Bioinformatics Advances, 2023, vbad001
https://doi.org/10.1093/bioadv/vbad001
Advance Access Publication Date: 11 January 2023
Review

Data and text mining

Applications of transformer-based language models in bioinformatics: a survey

Shuang Zhang [1,†], Rui Fan[1,†], Yuti Liu [1,†], Shuang Chen [1,†], Qiao Liu [2] and Wanwen Zeng[1,2,*]

[1]College of Software, Nankai University, Tianjin 300350, China and [2]Department of Statistics, Stanford University, Stanford, CA 94305, USA

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first four authors should be regarded as Joint First Authors.

Associate Editor: Alex Bateman

Received on September 6, 2022; revised on December 15, 2022; editorial decision on January 6, 2023; accepted on January 9, 2023

Abstract

Summary: The transformer-based language models, including vanilla transformer, BERT and GPT-3, have achieved revolutionary breakthroughs in the field of natural language processing (NLP). Since there are inherent similarities between various biological sequences and natural languages, the remarkable interpretability and adaptability of these models have prompted a new wave of their application in bioinformatics research. To provide a timely and comprehensive review, we introduce key developments of transformer-based language models by describing the detailed structure of transformers and summarize their contribution to a wide range of bioinformatics research from basic sequence analysis to drug discovery. While transformer-based applications in bioinformatics are diverse and multifaceted, we identify and discuss the common challenges, including heterogeneity of training data, computational expense and model interpretability, and opportunities in the context of bioinformatics research. We hope that the broader community of NLP researchers, bioinformaticians and biologists will be brought together to foster future research and development in transformer-based language models, and inspire novel bioinformatics applications that are unattainable by traditional methods.

Contact: wanwen@stanford.edu

Supplementary information: Supplementary data are available at *Bioinformatics Advances* online.

1 Introduction

Bioinformatics, an interdisciplinary research field, has become one of the most influential areas of life science research in a profound way. It is characterized by the demand to develop and utilize computational tools and methods to analyze huge amounts of biomedical data and translate them into knowledge for developing downstream applications.

In recent years, natural language processing (NLP) (Nadkarni *et al.*, 2011; Supplementary Table S1), a branch of artificial intelligence, has been increasingly showing a substantial impact in bioinformatics research fields (Han and Kwoh, 2019), ranging from DNA/RNA sequence analysis to computational biology (Iuchi *et al.*, 2021; Zeng *et al.*, 2018). Specifically, NLP technologies, with the aim to grant computers the ability to understand words and texts from human beings (Tsujii, 2021), have the potential power to also understand biological languages. Language models enable computers to analyze the patterns of human language by predicting words (Adel *et al.*, 2018) (Fig. 1A) and are becoming one of the core technologies for many NLP tasks, including sentiment analysis (Schouten and Frasincar, 2016), machine translation (Bahdanau *et al.*, 2016) and text summarization (Nenkova and McKeown, 2012). The history of leveraging the power of neural networks (NNs) (Walczak and Cerpa, 2003) in NLP tasks can be tracked back two decades (Bengio *et al.*, 2003), where a series of word embedding technologies were proposed to provide a novel representation of text and achieved superior results (Blacoe and Lapata, 2012; Turian *et al.*, 2010). For example, Word2Vec (Le and Mikolov, 2014; Mikolov *et al.*, 2013a, b), which maps one-hot word vectors to distributed word vectors using a shallow neural network, is one of the most representative models. Word2vec can utilize either of two types of model architecture to produce these distributed representations of words: continuous bag-of-words (CBOW) or continuous skip-gram. CBOW predicts the current word based on the context while skip-gram predicts surrounding words given the current word (Fig. 1B). With the rapid development of deep learning technologies (LeCun *et al.*, 2015), language models in NLP have continuously made significant breakthroughs: conventional RNN-based models, including Bi-RNN (Schuster and Paliwal, 1997), LSTM (Hochreiter and Schmidhuber, 1997) and GRU (Cho *et al.*, 2014), attempt to encode the entire sequence into a finite length vector without paying more attention to those important works. Although these RNN-based models are able to learn long-term dependency, they greatly suffer from vanishing gradient and low-efficiency problems as they

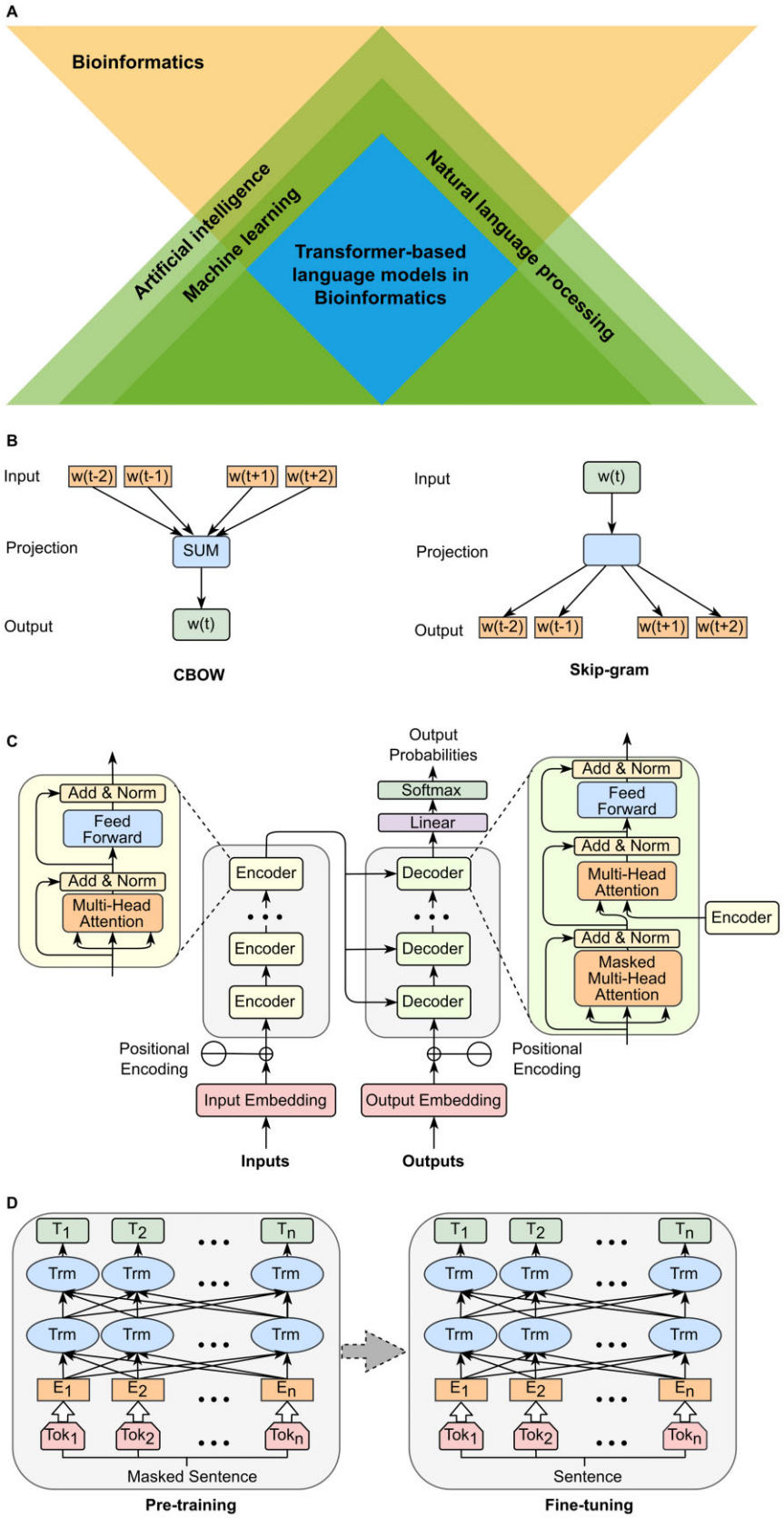

Fig. 1. The focus of this review article and some classic language models frameworks. (**A**) Relationships of artificial intelligence, machine learning, natural language processing, transformer-based language models and bioinformatics. The blue square denotes the focal point of this review article. (**B**) Two common models in Word2Vec: CBOW (Continuous Bag-of-Words Model) and Skip-gram (Continuous Skip-gram Model). (**C**) The structure of transformer model. (**D**) The structure of BERT model

sequentially process all past states and compress contextual information into a bottleneck with long input sequences (Bengio *et al.*, 1994; Pascanu *et al.*, 2013). For example, Seq2Seq (Sutskever *et al.*, 2014), the first encoder–decoder model in machine translation tasks, supports variable-length inputs and outputs but is still limited by its infrastructure LSTM. The Transformer (Vaswani *et al.*, 2017) model was then developed by Google, which completely abandoned RNN-based network structures, and only used the multi-head attention mechanism (Fig. 1C). Transformer does not rely on the past hidden states to capture the dependency on the previous words. Instead, transformer processes a sentence as a whole to allow for parallel computing and alleviates the vanishing gradient and performance degradation caused by long-term dependency. In this review article, we will focus on transformer-based language models.

In general, transformer-based language models fall into two categories: scratch-trained models and pre-trained models. The scratch-trained models directly train all model parameters from the beginning using task-specific datasets and often require many iterations to fully converge. For example, Transformer-XL (Dai *et al.*, 2019) uses relative positional encoding and segmented RNN mechanism to model long text; Sparse Transformers (Zhao *et al.*, 2019) uses only a small number of tokens in the computation of attention distribution to improve the concentration of attention mechanism; Reformer (Kitaev *et al.*, 2020) addresses the resource-hungry problem of the transformer by replacing dot-product attention and using reversible residual layers; Longformer (Beltagy *et al.*, 2020) proposes sliding windows, dilated sliding windows and global attention strategies to reduce the complexity of the model. On the other hand, transformer-based pre-trained models are trained from large amounts of unlabeled data and then fine-tuned for specific tasks. Pre-training learns general information from unlabeled data, speeds up the convergence rate of the target tasks and usually has better generalization than training parameters from scratch (Han *et al.*, 2021). For example, GPT-X (Brown *et al.*, 2020; Radford and Narasimhan, 2018; Radford *et al.*, 2019) proposes unsupervised pre-training and supervised fine-tuning for the first time; BERT (Devlin *et al.*, 2019) utilizes bi-directional transformers and mask mechanism (Fig. 1D) to achieve a deeper understanding of context than GPT; RoBERTa (Liu *et al.*, 2019b) uses dynamic masking and has a significant improvement over BERT in terms of model size and arithmetic power; XLNet (Yang *et al.*, 2019b), which is based on the Transformer-XL architecture, further introduces permutation language modeling as an improved training method; ERNIE (Zhang et al., 2019) adopts a continual learning mechanism, which consists of two parts: continual construction of pre-training tasks and incremental multi-task learning; ALBERT (Lan *et al.*, 2020) is a mini-model using cross-layer parameter sharing and paragraph continuity tasks; T5 (Raffel *et al.*, 2020) is a generic framework that converts all NLP tasks into Text-to-Text format. These two types of transformer-based language models show their strength in addressing key challenges and have become a quintessential choice in almost all NLP tasks (Casola *et al.*, 2022; Chaudhari *et al.*, 2021). These breakthroughs in methodologies and technologies have revolutionized the field of NLP, thus bringing the thoughts of applications in biological and biomedical research.

Although there are reviews of transformers in the general domain (Kalyan *et al.*, 2021b; Lin *et al.*, 2022; Qiu *et al.*, 2020) and a survey of transformer-based biomedical pre-trained language models (Kalyan *et al.*, 2021a), the applications of transformer-based language models in the latest bioinformatics research, such as spatial transcriptomics and multi-omics, have not yet been documented. In this review, we provide a comprehensive viewpoint of facilitating research in the field of NLP and the applications of transformers in bioinformatics. We revisit the basics of transformer-based language models, summarize the latest developments in the transformer-based language models and then review the applications of transformers in bioinformatics and biomedical downstream tasks such as sequence analysis, gene expression, proteomics, spatial transcriptomics, etc. Last but not least, we discuss the future challenges and opportunities in using and understanding multi-omics high-throughput sequencing data. We hope that transformer-based language models not only benefit the computer science community but also the broader community of bioinformaticians and biologists, and further provide insights for future bioinformatics research across multiple disciplines that are unattainable by traditional methods.

2 Basics of transformer-based language models

Language models are trained in a self-supervised fashion (Liu *et al.*, 2023). Compared to supervised learning (Hastie *et al.*, 2009), which usually needs human annotations, language models could use massive amounts of unannotated corpora from the internet, books, etc. Language models either take the next word as a natural label for the context in a sentence or artificially mask a known word and then predict it (Petroni *et al.*, 2019). The paradigm that uses the unstructured data itself to generate labels (e.g. the next word or the masked word in language models) and train supervised models (language models) to predict labels is called 'self-supervised learning' (Howard and Ruder, 2018). Specifically, because of their parallelism and the ability to extract correlation across the whole sequences, transformer-based models achieve state-of-the-art (SOTA) performance in a variety of important tasks such as machine translation and question answering (QA) (Pundge *et al.*, 2016). Since there are high similarities between human language and bioinformatics sequence data, transformer-based models are becoming one of the most promising models to tackle the sequence-based problems in bioinformatics (Ofer *et al.*, 2021).

The vanilla transformer model can be divided into two parts: encoder and decoder, which have similar basic architectures composed of a stack of identical blocks (Vaswani *et al.*, 2017). Each block consists of two kinds of sub-layers: the multi-head attention sub-layer and the position-wise feed-forward sub-layer. Both kinds of sub-layers are followed by layer normalization. A residual connection around every sub-layer will be applied in each block to speed up the training process. The following sections will describe each module that makes up the transformer model in detail.

2.1 Attention modules

The key innovation in transformer is the multi-head self-attention layer, which can relate all relevant tokens to better encode every word in the input sequence (Lin *et al.*, 2017). The self-attention layer takes a sequence of tokens as input (tokens equivalent to words in the language) and learns sequence-wide context information. Multi-head represents multiple simultaneous attention heads. Figure 2 shows the example process of a single attention head in calculating the first token T_1's output embedding in a sequence composed of four tokens.

Before calculating the attention function, each token embedding will be transformed into the corresponding query vector, the key vector of dimension d_k and the value vector of the dimension d_v by multiplying with three randomly initialized learnable parameter matrices W^Q, W^K and W^V. Then, the attention head will compute the dot products of the query with all keys and divide each by $\sqrt{d_k}$ and apply a softmax function to obtain the weights on these values (Vaswani *et al.*, 2017). Through this process, the attention function can be described as mapping a query vector and a set of key-value pairs to an output vector that contains information for the entire sequence. As is seen in Figure 2, the output of the attention function is the weighted sum of these values. The weight assigned to each value is computed by a compatibility function of the query with the corresponding key (Vaswani *et al.*, 2017).

In the parallel computation of the attention function, a set of query vectors is packed into a matrix Q. These key and value vectors are also packed together into matrices K and V. In practice, the attention function is computed as follow:

$$\text{Attention}(Q,\ K,\ V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{1}$$

When being generalized to multi-head attention with h heads, the results of multiple heads assigned different

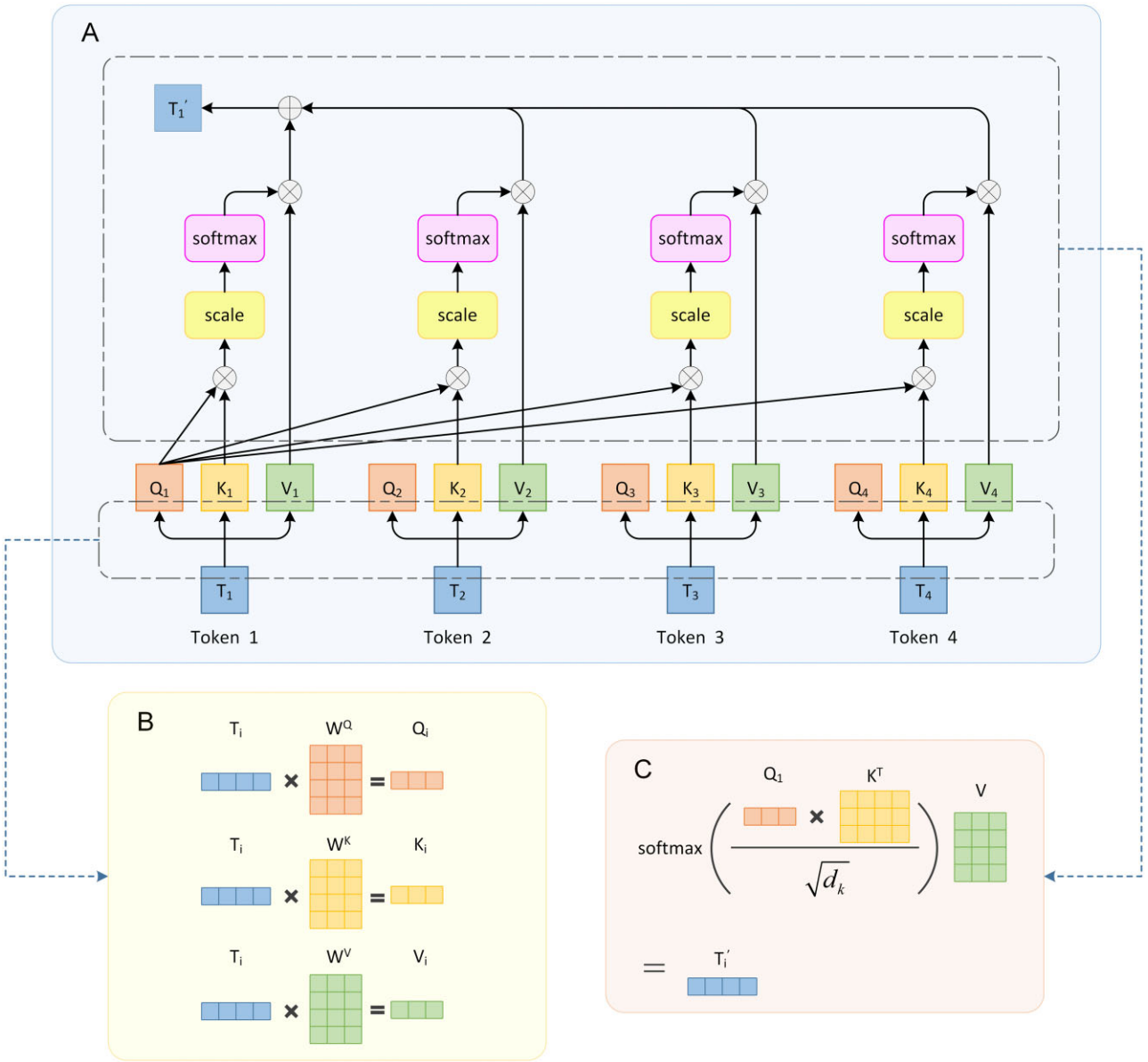

Fig. 2. The example illustration of calculating self-attention. (**A**) The process of computing the output embedding of token T_1 in a single attention head. $T_{i(i=1,2,3,4)}$ represents the embeddings corresponding to the ith token in the input sequence. T_1' is the output corresponding to T_1. Each embedding in the input sequence needs to be multiplied with the three parameter matrices W^Q, W^K and W^V, respectively to obtain the corresponding query vector, key vector and value vector. (**B**) The figure complements the process of generating the ith ($i = 1,2,3,4$) token's corresponding query vector Q_i, key vector K_i and value vector V_i. Each attention head has its own set of three learnable parameter matrices W^Q, W^K and W^V. (**C**) If the key vectors of all tokens are concatenated into a matrix K by row and all value vectors are concatenated into a matrix V by row, the process of calculating T_1' in part A can be expressed as the formula in part C using matrix operations, where K^T is the transpose of K and d_k is the dimension of the key vector

parameters W^Q, W^K and W^V are concatenated, and once again projected with parameter, resulting in the final values, as depicted as follows:

$$\begin{aligned}\text{MultiHead}(Q,\ K,\ V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \\ \text{where head}_i &= \text{Attention}\left(QW_i^Q,\ KW_i^K,\ VW_i^V\right).\end{aligned} \quad (2)$$

2.2 Position-wise feed-forward networks

Except for the attention sub-layer, each block of the encoder and decoder contains a fully connected feed-forward network (FFN) (Skansi, 2018), which is applied identically to each token. This layer consists of two linear transformations with rectified linear unit (ReLU) activation in the middle (Vaswani *et al.*, 2017), where W_1, b_1, W_2 and b_2 are learnable parameters.

$$\text{FFN}(x) = \max(0,\ xW_1 + b_1)W_2 + b_2. \quad (3)$$

2.3 Residual connection and layer normalization

Each encoder contains two residual connection and layer normalization layers, and they are applied on both multi-head self-attention and FFN. The calculation formulas are as follows:

$$\text{LayerNorm}(X + \text{MutiHeadAttention}(X)), \quad (4)$$

$$\text{LayerNorm}(X + \text{FeedForward}(X)). \quad (5)$$

X represents the input of multi-head self-attention or FFN, which is added to the output and forms a residual connection. For the deep network, the residual connection can help fend against vanishing and exploding gradients by keeping the original input

(Zhang *et al.*, 2018). Layer normalization can accelerate the training process of the model by normalizing the output of the former layers to make it converge faster (Ba *et al.*, 2016).

2.4 Position encodings

Since transformer uses pure self-attention without recurrence or convolution to capture connections between tokens, it cannot identify the order of the tokens in the sequence. Therefore, transformer adds position encodings to the input embeddings (Liu *et al.*, 2020) to reflect the absolute or relative position of the tokens in the sequence. The absolute position encoding informs the transformer architecture of the absolute position of each token in the input sequence, while the relative position encoding acts as a self-attention mechanism, informing the transformer architecture of the distance between two tokens (Ke *et al.*, 2021). The input for the first transformer encoder layer is the sum of the input embedding and the position encoding.

2.5 Encoder and decoder

Using the components above, the encoder encodes the input sequence and passes the output intermediate sequence to the decoder, and the decoder decodes the intermediate sequence and outputs the sequence we need. The encoder consists of several identical blocks consisting of one attention sub-layer and a feed-forward layer (Fig. 1C). The decoder inserts one more attention sub-layer between the original two sub-layers to perform multi-head attention over the output of the encoder stack (Fig. 1C).

Decoding the intermediate output of the encoder into a new sequence can be considered as a translation process. First, the decoder takes a special token 'BEGIN' as input, combining it with the encoder's output sequence to produce a vector after passing through the inner blocks of the decoder and a linear layer. The length of this vector is the size of the lexicon. Then, a softmax function is applied to the output vector to generate a probability distribution, and the token in the lexicon with the highest probability is the output, which is also the first token in the final output sequence (Fig. 3).

This output token will be appended to the sequence containing the 'BEGIN' token as the next round of the decoding process's input. This process will be repeated, appending the new output into the input sequence. To end the loop, an 'END' token is appended to the lexicon. The loop stops when the output token is 'END', resulting in the complete final output sequence. Because of the extra 'BEGIN' token, the decoder's input is shifted one position to the right (Fig. 4).

It is worth mentioning that when generating an output token, the input sequence only contains the tokens before it. When passing through the first attention layer, the queries, values and keys after this token will be masked and will not participate in the attention calculation. The decoder's input in the current round, which is the input of the previous round appending the output of the previous round, generates the vector of the corresponding position after passing through the masked self-attention layer. This vector will be multiplied by a transition matrix to obtain the query matrix of the second attention layer, which is also called the 'cross-attention layer' (Fig. 5).

In the cross-attention layer, the key matrices and value matrices in the attention function are provided by the output sequence of the encoder, while the query matrix is transformed from the output of the masked attention layer. Calculating cross-attention is the same as self-attention, except that the source of the query matrix is different. The output of the cross-attention layer also goes through a feed-forward layer. After that, it will be fed into the last linear layer and the softmax function to produce the final output of the round.

3 Bioinformatics applications of transformer-based language models

This section summarizes and compares representative works in different fields of bioinformatics applications (Table 1), lists important works related to transformer (Fig. 6) and identifies their main focuses and benefits, e.g. improving model accuracy, reproducibility and interpretability. The number of transformer-based applications over the past 3 years (Fig. 7) suggests a growing interest in the field of bioinformatics.

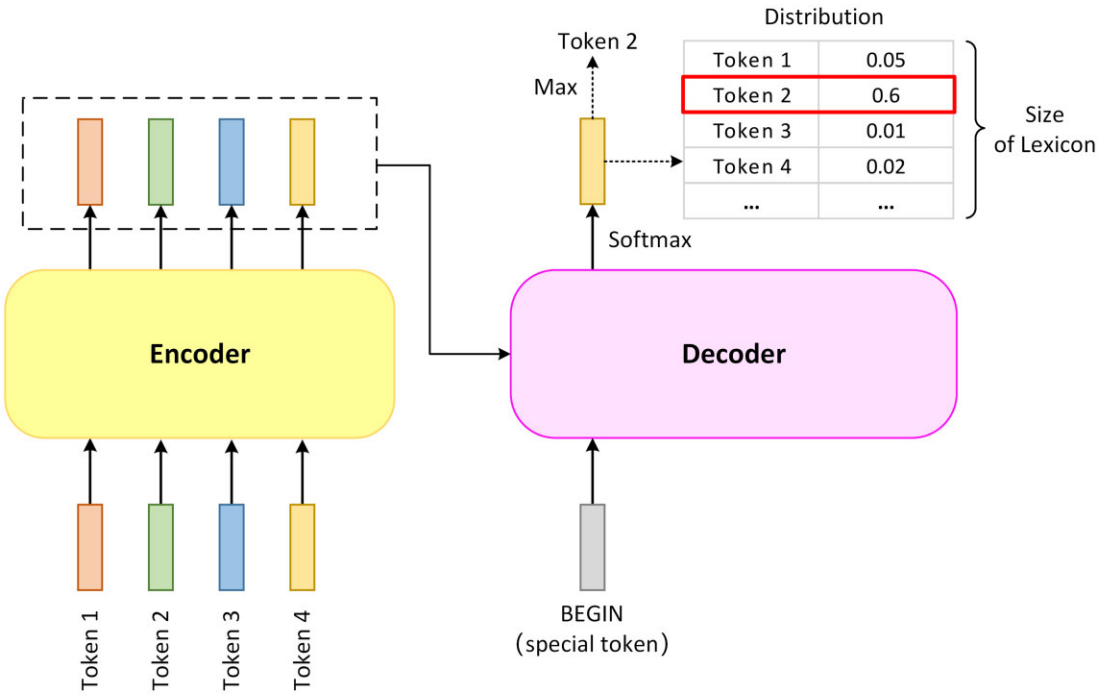

Fig. 3. The first step of the decoding process. The decoder predicts which token to output with its input and the output of the encoder. The decoder takes a special token 'BEGIN' as input, combining it with the encoder's output to generate the probability distribution vector. The length of this vector is the size of the lexicon, and each dimension of the output probability distribution vector represents the probability of a certain token in the lexicon. The output vector is then applied to a softmax function to generate a probability distribution, and the token in the lexicon with the highest probability is the corresponding output, which is also the first token in the final output sequence

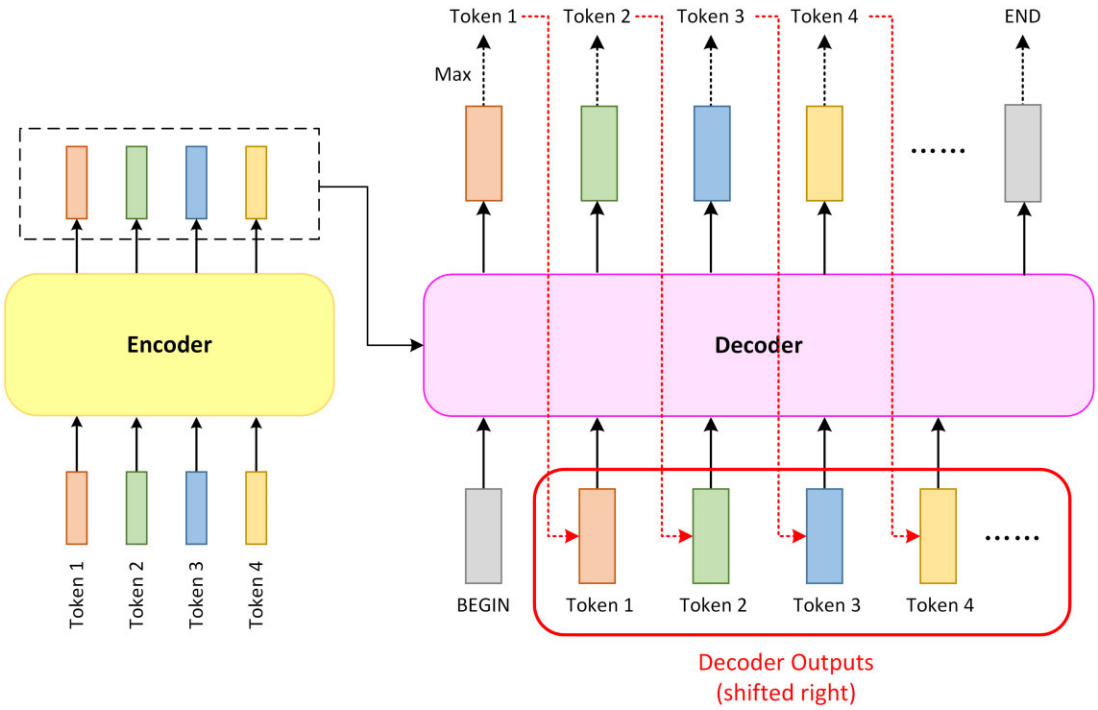

Fig. 4. The process of decoder looping to produce the final output one by one, also known as the autoregressive process. In the same way as the first decoding round shown in Figure 3, each round decoder will generate a predicted probability distribution vector. Predicted tokens generated by 'BEGIN' in the first round will be appended to the sequence containing the 'BEGIN' token as the second round of the decoding process's input. This process will be repeated, adding the new output into the input sequence. Because of the extra 'BEGIN' token, the decoder's input is shifted one position to the right. In each decoding round, only the predicted tokens already decoded so far together with the special token 'BEGIN' are received as input to generate the new predicted token. It is worth mentioning these predicted tokens are not always correct, and a misprediction in the current round may affect the decoding correctness in the subsequent rounds. To end the loop, an 'END' token is added to the lexicon. When the output token is 'END' the loop stops, resulting in the complete final output sequence

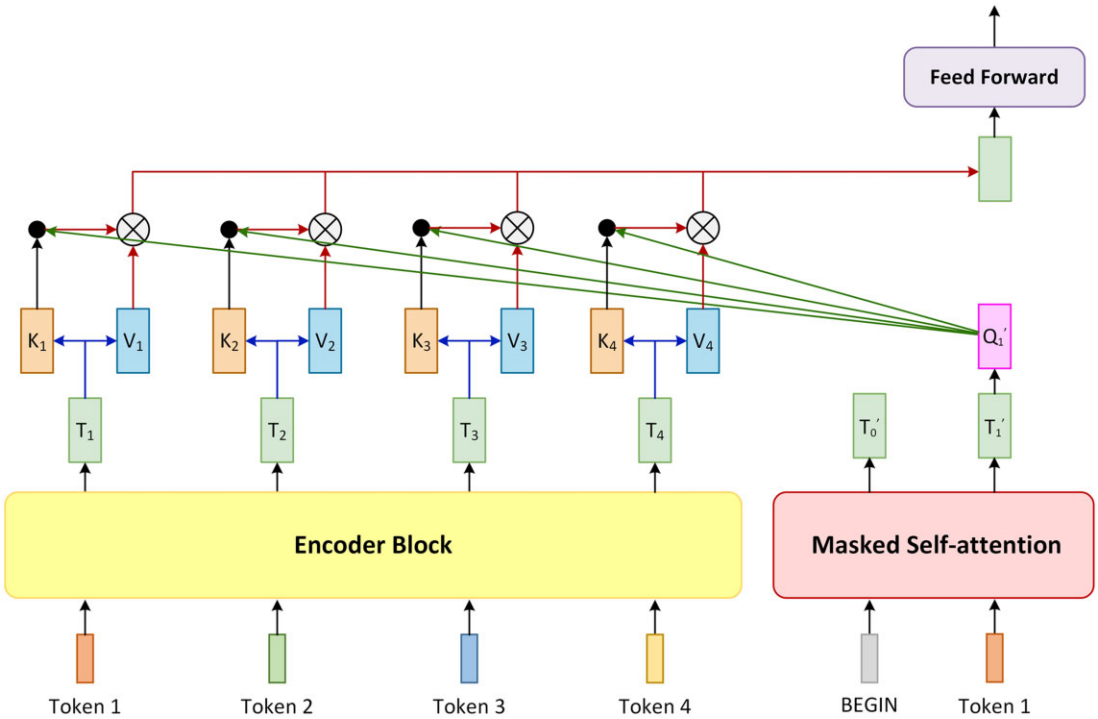

Fig. 5. Structure of the cross-attention layer. The encoder block in this figure refers to a certain block in encoder whose output participate in cross-attention with the decoder. Masked self-attention refers to the first attention sub-layer in decoder block. $T_{i(i=1,2,3,4)}$ is the ith token's output of the encoder block shown in this figure and also the ith token's input of next encoder block. $K_{i(i=1,2,3,4)}$ and $V_{i(i=1,2,3,4)}$ are the key matrix and the value matrix of T_i. $Q_1^{'}$ is the corresponding query matrix of $T_1^{'}$, which is the first token's output of masked self-attention. Cross-attention uses the decoder's query and the encoder's keys and values to calculate the attention function, and the output of cross-attention will be fed into the feed-forward layer in decoder block

Table 1. Summary and comparison of the representative applications of transformer-based language models in different fields of bioinformatics

Field	Paper	Pre-trained model? (Y/N)	Main focus	Data repositories address
Sequence analysis	Ji *et al.* (2021)	Y	Novel pre-trained bi-directional encoder representations that achieved state-of-the-art results in predicting promoters and identifying TFBSs	https://github.com/jeryj1993/DNABERT
	Lee *et al.* (2021)	Y	A transformer architecture based on BERT and 2D CNN to identify DNA enhancers from sequence information	https://github.com/khanhlee/bert-enhancer
	Zhang *et al.* (2021b)	Y	A transformer architecture based on BERT and stacking ensemble to identify RNA N7-Methylguanosine sites from sequence information	NA
	Charoenkwan *et al.* (2021)	Y	A bi-directional encoder representation from BERT-based model for improving the prediction of bitter peptides from the original amino acid sequence	http://pmlab.pythonanywhere.com/BERT4Bitter
	Qiao *et al.* (2022)	Y	Prediction of lysine crotonylation sites by a transfer learning method with pre-trained BERT models	http://zhulab.org.cn/BERT-Kcr_models/
Genome analysis	Clauwaert *et al.* (2021)	N	A prokaryotic genome annotation method based on the transformer-XL neural network framework for identifying TSSs in Escherichia coli	https://github.com/jdcla/DNA-transformer
	Raad *et al.* (2022)	N	A full end-to-end deep model based on transformer for prediction of pre-miRNAs in genome-wide data	https://github.com/sinc-lab/miRe2e
	Chen *et al.* (2022b)	N	Prediction of EPI in different cell types by capturing large genome contexts	https://github.com/biomed-AI/TransEPI
	Baid *et al.* (2022)	N	A gap-aware transformer–encoder for sequence correction trained by an alignment-based loss	https://github.com/google/deepconsensus
	Mo *et al.* (2021)	Y	Prediction of interactions between regulatory elements by pre-training large-scale genomic data in a multi-modal and a self-supervised manner	NA
Gene expression	Avsec *et al.* (2021)	N	A portmanteau of enhancer and transformer to predict gene expression and chromatin states from DNA sequences	https://github.com/deepmind/deepmind-research/tree/master/enformer
	Khan and Lee (2021)	N	Transformer for the gene expression-based classification of lung cancer subtypes that solved the complexity of high-dimensional gene expression through a multi-headed self-attention module	NA
	Yang *et al.* (2022)	Y	Single-cell bi-directional encoder representations from transformers for cell type annotation, new cell type discovery, handling of batch effects, and improving model interpretability.	https://github.com/TencentAILabHealthcare/scBERT
Proteomics	Cao and Shen (2021)	N	A high-throughput Transformer-based protein function annotator with both accuracy and generalizability	https://github.com/Shen-Lab/TALE

(continued)

Table 1. (continued)

Field	Paper	Pre-trained model? (Y/N)	Main focus	Data repositories address
	Rao *et al.* (2020)	Y	An alternative to MSA to predict inter-residue correlations in an end-to-end manner with Transformer protein language models	https://github.com/facebookresearch/esm
	Rives *et al.* (2021)	Y	Learning protein biological structure and function from UniRef dataset using pre-trained Transformer	https://github.com/facebookresearch/esm
	Zhang *et al.* (2021a)	N	Jointly considering information of all homologous sequences in MSA to capture global co-evolutionary patterns	https://github.com/microsoft/ProteinFolding/tree/main/coevolution_transformer
	Elnaggar *et al.* (2022)	Y	Understanding the language of life with transformer-based protein language models through self-supervised learning	https://github.com/agemagician/ProtTrans
	Brandes *et al.* (2022)	Y	A self-supervised deep language model specifically designed for proteins to capture local and global representations of proteins in a natural way	https://github.com/nadavbra/protein_bert
	Park *et al.* (2022)	Y	A sequence-based pre-trained BERT model improved linear and structural epitope prediction by learning long-distance protein interactions effectively	NA
	Ferruz *et al.* (2022)	Y	A pre-trained GPT-based model to generate sequences similar to natural proteins from scratch	https://huggingface.co/docs/transformers/main_classes/trainer
	Castro *et al.* (2022)	N	An autoencoder based on transformer with a highly structured latent space trained to jointly generate sequences and predict fitness	https://github.com/KrishnaswamyLab/ReLSO-Guided-Generative-Protein-Design-using-Regularized-Transformers
Multi-omics	Tao *et al.* (2020)	Y	Prediction of multiple cancer phenotypes based on somatic genomic alterations via the genomic impact transformer	https://github.com/yifengtao/genome-transformer
	Jurenaite *et al.* (2022)	Y	Applying Transformer-based deep neural network on mutomes and transcriptome counting for tumor type classification	https://github.com/danilexn/nebis
	Kaczmarek *et al.* (2021)	N	The use of graph transformer network (GTN) for cancer classification and interpretation	NA
	Ma *et al.* (2021)	N	Utilizing the heterogeneous graph transformer framework to infer cell type-specific single-cell biological networks	https://github.com/OSU-BMBL/deepmaps
Spatial transcriptomics	Pang *et al.* (2021)	N	Usage of Vision Transformer (ViT) to predict super-resolution gene expression from histology images in tumors	https://github.com/maxpmx/HisToGene
	Zeng *et al.* (2022)	N	Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks	https://github.com/biomed-AI/Hist2ST
Biomedical informatics	Lee *et al.* (2020)	Y	The first pre-trained biomedical language representation model for biomedical text mining	https://github.com/dmis-lab/biobert

(continued)

Table 1. (continued)

Field	Paper	Pre-trained model? (Y/N)	Main focus	Data repositories address
	Rasmy *et al.* (2021)	Y	Pre-training contextualized embeddings on large-scale structured electronic health records for disease prediction that used the International Classification of Diseases (ICD) codes	https://github.com/ZhiGroup/Med-BERT
	Wang *et al.* (2021)	Y	An innovative AlBERT-based causal inference model of clinical events	https://github.com/XingqiaoWang/DeepCausalPV-master
	Chen *et al.* (2021a)	Y	A powerful alternative to mainstream medical image segmentation methods that combined transformer and U-Net	https://github.com/Beckschen/TransUNet
	Chen *et al.* (2021b)	Y	Using ViT for the first time in self-supervised volumetric medical image registration	https://bit.ly/3bWDynR
Drug discovery	Wang *et al.* (2019)	Y	A pioneer pre-training method for molecular property prediction by pre-trained on unlabeled SMILES strings	https://github.com/uta-smile/SMILES-BERT
	Rong *et al.* (2020)	Y	A new GNN/Transformer architecture that learned rich molecular structure and semantic information from large amounts of unlabeled data	https://github.com/tencent-ailab/grover
	Chithrananda *et al.* (2020)	Y	Utilizing RoBERTa-based transformer for molecular property prediction	https://huggingface.co/seyonec
	Wu *et al.* (2022)	Y	Presenting new pre-training strategies that allowed the model to extract molecular features directly from SMILES	https://github.com/wzxxxx/Knowledge-based-BERT
	Li *et al.* (2022)	Y	A novel knowledge-guided pre-training framework of graph transformer for molecular property prediction	https://github.com/lihan97/KPGT
	Huang *et al.* (2021)	N	Improving the prediction accuracy of DTI by knowledge-inspired representation, interaction modeling modules and an augmented transformer encoder	https://github.com/kexinhuang12345/moltrans
	Kalakoti *et al.* (2022)	N	A modular framework that employing transformer-based language models for DTI prediction	https://github.com/TeamSundar/transDTI
	Jiang *et al.* (2022)	N	An end-to-end deep transformer-based learning model that used cancer cell transcriptome information and chemical substructures of drugs to predict drug response	https://github.com/jiangli kun/DeepTTC
	Bagal *et al.* (2022)	Y	A small version of the GPT model for molecular generation	https://github.com/devalab/molgpt
	Grechishnikova (2021)	N	A *de novo* drug generation model based on transformer architecture	https://github.com/dariagrechishnikova/molecule_structure_generation

Note: The papers are sorted by their appearance in this review and divided into different categories based on their research field.

3.1 Sequence analysis

Biological sequence analysis, including DNA, RNA and protein sequence analysis, represents one of the fundamental applications of computational methods in molecular biology. Traditional sequence analysis methods rely heavily on k-mers frequency (Koonin and Galperin, 2003b), which is not able to capture distant semantic relationships of gene regulatory code. Deep learning models like CNN also have problems capturing semantic dependency within long-range contexts (Tang *et al.*, 2018), as their capability to extract local features is limited by the filter size. The RNN-based models (e.g. LSTM and GRU) are developed to capture long-range dependency; however, it is difficult for them to perform large-scale learning due to their limited degree of parallelization. In addition, existing models generally require large amounts of labeled data, which is difficult to obtain in bioinformatics research (Butte, 2001).

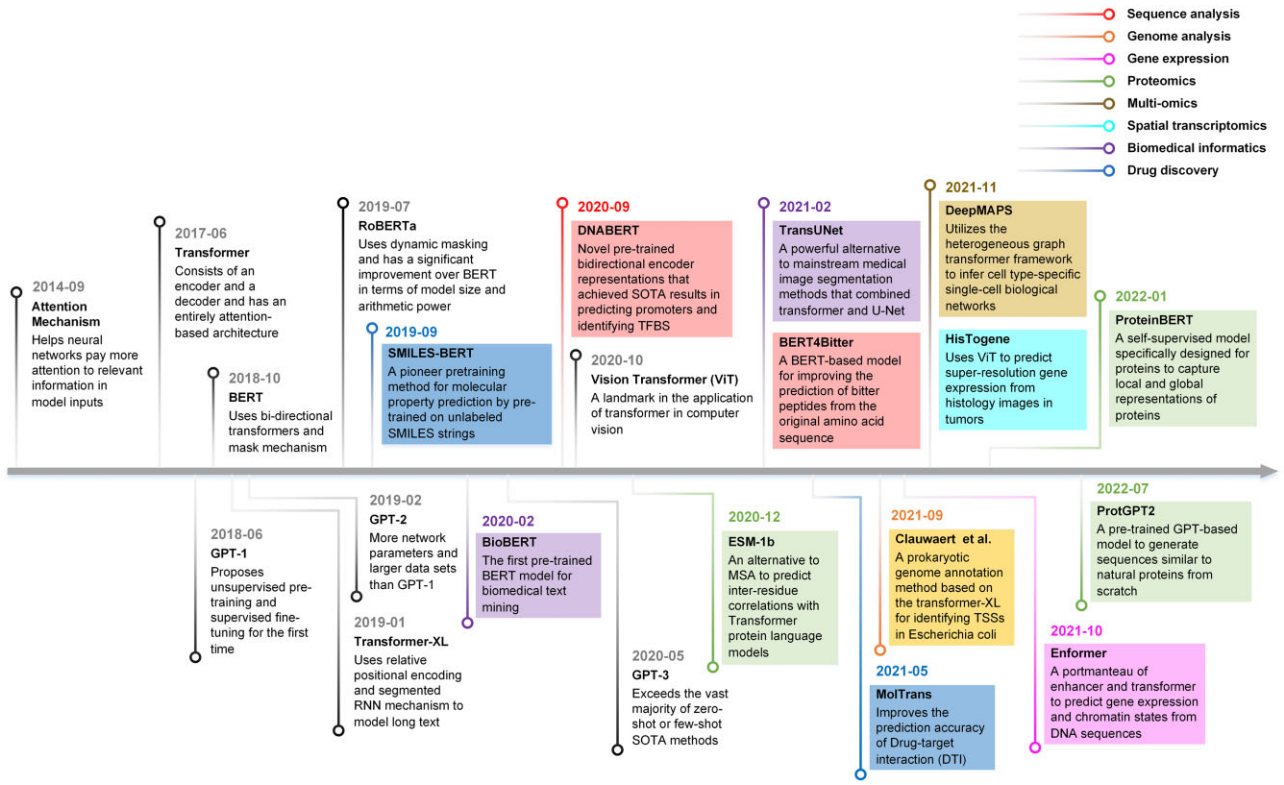

Fig. 6. An overview of important works related to TRANSFORMER. Different bioinformatics application models are represented chronologically by different colored lines. Following the prominent progress in the past years, Transformer in bioinformatics will embrace great advancement in the upcoming years. SOTA, state-of-the-art

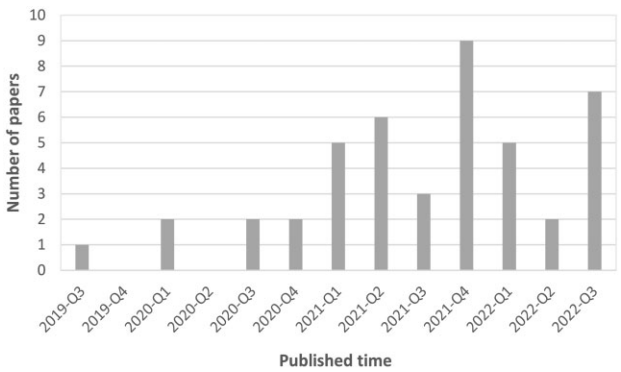

Fig. 7. Distribution of selected papers published in recent years. Most papers (84.1%) were published after 2021, with the highest number of publications registered in 2021 (23 papers). Qx, xth quarter of the year

Considering the large amount of unlabeled genomic sequences, transformer-based pre-trained language models are well-suited for DNA sequence analysis and have received increasing attention for their significant improvement over other traditional or deep learning models. DNABERT (Ji *et al.*, 2021), a novel pre-trained bi-directional coding representation, used tokenized k-mer sequences as input for the BERT model (Fig. 8A). DNABERT utilized context information in DNA sequences and achieved state-of-the-art results in downstream tasks such as predicting promoters and identifying transcription factor binding sites (TFBSs). Another example is to use the multi-language model based on BERT by converting DNA sequences into a numerical matrix of constant size for the prediction of enhancers (Lee *et al.*, 2021). Compared with the most advanced features in bioinformatics, BERT-based features increased the sensitivity, specificity, accuracy and Matthews correlation coefficient (MCC) by 5–10%.

Compared with DNA sequences, RNA sequences provide additional transcription information. While traditional methods still rely on manually curated RNA sequence features, deep learning models enable automatic feature extraction (Urda *et al.*, 2017). BERT-m7G was a transformer model based on BERT and used a stacking ensemble to identify RNA N7-methylguanosine (m7G) sites from RNA sequence information (Zhang *et al.*, 2021b). N7-methylguanosine is one of the most prevalent RNA post-transcriptional modifications and plays an important role in the regulation of gene expression. The experimental results showed that the identification performance of BERT-m7G obviously exceeded the existing prediction methods, with the accuracy increasing by 3–20.7% and the MCC improving by 0.06–0.415.

Protein sequence analysis can be regarded as an extension of DNA sequence analysis (von Heijne, 1992), but it is much more complicated than DNA sequence analysis because polymers are composed of 20 amino acids (Karlin and Ghandour, 1985). The analysis of protein sequences can better capture the relationships between protein sequences and the spatial structure of proteins and provide a theoretical basis for further study on protein function and structure (Findlay *et al.*, 1995; Ponting and Birney, 2005). For example, bitter peptides are oligopeptides with a bitter taste usually produced during food fermentation and protein hydrolysis (Karametsi *et al.*, 2014), which are useful for drug development since diluting the bitterness of drugs can increase patients' willingness to take medicine. BERT4Bitter was proposed to predict bitter peptides directly from the original amino acid sequence without using any structural information (Charoenkwan *et al.*, 2021). It was the first study to identify bitter peptides using the NLP-inspired model and feature encoding. In another study, Qiao *et al.* (2022) established a more effective predictor for protein lysine crotonylation sites (Kcr), which is one of the most important post-translational modifications, by pre-training BERT model. The authors converted each amino acid into a word as the input to the pre-trained BERT model. The features encoded by BERT were extracted and then fed to the BiLSTM network (Zeng *et al.*, 2016) to construct the final model.

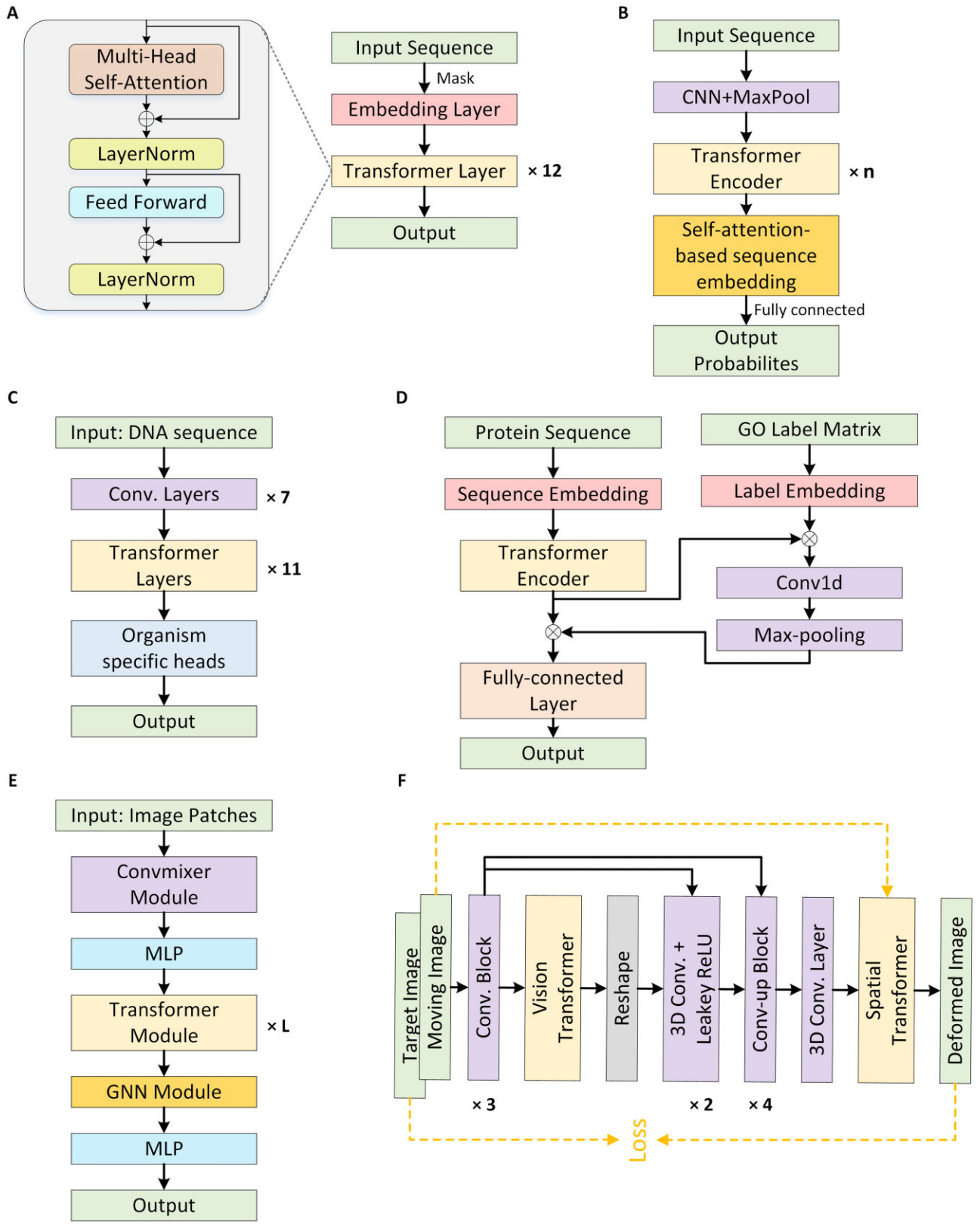

Fig. 8. Several typical models of Transformer applied to bioinformatics including the frameworks of (**A**) DNABERT, (**B**) TransEPI, (**C**) Enformer, (**D**) TALE, (**E**) Hist2ST and (**F**) ViT-V-Net

3.2 Genome analysis

Although sequence analysis contributes significantly to biological discovery, genome analysis is also essential to capture the full repertoire of information encoded in the genome (Koonin and Galperin, 2003a). Genome analysis explains the appearance of tumors or phenotypes from the DNA level, including gene mutations, deletions, amplifications (Feuk *et al.*, 2006) and epigenetic modifications (e.g. DNA methylation) (Nikpay *et al.*, 2015; Portela and Esteller, 2010).

Several scratch-trained methods based on the Transformer model have been developed to this end. For example, Clauwaert *et al.* (2021) proposed a prokaryotic genome annotation method based on the Transformer-XL neural network framework, which was designed to identify transcription start sites (TSSs) for the transcription process in *Escherichia coli*. Beyond the application to genome annotation, some studies also applied Transformers to the prediction of small-RNA sequences. For example, MiRe2e, a full transformers-based end-to-end deep model, was developed to predict pre-miRNAs (Raad *et al.*, 2022). MiRe2e showed its advantages in two aspects: (i) It can receive raw genome-wide data without any preprocessing or secondary structure prediction; (ii) It identified all pre-miRNA sequences in the genome with high accuracy and recall. In another study, TransEPI (Chen *et al.*, 2022b) was developed based on enhancer–promoter interaction (EPI) datasets derived from Hi-C or ChIA-PET data to predict EPI in different cell types by capturing large genome contexts (Fig. 8B). This model not only achieved state-of-the-art results on experimental datasets [the area under the precision-recall curve (auPRC) of TransEPI increased by an average of 28.1% compared to the second-best baseline] but has also been extended to the interpretation of disease-related non-coding mutations. Last but not least, Google's Andrew Carroll research group recently developed DeepConsensus, which uses the alignment-based loss to train gap-aware transformer-encoders for sequence correction (Baid *et al.*, 2022). Compared to methods using pbccs (standard approach to consensus generation), DeepConsensus reduced errors in reads (small genome fragments from sequencing sampling) by 41.9%, and improved the adjacency, completeness and correctness of genome assembly.

In addition, the Transformer-based pre-trained models were also used to predict the interactions between regulatory elements. One example is GeneBERT (Mo *et al.*, 2021). It was proposed to address the problem that traditional methods rarely consider the interactions among multiple regulatory elements in the regulatory genome. GeneBERT was pre-trained using large-scale genomic data in a multi-modal and self-supervised manner, in which three pre-training tasks: sequence pre-training, region pre-training and sequence-region matching, were proposed to improve the robustness and generalization ability of the model.

3.3 Gene expression

Gene expression data (Brazma and Vilo, 2000), like RNA-sequencing (RNA-seq) and single-cell RNA-seq (scRNA-seq) (Kolodziejczyk *et al.*, 2015; Ozsolak and Milos, 2011), has been extensively studied to better understand complex diseases and to identify biomarkers that can guide therapeutic decision-making (Goeman and Bühlmann, 2007). They have substantial applications in clinical medical diagnosis, drug efficacy judgment and revealing the mechanism of disease (Rotter *et al.*, 2010; Rung and Brazma, 2013).

To examine how non-coding DNA determines gene expression in different cell types, DeepMind proposed a noteworthy model Enformer (Avsec *et al.*, 2021). Due to the limitations of previous convolutional operations in modeling the effects of distal enhancers and insulators on gene expression, Enformer introduced the transformer structure (Fig. 8C), greatly increasing the receptive field of the network (from 20 to 100 kb). Enformer not only greatly improved the accuracy of predicting gene expression from DNA sequences, with the mean correlation increasing from 0.81 to 0.85, but also represented an important step forward in human understanding of the complexity of genome sequences. Furthermore, Enformer predicted EPI directly from DNA sequences by leveraging the self-attention mechanism and provided a more accurate prediction of mutation effects through direct mutation analysis and population eQTL studies (Liu *et al.*, 2022).

In addition to predicting the effect of non-coding DNA on gene expression, transformer-based models have been widely used to predict cancer subtypes according to gene expression data. Gene transformer used the multi-headed self-attention module to solve the complexity of high-dimensional gene expression for joint classification of lung cancer subtypes (Khan and Lee, 2021). Compared with traditional classification algorithms, the proposed model achieved an overall performance improvement in all evaluation metrics, with 100% accuracy and zero false-negative rates on most datasets.

scRNA-seq is a revolutionary technology in the life science field. One of the latest studies innovatively proposed scBERT model for single-cell annotation (Yang *et al.*, 2022). It was the first time to apply Transformer in scRNA-seq data analysis. Following BERT's pre-training and fine-tuning paradigm, scBERT reused large-scale unlabeled scRNA-seq data to accurately capture the expression information of a single gene and the gene–gene interactions and revealed single-cell type annotation with high interpretability, generalization and stability.

3.4 Proteomics

The essential task of proteomics is to understand protein dynamics in complex systems and diseases (Larance and Lamond, 2015; Rix and Superti-Furga, 2009). Protein sequences can be viewed as a concatenation of letters from the amino acids, analogously to human languages. These letters form secondary structural elements ('words'), which assemble to form domains ('sentences') that undertake a function ('meaning') (Ofer *et al.*, 2021). With the extraordinary advances in the NLP field in understanding and generating language with human-like capabilities, some language models open a new door to figuring out protein-related problems from sequences alone, such as protein sequence representation, post-translational modifications, protein function annotation and protein design.

Especially, transformer has served as a key technique for addressing various aspects of proteomics data analysis. The work of Cao and Shen (2021) exemplified the application of transformer to protein function annotation, a critical step in identifying the overall functional distribution of differentially expressed proteins. Specifically, the model obtained embedding by using sequence inputs, hierarchical function labels and their joint similarity to measure the contribution of each amino acid to each label. The final model was shown to be a high-throughput protein function annotator with high accuracy and generalizability (Fig. 8D).

The measurement of amino acid proximity of proteins is called the inter-residue contact map, which well characterizes the structural information of proteins. Most of the top-performing models for protein contact prediction use multiple sequence alignment (MSA), which improves protein 3D structure prediction by analyzing residue co-evolution information in sequences. Facebook AI Research proposed ESM-1b (Rao *et al.*, 2020), a method alternative to MSA using the transformer to predict inter-residue correlations in an unsupervised manner. Subsequently, they applied ESM-1b to the UniRef dataset (250M protein sequence) for biochemical properties analysis, secondary and tertiary structure prediction and mutation analysis to fully explore the rich information contained in protein sequences (Rives *et al.*, 2021). Since the prevalence of non-homologous residues and gaps in MSA may lead to erroneous estimation of residue co-evolution, Co-evolution Transformer (CoT) was proposed to reduce the impact of non-homologous information (Zhang *et al.*, 2021a). CoT selectively aggregated features from different homologous sequences by assigning smaller weights to non-homologous sequences or residue pairs. By jointly considering the information of all homologous sequences in MSA, CoT was able to capture global co-evolutionary patterns.

There are some important works related to protein sequence embedding in recent years (Alley *et al.*, 2019; Elnaggar *et al.*, 2022; Heinzinger *et al.*, 2019; Unsal *et al.*, 2022). Elnaggar *et al.* (2022) proposed to make transformer-based protein language models capture constraints relevant for protein structure and function by transfer learning (using trained embeddings as input to subsequent

supervised training). The researchers trained two auto-regressive models (Transformer-XL and XLNet) and four auto-encoder models (BERT, ALBERT, ELECTRA and T5) on large-scale protein sequences and tested both residue-level (3-state accuracy Q3 = 81–87%) and protein-level (10-state accuracy: Q10 = 81%, 2-state accuracy Q2 = 91%) prediction tasks using the embeddings obtained from the language models above, and found that ProtT5 fine-tuned on UniRef50 without MSA outperformed ESM-1b and achieved the best performance.

Other transformer-based pre-trained models have also been widely used in proteomics research. ProteinBERT is a model specifically designed for proteins (Brandes *et al.*, 2022). The pre-training scheme combined language modeling with gene ontology (GO) (Ashburner *et al.*, 2000; Stevens, 2000) annotation prediction. ProteinBERT aimed to capture local and global representations of proteins in a natural way, which allowed end-to-end processing of these types of input and output, making the model efficiently and flexibly adapt to long sequences. EpiBERTope (Park *et al.*, 2022) is a sequence-based pre-trained BERT model to predict both linear and structural epitopes. Epitopes are immunogenic regions of antigens that can be recognized by antibodies in a highly specific manner and trigger immune responses. EpiBERTope used a multi-headed attention mechanism to construct global dependencies for each amino acid in the protein sequences. In the fine-tuning stage, both linear and structural epitopes datasets were the input of EpiBERTope.

Beyond the applications mentioned above, transformer-based generative models began to be used for protein design in recent studies. Inspired by generative transformer-based language models (such as the GPT-X family), ProtGPT2 (Ferruz *et al.*, 2022) could generate sequences similar to natural proteins from scratch and thereby possesses the potential to solve many biomedical and environmental problems. Castro *et al.* (2022) proposed Regularized Latent Space Optimization (ReLSO), which combined the powerful encoding ability of the model with the capacity to generate low-dimensional latent representations with rich information. By simultaneously optimizing protein sequence generation and fitness landscape (Romero and Arnold, 2009) prediction, a latent space that contained rich information about sequence and fitness was explicitly created. In addition, the authors mentioned that ReLSO-like structures could be applied to other biomolecules such as DNA and RNA.

3.5 Multi-omics

The multi-omics analysis aims to better understand biological regulation by combining different types of omics data (Yang *et al.* 2019a). With the development of high-throughput sequencing technology, there is a growing interest in combining genomics with transcriptomics, proteomics and metabolomics together to understand the disease pathways and processes as a single type of omics data cannot capture the entire landscape of the complex biological networks (Castro-Vega *et al.*, 2015; Kang *et al.*, 2022).

The transformer-based model provides a new perspective for the analysis of various omics data in terms of diseases, while most conventional methods rarely take the relationships between different omics levels into account. To this end, Tao *et al.* (2020) proposed the genomic impact transformer (GIT). The GIT fine-tuned gene embeddings that were pre-trained by the 'Gene2Vec' algorithm in order to infer how somatic genomic alterations (SGAs) affect the function of cellular signaling systems and thus cause cancer by modeling the statistical relationship between SGAs events and tumor differentially expressed genes (DEGs). A recent article presented SetQuence and SetOmic (Jurenaite *et al.*, 2022), which applied transformer-based deep neural networks on mutome and transcriptome together, showing superior accuracy and robustness over previous baselines (including GIT) on tumor classification tasks.

Several applications in multi-omics made use of graph transformer networks (GTN) (Yun *et al.*, 2019). For instance, a novel method for cancer classification and interpretation (Kaczmarek *et al.*, 2021) could correctly model and interpret the interaction and biological communication between miRNAs and mRNAs to discover important miRNA-mRNA cancer pathways. Notably, although GTN was not superior to other baselines like GCN (Zitnik *et al.*, 2018), SVM (Cortes and Vapnik, 1995) and MLP (Kothari and Oh, 1993), it provided a high degree of interpretation of the results, as the attention of GTN could identify potential targeting pathways and biomarkers, which is almost impossible to be achieved by other models. DeepMAPS was a deep learning-based single-cell multi-omics data analysis platform that utilized the heterogeneous graph transformer framework to infer cell type-specific single-cell biological networks (Ma *et al.*, 2021). DeepMAPS can include all cells and genes in a heterogeneous graph to infer cell–cell, gene–gene and cell–gene relationships simultaneously.

3.6 Spatial transcriptomics

Spatially resolved transcriptomics has experienced significant progress in the biomedical research field with advances in imaging and next-generation sequencing technology (Reis-Filho, 2009). The relationship between cells and their relative positions in tissue samples is crucial for identifying intercellular communication networks and global transcriptional patterns, and understanding disease pathology. While single-cell transcriptome sequencing techniques address the issue of cell heterogeneity and allow us to identify cellular variants that play key roles in diseases (Faridani *et al.*, 2016), they cannot be targeted to specific spatial positions, resulting in the exploration of cell functions that are not yet particularly precise. Spatial transcriptomics not only provides information on the transcriptome data of the subject, but also locates its spatial location in the tissue, which is of great significance and thus provides a tremendous opportunity for many research fields such as oncology, neuroscience, immunology and developmental biology (Chen *et al.*, 2022a).

Transformer-based language models have been applied on this front to predict cell composition and gene expression in different areas of tissue. One example is HisTogene (Pang *et al.*, 2021), which employed Vision Transformer (ViT) (Dosovitskiy *et al.*, 2021), a state-of-the-art method for image recognition, to predict super-resolution gene expression from hematoxylin and eosin (H&E)-stained histology images. The model demonstrated favorable performance across datasets of 32 HER2+ breast cancer samples both in gene expression prediction and clustering tissue regions using the predicted expression. Based on this study, to capture 2D visual features of histology images and better highlight the explicit neighborhood relationships of image patches, the Hist2ST (Zeng *et al.*, 2022) model was developed for predicting RNA-seq expression from histology images (Fig. 8E). The model cropped histology images into patches at sequencing spots, learned 2D features in the image patches by convolutional operations and then captured global spatial dependencies between features using the transformer module while capturing explicit neighborhood relationships by graph neural networks (GNN) (Scarselli *et al.*, 2009). This study also proposed a self-distillation mechanism to mitigate the effects of small spatial transcriptomics data effectively.

3.7 Biomedical informatics

Biomedical informatics uses theories and techniques of computer science and other related disciplines' research methods for developing innovative research and application in biomedical and clinical medicine (Boguski and McIntosh, 2003; Sarkar, 2010). The success of transformer-based language models has led researchers to focus on biomedical text and medical image processing, which again shows the superior performance of the Transformer.

One of the applications in biomedical text processing is BioBERT (Lee *et al.*, 2020), the first pre-trained BERT model for biomedical corpora. BioBERT initialized weights from general domain pre-trained BERT, trained on a large-scale biomedical corpus and fine-tuned on biomedical text mining tasks including NER (Marrero *et al.*, 2013), RE (Zhang *et al.*, 2017) and QA (Calijorne Soares and Parreiras, 2020). To enable deep learning models to predict disease status using limited training data, another study proposed Med-BERT (Rasmy *et al.*, 2021), a contextualized embedding model for pre-training on structured electronic health records (EHRs) data. In contrast to other medical pre-trained models that

were trained on free text, this model was characterized by using the International Classification of Diseases (ICD) codes. After fine-tuning experiments on pancreatic cancer prediction and heart failure prediction in diabetic patients, Med-BERT was validated to be generalized on different sizes of fine-tuned training samples, which can better meet disease prediction research with small training datasets. Another promising application based on biomedical text data is an ALBERT-based model called InferBERT to predict clinical events and infer the causality (Wang *et al.*, 2021), which is a prerequisite for deployment in drug safety. As evaluated on two FDA Adverse Event Reporting System cases, the results showed that the number of causal factors identified by InferBERT for analgesics-related acute liver failure and Tramadol-related mortalities was 1.87 and 1.16 times higher than the second-best baseline, respectively.

Transformer has not only dominated the NLP field but has recently revolutionized the computer vision field (Han *et al.*, 2023; Khan *et al.*, 2022). Specifically, ViT applied Transformer to image classification tasks and achieved SOTA performance with less computational expense than other methods (Dosovitskiy *et al.*, 2021). Subsequent to this progress, TransUNet pioneered the pre-trained ViT for 2D medical image segmentation (Chen *et al.*, 2021a). It not only encoded image features as sequences to extract global context but also exploited low-level details for precise localization through a U-Net (Ronneberger *et al.*, 2015) hybrid network design. As a powerful alternative to mainstream medical image segmentation methods based on fully convolutional neural networks, TransUNet outperformed prior tools on tasks such as synapse multi-organ segmentation and cardiac segmentation, e.g. average dice score gained a range from 1.91% to 8.67%. ViT-V-Net (Chen *et al.*, 2021b) used ViT for the first time in self-supervised volumetric medical image (i.e. 3D images) registration (Fig. 8F). Combining the advantages of Transformer and V-Net (Milletari *et al.*, 2016), the network learned long-distance relationships between points in images while maintaining the flow of localization information.

3.8 Drug discovery

Despite progress in technology and enhanced knowledge of human disease, the translation of these advances into therapeutic benefits has been far slower than expected. The challenges facing the global pharmaceutical industry are multifold, including high attrition rates, increased time to bring new drugs to the market and changing regulatory requirements, which can all contribute to higher costs. A key issue in the early stage of drug design and discovery is the prediction of molecular properties and interactions (Lo *et al.*, 2018). While deep learning models have been widely applied to this end (Feinberg *et al.*, 2018; Liu *et al.*, 2019a; Wu *et al.*, 2018), the scarcity of labeled data remains a fundamental obstacle to accurate and efficient molecular property prediction. For this reason, large amounts of unlabeled data have been considered to improve the prediction performance on small-scale labeled data with the strength of transformer-based self-supervised pre-training.

Several momentous pre-training methods for molecular property prediction have been proposed, including SMILES-BERT (Wang *et al.*, 2019), GROVER (Rong *et al.*, 2020), ChemBERTa (Chithrananda *et al.*, 2020), K-BERT (Wu *et al.*, 2022) and KPGT (Li *et al.*, 2022). SMILES-BERT was pre-trained on large-scale unlabeled data by a Masked SMILES Recovery task by converting molecular formulas into SMILES strings (a kind of single-line text representation for the structure of molecular compounds) as input sequences (Wang *et al.*, 2019). The pre-trained model was fine-tuned with the labeled datasets and achieved excellent results on many datasets. However, SMILES-BERT lacks model interpretability since SMILES is not topology-aware and cannot explicitly encode the structural information of molecules. GROVER integrated Dynamic Message Passing Networks (Gilmer *et al.*, 2020) from GNNs and long-range residual connection into Transformer architecture to provide a more expressive molecular encoder and demonstrated clear improvement in molecular classification and regression tasks (Rong *et al.*, 2020). ChemBERTa utilized RoBERTa-based Transformer and evaluated the model with ROC–AUC metrics for MoleculeNet tasks (Chithrananda *et al.*, 2020). Although the experimental result was not state-of-the-art, ChemBERTa could scale the pre-training dataset well, with powerful downstream performance and practical attention-based visualization modality. K-BERT (Wu *et al.*, 2022) presented new pre-training strategies that allowed the model to extract molecular features directly from SMILES. The atomic feature prediction task enabled K-BERT to learn the initial atomic information that was extracted manually in graph-based approaches, the molecular feature prediction task enabled K-BERT to learn the molecular descriptor/fingerprint information that was extracted manually in descriptor-based approaches, and the contrastive learning task enabled K-BERT to better 'understand' SMILES through making the embeddings of different SMILES of the same molecule more similar. To alleviate the issues of the unclear definition of pre-training tasks and limited model capacity, Li *et al.* (2022) introduced KPGT, i.e. Knowledge-guided Pre-training of Graph Transformer for molecular graph representation learning and achieved state-of-the-art performance. KPGT proposed the Line Graph Transformer, which is a high-capacity model to emphasize the importance of chemical bonds and model the structural information of molecular graphs as line graphs. A knowledge-guided pre-training strategy based on generative self-supervised learning was then designed to exploit the molecular descriptors/fingerprints to guide the model to obtain plentiful structural and semantic information from large-scale unlabeled molecular graphs.

In addition to its role in molecular property prediction, transformer has been used in a wide range of applications to predict the interaction between biomolecules and compounds, e.g. drug–targeting interaction (DTI), which is a fundamental task for in silico drug discovery. Huang *et al.* (2021) proposed Molecular Interaction Transformer (MolTrans) to improve the accuracy of DTI prediction. With knowledge-inspired representation, interaction modeling modules and an augmented transformer encoder, MolTrans could extract semantic relationships between substructures from large amounts of unlabeled biomedical data. A recent study presented TransDTI (Kalakoti *et al.*, 2022), a modular framework that employs transformer-based language models to predict DTIs. TransDTI outperformed other descriptors and existing models including MolTrans. More recently, DeepTTA was released, which used cancer cell transcriptome information and chemical substructures of drugs to predict drug response (Jiang *et al.*, 2022). The model utilized transformers to mine drug features from substructures and a four-layer neural network to predict the transcriptomic data of anticancer drug response, making it easier to find effective cancer therapeutic drugs.

The generative models can produce molecules similar to but different from those in the training set by learning the distribution of the molecules in the training set. Another important development is that the transformer-based generative modeling brings new ideas to drug design. MolGPT is a small version of the GPT model for molecular generation (Bagal *et al.*, 2022). The model used masking self-attention mechanisms to make it easier to capture the long-range dependencies. In order to reduce the dependence on prior knowledge, such as the physical and chemical characteristics of proteins in the process of drug discovery, Grechishnikova (2021) proposed a *de novo* drug generation model based on transformer architecture. The goal of this model is to generate realistic lead compounds only using the amino acid sequence information of the target protein.

4 Challenges and opportunities

In this subsection, we discuss several key challenges and opportunities when applying transformer-based language models in bioinformatics research.

4.1 Heterogeneous training data

The rapid development of various types of omics technologies represented by high-throughput sequencing and mass spectrometry (Noor *et al.*, 2021) has made bioinformatics research obtain powerful data as input, with the result that the input of transformer in bioinformatics is not the same as it was originally applied in NLP. Instead,

there is heterogeneous information, including text, code, graphs, etc. To fully capture the information in these heterogeneous data, both in-depth data preprocessing and model adaption may be needed. For instance, biological sequence and genomic feature information is generally textual, e.g. in FASTQ, BED and SRA formats. Such data can be directly fed to the transformer by word embedding or character embedding techniques (Chen *et al.*, 2022b; Ji *et al.*, 2021; Rives *et al.*, 2021); patient visit information (including disease, medication and clinical records) is represented as sequences of codes, such as EHR, ICD, where the code sequences are mapped to vector sequences in the application (Li *et al.*, 2020; Meng *et al.*, 2021; Rasmy *et al.*, 2021); the biomedical field involves images that are generally reshaped into sequences of patches for tokenization and mapped into a latent space using a trainable linear projection (Chen *et al.*, 2021a, b).

Furthermore, much more attention should be paid to multimodal learning (MML). Recently, the studies of MML with Transformer have made great progress in the field of NLP and computer vision (Chen *et al.*, 2020; Lu *et al.*, 2019; Zheng *et al.*, 2021). Since Transformer can work in a mode-independent manner, it can extract and related information from multimodal data by fusion (or alignment) of the input token embeddings of self-attention (Radford *et al.*, 2021; Xu *et al.*, 2022). Making use of biomedical codes, medical images, waveforms and genomics in pre-training models would be beneficial but requires in-depth studies of multimodal transformers.

4.2 Computational expense

The large amount of high-throughput sequencing data has led to the fact that many labs currently spend more on storage and computation, and the calculation and mining of massive amounts of data have become a major bottleneck for downstream studies. The powerful performance of the transformer comes largely from self-attention, which leads to the huge computational expense and makes transformer unable to model long sequences. Many efforts have been made to improve the transformer for this problem:

1. Improvements based on recursive connection: Transformer-XL (Dai *et al.*, 2019) proposed segment-level RNN mechanism and relative positional encoding to model long-distance dependence.
2. Improvements based on sparse attention: For example, Longformer (Beltagy *et al.*, 2020) proposed sliding windows, dilated sliding windows and global attention strategies to reduce the complexity of the model; Big Bird (Zaheer *et al.*, 2020) added random attention and introduced prior knowledge to limit the scope of attention and enhance efficiency; Reformer (Kitaev *et al.*, 2020) computed the Q and K matrices using the same linear layer parameters and calculated the attention score separately for each query, changing the storage expense to the square root level of the original.
3. Improvements based on low-rank decomposition: Linformer (Wang *et al.*, 2020) proposed singular value decomposition of the calculated attention matrix to transform the complexity from square to linear.
4. Improvements based on linear attention: Such as Linear Transformer (Katharopoulos *et al.*, 2020) and Performer (Choromanski *et al.*, 2021) replaced softmax with other mappings, making the multiplication complexity of Q, K and V matrices $O(N)$.

In addition, Zhang *et al.* (2020) proposed Scale-dot Product Attention for dimensionality in TensorCoder, which reduced the computational expense from $O(N^{2d})$ to $O(Nd^2)$. When the sequence length (N) is greater than the word vector dimension (d), it can reduce the costs. Given the increasing volume of data and the complexity of analysis, developing more efficient transformer models and architectures will be another crucial direction not only for machine learning but also for bioinformatics research.

4.3 Model interpretability

A common criticism of deep learning models is their lack of interpretability. However, the model interpretability analysis is particularly vital when the dimension of original features is too high. Especially in the field of bioinformatics, gaining insight from the model is critical since having an interpretable model of a biological system may lead to hypotheses that can be validated experimentally. The self-attention mechanism in Transformer has notable advantages in this direction. For example, through the analysis of attention maps, DNABERT (Ji *et al.*, 2021) could visualize important areas that contributed to model decision-making, thereby improving the interpretability of the model. Expect for prediction, DNABERT could directly rank the importance of the input nucleotide molecules and analyze the relationship between the input sequence contexts, resulting in better visualization information and accurate motifs extraction. Most of the attention heads of the Transformer-XL-based network architecture (Clauwaert *et al.*, 2021) could successfully identify and characterize transcription factors' binding sites and consensus sequences, which showed that transformer has unique potential for genome annotation tasks and biological significance extraction. Reflecting the contribution of each gene and the interaction between gene pairs by self-attention mechanism, scBERT (Yang *et al.*, 2022) can obtain the top attention genes corresponding to a specific cell type, which is important for cell type annotation. The attention mechanism in DeepMAPS enhanced biological interpretability by fully capturing complex molecular mechanisms and cellular heterogeneity (Ma *et al.*, 2021). And the attention of GTN could identify potential miRNA-mRNA targeting pathways and biomarkers, which is not easy or even impossible to be achieved by other models (Kaczmarek *et al.*, 2021). Interpretability makes the model itself, rather than results or data, become the source of knowledge. How to better utilize the self-attention mechanisms to demonstrate the biological insight behind the models will become one of the most desirable improvements in transformer-based applications in bioinformatics.

5 Conclusion

The recent development of transformer-based language models has substantially enriched the NLP field with novel architectures of self-attention that can greatly improve model accuracy, efficiency and interpretability. As a new potential force, transformer-based models have brushed up on SOTA performance with a large margin in most bioinformatics tasks. For example, the precision of GeneBERT in promoter classification, TFBS classification and disease risks estimation tasks was 0.130, 0.674 and 0.510 higher than that of the second-best method, respectively; the accuracy of scBERT in the prediction of novel and known cell types increased by 0.155 and 0.158, respectively; ESM-1b increased precision on secondary structure and contact predictions by 0.092 and 0.279; InferBERT almost doubled the number of identified causal factors on acute liver failure (from 23 to 43). Although several models did not reach SOTA in terms of evaluation metrics, such as GTN and ChemBERTa, they also made significant breakthroughs, and they were still innovative for other properties, such as the robustness to high-dimensional, small sample size and heterogeneous data.

Nevertheless, the development and application of transformers in bioinformatics are still in their infancy. There are many directions for further exploration, such as developing better pre-training methods, improving model flexibility, standardizing benchmarks and mitigating bias. Research in these directions will improve the analysis and interpretation of transformer-based models, and help the research community to utilize various biological data effectively. We hope this review article sparks thoughts on transformer-based language models across multiple disciplines and will inspire future research and applications that revolutionize biological and biomedical research and open up new avenues for the diagnosis and treatment of human diseases.

Funding

This work was supported by National Natural Science Foundation of China [62003178]; and the National Key Research and Development Program of China [2021YFF1001000].

Conflict of Interest: none declared.

References

Adel,H. *et al.* (2018) Overview of character-based models for natural language processing. In: Gelbukh,A. (ed.) *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science*. Springer International Publishing, Cham, pp. 3–16.

Alley,E.C. *et al.* (2019) Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods*, **16**, 1315–1322.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Avsec,Ž. *et al.* (2021) Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, **18**, 1196–1203.

Ba,J.L. *et al.* (2016) Layer normalization. arXiv, arXiv:1607.06450v1, https://arxiv.org/abs/1607.06450v1, preprint: not peer reviewed.

Bagal,V. *et al.* (2022) MolGPT: molecular generation using a transformer-decoder model. *J. Chem. Inf. Model.*, **62**, 2064–2076.

Bahdanau,D. *et al.* (2016) Neural machine translation by jointly learning to align and translate. arXiv, arXiv:1409.0473v7, https://arxiv.org/abs/1409.0473v7, preprint: not peer reviewed.

Baid,G. *et al.* (2022) DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nat. Biotechnol.*, 1–7.

Beltagy,I. *et al.* (2020) Longformer: the long-document transformer. arXiv, arXiv:2004.05150v2, https://arxiv.org/abs/2004.05150v2, preprint: not peer reviewed.

Bengio,Y. *et al.* (2003) A neural probabilistic language model. *J. Mach. Learn. Res.*, **3**, 1137–1155.

Bengio,Y. *et al.* (1994) Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.*, **5**, 157–166.

Blacoe,W. and Lapata,M. (2012) A comparison of vector-based representations for semantic composition. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, Jeju Island, Korea, pp. 546–556.

Boguski,M.S. and McIntosh,M.W. (2003) Biomedical informatics for proteomics. *Nature*, **422**, 233–237.

Brandes,N. *et al.* (2022) ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, **38**, 2102–2110.

Brazma,A. and Vilo,J. (2000) Gene expression data analysis. *FEBS Lett.*, **480**, 17–24.

Brown,T. *et al.* (2020) Language models are few-shot learners. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vol. 33, Curran Associates Inc., Vancouver, Canada, pp. 1877–1901.

Butte,A.J. (2001) Challenges in bioinformatics: infrastructure, models and analytics. *Trends Biotechnol.*, **19**, 159–160.

Calijorne Soares,M.A. and Parreiras,F.S. (2020) A literature review on question answering techniques, paradigms and systems. *J. King Saud Univ. Comput. Inf. Sci.*, **32**, 635–646.

Cao,Y. and Shen,Y. (2021) TALE: transformer-based protein function annotation with joint sequence-label embedding. *Bioinformatics*, **37**, 2825–2833.

Casola,S. *et al.* (2022) Pre-trained transformers: an empirical comparison. *Mach. Learn. Appl.*, **9**, 100334.

Castro,E. *et al.* (2022) Transformer-based protein generation with regularized latent space optimization. *Nat. Mach. Intell.*, 1–12.

Castro-Vega,L.J. *et al.* (2015) Multi-omics analysis defines core genomic alterations in pheochromocytomas and paragangliomas. *Nat. Commun.*, **6**, 6044.

Charoenkwan,P. *et al.* (2021) BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. *Bioinformatics*, **37**, 2556–2562.

Chaudhari,S. *et al.* (2021) An attentive survey of attention models. *ACM Trans. Intell. Syst. Technol*, **12**, 53:1–53:32.

Chen,J. *et al.* (2022a) A comprehensive comparison on cell-type composition inference for spatial transcriptomics data. *Brief. Bioinform.*, **23**, bbac245.

Chen,J. *et al.* (2021a) TransUNet: transformers make strong encoders for medical image segmentation. arXiv, arXiv:2102.04306v1, https://arxiv.org/abs/2102.04306v1, preprint: not peer reviewed.

Chen,J. *et al.* (2021b) ViT-V-Net: vision transformer for unsupervised volumetric medical image registration. arXiv, arXiv:2104.06468v1, https://arxiv.org/abs/2104.06468v1, preprint: not peer reviewed.

Chen,K. *et al.* (2022b) Capturing large genomic contexts for accurately predicting enhancer-promoter interactions. *Brief. Bioinform.*, **23**, bbab577.

Chen,Y.-C. *et al.* (2020) UNITER: UNiversal Image-TExt Representation learning. In: Vedaldi,A. *et al.* (ed.) *Computer Vision – ECCV 2020, Lecture Notes in Computer Science*. Springer International Publishing, Cham, pp. 104–120.

Chithrananda,S. *et al.* (2020) ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv, arXiv:2010.09885v2, https://arxiv.org/abs/2010.09885v2, preprint: not peer reviewed.

Cho,K. *et al.* (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pp. 1724–1734.

Choromanski,K. *et al.* (2021) Rethinking attention with performers. arXiv, arXiv:2009.14794v1, https://arxiv.org/abs/2009.14794v1, preprint: not peer reviewed.

Clauwaert,J. *et al.* (2021) Explainability in transformer models for functional genomics. *Brief. Bioinform.*, **22**, bbab060.

Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.

Dai,Z. *et al.* (2019) Transformer-XL: attentive language models beyond a fixed-length context. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pp. 2978–2988.

Devlin,J. *et al.* (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.

Dosovitskiy,A. *et al.* (2021) An image is worth 16x16 words: transformers for image recognition at scale. arXiv, arXiv:2010.11929v2, https://arxiv.org/abs/2010.11929v2, preprint: not peer reviewed.

Elnaggar,A. *et al.* (2022) ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, **44**, 7112–7127.

Faridani,O.R. *et al.* (2016) Single-cell sequencing of the small-RNA transcriptome. *Nat. Biotechnol.*, **34**, 1264–1266.

Feinberg,E.N. *et al.* (2018) PotentialNet for molecular property prediction. *ACS Cent. Sci.*, **4**, 1520–1530.

Ferruz,N. *et al.* (2022) ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.*, **13**, 4348.

Feuk,L. *et al.* (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 85–97.

Findlay,J.B.C. *et al.* (1995) Protein sequence analysis, storage and retrieval. In: Atassi,M.Z. and Appella,E. (ed.) *Methods in Protein Structure Analysis*. Springer US, Boston, MA, pp. 465–472.

Gilmer,J. *et al.* (2020) Message passing neural networks. In: Schütt,K.T. *et al.* (ed.) *Machine Learning Meets Quantum Physics, Lecture Notes in Physics*. Springer International Publishing, Cham, pp. 199–214.

Goeman,J.J. and Bühlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Grechishnikova,D. (2021) Transformer neural network for protein-specific de novo drug generation as a machine translation problem. *Sci. Rep.*, **11**, 321.

Han,K. *et al.* (2023) A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.*, **45**, 87–110.

Han,X. *et al.* (2021) Pre-trained models: past, present and future. *AI Open*, **2**, 225–250.

Han,X. and Kwoh,C.K. (2019) Natural language processing approaches in bioinformatics. In: Ranganathan,S. (ed.) *Encyclopedia of Bioinformatics and Computational Biology*. Academic Press, Oxford, pp. 561–574.

Hastie,T. *et al.* (2009) Overview of supervised learning. In: Hastie,T. *et al.* (ed.) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, NY, pp. 9–41.

Heinzinger,M. *et al.* (2019) Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, **20**, 723.

Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural Comput.*, **9**, 1735–1780.

Howard,J. and Ruder,S. (2018) Universal language model fine-tuning for text classification. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, pp. 328–339.

Huang,K. *et al.* (2021) MolTrans: molecular interaction transformer for drug-target interaction prediction. *Bioinformatics*, **37**, 830–836.

Iuchi,H. *et al.* (2021) Representation learning applications in biological sequence analysis. *Comput. Struct. Biotechnol. J.*, **19**, 3198–3208.

Ji,Y. *et al.* (2021) DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, **37**, 2112–2120.

Jiang,L. *et al.* (2022) DeepTTA: a transformer-based model for predicting cancer drug response. *Brief. Bioinform.*, **23**, bbac100.

Jurenaite,N. *et al.* (2022) SetQuence & SetOmic: deep set transformer-based representations of cancer multi-omics. In: *2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Ottawa, ON, Canada, pp. 1–9.

Kaczmarek,E. *et al.* (2021) Multi-omic graph transformers for cancer classification and interpretation. In: *Biocomputing 2022*. World Scientific, Kohala Coast, Hawaii, USA, pp. 373–384.

Kalakoti,Y. *et al.* (2022) TransDTI: transformer-based language models for estimating DTIs and building a drug recommendation workflow. *ACS Omega*, **7**, 2706–2717.

Kalyan,K.S. *et al.* (2021a) AMMU: a survey of transformer-based biomedical pretrained language models. *J. Biomed. Inform.*, **126**, 103982.

Kalyan,K.S. *et al.* (2021b) AMMUS: a survey of transformer-based pretrained models in natural language processing. arXiv, arXiv:2108.05542v2, https://arxiv.org/abs/2108.05542v2, preprint: not peer reviewed.

Kang,M. *et al.* (2022) A roadmap for multi-omics data integration using deep learning. *Brief. Bioinform.*, **23**, bbab454.

Karametsi,K. *et al.* (2014) Identification of bitter peptides in aged cheddar cheese. *J. Agric. Food Chem.*, **62**, 8034–8041.

Karlin,S. and Ghandour,G. (1985) Comparative statistics for DNA and protein sequences: single sequence analysis. *Proc. Natl. Acad. Sci. USA*, **82**, 5800–5804.

Katharopoulos,A. *et al.* (2020) Transformers are RNNs: fast autoregressive transformers with linear attention. In: *Proceedings of the 37th International Conference on Machine Learning*, PMLR, Online, pp. 5156–5165.

Ke,G. *et al.* (2021) Rethinking positional encoding in language pre-training. arXiv, arXiv:2006.15595v4, https://arxiv.org/abs/2006.15595v4, preprint: not peer reviewed.

Khan,A. and Lee,B. (2021) Gene transformer: transformers for the gene expression-based classification of lung cancer subtypes. arXiv. arXiv: 2108.11833v3, https://arxiv.org/abs/2108.11833v3, preprint: not peer reviewed.

Khan,S. *et al.* (2022) Transformers in vision: a survey. *ACM Comput. Surv.*, **54**, 200:1-200:41.

Kitaev,N. *et al.* (2020) Reformer: the efficient transformer. arXiv, arXiv: 2001.04451v2, https://arxiv.org/abs/2001.04451v2, preprint: not peer reviewed.

Kolodziejczyk,A.A. *et al.* (2015) The technology and biology of single-cell RNA sequencing. *Mol. Cell.*, **58**, 610–620.

Koonin,E.V. and Galperin,M.Y. (2003a) Genome annotation and analysis. In: Koonin,E.V. and Galperin,M.Y. (ed.) *Sequence — Evolution — Function: Computational Approaches in Comparative Genomics*. Springer US, Boston, MA, pp. 193–226.

Koonin,E.V. and Galperin,M.Y. (2003b) Principles and methods of sequence analysis. In: Koonin,E.V. and Galperin,M.Y. (ed.) *Sequence — Evolution — Function: Computational Approaches in Comparative Genomics*. Springer US, Boston, MA, pp. 111–192.

Kothari,S.C. and Oh,H. (1993) Neural networks for pattern recognition. In: Yovits,M.C. (ed.) *Advances in Computers*. Academic Press, San Diego, CA, USA, pp. 119–166.

Lan,Z. *et al.* (2020) ALBERT: a lite BERT for self-supervised learning of language representations. arXiv, arXiv:1909.11942v6, https://arxiv.org/abs/1909.11942v6, preprint: not peer reviewed.

Larance,M. and Lamond,A.I. (2015) Multidimensional proteomics for cell biology. *Nat. Rev. Mol. Cell Biol.*, **16**, 269–280.

Le,Q. and Mikolov,T. (2014) Distributed representations of sentences and documents. In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14. JMLR.org, Beijing, China, p. II-1188-II–1196.

LeCun,Y. *et al.* (2015) Deep learning. *Nature*, **521**, 436–444.

Lee,J. *et al.* (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**, 1234–1240.

Lee,K. *et al.* (2021) A transformer architecture based on BERT and 2D convolutional neural network to identify DNA enhancers from sequence information. *Brief. Bioinform.*, **22**, bbab005.

Li,H. *et al.* (2022) KPGT: knowledge-guided pre-training of graph transformer for molecular property prediction. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22. Association for Computing Machinery, New York, NY, USA, pp. 857–867.

Li,Y. *et al.* (2020) BEHRT: transformer for electronic health records. *Sci. Rep.*, **10**, 7155.

Lin,T. *et al.* (2022) A survey of transformers. *AI Open*, **3**, 111–132.

Lin,Z. *et al.* (2017) A structured self-attentive sentence embedding. arXiv, arXiv:1703.03130v1, https://arxiv.org/abs/1703.03130v1, preprint: not peer reviewed.

Liu,C. *et al.* (2022) eQTLs play critical roles in regulating gene expression and identifying key regulators in rice. *Plant Biotechnol. J.*, **20**, 2357–2371.

Liu,K. *et al.* (2019a) Chemi-Net: a molecular graph convolutional network for accurate drug property prediction. *Int. J. Mol. Sci.*, **20**, E3389.

Liu,Q. *et al.* (2020) A survey on contextual embeddings. arXiv, arXiv: 2003.07278v2, https://arxiv.org/abs/2003.07278v2, preprint: not peer reviewed.

Liu,X. *et al.* (2023) Self-Supervised Learning: Generative or Contrastive. *IEEE Trans. Knowl. Data Eng.*, **35**, 857–876.

Liu,Y. *et al.* (2019b) RoBERTa: a robustly optimized BERT pretraining approach. arXiv, arXiv:1907.11692v1, https://arxiv.org/abs/1907.11692v1, preprint: not peer reviewed.

Lo,Y.-C. *et al.* (2018) Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today.*, **23**, 1538–1546.

Lu,J. *et al.* (2019) ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., Vancouver, Canada.

Ma,A. *et al.* (2021) Biological network inference from single-cell multi-omics data using heterogeneous graph transformer. bioRxiv, doi: 10.1101/2021.10.31.466658, preprint: not peer reviewed.

Marrero,M. *et al.* (2013) Named entity recognition: fallacies, challenges and opportunities. *Comput. Stand. Interfaces*, **35**, 482–489.

Meng,Y. *et al.* (2021) Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE J. Biomed. Health Inform.*, **25**, 3121–3129.

Mikolov,T. *et al.* (2013a) Distributed representations of words and phrases and their compositionality. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13. Curran Associates Inc., Red Hook, NY, USA, pp. 3111–3119.

Mikolov,T. *et al.* (2013b) Efficient estimation of word representations in vector space. In: *Proceedings of Workshop at ICLR*, 2013, Scottsdale, Arizona.

Milletari,F. *et al.* (2016) V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: *2016 Fourth International Conference on 3D Vision (3DV)*, Stanford, California, USA, pp. 565–571.

Mo,S. *et al.* (2021) Multi-modal self-supervised pre-training for regulatory genome across cell types. arXiv, arXiv:2110.05231v2, https://arxiv.org/abs/2110.05231v2, preprint: not peer reviewed.

Nadkarni,P.M. *et al.* (2011) Natural language processing: an introduction. *J. Am. Med. Inform. Assoc.*, **18**, 544–551.

Nenkova,A. and McKeown,K. (2012) A survey of text summarization techniques. In: Aggarwal,C.C. and Zhai,C. (ed.) *Mining Text Data*. Springer US, Boston, MA, pp. 43–76.

Nikpay,M. *et al.* (2015) A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.*, **47**, 1121–1130.

Noor,Z. *et al.* (2021) Mass spectrometry–based protein identification in proteomics—a review. *Brief. Bioinform.*, **22**, 1620–1638.

Ofer,D. *et al.* (2021) The language of proteins: NLP, machine learning & protein sequences. *Comput. Struct. Biotechnol. J.*, **19**, 1750–1758.

Ozsolak,F. and Milos,P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, **12**, 87–98.

Pang,M. *et al.* (2021) Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. bioRxiv, doi: 10.1101/2021.11.28.470212v1, preprint: not peer reviewed.

Park,M. *et al.* (2022) EpiBERTope: a sequence-based pre-trained BERT model improves linear and structural epitope prediction by learning long-distance protein interactions effectively. bioRxiv, doi: 10.1101/2022.02.27.481241, preprint: not peer reviewed.

Pascanu,R. *et al.* (2013) On the difficulty of training recurrent neural networks. In: *Proceedings of the 30th International Conference on*

International Conference on Machine Learning - Volume 28, ICML'13. JMLR.org, Atlanta, USA, pp. III-1310-III–1318.
Petroni,F. *et al.* (2019) Language models as knowledge bases? In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, pp. 2463–2473.
Ponting,C.P. and Birney,E. (2005) Protein sequence analysis and domain identification. In: Walker,J.M. (ed.) *The Proteomics Protocols Handbook*. Springer Protocols Handbooks, Humana Press, Totowa, NJ, pp. 527–541.
Portela,A. and Esteller,M. (2010) Epigenetic modifications and human disease. *Nat. Biotechnol.*, **28**, 1057–1068.
Pundge,A.M. *et al.* (2016) Question answering system, approaches and techniques: a review. *Int. J. Comput. Appl. A*, **141**, 34–39.
Qiao,Y. *et al.* (2022) BERT-Kcr: prediction of lysine crotonylation sites by a transfer learning method with pre-trained BERT models. *Bioinformatics*, **38**, 648–654.
Qiu,X. *et al.* (2020) Pre-trained models for natural language processing: a survey. *Sci. China Technol. Sci.*, **63**, 1872–1897.
Raad,J. *et al.* (2022) miRe2e: a full end-to-end deep model based on transformers for prediction of pre-miRNAs. *Bioinformatics*, **38**, 1191–1197.
Radford,A. *et al.* (2019) Language models are unsupervised multitask learners. *Technical report*.
Radford,A. *et al.* (2021) Learning transferable visual models from natural language supervision. In: *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Virtual, pp. 8748–8763.
Radford,A. and Narasimhan,K. (2018) Improving language understanding by generative pre-training.
Raffel,C. *et al.* (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, **21**, 1−67.
Rao,R. *et al.* (2020) Transformer protein language models are unsupervised structure learners. bioRxiv, doi: 10.1101/2020.12.15.422761, preprint: not peer reviewed.
Rasmy,L. *et al.* (2021) Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.*, **4**, 86.
Reis-Filho,J.S. (2009) Next-generation sequencing. *Breast Cancer Res.*, **11**, S12.
Rives,A. *et al.* (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. USA*, **118**, e2016239118.
Rix,U. and Superti-Furga,G. (2009) Target profiling of small molecules by chemical proteomics. *Nat. Chem. Biol.*, **5**, 616–624.
Romero,P.A. and Arnold,F.H. (2009) Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.*, **10**, 866–876.
Rong,Y. *et al.* (2020) Self-supervised graph transformer on large-scale molecular data. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems - Volume 33*, NIPS'20. Curran Associates Inc., Vancouver, Canada, pp. 12559–12571.
Ronneberger,O. *et al.* (2015) U-Net: convolutional networks for biomedical image segmentation. In: Navab,N. *et al.* (ed.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Lecture Notes in Computer Science*. Springer International Publishing, Cham, pp. 234–241.
Rotter,A. *et al.* (2010) Gene expression data analysis using closed itemset mining for labeled data. *OMICS*, **14**, 177–186.
Rung,J. and Brazma,A. (2013) Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.*, **14**, 89–99.
Sarkar,I.N. (2010) Biomedical informatics and translational medicine. *J. Transl. Med.*, **8**, 22.
Scarselli,F. *et al.* (2009) The graph neural network model. *IEEE Trans. Neural Netw.*, **20**, 61–80.
Schouten,K. and Frasincar,F. (2016) Survey on aspect-level sentiment analysis. *IEEE Trans. Knowl. Data Eng.*, **28**, 813–830.
Schuster,M. and Paliwal,K.K. (1997) Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, **45**, 2673–2681.
Skansi,S. (2018) Feedforward neural networks. In: Skansi,S. (ed.) *Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence, Undergraduate Topics in Computer Science*. Springer International Publishing, Cham, pp. 79–105.
Stevens,R. *et al.* (2000) Ontology-based knowledge representation for bioinformatics. *Brief. Bioinform.*, **1**, 398–414.
Sutskever,I. *et al.* (2014) Sequence to sequence learning with neural networks. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14. MIT Press, Cambridge, MA, USA, pp. 3104–3112.
Tang,G. *et al.* (2018) Why self-attention? A targeted evaluation of neural machine translation architectures. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, pp. 4263–4272.
Tao,Y. *et al.* (2020) From genome to phenome: predicting multiple cancer phenotypes based on somatic genomic alterations via the genomic impact transformer. In: *Pacific Symposium on Biocomputing*. Vol. **25**, Big Island of Hawaii, USA, pp. 79–90.
Tsujii,J. (2021) Natural language processing and computational linguistics. *Comput. Linguist.*, **47**, 707–727.
Turian,J. *et al.* (2010) Word representations: a simple and general method for semi-supervised learning. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10. Association for Computational Linguistics, Uppsala, Sweden, pp. 384–394.
Unsal,S. *et al.* (2022) Learning functional properties of proteins with language models. *Nat. Mach. Intell.*, **4**, 227–245.
Urda,D. *et al.* (2017) Deep learning to analyze RNA-seq gene expression data. In: Rojas,I. *et al.* (ed.) *Advances in Computational Intelligence, Lecture Notes in Computer Science*. Springer International Publishing, Cham, pp. 50–59.
Vaswani,A. *et al.* (2017) Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17. Curran Associates Inc., Red Hook, NY, USA, pp. 6000–6010.
von Heijne,G. (1992) Computer analysis of DNA and protein sequences. In: Christen,P. and Hofmann,E. (ed.) *EJB Reviews 1991*. Springer, Berlin, Heidelberg, pp. 85–88.
Walczak,S. and Cerpa,N. (2003) Artificial neural networks. In: Meyers,R.A. (ed.) *Encyclopedia of Physical Science and Technology*. 3rd edn. Academic Press, New York, pp. 631–645.
Wang,S. *et al.* (2020) Linformer: self-attention with linear complexity. arXiv, arXiv:2006.04768v3, https://arxiv.org/abs/2006.04768v3, preprint: not peer reviewed.
Wang,S. *et al.* (2019) SMILES-BERT: large scale unsupervised pre-training for molecular property prediction. In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. ACM, Niagara Falls, NY, USA, pp. 429–436.
Wang,X. *et al.* (2021) InferBERT: a transformer-based causal inference framework for enhancing pharmacovigilance. *Front. Artif. Intell.*, **4**.
Wu,Z. *et al.* (2022) Knowledge-based BERT: a method to extract molecular features like computational chemists. *Brief. Bioinform.*, **23**, bbac131.
Wu,Z. *et al.* (2018) MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.*, **9**, 513–530.
Xu,P. *et al.* (2022) Multimodal learning with transformers: a survey. arXiv, arXiv:2206.06488v1, https://arxiv.org/abs/2206.06488v1, preprint: not peer reviewed.
Yang,F. *et al.* (2022) scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat. Mach. Intell.*, **4**, 852–866.
Yang,P. *et al.* (2019a) Multi-omic profiling reveals dynamics of the phased progression of pluripotency. *Cell Syst.*, **8**, 427–445.e10.
Yang,Z. *et al.* (2019b) XLNet: generalized autoregressive pretraining for language understanding. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA.
Yun,S. *et al.* (2019) Graph transformer networks. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems - Volume 32*. Curran Associates Inc., Vancouver, Canada.
Zaheer,M. *et al.* (2020) Big bird: transformers for longer sequences. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems - Volume 33*. Curran Associates Inc., Vancouver, Canada, pp. 17283–17297.
Zeng,W. *et al.* (2018) Prediction of enhancer-promoter interactions via natural language processing. *BMC Genomics*, **19**, 84.
Zeng,Y. *et al.* (2016) A convolution BiLSTM neural network model for Chinese event extraction. In: Lin,C.-Y. *et al.* (ed.) *Natural Language Understanding and Intelligent Applications, Lecture Notes in Computer Science*. Springer International Publishing, Cham, pp. 275–287.
Zeng,Y. *et al.* (2022) Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Brief. Bioinform.*, **23**, bbac297.
Zhang,H. *et al.* (2021a) Co-evolution transformer for protein contact prediction. In: *Proceedings of the 35th International Conference on Neural Information Processing Systems - Volume 34*. Curran Associates Inc., Virtual, pp. 14252–14263.
Zhang,K. *et al.* (2018) Residual networks of residual networks: multilevel residual networks. *IEEE Trans. Circuits Syst. Video Technol.*, **28**, 1303–1314.

Zhang,L. *et al.* (2021b) BERT-m7G: a transformer architecture based on BERT and stacking ensemble to identify RNA N7-Methylguanosine sites from sequence information. *Comput. Math. Methods Med.*, **2021**, 7764764.

Zhang,Q. *et al.* (2017) A review on entity relation extraction. In: *2017 Second International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, Harbin, China, pp. 178–183.

Zhang,S. *et al.* (2020) TensorCoder: dimension-wise attention via tensor representation for natural language modeling. arXiv, arXiv:2008.01547v2, https://arxiv.org/abs/2008.01547v2, preprint: not peer reviewed.

Zhang,Z. *et al.* (2019) ERNIE: enhanced language representation with informative entities. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pp. 1441–1451.

Zhao,G. *et al.* (2019) Explicit sparse transformer: concentrated attention through explicit selection. arXiv, arXiv:1912.11637v1, https://arxiv.org/abs/1912.11637v1, preprint: not peer reviewed.

Zheng,R. *et al.* (2021) Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In: *Proceedings of the 38th International Conference on Machine Learning*, *PMLR*, Virtual, pp. 12736–12746.

Zitnik,M. *et al.* (2018) Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, **34**, i457–i466.