JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Application of Transformers in Cheminformatics

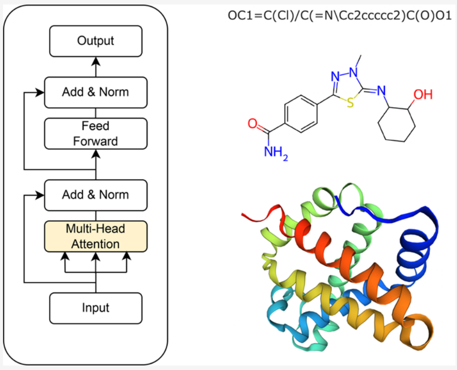
Kha-Dinh Luong* and Ambuj Singh*

Cite This: J. Chem. Inf. Model. 2024, 64, 4392−4409

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: By accelerating time-consuming processes with high efficiency, computing has become an essential part of many modern chemical pipelines. Machine learning is a class of computing methods that can discover patterns within chemical data and utilize this knowledge for a wide variety of downstream tasks, such as property prediction or substance generation. The complex and diverse chemical space requires complex machine learning architectures with great learning power. Recently, learning models based on transformer architectures have revolutionized multiple domains of machine learning, including natural language processing and computer vision. Naturally, there have been ongoing endeavors in adopting these techniques to the chemical domain, resulting in a surge of publications within a short period. The diversity of chemical structures, use cases, and learning models necessitate a comprehensive summarization of existing works. In this paper, we review recent innovations in adapting transformers to solve learning problems in chemistry. Because chemical data is diverse and complex, we structure our discussion based on chemical representations. Specifically, we highlight the strengths and weaknesses of each representation, the current progress of adapting transformer architectures, and future directions.



KEYWORDS: *cheminformatics, machine learning, chemical representations, transformer, graphs, sequences*

1. INTRODUCTION

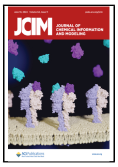
Chemical experiments are often expensive and time-consuming, requiring domain expertise, sophisticated equipment, and laborious operations. The long history of chemical sciences results in ample documented experimental data suitable as inputs for machine learning (ML) algorithms. By discovering useful chemical patterns within the data, ML can automate costly processes and accelerate scientific advancement. Most notably, predictive models spur the development of quantitative structure−activity relationships (QSARs) by utilizing curated labeled data sets of chemical and physical properties. Chemical generative models trained on large molecular databases help discover novel chemical structures, which is of interest in drug discovery. Utilization of ML techniques also facilitates automation in other tasks such as synthesis, experimental planning, and physical simulation.

Structural data such as chemical compounds is relatively new to ML. Earlier attempts transformed chemical structures into vectorized representations fitting traditional methods such as Support Vector Machine or Decision Tree. More recently, advanced learning architectures capable of encoding structural dependencies like convolutional neural networks (CNNs) or graph neural networks (GNNs) have been applied to learning on chemical data.[1−3] Nevertheless, chemical structures are challenging for ML. Unlike image or text which has uniform structural patterns, chemical data contain irregular connectivities between information-rich components. As a result, there is a crucial need for complex ML methods with greater learning power to capture sophisticated chemical patterns.

Since their introduction in 2017,[4] transformer architectures have revolutionized learning in multiple ML domains.[5,6] Consequently, there is growing interest in developing transformer-based learning models for chemical structures. Thanks to the attention mechanism, transformers can capture long-range structural dependency, which is highly beneficial in learning complex chemical interactions. Given the diversity and structural specificity of the chemical space, this emerging field remains open to numerous challenges. Nevertheless, its potential is significant and has attracted increasing attention from the research community, leading to a substantial body of literature in a short time. To further facilitate research in this direction, we provide a comprehensive survey of existing works.

Our review of transformers in the chemical domain is structured as follows. Section 2 explains the transformer architectures and highlights related studies. After that, the review is organized around the representations of chemical data. Specifically, sections 3 and 4 delve into sequence-based

Received: December 28, 2023
Revised: April 5, 2024
Accepted: May 6, 2024
Published: May 30, 2024



https://doi.org/10.1021/acs.jcim.3c02070

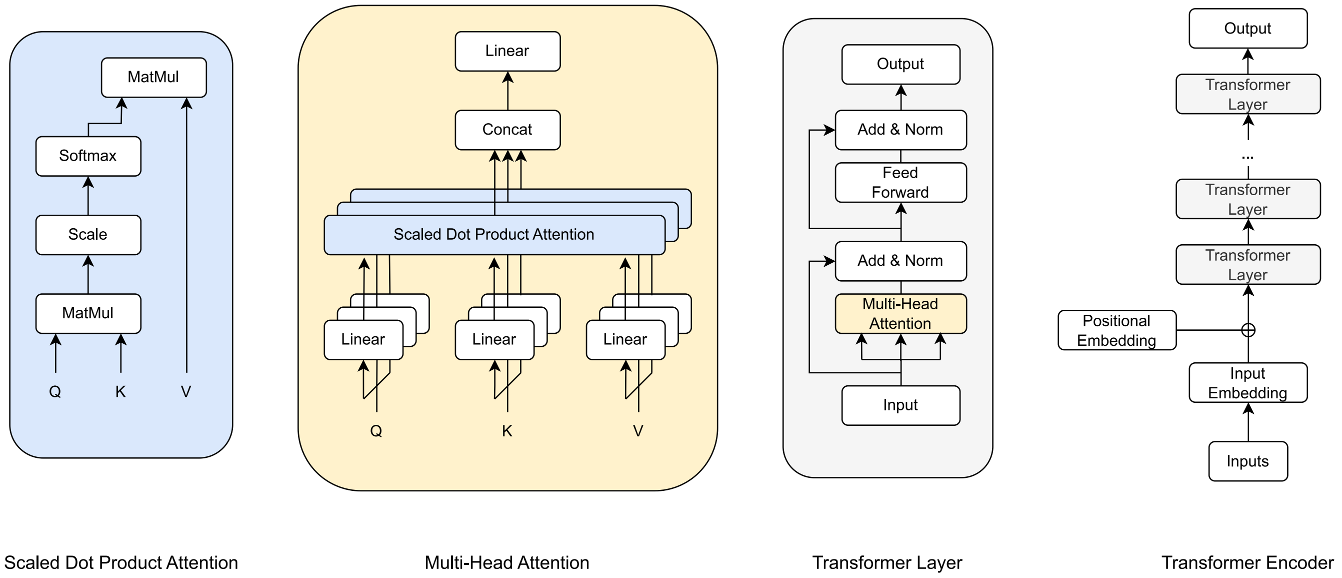


Figure 1. Major components of a transformer network.

representations, while section 5 explores graph-based representations. Section 6 explores the application side of transformers in cheminformatics and section 7 discusses future directions. Our objective is to elucidate the strengths and limitations inherent in different data representations as well as their unique challenges. Crucially, we highlight the innovative approaches that have been proposed to overcome these challenges. Throughout this paper, we aspire to present a valuable entry point for researchers from diverse communities, offering an insightful introduction to this intriguing area of research.

2. BACKGROUND

2.1. Transformer Architecture. Transformers are the latest advancement in deep learning, building on the success of other well-established architectures such as convolutional neural networks (CNNs) for computer vision (CV) and recurrent neural networks (RNNs) for natural language processing (NLP).[4] The fundamental operation of transformers is simple. At each iteration or layer of the algorithm, the embedding of each element undergoes updates through referencing and combining with the embeddings of other elements. These elements can take various forms, such as tokens from a sentence, pixels from an image, or nodes from a graph. The major components and flows in a transformer network are illustrated in Figure 1.

Referencing in transformers is executed using the multihead-attention mechanism, which employs the query-keyword-value (QKV) model.[4] This nomenclature draws inspiration from information retrieval, which focuses on assessing the relevance of a query to a given keyword. The objective of a transformer model is slightly different: we seek to determine the degree of attention, or weighting, that an element should assign when referencing another element.

Let $X^{(i)} \in \mathbb{R}^{N\times D}$ be the embedding of elements after the ith layer, where N is the number of elements and D is the input embedding size. We define three matrices $Q \in \mathbb{R}^{N\times D_k}$, $K \in \mathbb{R}^{N\times D_k}$, and $V \in \mathbb{R}^{N\times D}$ as

$$\begin{aligned} Q &= X^{(i)}W_Q \\ K &= X^{(i)}W_K \\ V &= X^{(i)}W_V, \end{aligned} \tag{1}$$

where D_k is the intermediate embedding size and $W_Q \in \mathbb{R}^{D\times D_k}$, $W_K \in \mathbb{R}^{D\times D_k}$, and $W_V \in \mathbb{R}^{D\times D}$ are separate projection heads. Let the attention matrix be $A = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right)$, the single-head attention is calculated as

$$\mathrm{attn}(X^{(i)}) = AV = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right)V \tag{2}$$

The attention matrix A contains pairwise attention weight between elements. Weights on the diagonal are regarded as self-attentions. Consequently, multihead attention concatenates multiple single-head attentions:

$$\mathrm{multihead}(X^{(i)}) = \mathrm{concat}(\mathrm{attn}_1(X^{(i)}), \mathrm{attn}_2(X^{(i)}), ..., \mathrm{attn}_H(X^{(i)})) \tag{3}$$

The embeddings of the elements are updated with residual connections that add the embeddings of the previous output:

$$X^{(i+1)} = \mathrm{MLP}_i(\mathrm{multihead}(X^{(i)}) + X^{(i)}) \tag{4}$$

The attention mechanism allows each element, a word token or an atom, to reference any other elements within the same sentence or molecule. This is usually referred to as global attentions. However, if the sentence is too long or the molecule is too large, calculating the whole matrix of pairwise attentions is computationally costly. In these cases, the attention can be constrained to the surrounding neighborhood of each element, i.e, local attention. Often, additional location context is provided via positional encodings, which distinguishes elements based on their positions within the data. For example, a word token may attend differently to tokens within the same sentence than tokens from other sentences. In the original transformer,[4] absolute positional encodings are added to the element embeddings $X^{(0)}$ before feeding them into the network.

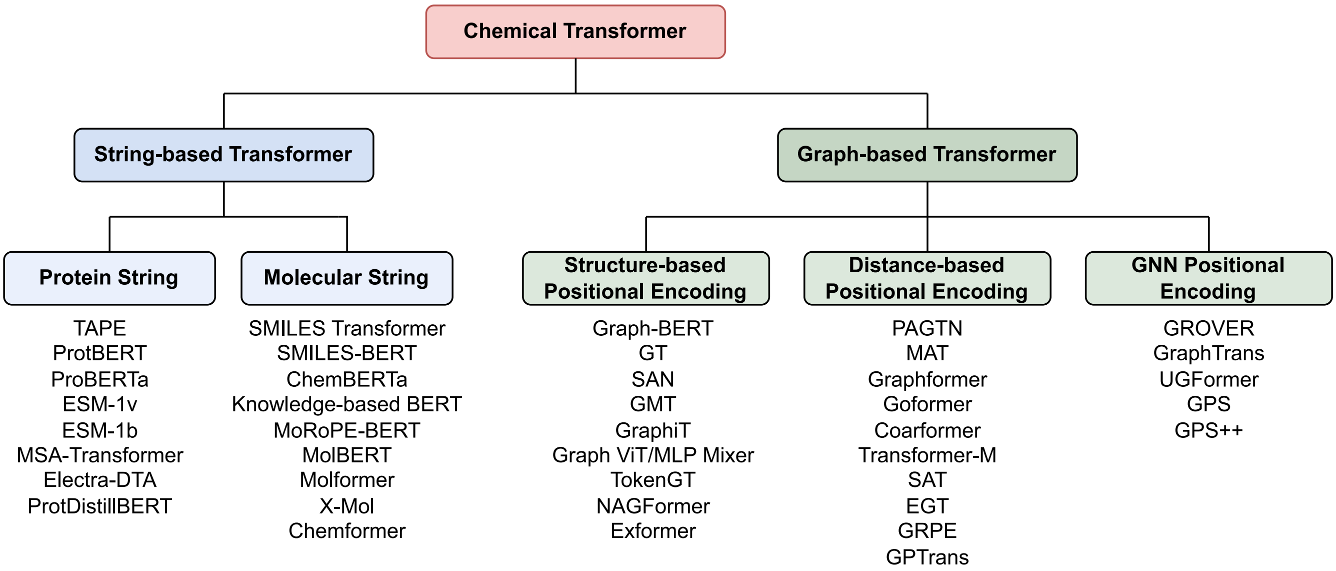


Figure 2. An overview of the chemical transformer landscape with a sample of methods in each category. Structure representations can be broadly categorized into either string-based or graph-based.

Recently, relative positional encodings have been a promising approach that offers generalization to data of unseen lengths and sizes.[7] Positional encodings are an essential part and a major challenge when adapting transformers to structural data, which we discuss more in section 5.2.

The attention mechanism, coupled with global referencing, bestows transformers with distinct advantages over other deep learning architectures like CNNs. Unlike convolutional kernels, which are static with fixed kernel weights that do not adapt to each pixel and its surrounding context, attention matrices offer dynamic learning, taking into account pairwise contextual relationships. Furthermore, the use of global attention enables an exceptionally broad perception field, surpassing the typical reach of convolutional kernels. This expanded perception field allows collecting and updating information from any element in a single step, marking a significant step up from sequential models such as recurrent neural networks (RNNs) or long short-term memory networks (LSTMs), which require as many steps as the length of the input sequence to process all elements. As a result, transformers, with the attention mechanism, are revolutionary models that consistently achieve and maintain leading positions in multiple domains of ML.

2.2. Transformers in Text and Image Processing. The original transformer architecture was designed for machine translation tasks. However, since the introduction of BERT,[8] a groundbreaking transformer-based design, transformers have solidified their position as preferred architectures for a wide array of language-related learning problems. Numerous transformer-based models have since emerged, addressing challenges related to performance, generalizability, and scalability.[9,10] Currently, modern transformer models with billions of parameters can be effectively and efficiently trained, leading to the development of large language models (LLMs).[11,12] These pretrained LLMs possess a remarkable ability to learn from vast corpora of text using appropriate self-supervised learning strategies. Once pretrained, these models become powerful general language understanding systems, capable of performing human-like tasks, such as translation, summarization, text generation, and question answering, with human-level proficiency.[12] When fine-tuned, an LLM easily outperforms other nonpretrained models on predictive tasks, solidifying the status of transformer-based models as state-of-the-art in NLP. Furthermore, their influence extends to various other domains of machine learning.

Transformer-based models have achieved remarkable success in computer vision (CV) and have been applied to a broad spectrum of tasks, including object detection, tracking, segmentation, and image classification.[13] For a long time, variations of CNNs dominated the CV landscape. Transitioning from CNNs to vision-based transformers entails forfeiting some inductive bias, which is characteristic of convolutional kernels that discover generally reusable patterns, in favor of enhanced learning capabilities facilitated by pairwise context-based attentions. ViT is one of the most prominent vision-based transformers, adopting the same underlying principles as its text-based counterpart.[14] Instead of processing word tokens, ViT divides images into gridlike patches and arranges them into an input sequence. Being a complex model, ViT is harder to train than CNNs and can outperform CNNs on large data sets but not on tasks with limited data. Recent variations of ViT, such as XCiT, PiT, LV-ViT, and DeiT, have made strides in closing this performance gap on smaller data sets by incorporating features like feature channel attentions, hierarchical feature mapping, auxiliary supervision, or knowledge distillation.[15−18] For more comprehensive discussions on transformers in language and vision, we direct readers to existing review papers.[13]

2.3. Machine Learning on Chemical Structures. Compared to images and texts, chemical data exhibits greater structural complexity, posing significant challenges for ML applications. For that reason, any ML pipeline must start with identifying an appropriate representation, a decision influenced by the type of structure, the downstream task, and the chosen learning model.

For organic molecules, early efforts predominantly relied on fixed hand-crafted representations, such as descriptors and fingerprints. Descriptors are numerical or quantitative representations of various physicochemical and structural properties of the molecules. Existing descriptors can be broadly categorized into constitutional descriptors, topological and geometrical descriptors, and physicochemical descriptors. On the other hand, fingerprints are binary vectors indicating the presence or absence of predefined substructures. Fragment descriptors are similar to fingerprints as they record occurences of substructures in molecules; however, they can be nonbinary. While these vector representations are straightforward and well-suited for traditional shallow learning models, they are ineffective at capturing structural information. Consequently, there has been a growing interest in adopting string-based molecular representations like SMILES and InChI, inspired by the success of language learning models. Such string-based representations naturally lend themselves to describing biological structures like proteins and peptides, which are amino acid sequences. However, for molecular structures, the transition to string-based representations requires the incorporation of additional syntax to convert structural connectivity into text, adding complexity to the learning process. As a result, there has been a push for graph-based representation of molecules, coupled with modern graph-based learning algorithms, such as GNNs. Graphs also provide means to represent higher order protein structures and lattice patterns in material science.[19,20]

The rising popularity of transformers has led to a surge in research efforts aimed at adapting these models for learning with chemical structures, mostly with string-based and graph-based representations (Figure 2). The main objective of this review is to provide a comprehensive overview of these adaptations. We begin by exploring protein sequences, which resemble text data and are the most straightforward to adapt to text-based transformers.

3. PROTEIN SEQUENCES AND GENOMICS SEQUENCES

In this section, we review the adaptation of transformers to protein and genomics data. The representations of these data types can be easily expressed in sequential forms, making the application of transformers, developed for processing texts, more straightforward.

3.1. Protein Sequences.

The structure of protein sequences can be likened to a complex language, where the arrangement of amino acids dictates various levels of structural folding and an array of biological functions and properties. Much like the protein space itself, these characteristics span a wide spectrum, encompassing local properties to global attributes and intrastructural states to interstructural interactions. However, determining these properties can be cumbersome. Laboratory experiments are time-consuming, and computational simulations of protein interactions may extend over hours or even days. Consequently, the prediction of protein properties and interactions from sequence information is a critical challenge that has garnered substantial interest from ML practitioners.

Given the specific challenges, transformers have proven highly adept at handling protein sequences. First and foremost, the prediction of numerous protein properties requires understanding the dependencies and interactions between components of sequences across multiple scales. The global attention mechanism, an integral feature of transformers, is tailored precisely for capturing this intricate information. Second, protein sequences closely resemble text data, an area where transformers have demonstrated remarkable success. By conceptualizing each amino acid as a word, these sequences form sentences, while higher-order protein structures can be likened to paragraphs. Lastly, the breakthroughs achieved by transformer-based models in text analysis are a result of innovative pretraining on vast text corpora. Notably, there are now open databases housing tens of millions of protein sequences, serving as invaluable data sources that allow adapting pretraining techniques that have revolutionized NLP to the realm of protein research, giving rise to protein language models (PLMs).

Exploiting the growing availability of enormous protein databases, many existing works pretrain large transformer-based language models to capture the biological syntax of amino acid sequences. TAPE and ESM-1b are among the first models to utilize transformers in learning PLMs.[21,22] TAPE pretrains the original transformer on Pfam,[23] a database containing thirty-one million proteins, using text-based pretraining tasks such as next-token prediction and masked-token prediction. To encode protein specificities, TAPE further supervised pretrains the model on contact prediction and remote homology prediction. ESM-1b extends training to 250 million protein sequences and conducts evaluations on a variety of prediction tasks. Given large-scale data, these papers show the superiority of transformers over earlier sequential learning models such as RNN and LSTM. In PRoBERTa, the authors optimized the RoBERTa model for proteins by pretraining using the masked-token prediction task with byte-pair encoding.[24] Another work modifies the pretraining of RoBERTa and Longformer[25] by injecting binding protein pairs into the training set, improving downstream performance on binding prediction tasks.[26] ProteinBERT enriches the training of BERT by adding a novel gene ontology prediction task on top of the bidirectional language modeling.[27] ProtTrans conducts large-scale pretraining of transformers on proteins via a combined data set of 393 billion structures.[28] The authors experimented with a variety of transformer architectures from NLP, producing ProtBERT, ProtAlbert, ProtT5, ProtElectra, and ProtXLNet.[9,29−31] They found that the embeddings produced by these pretrained PLMs serve as competitive initial network weights for smaller predictive models even without training. ProtDistillBERT applies a distillation technique to reduce the size of ProtBERT by half while maintaining most of its performance.[32] Interestingly, MSA-transformer introduces a protein-specific transformer architecture that takes in a set of proteins, represented as a multiple sequence alignment, and performs novel row and column attentions on the alignment.[33] Overall, the introduced protein foundational models based on transformers are important artifacts that can be transferred to a wide range of ML use cases.

Following the success of PLMs, transformers have emerged as a powerful means to tackle many ML problems on proteins. CollagenTransformer uses transformer-based models to predict the thermal stability of collagen triple helices.[34] With only hundreds of downstream data points, it would be infeasible to train transformers directly from scratch. The authors exploited the pretrained ProtBERT model for the task, outperforming smaller transformer models trained directly on collagen data. Another work deploys PLMs to zero-shot prediction of the effects of mutation on the functionality of proteins.[35] The authors introduced ESM-1v, an extension of ESM-1b, and compared it with other transformer-based PLMs such as TAPE and ProtBERT. On the basis of the BERT architecture, MutFormer is developed and pretrained on pairs of reference and mutated proteins.[36] The model is finetuned for the prediction of missense mutation, achieving competitive performances. For drug-binding affinity prediction, ELECTRA-DTA pretrains an ELECTRA-based model to capture contextual information on protein sequences and further stacked a CNN block to capture geometrical features from the learned representation.[30] The authors applied ELECTRA-DTA to drug repurposing and target selection for COVID-19, illustrating the capability of the model and protein transformers in general to tackle urgent emerging problems.

Arguably one of the most challenging yet crucial endeavors in chem- and bioinformatics is protein structure prediction (PSP). PSP is a long-standing problem, stemming from the interdependence between structure and functionality. The environment and the internal interactions of the 20 amino acids constituting protein sequences determine the structure of proteins. In their working environment, most proteins reliably fold into the low-energy conformation that allow them to perform their functions. Predicting the stable conformations is not only essential in understanding the characteristics of existing proteins but also in designing new proteins. As a result, performances on PSP is the standard in evaluating many PLMs.[21,22,28] Spot-Contact-Single adapts the PLM ESM-1b to predict contact maps of protein sequences.[37] The method notably outperforms previous models on sequences in which homologous information is limited. RGN2[38] and trRosettaX-Single[39] are recent PSP models on single-sequence inputs that outperform AlphaFold2[40] and RoseTTAFold,[41] ground-breaking graph-based PSP models that we will discuss in section 7.2, on orphan proteins and human designed proteins. In the RGN2 framework, the authors pretrained another variation of PLM called AminoBert with novel self-supervised tasks and sequence representation.[38] trRosettaX-Single is developed upon ESM-1b and is 2 times faster than AlphaFold2.[39]

3.2. Genomics Sequences.

Genomics sequences are another major type of biological encoding. Many discussions regarding learning on protein sequences can be applied to DNA or RNA sequences as they share significant similarities. Both can be considered languages in which sequences are constructed via constitutional units, of which the ordering determines the biological semantics and functions. However, compared to learning on protein sequences, learning transformer-based foundational models on genomics data is still an under-explored area with a potential for wide applications. As a language, genomics sequences convey rich semantic information, including those closely related to natural language such as polysemy and distant semantic relationships.[42]

The tokenization of genomics sequences is different from that of protein sequences as each token is a nucleotid base concatenated with a number of the trailing bases. Such token is called a *k*-mer if there are *k* trailing bases. Since the number of unique bases in a DNA is extremely limited, such representation helps diversify the tokens with contextual information. DNABert is the first language model on DNA sequences with the *k*-mer representations.[42] The authors finetuned the model on multiple downstream scenarios and obtained better performances compared to those of other architectures such as CNN and LSTM. Enformer further demonstrates the superiority of transfromer architectures by training a model with a perception field of up to 100 kilobases, compared to only 20 kilobases in previous CNN models.[43] MoDNA extends the self-supervised learning beyond *k*-mer to include repetitive motifs, improving the learned embeddings with biologically inspired patterns.[44] Most recently, Nucleotide Transformer and DNABert-2 further introduce foundational models on genomics data.[45,46] DNABert-2, in particular, replaces the *k*-mer representation with byte-pair tokenization, improving the efficiency and efficacy of learning embeddings.[46]

Even though transformers have been remarkable in addressing challenges related to learning from sequential protein or genomics data, it is important to note that not all types of substances can be approached similarly. While proteins or DNAs can be conveniently represented as sequences that neatly align with well-established NLP frameworks, the same does not hold for other complex structures like molecules or material lattices and adaptation in terms of the representation is often required. In the following sections, we discuss the current progress of applying transformers on inherently nonsequential structures.

4. MOLECULAR STRINGS

Molecules constitute a major portion of the chemical space and are building blocks for larger chemical and biological structures. As a type of data, molecules are rich in information, not only from the chemical specificity of their components, such as atoms, bonds, and functional groups, but also from the connectivity and interaction between these components. As molecular structures significantly influence chemical and physical properties, preserving structural information in molecule representations is essential for extracting predictive learning patterns. Fortunately, this challenge has been a topic of interest among researchers for quite some time.[47] However, its motivation was not rooted in machine learning but stemmed from the necessity to effectively document chemical knowledge in written text. Because of the vastness of the chemical space, it is infeasible to assign unique names to each substance. Instead, given a system of rules and syntax, each molecule can be represented as an identifying string, which in turn, can be utilized to reconstruct the molecule itself. The possibility of reconstruction means that these strings implicitly encapsulate structural information. When these representations are coupled with text-based transformer models borrowed from NLP, the prospects for advancing molecular learning become promising.

4.1. String-Based Molecular Representations.

Numerous string-based representations have been devised for molecules, each comprising distinct sets of rules to transform molecular structures into one-dimensional strings. For instance, the IUPAC nomenclature defines a method to name organic compounds by building upon a predefined vocabulary of common substructures and functional groups.[48] Naming becomes increasingly intricate as molecules expand in size, which requires more compact string representations, such as the international chemical identifier (InCHI) and the simplified molecular-input line-entry system (SMILES).[49,50] Among these, SMILES is the most frequently used in text-based ML on molecules, thanks to its simple and compact expression. Since its inception in 1988, SMILES has been the standard string-based molecular representation in computational chemistry. Its ubiquity led to the development of other variations, such as BigSMILES and DeepSMILES.[51,52]

While widely adopted, SMILES is not without its limitations. Particularly, SMILES representation is not unique for each molecule. Quite commonly, multiple SMILES strings, even considerably different ones, correspond to the same molecule. Conversely, not every SMILES string translates to a valid molecule. A large portion of the SMILES space consists of such invalid SMILES.[53] Additionally, structural differences between molecules do not translate to equivalent string edit distances in their respective SMILES representations. In other words, a minor change to a molecule may lead to a drastically different SMILES string. These phenomena pose a notable challenge to learning with SMILES as QSAR relies on using structural similarity to infer characteristic similarity. Recently, self-referencing embedded strings (SELFIES) have emerged as an innovation that mitigates some of the existing issues with SMILES.[53] Every SELFIES corresponds to a valid molecule, ensuring the robustness of the string representation space.

4.2. Transformers on Molecular Strings.

String-based representations of molecules recorded in existing massive accumulation of chemical documentation are valuable resources for ML, especially text-based transformer models that are extremely data-hungry. SMILES Transformer uses an encoder-decoder pipeline to pretrain a sequence-to-sequence SMILES language model.[54] The encoder maps an input SMILES string into a latent encoding, and the decoder reconstructs the original string from this latent encoding. This encoding, which the authors called ST Fingerprint, is utilized as the input feature for shallow predictors such as support vector machines. SMILES-BERT adopts the masked language model from BERT for pretraining on SMILES, treating each SMILES string as a sentence and each atom symbol as a word token.[55] The entire pretrained model is finetuned for downstream tasks. Compared to earlier featurizations based on molecular fingerprints, both ST fingerprint and SMILES-BERT exhibit a marked improvement in predictive quality across a range of chemical tasks. Such results motivated later works that extend pretraining to more variations of transformers, domain-specific pretext tasks, and larger-scale training data.

MolBERT pretrains BERT on SMILES with masked token prediction and auxiliary chemistry-relevant predictive tasks.[56] These tasks include SMILES equivalent prediction, which checks whether a pair of SMILES strings encode the same molecule, and molecule descriptor prediction, which estimates various physical properties. Instead of predicting SMILES equivalence directly, knowledge-based BERT follows a contrastive setting in which embeddings of similar examples are enforced to be closer in the latent space while those of different examples are pushed apart.[57] In this context, SMILES permutations of the same molecule form similar, or positive, pairs of examples. The authors also performed atom feature prediction from atom token embeddings and global property prediction from SMILES sequence embeddings. Interestingly, PolyBERT pretrains a language model on polymers represented as SMILES strings.[58] TransPolymer takes a step further and defines a novel string-based representation that contains the polymer SMILES, the copolymer SMILES, and other condition parameters such as ratio and temperature.[59]

Other studies investigate the influence of large-scale pretraining on string-based molecular learning models. ChemBERTa adopts the RoBERTa pipeline for pretraining 10 million molecular strings, represented as either SMILES or SELFIES.[60] Even though ChemBERTa was outperformed by strong existing baselines, the authors demonstrated improvement in performance with more pretraining data, strengthening the positive effect of large-scale pretraining. Recently, ChemBERTa-2, pretrained on 77 million molecular strings via a more optimized pipeline, has demonstrated competitive performance versus contemporary state-of-the-art methods on MoleculeNet benchmarks.[61] To enforce understanding of molecular structural syntax, Chemformer pretrains BART on 100 million molecules via an autoregressive SMILES reconstruction task.[62] Other works push the scale of the pretraining to over a billion molecular strings. X-MOL generatively pretrains various language models, BERT, RoBERTa, XLNet, T5, and ERNIE on a colossal data set of 1.1 billion molecules.[63] Similarly, MolFormer also pretrains on 1.1 billion SMILES and introduces a masked language prediction task coupled with rotary positional encoding and a novel linear attention.[64] These large-scale pretrained models outperform strong baselines on various downstream predictive tasks, confirming the effectiveness of transformer architectures in learning text-based chemical syntax.

Besides increasing the scale of the pretraining, other related works employ innovative mechanisms to elevate the performance of transformer models. MTL-BERT uses SMILES enumeration to tackle the data scarcity problem, especially during downstream multitask finetuning in which the authors comprehensively evaluated their models on 60 prediction tasks.[65] MolRoPE-BERT incorporates rotary positional embeddings into pretraining BERT.[66] Interestingly, Mol-BERT and FP-BERT utilize the masked token prediction from BERT to learn rich fingerprint embeddings. Even though they do not pretrain directly on molecular strings, these methods are inspired by Mol2Vec, which, in turn, was inspired by Word2Vec, a famous deep self-supervised pretraining method from NLP.[67−69]

5. MOLECULAR GRAPHS

Though interesting, string-based molecular representation learning has several shortcomings. Competitive performance relies on immensely powerful transformer-based architectures, enormous unlabeled data for pretraining, and huge computing resources. The additional syntax in constructing molecular strings complicates learning, and string-based representations are not directly topologically aware. Graphs, on the other hand, can represent molecular connectivity explicitly. Incorporating chemical properties into node and edge featurizations is more straightforward in graphs compared to strings and fingerprints. As a result, graphs can be considered the most natural way to represent molecular structures. Learning on molecular graphs is an active research area in chemical ML. A large portion of state-of-the-art methods are graph-based models, such as GNNs and graph transformers.[70,71] In this section, we review existing innovations in developing transformer-based learning models for chemical structures formulated as graphs.

5.1. Graph Transformers and Positional Encodings.

In an ideal scenario, it can be expected that combining the most natural molecular representations, graphs, with the most expressive learning models, transformers, would lead to a major leap in performance. However, adapting transformer architectures to graphs is a more complicated process than that of images or texts. If we consider texts as line graphs and images as grid graphs, then graphs are their generalization. Unlike line graphs and grid graphs, general graphs, including molecular graphs, do not always have uniform graph connectivity and pivoting points based on which one can define positional information. As a result, an essential problem in developing transformer-based models on graphs is effectively defining and learning graph positional encodings. In that regard, we can categorize existing techniques according to how they approach handling this problem. Specifically, we break down existing graph transformers on molecules into groups that use structure-based positional encoding, distance-based positional encoding, and positional encoding via GNNs. Despite our nomenclature, both structural-based and distance-based positional encodings capture various degrees of graph structure. While distance-based encoding focuses more on the immediate relative structural context and distance between two particular nodes, structure-

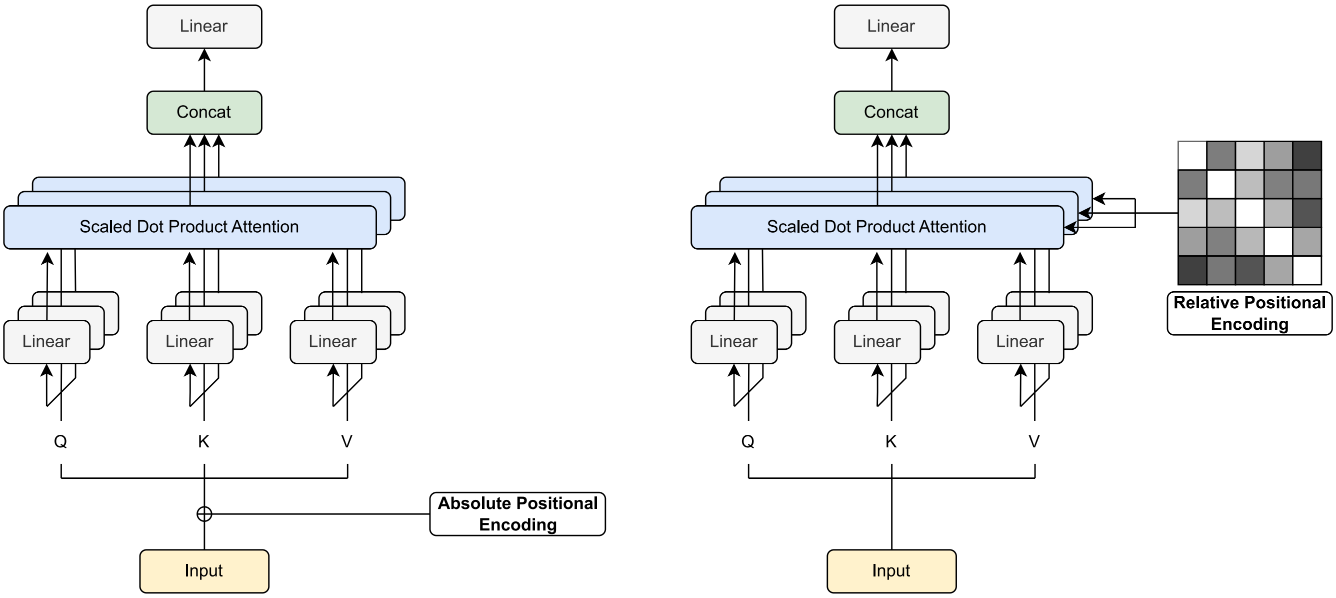

Linear
Concat
Scaled Dot Product Attention
Linear
Linear
Linear
Q
K
V
Absolute Positional Encoding
Input
Linear
Concat
Scaled Dot Product Attention
Relative Positional Encoding
Linear
Linear
Linear
Q
K
V
Input

Figure 3. (left) Absolute positional encoding and (right) relative positional encoding. Absolute structural information can be added directly to the input node and edge embeddings. On the other hand, a common way to handle relative positional encodings is to aggregate the embedded pairwise distances to the pairwise attentions.

based encoding encompasses the overall graph structure and the position of nodes within the context of the whole graph. Interestingly, structure-based encoding and distance-based encoding are analogous to absolute positional encoding and relative positional encoding, two common encoding paradigms for transformers on text, respectively. Figure 3 illustrates some common ways of performing absolute positional encodings and relative positional encodings.

5.2. Graph Transformers with Strutural Positional Encoding. Structure-based graph transformers enrich the node features and, optionally, the edge features with structural information before passing them into the network. These enrichments enhance node embeddings with graph structural information and positional information with respect to the whole graph. Essentially, nodes that are closer in the graph have similar positional encodings and vice versa. For example, Graph-BERT orders nodes via the PageRank algorithm, which obtains the ordering scores with a closed-form formula involving the adjacency matrix.[72,73] Positional encodings are obtained from this ordering of nodes, much like in a vanilla transformer. Such encodings provide a canonical way to process a sequence of nodes; however, they fail to encode the graph structure. For this reason, the authors enhanced node embeddings with graph role embedding based on the Weisfeiler-Lehman (WL) algorithm, a standard procedure for graph isomorphism testing.[74]

Other works use the eigenvectors of the graph Laplacian matrix, which was shown to empirically perform better than WL-encoding,[75] to capture structural and positional information.[75−79] GT adds the Laplacian embeddings to the node features at the input layer.[75] SAN further processes the Laplacian eigenvectors ordered by their corresponding eigenvalues via a series of linear projections and transformers before adding them to the node features.[76] TokenGT employs a simple yet effective strategy that uses various concatenation of Laplacian embeddings to node and edge features.[78] Graph ViT/MLP Mixer sequentially processes graph at node levels and subgraph levels.[77] Laplacian embeddings are added to the nodes before being processed by a GNN to obtain subgraph embeddings. After that, the subgraph embeddings are enhanced with relative position embeddings obtained from the subgraph adjacency matrix weighted by the number of edges connecting pairs of subgraphs. The subgraphs are passed into a series of transformer layers and a pooling layer to obtain the final graph embedding. NAGFormer represents each node as a series of k-hop neighborhood embeddings with various radii.[79] The Laplacian eigenvectors are concatenated to the node features before neighborhood extraction and projection. Besides the eigenvectors of the Laplacian matrix, other structure-based encoding approaches have been proposed. Graph Multiset Transformer learns to cluster nodes via the attention mechanism and uses the cluster identity as the positional encoding.[80] Exphormer introduces attention via edges of random expander graphs generated from the input graphs.[81]

A major problem of many structure-based positional encodings, especially those relying on spectral decompositions of Laplacian matrices, is the limited transferability of these encodings between different graphs. Such a drawback is critical for inductive learning scenarios such as property prediction on molecular graphs. To tackle this problem, SAT uses random walk positional encoding, which is more transferable than Laplacian-based encoding.[82,83] GraphiT improves transferability by exploring graph kernels defined through regularizations applied to the Laplacian spectrum and using the Gram matrix of these kernels to directly bias the attention scores.[84] While a majority of structural encoding methods define absolute positional encoding, such a matrix acts as relative positional encoding. Biasing the attention matrix with relative positional encoding is a commonly employed strategy by distance-based methods, which we review in the next section.

5.3. Graph Transformers with Relative Distance Encoding. Even though relative distance encodings may not capture the overall graph structure as effectively as numerous spectral methods, their strengths reside in their simplicity and transferability. Concretely, Laplacian eigenvectors are graph-specific, and vectors from different graphs are not comparable. On the other hand, relative node-to-node distance metrics such as shortest path distance are consistent across various graphs. This property is particularly suitable for inductive learning, which may explain the recent superior performances of graph transformers with relative distance encodings on molecular prediction tasks. Additionally, the limited encoding of overall structural information may not be detrimental to these tasks since molecular graphs are quite simple. The node degrees are low due to valency, and cycles often come from relatively large aromatic rings.

PATGN encodes both the distance and the collection of edges within the shortest path between pairs of nodes.[85] The encodings are transformed into a matrix of shortest-path features. Each path feature is concatenated to the features of the corresponding nodes to calculate the pairwise attention score. MAT biases the attention matrix with 3D interatomic distances.[86] R-MAT extends MAT by applying the radial basis function[87] on the 3D distance and encoding the shortest path distance between nodes.[88] Graphormer trains a graph transformer on PCQM4M,[89] a large predictive data set with 3.8 M molecules, and achieves a better performance over GNNs.[90] While Graphormer also relies on shortest path encoding, unlike PATGN, the method encodes both the bonds and the bond order within each path. GRPE also encodes the shortest path; however, the model learns shortest path embeddings and uses this information to update the query and key features before calculating the attention matrix.[91] While Graphormer uses the same relative positional bias for all layers, EGT updates the pairwise relative positional embeddings after each layer, starting with the embedding of the adjacency matrix at the input layer.[92] GPTrans also updates the edge embeddings, employing node-to-node, edge-to-node, and node-to-edge attention.[93] Interestingly, MolFormer defines motif nodes that connect to corresponding atom nodes that appear in the motif.[94] The authors define attention between motif nodes and the connected atom nodes and use 3D absolute distance for relative positional encoding.

Apparently, the ever-growing interest in applying transformers to chemical graphical data has resulted in a wide variety of methods. Traversing the graph transformer landscape has become more challenging due to the sheer diversity of positional encodings, be it absolute or relative, structure-based or distance-based. GPS introduces a modular and unified framework for experimenting with various types of positional encodings.[95] The authors also proposed a novel network layer that is a hybrid combination of the message-passing layers of GNNs and transformer layers. Such a combination is beneficial because it can leverage both the local structural encoding power of GNNs and the long-range reference capability of transformers. We review the methods that follow this line of ideas in the next subsection.

5.4. Combining Graph Neural Networks and Graph Transformers. Methods reviewed in this section do not necessarily require positional or structural encoding because such information can be captured by graph neural networks (GNNs). After being processed by a k-layer GNN, a node embedding would encode the structural information on its k-hop neighborhood. Note that GNNs often suffer from oversmoothing, a phenomenon associated with deeper GNNs. When a GNN has too many layers, the final node embeddings get saturated with information that embeddings of different nodes appear indistinguishable.[96,97] Such oversmoothed embeddings may limit the quality of structural encoding. However, given enough structural diversity within a molecular graph, a moderately deep GNN can be quite useful at encoding graph positional information. When positional encoding is not explicitly defined, the network can learn to extract useful structural information without relying on hard-coded inductive bias from practitioners. Additionally, relative inductive bias may not be as useful for long-range dependencies as it is for learning short-range patterns.[98]

In place of positional encoding, several methods rely entirely on GNNs for capturing positional and structural information. In GraphTrans, a transformer module follows a GNN module.[99] The output of the GNN module is embeddings that encode the neighborhood and structural information surrounding each node. The concatenation of these embeddings and the original node features form the inputs to the transformer module. Essentially, the embeddings produced by the GNN module replace positional encodings. Similarly, GROVER uses a dynamic message-passing network that captures neighborhoods surrounding a node with randomly sampled size before passing the output embeddings into a transformer.[100]

Instead of stacking two distinct blocks of GNN and transformer, other methods merge GNN layers and transformer layers to form hybrid layers, leveraging the advantages of both types. For example, in UGFormer, a network layer consists of a transformer layer followed by a GNN layer.[101] Graph connectivity information is only included during the pass through the GNN layer. In contrast, GPS adopts an inverse approach in which a transformer layer follows a GNN layer. Similar to GraphTrans and GROVER, the GNN layers encode structural information. However, interweaving the GNN layers in between the transformer layers minimizes the chance of oversmoothing. The authors of GPS++ further capitalized on this idea and achieved first place on the large-scale PCQM4M benchmark for learning on molecular graphs.[102]

6. APPLICATIONS OF TRANSFORMERS IN CHEMINFORMATICS

Our review has focused so far on the transformer architecture and its adaptation in processing various chemical representations. For the purpose of comprehensiveness, in this section, we shift the focus to the application side of cheminformatics. We review several key areas in which machine learning has played a pivotal role with the focus on the adaptation of transformers within these domains.

6.1. Property Prediction. An important part of cheminformatics is the analysis of chemical activities and properties based on structural patterns, as per the development of quantitative structure−activity relationship (QSAR) and quantitative structure−property relationship (QSPR) models.[103] These models assume a correlation between structural similarity and characteristics similarity among chemical compounds. As a result, machine learning has naturally been useful analytical tools for the development of such models. Recently, the adaptation of expressive learning architectures such as transformers has become prominent for predicting a wide array of chemical properties and activities.

The availability of benchmark databases such as MoleculeNet[104] and Open Graph Benchmark (OGB)[89] has played a pivotal role in advancing the development of property predictive models, especially transformer-based models spanning various modalities of molecular representations.[90,105] Notably, OGB hosts PCQM4M and PCQM4Mv2, some of the largest labeled molecular data sets containing more than 3.8 million compounds. At the moment, graph-based transformer models such as Graphormer,[90] EGT,[92] and Uni-Mol,[106] achieve the most competitive performance on these large scale challenges. In particular, Graphormer won the 2021 KDD Cup on PCQM4M and EGT currently tops the leaderboard on PCQM4Mv2.[107,108] Most recently, long-range graph benchmark[109] introduces several data sets on macro-molecular structures, which require the understanding of long-range dependencies among graph components. Benchmarking results on these data sets showed that transformer models outperform other learning architectures, confirming the advantage of transformers and the attention mechanism in learning complex structural interactions.

Beyond benchmark data sets, transformer-based models have been successfully applied to tackle various QSAR/QSPR problems, such as predicting solubility[110] and toxicity.[111] For example, Riedl et al.[112] finetuned a pretrained language model on SMILES to predict the fraction unbound in human plasma, an important parameter in ADMET. With a similar SMILES-based model, transformer-CNN[110] predicts AMES mutagenicity and aqueous solubility. Cremer et al.[111] applied TorchMD-NET, an equivariant graph transformer, on predicting drug toxicity.

One important QSAR usage of learning models is the prediction of drug-target affinity prediction (DTA). DTA is a crucial step in computer-aided drug discovery since it helps identifying appropriate drug targets, usually physiological proteins, and the design of molecules capable of modulating their activity. Transformers are well-suited for this task as their attention mechanism not only allow effective capturing of drug-target interactions and long-range structural information, but also result in more interpretability. For that reason, there has been increasing adaptations of transformers for DTA. For example, DTITR[113] employs transformer encoders to process SMILES strings and protein sequences, followed by a cross-attention module to produce molecule-protein interaction embeddings. Another work by Kang et al.[114] adapts additional BERT-like pretrainings on the transformer encoders separately for either SMILES strings or protein sequences. TEFDTA[115] extends the analysis to bonded (valence) interactions. The authors represented molecules as MACCS fingerprints. A transformer encoder processes these MACCS vectors while a 1D convolutional neural network processes the protein sequences. TAG-DTA[116] is another transformer-based model on SMILES strings and protein sequences in which an auxiliary binding pocket prediction task is learned in parallel in order to condition and guide the main DTA task. Other works look into interpretability. MolTrans[117] is a popular model that works on SMILES and protein subsequences. The authors computationally mined common SMILES and protein subsequences from large databases and decomposed input SMILES or proteins based on the mined vocabulary of subsequences. These sequences of substructures are processed by transformer encoders and pairwise interaction scores between each pair of protein substructure, and a molecular substructure is obtained via dot products of the embeddings. The resulting interaction map casts light on the engagement between drugs and targets. HoTS[118] pretrains the encoders on complexes of protein—ligand interaction. Evaluated interpretability of downstream DTA predictions are obtained from the attention scores. To better capture structural information, which is essential for DTA, several recent methods such as GTAMP-DTA,[119] AttentionMGT-DTA,[120] and TransVAE-DTA[121] turn to graph representations, representing molecules and proteins as molecular graph and protein pocket graphs and processing them via graph transformers. These models obtain strong performances on DTA predictions.

6.2. Structure Generation. Generative models are at the forefront of AI research, encompassing diverse domains including chemical structures. Out of the 10^{23} to 10^{60} possible druglike molecules,[122,123] only a minuscule portion has been synthesized,[124] explaining the strong need for the discovery and generation of new drug candidates. As a result, generative models have been developed in parallel with predictive models as a major part of machine learning in cheminformatics. Coupling generative and predictive methods allows the discovery of chemically valid molecules with one or multiple desired properties. Existing generative models span a wide variety of paradigms, including variation encoders, generative adversarial networks, flow-based models, transformer-based models, and diffusion.[125] Given the prevalence of transformers in the generative AI landscape, it is compelling to explore their potential in discovering novel chemical structures.

Text-based representations such as SMILES are common choices for chemical generative models because of the natural transferability of successful transformer-based language models. For instance, GMTransformer[126] autoregressively generates molecular strings via pretraining a transformer network with a novel blank filling language model. The authors experimented with various string-based representations, including SMILES, SELFIES, and DeepSMILES. Regarding generation of molecules with target properties, MCMG[127] trains a conditional transformer with reinforcement learning and knowledge distillation. The model effectively generates molecules with multiple desired properties. In a different flavor, CMGN[128] autoregressively reconstruct molecules from fragments, conditioning on certain properties. RegressionTransformer[129] is another conditional transformer model that not only generates molecules with high-quality continuous attributes but also outperforms multiple baselines on regression tasks. Besides string-based representations, MolFormer[94] explores the generation on molecular graphs.

Several methods utilize the ability of transformer in interdomain translation to train generative models. Grechishnikova[130] trained a transformer model to translate from amino acid sequences to SMILES, effectively generating molecules for given target proteins. TransAntivirus[131] translates IUPAC nomenclatures into SMILES. This translation includes select-and-replace edit that transform input molecules into ones with desired properties, with the application in discovering antiviral compounds. In general, translation is a powerful mechanism of transformer architectures with, beyond structure generation, many important use cases in cheminformatics, which we review in the following section.

6.3. Chemical Translations. The original transformer was developed for the machine translation task,[4] which converts texts from one language to those with the same meaning in another language. However, this mechanism can be applied to use cases outside of the ordinary language translation. An example is in the question answering task in which the input question is “translated” into the appropriate answer.[132] In general, translation via transformer can be extended to mapping between multimodal data or data distributions. This usage is especially applicable for chemical data in which multiple representations often exist for the same compound. For instance, the transformer model on natural language text can be readily adapted to translating from SMILES to IUPAC[133] or from InCHI to IUPAC.[134] Notably, pretraining a transformer model on the SMILES-to-IUPAC translation task has been found to result in better performance when finetuned on downstream tasks, such as binding affinity prediction.[135] Another interesting application is in optical chemical structure recognition, in which the chemical structure information on compounds from scientific records and publications is converted into machine readable formats. Several works formulate this task as an image-to-text translation with the text being any string-based representation of chemical structures.[136−138] For instance, SwinOCSR[136] uses a Swin Transformer for image to SMILES translation. Other similar models include Image2Smiles[137] and DECIMER.[138] Interestingly, MassGenie[139] predicts chemical structures from mass spectroscopy by translating the mass spectroscopy of molecular fragments into the original molecule.

Analogical to the question answering task, in cheminformatics, translation can be applied to generate “answers” to a given “question”, both represented as chemical structures. For example, to generate druglike molecules that target specific proteins, several works translate input amino acid sequences into SMILES strings of molecule with appropriate binding activities.[127,130] To optimize existing molecules in order to obtain certain desired properties, translation models are trained to make adjustment to input molecules or scaffold. He et al.[140] trained a conditional transformer model that converts an aggregated input of the source molecule SMILES string and the target property into an output SMILES string of a molecule possessing the property. Similarity constraints ensure that the output molecule closely resemble the source molecule. DeepHop[141] uses translation to perform scaffold hopping in which the output molecule is novel in terms of scaffold as compared to the input molecule while maintaining similar bioactivity, which is ensured via enforcing 3D structural similarity. MetaTrans[142] translate an input molecule into its products after going through various metabolism procedures, effectively predicting metabolism outcomes of drugs.

6.4. Chemical Reactions. Chemical reaction analysis is another pivotal domain within cheminformatics where computational methodologies and learning techniques have made significant strides. Both the forward problem, entailing the prediction of reaction outcomes, and the backward problem, involving retrosynthetic analysis, have witnessed substantial advancements aided by modern learning architectures. While earlier approaches rely upon rule-based procedures and reaction templates, of which the extraction from existing literature is a laborous and tricky process, recent learning-based methods explore template-free modeling of chemical reactions. In these template-free models, chemical reaction information is automatically learned from data and stored as embeddings in deep neural networks, resulting in more flexibility and robustness of both the forward and backward analysis. The rise of transformers has catalyzed modeling chemical reactions as a translation problem. Intuitively, this formulation fits the idea of chemical reaction being a transition of a chemical system from the reactants to the products and vice versa.

A prime example of formulating reactions as translations is the Molecule Transformer,[143] a popular translational model for predicting reaction outcomes. The authors modeled the mapping from the reactants to the products as a SMILES-to-SMILES translation. Molecule Transformer outperforms previous baselines by large margins on various experimental settings. Jaume-Santero et al.[144] further investigated the effects of different training parameters, including representations (SMILES or SELFIES), tokenization (atom or byte-pair encoding), pretraining, data augmentations, and predictive tasks, on the performance of Molecule Transformer. Pesciullesi et al.[145] developed finetuning strategies for Molecule Transformer and evaluated the model on predicting regio-stereo-selective reactions on carbohydrates. Andronov et al.[146] used the translation formulation to predict reaction reagents. In this case, the input is the whole reaction string instead of just the reactants. A few recent works incorporate both the graph and string representations into the translation process.[147,148] Graph2-SMILES[147] translates the molecular graphs of reactants into the SMILES strings of products. The authors replaced the transformer encoder with a graph neural network (GNN). Instead of GNN, SeqAGraph[148] develops a novel transformer-based graph encoder that assigns node ordering on the graph atoms based on the corresponding SMILES string. The whole model is trained for both forward and backward (retrosynthesis) prediction problems.

Similarly, translational transformers are also popular for the backward retrosynthetic analysis of chemical compounds. For instance, Karpov et al.[149] trained a translational transformer model that converts the SMILES string of a compound into another SMILES string of its constitutional reactants. Schwaller et al.[150] coupled the single-step backward prediction with hyper graph representations of chemical reactions to discover synthetic pathways. SCROP[151] couples the translational transformer with a neural network-based syntax corrector, significantly reducing the number of chemically invalid candidate precursors. Tetko et al.[152] employed data augmentation on the input SMILES strings, noticeably improving the performance of their model across multiple metrics. RetroPrime[153] disects a retrosynthesis step into 2 steps: generating sythons and adding leaving groups. The authors developed 2 separate translational models for these steps. Recent methods on retrosynthesis also incorporate molecular graph information on top of the SMILES representation. For example, RetroFormer[154] has separate attention heads for the global attentions on SMILES strings and the local attentions that encode neighborhood connectivities on molecular graphs. The aggregated embeddings of these attentions serves as the input to the decoder, which outputs SMILES strings. G2GT[155] takes a step further and formulate retrosynthesis as a graph-to-graph translation problem in which they utilize the Graphormer architecture[90] as the encoder and the decoder. The method outperforms other template-free transformer baselines on top-1 accuracy.

Transformer-based models also have other interesting applications within the chemical reaction domain. For example, Wang et al.[156] introduced the reaction generation task and trained the Transformer-XL to generate novel reactions within the same reaction class. Thousands of generated reactions were assessed and confirmed by chemists, and the whole process from training to confirmation took only 15 days. GraphormerMapper[157] trains a atom-mapping model based on the BERT architecture with the encoder replaced by a Graphormer encoder.[90] The model effectively maps corresponding atoms between the reactants and the resulting molecule from chemical reactions, outperforming the state-of-the-art atom mapping algorithm. In a different flavor, Schwaller et al.[158] used the attention weights of a pretrained transformer model to capture the atom mapping. Another work by Schwaller et al.[159] predicts and analyzes reaction classes. They trained a SMILES-to-SMILES translational transformer model in an unsupervised manner and a BERT-based encoder-only model in an supervised manner on reaction classification. Both models achieve remarkable classification performance. Additionally, the authors found that the embeddings learned by these models, which they termed reaction fingerprints, effectively capture the clustering of various chemical reaction classes with granular details and differences.

7. FUTURE DIRECTIONS

We conclude the review on the application of transformer-based architectures in cheminformatics by discussing interesting research directions. These directions encompass novel applications and developments of models capable of encoding domain-specific, geometric, and multimodal information. Such approaches potentially achieve improved performances on existing tasks and solving more complex ones.

7.1. Transformers for Molecular Dynamics Simulations. We have witnessed rapid progress on the analysis of proteins, especially on crucial tasks such as protein structure and protein-drug affinity predictions. However, as encouraging as they are, these tasks present only interactions under fixed conformations. Existing learning frameworks provide limited capability in reasoning about dynamic chemical/biological configurations. One important use case of such reasoning is in targeting protein misfolding and degenerative diseases.[160] Degenerative diseases such as Alzheimer's and Parkinson's are extremely common, affecting millions of people worldwide, and are directly linked to the misfolding of certain proteins.[161,162] Developing prevention and treatment for these diseases requires understanding the effects of drugs on flexible protein conformations.

Molecular dynamics (MD) simulations can cast light on interactions and changes in protein structures; however, they are costly in terms of time and computation and not suitable for large scale screening of drugs. Machine learning has the potential to automate and speed up parts or the whole process. Moreover, transformer-based models are particularly suitable for simulations since they can effectively capture multiscale interactions between drugs and protein sequences amd handle multimodal data. For example, Wang et al.[163] formulated molecular dynamics simulations as a generative problem with the goal of discovering novel conformation. A transformer encoder-decoder network is trained to predict propagating frames of protein complexes with data obtained from MD simulations. Despite its importance, this direction is still quite unexplored with only a few other works with similar ideas.[164−166] A general approach toward generative models using MD simulation is still much desired.

7.2. Equivariant Graph Transformers. Despite their usual 2D visualization, molecules do not lie on the 2D space as planar graphs. Instead, a large portion of their chemical properties depend on their 3D geometrical alignments and interactions. As a result, more works have investigated the incorporation of 3D geometric features into the learning of molecular graphs. However, such a process is not straightforward. Due to translations and rotations, there can be an intractable number of equivalent 3D atomic coordinates to the same molecule. Translation and rotation form the SE(3) group in the 3D Euclidean space, and as such, geometric learning on molecules requires a certain degree of equivariant or invariant with respect to SE(3) transformations.

Before transformers, several GNNs have experimented with equivariant convolution. Inspired by the use of Clebsch-Gordan coefficients and the spherical harmonics in Tensor Field Networks,[167] SE(3)-Transformer connects SE(3) equivariant and the attention mechanism.[168] In particular, the method achieves roto-translational equivariance within the node embeddings and invariance with respect to the attention weights. TorchMD-Net obtains rotational equivariance by encoding interatomic distances via the radial basis functions.[169] GPS++ encodes interatomic distances with Gaussian kernels.[102] Besides rotation and translation, Equiformer extends equivariance to inversion. The authors replaced the dot product attention with multilayer perceptron attention and nonlinear messages, leading to higher expressive power.[170]

By extending geometric learning to graph transformers, equivariant transformers allow fuller utilization of the long-range referencing via the attention mechanism. Such property is especially beneficial to tasks in which 3D geometric information is important, such as energy, binding affinity, and protein folding prediction. For example, RoseTTAFold, a recent remarkable work on protein folding prediction uses SE(3)-Transformer as the backbone of the pipeline.[41] Similarly, AlphaFold2, a revolutionizing model for protein folding prediction, is backed by an original equivariant graph transformer layer called Evoformer and novel triangle attentions.[40] AlphaFold2 was the top performer on the CASP14 protein folding challenge and is the state-of-the-art model. These facts signify the importance of equivariant transformers as a research direction in geometric learning and ML within the chemical domain.

7.3. Graph Transformers for Infinitely Repetitive Patterns. Graph-based learning on material structures is an interesting direction that has recently captured increasing attention. These graphs consist of local connectivity patterns between atoms within a unit cell and global connectivity patterns among an infinitely repetitive lattice of unit cells. These properties pose a significant challenge in constructing graphical representations for such data. With the availability of larger material data sets such as the Material Project or the Cambridge Structural Database, there is an interest in applying complex learning models to solve learning problems in this data domain.[171,172]

CGCNN is an earlier work that tackles this problem.[173] For any crystal lattice, the authors constructed a compact graph in which nodes are atoms in the unit cell and edges represent both intercell and intracell connectivities. Since lattice data does not often contain explicit bonding information, connectivities are determined via a distance threshold. Matformer further capitalizes on this graph construction and rigorously proves its periodic invariance.[174] They enriched the attention mechanism with edge distances featurized via radial basis functions, achieving roto-translational and reflective invariance. MOFNet works on metal−organic frameworks (MOF) and only applies transformer-based embedding on the local graph representation of the unit cell.[175] For capturing global lattice patterns, MOFNet incorporates features such as crystal density, porous volume, gravimetric surface area, etc. Xtal2DoS learns to predict the density of states of crystals.[176] The authors used GAT, an attention-based GNN,[177] to embed both local and global patterns, then employed a transformer-based model to decode the embeddings to density sequences. To further advance performance, other methods attempt pretraining MOF on large data sets. In a supervised manner, MOFTransformer pretrains a standard transformer model to predict the topology, the void fraction, the metal cluster, and the organic linker of more than 1 million hypothetical MOF.[178] The pretrained model captures both local and global features of MOF lattices and obtains competitive results on downstream tasks such as gas absorption prediction. Instead of supervised pretraining, MOFormer pretrains in a self-supervised manner.[179] The authors used GCGNN to learn node embeddings that capture the graph connectivity of unit cells and a text-based transformer that processes the string-based representations of MOFs. The framework then contrastively enforces the correlation between both embeddings.

7.4. Multimodal Pretraining with Large Chemical Knowledge Bases. Despite the availability of large chemical knowledge bases,[124] most existing methods only extract patterns from molecular structures, neglecting an abundance of corresponding chemical information stored as text descriptions. The main challenge of exploiting these chemical corpora is the multimodality of the data, i.e, graphs and SMILES strings versus natural language texts. Many recent works take on the challenge by leveraging the powerful processing ability of transformer-based language models.

Joint learning of SMILES strings and texts is an apparent direction because both modalities can share the same learning architecture and utilize NLP learning techniques. For example, KV-PLM[180] learn to process both SMILES strings and text descriptions using the masked token prediction task and the BERT model.[8] The model processes molecule text descriptions with masked tokens being either byte-pair encoded SMILES strings placed next to the substance name in the text or other randomly selected words. The pretrained model then performs tasks such as property or reaction prediction using SMILES strings as inputs. Similarly, MolT5[181] trains a single medium-size T5 model for both text descriptions and SMILES strings using the replace-disrupted-span objective. Downstream tasks include molecule captioning which translates SMILES into text and text-to-molecule generation which outputs SMILES according to textual descriptions. MolXPT[182] replaces substance names with SMILES strings and pretrains on the combined corpus of texts, texts with wrapped SMILES, and SMILES strings. The downstream predictions on molecular properties are obtained via prompting. MolReGPT[183] employs LLM to perform molecule and text translation.

Other works expand the multimodal learning to other molecular representations. Text2Mol[184] uses molecular graphs with Mol2Vec[68] embeddings as node features. The embeddings of these graphs produced by a GNN are contrastively pretrained against the text embeddings produced by SciBERT,[185] a language model trained on scientific texts. Following a similar approach contrasting texts against molecular graphs, MoMu[186] extensively evaluates the pretrained model on a variety of challenging downstream tasks, such as cross-modal retrieval, molecule captioning, zero-shot text-to-molecule generation, and property prediction. Interestingly, CLAMP[187] finds that traditional fingerprints work better than SMILES or graphs in representing molecules for contrastively pretraining against texts. Finally, GIT-Mol[188] pretrains on a wide range of modality including texts, images, and graphs.

7.5. Large Language Models. Recently, the most important artifacts built upon the transformer architecture are undoubtedly large language models (LLMs). LLMs such as GPT, Llama, Claude, and Falcon have significantly impacted various fields, including technology, education, creativity, business, healthcare and the society at large.[189] The revolutional power of LLMs comes from their ability to efficiently process enormous text corpus, capture the underlying associations/patterns, and reproduce the information via an intuitive interface with human-like communication. Driven by this wave of success, there is an rising interest in applying LLMs to assist scientific research. Most recent findings on the usage of LLMs in cheminformatics show that this is a promising direction.

The most straightforward application is using LLMs as property predictors via smart prompting.[190] In ChemLLMBench,[191] the authors created a testbed for benchmarking LLMs with eight chemical property prediction tasks. Jablonka et al.[192] finetune GPT-3 model on molecular classification and regression tasks. More importantly, the authors attempt inverse designing by training the LLM to generate molecular photoswitches with a desired range of wavelength. Although most generated molecules belong to the training set, a considerable portion of them are novel and do not exist in PubChem. Further evaluation reveals that the mean absolute error on the transition wavelengths of these molecules is remarkably around 10 percents of the desired values. ChemCrow[193] integrates existing chemical tools with LLMs to improve chemical research. In particular, chemical tools can help with input processing and output correction. Using such hybrid approach enhances the overall performance of various chemical tasks, including property prediction, structure generation, and prediction outcome forecasting. Interestingly, White et al.[194] formulate chemistry problems as coding tasks, on which LLMs have been shown to perform well. More specifically, the inputs are incomplete code with instructions as comments and LLMs are asked to complete the code to produce a function that executes certain chemical calculations.

As a new direction, applying LLMs on chemical tasks still requires further research and development. Encouragingly, the scientific community has responded with enthusiasm. In a relatively short period, numerous benchmarks and competitions have been established to expedite progress in this area.[191,195] We anticipate further advancements that will significantly enhance computational methods in cheminformatics.

8. SUMMARY AND CONCLUSION

Machine learning models play an important part in many modern chemical pipelines as they can efficiently assist or even replace expensive and time-consuming chemical experiments. This trend is assisted by the growing availability of large chemical databases and the rapid development of machine learning methods. Even though learning algorithms are traditionally developed for tabular and vectorized data, there has been a growing number of methods geared toward structural data, such as graph neural networks. These methods have led to a surge of applications and advancements in terms of performance on multiple chemical learning tasks. This fact confirms the effectiveness of powerful structural learning architectures on complex data in the chemical domain. For that reason, transformer models, which recently revolutionized learning in natural language processing and computer vision, gained the attention of researchers as a potential solution to chemical learning problems. In this paper, we reviewed recent efforts in applying transformer architectures to learning in the chemical domain.

Many methods utilize the 2D sequential representations of chemical structures to fit the input configuration of transformers. Examples include string-based amino acid chains, SMILES, and SELFIES. Such setup conveniently benefits from the established transformer models developed for text processing. Since text data are simple to obtain and process, large-scale self-supervised learning is feasible, resulting in multiple foundational models for string-based chemical data. Other works adapt and develop novel transformer architectures fitting the intrinsic representation and properties of chemical data, often in the form of graphs. A wide variety of graph transformers have been developed for this purpose, taking into account multiple geometric and chemical characteristics of structures. Interestingly, several research directions focus on domain-specific characteristics, such as equivariant transformers that are invariant to equivalent 3D configurations or graph transformers that process infinitely repetitive patterns to learn on material lattices.

Overall, transformer models are highly capable learning architectures, and the existing methods we review show the potential of transformers on chemical data. The chemical domain is vast and diverse with a variety of structures, a wide range of problems, and a diversity of physical and chemical characteristics that can be exploited for learning. As research efforts continue to rapidly expand across multiple scientific communities, we have great confidence that even more exciting results await us in the near future.

■ AUTHOR INFORMATION

Corresponding Authors

Kha-Dinh Luong − *Department of Computer Science, University of California Santa Barbara, Santa Barbara, CA 93106, United States*; orcid.org/0009-0003-6919-4528; Email: vluong@ucsb.edu

Ambuj Singh − *Department of Computer Science, University of California Santa Barbara, Santa Barbara, CA 93106, United States*; Email: ambuj@cs.ucsb.edu

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jcim.3c02070

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the BioPACIFIC Materials Innovation Platform of the National Science Foundation under Award DMR-1933487.

■ REFERENCES

(1) Shen, J.; Nicolaou, C. A. Molecular property prediction: recent trends in the era of artificial intelligence. *Drug Discovery Today: Technologies* **2019**, *32*, 29−36.

(2) Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learning Syst.* **2022**, *33*, 6999.

(3) Zhang, S.; Tong, H.; Xu, J.; Maciejewski, R. Graph convolutional networks: a comprehensive review. *Computational Social Networks* **2019**, *6*, 1−23.

(4) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. *Attention is All You Need.* Proceedings of the 31st International Conference on Neural Information Processing Systems; Red Hook, NY, 2017; pp 6000−6010.

(5) Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M. *Transformers: State-of-the-art natural language processing. Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*; Association for Computational Linguistics, 2020; pp 38−45, .

(6) Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; Yang, Z.; Zhang, Y.; Tao, D. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence* **2023**, *45*, 87−110.

(7) Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-Attention with Relative Position Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* **2018**, *2*, 464−468. Short Papers)

(8) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers)*; Minneapolis, MN, 2019; pp 4171−4186.

(9) Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.* 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26−30, 2020.

(10) Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*, 2019.

(11) Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023.

(12) Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language models are few-shot learners. In *Advances in neural information processing systems*; NeurIPS, 2020; Vol. *33*, pp 1877−1901.

(13) Khan, S.; Naseer, M.; Hayat, M.; Zamir, S. W.; Khan, F. S.; Shah, M. Transformers in Vision: A Survey. *ACM Computing Survey* **2022**, *54*, 1−41.

(14) Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; Houlsby, N. *An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations.* 2021.

(15) El-Nouby, A.; Touvron, H.; Caron, M.; Bojanowski, P.; Douze, M.; Joulin, A.; Laptev, I.; Neverova, N.; Synnaeve, G.; Verbeek, J.; Jegou, H. XCiT: Cross-Covariance Image Transformers. In *Advances in Neural Information Processing Systems*; NeurIPS, 2021.

(16) Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; Oh, S. J. *Rethinking Spatial Dimensions of Vision Transformers. International Conference on Computer Vision.* ICLR, 2021.

(17) Jiang, Z.-H.; Hou, Q.; Yuan, L.; Zhou, D.; Shi, Y.; Jin, X.; Wang, A.; Feng, J. All Tokens Matter: Token Labeling for Training Better Vision Transformers. In *Advances in Neural Information Processing Systems*; NeurIPS, 2021; pp 18590−18602.

(18) Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jegou, H. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*; Proceedings of the 38th International Conference on Machine Learning, 2021; pp 10347−10357.

(19) Camoglu, O.; Kahveci, T.; Singh, A. K. PSI: indexing protein structures for fast similarity search. *Bioinformatics* **2003**, *19*, i81−i83.

(20) Cheng, G.; Gong, X.-G.; Yin, W.-J. Crystal structure prediction by combining graph network and optimization algorithm. *Nature communications* **2022**, *13*, 1492.

(21) Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, X.; Canny, J.; Abbeel, P.; Song, Y. S. Evaluating Protein Transfer Learning with TAPE. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, NeurIPS, 2019.

(22) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*; 2021, 118, e2016239118.

(23) Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G. A; Sonnhammer, E. L L; Tosatto, S. C E; Paladin, L.; Raj, S.; Richardson, L. J; Finn, R. D; Bateman, A. Pfam: The protein families database in 2021. *Nucleic acids research* **2021**, *49*, D412−D419.

(24) Nambiar, A.; Heflin, M.; Liu, S.; Maslov, S.; Hopkins, M.; Ritz, A. Transforming the language of life: transformer neural networks for protein prediction tasks. *Proceedings of the 11th ACM international conference on bioinformatics, computational biology and health informatics*; ACM, 2020; pp 1−8.

(25) Beltagy, I.; Peters, M. E.; Cohan, A. Longformer: The Long-Document Transformer. *arXiv:2004.05150*, **2020**.

(26) Filipavicius, M.; Manica, M.; Cadow, J.; Martinez, M. R. Pre-training protein language models with label-agnostic binding pairs enhances performance in downstream tasks. *arXiv:2012.03084*, **2020**.

(27) Brandes, N.; Ofer, D.; Peleg, Y.; Rappoport, N.; Linial, M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **2022**, *38*, 2102−2110.

(28) Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Yu, W.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; Bhowmik, D.; Rost, B. ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, 2021; p 1, .

(29) Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* **2020**, *21*, 5485−5551.

(30) Wang, J.; Wen, N.; Wang, C.; Zhao, L.; Cheng, L. ELECTRA-DTA: a new compound-protein binding affinity prediction model based on the contextualized sequence encoding. *Journal of cheminformatics* **2022**, *14*, 1−14.

(31) Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*; NeurIPS, 2019; Vol. 32.

(32) Geffen, Y.; Ofran, Y.; Unger, R. DistilProtBert: a distilled protein language model used to distinguish between real proteins and their randomly shuffled counterparts. *Bioinformatics* **2022**, *38*, ii95−ii98.

(33) Rao, R. M.; Liu, J.; Verkuil, R.; Meier, J.; Canny, J.; Abbeel, P.; Sercu, T.; Rives, A. MSA Transformer. *Proceedings of the 38th International Conference on Machine Learning*; ICML, 2021; pp 8844−8856.

(34) Khare, E.; Gonzalez-Obeso, C.; Kaplan, D. L.; Buehler, M. J. CollagenTransformer: end-to-end transformer model to predict thermal stability of collagen triple helices using an NLP approach. *ACS Biomaterials Science & Engineering* **2022**, *8*, 4301−4310.

(35) Meier, J.; Rao, R.; Verkuil, R.; Liu, J.; Sercu, T.; Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems*, NeurIPS, 2021; Vol. *34*, pp 29287−29303.

(36) Jiang, T. T.; Fang, L.; Wang, K. Deciphering "the language of nature": A transformer-based language model for deleterious mutations in proteins. *Innovation* **2023**, *4*, No. 100487.

(37) Singh, J.; Litfin, T.; Singh, J.; Paliwal, K.; Zhou, Y. SPOT-Contact-LM: improving single-sequence-based prediction of protein contact map using a transformer language model. *Bioinformatics* **2022**, *38*, 1888−1894.

(38) Chowdhury, R.; Bouatta, N.; Biswas, S.; Floristean, C.; Kharkar, A.; Roy, K.; Rochereau, C.; Ahdritz, G.; Zhang, J.; Church, G. M.; Sorger, P. K.; AlQuraishi, M. Single-sequence protein structure prediction using a language model and deep learning. *Nat. Biotechnol.* **2022**, *40*, 1617−1623.

(39) Wang, W.; Peng, Z.; Yang, J. Single-sequence protein structure prediction using supervised transformer protein language models. *Nature Computational Science* **2022**, *2*, 804−814.

(40) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.

(41) Baek, M.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871−876.

(42) Ji, Y.; Zhou, Z.; Liu, H.; Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **2021**, *37*, 2112−2120.

(43) Avsec, Ž.; Agarwal, V.; Visentin, D.; Ledsam, J. R.; Grabska-Barwinska, A.; Taylor, K. R.; Assael, Y.; Jumper, J.; Kohli, P.; Kelley, D. R. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **2021**, *18*, 1196−1203.

(44) An, W.; Guo, Y.; Bian, Y.; Ma, H.; Yang, J.; Li, C.; Huang, J. MoDNA: motif-oriented pre-training for DNA language model. *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*; ACM, 2022; pp 1−5.

(45) Dalla-Torre, H.; Gonzalez, L.; Mendoza-Revilla, J.; Carranza, N. L.; Grzywaczewski, A. H.; Oteri, F.; Dallago, C.; Trop, E.; de Almeida, B. P.; Sirelkhatim, H. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, **2023**, .

(46) Zhou, Z.; Ji, Y.; Li, W.; Dutta, P.; Davuluri, R.; Liu, H. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv:2306.15006*, **2023**.

(47) Krenn, M.; et al. SELFIES and the future of molecular string representations. *Patterns* **2022**, *3*, No. 100588.

(48) McNaught, A. D.; Wilkinson, A. *Compendium of chemical terminology*; Blackwell Science: Oxford, 1997; Vol. 1669.

(49) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC international chemical identifier. *Journal of cheminformatics* **2015**, *7*, 1−34.

(50) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988**, *28*, 31−36.

(51) O'Boyle, N.; Dalke, A. DeepSMILES: an adaptation of SMILES for use in machine-learning of chemical structures. *ChemRxiv* **2018**, DOI: 10.26434/chemrxiv.7097960.v1.

(52) Lin, T.-S.; Coley, C. W.; Mochigase, H.; Beech, H. K.; Wang, W.; Wang, Z.; Woods, E.; Craig, S. L.; Johnson, J. A.; Kalow, J. A.; Jensen, K. F.; Olsen, B. D. BigSMILES: a structurally-based line notation for describing macromolecules. *ACS central science* **2019**, *5*, 1523−1531.

(53) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology* **2020**, *1*, No. 045024.

(54) Honda, S.; Shi, S.; Ueda, H. R. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv:1911.04738*, **2019**.

(55) Wang, S.; Guo, Y.; Wang, Y.; Sun, H.; Huang, J. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*; ACM, 2019; pp 429−436.

(56) Li, J.; Jiang, X. Mol-BERT: an effective molecular representation with BERT for molecular property prediction. *Wireless Communications and Mobile Computing* **2021**, *2021*, 1−7.

(57) Wu, Z.; Jiang, D.; Wang, J.; Zhang, X.; Du, H.; Pan, L.; Hsieh, C.-Y.; Cao, D.; Hou, T. Knowledge-based BERT: a method to extract molecular features like computational chemists. *Briefings in Bioinformatics* **2022**, *23*, No. bbac131.

(58) Kuenneth, C.; Ramprasad, R. polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nat. Commun.* **2023**, *14*, 4099.

(59) Xu, C.; Wang, Y.; Barati Farimani, A. a Transformer-based language model for polymer property predictions. *npj Computational Materials* **2023**, *9*, 64.

(60) Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. *arXiv:2010.09885*, **2020**.

(61) Ahmad, W.; Simon, E.; Chithrananda, S.; Grand, G.; Ramsundar, B. Chemberta-2: Towards chemical foundation models. *arXiv:2209.01712*, 2022.

(62) Irwin, R.; Dimitriadis, S.; He, J.; Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology* **2022**, *3*, No. 015022.

(63) Xue, D.; Zhang, H.; Chen, X.; Xiao, D.; Gong, Y.; Chuai, G.; Sun, Y.; Tian, H.; Wu, H.; Li, Y.; Liu, Q. X-MOL: large-scale pre-training for molecular understanding and diverse molecular analysis. *Sci. Bull.* **2022**, *67*, 899.

(64) Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence* **2022**, *4*, 1256−1264.

(65) Zhang, X.-C.; Wu, C.-K.; Yi, J.-C.; Zeng, X.-X.; Yang, C.-Q.; Lu, A.-P.; Hou, T.-J.; Cao, D.-S. Pushing the Boundaries of Molecular Property Prediction for Drug Discovery with Multitask Learning BERT Enhanced by SMILES Enumeration. *Research* **2022**, *2022*, 0004.

(66) Liu, Y.; Zhang, R.; Li, T.; Jiang, J.; Ma, J.; Wang, P. MolRoPE-BERT: An enhanced molecular representation with Rotary Position Embedding for molecular property prediction. *Journal of Molecular Graphics and Modelling* **2023**, *118*, No. 108344.

(67) Wen, N.; Liu, G.; Zhang, J.; Zhang, R.; Fu, Y.; Han, X. A fingerprints based molecular property prediction method using the BERT model. *Journal of Cheminformatics* **2022**, *14*, 1−13.

(68) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27−35.

(69) Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv:1301.3781*, **2013**.

(70) Wieder, O.; Kohlbacher, S.; Kuenemann, M.; Garon, A.; Ducrot, P.; Seidel, T.; Langer, T. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies* **2020**, *37*, 1−12.

(71) Müller, L.; Galkin, M.; Morris, C.; Rampášek, L. Attending to graph transformers. *arXiv:2302.04181*, **2023**.

(72) Zhang, J.; Zhang, H.; Xia, C.; Sun, L. Graph-bert: Only attention is needed for learning graph representations. *arXiv:2001.05140*, 2020.

(73) Page, L. *The pagerank citation ranking: Bringing order to the web*. Stanford Digital Library Technologies Project, 1998.

(74) Huang, N. T.; Villar, S. A short tutorial on the weisfeiler-lehman test and its variants. *ICASSP 2021−2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE, 2021; pp 8533−8537.

(75) Dwivedi, V. P.; Bresson, X. A generalization of transformer networks to graphs. *arXiv:2012.09699*, **2020**.

(76) Kreuzer, D.; Beaini, D.; Hamilton, W.; Létourneau, V.; Tossou, P. Rethinking graph transformers with spectral attention. *Advances in Neural Information Processing Systems*; NeurIPS, 2021; Vol. *34*, pp 21618−21629.

(77) He, X.; Hooi, B.; Laurent, T.; Perold, A.; LeCun, Y.; Bresson, X. A generalization of vit/mlp-mixer to graphs. International Conference on Machine Learning. ICML, 2023; pp 12724−12745.

(78) Kim, J.; Nguyen, D.; Min, S.; Cho, S.; Lee, M.; Lee, H.; Hong, S. Pure transformers are powerful graph learners. *Advances in Neural Information Processing Systems*; NeurIPS, 2022; Vol. *35*, pp 14582−14595.

(79) Chen, J.; Gao, K.; Li, G.; He, K. NAGphormer: A tokenized graph transformer for node classification in large graphs. *The Eleventh International Conference on Learning Representations*, 2022.

(80) Baek, J.; Kang, M.; Hwang, S. J. Accurate Learning of Graph Representations with Graph Multiset Pooling. *International Conference on Learning Representations*, 2020.

(81) Shirzad, H.; Velingker, A.; Venkatachalam, B.; Sutherland, D. J.; Sinop, A. K. Exphormer: Sparse transformers for graphs. *International Conference on Machine Learning*, 2023.

(82) Chen, D.; O'Bray, L.; Borgwardt, K. Structure-aware transformer for graph representation learning. *International Conference on Machine Learning*, 2022; pp 3469−3489.

(83) Dwivedi, V. P.; Luu, A. T.; Laurent, T.; Bengio, Y.; Bresson, X. Graph Neural Networks with Learnable Structural and Positional Representations. *International Conference on Learning Representations*, 2022.

(84) Mialon, G.; Chen, D.; Selosse, M.; Mairal, J. Graphit: Encoding graph structure in transformers. arXiv:2106.05667, **2021**.

(85) Chen, B.; Barzilay, R.; Jaakkola, T. Path-augmented graph transformer network. arXiv:1905.12712, **2019**.

(86) Maziarka, Ł.; Danel, T.; Mucha, S.; Rataj, K.; Tabor, J.; Jastrzębski, S. Molecule attention transformer. *arXiv:2002.08264*, **2020**.

(87) Gasteiger, J.; Groß, J.; Günnemann, S. *Directional Message Passing for Molecular Graphs*; International Conference on Learning Representations (ICLR), 2020.

(88) Maziarka, Ł.; Majchrowski, D.; Danel, T.; Gaiński, P.; Tabor, J.; Podolak, I.; Morkisz, P.; Jastrzębski, S. Relative molecule self-attention transformer. *arXiv:2110.05841*, 2021.

(89) Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; Leskovec, J. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*; NeurIPS, 2020; Vol. *33*, pp 22118−22133.

(90) Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; Liu, T.-Y. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems*, 2021; Vol. *34*, pp 28877−28888.

(91) Park, W.; Chang, W.-G.; Lee, D.; Kim, J.; Hwang, S. GRPE: Relative Positional Encoding for Graph Transformer. *ICLR2022 Machine Learning for Drug Discovery*, 2022.

(92) Hussain, M. S.; Zaki, M. J.; Subramanian, D. Global self-attention as a replacement for graph convolution. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*; ACM, 2022; pp 655−665.

(93) Chen, Z.; Tan, H.; Wang, T.; Shen, T.; Lu, T.; Peng, Q.; Cheng, C.; Qi, Y. Graph Propagation Transformer for Graph Representation Learning. *arXiv:2305.11424*, **2023**.

(94) Wu, F.; Radev, D.; Li, S. Z. Molformer: Motif-based transformer on 3d heterogeneous molecular graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI, 2023; pp 5312−5320.

(95) Rampášek, L.; Galkin, M.; Dwivedi, V. P.; Luu, A. T.; Wolf, G.; Beaini, D. Recipe for a general, powerful, scalable graph transformer*Advances in Neural Information Processing Systems*; NeurIPS, 2022; Vol. *35*, pp 14501−14515.

(96) Alon, U.; Yahav, E. On the Bottleneck of Graph Neural Networks and its Practical Implications. *International Conference on Learning Representations*, 2020.

(97) Li, Q.; Han, Z.; Wu, X.-M. Deeper insights into graph convolutional networks for semi-supervised learning. *Proceedings of the AAAI conference on artificial intelligence*. AAAI, 2018; 1, 1, .

(98) Srinivas, A.; Lin, T.-Y.; Parmar, N.; Shlens, J.; Abbeel, P.; Vaswani, A. Bottleneck Transformers for Visual Recognition. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE, 2021; pp 16514−16524.

(99) Wu, Z.; Jain, P.; Wright, M.; Mirhoseini, A.; Gonzalez, J. E.; Stoica, I. Representing long-range context for graph neural networks with global attention. *Advances in Neural Information Processing Systems*; NeurIPS, 2021; Vol. *34*, pp 13266−13279.

(100) Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; Huang, J. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*; NeurIPS, 2020; Vol. *33*, pp 12559−12571.

(101) Nguyen, D. Q.; Nguyen, T. D.; Phung, D. *Universal Graph Transformer Self-Attention Networks*. Companion Proceedings of the Web Conference 2022: New York, NY, USA, 2022; pp 193−196.

(102) Masters, D.; Dean, J.; Klaser, K.; Li, Z.; Maddrell-Mander, S.; Sanders, A.; Helal, H.; Beker, D.; Rampášek, L.; Beaini, D. Gps++: An optimized hybrid mpnn/transformer for molecular property prediction. *arXiv:2212.02229*, **2022**.

(103) Toropov, A. A.; Toropova, A. P. QSPR/QSAR: State-of-art, weirdness, the future. *Molecules* **2020**, *25*, 1292.

(104) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* **2018**, *9*, 513−530.

(105) Zhong, Z.; Zhou, K.; Mottin, D. Benchmarking Large Language Models for Molecule Prediction Tasks. *arXiv:2403.05075*, **2024**.

(106) Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; Ke, G. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. *The Eleventh International Conference on Learning Representations*, 2023.

(107) *OGB-LSC @ KDD Cup 2021 — ogb.stanford.edu*. https://ogb.stanford.edu/kddcup2021/results, 2021 (accessed 01-04-2024).

(108) *OGB-LSC Leaderboards — ogb.stanford.edu*. https://ogb.stanford.edu/docs/lsc/leaderboards (accessed 01-04-2024).

(109) Dwivedi, V. P.; Rampášek, L.; Galkin, M.; Parviz, A.; Wolf, G.; Luu, A. T.; Beaini, D. Long range graph benchmark. *Advances in Neural Information Processing Systems*, NeurIPS, 2022; Vol. *35*, pp 22326−22340.

(110) Karpov, P.; Godin, G.; Tetko, I. V. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *Journal of cheminformatics* **2020**, *12*, 1−12.

(111) Cremer, J.; Medrano Sandonas, L.; Tkatchenko, A.; Clevert, D.-A.; De Fabritiis, G. Equivariant graph neural networks for toxicity prediction. *Chem. Res. Toxicol.* **2023**, *36*, 1561−1573.

(112) Riedl, M.; Mukherjee, S.; Gauthier, M. Descriptor-Free Deep Learning QSAR Model for the Fraction Unbound in Human Plasma. *Mol. Pharmaceutics* **2023**, *20*, 4984−4993.

(113) Monteiro, N. R.; Oliveira, J. L.; Arrais, J. P. DTITR: End-to-end drug−target binding affinity prediction with transformers. *Computers in Biology and Medicine* **2022**, *147*, No. 105772.

(114) Kang, H.; Goo, S.; Lee, H.; Chae, J.-W.; Yun, H.-Y.; Jung, S. Fine-tuning of BERT model to accurately predict drug−target interactions. *Pharmaceutics* **2022**, *14*, 1710.

(115) Li, Z.; Ren, P.; Yang, H.; Zheng, J.; Bai, F. TEFDTA: a transformer encoder and fingerprint representation combined prediction method for bonded and non-bonded drug−target affinities. *Bioinformatics* **2024**, *40*, No. btad778.

(116) Monteiro, N. R.; Oliveira, J. L.; Arrais, J. P. TAG-DTA: Binding-region-guided strategy to predict drug-target affinity using transformers. *Expert Systems with Applications* **2024**, *238*, No. 122334.

(117) Huang, K.; Xiao, C.; Glass, L. M.; Sun, J. MolTrans: molecular interaction transformer for drug−target interaction prediction. *Bioinformatics* **2021**, *37*, 830−836.

(118) Lee, I.; Nam, H. Sequence-based prediction of protein binding regions and drug−target interactions. *Journal of cheminformatics* **2022**, *14*, 5.

(119) Tian, C.; Wang, L.; Cui, Z.; Wu, H. GTAMP-DTA: Graph transformer combined with attention mechanism for drug-target binding affinity prediction. *Computational Biology and Chemistry* **2024**, *108*, No. 107982.

(120) Wu, H.; Liu, J.; Jiang, T.; Zou, Q.; Qi, S.; Cui, Z.; Tiwari, P.; Ding, Y. AttentionMGT-DTA: A multi-modal drug-target affinity prediction using graph transformer and attention mechanism. *Neural Networks* **2024**, *169*, 623−636.

(121) Zhou, C.; Li, Z.; Song, J.; Xiang, W. TransVAE-DTA: Transformer and variational autoencoder network for drug-target binding affinity prediction. *Computer Methods and Programs in Biomedicine* **2024**, *244*, No. 108003.

(122) Ertl, P. Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *Journal of chemical information and computer sciences* **2003**, *43*, 374−380.

(123) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Medicinal research reviews* **1996**, *16*, 3−50.

(124) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A; Thiessen, P. A; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E PubChem 2023 update. *Nucleic acids research* **2023**, *51*, D1373−D1380.

(125) Pang, C.; Qiao, J.; Zeng, X.; Zou, Q.; Wei, L. Deep generative models in de novo drug molecule generation. *J. Chem. Inf. Model.* **2024**, *64*, 2174.

(126) Wei, L.; Fu, N.; Song, Y.; Wang, Q.; Hu, J. Probabilistic generative transformer language models for generative design of molecules. *Journal of Cheminformatics* **2023**, *15*, 88.

(127) Wang, J.; Hsieh, C.-Y.; Wang, M.; Wang, X.; Wu, Z.; Jiang, D.; Liao, B.; Zhang, X.; Yang, B.; He, Q.; Cao, D.; Chen, X.; Hou, T. Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning. *Nature Machine Intelligence* **2021**, *3*, 914−922.

(128) Yang, M.; Sun, H.; Liu, X.; Xue, X.; Deng, Y.; Wang, X. CMGN: a conditional molecular generation net to design target-specific molecules with desired properties. *Briefings in Bioinformatics* **2023**, *24*, No. bbad185.

(129) Born, J.; Manica, M. Regression transformer enables concurrent sequence regression and generation for molecular language modelling. *Nature Machine Intelligence* **2023**, *5*, 432−444.

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

(131) Mao, J.; Wang, J.; Zeb, A.; Cho, K.-H.; Jin, H.; Kim, J.; Lee, O.; Wang, Y.; No, K. T. Transformer-based molecular generative model for antiviral drug design. *J. Chem. Inf. Model.* **2024**, *64*, 2733.

(132) Bao, J.; Duan, N.; Zhou, M.; Zhao, T. Knowledge-based question answering as machine translation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)*; ACL, 2014; pp 967−976.

(133) Krasnov, L.; Khokhlov, I.; Fedorov, M. V.; Sosnin, S. Transformer-based artificial neural networks for the conversion between chemical notations. *Sci. Rep.* **2021**, *11*, No. 14798.

(134) Handsel, J.; Matthews, B.; Knight, N. J.; Coles, S. J. Translating the InChI: adapting neural machine translation to predict IUPAC names from a chemical identifier. *Journal of cheminformatics* **2021**, *13*, 1−11.

(135) Morris, P.; St Clair, R.; Hahn, W. E.; Barenholtz, E. Predicting binding from screening assays with transformer network embeddings. *J. Chem. Inf. Model.* **2020**, *60*, 4191−4199.

(136) Xu, Z.; Li, J.; Yang, Z.; Li, S.; Li, H. SwinOCSR: end-to-end optical chemical structure recognition using a Swin Transformer. *Journal of Cheminformatics* **2022**, *14*, 41.

(137) Khokhlov, I.; Krasnov, L.; Fedorov, M. V.; Sosnin, S. Image2SMILES: Transformer-based molecular optical recognition engine. *Chemistry-Methods* **2022**, *2*, No. e202100069.

(138) Rajan, K.; Zielesny, A.; Steinbeck, C. DECIMER 1.0: deep learning for chemical image recognition using transformers. *Journal of Cheminformatics* **2021**, *13*, 1−16.

(139) Shrivastava, A. D.; Swainston, N.; Samanta, S.; Roberts, I.; Wright Muelas, M.; Kell, D. B. MassGenie: A transformer-based deep learning method for identifying small molecules from their mass spectra. *Biomolecules* **2021**, *11*, 1793.

(140) He, J.; Nittinger, E.; Tyrchan, C.; Czechtizky, W.; Patronov, A.; Bjerrum, E. J.; Engkvist, O. Transformer-based molecular optimization beyond matched molecular pairs. *Journal of cheminformatics* **2022**, *14*, 18.

(141) Zheng, S.; Lei, Z.; Ai, H.; Chen, H.; Deng, D.; Yang, Y. Deep scaffold hopping with multimodal transformer neural networks. *Journal of cheminformatics* **2021**, *13*, 1−15.

(142) Litsa, E. E.; Das, P.; Kavraki, L. E. Prediction of drug metabolites using neural machine translation. *Chemical science* **2020**, *11*, 12777−12788.

(143) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science* **2019**, *5*, 1572−1583.

(144) Jaume-Santero, F.; Bornet, A.; Valery, A.; Naderi, N.; Vicente Alvarez, D.; Proios, D.; Yazdani, A.; Bournez, C.; Fessard, T.; Teodoro, D. Transformer performance for chemical reactions: Analysis of different predictive and evaluation scenarios. *J. Chem. Inf. Model.* **2023**, *63*, 1914−1924.

(145) Pesciullesi, G.; Schwaller, P.; Laino, T.; Reymond, J.-L. Transfer learning enables the molecular transformer to predict regio-and

stereoselective reactions on carbohydrates. *Nat. Commun.* **2020**, *11*, 4874.

(146) Andronov, M.; Voinarovska, V.; Andronova, N.; Wand, M.; Clevert, D.-A.; Schmidhuber, J. Reagent prediction with a molecular transformer improves reaction data quality. *Chemical Science* **2023**, *14*, 3235−3246.

(147) Tu, Z.; Coley, C. W. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *J. Chem. Inf. Model.* **2022**, *62*, 3503−3513.

(148) Hu, H.; Jiang, Y.; Yang, Y.; Chen, J. X. Enhanced Template-Free Reaction Prediction with Molecular Graphs and Sequence-based Data Augmentation. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*; ACM, 2023; pp 813−822.

(149) Karpov, P.; Godin, G.; Tetko, I. V. A transformer model for retrosynthesis. *International Conference on Artificial Neural Networks*; European Neural Network Society, 2019; pp 817−830.

(150) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hypergraph exploration strategy. *Chemical science* **2020**, *11*, 3316−3325.

(151) Zheng, S.; Rao, J.; Zhang, Z.; Xu, J.; Yang, Y. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *J. Chem. Inf. Model.* **2020**, *60*, 47−55.

(152) Tetko, I. V.; Karpov, P.; Van Deursen, R.; Godin, G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nat. Commun.* **2020**, *11*, 5575.

(153) Wang, X.; Li, Y.; Qiu, J.; Chen, G.; Liu, H.; Liao, B.; Hsieh, C.-Y.; Yao, X. Retroprime: A diverse, plausible and transformer-based method for single-step retrosynthesis predictions. *Chemical Engineering Journal* **2021**, *420*, No. 129845.

(154) Wan, Y.; Hsieh, C.-Y.; Liao, B.; Zhang, S. Retroformer: Pushing the limits of end-to-end retrosynthesis transformer. *International Conference on Machine Learning*; ICML, 2022; pp 22475−22490.

(155) Bagal, V.; Aggarwal, R.; Vinod, P.; Priyakumar, U. D. MolGPT: molecular generation using a transformer-decoder model. *J. Chem. Inf. Model.* **2022**, *62*, 2064−2076.

(156) Wang, X.; Yao, C.; Zhang, Y.; Yu, J.; Qiao, H.; Zhang, C.; Wu, Y.; Bai, R.; Duan, H. From theory to experiment: transformer-based generation enables rapid discovery of novel reactions. *Journal of Cheminformatics* **2022**, *14*, 60.

(157) Nugmanov, R.; Dyubankova, N.; Gedich, A.; Wegner, J. K. Bidirectional graphormer for reactivity understanding: neural network trained to reaction atom-to-atom mapping task. *J. Chem. Inf. Model.* **2022**, *62*, 3307−3315.

(158) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances* **2021**, *7*, No. eabe4166.

(159) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the space of chemical reactions using attention-based neural networks. *Nature machine intelligence* **2021**, *3*, 144−152.

(160) Reynaud, E. Protein misfolding and degenerative diseases. *Nature Education* **2010**, *3*, 28.

(161) Breydo, L.; Wu, J. W.; Uversky, V. N. α-Synuclein misfolding and Parkinson's disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* **2012**, *1822*, 261−285.

(162) Selkoe, D. J. Cell biology of protein misfolding: the examples of Alzheimer's and Parkinson's diseases. *Nature cell biology* **2004**, *6*, 1054−1061.

(163) Wang, J.; Wang, X.; Chu, Y.; Li, C.; Li, X.; Meng, X.; Fang, Y.; No, K. T.; Mao, J.; Zeng, X. Exploring the conformational space of protein-protein complex with transformer-based generative model. *bioRxiv* **2024**, 2024−02, DOI: 10.1101/2024.02.24.581708.

(164) Schwing, G.; Palese, L. L.; Fernández, A.; Schwiebert, L.; Gatti, D. L. Molecular dynamics without molecules: searching the conformational space of proteins with generative neural networks. *arXiv:2206.04683*, **2022**.

(165) Zeng, W.; Cao, S.; Huang, X.; Yao, Y. A note on learning rare events in molecular dynamics using lstm and transformer. *arXiv:2107.06573*, **2021**.

(166) Chennakesavalu, S.; Rotskoff, G. M. Data-Efficient Generation of Protein Conformational Ensembles with Backbone-to-Side-Chain Transformers. *J. Phys. Chem. B* **2024**, *128*, 2114.

(167) Thomas, N.; Smidt, T.; Kearnes, S.; Yang, L.; Li, L.; Kohlhoff, K.; Riley, P. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv:1802.08219*, **2018**.

(168) Fuchs, F.; Worrall, D.; Fischer, V.; Welling, M. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in neural information processing systems*, NeurIPS, 2020; Vol. *33*, pp 1970−1981.

(169) Thölke, P.; Fabritiis, G. D. Equivariant Transformers for Neural Network based Molecular Potentials. *International Conference on Learning Representations*, 2022.

(170) Liao, Y.-L.; Smidt, T. Equiformer: Equivariant Graph Attention Transformer for 3D Atomistic Graphs. *The Eleventh International Conference on Learning Representations*, 2022.

(171) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL materials* **2013**, *1*.1

(172) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge structural database. *Acta Crystallographica Section B: Structural Science Crystal Engineering and Materials* **2016**, *72*, 171−179.

(173) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters* **2018**, *120*, No. 145301.

(174) Yan, K.; Liu, Y.; Lin, Y.; Ji, S. Periodic graph transformers for crystal material property prediction. *Advances in Neural Information Processing Systems*, 2022; Vol. *35*, pp 15066−15080.

(175) Chen, P.; Jiao, R.; Liu, J.; Liu, Y.; Lu, Y. Interpretable graph transformer network for predicting adsorption isotherms of metal−organic frameworks. *J. Chem. Inf. Model.* **2022**, *62*, 5446−5456.

(176) Bai, J.; Du, Y.; Wang, Y.; Kong, S.; Gregoire, J.; Gomes, C. Xtal2DoS: Attention-based crystal to sequence learning for density of states prediction, *arXiv:2302.01486*, **2023**.

(177) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. *International Conference on Learning Representations*, 2018.

(178) Kang, Y.; Park, H.; Smit, B.; Kim, J. A multi-modal pre-training transformer for universal transfer learning in metal−organic frameworks. *Nature Machine Intelligence* **2023**, *5*, 309−318.

(179) Cao, Z.; Magar, R.; Wang, Y.; Barati Farimani, A. MOFormer: self-supervised transformer model for metal−organic framework property prediction. *J. Am. Chem. Soc.* **2023**, *145*, 2958−2967.

(180) Zeng, Z.; Yao, Y.; Liu, Z.; Sun, M. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nat. Commun.* **2022**, *13*, 862.

(181) Edwards, C.; Lai, T.; Ros, K.; Honke, G.; Cho, K.; Ji, H. Translation between Molecules and Natural Language. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*; Abu Dhabi, United Arab Emirates, 2022; pp 375−413.

(182) Liu, Z.; Zhang, W.; Xia, Y.; Wu, L.; Xie, S.; Qin, T.; Zhang, M.; Liu, T.-Y. *MolXPT: Wrapping Molecules with Text for Generative Pre-training*. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers). Toronto, Canada, 2023; pp 1606−1616.

(183) Li, J.; Liu, Y.; Fan, W.; Wei, X.-Y.; Liu, H.; Tang, J.; Li, Q. *Empowering Molecule Discovery for Molecule-Caption Translation with Large Language Models: A ChatGPT Perspective.* arXiv:2306.06615, **2023**.

(184) Edwards, C.; Zhai, C.; Ji, H. Text2Mol: Cross-Modal Molecule Retrieval with Natural Language Queries. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*; Online and Punta Cana, Dominican Republic, 2021; pp 595−607.

(185) Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*; EMNLP-IJCNLP: Hong Kong, China, 2019; pp 3615−3620.

(186) Su, B.; Du, D.; Yang, Z.; Zhou, Y.; Li, J.; Rao, A.; Sun, H.; Lu, Z.; Wen, J.-R. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv:2209.05481*, **2022**.

(187) Seidl, P.; Vall, A.; Hochreiter, S.; Klambauer, G. Enhancing Activity Prediction Models in Drug Discovery with the Ability to Understand Human Language. *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

(188) Liu, P.; Ren, Y.; Ren, Z. *Git-mol: A multi-modal large language model for molecular science with graph, image, and text*. arXiv:2308.06911, **2023**.

(189) Bharathi Mohan, G.; Prasanna Kumar, R.; Vishal Krishh, P.; Keerthinathan, A.; Lavanya, G.; Meghana, M. K. U.; Sulthana, S.; Doss, S. An analysis of large language models: their impact and potential applications. *Knowledge and Information Systems* **2024**, 1−24.

(190) Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Nee- lakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems* **2020**, *33*, 1877−1901.

(191) Guo, T.; Nan, B.; Liang, Z.; Guo, Z.; Chawla, N.; Wiest, O.; Zhang, X.; et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems* **2023**, *36*, 59662−59688.

(192) Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence* **2024**, *6*, 161−169.

(193) M. Bran, A.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence* **2024**, *6*, 525−535.

(194) White, A. D.; Hocky, G. M.; Gandhi, H. A.; Ansari, M.; Cox, S.; Wellawatte, G. P.; Sasmal, S.; Yang, Z.; Liu, K.; Singh, Y.; Pena Ccoa, W. J. Assessment of chemistry knowledge in large language models that generate code. *Digital Discovery* **2023**, *2*, 368−376.

(195) Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; Bocarsly, J. D.; Bran, A. M.; Bringuier, S.; Bringuier, L. C.; Choudhary, K.; Circi, D.; et al. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery* **2023**, *2*, 1233−1250.