Received: 12 April 2024 | Revised: 23 June 2024 | Accepted: 2 July 2024

DOI: 10.1002/wcms.1725

OVERVIEW

WIREs COMPUTATIONAL MOLECULAR SCIENCE WILEY

Transformer technology in molecular science

Jian Jiang[1,2] | Lu Ke[1] | Long Chen[1] | Bozheng Dou[1] | Yueying Zhu[1] | Jie Liu[1] | Bengong Zhang[1] | Tianshou Zhou[3] | Guo-Wei Wei[2,4,5]

[1]Research Center of Nonlinear Science, School of Mathematical and Physical Sciences, Wuhan Textile University, Wuhan, China

[2]Department of Mathematics, Michigan State University, East Lansing, Michigan, USA

[3]Key Laboratory of Computational Mathematics, Guangdong Province, and School of Mathematics, Sun Yat-sen University, Guangzhou, China

[4]Department of Electrical and Computer Engineering, Michigan State University, East Lansing, Michigan, USA

[5]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan, USA

Correspondence
Jian Jiang, Research Center of Nonlinear Science, School of Mathematical and Physical Sciences, Wuhan Textile University, Wuhan, 430200, China.
Email: jjiang@wtu.edu.cn

Guo-Wei Wei, Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA.
Email: weig@msu.edu

Funding information
NASA, Grant/Award Number: 80NSSC21M0023; National Science Foundation, Grant/Award Numbers: DMS2052983, DMS-1761320, IIS-1900473; National Natural Science Foundation of China, Grant/Award Numbers: 11971367, 12271416, 11972266; NIH, Grant/Award Numbers: R01GM126189, R01AI164266, R35GM148196; MSU Foundation; BristolMyers Squibb, Grant/Award Number: 65109; Pfizer

Edited by: Peter R. Schreiner, Editor-in-Chief

Abstract

A transformer is the foundational architecture behind large language models designed to handle sequential data by using mechanisms of self-attention to weigh the importance of different elements, enabling efficient processing and understanding of complex patterns. Recently, transformer-based models have become some of the most popular and powerful deep learning (DL) algorithms in molecular science, owing to their distinctive architectural characteristics and proficiency in handling intricate data. These models leverage the capacity of transformer architectures to capture complex hierarchical dependencies within sequential data. As the applications of transformers in molecular science are very widespread, in this review, we only focus on the technical aspects of transformer technology in molecule domain. Specifically, we will provide an in-depth investigation into the algorithms of transformer-based machine learning techniques in molecular science. The models under consideration include generative pre-trained transformer (GPT), bidirectional and auto-regressive transformers (BART), bidirectional encoder representations from transformers (BERT), graph transformer, transformer-XL, text-to-text transfer transformer, vision transformers (ViT), detection transformer (DETR), conformer, contrastive language-image pre-training (CLIP), sparse transformers, and mobile and efficient transformers. By examining the inner workings of these models, we aim to elucidate how their architectural innovations contribute to their effectiveness in processing complex molecular data. We will also discuss promising

Abbreviations: AI, artificial intelligence; BART, bidirectional and auto-regressive transformers; BERT, bidirectional encoder representations from transformers; CLIP, contrastive language-image pre-training; CNN, convolutional neural network; DETR, detection transformer; DL, deep learning; GRU, gated recurrent unit; GNN, graph neural network; GPT, generative pre-trained transformer; SMILES, simplified molecular input line entry system; T5, text-to-text transfer transformer; LSTM, long short-term memory; ML, machine learning; MPP, molecular property prediction; NLP, natural language processing; PPIs, protein–protein interactions; ViT, vision transformers.

trends in transformer models within the context of molecular science, emphasizing their technical capabilities and potential for interdisciplinary research. This review seeks to provide a comprehensive understanding of the transformer-based machine learning techniques that are driving advancements in molecular science.

This article is categorized under:
Data Science > Chemoinformatics
Data Science > Artificial Intelligence/Machine Learning

KEYWORDS
biology, chemistry, machine learning, molecular science, transformer technology

1 | INTRODUCTION

In recent years, machine learning (ML), particularly deep learning (DL), has made significant strides across diverse research domains, encompassing science, engineering, technology, medicine, and industry.[1–4] This progress marks a crucial milestone in data-driven discovery. Transformer-based models, the foundational architecture behind large language models (LLMs) such as generative pre-trained transformer (GPT), have emerged as powerful tools in molecular science due to their distinctive architectural characteristics and their adeptness at handling complex data.[5,6] The field of biological molecular and chemical sciences, dedicated to exploring the essence of matter and revealing its composition, properties, and interactions, faces escalating challenges as the volume and complexity of data continue to grow.[7,8] Traditional research methods are encountering difficulties in coping with these challenges, and the advent of ML techniques, especially transformer-based technologies, has opened new avenues for advancement in molecular science.[9,10]

The success of transformers in molecular science can be attributed to several key factors. Firstly, their inherent strength in processing sequential data aligns seamlessly with the sequential nature of chemical and biological molecules, including DNA, RNA, and proteins.[11,12] Additionally, the attention mechanism within transformers facilitates the consideration of global context, empowering the model to capture long-range dependencies within intricate molecular structures.[13,14] Furthermore, the adaptability of transformers to diverse data representations, such as Simplified Molecular Input Line Entry System (SMILES) strings or molecular graphs, enhances their versatility in handling varied molecular information.[15,16] Chen et al. has developed a multiscale approach to convert three-dimensional (3D) molecular structural information into topological sequences suitable for transformers.[17]

Another critical factor contributing to the achievements of transformers in molecular science is their capacity for transfer learning and pre-training. Pre-training on extensive chemical and biological datasets allows the model to grasp general features,[18] and subsequent fine-tuning on specific tasks enhances its performance for specialized applications.[19] Transformers also exhibit robustness in handling variable sequence lengths,[20,21] a crucial capability in chemistry and biology where biological and chemical molecules can significantly vary in size and complexity.

The success of transformer-based models in molecular science extends beyond their technical capabilities to their potential for interdisciplinary collaboration. These models seamlessly integrate biological, physical, and chemical information, bridging gaps among different disciplines. The demonstrated state-of-the-art performance of transformers in various sequence-related tasks, combined with their ability to handle multimodal data, positions them as powerful tools for advancing molecular research and contributing to the interdisciplinary nature of chemical and biological domains.[22,23]

Overall, the transformer architecture is dynamic and adept at adapting to diverse domains and challenges, underscoring the versatility and potency of transformer-based algorithms across a broad spectrum of applications. In this review, we pay close attention on the technical aspects of transformer technology, and there is no space to discuss the widespread applications of transformer-based models. In particular, noteworthy algorithms within this framework include generative pre-trained transformer (GPT),[24,25] bidirectional encoder representations from transformers (BERT),[26] bidirectional and auto-regressive transformers (BART),[27,28] graph transformer,[29,30] transformer-XL,[31] text-to-text transfer transformer,[32,33] vision transformers (ViT),[34,35] detection transformer (DETR),[36] conformer,[37,38] contrastive language-image pre-training (CLIP),[39,40] sparse transformers,[41] and mobile and efficient transformers.[42] Despite

their prevalence, there is a lack of an organized taxonomy linking these techniques in existing literature. This review aims to fill this gap by conducting a survey that outlines the algorithms and methodologies of transformers in molecular science.

The remaining sections of this paper are organized as follows. Transformers preliminaries are given in Section 2. Section 3 provides a brief overview of several main methods used for biological and chemical molecular research. Finally, Section 4 offers an outlook on future developments.

2 | TRANSFORMER PRELIMINARIES

Transformer is a revolutionary neural network architecture proposed by Vaswani et al. in 2017.[43] Specifically, it is divided into two parts, the encoder and the decoder, both consisting of multiple identical blocks, each of which contains a multi-head attention sublayer and a position feed-forward sublayer. These sublayers are all interconnected via residual connections and layer normalization. Additionally, position encoding is added to the input embedding. In order to gain insight into the construction of the transformer, we next describe each of these key blocks.

2.1 | Self-attention and multi-head attention

In general, people focus more on the crucial aspects of their objectives. Because of this uniqueness, the attention mechanism will give more weight to important aspects while giving less or no weight to less important features.[44] Initially, self-attention was proposed by Vaswani et al. as an improved form of the attention mechanism that focuses more on the internal connections of the input.[43] Figure 1a depicts the self-attention process. In detail, there are typically n samples, and by initializing weights, each sample has its own key (K), query (Q), and value (V). Then the dot product calculation involves the query and other input keys for one input, resulting in its attention. Similarly, this calculation can be used by all inputs to obtain the attention score, resulting in a matrix. Finally, to obtain weighted values, the

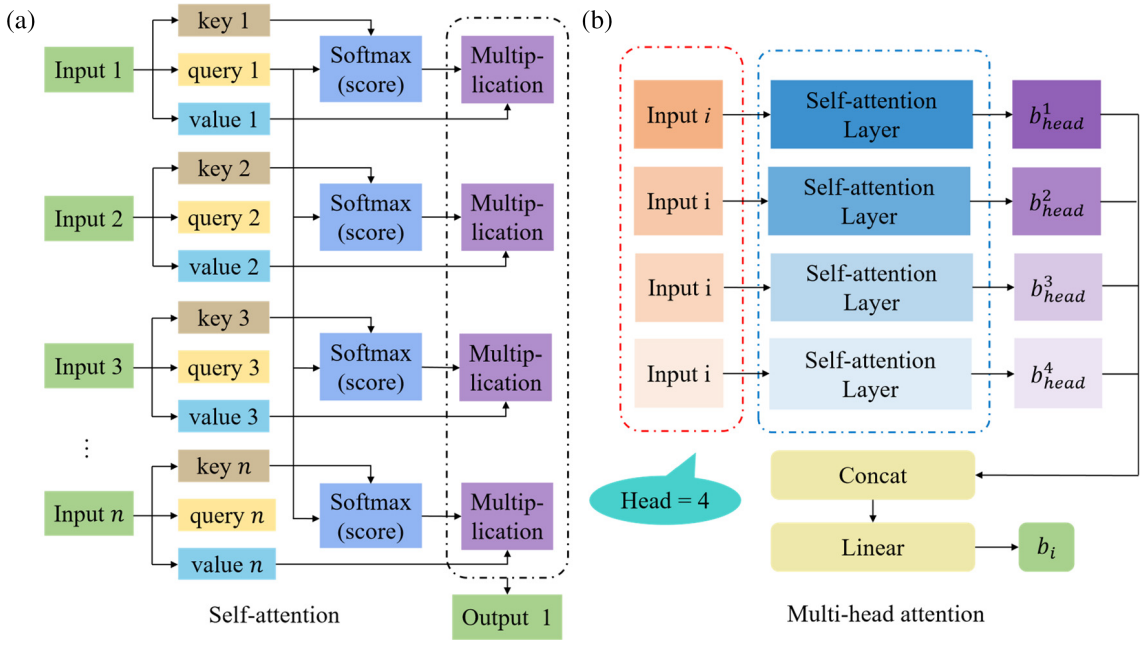

FIGURE 1 Self-attention and multi-head attention. (a) This subgraph represents the self-attention mechanism, and the first thing to determine is that each sample has K, Q, and V values. Taking sample 1 as an example, the output of sample 1 is obtained after a series of calculations with the values associated with other samples. Similar calculations are performed for the other samples. (b) The multi-head attention mechanism. Considering the computational process in subgraph a as a self-attention layer, there are four layers in this graph which means there are four heads. Each head performs the corresponding computation to get the final result.[43]

normalized attention score matrix by softmax is multiplied by value, and these weighted values are added to the final result. As a result, keep in mind that all variables are expressed in a matrix form as.

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

where d_k represents the dimension of the word vector. It can prevent softmax function calculation overflow.

Regrettably, the self-attention model concentrates too much on its own location when encoding information about its current location. Therefore, a multi-head attention mechanism emerged as the time required and here are some brief introductions.[45,46] From Figure 1b, we can easily summarize its mechanism based on self-attention. For a particular input, which is designated as input i, which can be viewed as a head. Consequently, As shown in Figure 1b, input i has multiple heads (here head = 4), and each head will receive the corresponding output called b_{head} through the self-attention mechanism. In the end, these b_{head} carry out vector end-to-end calculations and obtain b_i through linear transformation.

The process and parameters for the other inputs in the sequence are the same.

2.2 | Embedding and positional encoding

In order to facilitate computer recognition and localization of sequence information, embedding and positional encoding are both necessary preprocessing methods. In addition, embedding involves both input and output embedding. Their similarity lies in converting sequences into vectors, while the difference lies in the former being the embedding result of input data and the latter being the embedding result of ground truth. Directly utilizing the positional information of words is not possible in models like neural networks and transformers. Position encoding can incorporate position information into the vector representation of words, enabling the model to learn the order relationship of words in sentences.[47] In transformers, sine and cosine functions are employed to generate position encoding[43]:

$$\text{PE}_{(pos,2i)} = \sin\left(pos/10000_{2i/d\text{model}}\right) \tag{2}$$

and

$$\text{PE}_{(pos,2i+1)} = \cos\left(pos/10000_{2i/d\text{model}}\right). \tag{3}$$

In these two formulas for representing positional encoding, i is a parameter representing the dimension and d_{model} denotes the hidden layer dimension of the model. Meanwhile, pos is the position index of the element in the sequence, and PE stands for positional encoding.

2.3 | Position-wise feed-forward networks

Position-wise feed-forward networks are a layer of fully connected feed-forward neural networks in the transformer model.[48] They consist of two main operations: a fully connected hidden layer and an activation function. At each position, the input feature vectors pass through the fully connected layer and then undergo a nonlinear transformation via the activation function to generate a new feature representation.[49] In general, position feed-forward networks have the same structure at each location, but the weight parameters are independent.

Let the feature at each position of the input sequence be represented by $x = (x_1, x_2, ..., x_n)$, where n is the sequence length. The position feed-forward networks can be expressed as.[43]

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2, \tag{4}$$

where W_1, W_2, b_1 and b_2 are learnable parameters. Here, $\max(0,x)$ represents the activation function, usually the rectified linear unit (ReLU). Additionally, FFN stands for feed-forward network. It serves to transform the input vectors to improve the model representation.

2.4 | Residual connection and layer normalization

In the transformer architecture, residual connection and layer normalization are two important technologies that help improve model stability, efficiency, and performance. Each encoder consists of two residual connection and layer normalization layers, which are applied to both multi-head self-attention and feed-forward networks.[50] Given below is its calculation formula:

$$\mathrm{LayerNorm}(X + \mathrm{MutiHeadAttention}(X)), \tag{5}$$

$$\mathrm{LayerNorm}(X + \mathrm{FeedForward}(X)), \tag{6}$$

where X represents the input of the multi-head self-attention or feed-forward networks, which is added to the output to form the residual connection. The residual connection ensures that gradients pass smoothly through the network, while layer normalization helps to maintain a consistent activation distribution and avoid training instability. These techniques have been shown to be important for effective training of deep neural networks, and as such, they are key components of transformer models.

2.5 | Encoder and decoder

Like the earlier seq2seq model, the original transformer model uses an encoder/decoder architecture.[51] Each encoder consists of two main parts: the self-attention mechanism and the feed-forward neural networks. In contrast, each decoder consists of three main parts: a self-attention mechanism, a coding attention mechanism, and a feed-forward neural network. Specifically, the role of the encoder is to process the input sequence and generate a hidden representation summarizing the input information, while the decoder is to use this hidden representation to generate the desired output sequence. With end-to-end training of the encoder and decoder, this maximizes the likelihood of a correct output sequence for a given input sequence.[52]

3 | TRANSFORMER ALGORITHMS

3.1 | Generative pre-trained transformer (GPT)

GPT is a pre-trained language model proposed by the OpenAI team in 2018.[53] Thanks to its excellent natural language processing (NLP) capabilities, it has achieved remarkable success in the fields of text generation,[54,55] language translation,[56,57] and so on. Generally, GPT uses the decoder part of the transformer for text generation, focusing on predicting the next word in a sequence. Unlike models that use both the encoder and decoder, GPT does not utilize the encoder part, which is typically responsible for processing and understanding the input data in tasks like translation. Specifically, the core component of GPT is the decoder of the transformer model, stacked in multiple layers. GPT leverages pre-trained weights to generate human-like text, based on its ability to learn from vast text data. GPT2, for instance, stacks 12 layers of transformer decoders, modified with Layer Norm, and utilizes token embedding and positional encoding for input modeling. This structure enables GPT to generate coherent and contextually relevant responses. In addition to the field of NLP, it has many applications in image processing,[58,59] molecular generation,[60–64] drug design,[63,65–67] and protein engineering[68,69] due to its powerful generative and representation capabilities.

Nowadays, the study of protein–ligand binding is a popular area of research aimed at discovering molecules that bind to proteins. Hence, DrugGPT emerged and focused on chemical space exploration and the discovery of specific protein ligands, and it is a GPT-based ligand design strategy. Regarding the application of GPT on chemistry and biology, one can also refer to the following papers.[63,70–72]

The impact of GPT is rapidly growing in various fields, including mathematics.[73] GPT has a potential to revolutionize science and technology in the near future.

3.2 | Bidirectional encoder representations from transformers (BERT)

Introduced in 2018, BERT is a pre-trained model developed by Jacob Devlin and colleagues at Google AI Language.[74] It utilizes transformer encoders to understand language by considering both past and future contexts, surpassing earlier models like long short-term memory (LSTM) and gated recurrent unit (GRU).[75,76] BERT employs an attention mechanism to assign importance to words, undergoing pre-training for masked language modeling and next sentence prediction. It is further fine-tuned for specific NLP tasks, making it versatile for applications such as text prediction,[77] sentiment analysis,[78] question answering,[79] and text generation.[80] BERT's capabilities extend beyond NLP, finding utility in fields like molecular science, including applications in chemical and biological molecules as well as protein engineering.[81]

Molecular property prediction (MPP) stands as a crucial concern within the realms of drug design and substance discovery,[82,83] owing to its potential for enhancing chemical design, reducing research and development costs, and expediting the drug discovery process. In 2022, Wen et al. introduced a pioneering method for molecular sequence embedding and prediction, leveraging pre-training with a bidirectional transformer to generate a semantically enriched composite fingerprint representation, termed FP-BERT, followed by convolutional neural network (CNN) feature extraction and fully connected layers for classification or regression tasks, demonstrating exceptional performance across various MPP tasks.[84] The schematic representation of their algorithm's framework can be visualized in Figure 2.

Proteins are of paramount significance within the field of living organisms, as they play indispensable roles in virtually every facet of cellular processes and functions.[85] In the domain of bioinformatics, the identification of protein–protein interactions (PPIs) stands as a pivotal challenge. In 2023, Kanchan Jha and colleagues devised an AI model for PPI identification, utilizing PPI networks and protein sequences by graphically representing the PPI network, extracting features directly from protein sequences using a language model, encoding the network with Graph-BERT, and employing fully connected and softmax layers for classification, showing superior performance over existing models in

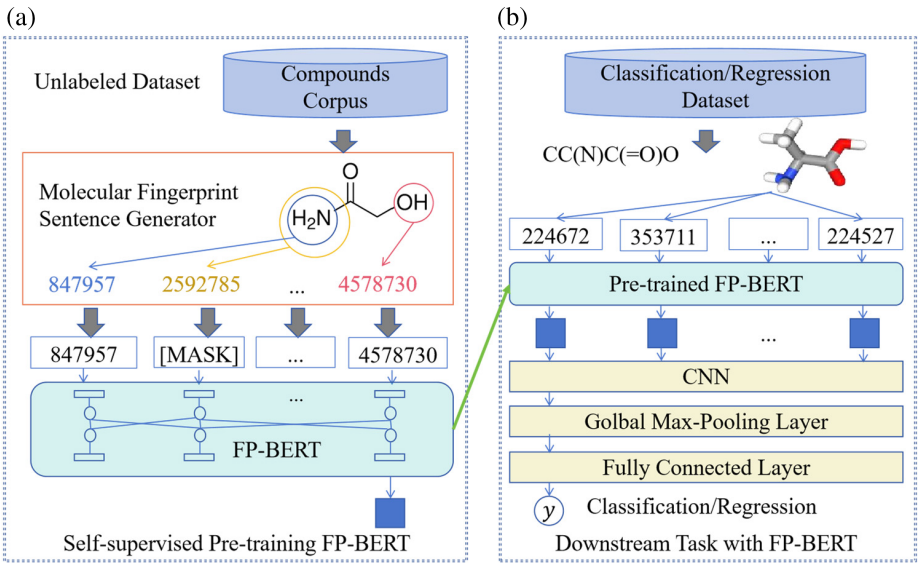

FIGURE 2 The architecture of FP-BERT based MPP model. The proposed MPP method consists of two parts. (a) The pre-trained FP-BERT model on the left. (b) The neural network for the downstream prediction task on the right.[84]

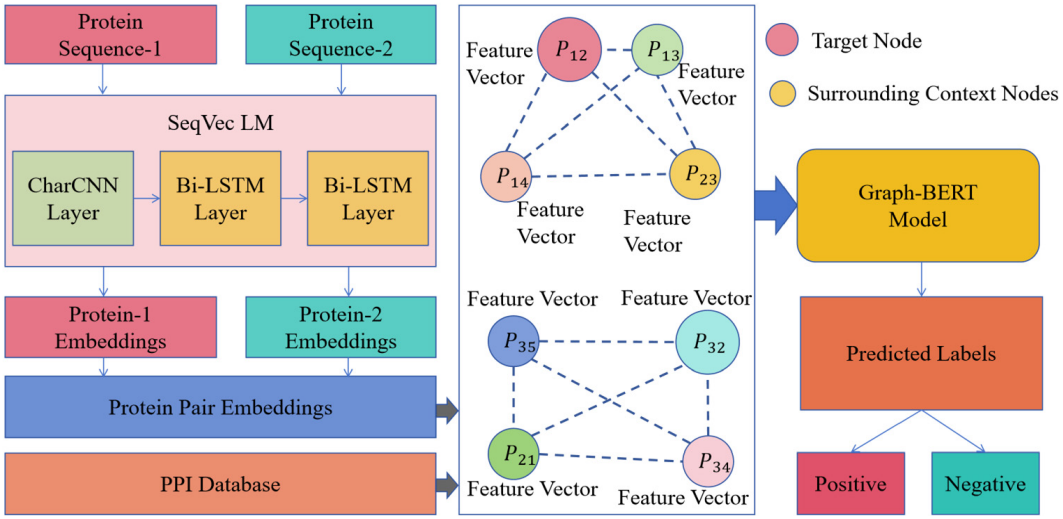

FIGURE 3 Protein–protein interaction recognition framework diagram. Predictions of protein–protein interactions were formulated using PPI networks and sequence-based features as a node classification problem.[86]

predicting protein interactions across multiple datasets.[86] The schematic representation of this methodology is depicted in Figure 3.

BERT finds extensive applications in molecular science, exemplified by SMILES-BERT[87] and Mol-BERT[88] for end-to-end molecular property prediction, sequence-based pre-trained BERT models for epitope prediction,[89] topological transformer for protein engineering,[81] enhancement of BERT for protein folding and design,[90] extracting predictive representations from hundreds of millions of molecules,[91] multi-transformers for virtual screening of biomolecular interaction,[92] and knowledge-based BERT for molecular feature extraction in drug discovery.[93] While BERT offers advantages such as reduced training time, low memory requirements, and high accuracy, it also presents limitations like limited contextual understanding, suboptimal text generation, and time-consuming fine-tuning. Nevertheless, its ongoing integration with other AI and ML technologies promises continued advancement in the realm of molecular science.

3.3 | Bidirectional and auto-regressive transformers (BART)

BART, proposed by Lewis et al. in 2019, is a pre-trained model integrating bidirectional and auto-regressive transformers, functioning as a denoising autoencoder for various tasks.[94] Its pre-training involves corrupting text data and reconstructing it via a sequence-to-sequence model. BART's architecture is based on a standard transformer neural network, making it versatile for natural language processing (NLP) tasks like language translation,[95] question answering,[96] summarization,[97] and paraphrasing.[98] Beyond NLP, BART proves useful in generating molecules,[99] drug discovery,[100] and protein sequence generation,[101] underscoring its versatility and significance in molecular science.

In recent years, machine learning has played a pivotal role in cheminformatics, notably accelerating drug discovery and aiding in the prediction of molecular properties and activities, crucial for prioritizing experimental work.[102] Dimitriadis et al. introduced a novel approach to multi-task regression in 2021, utilizing text-based transformer models trained on SMILES representations of molecules and emphasizing the significance of larger-scale pre-trained models across various chemical domains to enhance transformer model performance.[103] These pre-trained models were both based on the BART architecture, including the encoder and decoder, and the BART architecture is shown in Figure 4.

Recent research highlights the efficacy of integrating transformer models with SMILES for addressing cheminformatics challenges, although these models often require significant computational resources and are tailored to specific applications. To address this, Irwin et al. introduced the Chemformer model in 2022, based on the BART language model, which utilizes both encoder and decoder stacks to facilitate sequence-to-sequence tasks while minimizing

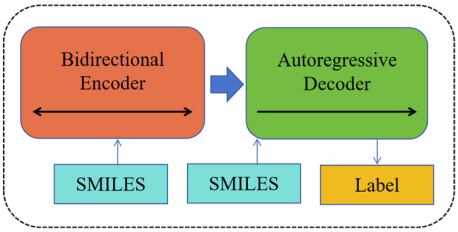

FIGURE 4 Encoder–decoder BART transformer model with masked inputs and auto-regressive generation of input sequences. During training, specific parts of the input sequence (molecules) are masked and the decoder tries to construct the entire original sequence in an auto-regressive manner.[103]

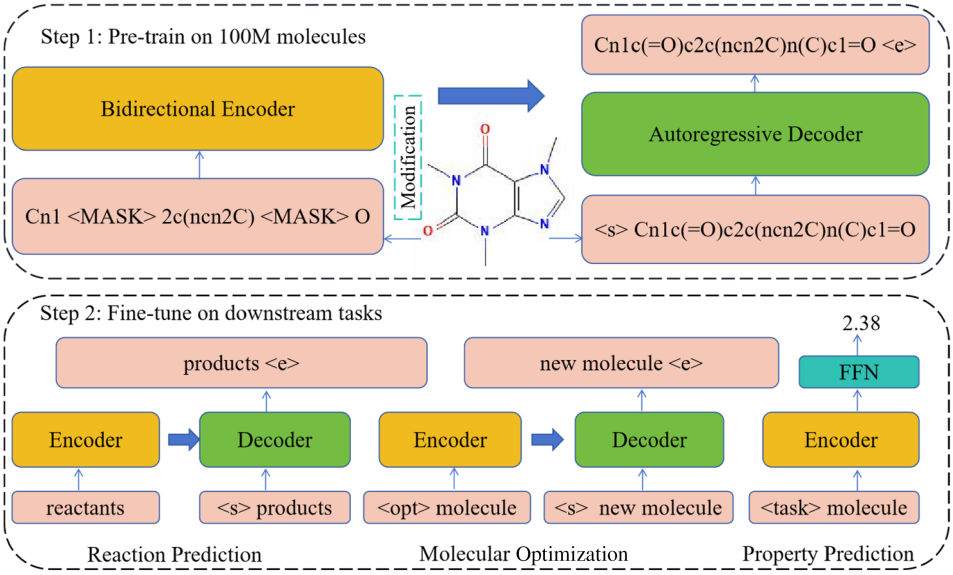

FIGURE 5 Schematic representation of the pre-training and fine-tuning procedure for downstream tasks. Here, we use the term "pre-training" to refer to training on a large dataset in an unsupervised manner before fine-tuning the model weights for a specific chemoinformatics task. The FFN in the figure refers to a feed-forward neural network—a sequence of fully connected artificial neuron layers with nonlinear activation.[104]

computational demands.[104] Figure 5 outlines how pre-training and downstream fine-tuning are applied to the Chemformer model.

BART demonstrates vast potential in molecular science, with applications like MS2Mol by Butler et al. for analyzing mass spectrometry data,[105] MolBART by Chilingaryan et al. for molecular feature learning,[106] and MegaMolBART by Hödl et al. for visualizing atom importance.[107] Despite its excellent fine-tuning capability, BART's resource-intensive training poses challenges, particularly for smaller organizations and researchers. Nonetheless, it signifies a significant advancement in AI and NLP, promising innovative applications in biomolecular and chemical realms as research and development efforts continue to evolve.

3.4 | Graph transformer

Graphs serve as fundamental data structures with applications spanning drug–protein interactions,[108] particle interactions,[109] protein flexibility,[110] and molecular graphs.[111] Graph neural networks (GNNs) have been popular for graph representation, particularly in biological and chemical domains, yet they struggle with capturing long-distance node interactions, affecting molecular property prediction (MPP) performance. Graph transformers, leveraging self-attentive mechanisms, excel in learning global context representations, thus gaining traction in molecular science for

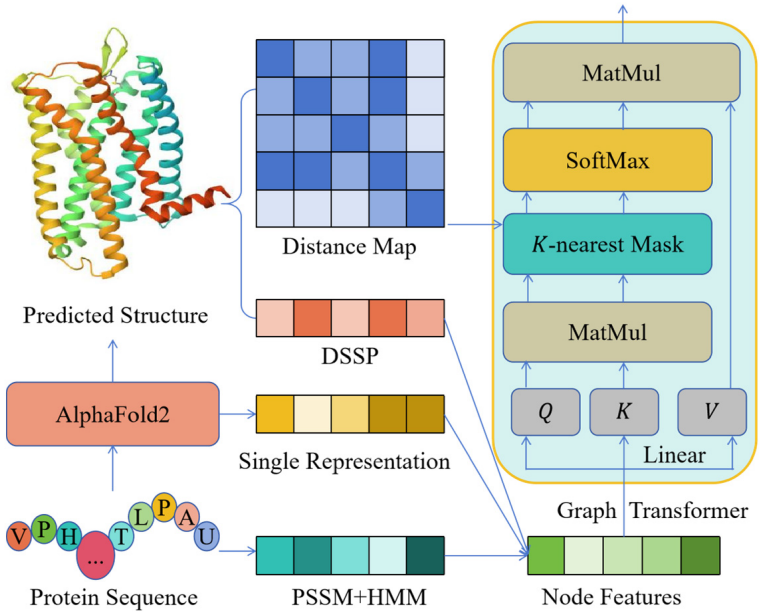

FIGURE 6 The overall architecture of GraphSite. First, protein sequences are fed into AlphaFold2 to generate a single protein representation and structure prediction. Next, distance maps and DSSP information are extracted from these predictions. Then, the single protein representation, DSSP information, and sequence-derived features such as position-specific scoring matrix (PSSM) and hidden Markov models (HMM) are integrated to form node feature vectors. Finally, using the distance graph, these node feature vectors are fed into a graph transformer model equipped with a *k*-nearest neighbor mask for learning patterns of DNA binding sites.[114]

their ability to model complex interactions among all nodes, demonstrating superior performance across various applications in molecular science.[112]

Protein–DNA interactions are crucial in biology, influencing vital processes like transcription and repair.[113] Accurate identification of binding sites is essential for understanding biological activities and advancing various fields such as drug discovery and gene regulation. In 2022, Yuan et al. introduced GraphSite, a method utilizing a graph transformation network incorporating protein structure information from AlphaFold2, which significantly improved DNA binding residue prediction from protein sequences.[114] The overall architecture of GraphSite is shown in Figure 6.

Predicting molecular properties is crucial in drug and material discovery, with deep learning models, particularly GNNs, widely used for learning molecular graph representations. However, supervised learning faces challenges due to limited labeled data and a vast chemical space. To overcome this, Li et al. introduced the knowledge-guided graph transformer pre-training (KPGT) method in 2022,[115] which outperformed current methods in various MPP tasks and offered a robust tool for molecular graph representation learning in drug design.

Graph transformers have advanced chemical and biological molecule studies, exemplified by path augmented graph transformer network (PAGTN) by Chen et al. for establishing long-range dependencies,[116] anchor-graph transformer by Jiang et al. for robust feature learning,[112] Hi-MGT by Tan et al. for toxicity identification,[117] edge-based graph transformer by Hu et al. for predicting drug combination effects,[118] and fingerprint-based multilevel graph transformer by Teng et al. for drug toxicity prediction.[119] Despite their advantages in overcoming input structure limitations and providing task-relevant information, traditional self-attentive modules in graph transformers result in high computational costs and sensitivity to graph noise, posing challenges for effective utilization in graph data representation and learning.

3.5 | Transformer-XL

Introduced by Google researchers in 2019, transformer-XL addresses long sequence data by leveraging the ability to capture distant dependencies in text,[120] exhibiting effectiveness in various domains such as document modeling,[121] long dialog systems,[122] sentiment analysis,[123] high-quality machine translation,[124] and even musical composition generation.[125] It has also proven effective in the field of molecular science.

For instance, Osipenko et al. utilized transformer-XL for small molecule retention time prediction, achieving 0.774 accuracy,[126] while Honda et al. applied it in drug design.[127] Additionally, Mandhana et al. employed transformer-XL with a self-attention mechanism for drug discovery, training on a dataset of 1.27 million molecules from the CHEMBL database.[128] Their model showed significant improvement in validity, uniqueness, novelty, KL divergence, and FCD score metrics compared to others.

Unlike traditional transformers, transformer-XL incorporates relative positional encoding, enhancing modeling of long sequences. Nayak et al. utilized transformer-XL for improved modeling of bond features between atom pairs, achieving superior performance in predicting molecular properties compared to other models (RMSE = 0.6) using 5328 molecular training examples from ESOL.[129]

Transformers can be applied to molecular generation and related studies, such as reaction generation tasks. Wang et al. utilized transformer-XL for a reaction generation task focused on Heck reactions, with a dataset of 8863 instances.[130] The model successfully generated over 2000 novel responses not present in the training set, with a feasibility rate of 47.76%. Verification of model feasibility and reactions is discussed in detail in the referenced article.

Relative positional encoding sets transformer-XL apart from other models, addressing limitations in modeling long sequences. Despite its effectiveness in capturing long-range dependencies, there are still challenges in handling long-term dependencies. Future enhancements of transformer-XL will focus on optimizing computational complexity and improving its ability to capture both short and long-term dependencies in text.

3.6 | Text-to-text transfer transformer

Introduced by the Google Research team in 2019, the text-to-text transfer transformer (T5) unifies NLP tasks into text-to-text mapping.[32] Notable for its permutation-based training, enhancing generalization, T5 excels in text categorization, generation, translation,[131–134] as well as information retrieval,[135] bioinformatics,[136] and finance applications.[137]

In the chemical domain, Lu et al. introduced T5Chem, a self-supervised pre-trained model trained on 97 million molecules from PubChem[138,139] and its flowchart is shown in Figure 7. Utilizing the USPTO 500 MT dataset for validation, T5Chem demonstrates superior performance across various chemical reaction prediction tasks, including reaction classification with 99.5% accuracy.

To design organic materials with desirable properties, Rothchild et al. introduced the C5T5 model, employing a self-supervised training approach and utilizing the IUPAC naming system for molecules.[140] They explored molecular optimization tasks related to drug discovery, aiming to modify molecular properties to meet various user requirements. Experimental results confirm the successful alteration of properties in the studied molecules.

In summary, as a text-to-text model, T5 exhibits remarkable versatility, given that both input and output consist of textual data. In the field of molecular science, one of its most distinctive attributes lies in its large-scale training methodology, rendering it highly proficient at capturing a wide range of language patterns.[141] However, the drawbacks of training models at scale also become apparent: high computational costs,[142] resource constraints,[143] and so forth. In the future, T5 will continue to incorporate new technologies to adapt to evolving molecular language forms, and multilingual capabilities are a new direction in its development.

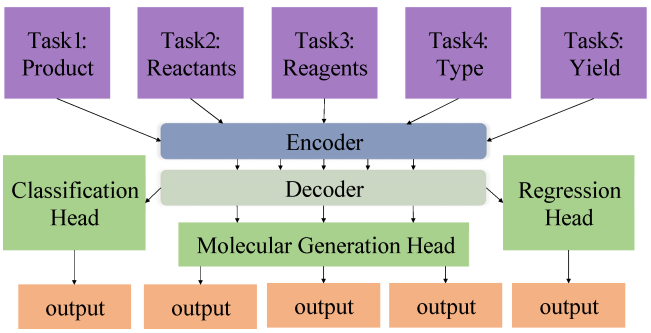

FIGURE 7 Flowchart of T5Chem, which is a multi-tasking model where there are five tasks. The similarity is that all of them have encoder and decoder structure, the difference is that the prompts and head types are different for different types of tasks.[139]

3.7 | Vision transformers (ViT)

Vision Transformer (ViT) pioneers the application of transformer architecture in image recognition, utilizing self-attention to capture inter-patch relationships.[144] Departing from CNNs, ViT partitions images into patches, treats them as sequential elements, and processes them using transformers. ViT exhibits outstanding performance in computer vision tasks, driving the adoption of transformer models across diverse domains, including NLP and multimodal learning.[145,146] Recently, the academic community in molecular science has also shown growing interest in ViT, a trend explored further in the subsequent discussion.

Organic molecules were not only essential as the basic components of life but also served as the crucial raw materials for living products such as drugs, food, cosmetics, and so forth. Allahyani et al. employed Wasserstein Generative Adversarial Network (WGAN) to generate conditional molecules, enhancing the dataset for subsequent experiments.[147] This generated dataset was accurately categorized using Vision Transformer (ViT) with 97% accuracy, showcasing a streamlined approach for organic molecule generation and prediction, thereby reducing time and costs in chemical experimentation.

Additionally, several studies have leveraged ViT in diverse applications: integrating ViT with microbial diagnostic techniques like surface-enhanced Raman scattering (SERS) facilitated rapid bacterial identification[148]; self-supervised visual transformers (ss-ViTs) accurately depicted stem cell phenotype heterogeneity[149]; ViT-WSI modeling accurately predicted molecular marker status in diagnostic gliomas using whole-slide images[150]; Zhang et al. developed the mass spectrum transformer (MST) alongside GNN in TransG-Net to predict molecular properties based on mass spectrometry data.[151] Despite their widespread applications, ViT models have limitations. They typically require larger datasets for training compared to CNNs, which can be challenging in data-scarce domains. ViTs are also computationally intensive, demanding significant resources for training and inference.

3.8 | Detection transformer (DETR)

DETR, introduced by Carion et al. in 2020, simplifies target detection by treating it as an ensemble prediction problem, utilizing an ensemble-based full loss function for training, and performing bipartite graph matching between predicted and real objects, eliminating the need for anchor boxes and nonmaximal suppression.[152] While primarily designed for computer vision tasks such as target detection and segmentation,[153] including instance and semantic segmentation,[154,155] DETR also finds applications in the molecular science domain, which are described in detail below.

Single-particle cryo-electron microscopy (cryo-EM) is a biostructuring technique that directly captures the structure of biomolecules without the need for crystallization.[156] To enhance cryo-EM particle selection accuracy, Zhang et al. developed the TransPicker framework, the first application of a transformer-based approach to cryo-EM particle picking, resulting in significant performance improvements and a substantial reduction in processing time.[157]

Most chemical literature conveys new molecules and reactions through 2D molecule images, lacking machine-readable descriptors like SMILES. Campos et al. introduced IMG2SMI, employing a deep residual network for image features and an encoder–decoder transformer for molecular descriptors, offering superior performance in molecular description generation compared to optical structure recognition application (OSRA)-based systems.[158]

Cell segmentation poses a challenge in biomedical object detection.[159] Prangemeier et al. introduced Cell-DETR, an attention-based model, combining CNN and transformer encoders for image feature extraction and attention-based processing, respectively.[160] The framework of the Cell-DETR model is shown in Figure 8. Specifically, Cell-DETR's prediction heads, comprising feed-forward neural networks (FFNN), produced bounding box and classification predictions, while its segmentation head, employing multi-head attention and a CNN decoder, generated segmentation maps for each object instance. This model excelled in systems and synthetic biology, accurately detecting cells and providing instance-level segmentation maps for morphology and fluorescence measurements, surpassing previous methods in yeast segmentation performance with comparable runtime.

DETR's strong performance stems from its end-to-end learning and contextual awareness, yet challenges like lengthy training and small target detection persist. Recent enhancements like efficient DETR,[161] deformable DETR,[162] and sparse DETR models[163] address these issues. Notably, UP-DETR improves detection accuracy on smaller datasets like PASCAL VOC.[164] While DETR's current use in molecular science is limited, it shows potential for analyzing

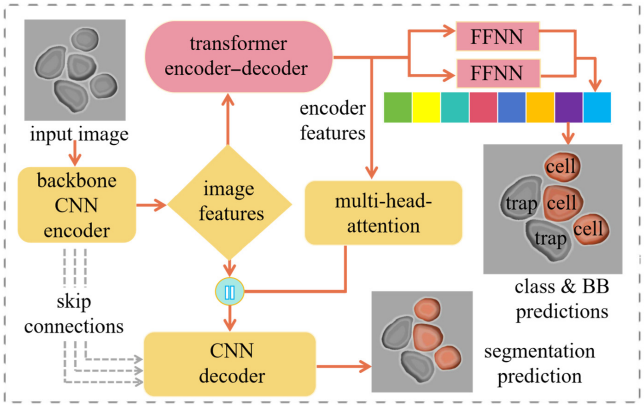

FIGURE 8 The framework of the Cell-DETR model.[160] Cell-DETR consists of a backbone CNN encoder, a transformer encoder–decoder, a bounding box and class prediction header, and a segmentation header.

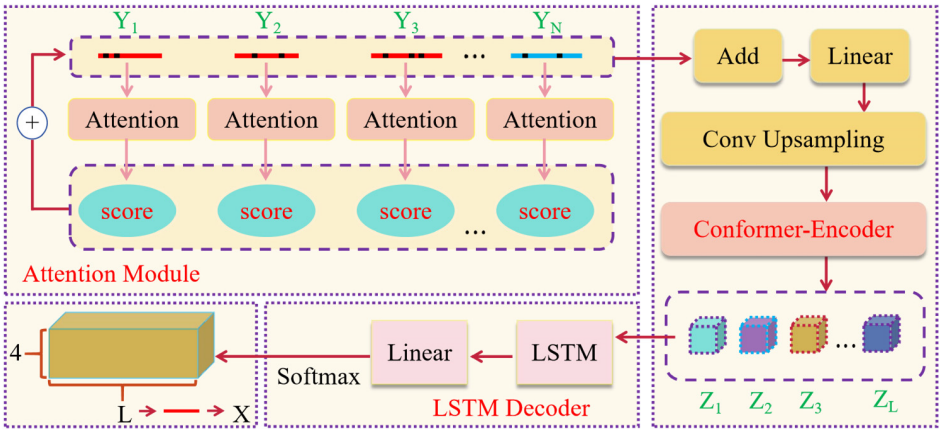

FIGURE 9 The RRCC-DNN module architecture.[171] RRCC-DNN consists of three key components: The Attention Module, the Conformer-Encoder, and the LSTM-Decoder.

cellular and molecular structures in microscopy and medical images, enhancing cell image analysis efficiency, and aiding in drug discovery by identifying compound features.

3.9 | Conformer

The Conformer model, proposed by Gulati et al., integrates convolutional modules into the transformer structure to enhance speech recognition by effectively capturing both local patterns and global information.[165] By combining the strengths of CNNs and transformers, Conformer excels in tasks requiring long-range dependencies and fine-grained patterns. It has been successfully applied in various domains such as speech recognition,[166] language segmentation,[167] speech translation,[168] text-to-speech (TTS),[169] and image processing.[170] The applications of Conformer in molecular science are further detailed below.

Recently, Qin et al. introduced the RRCC-DNN model to address the impact of contaminating sequences during DNA storage on multi-read reconstruction.[171] By integrating self-attention and a conformer block, this model effectively handles noisy reads and outlier sequences containing IDS errors. Structured on an encoder–decoder architecture, the model comprises three key components, as illustrated in Figure 9: an Attention module to mitigate contaminating sequences, a Conformer-Encoder to correct IDS errors, and a single-layer LSTM-Decoder for reference DNA prediction. Tsui et al. proposed NG-DTA, a DL method for predicting drug-target affinity (DTA), leveraging conformer architecture.[172] The NG-DTA model comprises two GNNs for drug molecular graph embeddings, followed by graph convolutional networks (GCN) and local extrema convolution (LEConv). Three conformer blocks process the sequence of

3-gram graph embeddings, with final predictions made by connecting protein and drug embeddings through attentional weighting. Comparative testing on Davis and Kiba datasets demonstrated that incorporating n-gram molecular subgraphs of proteins enhanced DL model performance in DTA prediction, especially with increased protein samples and utilization of LEConv for capturing local graph information.

The conformer model blends self-attention and convolutional structures, effectively handling global dependencies and local features in sequence data processing. Its self-attention mechanism enables efficient handling of long sequences, addressing issues like gradient vanishing or explosion common in traditional RNNs. While promising for tasks like sequence annotation or classification in bioinformatics, conformer models entail computational costs due to their complexity and sensitivity to parameters, warranting careful consideration in usage based on data characteristics.

3.10 | Contrastive language-image pre-training (CLIP)

Contrastive language-image pre-training (CLIP), a multimodal learning transformer model proposed by OpenAI, efficiently learns visual concepts through natural language supervision.[173] It comprises a CNN for images and a transformer for text, enabling prediction of relevant text segments from images with natural language cues, akin to zero-shot functionality seen in GPT models.[39] Its introduction has expanded possibilities for multimodal learning, including applications in molecular science,[174,175] offering a new approach for tackling various challenges related to molecular structure, properties, and reactions.

Functional peptides have the potential to target difficult-to-drug targets for therapy.[176] Palepu et al. introduced a sequence-based peptide design framework utilizing the CLIP model in 2022, enabling the targeting of challenging therapy targets.[177] They developed a streamlined inference process, Cut&CLIP, leveraging existing experimental binding proteins to rapidly select peptides for validation. By linking candidate peptides with E3 ubiquitin ligase domains, they successfully achieved intracellular degradation of pathogenic protein targets in human cells, laying groundwork for enhancing peptide design technology and clinical applications. Figure 10 shows the process they used to train peptide-protein pairs with the CLIP model.

To gain a deeper understanding of biological mechanisms and precision medicine, we need to have a clear knowledge of genes and their relationships in biological pathways. He et al. introduced pathCLIP, a graph-text contrastive learning framework in 2023, inspired by CLIP, to discern genes and their relationships from images and texts, thus enhancing gene identification accuracy through contrastive learning.[178] Evaluation revealed pathCLIP's superior performance in gene identification tasks, with notable improvements in accuracy, recall, and F1 score. Furthermore, validation using pathway data from PubMed showcased pathCLIP's multimodal capability, surpassing models reliant on single modality inputs, underscoring its potential for gene identification and pathway analysis.

Contrastive learning, a potent approach for acquiring nuanced data representations, was leveraged by Sanchez et al. in their 2023 framework, CLOOME.[179] Integrating CLIP and CLOOB (contrastive leave-one-out boost) methods, they facilitated zero-shot transfer learning for multimodal data, yielding impressive outcomes. By embedding biological images and chemical structures into a unified space, they harnessed a multimodal contrast learning strategy, revealing the bio-image encoder's versatility in diverse drug discovery predictions like activity, classification, and mechanism identification.

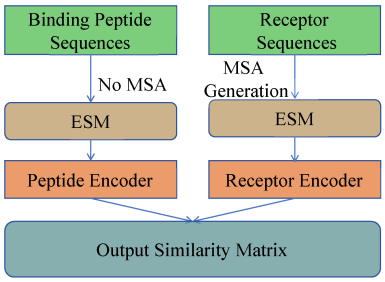

FIGURE 10 CLIP training process for peptide–protein pairs. The peptide and receptor encoders are jointly trained on the ESM embedding and finally the similarity matrix is obtained.[177]

CLIP, a pioneering deep-learning model, revolutionizes multimodal learning, fostering profound insights into molecular structures by establishing a unified language-image space. Its standout features include robust generalization across diverse tasks in the field of molecular science, as well as multi-task capability, offering a versatile tool for comprehensive research. Yet, challenges persist, notably the demand for extensive data and the model's broad focus. Optimizing CLIP for specialized applications in molecular science domain is crucial for unlocking its full potential.

3.11 | Sparse transformers

In 2019, OpenAI introduced sparse transformers, a neural network architecture optimized for sequence data processing, offering superior performance and efficiency in various deep learning tasks.[180] Notably, sparse transformers employ sparse attention matrices, reducing complexity to $O(n\sqrt{n})$, and incorporate architectural enhancements for deeper networks, memory efficiency, and accelerated training.[181,182] These advances enable effective utilization across diverse domains like image,[183] audio,[184] and text processing.[185] Given the intricate nature of molecular structures, characterized by numerous atoms and bonds, sparse transformers hold significant promise and applicability in molecular science. We'll further explore their specific applications in this domain.

Predicting drug-target interactions (DTI) computationally is crucial in drug development, aiming to expedite the process and reduce costs.[186] In 2021, Kim et al. proposed an interpretable framework to identify interaction sites, employing a gated cross-attention mechanism to emphasize drug and target features and foster explicit interactions between them.[187] The gated cross-attention network is shown in Figure 11. Their findings underscore the sensitivity of gated cross-attention to mutations, offering valuable insights for developing drugs targeting mutant proteins.

Sparse transformers are integral to the field of molecular science, offering key advantages. They excel in conserving memory and processing longer sequences efficiently, while also demonstrating excellent computational speed for large-scale networks.[188] Moreover, they provide flexibility in implementing various sparse methods without sacrificing performance. However, introducing sparsity may impact the model's ability to capture dependencies between specific tokens. Achieving a balance between processing speed and information flow is crucial when implementing sparsity in sparse transformers. Overall, with ongoing research and development, sparse transformers have the potential to significantly enhance sequence processing tasks, improving the efficiency of AI models.

3.12 | Mobile and efficient transformers

Transformer models, prevalent in NLP, face constraints in mobile applications due to high computational demands.[102] To address this, researchers have devised mobile-efficient variants through techniques like compression, pruning, and distillation.[189] Examples include TinyBERT, MobileBERT, and AutoTinyBERT,[190–192] which offer smaller sizes, reduced computational costs, and maintained performance, broadening the transformer models' applicability. These mobile-efficient models find increasing use in molecular science, showcasing notable performance.

Combining multiple drugs for treating complex diseases has become a significant therapeutic strategy due to their synergistic effects.[193,194] Zhang et al. introduced Molormer in 2022, a DDI prediction method utilizing a lightweight

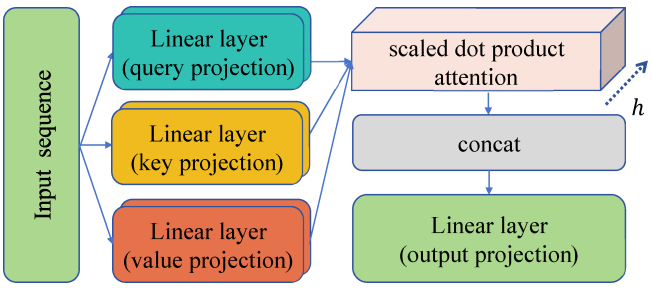

FIGURE 11 Overview of the gated cross-attention (GCA) network and detailed procedures when obtaining protein attention (right). GCA uses multi-head gated attention to explicitly construct interactions between drug and target features that are obtained from each feature extractor.[187]

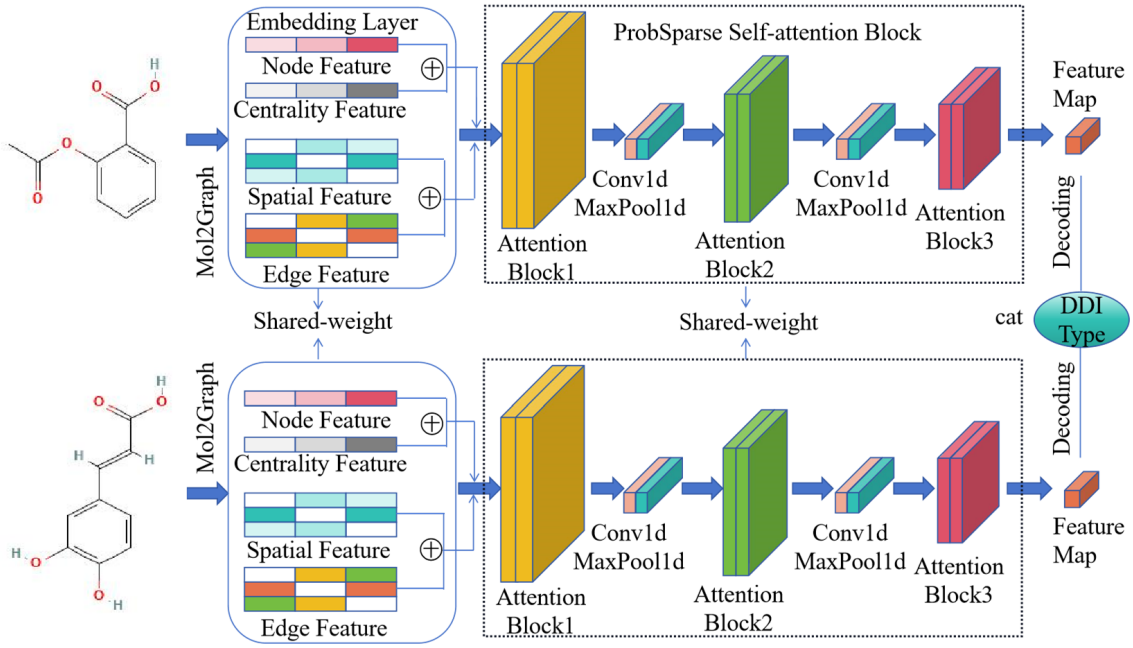

FIGURE 12 The overview of Molormer. In Molormer, the original drug graph goes through an embedding layer to form four feature embeddings representing the spatial structure of the drug. These features are processed by two ProbSparse self-attentive blocks, connected and fed to the decoder for final prediction.[195]

attention mechanism, and achieved improved performance while minimizing computational costs, as validated through experiments outperforming existing methods on multilabel DDI datasets.[195] The general framework of Molormer is shown in Figure 12.

In human genetics, interpreting non-coding genomes poses challenges due to limited annotation of active elements.[196] Yang et al. introduced LOGO (Language of Genomes) in 2022, a lightweight pre-trained language model with two self-attentive layers trained on unlabeled human reference genomes.[197] LOGO was fine-tuned for sequence labeling tasks and extended with input coding strategies, including variant prioritization and a convolutional module for enhanced performance, highlighting its accuracy, efficiency, scalability, and robustness.

DL has demonstrated potential in genome sequence analysis, particularly in motif identification and variant calling.[198] Wichmann et al. introduced MetaTransformer in 2023, a DL tool for macrogenome analysis leveraging a self-attention mechanism.[199] The model's transformer encoder architecture facilitated efficient parallel computation and improved computational speed, with the study exploring various embedding schemes to enhance performance and reduce memory consumption. Results indicated significant performance improvements in processing macrogenome sequencing data using self-attention models and embedding schemes in DL.

Mobile and efficient transformers have demonstrated effectiveness in molecular level applications, offering advantages over traditional transformer models. They excel in resource-constrained environments like mobile devices and edge computing due to their smaller model sizes and faster computational speed.[200] However, their reduced model size may impact accuracy and contextual understanding compared to larger models.[201] Achieving a balance between model size and performance optimization is crucial for their widespread adoption in molecular science as mobile technology advances.

4 | OUTLOOK

In this review, we outline the objectives and methodology for investigating transformer-based machine learning techniques in molecular science research. State-of-the-art technologies like ChatGPT and multifunctional transformers have deepened our understanding of molecular structures, sequences, and functions. Additionally, molecular-aware transformers can improve the effectiveness in handling diverse data types and predicting protein properties. The emerging

superintelligence will revolutionize the application of AI and ML in complex molecular systems and accelerate progress in drug development, chemical synthesis, protein engineering, and beyond.

4.1 | Integration of ChatGPT into molecular science

Transformers enable the development of large language models like ChatGPT and chatbots by utilizing self-attention mechanisms to efficiently process and generate text, leading to advanced capabilities in language understanding and generation. Consequently, the emergence of chatbots has heralded a paradigmatic shift in the landscape of natural language processing and comprehension. Particularly within the domain of molecular science, ChatGPT's discerning capability to comprehend and generate text akin to human language assumes a paramount role as a potent instrument for facilitating seamless communication among researchers. This proficiency transcends mere articulation of ideas, extending to nuanced expression of hypotheses and sharing profound insights within the scientific community.

The unique characteristics and advantages of ChatGPT enable it to assist researchers in analyzing complex datasets, conducting literature reviews, and synthesizing information from diverse sources with remarkable efficiency. In chemistry, ChatGPT could accelerate drug discovery processes by facilitating analysis and generation of large-scale chemical data and predicting molecular druggability properties. Similarly, in biology, it could aid in interpreting genomic and proteomic data, identifying spatiotemporal patterns, and generating hypotheses for further experimentation.

4.2 | Multifunction transformers and molecular feature learning

The advent of multifunction transformers represents a seminal breakthrough in the field of molecular sciences, offering a sophisticated means of extracting intricate features from molecular structures and sequences. These transformative models showcase an extraordinary proficiency in discerning subtle patterns and correlations embedded within expansive datasets, thereby providing researchers with an unprecedented tool set to unveil concealed insights into the intricacies of molecular behavior. The implications of this advanced feature learning extend across a spectrum of critical domains within scientific inquiry. In the realm of drug discovery, for instance, multifunction transformers play an instrumental role in elucidating the nuanced relationships between molecular structures and pharmacological activities, thus expediting the identification of potential therapeutic agents. Furthermore, in the domain of materials science, the capacity of these transformers to decipher complex molecular features holds significant promise for the rational design and optimization of novel materials with tailored properties. This section intricately explores the far-reaching consequences of multifunction transformers' prowess in molecular feature learning, shedding light on their transformative potential across diverse scientific disciplines and underscoring their role as catalysts for innovation in molecular research.

4.3 | Comprehensive molecular-aware transformers

The advent of comprehensive molecular-aware transformers marks a profound paradigm shift in the comprehension of molecular entities within scientific discourse. The molecular awareness indicates that these transformative models could identify a diverse array of data types within the molecular field, namely text, sequence, structure, image, energy, molecular dynamics, and function, which may usher in a new era of understanding and transcend traditional boundaries. The significance of these molecular-aware transformers lies in their ability to provide researchers with a holistic perspective on complex molecular systems, bridging the gaps between disparate forms of data and offering a unified framework for analysis. This comprehensive approach holds the promise of significantly augmenting the precision and scope of predictions, simulations, and analyses across a multitude of molecular disciplines. Within this context, the review meticulously scrutinizes the impact of such holistic models on the accuracy of predictions, demonstrating their capacity to unravel intricate relationships within molecular structures and functions. Furthermore, the integration of diverse data types empowers researchers to derive nuanced insights, uncover hidden patterns, and make informed decisions in various scientific pursuits, from drug discovery and materials science to bioinformatics and beyond. As such, the exploration of comprehensive molecular-aware transformers in this review illuminates their potential to redefine

the boundaries of molecular research, fostering a more interconnected and insightful understanding of the multifaceted world of molecular entities.

4.4 | Self-assessment transformers and superintelligence

The conceptualization of self-assessment transformers introduces a captivating and innovative dimension to the field of molecular sciences, signifying a paradigmatic shift in the capabilities of intelligent models. These sophisticated transformers exhibit a unique prowess, extending beyond conventional learning mechanisms by assimilating insights not only from correlated data but also from seemingly disparate and unrelated sources. The depth of their intelligence lies in their capacity to discern patterns and derive meaningful conclusions from information that may not conventionally appear pertinent to molecular structures or functions.

This review delves into the intriguing prospect of superintelligence within the realm of molecular self-assessment. The exploration of this concept is not merely confined to the acquisition of knowledge within the molecular sciences; rather, it transcends disciplinary boundaries by investigating how these transformers can autonomously glean valuable insights from a diverse spectrum of fields. The potential for superintelligence in molecular self-assessment holds the promise of contributing to unforeseen and groundbreaking advancements in molecular research. By leveraging insights derived from seemingly unrelated sources, self-assessment transformers have the potential to uncover novel connections, propose innovative hypotheses, and catalyze paradigm shifts in our understanding of molecular phenomena. The review meticulously scrutinizes this transformative potential, shedding light on the prospect of these intelligent models as catalysts for serendipitous discoveries and novel breakthroughs that could redefine the landscape of molecular sciences in unprecedented ways.

The four aspects discussed above, such as the chatbots, multifunction transformers, molecular-aware capabilities, and self-assessment intelligence are the potential development directions of transformer models in the future. As transformers continue to reshape the landscape of molecular science, the combination of these aspects promises a future where our understanding of molecular entities is not only profound but also dynamic and adaptive. This review underscores the transformative potential of these technologies and sets the stage for a collaborative and innovative future in molecular science research.

AUTHOR CONTRIBUTIONS

Jian Jiang: Data curation (equal); formal analysis (equal); methodology (equal); writing – original draft (equal); writing – review and editing (equal). **Lu Ke:** Data curation (equal); formal analysis (equal); writing – original draft (equal). **Long Chen:** Data curation (equal); formal analysis (equal); writing – original draft (equal). **Bozheng Dou:** Data curation (equal); formal analysis (equal); writing – original draft (equal). **Yueying Zhu:** Data curation (equal); formal analysis (equal). **Jie Liu:** Data curation (equal); formal analysis (equal). **Bengong Zhang:** Data curation (equal); formal analysis (equal); funding acquisition (equal). **Tianshou Zhou:** Project administration (equal); supervision (supporting); writing – review and editing (supporting). **Guo-Wei Wei:** Conceptualization (lead); funding acquisition (lead); resources (lead); supervision (lead); writing – review and editing (lead).

FUNDING INFORMATION

This study was supported by the NIH grants R01GM126189, R01AI164266, and R35GM148196; National Science Foundation grants DMS2052983, DMS-1761320, and IIS-1900473; NASA grant 80NSSC21M0023; MSU Foundation; BristolMyers Squibb 65109; Pfizer; National Natural Science Foundation of China under Grant Nos. 11971367, 12271416, and 11972266.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Jie Liu https://orcid.org/0000-0002-4828-3244
Guo-Wei Wei https://orcid.org/0000-0001-8132-5998

RELATED WIREs ARTICLES

Efficient and enhanced sampling of drug-like chemical space for virtual screening and molecular design using modern machine learning methods

REFERENCES

1. Carleo G, Cirac I, Cranmer K, Daudet L, Schuld M, Tishby N, et al. Machine learning and the physical sciences. Rev Mod Phys. 2019; 91(4):045002.
2. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15(141):20170387.
3. Fan C, Sun Y, Zhao Y, Song M, Wang J. Deep learning-based feature engineering methods for improved building energy prediction. Appl Energy. 2019;240:35–45.
4. Hinton G. Deep learning—a technology with the potential to transform health care. JAMA. 2018;320(11):1101–2.
5. Choi SR, Lee M. Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. Biology. 2023; 12(7):1033.
6. Wei L, Fu N, Song Y, Wang Q, Hu J. Probabilistic generative transformer language models for generative design of molecules. J Chem. 2023;15(1):88.
7. Dou B, Zhu Z, Merkurjev E, Ke L, Chen L, Jiang J, et al. Machine learning methods for small data challenges in molecular science. Chem Rev. 2023;123(13):8736–80.
8. Wang H, Fu T, Du Y, Gao W, Huang K, Liu Z, et al. Scientific discovery in the age of artificial intelligence. Nature. 2023;620(7972): 47–60.
9. Chandra A, Tünnermann L, Löfstedt T, Gratz R. Transformer-based deep learning for predicting protein properties in the life sciences. Elife. 2023;12:e82819.
10. Liu C, Sun Y, Davis R, Cardona ST, Hu P. ABT-MPNN: an atom-bond transformer-based message-passing neural network for molecular property prediction. J Chem. 2023;15(1):29.
11. Sadad T, Aurangzeb RA, Safran M, Imran, Alfarhood S, Kim J. Classification of highly divergent viruses from DNA/RNA sequence using transformer-based models. Biomedicine. 2023;11(5):1323.
12. Zhao H, Zhu B, Jiang T, Cui Z, Wu H. A transformer-based deep learning approach with multi-layer feature processing for accurate prediction of protein–DNA binding residues. International conference on intelligent computing. Berlin: Springer; 2023. p. 556–67.
13. Pan J. Large language model for molecular chemistry. Nat Comput Sci. 2023;3(1):5.
14. Shen X, Han D, Guo Z, Chen C, Hua J, Luo G. Local self-attention in transformer for visual question answering. Appl Intell. 2023; 53(13):16706–23.
15. Deng J, Yang Z, Wang H, Ojima I, Samaras D, Wang F. A systematic study of key elements underlying molecular property prediction. Nat Commun. 2023;14(1):6395.
16. Monteiro NR, Pereira TO, Machado ACD, Oliveira JL, Abbasi M, Arrais JP. Fsmddtr: end-to-end feedback strategy for multi-objective de novo drug design using transformers. Comput Biol Med. 2023;164:107285.
17. Chen D, Liu J, Wei G-W. Multiscale topology-enabled structure-to-sequence transformer for protein–ligand interaction predictions. Nat Mach Intell. 2024;6:799–810. https://doi.org/10.1038/s42256-02400855-1
18. Hu S, Liu J, Yang R, Wang Y, Wang A, Li K, et al. Exploring the applicability of transfer learning and feature engineering in epilepsy prediction using hybrid transformer model. IEEE Trans Neural Syst Rehabil Eng. 2023;31:1321–32.
19. Chen D, Gao K, Nguyen DD, Chen X, Jiang Y, Wei G-W, et al. Algebraic graph-assisted bidirectional transformers for molecular property prediction. Nat Commun. 2021;12(1):3521.
20. Shi Z, Zhang H, Chang K-W, Huang M, Hsieh C-J. Robustness verification for transformers; 2020. arXiv preprint arXiv:2002.06622.
21. Zhou D, Yu Z, Xie E, Xiao C, Anandkumar A, Feng J, et al. Understanding the robustness in vision transformers. International conference on machine learning. Baltimore, MA: PMLR; 2022. p. 27378–94.
22. Buehler MJ. Multiscale modeling at the interface of molecular mechanics and natural language through attention neural networks. Acc Chem Res. 2022;55(23):3387–403.
23. Rafiei F, Zeraati H, Abbasi K, Ghasemi JB, Parsaeian M, Masoudi-Nejad A. Deeptrasynergy: drug combinations using multimodal deep learning with transformers. Bioinformatics. 2023;39(8):btad438.
24. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of generative pretrained transformer 3 (GPT-3) in healthcare delivery. NPJ Digit Med. 2021;4(1):93.
25. Zhang Y, Pei H, Zhen S, Li Q, Liang F. Chat generative pre-trained transformer (ChatGPT) usage in healthcare. Gastroenterol Endosc. 2023;1(3):139–43.
26. Ganesh P, Chen Y, Lou X, Khan MA, Yang Y, Sajjad H, et al. Compressing large-scale transformer-based models: a case study on BERT. Trans Assoc Comput Linguist. 2021;9:1061–80.

27. Alokla A, Gad W, Nazih W, Aref M, Salem A-b. Pseudocode generation from source code using the BART model. Mathematics. 2022; 10(21):3967.
28. Karnyoto AS, Sun C, Liu B, Wang X. TB-BCG: topic-based BART counterfeit generator for fake news detection. Mathematics. 2022; 10(4):585.
29. Cai D, Lam W. Graph transformer for graph-to-sequence learning. Proceedings of the AAAI conference on artificial intelligence; Washington, DC: AAAI Press; 2020. p. 7464–71.
30. Zhang Z, Liu Q, Hu Q, Lee C-K. Hierarchical graph transformer with adaptive node sampling. Adv Neural Inform Process Syst. 2022; 35:21171–83.
31. Zhang X, Yang S, Duan L, Lang Z, Shi Z, Sun L. Transformer-XL with graph neural network for source code summarization. IEEE international conference on systems, man, and cybernetics (SMC). Melbourne: IEEE; 2021. p. 3436–41.
32. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(1):5485–551.
33. Rodriguez-Torrealba R, Garcia-Lopez E, Garcia-Cabot A. End-to-end generation of multiple choice questions using text-to-text transfer transformer models. Expert Syst Appl. 2022;208:118258.
34. Fan H, Xiong B, Mangalam K, Li Y, Yan Z, Malik J, et al. Multiscale vision transformers. Proceedings of the IEEE/CVF international conference on computer vision; Washington, DC: IEEE; 2021. p. 6824–35.
35. Ranftl R, Bochkovskiy A, Koltun V. Vision transformers for dense prediction. Proceedings of the IEEE/CVF international conference on computer vision; Washington, DC: IEEE; 2021. p. 12179–88.
36. Fang Y, Liao B, Wang X, Fang J, Qi J, Wu R, et al. You only look at one sequence: rethinking transformer in vision through object detection. Adv Neural Inform Process Syst. 2021;34:26183–97.
37. Peng Z, Huang W, Gu S, Xie L, Wang Y, Jiao J, et al. Conformer: local features coupling global representations for visual recognition. Proceedings of the IEEE/CVF international conference on computer vision; Washington, DC: IEEE; 2021. p. 367–76.
38. Vyas P, Kuznetsova A, Williamson DS. Optimally encoding inductive biases into the transformer improves end-to-end speech translation. Interspeech. Brno: ISCA; 2021;2287–91.
39. Pan X, Ye T, Han D, Song S, Huang G. Contrastive language-image pre-training with knowledge graphs. Adv Neural Inform Process Syst. 2022;35:22895–910.
40. Zhou J, Dong L, Gan Z, Wang L, Wei F. Non-contrastive learning meets language-image pre-training. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; Washington, DC: IEEE; 2023. p. 11028–38.
41. Yun C, Chang Y-W, Bhojanapalli S, Rawat AS, Reddi S, Kumar S. O (n) connections are expressive enough: universal approximability of sparse transformers. Adv Neural Inform Process Syst. 2020;33:13783–94.
42. Pan J, Bulat A, Tan F, Zhu X, Dudziak L, Li H, et al. EDGEVITS: competing light-weight cnns on mobile devices with vision transformers. European conference on computer vision. Berlin: Springer; 2022. p. 294–311.
43. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inform Process Syst. 2017;30:1–11.
44. Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. Neurocomputing. 2021;452:48–62.
45. Tao C, Gao S, Shang M, Wu W, Zhao D, Yan R. Get the point of my utterance! Learning towards effective responses with multi-head attention mechanism. Proceedings of the twenty-seventh international joint conference on artificial intelligence. Stockholm, Sweden: International Joint Conference on Artificial Intelligence Organization (IJCAI); 2018. p. 4418–24.
46. Voita E, Talbot D, Moiseev F, Sennrich R, Titov I. Analyzing multi-head self-attention: specialized heads do the heavy lifting, the rest can be pruned; 2019. arXiv preprint arXiv:1905.09418.
47. Jacovi A, Shalom OS, Goldberg Y. Understanding convolutional neural networks for text classification; 2018. arXiv preprint arXiv: 1809.08037.
48. Skansi S. Introduction to deep learning: from logical calculus to artificial intelligence. Berlin: Springer; 2018.
49. Lu S, Wang M, Liang S, Lin J, Wang Z. Hardware accelerator for multi-head attention and position-wise feed-forward in the transformer. IEEE 33rd international system-on-chip conference (SOCC). Washington, DC: IEEE; 2020. p. 84–9.
50. Zhang S, Fan R, Liu Y, Chen S, Liu Q, Zeng W. Applications of transformer-based language models in bioinformatics: a survey. Bioinform Adv. 2023;3(1):vbad001.
51. Zheng Z, Zhong Y, Tian S, Ma A, Zhang L. Changemask: deep multi-task encoder-transformer-decoder architecture for semantic change detection. ISPRS J Photogramm Rem Sens. 2022;183:228–39.
52. Li Y, Cai W, Gao Y, Li C, Hu X. More than encoder: introducing transformer decoder to upsample. IEEE international conference on bioinformatics and biomedicine (BIBM). Washington, DC: IEEE; 2022. p. 1597–602.
53. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. OpenAI blog; 2018.
54. Mager M, Astudillo RF, Naseem T, Sultan MA, Lee Y-S, Florian R, et al. Gpttoo: a language-model-first approach for AMR-to-text generation; 2020. arXiv preprint arXiv:2005.09123.
55. Qu Y, Liu P, Song W, Liu L, Cheng M. A text generation and prediction system: pretraining on new corpora using BERT and GPT-2. IEEE 10th international conference on electronics information and emergency communication (ICEIEC). Washington, DC: IEEE; 2020. p. 323–6.
56. Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art. 2023;6(1):9.

57. Sawai R, Paik I, Kuwana A. Sentence augmentation for language translation using GPT-2. Electronics. 2021;10(24):3082.
58. Bahani M, El Ouaazizi A, Maalmi K. The effectiveness of T5, GPT-2, and BERT on text-to-image generation task. Pattern Recogn Lett. 2023;173:57–63.
59. Lecler A, Duron L, Soyer P. Revolutionizing radiology with gpt-based models: current applications, future possibilities and limitations of ChatGPT. Diagn Interv Imaging. 2023;104(6):269–74.
60. Bagal V, Aggarwal R, Vinod P, Priyakumar UD. Molgpt: molecular generation using a transformer-decoder model. J Chem Inf Model. 2021;62(9):2064–76.
61. Li J, Liu Y, Fan W, Wei X-Y, Liu H, Tang J, et al. Empowering molecule discovery for molecule-caption translation with large language models: a ChatGPT perspective; 2023. arXiv preprint arXiv:2306.06615.
62. Maeda K, Kurata H. Automatic generation of SBML kinetic models from natural language texts using GPT. Int J Mol Sci. 2023;24(8):7296.
63. Wang R, Feng H, Wei G-W. ChatGPT in drug discovery: a case study on anticocaine addiction drug development with chatbots. J Chem Inf Model. 2023;63(22):7189–209.
64. Wang Y, Zhao H, Sciabola S, Wang W. cmolgpt: a conditional generative pre-trained transformer for target-specific de novo molecular generation. Molecules. 2023;28(11):4430.
65. Goel M, Aggarwal R, Sridharan B, Pal PK, Priyakumar UD. Efficient and enhanced sampling of drug-like chemical space for virtual screening and molecular design using modern machine learning methods. WIREs Comput Mol Sci. 2023;13(2):e1637.
66. Haroon S, Hafsath C, Jereesh A. Generative pre-trained transformer (GPT) based model with relative attention for de novo drug design. Comput Biol Chem. 2023;106:107911.
67. Wang X, Gao C, Han P, Li X, Chen W, Rodríguez Patón A, et al. Petrans: de novo drug design with protein-specific encoding based on transfer learning. Int J Mol Sci. 2023;24(2):1146.
68. Hobbs HT, Liu CC. Learning to read and write in the language of proteins. GEN Biotechnology. 2023;2(2):92–4.
69. Rehana H, Çam NB, Basmaci M, He Y, Özgür A, Hur J. Evaluation of GPT and BERT-based models on identifying protein–protein interactions in biomedical text; 2023. arXiv preprint arXiv:2303.17728.
70. Jablonka KM, Schwaller P, Smit B. Is GPT-3 all you need for machine learning for chemistry? AI for accelerated materials design NeurIPS 2022 work. New Orleans, Louisiana: Neural Information Processing Systems (NeurIPS); 2022.
71. Jubair S, Tucker JR, Henderson N, Hiebert CW, Badea A, Domaratzki M, et al. Gptransformer: a transformer-based deep learning method for predicting fusarium related traits in barley. Front Plant Sci. 2021;12:761402.
72. Xu X, Xu T, Zhou J, Liao X, Zhang R, Wang Y, et al. Ab-gen: antibody library design with generative pre-trained transformer and deep reinforcement learning. Genomics Proteomics Bioinformatics. 2023;21:1043–53.
73. Liu J, Shen L, Wei G-W. ChatGPT for computational topology; 2023. arXiv preprint arXiv:2310.07570.
74. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding; 2018. arXiv preprint arXiv:1810.04805.
75. Na K-I, Kim U-H, Kim J-H. SPU-BERT: faster human multi-trajectory prediction from socio-physical understanding of BERT. Knowl Based Syst. 2023;274:110637.
76. Zhao A, Yu Y. Knowledge-enabled BERT for aspect-based sentiment analysis. Knowl Based Syst. 2021;227:107220.
77. Saga T, Tanaka H, Iwasaka H, Nakamura S. Multimodal prediction of social responsiveness score with BERT-based text features. IEICE Trans Inform Syst. 2022;105(3):578–86.
78. Xu H, Liu B, Shu L, Yu PS. BERT post-training for review reading comprehension and aspect-based sentiment analysis; 2019. arXiv preprint arXiv:1904.02232.
79. Wang Z, Ng P, Ma X, Nallapati R, Xiang B. Multi-passage BERT: a globally normalized BERT model for open-domain question answering; 2019. arXiv preprint arXiv:1908.08167.
80. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. Bertscore: evaluating text generation with BERT; 2019. arXiv preprint arXiv:1904.09675.
81. Qiu Y, Wei G-W. Persistent spectral theory-guided protein engineering. Nat Comput Sci. 2023;3(2):149–63.
82. Ramsay RR, Popovic-Nikolic MR, Nikolic K, Uliassi E, Bolognesi ML. A perspective on multi-target drug discovery and design for complex diseases. Clin Transl Med. 2018;7(1):1–14.
83. Xie A, Zhang Z, Guan J, Zhou S. Self-supervised learning with chemistry-aware fragmentation for effective molecular property prediction. Brief Bioinform. 2023;24:bbad296.
84. Wen N, Liu G, Zhang J, Zhang R, Fu Y, Han X. A fingerprints based molecular property prediction method using the BERT model. J Chem. 2022;14(1):1–13.
85. Olaya-Abril A, Jiménez-Munguía I, Gómez-Gascón L, Rodríguez-Ortega MJ. Surfomics: shaving live organisms for a fast proteomic identification of surface proteins. J Proteomics. 2014;97:164–76.
86. Jha K, Karmakar S, Saha S. Graph-BERT and language model-based framework for protein–protein interaction identification. Sci Rep. 2023;13(1):5663.
87. Wang S, Guo Y, Wang Y, Sun H, Huang J. Smiles-BERT: large scale unsupervised pretraining for molecular property prediction. Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics; New York, NY: Association for Computing Machinery (ACM); 2019. p. 429–36.

88. Li J, Jiang X. MOL-BERT: an effective molecular representation with BERT for molecular property prediction. Wirel Commun Mobile Comput. 2021;2021:1–7.
89. Park M, Seo S-w, Park E, Kim J. Epibertope: a sequence-based pre-trained BERT model improves linear and structural epitope prediction by learning long-distance protein interactions effectively. bioRxiv. 2022;2022–02. https://doi.org/10.1101/2022.02.27.481241
90. Dumortier B, Liutkus A, Carré C, Krouk G. PETRIBERT: augmenting BERT with tridimensional encoding for inverse protein folding and design. bioRxiv. 2022;2022-08. https://doi.org/10.1101/2022.08.10.503344
91. Chen D, Zheng J, Wei G-W, Pan F. Extracting predictive representations from hundreds of millions of molecules. J Phys Chem Lett. 2021;12(44):10793–801.
92. Shen L, Feng H, Qiu Y, Wei G-W. SVSBI: sequence-based virtual screening of biomolecular interactions. Commun Biol. 2023;6(1):536.
93. Wu Z, Jiang D, Wang J, Zhang X, Du H, Pan L, et al. Knowledge-based BERT: a method to extract molecular features like computational chemists. Brief Bioinform. 2022;23(3):bbac131.
94. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension; 2019. arXiv preprint arXiv:1910.13461.
95. Liu Y, Gu J, Goyal N, Li X, Edunov S, Ghazvininejad M, et al. Multilingual denoising pre-training for neural machine translation. Trans Assoc Comput Linguist. 2020;8:726–42.
96. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer; 2020. arXiv preprint arXiv:2004.05150.
97. La Quatra M, Cagliero L. BART-IT: an efficient sequence-to-sequence model for Italian text summarization. Future Internet. 2022;15(1):15.
98. Xiong H, Yan Z, Wu C, Lu G, Pang S, Xue Y, et al. BART-based contrastive and retrospective network for aspect-category-opinion-sentiment quadruple extraction. Int J Mach Learn Cybern. 2023;14:1–13.
99. Yoshikai Y, Mizuno T, Nemoto S, Kusuhara H. Difficulty in learning chirality for transformer fed with smiles; 2023. arXiv preprint arXiv:2303.11593.
100. Sai Prakash MV, Nareddy SR, Parab G, Venkatesan V, Vaddina V, Gopalakrishnan S. Synergistic fusion of graph and transformer features for enhanced molecular property prediction. bioRxiv. 2023;2023-08. https://doi.org/10.1101/2023.08.28.555089
101. Fu N, Wei L, Song Y, Li Q, Xin R, Omee SS, et al. Material transformers: deep learning language models for generative materials design. Mach Learn Sci Technol. 2023;4(1):015001.
102. Liu X, Ye K, van Vlijmen HW, IJzerman AP, van Westen GJ. DRUGEX V3: scaffold constrained drug design with graph transformer-based reinforcement learning. J Chem. 2023;15(1):24.
103. Dimitriadis S. Multi-task regression QSAR/QSPR prediction utilizing text-based transformer neural network and single-task using feature-based models. Linkoping, Sweden: Linkoping University; 2021.
104. Irwin R, Dimitriadis S, He J, Bjerrum EJ. Chemformer: a pre-trained transformer for computational chemistry. Mach Learn Sci Technol. 2022;3(1):015022.
105. Butler T, Frandsen A, Lightheart R, Bargh B, Taylor J, Bollerman T, et al. Ms2mol: a transformer model for illuminating dark chemical space from mass spectra. ChemRxiv. 2023. https://doi.org/10.26434/chemrxiv-2023-vsmpx-v2
106. Chilingaryan G, Tamoyan H, Tevosyan A, Babayan N, Khondkaryan L, Hambardzumyan K, et al. MolBART: generative masked language models for molecular representations; Proceedings of the Eleventh International Conference on Learning Representations (ICLR) 2023. Kigali, Rwanda: ICLR; 2023.
107. Hödl S, Robinson W, Bachrach Y, Huck W, Kachman T. Explainability techniques for chemical language models; 2023. arXiv preprint arXiv:2305.16192.
108. Pichler WJ. The important role of non-covalent drug–protein interactions in drug hypersensitivity reactions. Allergy. 2022;77(2):404–15.
109. Likos CN. Effective interactions in soft condensed matter physics. Phys Rep. 2001;348(4–5):267–439.
110. Bramer D, Wei G-W. Multiscale weighted colored graphs for protein flexibility and rigidity analysis. J Chem Phys. 2018;148(5):054103.
111. Das KC, Mondal S. On neighborhood inverse sum indeg index of molecular graphs with chemical significance. Inform Sci. 2023;623:112–31.
112. Jiang B, Xu F, Zhang Z, Tang J, Nie F. Agformer: efficient graph representation with anchor-graph transformer; 2023. arXiv preprint arXiv:2305.07521.
113. Zhang Y, Bao W, Cao Y, Cong H, Chen B, Chen Y. A survey on protein–DNA-binding sites in computational biology. Brief Funct Genomics. 2022;21(5):357–75.
114. Yuan Q, Chen S, Rao J, Zheng S, Zhao H, Yang Y. Alphafold2-aware protein–DNA binding site prediction using graph transformer. Brief Bioinform. 2022;23(2):bbab564.
115. Li H, Zhao D, Zeng J. KPGT: knowledge-guided pre-training of graph transformer for molecular property prediction. Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining; New York, NY: Association for Computing Machinery (ACM); 2022. p. 857–67.
116. Chen B, Barzilay R, Jaakkola T. Path-augmented graph transformer network; 2019. arXiv preprint arXiv:1905.12712.
117. Tan Z, Zhao Y, Zhou T, Lin K. HI-MGT: a hybrid molecule graph transformer for toxicity identification. J Hazard Mater. 2023;457:131808.
118. Hu J, Zhang X, Shang D, Ouyang L, Li Y, Xiong D. Egtsyn: edge-based graph transformer for anti-cancer drug combination synergy prediction; 2023. arXiv preprint arXiv:2303.10312.

119. Teng S, Yin C, Wang Y, Chen X, Yan Z, Cui L, et al. Molfpg: multi-level fingerprint based graph transformer for accurate and robust drug toxicity prediction. Comput Biol Med. 2023;164:106904.
120. Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R. Transformer-XL: attentive language models beyond a fixed-length context; 2019. arXiv preprint arXiv:1901.02860.
121. Peric L, Mijic S, Stammbach D, Ash E. Legal language modeling with transformers. Proceedings of the fourth workshop on automated semantic analysis of information in legal text (ASAIL 2020) held online in conjunction with the 33rd international conference on legal knowledge and information systems (JURIX 2020) December 9, 2020. Volume 2764. Stroudsburg, PA: CEUR-WS; 2020.
122. Bonetta G, Cancelliere R, Liu D, Vozila P. Retrieval-augmented transformer-XL for close domain dialog generation; 2021. arXiv preprint arXiv:2105.09235.
123. Habbat N, Anoun H, Hassouni L. Combination of GRU and CNN deep learning models for sentiment analysis on French customer reviews using xlnet model. IEEE Eng Manag Rev. 2022;51(1):41–51.
124. Yu L, Sartran L, Stokowiec W, Ling W, Kong L, Blunsom P, et al. Better document level machine translation with Bayes' rule. Trans Assoc Comput Linguist. 2020;8:346–60.
125. Wu X, Wang C, Lei Q. Transformer-XL based music generation with multiple sequences of time-valued notes; 2020. arXiv preprint arXiv:2007.07244.
126. Osipenko S, Botashev K, Nikolaev E, Kostyukevich Y. Transfer learning for small molecule retention predictions. J Chromatogr A. 2021;1644:462119.
127. Honda S, Shi S, Ueda HR. Smiles transformer: pre-trained molecular fingerprint for low data drug discovery; 2019. arXiv preprint arXiv:1911.04738.
128. Mandhana V, Taware R. De novo drug design using self attention mechanism. Proceedings of the 35th annual ACM symposium on applied computing; New York, NY: Association for Compu.ting Machinery (ACM); 2020. p. 8–12.
129. Nayak P, Silberfarb A, Chen R, Muezzinoglu T, Byrnes J. Transformer based molecule encoding for property prediction; 2020. arXiv preprint arXiv:2011.03518.
130. Wang X, Yao C, Zhang Y, Yu J, Qiao H, Zhang C, et al. From theory to experiment: transformer-based generation enables rapid discovery of novel reactions. J Chem. 2022;14(1):1–14.
131. Bird JJ, Ekárt A, Faria DR. Chatbot interaction with artificial intelligence: human data augmentation with T5 and language transformer ensemble for text classification. J Ambient Intell Humaniz Comput. 2023;14(4):3129–44.
132. Chi Z, Dong L, Ma S, Mao SHX-L, Huang H, Wei F. MT6: multilingual pretrained text-to-text transformer with translation pairs; 2021. arXiv preprint arXiv:2104.08692.
133. Halder K, Akbik A, Krapac J, Vollgraf R. Task-aware representation of sentences for generic text classification. Proceedings of the 28th international conference on computational linguistics; New York, NY: International Committee on Computational Linguistics (ICCL); 2020. p. 3202–13.
134. Nagoudi EMB, Chen W-R, Abdul-Mageed M, Cavusogl H. Indt5: a text-to-text transformer for 10 indigenous languages; 2021. arXiv preprint arXiv:2104.07483.
135. Zhuang S, Li H, Zuccon G. Deep query likelihood model for information retrieval. Advances in information retrieval: 43rd European conference on IR research, ECIR 2021, Virtual event, March 28–April 1, 2021, Part II 43. Berlin: Springer; 2021. p. 463–70.
136. Fenoy E, Edera AA, Stegmayer G. Transfer learning in proteins: evaluating novel protein learned representations for bioinformatics tasks. Brief Bioinform. 2022;23(4):bbac232.
137. Orzhenovskii M. T5-long-extract at FNS-2021 shared task. Proceedings of the 3rd financial narrative processing workshop; Stroudsburg, PA: Association for Computational Linguistics (ACL); 2021. p. 67–9.
138. Lu J. Integrating machine learning into synthetic organic chemistry [PhD thesis]. New York University; 2022.
139. Lu J, Zhang Y. Unified deep learning model for multitask reaction predictions with explanation. J Chem Inf Model. 2022;62(6):1376–87.
140. Rothchild D, Tamkin A, Yu J, Misra U, Gonzalez J. C5T5: controllable generation of organic molecules with transformers; 2021. arXiv preprint arXiv:2108.10307.
141. Diao S, Zhou W, Zhang X, Wang J. Write and paint: generative vision-language models are unified modal learners. The eleventh international conference on learning representations. Kigali, Rwanda: International Conference on Learning Representations (ICLR); 2023.
142. Ciosici MR, Derczynski L. Training a T5 using lab-sized resources; 2023. arXiv preprint arXiv:2208.12097.
143. Mastropaolo A, Scalabrino S, Cooper N, Palacio DN, Poshyvanyk D, Oliveto R, et al. Studying the usage of text-to-text transfer transformer to support code-related tasks. IEEE/ACM 43rd international conference on software engineering (ICSE). Washington, DC: IEEE; 2021. p. 336–47.
144. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16×16 words: Transformers for image recognition at scale. International conference on learning representations. Vienna, Austria: International Conference on Learning Representations (ICLR); 2021.
145. Dou Z-Y, Xu Y, Gan Z, Wang J, Wang S, Wang L, et al. An empirical study of training end-to-end vision-and-language transformers. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; Washington, DC: IEEE; 2022. p. 18166–76.
146. Yao J, Zhang B, Li C, Hong D, Chanussot J. Extended vision transformer (exvit) for land use and land cover classification: a multimodal deep learning framework. IEEE Trans Geosci Remote Sens. 2023;61:1–15.

147. Allahyani R, Alsowat A, Alsulami R, Almaqati S, Alafif T, Hawsawi M, et al. Learning to generate and predict new conditional small organic molecules. 2023. https://www.researchgate.net/publication/373739192
148. Tseng Y-M, Chen K-L, Chao P-H, Han Y-Y, Huang N-T. Deep learning-assisted surface enhanced Raman scattering for rapid bacterial identification. ACS Appl Mater Interfaces. 2023;15:26398–406.
149. Pfaendler R, Hanimann J, Lee S, Snijder B. Self-supervised vision transformers accurately decode cellular state heterogeneity. bioRxiv. 2023;2023-01. https://doi.org/10.1101/2023.01.16.524226
150. Wu B, Xu C, Dai X, Wan A, Zhang P, Yan Z, et al. Visual transformers: token-based image representation and processing for computer vision; 2020. arXiv preprint arXiv:2006.03677.
151. Zhang T, Chen S, Wulamu A, Guo X, Li Q, Zheng H. Transg-net: transformer and graph neural network based multi-modal data fusion network for molecular properties prediction. Appl Intell. 2023;53(12):16077–88.
152. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. European conference on computer vision. Berlin: Springer; 2020. p. 213–29.
153. Zhang G, Luo Z, Cui K, Lu S, Xing EP. Meta-DETR: image-level few-shot detection with inter-class correlation exploitation. IEEE Trans Pattern Anal Mach Intell. 2022;45:1–12.
154. Dersch S, Schöttl A, Krzystek P, Heurich M. Towards complete tree crown delineation by instance segmentation with mask R-CNN and DETR using UAV-based multispectral imagery and lidar data. ISPRS Open J Photogramm Rem Sens. 2023;8:100037.
155. He H, Cai J, Pan Z, Liu J, Zhang J, Tao D, et al. Dynamic focus-aware positional queries for semantic segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; Washington, DC: IEEE; 2023. p. 11299–308.
156. Lyumkis D. Challenges and opportunities in cryo-EM single-particle analysis. J Biol Chem. 2019;294(13):5181–97.
157. Zhang C, Li H, Wan X, Chen X, Yang Z, Feng J, et al. Transpicker: a transformer-based framework for particle picking in cryoEM micrographs. IEEE international conference on bioinformatics and biomedicine (BIBM). Washington, DC: IEEE; 2021. p. 1179–84.
158. Campos D, Ji H. Img2smi: translating molecular structure images to simplified molecular input line-entry system; 2021. arXiv preprint arXiv:2109.04202.
159. Haque IRI, Neubert J. Deep learning approaches to biomedical image segmentation. Inform Med Unlocked. 2020;18:100297.
160. Prangemeier T, Reich C, Koeppl H. Attention-based transformers for instance segmentation of cells in microstructures. IEEE international conference on bioinformatics and biomedicine (BIBM). Washington, DC: IEEE; 2020. p. 700–7.
161. Yao Z, Ai J, Li B, Zhang C. Efficient DETR: improving end-to-end object detector with dense prior; 2021. arXiv preprint arXiv: 2104.01318.
162. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: deformable transformers for end-to-end object detection; 2020. arXiv preprint arXiv:2010.04159.
163. Roh B, Shin J, Shin W, Kim S. Sparse DETR: efficient end-to-end object detection with learnable sparsity; 2021. arXiv preprint arXiv: 2111.14330.
164. Dai Z, Cai B, Lin Y, Chen J. UP-DETR: unsupervised pre-training for object detection with transformers. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; Washington, DC: IEEE; 2021. p. 1601–10.
165. Gulati A, Qin J, Chiu C-C, Parmar N, Zhang Y, Yu J, et al. Conformer: convolution-augmented transformer for speech recognition; 2020. arXiv preprint arXiv:2005.08100.
166. Burchi M, Timofte R. Audio-visual efficient conformer for robust speech recognition. Proceedings of the IEEE/CVF winter conference on applications of computer vision; Washington, DC: IEEE; 2023. p. 2258–67.
167. Li C, Wang Y, Deng F, Zhang Z, Wang X, Wang Z. EAD-conformer: a conformer-based encoder-attention-decoder-network for multi-task audio source separation. ICASSP 2022—IEEE international conference on acoustics, speech and signal processing (ICASSP). Washington, DC: IEEE; 2022. p. 521–5.
168. Kumar LA, Renuka DK, Priya VH, Sudarshan S. Spoken language translation using conformer model. International conference on intelligent systems for communication, IoT and security (ICISCoIS). Washington, DC: IEEE; 2023. p. 466–71.
169. Choi Y, Jang J, Koo M-W. A Korean menu-ordering sentence text-to-speech system using conformer-based fastspeech2. J Acoust Soc Korea. 2022;41(3):359–66.
170. Deng Z, Yu W, Che L, Chen S, Zhang Z, Shang J, et al. Text to image generation with conformer-GAN. International conference on neural information processing. Berlin: Springer; 2023. p. 3–14.
171. Qin Y, Zhu F, Xi B. Robust multi-read reconstruction from contaminated clusters using deep neural network for DNA storage; 2022. arXiv preprint arXiv:2210.11106.
172. Tsui L-I, Hsu T-C, Lin C. NG-DTA: drug-target affinity prediction with N-gram molecular graphs. 45th annual international conference of the IEEE Engineering in Medicine & Biology Society (EMBC). Washington, DC: IEEE; 2023. p. 1–4.
173. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. International conference on machine learning. Baltimore, MA: PMLR; 2021. p. 8748–63.
174. Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. Nat Biomed Eng. 2022;6(12):1346–52.
175. Pei X, Zuo K, Li Y, Pang Z. A review of the application of multi-modal deep learning in medicine: bibliometrics and future directions. Int J Comput Intell Syst. 2023;16(1):44.
176. Feng D, Liu L, Shi Y, Du P, Xu S, Zhu Z, et al. Current development of bicyclic peptides. Chin Chem Lett. 2023;34(6):108026.
177. Palepu K, Ponnapati M, Bhat S, Tysinger E, Stan T, Brixi G, et al. Design of peptide-based protein degraders via contrastive deep learning. bioRxiv. 2022;2022-05. https://doi.org/10.1101/2022.05.23.493169

178. He F, Liu K, Yang Z, Chen Y, Hammer R, Xu D, et al. pathclip: detection of genes and gene relations from biological pathway figures through image-text contrastive learning. bioRxiv. 2023;2023-10. https://doi.org/10.1109/JBHI.2024.3383610
179. Sanchez-Fernandez A, Rumetshofer E, Hochreiter S, Klambauer G. Cloome: contrastive learning unlocks bioimaging databases for queries with chemical structures. Nat Commun. 2023;14(1):7339.
180. Child R, Gray S, Radford A, Sutskever I. Generating long sequences with sparse transformers; 2019. arXiv preprint arXiv:1904.10509.
181. Fedus W, Zoph B, Shazeer N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. J Mach Learn Res. 2022;23(1):5232–70.
182. Lin T, Wang Y, Liu X, Qiu X. A survey of transformers. AI Open. 2022;3:111–32.
183. Yang M, Liu Z, Dong W, Wu Y. Sstnet: saliency sparse transformers network with tokenized dilation for salient object detection. IET Image Process. 2023;17:3759–76.
184. Zhang Z, He B, Zhang Z. Transmask: a compact and fast speech separation model based on transformer. ICASSP 2021—IEEE international conference on acoustics, speech and signal processing (ICASSP). Washington, DC: IEEE; 2021. p. 5764–8.
185. Ren H, Dai H, Dai Z, Yang M, Leskovec J, Schuurmans D, et al. Combiner: full attention transformer with sparse computation cost. Adv Neural Inform Process Syst. 2021;34:22470–82.
186. Chen L, Jiang J, Dou B, Feng H, Liu J, Zhu Y, et al. Machine learning study of the extended drug–target interaction network informed by pain related voltage gated sodium channels. Pain. 2024;165(4):908–21.
187. Kim Y, Shin B. An interpretable framework for drug-target interaction with gated cross attention. Machine learning for healthcare conference. Baltimore, MA: PMLR; 2021. p. 337–53.
188. Fang C, Zhou A, Wang Z. An algorithm–hardware co-optimized framework for accelerating N: M sparse transformers. IEEE Trans Very Large Scale Integr Syst. 2022;30(11):1573–86.
189. Wu Z, Liu Z, Lin J, Lin Y, Han S. Lite transformer with long-short range attention. 2020. https://doi.org/10.48550/arXiv.2004.11886
190. Jiao X, Yin Y, Shang L, Jiang X, Chen X, Li L, et al. Tinybert: distilling BERT for natural language understanding; 2019. arXiv preprint arXiv:1909.10351.
191. Sun Z, Yu H, Song X, Liu R, Yang Y, Zhou D. Mobilebert: a compact task-agnostic BERT for resource-limited devices; 2020. arXiv preprint arXiv:2004.02984.
192. Yin Y, Chen C, Shang L, Jiang X, Chen X, Liu Q. Autotinybert: automatic hyperparameter optimization for efficient pre-trained language models; 2021. arXiv preprint arXiv:2107.13686.
193. Lin S, Wang Y, Zhang L, Chu Y, Liu Y, Fang Y, et al. MDF-SA-DDI: predicting drug–drug interaction events based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism. Brief Bioinform. 2022;23(1):bbab421.
194. Xuan P, Li P, Cui H, Wang M, Nakaguchi T, Zhang T. Learning multi-types of neighbor node attributes and semantics by heterogeneous graph transformer and multi-view attention for drug-related side-effect prediction. Molecules. 2023;28(18):6544.
195. Zhang X, Wang G, Meng X, Wang S, Zhang Y, Rodriguez-Paton A, et al. Molormer: a lightweight self-attention-based method focused on spatial structure of molecular graph for drug–drug interactions prediction. Brief Bioinform. 2022;23(5):bbac296.
196. Nabi A, Dilekoglu B, Adebali O, Tastan O. Discovering misannotated LNCRNAS using deep learning training dynamics. Bioinformatics. 2023;39(1):btac821.
197. Yang M, Huang L, Huang H, Tang H, Zhang N, Yang H, et al. Integrating convolution and self-attention improves language model of human genome for interpreting noncoding regions at base-resolution. Nucleic Acids Res. 2022;50(14):e81.
198. Clauwaert J, Menschaert G, Waegeman W. Explainability in transformer models for functional genomics. Brief Bioinform. 2021;22(5):bbab060.
199. Wichmann A, Buschong E, Müller A, Jünger D, Hildebrandt A, Hankeln T, et al. Metatransformer: deep metagenomic sequencing read classification using self-attention models. NAR Genom Bioinform. 2023;5(3):lqad082.
200. Kwon E, Song H, Park J, Kang S. Mobile accelerator exploiting sparsity of multi-heads, lines, and blocks in transformers in computer vision. Design, automation & test in Europe conference & exhibition (DATE). Washington, DC: IEEE; 2023. p. 1–6.
201. Reza S, Ferreira MC, Machado JJM, Tavares JMR. A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks. Expert Syst Appl. 2022;202:117275.

How to cite this article: Jiang J, Ke L, Chen L, Dou B, Zhu Y, Liu J, et al. Transformer technology in molecular science. WIREs Comput Mol Sci. 2024;14(4):e1725. https://doi.org/10.1002/wcms.1725