate network value to the input, with the magnitude of the score showing the amount of contribution. Two state-of-the-art approaches in model interpretation for transcription factor motif-fed algorithms include DeepLift (Shrikumar et al., 2016) and TF-MoDISco (Shrikumar et al., 2020). DeepLift decomposes the output of a neural network on a specific input by backpropagating the contributions of all neurons in the network to every feature of the input. In this way, it quantifies the importance of each nucleotide in a sequence for the model prediction. On the other hand, TF-MoDISco is suitable for motif interpretation and discovery and uses all neurons of a network to process sequence importance scores. In doing so, it clusters the most important nucleotides from different sequences into motifs.

5 Enhancer identification and machine learning tools in health and disease

Predicting enhancers and deciphering their regulatory code (Kim and Wysocka, 2023) could have far-reaching ramifications not only for discovering fundamental regulatory mechanisms (Kim et al., 2021) but also for diagnosing and treating various diseases (Karczewski and Snyder, 2018). A salient example of the former includes the application of chromBPnet, a convolutional neural network, in characterizing the regulatory syntax of enhancers that drive the reprogramming of human fibroblasts to pluripotent cells (Nair et al., 2023). ChromBPnet was used to infer putatively bound tf motifs that influence chromatin accessibility, thus offering insights on transcriptional regulators that drive cellular reprogramming. Moreover, genomics-based enhancer identification frequently in conjunction with machine learning tools has been used to subtype, stage and predict drug responses for various types of cancer such as colon, Ewing sarcoma and hematological malignancies (Akhtar-Zaidi et al., 2012; Riggi et al., 2014; Morgan and Shilatifard, 2015; Ntziachristos et al., 2016; Sur and Taipale, 2016; Morrow et al., 2018; Mack et al., 2019; Cejas et al., 2019; Shang et al., 2019; Yao et al., 2021; Zhu et al., 2024) and other conditions, such as heart failure (Spurrell et al., 2022) and Alzheimer's disease (Berson et al., 2023).

Sepsis is another field of applicable but understudied enhancer discovery, as a condition characterized by important morbidity and mortality (Giamarellos-Bourboulis et al., 2024b), associated with a highly heterogenous but always dysregulated host immune response. Several sepsis subtypes have been identified among patients presenting with immunoparalysis (Cheng et al., 2016), those with macrophage activation-like syndrome (Karakike and Giamarellos-Bourboulis, 2019), as well as the novel category of patients with high interferon gamma and CXCL9 levels (Giamarellos-Bourboulis et al., 2024a). Other studies have applied unsupervised clustering to identify three or four distinct sepsis endotypes (Scicluna et al., 2017; Sweeney et al., 2018). Heterogeneity is addressed in other machine learning approaches, interrogating the transcriptome as conditioned by the pathogen or the predominating immune response (Komorowski et al., 2022); for example, bvnGPS2 has been trained with transcriptomic data to discriminate between bacterial and viral infections (Xie et al., 2024). Other deep learning models (Zhang et al., 2020; Davenport et al., 2016) trained with transcriptomic data from sepsis patients derived two classes of patients, with class 1 characterized by immunosuppression and higher mortality rates.

The large-scale application of epigenomics methodologies in sepsis that are needed to train machine learning algorithms is still at a rudimentary stage. MAGICAL (Chen et al., 2023), a hierarchical Bayesian framework, using single-cell RNA-seq and single-cell ATAC-seq data from peripheral blood mononuclear cells, was able to identify epigenetic circuit biomarkers distinctive of methicillin-susceptible or -resistant *Staphylococcus aureus* bloodstream infection. Nevertheless, there is an urgent need to produce large-scale epigenomics datasets from patients with sepsis, as epigenomics datasets, at least in other experimental systems (Dinh et al., 2020), have been found to have superior performance compared to transcriptomics data in stratifying patients and predicting responses to therapy.

Moreover, other immune response mechanisms, such as tolerance and resilience, are underrepresented in current subtyping efforts, and corresponding signatures are still missing (Shankar-Hari et al., 2024). Putative enhancers marked by H3K4me1 and H3K27ac that are activated in response to lipopolysaccharide-treated monocytes (Saeed et al., 2014) significantly overlapped sepsis-associated single nucleotide polymorphisms (SNPs) that were differentially expressed among two sepsis classes (Davenport et al., 2016). Moreover, there is evidence that histone modifications can act as a reservoir of epigenetic memory in immune tolerance and trained immunity (Netea et al., 2020). For a comprehensive review of epigenetic markers in sepsis, interested readers are referred to a recent review (Binnie et al., 2020). Finally, although promising immunotherapies have recently become available (Kyriazopoulou et al., 2021; Karakike et al., 2022), there is an urgent need for a deeper, holistic understanding of the immune dysregulation during sepsis and the development of novel biomarkers besides the classic markers already used in clinical practice (Karakike et al., 2024).

6 Discussion

Machine learning algorithms for genomics are evolving at a fast pace and as described above, have found applications in identifying enhancers, predicting their most crucial motifs and regulatory logic, as well as assessing the effect of variants on their function. In the latter application, state-of-the-art algorithms such as Enformer (Avsec et al., 2021) trained with large-scale structural (ChIP-seq, ATAC-seq) and functional epigenomics data such as massively parallel reporter assays (Arnold et al., 2013) and SELEX-like assays (Yan et al., 2021) could aid in assessing variants, for example, in immune conditions, such as sepsis where datasets are more scarce (Stankey and Lee, 2023). Nevertheless, machine learning algorithms are faced with challenges in their application, such as overfitting, meaning they do not generalize well to unseen examples. This could be due to the limited number of validated ground truth enhancers, the scarcity of large-scale training data beyond a small number of well-studied cell lines, and the presence of biological and technical variation in training datasets (Li et al., 2023).

The problem of data insufficiency could be ameliorated by using pre-trained self-supervised models adapted from the field of natural language processing, where models such as BERT (Devlin et al., 2019) and GPT (Brown et al., 2020) have achieved great success. For example, DNA-BERT (Ji et al., 2021) is a pre-trained bidirectional encoder that tokenizes the genome into k-mers, has cross-species transfer learning capabilities, and performs on par with other algorithms in tasks such as variant interpretation and transcription factor binding site prediction using only small amounts of task-specific labeled data. Other recently developed foundation models include Nucleotide Transformer (Dalla-Torre et al., 2025) and Alpha Genome (Avsec et al., 2025). Nevertheless, a comparison of pre-trained genomic language models with supervised tools such as Enformer did not find a clear advantage for self-supervised models in a variety of genomics tasks (Tang et al., 2025). Another challenge in clinical settings such as sepsis is that the presence of baseline datasets before overt disease onset (Garcia Lopez et al., 2024) is very rare. Perhaps in this scenario, machine learning algorithms could be used to impute missing datasets (Schreiber et al., 2023).

With more sophisticated and efficient algorithms trained with well-validated ground truth datasets, the stage is set for machine learning applications to delve into biological mechanisms and provide groundbreaking insights for the betterment of human health.

Author contributions

SF: Writing – original draft, Writing – review and editing. VB: Writing – review and editing. IS: Writing – review and editing. PK: Writing – original draft. AS: Writing – original draft. EK: Writing – review and editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2025.1603687/full#supplementary-material

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). Tensorflow: large-scale machine learning on heterogeneous distributed systems. Available online at: https://arxiv.org/abs/1603.04467.

Adam, P., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). "Automatic differentiation in PyTorch," in *Presented at 31st conference on neural information processing systems (NIPS 2017)*.

Agelopoulos, M., Foutadakis, S., and Thanos, D. (2021). The causes and consequences of spatial organization of the genome in regulation of gene expression. *Front. Immunol.* 12, 682397. doi:10.3389/fimmu.2021.682397

Akhtar-Zaidi, B., Cowper-Sal-lari, R., Corradin, O., Saiakhova, A., Bartels, C. F., Balasubramanian, D., et al. (2012). Epigenomic enhancer profiling defines a signature of colon cancer. *Science* 336 (6082), 736–739. doi:10.1126/science.1217277

Alexander, J. M., Guan, J., Li, B., Maliskova, L., Song, M., Shen, Y., et al. (2019). Live-cell imaging reveals enhancer-dependent *Sox2* transcription in the absence of enhancer proximity. *Elife* 8, e41769. doi:10.7554/eLife.41769

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33 (8), 831–838. doi:10.1038/nbt.3300

Angeloni, A., and Bogdanovic, O. (2019). Enhancer DNA methylation: implications for gene regulation. *Essays Biochem.* 63 (6), 707–715. doi:10.1042/EBC20190030

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339 (6123), 1074–1077. doi:10.1126/science.1232542

Arnosti, D. N., and Kulkarni, M. M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J. Cell Biochem.* 94 (5), 890–898. doi:10.1002/jcb.20352

Avsec, Ž., Kreuzhuber, R., Israeli, J., Xu, N., Cheng, J., Shrikumar, A., et al. (2019). The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.* 37 (6), 592–600. doi:10.1038/s41587-019-0140-0

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., et al. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* 18 (10), 1196–1203. doi:10.1038/s41592-021-01252-x

Avsec, Ž., Latysheva, N., Cheng, J., Novati, G., Taylor, K. R., Ward, T., et al. (2025). *AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model*. BioRxiv. doi:10.1101/2025.06.25.661532

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27 (2 Pt 1), 299–308. doi:10.1016/0092-8674(81)90413-x

Barral, A., and Déjardin, J. (2023). The chromatin signatures of enhancers and their dynamic regulation. *Nucleus* 14 (1), 2160551. doi:10.1080/19491034.2022.2160551

Baysoy, A., Bai, Z., Satija, R., and Fan, R. (2023). The technological landscape and applications of single-cell multi-omics. *Nat. Rev. Mol. Cell Biol.* 24 (10), 695–713. doi:10.1038/s41580-023-00615-w

Benabdallah, N. S., Williamson, I., Illingworth, R. S., Kane, L., Boyle, S., Sengupta, D., et al. (2019). Decreased enhancer-promoter proximity Accompanying enhancer activation. *Mol. Cell* 76 (3), 473–484. doi:10.1016/j.molcel.2019.07.038

Berson, E., Sreenivas, A., Phongpreecha, T., Perna, A., Grandi, F. C., Xue, L., et al. (2023). Whole genome deconvolution unveils Alzheimer's resilient epigenetic signature. *Nat. Commun.* 14 (1), 4947. doi:10.1038/s41467-023-40611-4

Binnie, A., Tsang, J. L. Y., Hu, P., Carrasqueiro, G., Castelo-Branco, P., and Dos Santos, C. C. (2020). Epigenetics of sepsis. *Crit. Care Med.* 48 (5), 745–756. doi:10.1097/CCM.0000000000004247

Boettiger, A., and Murphy, S. (2020). Advances in chromatin imaging at kilobase-scale resolution. *Trends Genet.* 36 (4), 273–287. doi:10.1016/j.tig.2019.12.010

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). *Language models are few-shot learners*. Preprint at arXiv. doi:10.48550/arXiv.2005.14165

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10 (12), 1213–1218. doi:10.1038/nmeth.2688

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., et al. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523 (7561), 486–490. doi:10.1038/nature14590

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? *Mol. Cell* 49 (5), 825–837. doi:10.1016/j.molcel.2013.01.038

Canver, M. C., Smith, E. C., Sher, F., Pinello, L., Sanjana, N. E., Shalem, O., et al. (2015). BCL11A enhancer dissection by Cas9-mediated *in situ* saturating mutagenesis. *Nature* 527 (7577), 192–197. doi:10.1038/nature15521

Catarino, R. R., and Stark, A. (2018). Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes Dev.* 32 (3-4), 202–223. doi:10.1101/gad.310367.117

Cejas, P., Drier, Y., Dreijerink, K. M. A., Brosens, L. A. A., Deshpande, V., Epstein, C. B., et al. (2019). Enhancer signatures stratify and predict outcomes of non-functional pancreatic neuroendocrine tumors. *Nat. Med.* 25 (8), 1260–1265. doi:10.1038/s41591-019-0493-4

Chen, X., Wang, Y., Cappuccio, A., Cheng, W. S., Zamojski, F. R., Nair, V. D., et al. (2023). Mapping disease regulatory circuits at cell-type resolution from single-cell multiomics data. *Nat. Comput. Sci.* 3 (7), 644–657. doi:10.1038/s43588-023-00476-5

Chen, Z., Snetkova, V., Bower, G., Jacinto, S., Clock, B., Dizehchi, A., et al. (2024). Increased enhancer-promoter interactions during developmental enhancer activation in mammals. *Nat. Genet.* 56 (4), 675–685. doi:10.1038/s41588-024-01681-2

Cheng, S. C., Scicluna, B. P., Arts, R. J., Gresnigt, M. S., Lachmandas, E., Giamarellos-Bourboulis, E. J., et al. (2016). Broad defects in the energy metabolism of leukocytes underlie immunoparalysis in sepsis. *Nat. Immunol.* 17 (4), 406–413. doi:10.1038/ni.3398

Chollet, F. (2015). Keras. Available online at: https://github.com/fchollet/keras.

Clough, E., Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., et al. (2024). NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. *Nucleic Acids Res.* 52 (D1), D138–D144. doi:10.1093/nar/gkad965

Core, L. J., Waterfall, J. J., and Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* 322 (5909), 1845–1848. doi:10.1126/science.1162228

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A. H., Oteri, F., et al. (2025). Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* 22 (2), 287–297. doi:10.1038/s41592-024-02523-z

Davenport, E. E., Burnham, K. L., Radhakrishnan, J., Humburg, P., Hutton, P., Mills, T. C., et al. (2016). Genomic landscape of the individual host response and outcomes in sepsis: a prospective cohort study. *Lancet Respir. Med.* 4 (4), 259–271. doi:10.1016/S2213-2600(16)00046-1

de Almeida, B. P., Reiter, F., Pagani, M., and Stark, A. (2022). DeepSTARR predicts enhancer activity from DNA sequence and enables the *de novo* design of synthetic enhancers. *Nat. Genet.* 54 (5), 613–624. doi:10.1038/s41588-022-01048-5

de Almeida, B. P., Schaub, C., Pagani, M., Secchia, S., Furlong, E. E. M., and Stark, A. (2024). Targeted design of synthetic enhancers for selected tissues in the Drosophila embryo. *Nature* 626 (7997), 207–211. doi:10.1038/s41586-023-06905-9

Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P. D., et al. (2012). Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* 149 (6), 1233–1244. doi:10.1016/j.cell.2012.03.051

Deng, Y., Bartosovic, M., Ma, S., Zhang, D., Kukanja, P., Xiao, Y., et al. (2022). Spatial profiling of chromatin accessibility in mouse and human tissues. *Nature* 609 (7926), 375–383. doi:10.1038/s41586-022-05094-1

Deshpande, D., Chhugani, K., Chang, Y., Karlsberg, A., Loeffler, C., Zhang, J., et al. (2023). RNA-seq data science: from raw data to effective interpretation. *Front. Genet.* 14, 997383. doi:10.3389/fgene.2023.997383

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). *BERT: pre training of deep bidirectional Transformers for language understanding*. Minneapolis, Minnesota: Association for Computational Linguistics, 4171–4186.

Dinh, T. A., Sritharan, R., Smith, F. D., Francisco, A. B., Ma, R. K., Bunaciu, R. P., et al. (2020). Hotspots of Aberrant enhancer activity in Fibrolamellar Carcinoma reveal Candidate Oncogenic Pathways and Therapeutic Vulnerabilities. *Cell Rep.* 31 (2), 107509. doi:10.1016/j.celrep.2020.03.073

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., et al. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485 (7398), 376–380. doi:10.1038/nature11082

Du, M., Stitzinger, S. H., Spille, J. H., Cho, W. K., Lee, C., Hijaz, M., et al. (2024). Direct observation of a condensate effect on super-enhancer controlled gene bursting. *Cell* 187 (10), 2595–2598. doi:10.1016/j.cell.2024.04.001

Eling, N., Morgan, M. D., and Marioni, J. C. (2019). Challenges in measuring and understanding biological noise. *Nat. Rev. Genet.* 20 (9), 536–548. doi:10.1038/s41576-019-0130-6

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489 (7414), 57–74. doi:10.1038/nature11247

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* 20 (7), 389–403. doi:10.1038/s41576-019-0122-6

Fadri, M. T. M., Lee, J. B., and Keung, A. J. (2024). Summary of ChIP-seq methods and Description of an optimized ChIP-seq Protocol. *Methods Mol. Biol.* 2842, 419–447. doi:10.1007/978-1-0716-4051-7_22

Fudenberg, G., Kelley, D. R., and Pollard, K. S. (2020). Predicting 3D genome folding from DNA sequence with Akita. *Nat. Methods* 17 (11), 1111–1117. doi:10.1038/s41592-020-0958-x

Fulco, C. P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S. R., Perez, E. M., et al. (2016). Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* 354 (6313), 769–773. doi:10.1126/science.aag2445

Furlong, E. E. M., and Levine, M. (2018). Developmental enhancers and chromosome topology. *Science* 361 (6409), 1341–1345. doi:10.1126/science.aau0320

Garcia Lopez, A., Schäuble, S., Sae-Ong, T., Seelbinder, B., Bauer, M., Giamarellos-Bourboulis, E. J., et al. (2024). Risk assessment with gene expression markers in sepsis development. *Cell Rep. Med.* 5 (9), 101712. doi:10.1016/j.xcrm.2024.101712

Ghandi, M., Mohammad-Noori, M., Ghareghani, N., Lee, D., Garraway, L., and Beer, M. A. (2016). gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics* 32 (14), 2205–2207. doi:10.1093/bioinformatics/btw203

Giamarellos-Bourboulis, E. J., Antonelli, M., Bloos, F., Kotsamidi, I., Psarrakis, C., Dakou, K., et al. (2024a). Interferon-gamma driven elevation of CXCL9: a new sepsis endotype independently associated with mortality. *EBioMedicine* 109, 105414. doi:10.1016/j.ebiom.2024.105414

Giamarellos-Bourboulis, E. J., Aschenbrenner, A. C., Bauer, M., Bock, C., Calandra, T., Gat-Viks, I., et al. (2024b). The pathophysiology of sepsis and precision-medicine-based immunotherapy. *Nat. Immunol.* 25 (1), 19–28. doi:10.1038/s41590-023-01660-5

Giese, K., Kingsley, C., Kirshner, J. R., and Grosschedl, R. (1995). Assembly and function of a TCR alpha enhancer complex is dependent on LEF-1-induced DNA bending and multiple protein-protein interactions. *Genes Dev.* 9 (8), 995–1008. doi:10.1101/gad.9.8.995

GTEx Consortium (2013). The genotype-Tissue expression (GTEx) project. *Nat. Genet.* 45 (6), 580–585. doi:10.1038/ng.2653

Harris, H. L., Gu, H., Olshansky, M., Wang, A., Farabella, I., Eliaz, Y., et al. (2023). Chromatin alternates between A and B compartments at kilobase scale for subgenic organization. *Nat. Commun.* 14 (1), 3303. doi:10.1038/s41467-023-38429-1

Heidersbach, A. J., Dorighi, K. M., Gomez, J. A., Jacobi, A. M., and Haley, B. (2023). A versatile, high-efficiency platform for CRISPR-based gene activation. *Nat. Commun.* 14 (1), 902. doi:10.1038/s41467-023-36452-w

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39 (3), 311–318. doi:10.1038/ng1966

Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., et al. (2023). Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* 24 (8), 550–572. doi:10.1038/s41576-023-00586-w

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., et al. (2013). Super-enhancers in the control of cell identity and disease. *Cell* 155 (4), 934–947. doi:10.1016/j.cell.2013.09.053

Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K., and Sharp, P. A. (2017). A phase separation model for transcriptional control. *Cell* 169 (1), 13–23. doi:10.1016/j.cell.2017.02.007

Hsieh, T. S., Cattoglio, C., Slobodyanyuk, E., Hansen, A. S., Darzacq, X., and Tjian, R. (2022). Enhancer-promoter interactions and transcription are largely maintained upon acute loss of CTCF, cohesin, WAPL or YY1. *Nat. Genet.* 54 (12), 1919–1932. doi:10.1038/s41588-022-01223-8

Hung, T. C., Kingsley, D. M., and Boettiger, A. N. (2024). Boundary stacking interactions enable cross-TAD enhancer-promoter communication during limb development. *Nat. Genet.* 56 (2), 306–314. doi:10.1038/s41588-023-01641-2

Inoue, F., and Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. *Genomics* 106 (3), 159–164. doi:10.1016/j.ygeno.2015.06.005

Inoue, F., Kircher, M., Martin, B., Cooper, G. M., Witten, D. M., McManus, M. T., et al. (2017). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* 27 (1), 38–52. doi:10.1101/gr.212092.116

Jindal, G. A., and Farley, E. K. (2021). Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Dev Cell.* 56(5), 575–587. doi:10.1016/j.devcel.2021.02.016

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. (2021). DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* 37 (15), 2112–2120. doi:10.1093/bioinformatics/btab083

John, S., Sabo, P. J., Canfield, T. K., Lee, K., Vong, S., Weaver, M., et al. (2013). Genome-scale mapping of DNase I hypersensitivity. *Curr. Protoc. Mol. Biol.* Chapter 27, Unit 21.27. doi:10.1002/0471142727.mb2127s103

Karakike, E., and Giamarellos-Bourboulis, E. J. (2019). Macrophage activation-like syndrome: a distinct entity leading to early Death in sepsis. *Front. Immunol.* 10, 55. doi:10.3389/fimmu.2019.00055

Karakike, E., Dalekos, G. N., Koutsodimitropoulos, I., Saridaki, M., Pourzitaki, C., Papathanakos, G., et al. (2022). ESCAPE: an open-label trial of Personalized immunotherapy in critically lll COVID-19 patients. *J. Innate Immun.* 14 (3), 218–228. doi:10.1159/000519090

Karakike, E., Metallidis, S., Poulakou, G., Kosmidou, M., Gatselis, N. K., Petrakis, V., et al. (2024). Clinical Phenotyping for Prognosis and immunotherapy Guidance in bacterial sepsis and COVID-19. *Crit. Care Explor* 6 (9), e1153. doi:10.1097/CCE.0000000000001153

Karczewski, K. J., and Snyder, M. P. (2018). Integrative omics for health and disease. *Nat. Rev. Genet.* 19 (5), 299–310. doi:10.1038/nrg.2018.4

Kelley, D. R., Snoek, J., and Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* 26 (7), 990–999. doi:10.1101/gr.200535.115

Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., and Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* 28 (5), 739–750. doi:10.1101/gr.227819.117

Kessler, S., Minoux, M., Joshi, O., Ben Zouari, Y., Ducret, S., Ross, F., et al. (2023). A multiple super-enhancer region establishes inter-TAD interactions and controls Hoxa function in cranial neural crest. *Nat. Commun.* 14 (1), 3242. doi:10.1038/s41467-023-38953-0

Kim, S., and Wysocka, J. (2023). Deciphering the multi-scale, quantitative cis-regulatory code. *Mol. Cell* 83 (3), 373–392. doi:10.1016/j.molcel.2022.12.032

Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465 (7295), 182–187. doi:10.1038/nature09033

Kim, D. S., Risca, V. I., Reynolds, D. L., Chappell, J., Rubin, A. J., Jung, N., et al. (2021). The dynamic, combinatorial cis-regulatory lexicon of epidermal differentiation. *Nat. Genet.* 53 (11), 1564–1576. doi:10.1038/s41588-021-00947-3

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., et al. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161 (5), 1187–1201. doi:10.1016/j.cell.2015.04.044

Komorowski, M., Green, A., Tatham, K. C., Seymour, C., and Antcliffe, D. (2022). Sepsis biomarkers and diagnostic tools with a focus on machine learning. *EBioMedicine* 86, 104394. doi:10.1016/j.ebiom.2022.104394

Kouzarides, T. (2007). Chromatin modifications and their function. *Cell* 128 (4), 693–705. doi:10.1016/j.cell.2007.02.005

Kulkarni, M. M., and Arnosti, D. N. (2003). Information display by transcriptional enhancers. *Development* 130 (26), 6569–6575. doi:10.1242/dev.00890

Kyriazopoulou, E., Poulakou, G., Milionis, H., Metallidis, S., Adamis, G., Tsiakos, K., et al. (2021). Early treatment of COVID-19 with anakinra guided by soluble urokinase plasminogen receptor plasma levels: a double-blind, randomized controlled phase 3 trial. *Nat. Med.* 27 (10), 1752–1760. doi:10.1038/s41591-021-01499-z

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., et al. (2018). The human transcription factors. *Cell* 172 (4), 650–665. doi:10.1016/j.cell.2018.01.029

Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., et al. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* 12 (14), 1725–1735. doi:10.1093/hmg/ddg180

Levy, J. J., Titus, A. J., Petersen, C. L., Chen, Y., Salas, L. A., and Christensen, B. C. (2020). MethylNet: an automated and modular deep learning approach for DNA methylation analysis. *BMC Bioinforma.* 21 (1), 108. doi:10.1186/s12859-020-3443-8

Li, Z., Gao, E., Zhou, J., Han, W., Xu, X., and Gao, X. (2023). Applications of deep learning in understanding gene regulation. *Cell Rep. Methods* 3 (1), 100384. doi:10.1016/j.crmeth.2022.100384

Long, H. K., Osterwalder, M., Welsh, I. C., Hansen, K., Davies, J. O. J., Liu, Y. E., et al. (2020). Loss of Extreme long-range enhancers in human neural crest drives a Craniofacial disorder. *Cell Stem Cell* 27 (5), 765–783. doi:10.1016/j.stem.2020.09.001

Mack, S. C., Singh, I., Wang, X., Hirsch, R., Wu, Q., Villagomez, R., et al. (2019). Chromatin landscapes reveal developmentally encoded transcriptional states that define human glioblastoma. *J. Exp. Med.* 216 (5), 1071–1090. doi:10.1084/jem.20190196

Mahat, D. B., Kwak, H., Booth, G. T., Jonkers, I. H., Danko, C. G., Patel, R. K., et al. (2016). Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* 11 (8), 1455–1476. doi:10.1038/nprot.2016.086

Marinov, G. K., Shipony, Z., Kundaje, A., and Greenleaf, W. J. (2023). Genome-wide mapping of active regulatory elements using ATAC-seq. *Methods Mol. Biol.* 2611, 3–19. doi:10.1007/978-1-0716-2899-7_1

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* 30 (3), 271–277. doi:10.1038/nbt.2137

Misteli, T. (2020). The self-Organizing genome: principles of genome architecture and function. *Cell* 183 (1), 28–45. doi:10.1016/j.cell.2020.09.014

Morgan, M. A., and Shilatifard, A. (2015). Chromatin signatures of cancer. *Genes Dev.* 29 (3), 238–249. doi:10.1101/gad.255182.114

Morrow, J. J., Bayles, I., Funnell, A. P. W., Miller, T. E., Saiakhova, A., Lizardo, M. M., et al. (2018). Positively selected enhancer elements endow osteosarcoma cells with metastatic competence. *Nat. Med.* 24 (2), 176–185. doi:10.1038/nm.4475

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5 (7), 621–628. doi:10.1038/nmeth.1226

Moses, L., and Pachter, L. (2022). Museum of spatial transcriptomics. *Nat. Methods* 19 (5), 534–546. doi:10.1038/s41592-022-01409-2

Nair, S., Ameen, M., Sundaram, L., Pampari, A., Schreiber, J., Balsubramani, A., et al. (2023). *Transcription factor stoichiometry, motif affinity and syntax regulate single-cell chromatin dynamics during fibroblast reprogramming to pluripotency*. bioRxiv. doi:10.1101/2023.10.04.560808

Netea, M. G., Domínguez-Andrés, J., Barreiro, L. B., Chavakis, T., Divangahi, M., Fuchs, E., et al. (2020). Defining trained immunity and its role in health and disease. *Nat. Rev. Immunol.* 20 (6), 375–388. doi:10.1038/s41577-020-0285-6

Ntziachristos, P., Abdel-Wahab, O., and Aifantis, I. (2016). Emerging concepts of epigenetic dysregulation in hematological malignancies. *Nat. Immunol.* 17 (9), 1016–1024. doi:10.1038/ni.3517

Parker, S. C., Stitzel, M. L., Taylor, D. L., Orozco, J. M., Erdos, M. R., Akiyama, J. A., et al. (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. U. S. A.* 110 (44), 17921–17926. doi:10.1073/pnas.1317023110

Popay, T. M., and Dixon, J. R. (2022). Coming full circle: on the origin and evolution of the looping model for enhancer-promoter communication. *J. Biol. Chem.* 298 (8), 102117. doi:10.1016/j.jbc.2022.102117

Quang, D., and Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* 44 (11), e107. doi:10.1093/nar/gkw226

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159 (7), 1665–1680. doi:10.1016/j.cell.2014.11.021

Rao, S. S. P., Huang, S. C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K. R., et al. (2017). Cohesin loss Eliminates all loop domains. *Cell* 171 (2), 305–320. doi:10.1016/j.cell.2017.09.026

Rickels, R., and Shilatifard, A. (2018). Enhancer logic and Mechanics in development and disease. *Trends Cell Biol.* 28 (8), 608–630. doi:10.1016/j.tcb.2018.04.003

Riggi, N., Knoechel, B., Gillespie, S. M., Rheinbay, E., Boulay, G., Suvà, M. L., et al. (2014). EWS-FLI1 utilizes divergent chromatin remodeling mechanisms to directly activate or repress enhancer elements in Ewing sarcoma. *Cancer Cell* 26 (5), 668–681. doi:10.1016/j.ccell.2014.10.004

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518 (7539), 317–330. doi:10.1038/nature14248

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* 4 (8), 651–657. doi:10.1038/nmeth1068

Saeed, S., Quintin, J., Kerstens, H. H., Rao, N. A., Aghajanirefah, A., Matarese, F., et al. (2014). Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. *Science* 345 (6204), 1251086. doi:10.1126/science.1251086

Sartorelli, V., and Lauberth, S. M. (2020). Enhancer RNAs are an important regulatory layer of the epigenome. *Nat. Struct. Mol. Biol.* 27 (6), 521–528. doi:10.1038/s41594-020-0446-0

Schreiber, J. M., Boix, C. A., Wook Lee, J., Li, H., Guan, Y., Chang, C. C., et al. (2023). The ENCODE Imputation Challenge: a critical assessment of methods for cross-cell type imputation of epigenomic profiles. *Genome Biol.* 24 (1), 79. doi:10.1186/s13059-023-02915-y

Schwessinger, R., Gosden, M., Downes, D., Brown, R. C., Oudelaar, A. M., Telenius, J., et al. (2020). DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat. Methods* 17 (11), 1118–1124. doi:10.1038/s41592-020-0960-3

Scicluna, B. P., van Vught, L. A., Zwinderman, A. H., Wiewel, M. A., Davenport, E. E., Burnham, K. L., et al. (2017). Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study. *Lancet Respir. Med.* 5 (10), 816–826. doi:10.1016/S2213-2600(17)30294-1

Shang, S., Yang, J., Jazaeri, A. A., Duval, A. J., Tufan, T., Lopes Fischer, N., et al. (2019). Chemotherapy-induced distal enhancers drive transcriptional programs to Maintain the chemoresistant state in ovarian cancer. *Cancer Res.* 79 (18), 4599–4611. doi:10.1158/0008-5472.CAN-19-0215

Shankar-Hari, M., Calandra, T., Soares, M. P., Bauer, M., Wiersinga, W. J., Prescott, H. C., et al. (2024). Reframing sepsis immunobiology for translation: towards informative subtyping and targeted immunomodulatory therapies. *Lancet Respir. Med.* 12 (4), 323–336. doi:10.1016/S2213-2600(23)00468-X

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. (2016). *Not just a black box: learning important features through propagating activation differences*. Preprint at arXiv.

Shrikumar, A., Tian, K., Avsec, Ž., Shcherbina, A., Banerjee, A., Sharmin, M., et al. (2020). Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5. 6.5. Available online at: http://arxiv.org/abs/1811.00416.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15 (8), 1034–1050. doi:10.1101/gr.3715005

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). *Deep inside convo lutional networks: visualising image classification models and saliency maps*. Preprint at arXiv. doi:10.48550/arXiv.1312.6034

Smaruj, P. N., Xiao, Y., and Fudenberg, G. (2025). Recipes and ingredients for deep learning models of 3D genome folding. *Curr. Opin. Genet. Dev.* 91, 102308. doi:10.1016/j.gde.2024.102308

Smith, G. D., Ching, W. H., Cornejo-Páramo, P., and Wong, E. S. (2023). Decoding enhancer complexity with machine learning and high-throughput discovery. *Genome Biol.* 24 (1), 116. doi:10.1186/s13059-023-02955-4

Spurrell, C. H., Barozzi, I., Kosicki, M., Mannion, B. J., Blow, M. J., Fukuda-Yuzawa, Y., et al. (2022). Genome-wide fetalization of enhancer architecture in heart disease. *Cell Rep.* 40 (12), 111400. doi:10.1016/j.celrep.2022.111400

Stankey, C. T., and Lee, J. C. (2023). Translating non-coding genetic associations into a better understanding of immune-mediated disease. *Dis. Model Mech.* 16 (3), dmm049790. doi:10.1242/dmm.049790

Stunnenberg, H. G., Hirst, M., and International Human Epigenome Consortium (2016). The international human epigenome Consortium: a blueprint for Scientific collaboration and discovery. *Cell* 167 (5), 1145–1149. doi:10.1016/j.cell.2016.11.007

Sur, I., and Taipale, J. (2016). The role of enhancers in cancer. *Nat. Rev. Cancer* 16 (8), 483–493. doi:10.1038/nrc.2016.62

Sweeney, T. E., Azad, T. D., Donato, M., Haynes, W. A., Perumal, T. M., Henao, R., et al. (2018). Unsupervised analysis of transcriptomics in bacterial sepsis across multiple datasets reveals three robust clusters. *Crit. Care Med.* 46 (6), 915–925. doi:10.1097/CCM.0000000000003084

Tan, J., Doing, G., Lewis, K. A., Price, C. E., Chen, K. M., Cady, K. C., et al. (2017). Unsupervised extraction of stable expression signatures from public Compendia with an ensemble of neural networks. *Cell Syst.* 5 (1), 63–71. doi:10.1016/j.cels.2017.06.003

Tan, J., Shenker-Tauris, N., Rodriguez-Hernaez, J., Wang, E., Sakellaropoulos, T., Boccalatte, F., et al. (2023). Cell-type-specific prediction of 3D chromatin organization enables high-throughput *in silico* genetic screening. *Nat. Biotechnol.* 41 (8), 1140–1150. doi:10.1038/s41587-022-01612-8

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6 (5), 377–382. doi:10.1038/nmeth.1315

Tang, Z., Somia, N., Yu, Y., and Koo, P. K. (2025). Evaluating the representational power of pre-trained DNA language models for regulatory genomics. *Genome Biol.* 26 (1), 203. doi:10.1186/s13059-025-03674-8

Taskiran, I. I., Spanier, K. I., Dickmänken, H., Kempynck, N., Pančíková, A., Ekşi, E. C., et al. (2024). Cell-type-directed design of synthetic enhancers. *Nature* 626 (7997), 212–220. doi:10.1038/s41586-023-06936-2

Thanos, D., and Maniatis, T. (1995). Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell* 83 (7), 1091–1100. doi:10.1016/0092-8674(95)90136-1

Toneyan, S., Tang, Z., and Koo, P. K. (2022). Evaluating deep learning for predicting epigenomic profiles. *Nat. Mach. Intell.* 4 (12), 1088–1100. doi:10.1038/s42256-022-00570-9

Uyehara, C. M., and Apostolou, E. (2023). 3D enhancer-promoter interactions and multi-connected hubs: organizational principles and functional roles. *Cell Rep.* 42 (4), 112068. doi:10.1016/j.celrep.2023.112068

Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., et al. (2020). Global reference mapping of human transcription factor footprints. *Nature* 583 (7818), 729–736. doi:10.1038/s41586-020-2528-x

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., et al. (2015). Enhancer evolution across 20 mammalian species. *Cell* 160 (3), 554–566. doi:10.1016/j.cell.2015.01.006

Wang, J., Ma, A., Chang, Y., Gong, J., Jiang, Y., Qi, R., et al. (2021). scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. *Nat. Commun.* 12 (1), 1882. doi:10.1038/s41467-021-22197-x

Wang, Y., Kong, S., Zhou, C., Wang, Y., Zhang, Y., Fang, Y., et al. (2024). A review of deep learning models for the prediction of chromatin interactions with DNA and epigenomic profiles. *Brief. Bioinform* 26 (1), bbae651. doi:10.1093/bib/bbae651

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., et al. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 153 (2), 307–319. doi:10.1016/j.cell.2013.03.035

Xie, J., Zheng, X., Yan, J., Li, Q., Jin, N., Wang, S., et al. (2024). Deep learning model to discriminate diverse infection types based on pairwise analysis of host gene expression. *iScience* 27 (6), 109908. doi:10.1016/j.isci.2024.109908

Yan, J., Qiu, Y., Ribeiro Dos Santos, A. M., Yin, Y., Li, Y. E., Vinckier, N., et al. (2021). Systematic analysis of binding of transcription factors to noncoding variants. *Nature* 591 (7848), 147–151. doi:10.1038/s41586-021-03211-0

Yang, R., Das, A., Gao, V. R., Karbalayghareh, A., Noble, W. S., Bilmes, J. A., et al. (2023). Epiphany: predicting Hi-C contact maps from 1D epigenomic signals. *Genome Biol.* 24 (1), 134. doi:10.1186/s13059-023-02934-9

Yao, Q., Epstein, C. B., Banskota, S., Issner, R., Kim, Y., Bernstein, B. E., et al. (2021). Epigenetic alterations in Keratinocyte Carcinoma. *J. Invest. Dermatol* 141 (5), 1207–1218. doi:10.1016/j.jid.2020.10.018

Yuan, D., Ahamed, A., Burgin, J., Cummins, C., Devraj, R., Gueye, K., et al. (2024). The European nucleotide archive in 2023. *Nucleic Acids Res.* 52 (D1), D92–D97. doi:10.1093/nar/gkad1067

Zhang, Z., Pan, Q., Ge, H., Xing, L., Hong, Y., and Chen, P. (2020). Deep learning-based clustering robustly identified two classes of sepsis with both prognostic and predictive values. *EBioMedicine* 62, 103081. doi:10.1016/j.ebiom.2020.103081

Zhou, J. (2022). Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat. Genet.* 54 (5), 725–734. doi:10.1038/s41588-022-01065-4

Zhou, J., and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* 12 (10), 931–934. doi:10.1038/nmeth.3547

Zhu, Y., Lee, H., White, S., Weimer, A. K., Monte, E., Horning, A., et al. (2024). Global loss of promoter-enhancer connectivity and rebalancing of gene expression during early colorectal cancer carcinogenesis. *Nat. Cancer* 5 (11), 1697–1712. doi:10.1038/s43018-024-00823-z