Pre-trained Language Models in Biomedical Domain: A Systematic Survey

BENYOU WANG, SRIBD & SDS, The Chinese University of Hong Kong, Shenzhen, China
QIANQIAN XIE, Department of Computer Science, University of Manchester, United Kingdom
JIAHUAN PEI, University of Amsterdam, Netherlands
ZHIHONG CHEN, SRIBD & SSE, The Chinese University of Hong Kong, Shenzhen, China
PRAYAG TIWARI, School of Information Technology, Halmstad University, Sweden
ZHAO LI, The University of Texas Health Science Center at Houston, USA
JIE FU, Mila, University of Montreal, Canada

Pre-trained language models (PLMs) have been the de facto paradigm for most natural language processing tasks. This also benefits the biomedical domain: researchers from informatics, medicine, and computer science communities propose various PLMs trained on biomedical datasets, e.g., biomedical text, electronic health records, protein, and DNA sequences for various biomedical tasks. However, the cross-discipline characteristics of biomedical PLMs hinder their spreading among communities; some existing works are isolated from each other without comprehensive comparison and discussions. It is nontrivial to make a survey that not only systematically reviews recent advances in biomedical PLMs and their applications but also standardizes terminology and benchmarks. This article summarizes the recent progress of pre-trained language models in the biomedical domain and their applications in downstream biomedical tasks. Particularly, we discuss the motivations of PLMs in the biomedical domain and introduce the key concepts of pre-trained language models. We then propose a taxonomy of existing biomedical PLMs that categorizes them from various perspectives systematically. Plus, their applications in biomedical downstream tasks are exhaustively discussed, respectively. Last, we illustrate various limitations and future trends, which aims to provide inspiration for the future research.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Natural language generation**; **Neural networks**; **Bio-inspired approaches;**

Additional Key Words and Phrases: Biomedical domain, pre-trained language models, natural language processing

This work is supported by Chinese Key-Area Research and Development Program of Guangdong Province (2020B0101350001), the Shenzhen Science and Technology Program (JCYJ20220818103001002), the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen, Shenzhen Key Research Project (C10120230151) and Shenzhen Doctoral Startup Funding (RCBS20221008093330065).
Authors' addresses: B. Wang, SRIBD & SDS, The Chinese University of Hong Kong, Shenzhen, China; email: wangbenyou@cuhk.edu.cn; Q. Xie (corresponding author), Department of Computer Science, University of Manchester, United Kingdom; email: qianqian.xie@manchester.ac.uk; J. Pei, University of Amsterdam, Netherlands; email: j.pei@uva.nl; Z. Chen, SRIBD & SSE, The Chinese University of Hong Kong, Shenzhen, China; email: zhihongchen@link.cuhk.edu.cn; P. Tiwari, School of Information Technology, Halmstad University, Sweden; email: prayag.tiwari@ieee.org; Z. Li, The University of Texas Health Science Center at Houston, USA; email: lizhao.informatics@gmail.com; J. Fu, Mila, University of Montreal, Canada; email: jie.fu@polymtl.ca.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
0360-0300/2023/10-ART55 $15.00
https://doi.org/10.1145/3611651

ACM Reference format:
Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. 2023. Pre-trained Language Models in Biomedical Domain: A Systematic Survey. *ACM Comput. Surv.* 56, 3, Article 55 (October 2023), 52 pages.
https://doi.org/10.1145/3611651

1 INTRODUCTION

As the principal method of communication, humans usually record information and knowledge in a format of *token* sequences, e.g., natural languages, time series, constructed knowledge base, and so on. For biomedical information and knowledge, tokens in sequences could be of various types, including words, disease codes, amino acids, and DNA. Tremendous biomedical information and knowledge in nature and human history are implicitly encapsulated in these natural token sequences in nature (a.k.a., data).

There exist many data that involve biomedical information with different abstraction degrees of biomedical knowledge. However, there is a tradeoff between the high abstraction degree and its scale. For data that explicitly convey biomedical knowledge (i.e., at a high abstraction degree), they are usually small-scaled; see biomedical knowledge bases and **electronic health record (EHR)** data (maybe in multi-modality). One example of data that may not directly convey biomedical knowledge could be protein and DNA sequences, since one can hardly know what a short protein or DNA sequence really means for humans, and it needs more effort for abstraction. Fortunately, these data are usually tremendous. In the current stage, existing work pays more attention to data at a high abstraction level (biomedical knowledge-intensive data, e.g., EHR, biomedical knowledge bases, and biomedical encyclopedia); however, it is usually relatively small scale. We argue that biomedical knowledge on various abstraction degrees should be paid attention to. To capture and mine the biomedical information and knowledge from various abstraction degrees, there is recently growing attention in the biomedical **natural language processing (NLP)** community to adopt PLMs; since PLMs could leverage these massive sequences without biomedical knowledge abstraction and human annotations, including but not limited to plain biomedical text, biomedical images, general text, protein sequences, and DNA sequences.

The biomedical NLP is a cross-discipline research direction from various communities such as bioinformatics, medicine, and computer science (especially a major frontier of artificial intelligence, i.e., NLP). The computational biology community [129] and biomedical informatics community [51] have made a substantial effort to make use of NLP tools for information mining and extraction of widespread-adopted electronic health records, medical scientific publications, medical WIKI pages, and so on. For many decades, NLP has been investigating various biomedical tasks [50, 52] such as classification, information extraction, question answering, drug discovery, and so on. Meanwhile, the approaches in the NLP community are changing rapidly, as one can witness exponentially increasing submitted papers in top conferences like ACL, EMNLP, and NAACL. Tailoring these NLP approaches that have been evidenced effectively in the NLP community to a specific biomedical domain is beneficial.

Unfortunately, there is usually a delay for newly proposed NLP approaches being applied to the biomedical domain as seen in Figure 1. Especially, since the adoption of various pre-trained language models (e.g., ELMo [208], GPT [216], BERT [59], XLNET [104], RoBERTa [168], T5 [217], and ELECTRA [49]) [213] have nearly shifted the paradigm in NLP, their biomedical variants trained using biomedical data come sooner or later. With this hot trend of the biomedical pre-trained language model, this survey aims to bridge the gap between pre-trained language models and their applications in the biomedical domain.

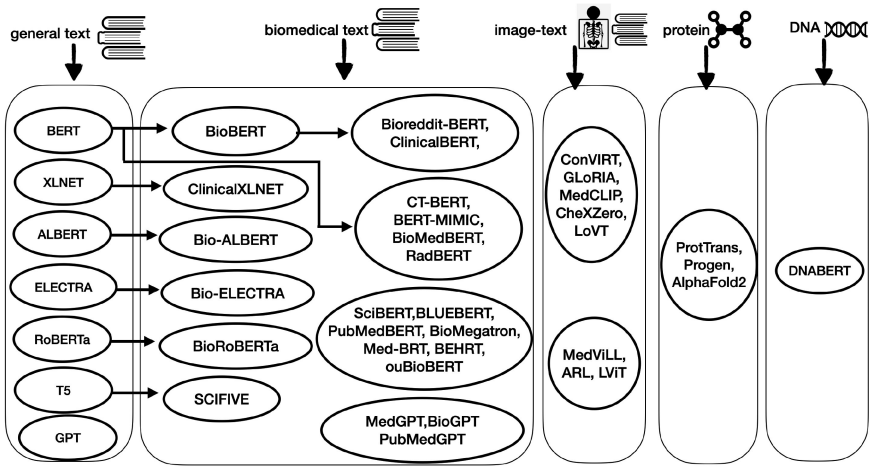

Fig. 1. Overview of selected released biomedical pre-trained language models. One can see a more detailed list in Section 3. Note that there is a BERT-like language model embedded in the overall architecture of AlphaFold 2.

Motivation of pre-trained language models in the biomedical domain. The current NLP paradigm is gradually shifting to a two-stage (pre-training and fine-tuning) paradigm, thanks to recently proposed pre-trained language models. Compared to the previous paradigm with purely supervised learning that relies on feature engineering or neural network architecture engineering [165], the current two-stage paradigm is more friendly to the scenario when supervised data are limited while large-scaled unsupervised data are tremendous. Fortunately, the biomedical domain is a typical case of such a scenario.

The motivation to use pre-trained language models in the biomedical domain is pretty straightforward. First, annotated data in the biomedical domain are usually not large scale. Therefore, a well-trained pre-trained language model is more crucial to provide a richer feature extractor, which may slightly reduce the dependence on annotated data. Second, the biomedical domain is more knowledge-intensive than the general domain. At the same time, pre-trained language models could serve as an easily used soft knowledge base [209] that captures implicit knowledge from large-scale plain documents without human annotations. More recently, GPT3 has been shown to have the potential to "remember" many complicated common knowledge [33]. Last, large-scaled biomedical corpora and biomedical sequences (including proteins and DNAs), which are previously thought as difficult to handle, can be effectively handled by pre-trained language models (especially transformer networks).

As shown in Figure 2, in recent three years, we have witnessed a rapid development of pre-trained language models (e.g., ELMo [208], GPT [216], BERT [59], XLNet [104], RoBERTa [168], T5 [217], and ELECTRA [49]) in the general NLP domain. Following these progresses, there are efforts to tailor these pre-trained language models to their corresponding biomedical variants via in-domain data. For example, BERT, the most typical pre-trained language, has many variants in the biomedical domain, e.g., Med-BERT [224], BioBERT [142], publicly available Clinical BERT Embeddings [8], SciBERT [18], ClinicalBERT [103], COVID-twitter-BERT [193], and so on. We draw an overview for these models in Figure 1. It shows that the extensions of general domain pre-trained language models to the biomedical domain attract great attention from researchers in both the NLP and bioinformatics communities. Interestingly, we can observe that once the general NLP community develops a new variant of PLM, it usually leads to a biomedical counterpart after

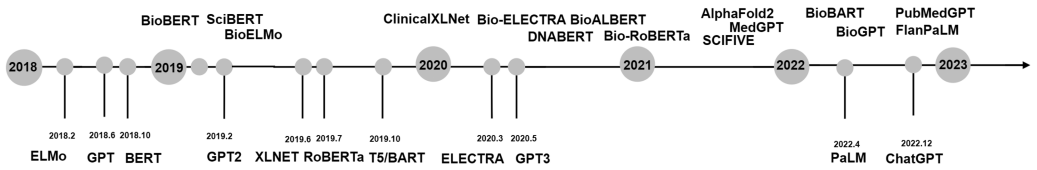

Fig. 2. Parallel development of general and biomedical pre-trained language models. The time is determined by the released date of the paper, for example, in arXiv. General pre-trained language models are shown below in the timeline, and biomedical pre-trained language models are shown above the timeline (refer to Table 4 for detailed dates).

some months. This parallel development between general PLMs and biomedical PLMs shows a strong demand and even a necessity to summarize the existing works, which could help beginners to start their contributions in this field easily.

Difference with existing surveys. There are a few reviews to summarize the NLP applications in the biomedical, clinical, bioinformatic domain, such as an early one [251] and recent ones [206, 295, 328]. They cover many general methods and applications of biomedical/clinical NLP. Specifically, Reference [251] mainly discuss either based on statistics-based NLP pipeline (including lexicon, co-occurrence patterns, syntactic/semantic parsing), or word embeddings based neural network approaches (it was mentioned that 60.8% of them are based on recurrent neural networks) [295] for NLP applications (e.g., information extraction, text classification, named entity recognition, and relation extraction.). Especially, two reviews [119, 123] discuss the word embeddings used in biomedical NLP.

All the above reviews made thorough summarization of existing work before the pre-trained language model era of NLP. The NLP techniques in these reviews are mainly about feature engineering or architecture engineering [165]. However, the NLP recently has been shifted to a pre-training and then fine-tuning paradigm with large-scale pre-trained language models (see existing surveys [27, 86, 165, 166, 213] for pre-trained language model in the general domain). Reference [27] called these pre-trained models as "foundation models" to underscore their critically central. We believe the biomedical NLP applications have benefited and will continually benefit from the development of pre-trained language models.

More recently, Reference [118] reviews biomedical textual pre-training, especially using BERT. The difference between Reference [118] and this review is that our article provides a more inclusive taxonomy of biomedical PLMs than Reference [118], which are threefold. First, biomedical PLMs summarized in our review are limited not only to that trained on texts like Reference [118] but also other data resources, including protein, DNA, and even biomedical text–image pairs. In general, any data that involve biomedical information could be used in biomedical PLMs. Second, in contrast to Reference [118], which only discusses Transformer-based pre-trained language models, this review also discusses **recurrent neural network– (RNN)** based language models (like ELMO [111], which is typically considered as the first pre-trained language model in NLP). We also summarize decoder involved *generative* pre-trained language models (like GPT [133] and T5 [210]), while Reference [118] mainly discusses encoder-based PLMs (BERT or BERT variants). Third, to the best of our knowledge, this is the first survey paper to discuss pre-trained *vision-language* models in the biomedical domain. Last, our article provides a more comprehensive overview of the applications of PLMs in the biomedical domain compared with Reference [118]. Except for biomedical NLP tasks such as natural language inference, text summarization [303], relation extraction, and so on, that are summarized in Reference [118], our article further reviews recent PLMs-based methods for event detection, dialogue systems, as well as protein and DNA sequence.

Moreover, compared with Reference [118] that only reviews recent methods of biomedical NLP tasks coarsely, we make a thorough categorization and discussion of PLMs-based methods for biomedical NLP tasks and their benchmark datasets. Our article also introduces competitions and venues such as shared tasks. Therefore, we believe there is a requirement for a more thorough survey paper to review the recent progress of pre-trained language models in the biomedical domain from a multi-scale perspective.

Contribution. The contributions of the article can be summarized as follows:

- We give a comprehensive review to summarize existing PLMs-based methods for the biomedical domain, which thoroughly categorizes and discusses biomedical data sources, biomedical PLMs, model variants, downstream tasks, and so on.
- We propose a taxonomy of biomedical PLMs, which classifies existing PLMs in the biomedical domain from various perspectives: training data sources, model architecture, and so on.
- We enumerate existing resources for PLMs and their detailed configuration, facilitating their spreading for beginners.
- We discuss the limitations of existing methods and prospect future trends.
- To the best of our knowledge, this is the first survey paper to summarize generative pre-trained language models, protein/DNA language models, pre-trained *vision-language* models in the biomedical domain.

How do we collect the papers? In this survey, we collected over 100 related papers. We used Google Scholar as the main search engine and also adopted MedPub, Web of Science, as an essential tool to discover related papers. In addition, we screened most of the related conferences and journals such as ACL, EMNLP, NAACL, AAAI, Bioinformatics, JAMIA, AMIA, and so on. The major keywords we used included medical pre-trained language model, clinical pre-trained language model, biological language model, and so on. Plus, we take Med-BERT [224], BioBERT [142], SciBERT [18], ClinicalBert [103], and COVID-twitter-BERT [193] as the seed papers to check papers that cited them.

Organization. The overall architecture of this article is shown in Figure 3. The article is organized as below: Section 2 introduces the general pre-trained language models including backbone networks, pre-training objective, pre-training corpora, fine-tuning, and categorization of PLMs. Section 3 introduces the pre-trained language models for the biomedical domain and proposes a taxonomy, including motivations for using PLMs, biomedical data sources, domain-specific pre-training, biomedical PLMs, and their categorization. Section 4 summarizes the applications of biomedical PLMs for various downstream tasks and categorizes existing methods for these tasks respectively. More discussions about limitations and future directions are in Section 5. We conclude in Section 6.

2 BACKGROUND: Pre-trained Language Models

PLMs have been widely used in natural language processing, and so on, due to their effectiveness to learn useful representations from unannotated data such as natural languages. In this article, we mainly discuss pre-trained language in sequential tokens.[1] We will introduce the textual pre-training in Section 2.2, one can read the review paper of PLMs in Reference [213] for more details. Thanks to the popularity of **contrastive language-image pre-training (CLIP)**, pre-trained language models are also usually jointly trained with a visual pre-trained model in the image-text pre-training scenario. We will also discuss visual pre-training in Section 2.3. Note that models in

[1]Tokens usually refers to words or subwords in NLP and also protein sequences in the biomedical domain.

Fig. 3. Architecture of this survey.

the visual pre-training usually treat image patches as visual tokens; this makes it language model-like pre-training. Therefore, we include visual pre-training models in this survey.

In this section, we will introduce the basic ingredients of pre-training models: the training objective with self-supervised tasks and corpora in Section 2.2 and Section 2.3 for text and images, respectively; basic neural network models in Section 2.1; and training paradigm in Section 2.4.

2.1 Backbone Networks in Language Models

The success of pre-trained language models is also attributed to the development of their base backbone network from **long short-term memory (LSTM)** [98] to Transformer [272]. Before Transformer was invented, LSTM was widely used as the base architecture of pre-trained language models such as ELMO. However, because of its recurrence structure, it is computationally expensive to scale up LSTM to be deeper in layers. To this end, Transformer is proposed and becomes the backbone of modern NLP. Transformers are better architecture can be attributed to (1) efficiency (a recurrent-free architecture that could compute the individual token in parallel) and (2) effectiveness (attention allows spatial interaction across tokens that dynamically depends on the input itself). In this section, we briefly introduce the two typical architectures in pre-trained language models, namely LSTM and Transformers.

2.1.1 Previous Backbone Networks in Texts.

LSTM. LSTM is an RNN) architecture for sequential modeling. Unlike standard feed-forward neural networks processing single data points (such as images), LSTM can deal with entire sequences of data (such as text, speech, or video). A common LSTM unit is composed of a cell, an

input gate, an output gate, and a forget gate. The cell learns hidden states over arbitrary time intervals, and the three gates regulate the flow of information into and out of the cell. LSTM networks are well suited for time-series data and were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs. Peters et al. [208] tried to adopt an LSTM network in pre-trained language, which naturally processes tokens sequentially.

2.1.2 *Previous Backbone Networks in Images.*

CNNs. **Convolutional neural networks (CNNs)** [141] are a type of neural networks that are particularly suited for vision tasks. Typically, CNNs are made up of four main types of layers: convolution, pooling, activation, and fully connected layers. The convolution layers are trainable filters that can learn to recognize patterns in images, such as edges, textures, and objects. The pooling layers are used to reduce the dimensionality of the data; the activation layers are used to introduce non-linearity to the network. The fully connected layers are used to make predictions based on the extracted features. Note that CNNs are also a good choice for language understanding [127].

2.1.3 *The Current Backbone Networks in Texts and Images.*

Transformer. The backbone of most pre-trained language models (e.g., BERT, its variants, GPT, and T5) is a neural network called "Transformer" building upon **self-attention networks (SANs)** and **feed-forward networks (FFNs)**. SAN is used to facilitate interaction between tokens, while FNN is used to refine the token presentation using non-linear transformation. Since Transformer has been the de facto backbone to replace recurrent and convolutional units, almost all language models adopt the Transformer as the backbone network. The Transformer is superior in terms of capacity and scalability thanks to (1) discarding recurrent units and process tokens more efficiently in parallel with the position embeddings [280, 281], (2) relieving saturation issue of expressive power with large-scale data and very deep layers due to the well-designed architecture including residual connections, layer normalization, and so on.

A Transformer layer consists of a SAN module and a FFN module. An input X[2] for SAN will be linearly Transformed into query, key, value, and output space $\{Q, K, V\}$ as follows:[3]

$$\begin{bmatrix} Q \\ K \\ V \end{bmatrix} = X \times \begin{bmatrix} \boldsymbol{W}^Q \\ \boldsymbol{W}^K \\ \boldsymbol{W}^V \end{bmatrix}. \tag{1}$$

The self-attention mechanism (a.k.a. Scaled Dot-Product Attention) is calculated as

$$\text{Attention(Q, K, V)} = \text{softmax}\left(\frac{QK}{\sqrt{d_k}}\right)V. \tag{2}$$

For a multi-head version of the self-attention mechanism, it linearly projects Q, K, V with h times using individual linear projections to smaller dimensions (e.g., $d_k = \frac{d_{\text{model}}}{h}$) instead of performing a single attention function with d_{model}-dimensional keys, values, and queries. Finally, the output of SAN is

$$\begin{aligned} \text{SAN}(X) &= [\text{head}_1; \ldots; \text{head}_h]\boldsymbol{W}^O \\ \text{head}_i &= \text{Attention}(Q_i, K_i, V_i), \end{aligned} \tag{3}$$

where $Q = [Q_1; \ldots Q_h]$, $K = [K_1; \ldots K_h]$, and $V = [V_1; \ldots V_h]$. The individual attention heads are independently calculated. A stack of many purely SAN layers is not expressive [63], since it is

[2] X is the word embedding of each individual input token that are tokenized using subword tokenization. Moreover, the input is usually concatenated with position embeddings [286] to perceive word order.
[3] For all linear transformation in this article, the bias term is in default omitted.

equivalent to a single linear transformation. To this end, a feed-forward network with non-linear activation is alternately used with each SAN layer,

$$\text{FFN}(X) = \delta(X\boldsymbol{W}^{\text{in}})\boldsymbol{W}^{\text{out}}. \tag{4}$$

Since some neurons after the activation function (e.g., δ is ReLU or GELU [94]) become inactivated (zero), d_{in} is usually bigger than d_{model} to avoid the low-rank bottleneck, typically, $d_{\text{in}} = 4 \times d_{\text{model}} = d_{\text{out}}$. Other tricks, such as layer normalization, residual connection, dropout, and weight decay, are also adopted to relieve the optimization and overfitting problems when it goes deeper, resulting in better stability when training large neural networks. It is generally believed that Transformer is better than LSTM in terms of generalization, since its performance usually does not get to saturation as early as LSTM. When models become large, the performance of the Transformer is consistently increasing when feeding more data while LSTM gets saturation if a certain amount of data is fed.

Interestingly, the computer vision [156] and computational biology communities also borrow some insights to design their models; see ViT [156] for vision and AlphaFold2 [116] for protein. In Table 1, we introduce some typical pre-trained language models in general NLP domains, based on these two backbone neural networks.

2.2 Pre-training for Texts

Previously, there were many typical methods to build token representation (e.g., word vectors) from plain corpora. For example, References [183, 205] build a one-to-one mapping between words and their vectors, which is called "static word embedding," since it is static and not related to word context. However, it is well known that words often express different meanings in different contexts. To achieve this, most recently many pre-trained language models [208] are proposed to learn "contextualized word embedding" that models the bi-directional contexts of words. For "contextualized word embedding," the vector for a word depends on its specific usage in a context. For example, the meanings of "`bank`" in "`river bank`" and in "`money bank`" are supposed to have some difference. Compared with "static word embedding," the "contextualized word embedding" largely improves the quality of word representation in various tasks [59].

A *language model* aims to assign a probability to a given piece of text (e.g., a sentence or an n-gram) [117] as follows:

$$\Theta : \mathbb{V}^N \rightarrow \mathbb{R}^+, \tag{5}$$

while in the scenario of natural language processing, a generally called *language model* is usually a *conditional language model* that assigns a probability to a next word w_n given some conditioning context (denoted as $[w_1, \ldots, w_{n-1}]$). A conditional language model is a generalization of *language model* in a sense the former could be obtained by dividing the probability of the concatenated sentence (i.e., $[w_1, \ldots, w_{n-1}, w_n]$) by that of the context, namely

$$P(w_n|w_1, \ldots, w_{n-1}, w_n) = \frac{\Theta(w_1, \ldots, w_{n-1})}{\Theta(w_1, \ldots, w_{n-1})}. \tag{6}$$

In the earliest, neural language models [20, 184] and their variants, such as Skip-Gram [183], CBow [183], and Glove [205], were the backbones of modern NLP to provide pre-trained word features. The pre-training task of classical neural language models [20] is the unidirectional language modeling that predicts the next word conditionally on history words. To learn better word embeddings, several classical models further improved the pre-training task. For example, the training objective of Skip-Gram [183] is predicting context words given the input word. CBow [183] aims

Table 1. Typical Ways for Word Vectors and Language Models

Model	Type	Architecture	Task	Loss function
NLM [20]	Static	One-layer MLP	$(a, b) \rightarrow c$ predicting the next word	$-\sum_{i=1}^{T} logp(x_i \vert \{x_1, \ldots, x_{i-1}\})$
Skip-Gram [183]	Static	One-layer MLP	$b \rightarrow c, \quad b \rightarrow a$ predicting neighboring words	$-\sum_{i=1}^{T} logp(\{x_{i-o}, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{i+o}\} \vert x_i)$, ($o$ is the window size)
CBow [183]	Static	One-layer MLP	$(a, c) \rightarrow b$ predicting central words	$-\sum_{i=1}^{T} logp(x_i \vert \{x_{i-o}, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{i+o}\})$, ($o$ is the window size)
Glove [205]	Static	One-layer MLP	$\vec{w_i}^T \vec{w_j} \propto logp(\#(w_i w_j))$ predicting the log co-occurrence count	$-\sum_{i=1, j=1}^{T} f(x_{ij})(\vec{w_i}^T \vec{w_j} + b_i + c_j - logx_{ij}), (x_{ij} = p(\#(w_i w_j)))$
ELMO [208]	Contextualized	LSTM	$(a, b, c, d) \rightarrow e, \quad (e, d, c, b) \rightarrow a$ bi-directional language model	$-\sum_{i=1}^{T} logp(x_i \vert \{x_1, \ldots, x_{i-1}\}) + logp(x_i \vert \{x_{i+1}, \ldots, x_T\})$
BERT [59], Roberta [168] ALBERT [140], XLNET [317]	Contextualized	Transformers or Transformer-XL	$(a, [\text{mask}], c) \rightarrow (\_, b, \_)$ predicting masked words	$-\sum_{x \in mask(x)} logp(x \vert \hat{X})$, $\hat{X}$ is the corrupted sentence with masks
Electra [49]	Contextualized	Transformer	$(a, b, c, d) \rightarrow (0, 1, 0, 1)$ replaced token prediction	$-\sum_{i=1}^{T} logp(b_i \vert \hat{X})$, b_i indicates whether x_i is replaced.
T5 [217] BART [144]	Contextualized	Transformers	$(a, b, c,) \rightarrow (d, e)$ predicting the sequence	$-\sum_{i=1}^{T} logp(y_i \vert X, y_0, \ldots, y_{i-1})$, X and $Y = \{y_1, \ldots, y_T\}$ are the input/output
GPT [216]	Contextualized	Transformers	$(a, b, c, d) \rightarrow e$ autoregressively predicting the next word	$-\sum_{i=1}^{T} logp(x_i \vert \{x_1, \ldots, x_{i-1}\})$, $\{x_1, \ldots, x_T\}$ is the sequence

$X = \{a, b, c, d, e\}$ is an example text sequence. ELMO, BERT, and GPT usually work on much longer sequences than NLMs, Skip-gram, and CBOW.

to predict the next word based on its bidirectional context words. The training task of Glove [205] is to predict the log co-occurrence of words. These models typically use shallow neural network architecture to conduct calculations between word vectors for efficient training.

Language models could be considered as an instance of self-supervision. Compared to data-hungry supervised learning, which usually needs annotations from humans, language models could make use of massive amounts and cheap plain corpora from the internet, books, and so on. In language models, a next word is a natural label for a context sentence as a next word prediction task, or one can artificially mask a known word and then predict it. The paradigm that uses the unstructured data itself to generate labels (for example, the next word or the masked word in language models) and train language models to predict labels thereof is called "self-supervision learning." Language model pre-training is therefore referred to as an "auxiliary task," in which the learned representations in language models can be used as an initial model for various downstream supervised tasks. The pre-training objective/task is critical for learning efficient representations that are generalizable and universal for downstream tasks.

Recently, efforts have been proposed to learn contextualized word representations based on deep neural networks, such as the pioneer method ELMO [208], GPT [216], and the breakthrough work, BERT [59]. Similarly to traditional neural language models, GPT uses the unidirectional language model task as the pre-training objective. ELMO proposed the pre-training task for bidirectional language modeling based on both the forward language model and backward language model task. The forward language model task aims to model the probability of the word given its previous words, while the backward language model task predicts the word based on its future words. To better model bi-directional contexts during pre-training, BERT proposed the **masked language model (MLM)** pre-training objective with the inspiration of the Cloze task. It randomly masks tokens of input sequences and aims to predict masked tokens with the masked text sequences. Different from ELMO, which concatenates the forward and backward language model, MLM can train the deep bidirectional contextual representations with only one language model. Based on MLM, Encoder-Decoder language models such as T5 [217] proposed the pre-training objective of generating the given sequences in an auto-regressive way taking the masked sequences as input. The language models based on the auto-regressive pre-training objective are more suitable for the text generation tasks such as abstractive summarization and question answering. The overview of pre-training tasks is shown in Table 1. Recently, Open AI have released many API services on their trained model, including GPT 3, InstuctGPT, Codex, and ChatGPT. Especially, ChatGPT could interact in a conversational manner, which makes it possible to answer follow-up questions, admit mistakes, challenge incorrect premises, and reject inappropriate requests.

These pre-training tasks in language modeling are sometimes called "pretext tasks." In conclusion, by pre-training multi-layer transforms in plain text using pretext tasks, it learns general text representation that can easily be adapted to downstream tasks.

Pre-training corpora. Except for the superior pre-training objective, it usually requires a large scale of raw texts to pre-training language models effectively. On the internet, unlabelled raw texts are abundant, ranging from news texts and web pages to online encyclopedias. The training corpora for pre-trained language models mainly include (1) online encyclopedias like Wikipedia,[4] which was widely used for training BERT and its variants; (2) existing books and stories that have been digitized, like BooksCorpus [345]; and (3) web texts extracted from online websites/URL, such as crawled online corpora.[5] PLMs trained by these corpora are usually able to capture the commonsense knowledge inherited in the raw training texts. For specific domains such as the biomedical domain, it therefore needs other efforts such as domain-specific pre-training with domain-specific texts to capture the domain knowledge (will further be introduced in the next section). Moreover, the vocabulary with limited words is unable to cover all words in the large-scale training texts. To address the out-of-vocabulary problem, they proposed to split words into sub-words to formulate the vocabulary via the Byte-Pair Encoding [237] or WordPiece [138] methods.

Table 2. Categories to Tailor Pre-trained Language Models

Category	Data	Task
Pre-training	General domain	Pre-training task
Domain adaption	Target domain	Pre-training task
Task adaption	General domain	Downstream task
Fine-tuning	Target domain	Downstream task

Representative PLMs. Pre-trained language models can generally be categorized into three principal types, based on whether the input or output constitutes a text sequence or label: Encoder-only, Decoder-only, and Encoder-Decoder models. Models such as BERT [59], RoBERTa [168], and ALBERT [140] fall under the Encoder-only category and are primarily utilized for text classification and sequence labeling tasks. RoBERTa [168] is a BERT variation that has undergone a more extended training phase and employs additional data. ALBERT [140] serves as a lightweight BERT variant but features shared weights and a factorized word embedding.

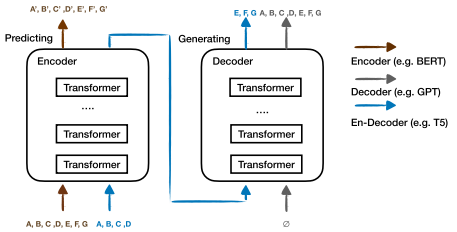

Fig. 4. The difference between Encoder, Decoder, and En-Decoder PLMs.

Pre-trained models equipped with the decoder such as GTP series, T5, **Bidirectional and Auto-Regressive Transformers (BART)**, could deal with generation-related tasks like translation, summarization, and language models.[6] See Figure 4 for the difference: an Encoder model predicts labels for each input tokens (in brownish yellow), a Decoder model generates a sequence of tokens w.r.t. a probability distribution (in blue), and an En-Decoder model predicts a new sequence conditioned on a given sequence (in grey), a.k.a. Seq2Seq.

Knowledge in PLMs. As a pioneer, LAMA [209] has explored the how much PLMs could capture factual and commonsense knowledge (in the format of triplets in knowledge bases). It concludes

[4]https://dumps.wikimedia.org/
[5]https://commoncrawl.org/
[6]XLNet [317] provides a generalization of autoregressive pre-training by leveraging bidirectional contexts to conduct masked word prediction akin to BERT. It could also deal with text generation.

that large PLMs (e.g., BERT-Large) can recall knowledge slightly better than small competitors and remarkably better than with non-neural and supervised alternatives [209]. However, Reference [34] revises the idea that PLMs can potentially be a reliable knowledge source. Cao et al. [34] claims that the way PLMs capture knowledge is vulnerable; it might overfit dataset artifacts and make use of answer leakage. In the biomedical domain, it needs more domain knowledge, and it is therefore more knowledge intensive than the general domain. Some existing work (e.g., Reference [108]) has explored injecting biomedical domain knowledge into PLMs.

2.3 Pre-training for Images

Deep neural networks have achieved excellent performance in the imaging domain on various vision tasks, e.g., image classification, object detection, and instance segmentation. One of the major reasons behind this is pre-training. However, different from language models in the NLP field, "pre-training" in the earliest means training vision models on large annotated image datasets, e.g., ImageNet [58]. Subsequently, different self-supervised learning approaches are proposed to overcome the shortcoming of supervised learning, e.g., generalization error and spurious correlations. Next, we detail different types of pre-training for images.

Supervised pre-training. In supervised pre-training, the most commonly used dataset is ImageNet, which contains over one million labeled images. Supervised pre-training [90, 136] involves training a deep learning model on the entire ImageNet dataset to learn generic features that can be useful for various downstream tasks. Once the model has been pre-trained on the large dataset, it can be fine-tuned on a smaller, task-specific dataset relevant to the specific task. This can help the model learn valuable features that can be generalized to different tasks at hand.

Contrastive self-supervised Learning. Different from supervised pre-training, contrastive self-supervised learning [43, 78, 89] is a method for representation learning without needing labeled data. It involves training a model to distinguish between different variations of a given input image. For example, the model might be trained to identify whether two images are a rotated version of the same image or whether they are two completely different images. By learning to predict these labels, the model can learn useful features that can be applied to various tasks, such as object detection and semantic segmentation.

Masked self-supervised Learning. Motivated by BERT in NLP, masked self-supervised learning has attracted attention in the computer vision field [16, 88, 305]. It is a type of generative pre-training approach. Models are trained to reconstruct images from incomplete data, in which part of the input image is removed or masked before it is fed into the model. This allows the model to learn the underlying structure of the image.

Contrastive language-image pre-training. CLIP [214] aims to train a vision model on a wide variety of image-text datasets. The model is trained to pair images and texts in a mini-batch through contrastive learning. CLIP showed excellent zero-shot transfer ability, where the pre-trained model can achieve comparable results with the original ResNet [90] on ImageNet in a zero-shot manner. One of the primary reasons is that texts provide rich, detailed information about the visual content of an image. For example, a text description of an image can include information about the objects and scenes depicted in the image, as well as their spatial relationships and attributes. This information can help a machine learning model to identify and understand an image's visual content. Additionally, texts can be easily generated and collected in large quantities, making them a convenient and scalable source of supervision for visual representation learning.

2.4 Fine-tuning Paradigm in PLMs

One challenge to use PLMs in downstream tasks is that there are two gaps between PLMs and downstream tasks, the *task gap* and *domain gap*. The *task gap* means the meta-task in PLMs (usually masked language model in BERT or causal language model in GPT) usually cannot directly be tailored to most downstream tasks (e.g., sentimental classification). The *domain gap* refers to the difference between the trained corpora in PLMs and the needed domain in a specific downstream task. The adaptation of both *task gap* and *domain gap* is crucial.

Adaption. To use the pre-trained language model in a downstream task, it is suggested to adopt both the domain and task adaption [79, 83, 228, 334]; see Table 2 for the difference. The domain adaption suggests continuing training pre-trained models trained from a general domain in the target domain, e.g., biomedical domain. Task adaption refers to fine-tuning on similar downstream tasks. In this article, without specifying, we mainly discuss the domain-adapted pre-trained models in various downstream tasks. Task adaption is not the main concern in this review. Take BERT as an example. BERT is first trained using **next-sentence predictions (NSP)** and masked language models in the pre-training phase. Such a pre-trained BERT will be used as the initial feature extractor. BERT with an additional classifier layer is then fine-tuned to optimize the objective of down-stream tasks (like MNLI [293], **named entity recognition (NER)** [266], and SQuAD [218]).

3 PLMS IN BIOMEDICAL DOMAIN

Recently, the pre-trained language models have been widely applied to various NLP tasks and achieved significant improvement in performance, because (1) pre-training on the huge text corpus can learn universal language representations and help with the downstream tasks; (2) pre-training provides a better model initialization, which usually leads to a better generalization performance and speeds up convergence on the target task; and (3) pre-training can be regarded as a kind of regularization to avoid overfitting on small data [213]. Self-supervised learning, on which pre-trained language models rely, usually adopts plain unstructured corpora in a format of a sequence of tokens. At first, most pre-trained language models focus on pre-training in general plain corpora from the Internet, like Wikipedia or crawled webpages. Except for the general domain, efforts have been proposed to extend PLMs in specific domains such as the following: Reference [71] trains CodeBERT in the programming language, and Reference [18] trains SciBERT on scientific publications and biological sequence. This article aims to discuss pre-trained language models in the biomedical domain. It is believed that the pre-trained language model can always benefit from more training corpora [79]. To achieve better performance in the domain-specific downstream tasks, it is also intuitive that the in-domain data pre-training is necessary.

We will first introduce the motivation of using pre-trained language models in the biomedical domain in the Section 3.1. Then we will illustrate the main components on tailoring PLMs to the biomedical domain, including the in-domain data in the Section 3.2 and the pre-training and fine-tuning strategy in the Section 3.3. Next, in Section 3.4, we will introduce existing pre-trained models in the biomedical domain, which are pre-trained from the in-domain data as introduced in Section 3.2. We will give an overview of these models and catagorization of them and discuss differences between them. We expect to help those in both the bioinformatics and computer science communities to get knowledge of the biomedical domain-specific pre-trained language model quickly.

3.1 Motivation

In the biomedical domain, the motivation for using pre-trained language models is manyfold.

- First, the biomedical domain involves biomedical data in the format of sequential tokens (like biomedical texts and the history of electronic health records) that usually lack annotations.

Table 3. Summary of Biomedical Data for Pre-training

dataset	types	size	characteristics
MIMIC III	EHR	58,976 hospital admissions for 38,597 patients	from Beth Israel Deaconess Medical Center in 2001-2012
CPRD	EHR	11.3M patients	anonymized medical records from 674 UK GP practices
BREATHE	Scientific Publications	6M articles and about 4 billion words	sources are diverse.
PubMed	Scientific Publications	35M citations and abstracts of biomedical literature	It provide only links to journal articles
COMETA in Reddit	Social Media	800K Reddit posts	68 health-themed subreddits with entity annotation
Tweets	Social Media	up-to-date	one could crawl real-time Tweets using its official API
UMLS	Knowledge Bases	2M names for 900K concepts	well-organized medical knowledge source
IU-Xray	image-Text Pairs	3,955 reports and 7,470 images	XML reports with findings, indications, comparisons, and so on
MIMIC-CXR	image-Text Pairs	77,110 images	images corresponding to 227,835 radiographic studies
ROCO	image-Text Pairs	81,000 radiology images and corresponding captions	figures and their corresponding captions in PubMed articles
MedICaT	image-Text Pairs	17,000 images includes captions	open-access biomedical papers and their captions

However, these sequential data were previously thought of as difficult to model. Thanks to pre-trained language models, it has been empirically demonstrated to train these sequential data in a self-supervised manner effectively. This would open a new door for processing biomedical data with pre-trained language models.

- Second, annotated data in the biomedical domain are usually limited at scale. Some extreme cases in machine learning are called "zero-shot" or "few-shot." More recently, language models such as GPT3 show that language models have the potential for few-shot learning and even zero-shot learning [33]. Therefore, a well-trained pre-trained language model in the biomedical domain is more crucial to provide a richer feature extractor, which may slightly reduce the dependence on annotated data.
- Plus, the biomedical domain is more knowledge intensive than the general domain, since most tasks may need domain expert knowledge, while pre-trained language models could serve as an easily used soft knowledge base [209] that captures implicit knowledge from large-scale plain biomedical corpora without human annotations. Recent study shows that GPT-3 [33] has the potential to "recall" intricate common knowledge and "reason" based on it [321].
- Last, beyond text, there exist various types of biological sequential data in the biomedical domain, like protein and DNA sequences. Using these data to train language models has shown great success in biological tasks like protein structure predictions. Therefore, it is expected that pre-trained language models could solve more challenging problems in biology.

3.2 Biomedical Data for Pre-training

Unstructured plain data for pre-trained language models mainly include electronic health records, scientific publications, social media text, biomedical image–text pairs, and other biological sequences like protein; see Table 3. An overview of EHR mining can be seen in References [67, 309], and Reference [77] discussed both health records and social media text. One can also check Reference [119] for some systematic overview of biomedical textual corpora.

3.2.1 Electronic Health Record. EHR is a collection of patient and population electronically stored health information in a digital format that may include demographics, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal statistics like age and weight, and billing information. One can check [249, 291] for details about EHR with deep learning. Assessing such records may be restricted to limited organizations, which hinders its widespread to the public. The reason may involve some privacy issues.

MIMIC III. The Medical Information Mart for Intensive Care III dataset [115][7] is one of the most popular EHR datasets, which consists of 58,976 unique hospital admissions from 38,597 patients

[7] https://mimic.mit.edu/

in the intensive care unit of the Beth Israel Deaconess Medical Center between 2001 and 2012. In addition, there are 2,083,180 de-identified notes associated with the admissions.

CPRD. The **Clinical Practice Research Datalink (CPRD)** [97] is the primary care database of anonymized medical records from 674 general physicians practices in the UK, which involves over 11.3 million patients. It consists of data on demographics, symptoms, tests, diagnoses, therapies, and health-related behaviors. It is also linked to secondary care (i.e., hospital episode statistics, or HES) and other health and administrative databases (e.g., office for national statistics" death registration). With 4.4 million active (alive, currently registered) patients meeting quality criteria, approximately 6.9% of the UK population are included, this shows that patients are broadly representative of the UK general population in terms of age, sex, and ethnicity. As a result, CPRD has been widely used across countries and spawned a lot of scientific research output.

3.2.2 Scientific Publications. Scientific publications are another source for biomedical pretrained language models, since we expect that biomedical knowledge may be encapsulated in scientific publications.

BREATHE. **Biomedical Research Extensive Archive To Help Everyone (BREATHE)**.[8] is a large and diverse dataset collection of biomedical research articles from leading medical archives. It contains titles, abstracts, and full-body texts. The dataset collection process was done with public APIs that were used when available. The primary advantage of the BREATHE dataset is its source diversity. BREATHE is from nine sources including BMJ, arXiv, medRxiv, bioRxiv, CORD-19, Springer Nature, NCBI, JAMA, and BioASQ [37]. BREATHE v1.0 contains more than 6M articles and about 4 billion words. BREATHE v2.0 is the most recent version.

PubMed. PubMed[9] is a free search engine accessing the MEDLINE database of references and abstracts on life sciences and biomedical topics primarily. PubMed comprises more than 32 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher websites. PubMed abstracts (PubMed) have 4.5B words, and **PubMed Central full-text articles (PMC)** have 13.5B words.

3.2.3 Social Media. Users post information on social media, which may contain biomedical information. We mainly introduce Reddit and Tweets as examples.

Reddit. Reddit is an American social news aggregation, web content rating, and discussion website. Registered members submit content to the site, such as links, text posts, images, and videos, and the content is voted up or down by other members. Posts are organized by subject into user-created boards called "communities" or "subreddits," which cover a variety of topics such as news, politics, religion, science, movies, video games, music, books, sports, fitness, cooking, pets, and image-sharing. Submissions with more up-votes appear toward the top of their subreddit and, if they receive enough up-votes, ultimately on the site's front page. Despite strict rules prohibiting harassment, Reddit's administrators have to moderate the communities and, on occasion, close them. COMETA corpus [17] crawled health-themed forums on Reddit using Pushshift (Baumgartner et al., 2020) and Reddit's own APIs.

Tweets. Twitter is an American micro-blogging and social networking service on which users post and interact with messages known as "tweets." Registered users can post, like, and retweet

[8]https://cloud.google.com/blog/products/ai-machine-learning/google-ai-community-used-cloud-to-help-biomedical-researchers
[9]https://pubmed.ncbi.nlm.nih.gov/

tweets. Tweets were originally restricted to 140 characters, but the limit was doubled to 280 for non-CJK languages in November 2017. Audio and video tweets remain limited to 140 s for most accounts. The COVID-twitter-BERT [193] is trained on a corpus of 160M tweets about the coronavirus collected through the Crowdbreaks platform [194] during the period from January 12 to April 16, 2020.

3.2.4 Online Medical Knowledge Sources. Other than unstructured text, there is some online medical knowledge source that is well organized. For example, **Unified Medical Language System (UMLS)** provides biomedical concepts that may benefit biomedical pre-trained language models.

UMLS. UMLS (http://umlsks.nlm.nih.gov) [25] is a repository of biomedical vocabularies developed by the US National Library of Medicine. The UMLS has over 2 million names for 900,000 concepts from more than 60 families of biomedical vocabularies, as well as 12 million relations among these concepts. These vocabularies include the NCBI taxonomy, the Medical Subject Headings, Gene Ontology, OMIM, and the Digital Anatomist Symbolic Knowledge Base. The UMLS knowledge sources are updated every quarter. In addition, all vocabularies are freely available for research purposes within an institution if a license agreement is signed.

3.2.5 Biomedical Image-Text Pairs. Besides texts, there are many medical texts paired with their corresponding images. These types of data are a good resource for learning the cross or joint representations of medical images and texts.

IU-Xray. IU-Xray [56] has a collection of chest X-ray images from the Indiana University hospital network. The data include two files: one for the images and the other for the XML reports of the radiography. Each report may have multiple images, typically having two views: frontal and lateral. The XML reports contain information such as findings, indications, comparisons, and impressions. In total, there are 3,955 reports and 7,470 images.

MIMIC-CXR. Medical Information Mart for Intensive Care Chest X-Ray [114] is a large publicly available dataset of chest radiographs with free-text radiology reports. It contains 377,110 images corresponding to 227,835 radiographic studies performed at the Beth Israel Deaconess Medical Center in Boston, Massachusetts.

ROCO. Radiology Objects in COntext [203] is a large-scale medical and multimodal imaging dataset from the articles of PubMed Central, an open-access biomedical literature database. They are figures and their corresponding captions in articles. It has over 81,000 radiology images (from various imaging modalities) and their corresponding captions.

MedICaT. MedICaT [256] is also a dataset of medical figure–caption pairs also extracted from PubMed Central. Different from ROCO, and 74% of its figures are compound figures, including several sub-figures. It contains more than 217,000 images from 131,000 open-access biomedical papers and includes captions, inline references, and manually annotated sub-figures and sub-captions.

3.2.6 Biological Sequences. Other than text, there are various types of biomedical token sequences, e.g., amino acids for proteins. The structure of each protein is fully determined by a sequence of amino acids [10]. These amino acids are from a limited-size amino acid vocabulary, of which 20 are commonly observed. This is similar to text that is composed of words in a lexicon vocabulary. In this subsection, we introduce a protein dataset called "Pfam" and a DNA sequence dataset from the Human Genome Project.

Pfam Protein Dataset. The Pfam database[10] is a large collection of protein families, in which each protein is represented by multiple sequence alignments using hidden Markov models. The newest version is Pfam 34.0, which was released in March 2021 and contains 19,179 families (or called "entries") and 645 clans.[11] The original purpose of the Pfam database is for the classification of protein families and domains. It creates the database using a semi-automated method of curating information on known protein families. Pfam 34.0 contains 47 million sequences, which could be used to train protein language models.

DNA Dataset. The DNA sequence is composed of a genomic sequence. The Human Genome Project was the international research effort to determine the DNA sequence of the entire human genome. Human Genome Project Results. In 2003, an accurate and complete human genome sequence was finished two years ahead of schedule and at a cost less than the original estimated budget. Reference [109] uses the reference human genome GRCh38.p13 primary assembly from GENCODE Release.[12] The total sequence length is about 3 billion.

3.3 How to Tailor PLMs to the Biomedical Domain

The pre-trained language model [59] is a new two-stage paradigm for NLP. In the first phase, it trains a language model (e.g., masked language model and casual language model) with a self-supervised meta-task in task-agnostic corpora. In the second phase, it fine-tunes the pre-trained language model to a (usually small-scaled) specific downstream task. To tailor pre-trained language models on the biomedical domain, References [79, 103, 142] have explored conducting the domain-specific adaptation on both the pre-training and fine-tuning stages. In the pre-training stage, the domain-specific adaption of existing efforts involves in the continual pre-training or training from scratch with a large scale of raw biomedical data. This yield many efficient foundation models in the biomedical domain such as BioBERT [142] and PubMedBERT [79], which can be directly used for downstream domain-specific tasks in the fine-tuning stage.

3.3.1 Biomedical Language Model Pre-training. One challenge in the biomedical domain is that medical jargon and abbreviations consist of many terms that are composed of Latin or Greek parts. Moreover, clinical notes have different syntax and grammar from books or encyclopedias. These lead to the semantic and domain-knowledge gap between the general pre-trained language models and the biomedical domain. Therefore, many existing approaches have investigated the biomedical language models pre-training on the basis of pre-trained language models in the general domain to tailor pre-trained language models to the biomedical domain.

Continual pre-training. The general way used by many methods [103, 142, 204] is to conduct the continual pre-training based on the general pre-trained language models such as BERT. They directly initialize the model with existing general PLMs and further pre-training it with the self-supervised task and domain-specific corpora, such as PubMed texts and MIMIC-III. The representative works include the BioBERT [142], which conducts continual pre-training based on the BERT with the PubMed abstracts and PubMed Central full-text articles; BlueBERT [204], which uses PubMed texts and MIMIC-III; and Clinical BERT [103], which further pre-trains BERT with clinical notes. In this case, they use the same vocabulary as the general PLMs, which cover words in a corpus of the general domain such as Wikipedia and BookCorpus. However, as mentioned

[10] http://pfam.xfam.org/

[11] Clans are the generated higher-level groupings of related entries in Pfam. A clan is a collection of entries that are related by sequence similarity, structure, or profile-HMM.

[12] https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39/

before, biomedical texts consist of many domain-specific terms. Using the same vocabulary as the general PLMs can be ineffective for modeling biomedical texts [79].

Pre-training from scratch. To conduct better pre-training for biomedical language models, some efforts [18, 79] have explored the way of pre-training from scratch. Different from the continual pre-training, they propose to build the new vocabulary from the raw biomedical training corpora. SciBERT [18] is the representative work that constructs the new vocabulary with the size of 30K and trains the model with the mix-domain corpora, where 18% of the training texts are from the computer science domain and 82% are from the biomedical domain. However, one recent work [79] has argued that the mixed domain pre-training does not make sense for the biomedical domain, since the target data of downstream applications in the biomedical domain is highly domain specific. Instead, they proposed the superior domain-specific pre-training from scratch that uses only biomedical corpora.

Summary. Our observation is that the core factors that affect the decision between training from scratch or continuously training are twofold: the scale of pre-training biomedical corpora and the domain specificity for biomedicine, where we need to make a tradeoff. Pre-training is, in general, data hungry, and one could fully leverage a large amount of biomedical corpora without inheriting parameters from a well-trained general PLM if there already exist enough biomedical corpora. Early work (e.g., Reference [322]) tends to continuously train biomedical PLMs from an initial BERT. Nowadays, it becomes more popular to directly train biomedical PLMs from scratch thanks to the large scale of collected data and adequate computing resources [170]. Interestingly, Reference [245] reused and tailored a giant general PLM (PaLM) to a clinical one, since giant models are economically expensive. We might expect some approach to decompose existing models and reuse part of them; afterward, one can inject biomedical modules into it.

3.3.2 Fine-tuning. Based on well-trained biomedical language models, one has to adapt them to downstream tasks. This is typically implemented to replace the mask language model prediction head and next sentence prediction head with a downstream prediction head, e.g., classification head, or sequence labeling heads.

Since the downstream tasks usually have much less training data than those used in pre-training, fine-tuning is an unstable process. Sun et al. [258] investigate different fine-tuning methods of BERT on the natural text classification tasks. Mosbach et al. [191] argues that the fine-tuning instability is due to vanishing gradients. Merchant et al. [180] observe that fine-tuning mainly modifies the top layers of BERT. Unfortunately, the solutions (e.g., hyper-parameters of which layer to fine-tune) proposed in those papers cannot be easily translated to other settings. To automate this process, automatic hyper-parameter tuning (e.g., Bayesian optimization [32, 269]) can come into help. Tinn et al. [264] systematically study fine-tuning stability in biomedical NLP. Particularly, it finds that freezing lower layers is beneficial for small models, while layerwise decay is beneficial for larger models. In most cases, it facilitates robust fine-tuning by using domain specific vocabulary and pre-training.

3.4 Biomedical Pre-trained Language Models

Based on the types of training corpora in the biomedical domain as introduced in Section 3.2, we mainly introduce two groups of biomedical pre-trained language models: biomedical textual language models and protein language models. Based on the types of training corpora in the biomedical domain as introduced in Section 3.2, we mainly introduce biomedical pre-trained language models in three scenarios: pure language models, vision-and-language modeling, and protein/DNA language models.

Table 4. Existing Textual Biomedical Pre-trained Models

Model	Corpora	Architecture	Size	Date	Link
BioBERT [142]	PubMed and PMC	BERT	base & large	2019.01	https://github.com/dmis-lab/biobert
BERT-MIMIC [244]	MIMIC III	BERT	base and large	2019.02	—
SciBERT [18]	Semantic Scholar papers	BERT	base	2019.03	https://github.com/allenai/SciBERT
BioELMo [111]	PubMed abstracts	ELMo	93.6 M	2019.04	https://github.com/Andy-jqa/bioelmo
Clinical BERT [8]	EHR (MIMIC-III)	BERT	base	2019.04	https://github.com/EmilyAlsentzer/clinicalBERT
Clinical BERT [103]	EHR (MIMIC-III)	BERT	base	2019.05	https://github.com/kexinhuang12345/clinicalBERT
BlueBERT [204]	PubMed+MIMIC-III	BERT	base & large	2019.05	https://github.com/ncbi-nlp/bluebert
G-BERT [238]	MIMIC III	BERT	—	2019.06	https://github.com/jshang123/G-Bert
BEHRT [152]	Clinical Practice Research Datalink	BERT	—	2019.07	https://github.com/deepmedicine/BEHRT
BioFLAIR [239]	PubMed abstracts	BERT	lagre	2019.08	https://github.com/zalandoresearch/flair
RadBERT [178]	RadCore radiology reports	BERT	—	2019.12	—
EhrBERT [146]	MADE corpus	BERT	base	2019.12	https://github.com/umassbento/ehrbert
Clinical XLNet [104]	EHR (MIMIC-III)	XLNET	base	2019.12	https://github.com/lindvalllab/clinicalXLNet
CT-BERT [193]	Tweets about the coronavirus	BERT	large	2020.05	https://github.com/digitalepidemiologylab/covid-twitter-bert
Med-BERT [224]	Cerner Health Facts (general EHR)	BERT	—	2020.05	https://github.com/ZhiGroup/Med-BERT
ouBioBERT [275]	PubMed	BERT	base	2020.05	https://github.com/sy-wada/blue_benchmark_with_transformers
Bio-ELECTRA [202]	PubMed	ELECTRA	base	2020.05	https://github.com/SciCrunch/bio_electra
BERT-XML	Anonymous Institution EHR system	BERT	small and base	2020.06	
PubMedBERT [79]	PubMed	BERT	base	2020.07	https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract
MCBERT [333]	Chinese social media, wiki and EHR	BERT	base	2020.08	https://github.com/alibaba-research/ChineseBLUE
BioALBERT [198]	PubMed and PMC	ALBERT	base & large	2020.09	https://github.com/usmaann/BioALBERT
BRLTM [179]	private EHR	BERT	customized	2020.09	https://github.com/lanyexiaosa/brltm
BioMegatron [243]	PubMed and PMC	BERT	0.3/0.8/1.2B	2020.10	https://ngc.nvidia.com/
ClinicalTransformer [315]	MIMIC III	[1]	base	2020.10	https://github.com/uf-hobi-informatics-lab/ClinicalTransformerNER
Bioreddit-BERT [17]	healththemed forums on Reddit	BERT	base	2020.10	https://github.com/cambridgeltl/cometa
BioRoBERTa [145]	PubMed, PMC, and MIMIC-III	RoBERTa	base & large	2020.11	https://github.com/facebookresearch/bio-lm
CODER [325]	UMLS Metathesaurus	BERT	base	2020.11	https://github.com/GanjinZero/CODER
bert-for-radiology [31]	daily clinical reports	BERT	—	2020.11	https://github.com/rAIdiance/bert-for-radiology
BioMedBERT [37]	BREATHE	BERT	large	2020.12	https://github.com/BioMedBERT/biomedbert
LBERT [288]	PubMed	BERT	base	2020.12	https://github.com/warikoone/LBERT
ELECTRAMED [186]	PubMed	ELECTRA	base	2021.04	https://github.com/gmpoli/electramed
SCIFIVE [210]	PubMed Abstract and PMC	T5	220/770M	2021.06	https://github.com/justinphan3110/SciFive
MedGPT [133]	King's College Hospital and MIMIC-III	GPY	customized	2021.07	https://pypi.org/project/medgpt/
Clinical-Longformer [154]	MIMIC-III	Longformer [19]	base	2022.01	https://github.com/luoyuanlab/Clinical-Longformer
Clinical-BigBird [326] [154]	MIMIC-III	BigBird	base	2022.01	https://github.com/luoyuanlab/Clinical-Longformer
BioLinkBERT [318]	PubMed with citation links	BERT	base& large	2022.03	https://github.com/michiyasunaga/LinkBERT
BioBART [323]	PubMed	BART	base & large	2022.04	https://github.com/GanjinZero/BioBART
BioGPT[170]	PubMed	GPT	GPT-2_{medium} [2]	2022.09	https://github.com/microsoft/BioGPT
PubMedGPT	PubMed	GPT	2.7B	2022.12	https://www.mosaicml.com/blog/introducing-pubmed-gpt
Flan-PaLM [245]	Instruction[3]	PaLM [48]	8B,62B and 540B	2022.12	unavailable
Med-PaLM 2 [246]	Instruction[4]	PaLM 2 [11]	8B,62B and 540B	2023.5	unavailable
HuatuoGPT [329]	Instruction + conversation	GPT (Bloom [233])	7B	2023.5	https://github.com/FreedomIntelligence/HuatuoGPT

The base setting is with 0.1B parameters, and the large setting is with 0.3B parameters. The date is based on the submission in arXiv or published date of the journal or conference proceeding.

[1]ClinicalTransformer [315] provides a series of biomedical models based on different architectures including BERT, RoBERTa, ALBERT, ELECTRA, DistilBERT, XLNet, Longformer, and DeBERTa.

[2]BioGPT adopts GPT-2_{medium} as the backbone network (24 layers, 1024 hidden size and 16 attention heads), resulting 347M 355M parameters in total. Its parameter size is close to BERT-large.

[3][245] adopts instruction prompt tuning on medical data. The details were not introduced.

[4]Instructions are from MedQA, MedMCQA, HealthSearchQA, LiveQA and MedicationQA.

3.4.1 Overview of Existing Biomedical Textual Language Models. Since BERT was released, various biomedical pre-trained language models have been proposed via continued training with in-domain corpora based on the BERT model or training from scratch. Table 4 presents existing pre-trained language models with used corpora, size, release date, and related web pages.

We introduce some representative pre-trained language models, including encoder-only pre-trained language models like BioBERT, ClinicalBERT, SciBERT, and COVID-twitter-BERT, decoder-only pre-trained language models like MedGPT, and encoder-decoder pre-trained language models like SCIFIVE.

- *BioBERT* [142] is initialized with the general BERT model and pre-trained on PubMed abstracts and PMC full-text articles.
- *ClinicalBERT* [103] is trained on clinical text from approximately 2M notes in the MIMIC-III database [115], a publicly available dataset of clinical notes.

- *SciBERT* [18] is trained on the large scale of scientific papers from a multi-domain based on the BERT. The training papers are from 1.14 M full-text papers in Semantic Scholar, in which 82% articles are from the biomedical domain.
- *COVID-twitter-BERT* [193] is a natural language model to analyze COVID-19 content on Twitter. The COVID-twitter-BERT model is trained on a corpus of 160M tweets about the coronavirus collected through the Crowdbreaks platform during the period from January 12 to April 16, 2020.
- *MedGPT* [133] is a GPT-like language model trained by patients' medical history in the format of EHRs. Given the sequence of past events, MedGPT aims to predict future events like a diagnosis of a new disorder or complications of an existing disorder.
- *SCIFIVE* [210] is a domain-specific T5 model that is pre-trained on large biomedical corpora. Like T5, SCIFIVE is a typical Seq2seq paradigm to transform an input sequence into an output sequence.

3.4.2 Discussions on Biomedical Pre-trained Language Models. Here we will discuss the listed models in various aspects as follows.

Training corpora: EHR, literature, social media, and so on, or the hybrid? Most pre-trained language models are based on scientific publications, e.g., PubMed, and EHR notes. Note that EHR datasets are usually relatively smaller than scientific publications datasets or Wikipedia. Hence pre-trained language models with only EHR datasets are typically trained from the initialization of well-trained BERT [8, 103], XLNET[104], and so on. Furthermore, some PLMs (e.g., BioRoBERTa [145]) adopt both scientific publications and EHRs. A few models such as CT-BERT and Bioreddit-BERT [17, 193] adopt social media, including Twitter and Reddit.

Extra features. EHR data usually have some extra meaningful features, e.g., disease codes, personal information of patients like age, gender. Such extra features can be embedded as dense vectors used in some models such as Med-BERT and BEHRT [152, 224] like word embedding, position embedding, and segment embedding.

Training from scratch or continue training. The standard approach to obtain a biomedical pre-trained model is to conduct continual pre-training from a general-domain pre-trained model like BERT [59], such as the BioBERT [322]. Specifically, this approach would initialize the model with the standard BERT model, including its word vocabulary, which is pre-trained by general Wikipedia and BookCorpus. Besides, some literature demonstrated that training from scratch may fully make use of in-domain data and reduce the negative effect from out-of-domain corpora, which may be beneficial for downstream tasks such as PubMedBERT [79].

Reusing existing vocabulary or building a new one. To make use of well-trained general pre-trained language models like BERT [59], one has to reuse its vocabulary [79]. However, biomedical NLP is more challenging than general NLP, because it involves jargon and abbreviations: Clinical notes have different syntax and grammar than books or encyclopedias. Moreover, a totally new vocabulary necessarily leads to training from scratch due to different vocabularies, which is computationally expensive.

Model size. Typically, big models usually have a bigger capacity that needs more data for training. However, the biomedical domain usually has as many corpora as the general domain. Thus, biomedical pre-trained language models are relatively smaller than general pre-trained language models. Another reason is that most of them are based on BERT or BERT-like encoder-based models, while pre-trained models with decoder architecture (e.g., GPT, T5) could be bigger than encoder-based pre-trained models. To the best of our knowledge, the biggest model is Biomegatron [243] with

1.2B parameters. Note that bigger models take longer for inference, which is unfriendly for those researchers without enough research computing resources.

Being publicly available. Thanks to the open source tradition of computer science, most models have web pages for downloading and documents for usage. Some of them standardized their model in huggingface (https://huggingface.co), which will largely be beneficial for widespread use. However, some models are not available to the public due to privacy issues even though data might have been anonymized [143].

Biomedical pre-trained language models in other languages. Most of the biomedical pre-trained language models are in English. However, there is an increasing need for biomedical pre-trained language models in other languages. There are typically two solutions: a multilingual solution or a purely second-language solution. The former may be beneficial for low-resource languages, and the latter is usually used in some rich-resource languages like Chinese [333].

3.5 Beyond Text: Biomedical Vision-and-Language Models

Biomedical data are inherently multi-modal. They include various types of data: text data, imaging data, tabular data, time-series data, and structured sequence data (e.g., proteins and DNA). Among them, the joint learning of text and imaging data are one of the most explored directions, and biomedical vision-and-language pre-training has emerged as an attractive direction in both artificial intelligence and clinical medicine. This is due to two facts: (i) From the technical perspective, computer vision and natural language processing have been the most popular directions in the past few years, and many models and algorithms have been proposed to process these two types of data, and (ii) from the data perspective, the text and imaging data are much easier to obtain in the medical domain, and, more importantly, they are always pair collected (e.g., radiology images and their corresponding diagnostic reports).

Most existing biomedical vision-and-language models are motivated by the success of the self-supervised pre-training recipe of SimCLR [43] in CV and BERT in NLP. Most recently, there have also been some studies [38, 39] applying the popular text-to-image diffusion models [219, 227, 231] to the medical domain. In this subsection, we summarize the existing biomedical vision-and-language models in Section 3.5.1 and describe them in detail.

3.5.1 Overview of Existing Biomedical Vision-and-Language Models. In biomedical vision-and-language pre-training, most existing studies could be categorized into two classes, i.e., dual-encoder and fusion encoder. These two types of models have different advantages and disadvantages. Dual-encoder models are able to capture the relationship between visual and linguistic elements in input by independently encoding each modality and then performing shallow iteration on the resulting vectors. This allows them to effectively learn representations that can be used for single-modal/cross-modal tasks, e.g., image classification, image captioning, and cross-modal retrieval. However, dual-encoder models are limited in their ability to fully capture the complex interactions between visual and linguistic elements, which can limit their performance on more challenging vision-and-language tasks.

However, fusion-encoder models aim to overcome this limitation by directly incorporating visual and linguistic elements into a single encoder. This allows them to capture more complex interactions between the two modalities, which can improve their performance on tasks that require a deeper understanding of the relationship between visual and linguistic elements. They jointly process these two modalities with an early interaction to learn multi-modal representations to solve those tasks requiring multi-modal reasoning, e.g., visual question answering. However, it can be

Table 5. Existing Biomedical Vision-and-language Pre-trained Models

Model	Date	Type	Image Encoder	Text Encoder	Fusion Module	Corpora	Downstream Datasets
UMRL [101]	2018.11	Dual-Encoder	DenseNet	GloVe	–	MIMIC-CXR	ICD-9-IT
ConVIRT [337]	2020.10	Dual-Encoder	ResNet	ClinicalBERT	–	MIMIC-CXR, RIH-BONE	CheXpert, COVIDx, MURA, RSNA
MulInfo [157]	2021.05	Dual-Encoder	ResNet	ClinicalBERT	–	MIMIC-CXR	Pathology9, EdemaSeverity
GLoRIA [105]	2021.10	Dual-Encoder	ResNet	BioClinicalBERT	–	CheXpert	CheXpert, RSNA, SIIM
LoVT [196]	2021.12	Dual-Encoder	ResNet	ClinicalBERT	–	MIMIC-CXR	COVID-Rural, NIH-CXR, Object CXR, SIIM
BioViL [26]	2022.04	Dual-Encoder	ResNet	CXR-BERT	–	MIMIC-CXR	MS-CXR, RSNA
BFSPR [236]	2022.05	Dual-Encoder	CLIP-Image	CLIP-Text	–	MIMIC-CXR	CheXpert, MIMIC-CXR, NIH-CXR, PadChest
CheXZero [265]	2022.09	Dual-Encoder	CLIP-Image	CLIP-Text	–	MIMIC-CXR	CheXpert, PadChest
MedCLIP [287]	2022.10	Dual-Encoder	ResNet/ViT	BioClinicalBERT	–	CheXpert, MIMIC-CXR	CheXpert, COVID, MIMIC-CXR, RSNA
MGCA [282]	2022.10	Dual-Encoder	ResNet/ViT	BioClinicalBERT	–	MIMIC-CXR	CheXpert, RSNA, SIIM
Analysis [195]	2022.11	Dual-Encoder	ResNet	ClinicalBERT	–	MIMIC-CXR	COVID-Rural, NIH-CXR, Object CXR, SIIM
Analysis [153]	2020.09	Fusion-Encoder	–	–	–	MIMIC-CXR	IU-Xray, MIMIC-CXR
Analysis [283]	2021.03	Fusion-Encoder	ResNet	BERT	Dual-Stream	MIMIC-CXR, NIH14-CXR, IU-Xray	MIMIC-CXR, NIH14-CXR, IU-Xray
Med-ViLL [190]	2021.05	Fusion-Encoder	ResNet	BERT	Single-Stream	MIMIC-CXR	MIMIC-CXR, IU-Xray, VQA-RAD
Berthop [189]	2021.08	Fusion-Encoder	ResNet	BlueBERT	Single-Stream	IU-Xray	IU-Xray
LViT [156]	2022.06	Fusion-Encoder	ViT	BERT	Single-Stream	QaTa-COV19, MoNuSeg	QaTa-COV19, MoNuSeg
M3AE [46]	2022.09	Fusion-Encoder	CLIP-Image	RoBERTa	Dual-Stream	MedICaT, ROCO	VQA-RAD, SLAKE, MedVQA-2019, MELINDA, ROCO
ARL [47]	2022.09	Fusion-Encoder	CLIP-Image	RoBERTa	Dual-Stream	MedICaT, MIMIC-CXR, ROCO	VQA-RAD, SLAKE, MedVQA-2019, MELINDA, ROCO

The date is based on the submission in arXiv or published data of the journal or conference proceeding.

more difficult to perform single-modal tasks, as the interactions between visual and linguistic elements are not as easily separated as they are in dual-encoder models. Table 5 presents existing dual-encoder and fusion-encoder vision-and-language models.

In addition to dual-encoder and fusion-encoder models, there are other approaches for biomedical vision-and-language pre-training. For example, motivated by the success of diffusion models [219, 227, 231] in the general domain, several medical text-to-image diffusion models [38, 39] have been proposed in the medical domains.

3.5.2 Dual-Encoder Vision-Language Models. Dual-encoder models encode images and texts separately to learn uni-modal/cross-modal representations following a shallow interaction layer (e.g., an image-text contrastive layer). The learned models can be transferred to many single-modal/cross-modal tasks, e.g., image classification and cross-modal retrieval tasks. Next, we detail some representative dual-encoder models:

- **ConVIRT** [337] is the first study to apply contrastive learning to images and texts, inspired by its success in the vision field. For the model architecture, it adopts ResNet and BERT as the vision encoder and the language encoder, respectively. Afterward, a bidirectional contrastive loss between two modalities is used to train these two encoders. It is found that the vision encoder can be used to perform the image classification tasks, requiring much fewer annotated training data as an ImageNet-initialized counterpart to achieve comparable or better performance.
- **GLoRIA** [105] proposed to perform the representation learning of medical images from global and local perspectives. Specifically, for global contrastive learning, it is similar to that of ConVIRT. For local contrastive learning, it uses an attention mechanism to learn local representations by matching the words in radiology reports and image sub-regions.
- **MedCLIP** [287] is trained on both image-text and image-label datasets. The core idea is to pre-compute the matching scores between an image and its text or an image and its label. Subsequently, the scores are used as the target to perform a learning procedure. It is observed that fewer data are required to the zero-shot disease classification.
- **CheXZero** [265] is initialized with the pre-trained CLIP model and pre-trained on the medical image-text dataset. With the strong backbone model and curated designs, CheXZero can achieve comparable results in disease classification tasks in a zero-shot manner.
- **LoVT** [196] is the first dual-encoder study targeting localized medical imaging tasks. It proposed a local contrastive loss to align local representations of sentences or image regions while encouraging spatial smoothness and sensitivity. This promotes its performance on many localized downstream tasks.

3.5.3 Fusion-Encoder Vision-Language Models. Fusion-encoder models encode images and texts and then exploit a fusion module to integrate the image and text features. For the fusion module, normally, there are two types: (i) single-stream, where the models use a single Transformer for early and unconstrained fusion between modalities, and (ii) dual-stream, the models adopt the co-attention mechanism to interact with different modalities. For fusion-encoder models, the most common objectives are masked language modeling and image-text matching. Similarly, we detail some representative fusion-encoder studies:

- **Li et al.** [153] adopted four general-domain pre-trained vision-and-language models (i.e., LXMERT [261], VisualBERT [150], UNITER [45], and PixelBERT [106]) to learn multi-modal representations from medical images and texts. The experimental results demonstrated their effectiveness of them in disease classification tasks.
- **MedViLL** [190] adopted a single BERT-based model and designed a masking scheme to improve both vision-language understanding tasks (e.g., disease classification, cross-modal retrieval, and visual question answering) and vision-language generation tasks (e.g., radiology report generation).
- **ARL** [47] proposed to integrate medical-domain knowledge bases (e.g., UMLS) into the fusion encoder. Medical knowledge is exploited from three perspectives: (i) aligning through knowledge, (ii) reasoning using knowledge, and (iii) learning from knowledge.
- **LViT** [156] is a vision-and-language fusion-encoder model for medical image segmentation. It leverages medical text annotation to improve the quality of generated segmentation results, especially in the semi-supervised setting.

3.5.4 Other Vision-Language Models. Besides the dual-encoder and fusion-encoder models, there are also some biomedical pre-trained models involving vision and language. We mainly introduce medical text-to-image diffusion models. Diffusion models are a type of generative model inspired by non-equilibrium thermodynamics. By defining a Markov chain of diffusion steps to add random noise to data slowly, the model aims to learn to reverse the diffusion process to construct desired data samples from the noise. Recently, different text-to-image diffusion models (e.g., DALLE-2 [219], Stable Diffusion [227], and Imagen [231]) have been proposed and achieved excellent performance on text-based image generation. In the medical domain, RoentGen [38, 39] investigated the adaptation of Stable Diffusion to the medical domain. In specific, they exploited chest X-ray images and their corresponding reports from the MIMIC-CXR dataset to train the model. Then they explored several adaptation approaches (i.e., partially fine-tuning or fully fine-tuning) and different text encoders for adaptation (e.g., domain-agnostic and domain-specific text encoder). The experiments demonstrated the effectiveness of the model with respect to image quality and clinical accuracy.

3.6 Beyond Text: Language Models for Proteins/DNA

Various biological sequences like proteins and DNA could also be treated like linguistic tokens in natural language. Therefore, many existing works explored training language models for these biological sequences. One crucial difference between language models for biological sequences and the counterparts for natural language is tokenization (see Section 3.6.1), which leads to different token vocabularies. Section 3.6.2 summarizes the existing language models for these biological sequences.

3.6.1 Tokenization for Proteins/DNAs. Like words in the text, biological sequences such as proteins and DNA sequences could also be modeled by language models, which typically aim to predict

the next token in a sequence. However, in contrast to that words are in a relatively big vocabulary (typically 10k–100k), and the vocabularies for biological sequences are usually small.

Tokenization in Proteins. Since the structure of a protein is fully determined by its amino acid sequence [10], one can represent a protein by its amino acid sequences. Roughly 500 amino acids have been identified in nature; however, only 20 amino acids are found to make up the proteins in the human body. The vocabulary of protein sequences consists of these 20 typical amino acids.

Tokenization in DNAs. The two DNA strands are known as polynucleotides, and they are composed of simpler monomeric units (a.k.a. nucleotides). Each nucleotide contains one of four nitrogen-containing nucleobases (i.e., cytosine [C], guanine [G], adenine [A], or thymine [T]). The two separate polynucleotides are bound together, according to deterministic base-pairing rules ([A] with [T] and [C] with [G]), with hydrogen bonds. Typically, existing work [109] usually adopts a "K-mer" representation for DNA sequences[13] for richer contextual information for DNAs. By doing so, the vocabulary size will increase to the $4^k + 5$, which is exponential to k and additionally pluses five special tokens ([CLS] , [SEP] , [PAD] , [MASK] , [UNK]).

3.6.2 *Language Models for Biological Sequences.*

Protein language models. Since the commonly found categories of amino acids are relatively small, namely 20. Initially, some work applied character-level language models to protein to deal with limited-size amino acids. In the beginning, there were many efforts to training RNN-based language models [7, 21] for protein sequences. References [92, 93] train a deep bi-directional model ELMo for proteins.[14] Other than those protein sequences, protein language models usually adopt additional features for proteins, e.g., global structural similarity between proteins and pairwise residue contact maps for each protein [21]. Later, Reference [222] introduces the Tasks Assessing Protein Embeddings, a suite of biologically relevant semi-supervised learning tasks. The authors also train language models based on LSTM, Transformer, and ResNet on the protein sequences. Bepler et al. [22] also proposed a novel framework based on the LSTM model to learn protein sequence embeddings. They have made their embeddings publicly available.[15] Reference [226] trains a contextual Transformer-based language model[16] on 250 million protein sequences. The representations learned by this LM encode multi-level information spanning from the biochemical properties of amino acids to the remote homology of proteins. Different from the above line of approaches, **multiple sequence alignments (MSA)** Transformer [223] fits a model separately to each family of proteins. ProtTrans [68] trains a variety of LM models with thousands of GPUs and also makes the trained models publicly available.[17] ProGen [174] is a generative LM trained on 280M protein sequences conditioned on taxonomic and keyword tags. ProteinLM [299] was recently proposed, which trained a large-scale pre-train model for evolutionary-scale protein sequences, and the trained model is available.[18] More recently, DeepMind developed Alphafold2 [116], which could predict protein structures with high accuracy in the challenging **14th Critical Assessment of protein Structure Prediction (CASP14)**. Most interestingly, there is an embedded protein language model in Alphafold2 that makes Alphafold2 feasible to make use of unlabelled protein data. In detail, Alphafold2 adopts an auxiliary BERT-like loss to predict pre-masked residues

[13] "K-mer" is like a k-size convolutional window for a sequence. For example, a DNA sequence ATGGCT will be tokenized to a sequence of 3-mers `{ATG TGC GGC GCT}` or to a sequence of 5-mers `{ATGGC TGGCT}`.
[14] https://github.com/Rostlab/SeqVec
[15] https://github.com/tbepler/protein-sequence-embedding-iclr2019
[16] The trained model and code are available at https://github.com/facebookresearch/esm
[17] https://github.com/agemagician/ProtTrans
[18] https://github.com/THUDM/ProteinLM

in MSAs. More recently, ProteinBERT [29] was proposed to use a novel task of Gene Ontology annotation prediction along with masked language modeling and it is also tailored to make the model highly efficient and flexible to very large sequence lengths.

DNA language models. Proteins are translated from DNA through the genetic code. There are 20 natural amino acids that are used to build the proteins that DNA encodes. Therefore, amino acids cannot be one-to-one mapped by only four nucleotides. Some work also explored the potential to build language models on DNA sequences. DNABERT [109] is a bidirectional encoder pre-trained on genomic DNA sequences with up- and downstream nucleotide contexts. Yamada et al. [311] pre-trains a BERT on RNA sequences and RNA-binding protein sequences. All the LMs remain largely the same as those used for human language data. Designing new architectures and pipelines tailored to protein/DNA sequences is a promising direction.

4 FINE-TUNING PLMS FOR BIOMEDICAL DOWNSTREAM TASKS

Similarly to the general domain, to evaluate the effectiveness and facilitate the research development of biomedical pre-trained language models, the **Biomedical Language Understanding Evaluation (BLUE)** benchmark has been proposed in Reference [204]. BLUE includes five text mining tasks in biomedical natural language processing, including sentence similarity, named entity recognition, relation extraction, text classification, and inference task. However, BLUE does not include some important biomedical application tasks, such as question answering, and it mixes the applications of clinical data and biomedical literature. To improve it, Gu et al. [79] proposed a novel benchmark, the **Biomedical Language Understanding & Reasoning Benchmark (BLURB)**. It includes NER, evidence-based medical information extraction (PICO), relation extraction, sentence similarity, document classification, and question-answering tasks. Moreover, some works proposed the benchmark in other languages, such as Chinese [333].

The development of biomedical pre-trained language models has greatly boosted the performance of these downstream tasks recently. In Table 6, we show the performance when directly fine-tuning different biomedical pre-trained language models for downstream tasks. All biomedical pre-trained language models significantly outperform PLMs in the general domain including BERT and RoBERTa. Especially for sentence similarity and question-answering tasks, the biomedical PLMs such as PubMedBERT and BioLinkBERT outperform BERT and RoBERTa by more than 10%. PubMedBERT conducts the domain-specific pre-training from scratch and consistently outperforms other biomedical PLMs such as BioBERT, ClinicalBERT, and BlueBERT in all tasks. Most recently, BioLinkBERT [318] further utilizes the citation links of documents from PubMed abstracts in the pre-training stage and has achieved the state-of-the-art (SOTA) performance on most tasks. Specifically, for the document level task such as the question-answering task, it outperforms PubMedBERT by 15% in the PubMedQA dataset and 4% in the BioASQ dataset. In the PubMedQA dataset and another document-level task, document classification, the BioGPT [170] achieves the new SOTA, which conducts the generative pre-training on PubMed abstracts from scratch like GPT.

Besides directly fine-tuning, there is other research exploring how to better leverage and improve PLMs for various downstream tasks. In the following, we will introduce the recent progress based on PLMs on these tasks (we show the example of each downstream task in Table 7) and other critical tasks in the biomedical domain.

4.1 Information Extraction

Information extraction plays a key role in automatically extracting structured biomedical information (entities, concepts, relations, and events) from unstructured biomedical text data ranging from biomedical literature and EHRs to biomedical-related social media corpus, and so on (see a latest

Table 6. The Performance of Different Biomedical Pre-trained Language Models on Downstream Tasks

	BERT	RoBERTa	BioBERT	SciBERT	ClinicalBERT	BlueBERT	PubMedBERT	BioM-ELECTRA	BioLinkBERT	BioGPT
NER										
BC5-chem [147]	89.99	89.43	92.85	92.51	90.80	91.19	93.33	93.1	**93.75**	–
BC5-disease [147]	79.92	80.65	84.70	84.70	83.04	83.69	85.62	85.2	**86.10**	–
NCBI-disease [62]	85.87	86.62	89.13	88.25	86.32	88.04	87.82	**88.4**	88.18	–
BC2GM [247]	81.23	80.90	83.82	83.36	81.71	81.87	84.52	–	**84.90**	–
JNLPBA [125]	77.51	77.86	78.55	78.51	78.07	77.71	**79.10**	–	79.03	–
PICO extraction										
EBM-PICO [201]	71.70	73.02	73.18	73.06	72.06	72.54	73.38	–	**73.97**	–
Relation extraction										
ChemProt [135]	71.54	72.98	76.14	75.00	72.04	71.46	77.24	**77.6**	77.57	–
DDI [96]	79.34	79.52	80.88	81.22	78.20	77.78	82.36	–	**82.72**	–
GAD [30]	79.61	80.63	82.36	81.34	80.48	79.15	83.96	–	**84.39**	–
Sentence similarity										
BIOSSES [248]	81.40	81.25	89.52	87.15	91.23	85.38	92.30	–	**93.25**	–
Document classification										
HoC [87]	80.12	79.66	81.54	81.16	80.74	80.48	82.32	–	84.35	**85.12**
Question answering										
PubMedQA [112]	49.96	52.84	60.24	51.40	49.08	48.44	55.84	–	70.20	**78.2**
BioASQ [199]	74.44	75.20	84.14	74.22	68.50	68.71	87.56	–	**91.43**	–
BLURB Score [79]	75.86	76.46	80.34	78.14	77.29	76.27	81.16	–	**83.39**	–

For all biomedical language models, we compare the F1 score of the base model on various tasks. The BLURB Score calculates the macro average of F1 test results on all tasks.

Table 7. Example for Each Downstream Task

Task	Input	Output	Example
Named Entity Recognition	Unannotated biomedical text	Annotated text with biomedical entities identified	E.g., identifying drug names, disease terms in text
Relation Extraction	Text with annotated entities	Text with relations between entities identified	E.g., recognizing drug-disease treatment relations
Event Extraction	Text with annotated entities and relations	Text with biomedical events identified	E.g., identifying gene-mutation-event in the literature
Text Classification	Biomedical text	Classified text into pre-defined categories	E.g., classifying medical reports based on disease types
Sentence Similarity	Pair of sentences	Similarity score between the sentences	E.g., measuring similarity between two medical findings
Question Answering	Question and context	Answer to the question based on context	E.g., answering clinical questions based on medical textbooks
Dialogue Systems	User input	System response	E.g., virtual health assistant responding to user health queries
Text Summarization	Long biomedical text	Short summary of the text	E.g., summarizing a medical research article
Natural Language Inference	Pair of sentences	Inference relation between the sentences	E.g., inferring medical conclusions from patient's symptoms

review in Reference [285] and study in Reference [85]). In the biomedical community, it generally refers to several important sub-tasks, including NER, relation extraction, and event extraction.

4.1.1 Named Entity Recognition. NER aims to identify the common biomedical concept mentions or entity names (such as genes, drug names, adverse effects, metabolites, and diseases) of biomedical texts. Singh et al. [230] proposed the first effort to investigate pre-training the bidirectional language model with the PubMed abstract dataset and then fine-tune the model for the supervised NER task. Compared with traditional neural network–based methods such as BiLSTM, it outperforms them by around 1% in the F1-score in datasets including NCBI-disease [62], BC5-disease [147], BC2GM [247], and JNLPBA [125] and requires less labelled training data to achieve comparable results. Several methods have shown that further pre-training the language models on the in-domain data can consistently improve the performance. For example, Zhu et al. [343][19] trained a domain-specific ELMo model in the mixture data of clinical reports and relevant Wikipedia pages that outperforms the previous SOTA method based on BiLSTM-CRF by 3.4% in F1-score in the i2b2 2010 [271] dataset. Si et al. [244] have shown that the BERT-large further pre-trained on the MIMIC-III achieves the best performance for the i2b2 2010 dataset and improves the performance by 5% over that of the traditional neural network method based on the GloVe embedding. Sheikhshab et al. [241] have shown that directly using the off-the-shelf ELMo embeddings has limited improvement on performance, while ELMo, continually pre-trained on the in-domain data, has significant improvement on the performance by 4% in the F1 score of the JNLPBA dataset. Gao et al. [73][20] investigated the pre-training and semi-supervised self-training

[19] https://github.com/noc-lab/clinical_concept_extraction
[20] https://code.ornl.gov/biomedner/biomedner

of BiLSTM-CRF and BlueBERT with the in-domain corpora such as MedMentions and SemMed. They evaluated these models on BioNER with limited labeled training data, and the BluBERT pre-trained on MedMentions has the best performance overall. Moreover, for the scenarios with very few labeled data, the semi-supervised self-training can significantly boost performance.

Some methods have explored how to utilize PLMs for BioNER with less time and computational consumption. Naseem et al. [198][21] proposed a lightweight domain-specific language model BioALBERT trained on the biomedical domain corpora for biomedical named entity recognition that captures inter-sentence coherence via the sentence-order-prediction task. For eight benchmark datasets, it outperforms the BioBERT by a significant margin, such as increasing the performance of the F1-score by 7.47% in the NCBI-disease dataset and 10.63% in the BC5CDR-disease dataset. Poerner et al. [211][22] proposed the time and memory saving domain-adaption method, training Word2Vec on target domain text and aligning them with the word vectors of existing PLMs, and thus propose the GreenBioBERT. On eight BioNER datasets, GreenBioBERT covers 60% results of BioBERT but only takes 2% of its cloud computing cost. Moreover, there are methods incorporating BioNER with the relation extraction task or modeling BioNER beyond the sequence labeling problem. Khan et al. [122] employed PLMs including BERT and BioBERT as the encoder for the multi-task learning of BioNER. They found that using BioBERT has moderately better performance than BERT, and it requires more training epochs for the BERT-based method to achieve comparable results. Giorgi et al. [76][23] proposed the end-to-end model for jointly extracting named entities and their relations using PLMs as the encoder. However, in the i2b2 2010 [271] dataset, it has worse performance than the method proposed by Si et al. [244] and BlueBERT. Sun et al. [259][24] proposed to model the BioNER as the machine reading comprehension problem to incorporate the prior knowledge flexibly and use PLMs as the text encoder. Among ClinicalBERT, BlueBERT, and BioBERT, the method based on BioBERT achieves the best performance. Tong et al. [267] proposed the auxiliary sentence-level prediction tasks, which can improve the F1 score by 3% in the low-resource scenario on three benchmark datasets compared with BioBERT. Banerjee et al. [14][25] formulated the BioNER as the knowledge-guided question-answer task, and it outperforms the SOTA by 1.78–12% in terms of *exact match F1* on 11 biomedical NER datasets.

4.1.2 Relation Extraction. **Biomedical relation extraction (BioRE)** aims to identify the relationship (semantic correlation) between biomedical entities mentioned (such as genes, proteins, and diseases) in texts and generally be considered as a classification problem to predict the possible relation type of two identified entities in a given sentence. Recently, PLMs have been widely explored in the BioRE. Wei et al. [289] conducted the first study that investigated fine-tuning BERT and combining additional BIO tag features for the clinical RE. It shows that the BERT-based model outperforms previous SOTA methods based on deep neural networks on n2c2 [95] and i2b2 [271] dataset. Similarly, Thillaisundaram et al. [263] adapted the SciBERT to the BioRE via fine-tuning the representation of the classification token (CLS). However, it only compared with and outperformed a simple sampling-based baseline. To further explore the potential of utilizing full information in the last layer to improve performance, Su et al. [255] proposed to utilize all outputs of the last layer when fine-tuning the BioBERT model on the BioRE task, which outperforms the BioBERT only using classification token on the DDI [96], PPI [134], and ChemProt [135] dataset. Su et al. [254] proposed to employ the contrastive learning to improve fine-tuning BERT model for

[21] https://github.com/usmaann/BioALBERT
[22] https://github.com/npoe/covid-qa
[23] https://github.com/bowang-lab/joint-ner-and-re
[24] https://github.com/CongSun-dlut/BioBERT-MRC
[25] https://github.com/kuntalkumarpal/KGQA

biomedical relation extraction, which outperforms directly fine-tuning BERT on the DDI, PPI and ChemProt datasets. Xue et al. [308] proposed to fine-tune BERT for joint entity and relation extraction in Chinese medical text, which outperforms the SOTA joint model based on Bi-LSTM by 1.6%. Chen et al. [44] combined BERT with the **one-dimensional convolutional neural network (1D-CNN)** for the medical relation extraction, which significantly outperforms the traditional 1D-CNN classifier. Lin et al. [160, 161] combined the global embeddings and multi-task learning to improve BERT on the clinical temporal relation extraction. Guan et al. [80] investigated several PLMs, including BERT, RoBERTa, ALBERT, XLNet, BioBERT, and ClinicalBERT, in predicting the relationships between clinical events and temporal expressions and found that RoBERTa generally has the best performance. To prevent private information leakage, Sui et al. [257] proposed the first privacy-preserving medical relation extraction method, FedED, based on BERT and federated learning, which achieved promising results on three benchmark datasets.

4.1.3 Event Extraction. Event extraction is another important task for mining structured knowledge from biomedical data, which aims to extract interactions between biological components (such as protein, gene, metabolic, drug, disease) and the consequences or effects of these interactions [9]. Similarly to BioRE, it is formulated into the multi-classification problem. Many efforts have been proposed to explore the application of PLMs in biomedical event extraction recently. Trieu et al. [268][26] proposed the model called DeepEventMine with the BERT-based encoder, which significantly outperforms the strong baseline based on CNN. Wadden et al. [276][27] explored combining the BERT model and graph propagation to capture long-range cross-sentence relationships, which have been proven to improve the performance of the model-based BERT alone. Ramponi et al. [221][28] modeled the biomedical event extraction as the sequence labeling problem, and proposed the model called BEESL with the BERT model as the encoder. It outperformed the baseline based on LSTM by 1.57% in the GENIA 2011 [126] benchmark. Wang et al. [284][29] formulated the biomedical event extraction as the multi-turn question-answering problem and utilized the question-answering system based on the SciBERT. The method can form event structures from the answers to multiple questions and achieves promising results on the GENIA 2011 [126] and Pathway Curation 2013 [212] datasets.

4.2 Text Classification

Text classification aims to classify biomedical texts into pre-defined categories, which play an important role in the statistical analysis, data management, retrieval of biomedical data, and so on. Fine-tuning pre-trained language models on biomedical text classification recently has attracted great attention. Gao et al. [72] investigated four methods of adapting the BERT model to handle input sequences up to approximately 400 words long for the clinical single-label and multi-label clinical document classification. However, they found that the BERT or BioBERT models generally have equal or worse performance for clinical data, such as the MIMIC-III clinical notes dataset, than a simple CNN model. They suggested that it may be because BERT or BioBERT models do not capture clinical domain knowledge, because they are trained on the general domain or biomedical literature datasets and cannot handle too-long sentences (longer than 512 tokens). Mascio et al. [176] made a comprehensive analysis of the performance of various word representation methods (such as Bag-of-Words, Word2Vec, GLoVe, FastText, BERT, and BioBERT) and classification approaches (Bi-LSTM, RNN, and CNN) on the electronic health records classification. They found that the

[26] https://github.com/aistairc/DeepEventMine
[27] https://github.com/dwadden/dygiepp
[28] https://github.com/cosbi-research/beesl
[29] https://github.com/WangXII/bio_event_qa

contextual embeddings from BERT and BioBERT generally outperform the traditional embeddings, and the traditional deep neural networks Bi-LSTM enriched with appropriate entity information and specific domain embeddings have better performance than BERT and BioBERT. Guo et al. [81] compared the performance of three PLMs including RoBERTa-base, BERTweet, and Clinical BioBERT on 25 social media classification datasets, in which six datasets are biomedical related. They found that RoBERTa-base and BERTweet outperform Clinical BioBERT, in which RoBERTa-base can capture general text semantic characteristics, while BERTweet captures more domain knowledge. Gutierrez et al. [84][30] also provided an analysis of traditional deep neural networks and fine-tuning PLMs including BERT and BioBERT on the performance of multi-label document classification on the COVID-19 dataset: LitCovid. They found that BERT and BioBERT models have better performance than traditional methods such as RNN, CNN, and Bi-LSTM in the datasets, and BioBERT outperforms BERT due to domain-specific pre-training.

4.3 Sentence Similarity

The semantic similarity task is generally formulated into the regression problem to predict the similarity score of each sentence pair. Recent works have focused on fine-tuning various PLMs for this task. Chen et al. [42][31] proposed the first pre-trained open set sentence embeddings in the biomedical domain called BioSentVec, which is trained on over 30 million documents from both biomedical literature such as PubMed and clinical notes such as the MIMIC-III Clinical Database. Compared with existing word embeddings and sentence encoder-based methods, it yields better performance on both sentence similarity and text classification tasks, due to better capturing the sentence-level semantic information. Chen et al. [40] empirically compared the performance of traditional deep learning methods such as random forest, RNN, and CNN with PLMs including BERT and BioBERT, which shows that PLMs are more effective. Chen et al. [41] further show the BioSentVec can improve the performance of traditional deep learning models by 2% F1 score. Yang et al. [316] explored three PLMs, including BERT, XLNet, and RoBERTa, for the clinical semantic textual similarity task, in which the XLNet achieves the best performance among the three models.

4.4 Question Answering

Biomedical question answering (BioQA) aims to extract or generate the natural language answers to the given questions and generally be formulated into the machine reading comprehension approach focusing on predicting the text span of answers with the given questions and passages containing the answers. Recently, the fine-tuning and transfer learning of PLMs have been widely explored in the task. Yoon et al. [320][32] applied the BioBERT to answer biomedical questions such as factoid, list, and Yes/No-type questions. They show that BioBERT fine-tuned with the question-answering datasets in both the general and biomedical domains and achieved the best performance in the 7th BioASQ Challenge. Jeong et al. [107][33] proposed to transfer the knowledge of **natural language inference (NLI)** to BioQA with BioBERT, which outperforms previous methods on Yes/No-, Factoid-, and List-type questions by 5.59%, 0.53%, and 13.58%, in the 7th BioASQ Challenge. Chakraborty et al. [37][34] proposed a novel language model, BioMedBERT, for **question answering (QA)** and information retrieval tasks, which is pre-trained on a large-scale biomedical literature dataset, BREATHE, based on BERT, and outperforms BERT in the BioQA. Kamath

[30] https://github.com/dki-lab/covid19-classification
[31] https://github.com/ncbi-nlp/BioSentVec
[32] https://github.com/dmis-lab/bioasq-biobert
[33] https://github.com/dmis-lab/bioasq8b
[34] https://github.com/BioMedBERT/biomedbert

et al. [120] compared the effectiveness of PLMs based on two different QA models, including the machine-reading comprehension and open question-answering method, and show that the question-answering model achieves better performance on the BioQA. Du et al. [66] utilized the BERT model as the encoder and then used the scaled dot-product attention mechanism to capture the interaction between the question and passage. The proposed method outperforms the best performance for factoid questions in 2016 and 2017 BioASQ-Task B. Zhou et al. [342] utilized the BioBERT and interactive Transformer model for both the recognizing question entailment and question answering task and showed significant improvements on the single task with the shared representations of both tasks. Similarly, Akdemir et al. [3] also explored multi-task learning to improve the performance of BioBERT on the BioQA task with the biomedical entity recognition task and show its improvements on the BioASQ 8B challenge. However, these models cannot detect multiple spans of the passage when there are multiple answers to the question. To solve the problem, Yoon et al. [319][35] reformulated the BioQA task as the sequence tagging problem to detect multiple entity spans simultaneously based on the BioBERT encoder, which achieves the BioASQ 7b and 8b list-type questions.

Some works tried to incorporate domain knowledge, such as biomedical-named entities, into PLMs. He et al. [91][36] proposed to infuse the domain knowledge of disease into a series of PLMs, including BERT, BioBERT, SciBERT, ClinicalBERT, BlueBERT, and ALBERT, to improve their performance. They found all these models can be improved by infusing the disease knowledge, and, for example, the accuracy of BioBERT on the CHQ dataset can be improved by nearly 4%. Rawat et al. [225][37] incorporated the medical entity information with entity embeddings and the auxiliary task on predicting the logical form of the question to improve the accuracy and generalization of the BERT model on answering questions, which improves the BERT model by 5% for the F1 score on the paraphrased question answering of the emrQA dataset. Kommaraju et al. [131] introduced the extra biomedical named entities prediction task to improve the BioBERT on Biomedical QA. They show that the BioBERT pre-trained by the prediction task outperforms the previous best model on the 7b-Phase B of the 7th BioASQ Task challenge.

Besides methods for biomedical literature corpora, other works have proposed question-answering models for unstructured EHRs. Soni et al. [250] investigated the performance of various PLMs, including BERT, BioBERT, ClinicalBERT, and XLNet, on clinical question answering and explored the fine-tuning methods with different datasets, including datasets in the general domain, biomedical, and clinical corpora. They find that fine-tuning the open-domain dataset SQuAD consistently improves the performance across all the model variants. Mairittha et al. [175] explored four different fine-tuned BERT models for personalized EHR question answering and show that the extended BioBERT-QA model pre-trained on unstructured EHR data achieves the best performance.

4.5 Dialogue Systems

The dialogue system aims to produce a proper response in either a selective [292, 338] or generative [167, 327, 340] way given a dialogue context for the biomedical goals of a user. The context includes historical utterances from users and systems, biomedical knowledge base, electronic health records of users, and so on. The format of a response could be various, e.g., a set of structured user goal data [290], a distribution of biomedical labels for diagnosis [163, 338], and natural language utterances [327]. For different types of contexts and responses, recent work focuses on end-to-end

[35] https://github.com/dmis-lab/SeqTagQA
[36] https://github.com/heyunh2015/diseaseBERT
[37] https://github.com/emrQA/bionlp_acl20

Dialogue System (DS) [306, 327] or parts of four typical DS modules, i.e., Natural Language Understanding [64, 242], Dialogue State Tracking [163, 290], Dialogue Policy Learning [290, 298], and Natural Language Generation [327]. Recently, PLMs are well known for natural language modeling, but it is nontrivial to pre-train on task datasets that are based on a specific domain [292]. To adapt PLMs to the medical domain, the dominant solution is to pre-train a language model on a large-scale general/medical corpus and then fine-tune the model with a medical dialogue dataset. Yan et al. [312][38] first explored fine-tuning PLMs, including BER-WWM, BERT-MED, MT5, and GPT2, on the M^2-MedDialog dataset for understanding the intents and slots of patients, in which MT5 achieves the best performance. Zeng et al. [327][39] pre-trained Transformer, BERT-GPT, and GPT on dialog datasets and other large-scale texts and then fine-tuned models on the Chinese MedDialog dataset for generating clinically correct and humanlike medical responses. BERT-GPT has been shown to have lower perplexity compared to both Transformer and GPT, while maintaining similar diversity metrics as Transformer. Shi et al. [242][40] show BERT has promising performance on the medical slot-filling task, and pre-trained embedding from BERT can further improve the performance of the weak supervision method. DialoGPT [340][41] is pre-trained based on GPT-2 [216] with a large in-domain dialogue dataset and is able to generate more relevant, informative and coherent responses compared with the strong baseline based on the sequence to sequence model. Li et al. [149][42] proposed the dialogue-adaptive pre-training objectives by considering dialogue-specific features, including coherence, specificity, and informativeness, which shows better performance than other language modeling objectives such as MLM and NSP. Different from using the accuracy, recall, and F1 score metrics used by previous tasks, the dialogue system task generally uses the machine translation metrics, including BLEU [340], METEOR [15], and NIST [61], to measure the similarity between generated responses and the ground truth based on n-gram matching. These metrics for evaluating generated responses are limited in that they only take into account shallow lexical overlaps and do not account for paraphrasing and terminology variations. To address this, some automatic metrics based on pre-trained language models have been developed, such as BERTScore [335], which calculates the similarity between two sentences using contextual embeddings from PLMs. However, these metrics have been shown to be inadequate in evaluating the faithfulness of generated responses. While there have been efforts to develop factual consistency metrics like BARTScore [324] in the general domain, there has been less focus on developing such metrics in the biomedical domain to evaluate factual correctness. Since the aforementioned methods utilized different datasets, it is hard to compare their performances directly. In summary, they have demonstrated that creating more effective pre-training tasks, incorporating task-specific information, and pre-training with large in-domain dialogue datasets are effective strategies for improving the performance of series PLMs.

4.6 Text Summarization

Automatic text summarization aims to automatically summarize the key information of single or multiple documents with shorter and more fluent texts, which greatly decreases the time-consuming of acquiring important information. Similarly to the general domain, existing methods can generally be classified into two categories: extractive summarization methods and abstractive summarization methods.

[38] https://github.com/yanguojun123/Medical-Dialogue
[39] https://github.com/UCSD-AI4H/Medical-Dialogue-System
[40] https://github.com/xmshi-trio/MSL
[41] https://github.com/microsoft/DialoGPT
[42] https://github.com/lockon-n/dapo

To explore the advanced PLMs in the text summarization of the biomedical domain, the domain knowledge is incorporated by existing methods via domain fine-tuning [173, 301]. For biomedical extractive summarization, Du et al. [65] proposed a novel model, BioBERTSum, which used the domain-aware pre-trained language model as the encoder and then fine-tuned it on the biomedical extractive summarization task. It outperforms SOTA extractive methods such as BERTSum. Xie et al. [300][43] proposed the knowledge infusion training framework to incorporate medical knowledge to improve a series of PLMs, including BERT, RoBERTa, BioBERT, and PubMedBERT. The PubMedBERT-based method has the best performance and outperforms other strong baselines such as BERTSum and MatchSum. Gharebagh et al. [75] utilized the domain knowledge: salient medical ontological terms to help the content selection of the SciBERT-based clinical abstractive summarization model, which improves SOTA results by around 2% in ROUGE metrics on two medical datasets, MIMIC-CXR [113] and OpenI [57]. Bishop et al. [23][44] proposed an unsupervised extractive summarization method for biomedical literature with T5 and BERTScore that achieves better performance than strong supervised methods such as BERTSum. Xie et al. [302][45] incorporated the neural topic model with hierarchical Transformer encoder based on PLMs, which significantly improved the performance of RoBERTa on long biomedical document summarization.

For abstractive summarization, Wallace et al. [277] utilized the BART as the encoder for generating biomedical evidence summary of multiple clinical trials. They found that the summarizers can produce fluent and relevant synopses, but the factual accuracy cannot be guaranteed. Deyoug et al. [60][46] investigated the BART model for the multi-document summarization on medical studies, which can generate coherent summaries that align with the reference summaries in evidence direction approximately 50% of the time. Guo et al. [82][47] proposed a novel task of plain language summarization task on the biomedical scientific reviews and explored pre-training BART model on general domain dataset CNN/DM and in-domain PubMed dataset. They found that BART pre-trained using CNN/DM and PubMed abstracts demonstrate the strongest ROUGE scores, whereas the BART model pre-trained only using PubMed abstracts has the lowest level of readability. Luo et al. [171][48] proposed the new task of readability controllable summarization for biomedical documents and explored the language model Longformer-Encoder-Decoder with advanced controllable techniques, including prompts and multi-head. They demonstrate that the method can generate fluent summaries, but it lacks the capability to effectively control for readability. Hu et al. [102] incorporated the additional knowledge with graph encoder and contrastive learning to enhance the performance of the BioBERT. The proposed method achieves state-of-the-art results in radiology report summarization. For the information acquisition of COVID-19-related scientific literature, Kieuvongngam et al. [124] proposed the BERT and GPT-2–based model for both extractive and abstractive summarization of COVID-19 research literature. There are some multi-document summarization systems for the information retrieval of COVID-19 research literature with the Siamese-BERT [69], BioBERT, and XLNet [253].

Similarly to the dialogue system task, the commonly used automatic metrics in the text summarization task, including ROUGE [162] and BERTScore [335], usually evaluate the relevance and similarity between the generated summaries and the gold summaries. Moreover, the factuality metrics have attracted much attention recently to evaluate the factual correctness of generated summaries [172, 301, 304]. Deyoung et al. [60] introduce the ΔEI metric to determine the degree of

[43] https://github.com/xashely/KeBioSum
[44] https://github.com/jbshp/gencomparesum
[45] https://github.com/xashely/GRETEL_extractive
[46] https://github.com/allenai/ms2/
[47] https://github.com/qiuweipku/Plain_language_summarization
[48] http://www.nactem.ac.uk/readability/

the factual accuracy of generated summaries in relation to the input medical studies. Zhang et al. [339] introduced the ChexBERT F1 score to evaluate the factual correctness between generated summaries and gold summaries of radiology reports.

4.7 Natural Language Inference

NLI, also known as text entailment, is a basic task for the natural language understanding of biomedical texts. It aims to infer the relation such as entailment, neutral, and contradiction, between two sentences, named the premise and hypothesis, which can further benefit biomedical downstream tasks such as commonsense comprehension, question answering, and evidence inference.

To facilitate the development of methods for text inference and entailment in the medical domain, participants in the MEDIQA 2019 shared task [2] investigated the SciBERT, BioBERT, and ClinicalBERT in the medical NLI task. Among these participants, Wu et al. [297][49] achieves the best performance with 98% accuracy in the REQ dataset [1], which ensembled results of different base models and incorporated the syntax information. Sharma et al. [240][50] incorporated the embedding of knowledge graph (UMLS) into the BioELMo to improve its performance, which shows an improvement of 0.8% regarding the accuracy to the base BioELMo model. Yadav et al. [310][51] a novel framework Sem-KGN for the medical textual entailment task, which infused the medical entity information from the medical knowledge bases into the BERT model. They show the medical knowledge information improves the SOTA language model ClinicalBERT by 8.27% on the REQ dataset. He et al. [91][52] proposed to infuse the domain knowledge of disease into a series of PLMs including BERT, BioBERT, SciBERT, ClinicalBERT, BlueBERT, and ALBERT, which improves performances of these models in all cases. Zhu et al. [344] utilized the neural architecture search to automatically find a better Transformer structure for language models, which improves the performance of the Chinese BERT-wwm-ext model [55] on two Chinese NLI datasets.

4.8 Proteins/DNAs Prediction

In this section, we only list some applications that have been well investigated or have potential, although there are much bigger spaces in biomedical domains to make use of PLMs.

4.8.1 Protein Structure Predictions. Proteins are essential to life, and knowing their structure can facilitate our understanding of their function. However, the structure of only a small fraction of proteins is known [116]. Predicting the three-dimensional structure of a protein is based solely on its amino acid sequence, a.k.a, "protein folding problem" [10]. To evaluate protein structure predictions, CASP uses proteins with recently solved structures that have not been deposited in the PDB or publicly disclosed; it, therefore, is a blind test for the participators, which is the gold-standard assessment for protein structure predictions [137, 192]. In CASP14, AlphaFold 2 [116], a model designed by DeepMind, achieves much better performance than other participating methods (e.g., template-based methods). The authors claim that AlphaFold 2 could provide precise estimates and could be confidently used for protein structure predictions with high reliability. However, predictions of existing methods, including AlphaFold 2, are more family specific than protein specific and rely on the evolutionary information captured in MSAs. To solve these issues, [346] proposed to use the attention head from the pre-trained protein language model ProtT5 without MSAs. Recently, Sturmfels et al. [252] presented a new biologically informed pre-training task: predicting

[49] https://github.com/ZhaofengWu/MEDIQA_WTMED
[50] https://github.com/soummyaah/KGMedNLI
[51] https://github.com/VishalPallagani/Medical-Knowledge-enriched-Textual-Entailment
[52] https://github.com/heyunh2015/diseaseBERT

protein profiles derived from multiple sequence alignments, which can improve the downstream protein structure prediction task.

4.8.2 DNA-related Applications. There are few works in DNA pre-training, among which DNABERT [109] is the representative one. DNABERT not only achieved SOTA performance on promoter prediction, splice sites, and transcription factor binding sites, but also identified functional genetic variants. Hong et al. [99] proposed to pre-train DNA vectors to encode enhancers and promoters and then incorporated the attention mechanism to predict long-range enhancer–promoter interactions. Yamada et al. [311] proposed a novel method based on BERT, in which BERT is pre-trained on the human reference genome, to predict the interactions between RNA sequences and RNA-binding proteins. Mock et al. [188] presented the BERTax based on BERT, for the taxonomic classification of DNA sequences.

5 DISCUSSION

5.1 Limitations and Concerns

In this subsection, we will mainly discuss the limitations of biomedical PLMs and raise some concerns about them.

Misinformation. The training corpora consist of EHR, and social media may include wrong information. Thus, pre-trained language models pre-trained on them may convey some misinformation [304]. Furthermore, the biomedical domain itself may have misclassified disease definitions during its development process. Misinformation has become much more serious in the biomedical domain than in the general domain, since this may lead to fatal biomedical decision-making consequences. However, researchers must be aware of the complexity of routinely collected electronic health records, including ways to manage variable completeness. We believe that the predictions from pre-trained language models should be artificially calibrated by biomedical experts before it is used by end-users like patients or the public.

Interpretation issues. Along with the power of neural networks, there is a growing concern about the interpretability of deep neural networks. While in the biomedical domain, the consequence of bad decisions/predictions may be deadly; thus, a well-interpreted model is more crucial. The interpretation in the biomedical domain may come from two aspects: (1) biomedical models should be easily understood, and the predictions could be simulated from the raw input, and (2) a (textual) reason should be provided for each prediction. The basic example of the former (a.k.a, transparency [164]) is decision trees that could clearly illustrate the decision path. However, such a transparency goal is hardly achieved in modern natural language processing, especially with pre-trained language models. More efforts could be made for the latter; one has to find some textual explanation for each prediction/decision, based on what doctors and patients could make their own decisions.

Identifying causalities from correlations. Similarly to interpretability, causality may provide the underlying explanation of the model decisions. Causality is crucial in many tasks of biomedical knowledge, e.g., diagnosis, pathology, or systems biology. Causal associations between biological entities, events, and processes are central to most claims of interest; see an early review from Reference [130]. With automatic causality recognition, it could suggest possible causal connections that may be beneficial for biomedical decisions, which hence greatly reduces the human workload [182].

Tradeoff between coverage or quality? There are no large-scaled and high-quality training corpora in the biomedical domain. This means one has to sacrifice its coverage to obtain a high-quality

vertical application or train a general model with large-scaled yet low-quality corpora. Pre-trained language models typically consist of many Transformer layers that have many parameters, which usually require a massive amount of plain text. This may lead to a general model with great coverage but a smaller proportion of high-quality expert knowledge.

Heterogeneous training data. For biomedical understandings, there is heterogeneous information, including tables, figures, graphs (fMRI), and so on. For example, tables and numbers are crucial in scientific literature. But most PLMs are unable to interpret tables and numbers well. To deeply capture the information in these heterogeneous data, both in-depth data prepossessing and model adaption may be needed. Especially, multi-domain pre-trained language models in biomedical should be paid much more attention.

Ethics and bias. With the rapid development of AI systems and applications in industrial products, it should be aware that they should not introduce any bias for special groups or populations [177], and some of the efforts were taken in the NLP field [24, 74, 260, 341]. This becomes more crucial in these sensitive environments in the biomedical domain that involves life-changing decisions, like surgery [229]. It should ensure that the decisions cannot reflect discriminatory or biased behavior toward specific groups or certain populations. A few works have quantified the ethics and bias issues in the domain of pre-trained language models. Reference [330] quantifies biases in clinical contextual word embeddings. The reason behind this may arise due to the training itself is biased with respect to various attributes like gender, language, race, age, ethnicity, and marital status. For example, in the MIMIC-III dataset [115], one can find (1) gender bias (males have more heart disease than females) and (2) ethnicity bias (black patients have fewer clinical studies than other groups) [118]. Considering the complexity of directly reducing biases in training corpora, existing works explore identifying bias by adversarial training [330] or data augmentation [185].

Privacy. Although most corpora used in biomedical pre-training like scientific publications and social medical are open access. Some EHRs are private, since some organizations do not want to expose their data. For example, clinical records may contain patient visits and medical history; these sensitive information may bring some physical and mental harm to patients if exposed [197]. Note that de-identification of these sensitive information in EHR records (like MIMIC III) is not always safe; recent works showed that there is data leakage from pre-trained models in the general domain, i.e., recovering **Personal Health Information (PHI)** from pre-trained models is possible [143]. Thus, we caution against the public release of pre-trained models if there's a risk of exposing PHI. Recently, Nakamura et al. [197] proposed a framework to assess the sensitive information from pre-trained biomedical language models using various attacks. Also, the federated learning [151, 313] framework may help when different organizations, and end-users could collaboratively learn a shared model while keeping training data private.

5.2 Future Trends

We suggest some future trends in this subsection.

Standardized benchmark. In general NLP fields, evaluation criteria and standard benchmarks are a driving force for the NLP community. For example, BERT [59] achieves excellent results on the GLUE and SQuAD benchmarks [218, 279], and these outstanding benchmark performances have made it shine across various NLP tasks. However, lacking an effective evaluation criterion is one of the bottlenecks of text generation [36]. In the biomedical domain, various pre-trained models and their fine-tuning applications have been proposed (as introduced in Section 3 and Section 4). However, they are generally not well compared. Although a few efforts have been made to standardize benchmarks for biomedical pre-trained models, which include but are not

limited to References [79, 332], this becomes much more difficult in a cross-discipline domain like the biomedical domain, since papers are usually from different communities like informatics, medicine, and computer science. An open standardized and well-categorized benchmark (like in References [145, 148]) should be proposed to make use of the advantages of each work and collaboratively push the development of biomedical NLP. This survey is the first step to introducing biomedical pre-trained language models and their applications in downstream tasks. More efforts are expected to be made to design fine-grained taxonomy and define each SOTA approach in various applications, based on what incremental work could be better evaluated.

Open culture. In general NLP fields, a lot of effort has been made to make better-available resources, including open source resources (released training data and models), and fairly implemented approaches. In addition, the open culture enables researchers to easily contribute to the community. For example, the NLP community has been largely developed thanks to the model collections [70, 294]. In addition, most accepted papers in top conferences tend to release codes, models, and data. Biomedical NLP fields also benefit greatly from such an open culture and standard systematic evaluations. For instance, pre-trained models in Huggingface[53] were used extensively in the biomedical domain.

Efficiency on pre-trained language models. Compared to previous SOTA methods training from scratch based on neural networks such as LSTM or CNN, before Transformer, pre-trained language models were much bigger in terms of model scale and much slower due to the increasing number of parameters. This is more expensive for deployment that requires more computing resources. One may have to refer to Reference [262] for efficient Transformers. For example, current work explores quantization [13, 336], weights pruning [100], and knowledge distillation [110, 232] for BERT. Therefore, in the biomedical domain, pre-training language models with lower computation complexity are a direction needed to pay more attention.

Generation-based PLMs are under-investigated. Most works focused on encoder-based models, and a few works involve decoder or encoder-decoder architectures that enable generations. This may be due to the fact that classification tasks may be widely used in downstream biomedical tasks. Very recently, Reference [133] proposed GPT models using temporal electronic health records, and Reference [210] trained a T5-based biomedical pre-trained model. We believe that generation-based PLMs (e.g., GPT, T5, and BART) have great potential in the biomedical domain, but it is currently under-investigated. Very recently, we have witnessed some work that uses large generation-based PLMs in the biomedical domain; see especially BIOGPT [170], PubMedGPT,[54] and Flan-PaLM [245].

Few-shot learning. Reference [207] evaluates the few-shot ability of LMs when held-out examples are unavailable for choosing hyperparameters or prompts. The study finds that LMs do not perform well compared to random selection and underperform when selections are based on held-out examples. In other words, previous methods overestimate the few-shot capability of LMs based on more realistic settings. This might be even worse for biomedical LMs.

In non-English or low-resource language. Most works in biomedical pre-trained language models use English corpora, and a few use Chinese [333], German [31], Japanese [121, 275], Spanish [4, 5, 169, 187], Korean [128], Russian [270], Italian [35], Arabic [12, 28], French [54], and Portuguese [234, 235]. For non-English biomedical tasks, there are two mainstream solutions: a

[53] https://huggingface.co/
[54] https://crfm.stanford.edu/2022/12/15/pubmedgpt.html

single non-English language paradigm and a multi-linguistic paradigm. The former uses a single language, while the latter uses multiple languages. The multi-linguistic paradigm could be more beneficial for low-resource languages, since biomedical knowledge itself is language independent, and information in a second language could be complementary.

Multi-modal pre-training. Multi-modal pre-training [215, 220, 278] has attracted much attention in image classification and generation tasks, because it only needs cheap but large-scale publicly available online resources. This shows great potential in machine learning, since less human annotation is needed. It is expected that various modalities could provide complementary information. For example, making use of biomedical codes, medical images, waveforms, and genomics in pre-training models would be beneficial but challenging due to its multi-modal nature.

Injecting biomedical knowledge in pre-trained language model. Before the pre-training age, some works [209] have explored injecting medical knowledge into embeddings that provide potentially better ML features. Recently, existing work claims that pre-trained language models could be a soft knowledge base that captures knowledge. Despite this, References [53, 307, 331] also tried to inject knowledge into pre-trained language models explicitly. In the biomedical domain, which is knowledge intensive, knowledge-injected models could have great potential in the future. For example, Reference [181] integrates domain knowledge (i.e., UMLS Metathesaurus) in pre-training via a knowledge augmentation strategy.

Interpretability in biomedical PLMs. Neural networks were criticized for having limited interpretability. Pre-trained language models are typically huge neural network models, which is more challenging in terms of interpretability. One may expect to understand the working mechanism related to the medical characteristics in pre-trained language models. For example, probing pre-trained language models have been widely used to understand pre-trained language models; see References [132, 159, 274, 296]. For biomedical pre-trained language models, Reference [6] aims to evaluate pre-trained language models about the disease knowledge. Reference [273] provides an exhaustive analysis of attention in protein Transformer models, offering numerous insightful findings to better understand their working mechanisms. Reference [111] conducts some probing experiments to determine what additional information is carried intrinsically by BioELMo and BioBERT. Another direction of interpretability in the biomedical field is to mine the causality (rather than correlation) due to its crucial relevance in establishing clinical interventions and public health policies. Correlation merely indicates a statistical relationship between two variables, which is valuable in generating hypotheses but provides limited insights into the underlying mechanisms. Conversely, causality moves beyond associative relationships to delineate direct cause–effect relationships. This deeper understanding is pivotal in biomedical research, as it provides the foundation for intervention studies and enables the development of effective treatments. Identifying a causal relationship, for instance, between a specific genetic mutation and a disease allows for targeted therapies and personalized medicine. Thus, while correlation provides a starting point for scientific exploration, it is the discernment of causality that truly advances biomedical knowledge and contributes to the development of life-saving interventions.

Dialogue-based medical consultation. Traditional medical consultation is used to obtain medical suggestions and treatment from clinicians. Recently, AI communities have tried to solve medical consultations through online ways using artificial intelligence tools, especially for pre-consultation and psychological treatment. Meanwhile, online medical consultation is another natural playground for current state-of-the-art AI algorithms under the "AI for science" trend. Some existing work formulates medical consultation as a question-answering task in the sense that it could leverage many existing question-answering pairs. However, medical consultation

is complicated in the sense that static and single-turn question-answering pairs could not solve individually dependent consultation; especially, medical consultations are more likely to be dependent on individual backgrounds, like historical diseases and treatment, genes, and dietary habits. We believe dialogue-based consultation systems could better fit medical scenarios than single-question-answering systems. Existing medical dialogue systems have shown some potential but also perform much worse than the expectation. Very recently, motivated by the great success of Open AI ChatGPT, which uses giant language models to meet human consultation needs, we believe using giant medical language models could largely improve the quality of medical consultation. More optimistically, we believe this might, at least to some extent, revolutionize the current medical industry; see References [139, 200] as some preliminary work.

Scale law in medical PLMs. Not only in dialogue systems, large-scale PLMs are as popular as it in the general domain. The reasons are twofold. First, the adaption of SOTA PLMs to the medical domain takes time, and it is usually more than half a year late after a general PLM is released. Second, non-generative language models are insensitive to huger scales, and their performance becomes saturated when they are beyond 24 layers (the scale of BERT-large). Meanwhile, most works use non-generative language models (e.g., BERT, RoBERTa, and Electra) in the biomedical domain, while very few generative language models are used. With larger language models, we might see some emergent abilities in medical applications. Fortunately, we have witnessed a preliminary sign that large language models are being used in medical/clinical tasks [158, 170, 245], especially BIOGPT [170], PubMedGPT,[55] and Flan-PaLM [245].

Data collection and sharing protocol. The need for data in biomedicine is tremendous, since data are the fuel for learning. The reasons that hinder medical data collection and sharing are manyfold. First, it has a legal risk regarding privacy issues, especially because this also involves cross-border or cross-organization data transfer. Second, an individual hospital might adopt different standards in terminology; this issue becomes more severe in developing countries than in developed countries. The merge between two data sources will be difficult due to the inconsistency of terminology. Therefore, it requires a well-defined protocol to deal with this, including solving terminology inconsistency and data privacy. From an NLP perspective, we need to normalize word terminology and data desensitization. For other perspectives, this needs some high-level data-sharing protocol, e.g., federated learning [314].

Dealing with long sequences. The computation of self-attentions in Transformers is quadratic to the length of sequences. This means the longer sequences would necessarily make Transformer-based PLMs much more time-consuming. Sequences in biomedicine are usually long; it varies from DNA/protein sequences to texts. First, DNA/protein sequences are long, especially for big protein sequences that have lengths that are longer than 4,096, i.e., the typical maximum sequence length in language models. Biomedical texts, including EHRs, biomedical encyclopedias, and biomedical literature, are usually longer than the general domain (e.g., the maximum sentence length used in GLUE is usually 128); for instance, there is usually text redundancy in clinical notes. Therefore, we need to design more efficient and effective models tailored to long sequences; see some existing recent works [154, 155].

6 CONCLUSION

This article systematically summarizes recent advances of pre-trained language models in the biomedical domain, including background, why and how pre-trained language models are used

[55] https://crfm.stanford.edu/2022/12/15/pubmedgpt.html

in the biomedical domain, existing biomedical pre-trained language models, data sources in the biomedical domain, and the application of pre-trained language models in various biomedical downstream tasks. Furthermore, we also discuss some limitations and future trends. Finally, we expect that the pre-trained language model in the general NLP domain could also help the specific biomedical domain.

REFERENCES

[1] Asma Ben Abacha and Dina Demner-Fushman. 2016. Recognizing question entailment for medical question answering. In *AMIA Annual Symposium Proceedings*, Vol. 2016. American Medical Informatics Association, 310.

[2] Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *BioNLP Workshop and Shared Task*. 370–379.

[3] Arda Akdemir and Tetsuo Shibuya. 2020. Transfer learning for biomedical question answering. In *CLEF (Working Notes)*.

[4] Liliya Akhtyamova. 2020. Named entity recognition in spanish biomedical literature: Short review and bert model. In *FRUCT*. IEEE, 1–7.

[5] Liliya Akhtyamova, Paloma Martínez, Karin Verspoor, and John Cardiff. 2020. Testing contextualized word embeddings to improve ner in spanish clinical case narratives. *IEEE Access* 8 (2020), 164717–164726.

[6] Israa Alghanmi, Luis Espinosa-Anke, and Steven Schockaert. 2021. Probing pre-trained language models for disease knowledge. arXiv:2106.07285. Retrieved from https://arxiv.org/abs/2106.07285

[7] Ethan C. Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M. Church. 2019. Unified rational protein engineering with sequence-only deep representation learning (unpublished).

[8] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical BERT embeddings. arXiv:1904.03323. Retrieved from http://arxiv.org/abs/1904.03323

[9] Sophia Ananiadou, Sampo Pyysalo, Jun'ichi Tsujii, and Douglas B. Kell. 2010. Event extraction for systems biology by text mining the literature. *Trends Biotechnol.* 28, 7 (2010), 381–390.

[10] Christian B. Anfinsen. 1973. Principles that govern the folding of protein chains. *Science* 181, 4096 (1973), 223–230.

[11] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. arXiv:2305.10403. Retrieved from https://arxiv.org/abs/2305.10403

[12] Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for Arabic language understanding. arXiv:2003.00104. Retrieved from https://arxiv.org/abs/2003.00104

[13] Haoli Bai, Wei Zhang, Lu Hou, Lifeng Shang, Jing Jin, Xin Jiang, Qun Liu, Michael Lyu, and Irwin King. 2020. Binarybert: Pushing the limit of bert quantization. arXiv:cs.CL/2012.15701. Retrieved from https://arxiv.org/abs/2012.15701

[14] Pratyay Banerjee, Kuntal Kumar Pal, Murthy Devarakonda, and Chitta Baral. 2021. Biomedical named entity recognition via knowledge guidance and question answering. *ACM Trans. Comput. Healthc.* 2, 4 (2021), 1–24.

[15] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. 65–72.

[16] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. BEiT: BERT pre-training of image transformers. In *ICLR*.

[17] Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. Cometa: A corpus for medical entity linking in the social media. arXiv:2010.03295. Retrieved from https://arxiv.org/abs/2010.03295

[18] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *EMNLP-IJCNLP*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3613–3618. https://doi.org/10.18653/v1/D19-1371

[19] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. arXiv:2004.05150. Retrieved from https://arxiv.org/abs/2004.05150

[20] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.* 3 (Feb. 2003), 1137–1155.

[21] Tristan Bepler and Bonnie Berger. 2018. Learning protein sequence embeddings using information from structure. In *ICLR*.

[22] Tristan Bepler and Bonnie Berger. 2019. Learning protein sequence embeddings using information from structure. arXiv:1902.08661. Retrieved from https://arxiv.org/abs/1902.08661

[23] Jennifer Bishop, Qianqian Xie, and Sophia Ananiadou. 2022. GenCompareSum: A hybrid unsupervised summarization method using salience. In *Proceedings of the 21st Workshop on Biomedical Language Processing*. 220–240.

[24] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. arXiv:2005.14050. Retrieved from https://arxiv.org/abs/2005.14050

[25] Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucl. Acids Res.* 32, suppl_1 (2004), D267–D270.

[26] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. 2022. Making the most of text semantics to improve biomedical vision–language processing. arXiv:2204.09817. Retrieved from https://arxiv.org/abs/2204.09817

[27] Rishi Bommasani and et al. 2021. On the opportunities and risks of foundation models. arXiv:cs.LG/2108.07258. Retrieved from https://arxiv.org/abs/2103.07258

[28] Nada Boudjellal, Huaping Zhang, Asif Khan, Arshad Ahmad, Rashid Naseem, Jianyun Shang, and Lin Dai. 2021. Abioner: A bert-based model for arabic biomedical named-entity recognition (unpublished).

[29] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. 2022. ProteinBERT: A universal deep-learning model of protein sequence and function. *Bioinformatics* 38, 8 (2022), 2102–2110.

[30] Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I. Furlong. 2015. Extraction of relations between genes and diseases from text and large-scale data analysis: Implications for translational research. *BMC Bioinf.* 16, 1 (2015), 1–17.

[31] Keno K. Bressem, Lisa C. Adams, Robert A. Gaudin, Daniel Tröltzsch, Bernd Hamm, Marcus R. Makowski, Chan-Yong Schüle, Janis L. Vahldiek, and Stefan M. Niehues. 2020. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. *Bioinformatics* 36, 21 (2020), 5255–5261.

[32] Eric Brochu, Vlad M. Cora, and Nando De Freitas. 2010. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv:1012.2599. Retrieved from https://arxiv.org/abs/1012.2599

[33] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. arXiv:2005.14165. Retrieved from https://arxiv.org/abs/2005.14165

[34] Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *ACL*. 1860–1874.

[35] Rosario Catelli, Francesco Gargiulo, Valentina Casola, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2020. Crosslingual named entity recognition for clinical de-identification applied to a covid-19 italian data set. *Appl. Soft Comput.* 97 (2020), 106779.

[36] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. arXiv:2006.14799. Retrieved from https://arxiv.org/abs/2006.14799

[37] Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco Mosconi. 2020. Biomedbert: A pre-trained biomedical language model for qa and ir. In *ICCL*. 669–679.

[38] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P. Langlotz, and Akshay Chaudhari. 2022. RoentGen: Vision-language foundation model for chest x-ray generation. arXiv:2211.12737. Retrieved from https://arxiv.org/abs/2211.12737

[39] Pierre Chambon, Christian Bluethgen, Curtis P. Langlotz, and Akshay Chaudhari. 2022. Adapting pretrained vision-language foundational models to medical imaging domains. arXiv:2210.04133. Retrieved from https://arxiv.org/abs/2210.04133

[40] Qingyu Chen, Jingcheng Du, Sun Kim, W. John Wilbur, and Zhiyong Lu. 2019. Evaluation of five sentence similarity models on electronic medical records. In *ACM-BCB*. 533–533.

[41] Qingyu Chen, Jingcheng Du, Sun Kim, W. John Wilbur, and Zhiyong Lu. 2020. Deep learning with sentence embeddings pre-trained on biomedical corpora improves the performance of finding similar sentences in electronic medical records. *BMC Med. Inf. Decis. Mak.* 20 (2020), 1–10.

[42] Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. Biosentvec: Creating sentence embeddings for biomedical texts. In *ICHI*. IEEE, 1–5.

[43] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*. PMLR, 1597–1607.

[44] Tao Chen, Mingfen Wu, and Hexi Li. 2019. A general approach for improving deep learning-based medical relation extraction using a pre-trained model and fine-tuning. *Database* (2019), baz116.

[45] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*. Springer, 104–120.

[46] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. 2022. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 679–689.
[47] Zhihong Chen, Guanbin Li, and Xiang Wan. 2022. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *Multimedia*. 5152–5161.
[48] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. arXiv:2204.02311. Retrieved from https://arxiv.org/abs/2204.02311
[49] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. arXiv:cs.CL/2003.10555. Retrieved from https://arxiv.org/abs/2003.10555
[50] Aaron M. Cohen and William R. Hersh. 2005. A survey of current work in biomedical text mining. *Brief. Bioinf.* 6, 1 (2005), 57–71.
[51] Jacques Cohen. 2004. Bioinformatics—An introduction for computer scientists. *ACM Comput. Surv.* 36, 2 (Jun. 2004), 122–158. https://doi.org/10.1145/1031120.1031122
[52] Kevin Bretonnel Cohen and Dina Demner-Fushman. 2014. *Biomedical Natural Language Processing*. Vol. 11. John Benjamins Publishing Company.
[53] Pedro Colon-Hernandez, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal. 2021. Combining pre-trained language models and structured knowledge. arXiv:2101.12294. Retrieved from https://arxiv.org/abs/2101.12294
[54] Jenny Copara, Julien Knafou, Nona Naderi, Claudia Moro, Patrick Ruch, and Douglas Teodoro. 2020. Contextualized french language models for biomedical named entity recognition. In *JEP/TALN/RÉCITAL*. ATALA, 36–48.
[55] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. arXiv:1906.08101. Retrieved from https://arxiv.org/abs/1906.08101
[56] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inf. Assoc.* 23, 2 (2016), 304–310.
[57] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inf. Assoc.* 23, 2 (2016), 304–310.
[58] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*. IEEE, 248–255.
[59] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. Retrieved from https://arxiv.org/abs/1810.04805
[60] Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Wang. 2021. MS^2: Multi-document summarization of medical studies. In *EMNLP*. 7494–7513.
[61] George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *HLTR*. 138–145.
[62] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *J. Biomed. Inf.* 47 (2014), 1–10.
[63] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. 2021. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. arXiv:2103.03404. Retrieved from https://arxiv.org/abs/a2103.03404
[64] Nan Du, Mingqiu Wang, Linh Tran, Gang Li, and Izhak Shafran. 2020. Learning to infer entities, properties and their relations from clinical conversations. In *EMNLP-IJCNLP*. Association for Computational Linguistics, 4978–4989.
[65] Yongping Du, Qingxiao Li, Lulin Wang, and Yanqing He. 2020. Biomedical-domain pre-trained language model for extractive summarization. *Knowl.-Bas. Syst.* 199 (2020), 105964.
[66] Yongping Du, Bingbing Pei, Xiaozheng Zhao, and Junzhong Ji. 2020. Deep scaled dot-product attention based domain adaptation model for biomedical question answering. *Methods* 173 (2020), 69–74.
[67] Marco Eichelberg, Thomas Aden, Jörg Riesmeier, Asuman Dogac, and Gokce B. Laleci. 2005. A survey and analysis of electronic healthcare record standards. *ACM Comput. Surv.* 37, 4 (Dec. 2005), 277–315. https://doi.org/10.1145/1118890.1118891
[68] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. 2020. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. arXiv:2007.06225. Retrieved from https://arxiv.org/abs/2007.06225
[69] Andre Esteva, Anuprit Kale, Romain Paulus, Kazuma Hashimoto, Wenpeng Yin, Dragomir Radev, and Richard Socher. 2020. Co-search: COVID-19 information retrieval with semantic search, question answering, and abstractive summarization. arXiv:2006.09595. Retrieved from https://arxiv.org/abs/2006.09595

[70] Yixing Fan, Liang Pang, JianPeng Hou, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2017. Matchzoo: A toolkit for deep text matching. arXiv:1707.07270. Retrieved from https://arxiv.org/abs/1707.07270

[71] Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, et al. 2020. Codebert: A pre-trained model for programming and natural languages. arXiv:2002.08155. Retrieved from https://arxiv.org/abs/2002.08155

[72] Shang Gao, Mohammed Alawad, Michael Todd Young, John Gounley, Noah Schaefferkoetter, Hong-Jun Yoon, Xiao-Cheng Wu, Eric B. Durbin, Jennifer Doherty, Antoinette Stroup, et al. 2021. Limitations of transformers on clinical text classification. *IEEE J. Biomed. Health Inf.* (2021).

[73] Shang Gao, Olivera Kotevska, Alexandre Sorokine, and J. Blair Christian. 2021. A pre-training and self-training approach for biomedical named entity recognition. *PLoS ONE* 16, 2 (2021), e0246310.

[74] Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L. Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Appl. Sci.* 11, 7 (2021), 3184.

[75] Sajad Sotudeh Gharebagh, Nazli Goharian, and Ross Filice. 2020. Attend to medical ontologies: Content selection for clinical abstractive summarization. In *ACL*. 1899–1905.

[76] John Giorgi, Xindi Wang, Nicola Sahar, Won Young Shin, Gary D. Bader, and Bo Wang. 2019. End-to-end named entity recognition and relation extraction using pre-trained language models. arXiv:1912.13415. Retrieved from https://arxiv.org/abs/1912.134.15

[77] Graciela Gonzalez-Hernandez, Abeed Sarker, Karen O'Connor, and Guergana Savova. 2017. Capturing the patient's perspective: A review of advances in natural language processing of health-related text. *Yrbk. Med. Inf.* 26, 1 (2017), 214.

[78] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* 33 (2020), 21271–21284.

[79] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* 3, 1 (2021), 1–23.

[80] Hong Guan, Jianfu Li, Hua Xu, and Murthy Devarakonda. 2020. Robustly pre-trained neural model for direct temporal relation extraction. arXiv:2004.06216. Retrieved from https://arxiv.org/abs/2004.06216

[81] Yuting Guo, Xiangjue Dong, Mohammed Ali Al-Garadi, Abeed Sarker, Cécile Paris, and Diego Mollá Aliod. 2020. Benchmarking of transformer-based pre-trained models on social media text classification datasets. In *Workshop of the Australasian Language Technology Association*. 86–91.

[82] Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2020. Automated lay language summarization of biomedical scientific reviews. arXiv:2012.12573. Retrieved from https://arxiv.org/abs/2012.12573

[83] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 8342–8360. https://doi.org/10.18653/v1/2020.acl-main.740

[84] Bernal Jimenez Gutierrez, Jucheng Zeng, Dongdong Zhang, Ping Zhang, and Yu Su. 2020. Document classification for covid-19 literature. In *EMNLP: Findings*. 3715–3722.

[85] Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by ChatGPT? An analysis of performance, evaluation criteria, robustness and errors. arXiv:2305.14450. Retrieved from https://arxiv.org/abs/2305.14450

[86] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre-trained models: Past, present and future. arXiv:cs.AI/2106.07139. Retrieved from https://arxiv.org/abs/2106.07139

[87] Douglas Hanahan and Robert A. Weinberg. 2000. The hallmarks of cancer. *Cell* 100, 1 (2000), 57–70.

[88] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *CVPR*. 16000–16009.

[89] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*. 9729–9738.

[90] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[91] Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020. Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition. In *EMNLP*. 4604–4614.

[92] Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. 2019. Modeling the language of life–deep learning protein sequences (unpublished).

[93] Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. 2019. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinf.* 20, 1 (2019), 1–17.

[94] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). arXiv:1606.08415. Retrieved from https://arxiv.org/abs/1606.08415

[95] Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *J. Am. Med. Inf. Assoc.* 27, 1 (2020), 3–12.

[96] María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *J. Biomed. Inf.* 46, 5 (2013), 914–920.

[97] Emily Herrett, Arlene M. Gallagher, Krishnan Bhaskaran, Harriet Forbes, Rohini Mathur, Tjeerd Van Staa, and Liam Smeeth. 2015. Data resource profile: Clinical practice research datalink (cprd). *Int. J. Epidemiol.* 44, 3 (2015), 827–836.

[98] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9, 8 (1997), 1735–1780.

[99] Zengyan Hong, Xiangxiang Zeng, Leyi Wei, and Xiangrong Liu. 2020. Identifying enhancer–promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics* 36, 4 (2020), 1037–1043.

[100] Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. Dynabert: Dynamic bert with adaptive width and depth. arXiv:cs.CL/2004.04037. Retrieved from https://arxiv.org/abs/2004.04037

[101] Tzu-Ming Harry Hsu, Wei-Hung Weng, Willie Boag, Matthew McDermott, and Peter Szolovits. 2018. Unsupervised multimodal representation learning across medical images and reports. arXiv:1811.08615. Retrieved from https://arxiv.org/abs/1811.08615

[102] Jinpeng Hu, Zhuo Li, Zhihong Chen, Zhen Li, Xiang Wan, and Tsung-Hui Chang. 2022. Graph enhanced contrastive learning for radiology findings summarization. In *ACL (Volume 1: Long Papers)*. 4677–4688.

[103] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv:1904.05342. Retrieved from http://arxiv.org/abs/1904.05342

[104] Kexin Huang, Abhishek Singh, Sitong Chen, Edward Moseley, Chin ying Deng, Naomi George, and Charlotta Lindvall. 2019. Clinical xlnet: Modeling sequential clinical notes and predicting prolonged mechanical ventilation. arXiv:1912.11975. Retrieved from https://arxiv.org/abs/1912.11975

[105] Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *IEEE/CVF ICCV*. 3942–3951.

[106] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. arXiv:2004.00849. Retrieved from https://arxiv.org/abs/2004.00849

[107] Minbyul Jeong, Mujeen Sung, Gangwoo Kim, Donghyeon Kim, Wonjin Yoon, Jaehyo Yoo, and Jaewoo Kang. 2020. Transferability of natural language inference to biomedical question answering. arXiv:2007.00217. Retrieved from https://arxiv.org/abs/2007.00217

[108] Kishlay Jha and Aidong Zhang. 2022. Continual knowledge infusion into pre-trained biomedical language models. *Bioinformatics* 38, 2 (2022), 494–502.

[109] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V. Davuluri. 2020. Dnabert: Pre-trained bidirectional encoder representations from transformers model for dna-language in genome (unpublished).

[110] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. arXiv:cs.CL/1909.10351. Retrieved from https://arxiv.org/abs/1909.10351

[111] Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. In *Workshop on Evaluating Vector Space Representations for NLP*. 82–89.

[112] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *EMNLP-IJCNLP*. 2567–2577.

[113] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* 6, 1 (2019), 317.

[114] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv:1901.07042. Retrieved from https://arxiv.org/abs/1901.07042

[115] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, H. Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Sci. Data* 3, 1 (2016), 1–9.

[116] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with alphafold. *Nature* (2021), 1.

[117] Dan Jurafsky. 2000. *Speech & Language Processing*. Pearson Education India.

[118] Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammu—A survey of transformer-based biomedical pretrained language models. arXiv:2105.00827. Retrieved from https://arxiv.org/abs/2105.00827

[119] Katikapalli Subramanyam Kalyan and S. Sangeetha. 2020. Secnlp: A survey of embeddings in clinical natural language processing. *J. Biomed. Inf.* 101 (2020), 103323.

[120] Sanjay Kamath, Brigitte Grau, and Yue Ma. 2019. How to pre-train your model? comparison of different pre-training models for biomedical question answering. In *ECML-PKDD*. Springer, 646–660.

[121] Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, Eiji Aramaki, and Kazuhiko Ohe. 2020. A clinical specific bert developed with huge size of japanese clinical narrative (unpublished).

[122] Muhammad Raza Khan, Morteza Ziyadi, and Mohamed AbdelHady. 2020. Mt-bioner: Multi-task learning for biomedical named entity recognition using deep bidirectional transformers. arXiv:2001.08904. Retrieved from https://arxiv.org/abs/2001.08904

[123] Faiza Khan Khattak, Serena Jeblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, and Frank Rudzicz. 2019. A survey of word embeddings for clinical text. *J. Biomed. Inf. X* 4 (2019), 100057.

[124] Virapat Kieuvongngam, Bowen Tan, and Yiming Niu. 2020. Automatic text summarization of covid-19 medical research articles using bert and gpt-2. arXiv:2006.01997. Retrieved from https://arxiv/org/abs/2006.01997

[125] Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at jnlpba. In *NLPBA/BioNLP*. Citeseer, 70–75.

[126] Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011. Overview of genia event task in bionlp shared task 2011. In *BioNLP Shared Task 2011 Workshop*. 7–15.

[127] Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*. 1746–1751.

[128] Young-Min Kim and Tae-Hoon Lee. 2020. Korean clinical entity recognition from diagnosis text using bert. *BMC Med. Inf. Decis. Mak.* 20, 7 (2020), 1–9.

[129] David T. Kingsbury. 1996. Computational biology. *ACM Comput. Surv.* 28, 1 (Mar. 1996), 101–103. https://doi.org/10.1145/234313.234358

[130] Samantha Kleinberg and George Hripcsak. 2011. A review of causal inference for biomedical informatics. *J. Biomed. Inf.* 44, 6 (2011), 1102–1112.

[131] Vaishnavi Kommaraju, Karthick Gunasekaran, Kun Li, Trapit Bansal, Andrew McCallum, Ivana Williams, and Ana-Maria Istrate. 2020. Unsupervised pre-training for biomedical question answering. arXiv:2009.12952. Retrieved from https://arxiv.org/abs/2009.12952

[132] Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. arXiv:2104.05882. Retrieved from https://arxiv.org/abs/2104.05882

[133] Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James Teo, and Richard Dobson. 2021. Medgpt: Medical concept prediction from clinical narratives. arXiv:2107.03134. Retrieved from https://arxiv.org/abs/2107.03134

[134] Martin Krallinger, Florian Leitner, Carlos Rodriguez-Penagos, and Alfonso Valencia. 2008. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biol.* 9, 2 (2008), 1–19.

[135] Martin Krallinger, Obdulia Rabal, Saber A. Akhondi, Martın Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, et al. 2017. Overview of the biocreative vi chemical-protein interaction track. In *BioCreative Challenge Evaluation Workshop*, Vol. 1. 141–146.

[136] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.

[137] Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. 2019. Critical assessment of methods of protein structure prediction (CASP)–Round XIII. *Proteins: Struct. Funct. Bioinf.* 87, 12 (2019), 1011–1020.

[138] Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *EMNLP: System Demonstrations*. 66–71.

[139] Tiffany H. Kung, Morgan Cheatham, Arielle Medinilla, ChatGPT, Czarina Sillos, Lorie De Leon, Camille Elepano, Marie Madriaga, Rimel Aggabao, Giezel Diaz-Candido, et al. 2022. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models (unpublished).

[140] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. arXiv:1909.11942. Retrieved from https://arxiv.org/abs/1909.11942

[141] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[142] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

[143] Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron C. Wallace. 2021. Does bert pretrained on clinical notes reveal sensitive data? arXiv:2104.07762. Retrieved from https://arxiv.org/abs/2104.07762

[144] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461. Retrieved from https://arxiv.org/abs/1910.13461

[145] Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Clinical Natural Language Processing Workshop*. 146–157.

[146] Fei Li, Yonghao Jin, Weisong Liu, Bhanu Pratap Singh Rawat, Pengshan Cai, and Hong Yu. 2019. Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: An empirical study. *JMIR Med. Inf.* 7, 3 (2019), e14830.

[147] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: A resource for chemical disease relation extraction (unpublished).

[148] Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023. Huatuo-26M, a large-scale chinese medical QA dataset. arXiv:2305.01526. Retrieved from https://arxiv.org/abs/2305.01526

[149] Junlong Li, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiang Zhou. 2020. Task-specific objectives of pre-trained language models for dialogue adaptation. arXiv:2009.04984. Retrieved from https://arxiv.org/abs/2009.04984

[150] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. arXiv:1908.03557. Retrieved from https://arxiv.org/abs/1908.03557

[151] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated learning: Challenges, methods, and future directions. *IEEE Sign. Process. Mag.* 37, 3 (2020), 50–60.

[152] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. Behrt: Transformer for electronic health records. *Sci. Rep.* 10, 1 (2020), 1–12.

[153] Yikuan Li, Hanyin Wang, and Yuan Luo. 2020. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In *BIBM*. IEEE, 1999–2004.

[154] Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-BigBird: Transformers for long clinical sequences. arXiv:2201.11838. Retrieved from https://arxiv.org/abs/2201.11838

[155] Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. 2023. A comparative study of pretrained language models for long clinical text. *J. Am. Med. Inf. Assoc.* 30, 2 (2023), 340–347.

[156] Zihan Li, Yunxiang Li, Qingde Li, You Zhang, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, and Qingqi Hong. 2022. LViT: Language meets vision transformer in medical image segmentation. arXiv:2206.14718. Retrieved from https://arxiv.org/abs/2206.14718

[157] Ruizhi Liao, Daniel Moyer, Miriam Cha, Keegan Quigley, Seth Berkowitz, Steven Horng, Polina Golland, and William M. Wells. 2021. Multimodal representation learning via maximization of local mutual information. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 273–283.

[158] Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? arXiv:2207.08143. Retrieved from https://arxiv.org/abs/2207.08143

[159] Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models. arXiv:2005.00683. Retrieved from https://arxiv.org/abs/2005.00683

[160] Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2019. A bert-based universal model for both within-and cross-sentence clinical temporal relation extraction. In *Clinical Natural Language Processing Workshop*. 65–71.

[161] Chen Lin, Timothy Miller, Dmitriy Dligach, Farig Sadeque, Steven Bethard, and Guergana Savova. 2020. A bert-based one-pass multi-task model for clinical temporal relation extraction. In *SIGBioMed Workshop on Biomedical Language Processing*. 70–75.

[162] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. 74–81.

[163] Xinzhu Lin, Xiahui He, Qin Chen, Huaixiao Tou, Zhongyu Wei, and Ting Chen. 2019. Enhancing dialogue symptom diagnosis with global attention and symptom graph. In *EMNLP-IJCNLP*. 5033–5042.

[164] Zachary C. Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[165] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv:2107.13586. Retrieved from https://arxiv.org/abs/2107.13586

[166] Qi Liu, Matt J. Kusner, and Phil Blunsom. 2020. A survey on contextual embeddings. arXiv:cs.CL/2003.07278. Retrieved from https://arxiv.org/abs/2003.07278

[167] Wenge Liu, Jianheng Tang, Jinghui Qin, Lin Xu, Zhen Li, and Xiaodan Liang. 2020. Meddg: A large-scale medical consultation dataset for building medical dialogue system. arXiv:2010.07497. Retrieved from https://arxiv.org/abs/2010.07497

[168] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692. Retrieved from https://arxiv.org/abs/1907.11692

[169] Pilar López-Úbeda, Manuel Carlos Díaz-Galiano, L. Alfonso Ureña-López, and M Teresa Martín-Valdivia. 2021. Pretrained language models to extract information from radiological reports. *CLEF eHealth* (2021).

[170] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinf.* 23, 6 (2022).

[171] Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. arXiv:2210.04705. Retrieved from https://arxiv.org/abs/2210.04705

[172] Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. arXiv:2303.15621. Retrieved from https://arxiv.org/abs/2303.15621

[173] Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. CitationSum: Citation-aware graph contrastive learning for scientific paper summarization. arXiv:2301.11223. Retrieved from https://arxiv.org/abs/2301.11223

[174] Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R. Eguchi, Po-Ssu Huang, and Richard Socher. 2020. Progen: Language modeling for protein generation. arXiv:2004.03497. Retrieved from https://arxiv.org/abs/2004.03497.

[175] Tittaya Mairittha, Nattaya Mairittha, and Sozo Inoue. 2020. Improving fine-tuned question answering models for electronic health records. In *UBICOMP*. 688–691.

[176] Aurelie Mascio, Zeljko Kraljevic, Daniel Bean, Richard Dobson, Robert Stewart, Rebecca Bendayan, and Angus Roberts. 2020. Comparative analysis of text classification approaches in electronic health records. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. 86–94.

[177] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* 54, 6 (2021), 1–35.

[178] Xing Meng, Craig H. Ganoe, Ryan T. Sieberg, Yvonne Y. Cheung, and Saeed Hassanpour. 2019. Self-supervised contextual language representation of radiology reports to improve the identification of communication urgency. arXiv:cs.LG/1912.02703. Retrieved from https://arxiv.org/abs/1912.02703

[179] Yiwen Meng, William Farran Speier, Michael K. Ong, and Corey Arnold. 2021. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE J. Biomed. Health Inf.* (2021), 1–1. https://doi.org/10.1109/jbhi.2021.3063721

[180] Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to bert embeddings during fine-tuning? arXiv:2004.14448. Retrieved from https://arxiv.org/abs/2004.14448

[181] George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alex Wong. 2020. Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. arXiv:2010.10391. Retrieved from https://arxiv.org/abs/2010.10391

[182] Claudiu Mihăilă, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2013. Biocause: Annotating and analysing causality in the biomedical domain. *BMC Bioinf.* 14, 1 (2013), 1–18.

[183] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781. Retrieved from https://arxiv.org/abs/1301.3781

[184] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. *NeuIPS* 26 (2013), 3111–3119.

[185] Joshua R. Minot, Nicholas Cheney, Marc Maier, Danne C. Elbers, Christopher M. Danforth, and Peter Sheridan Dodds. 2021. Interpretable bias mitigation for textual data: reducing gender bias in patient notes while maintaining classification performance. arXiv:2103.05841. Retrieved from https://arxiv.org/abs/2103.05841

[186] Giacomo Miolo, Giulio Mantoan, and Carlotta Orsenigo. 2021. Electramed: A new pre-trained language representation model for biomedical nlp. arXiv:cs.CL/2104.09585. Retrieved from https://arxiv.org/abs/2104.09585

[187] A. Miranda-Escalada, E. Farré, and M. Krallinger. 2020. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. In *IberLEF*.

[188] Florian Mock, Fleming Kretschmer, Anton Kriese, Sebastian Böcker, and Manja Marz. 2021. BERTax: Taxonomic classification of DNA sequences with Deep Neural Networks (unpublished).

[189] Masoud Monajatipoor, Mozhdeh Rouhsedaghat, Liunian Harold Li, C.-C. Jay Kuo, Aichi Chien, and Kai-Wei Chang. 2022. Berthop: An effective vision-and-language model for chest x-ray disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 725–734.

[190] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. 2022. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE J. Biomed. Health Inf.* (2022).

[191] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning bert: misconceptions, explanations, and strong baselines. arXiv:2006.04884. Retrieved from https://arxiv.org/abs/2006.04884

[192] John Moult, Jan T. Pedersen, Richard Judson, and Krzysztof Fidelis. 1995. A large-scale experiment to assess protein structure prediction methods.

[193] Martin Müller, Marcel Salathé, and Per Egil Kummervold. 2020. COVID-Twitter-BERT: A natural language processing model to analyse Covid-19 content on twitter. arXiv:2005.07503. Retrieved from https://arxiv.org/abs/2005.07503

[194] Martin M. Müller and Marcel Salathé. 2019. Crowdbreaks: Tracking health trends using public social media data and crowdsourcing. *Front. Publ. Health* 7 (2019), 81.

[195] Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2022. The role of local alignment and uniformity in image-text contrastive learning on medical images. arXiv:2211.07254. Retrieved from https://arxiv.org/abs/2211.07254

[196] Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. 2022. Joint learning of localized representations from medical images and reports. In *European Conference on Computer Vision*. Springer, 685–701.

[197] Yuta Nakamura, Shouhei Hanaoka, Yukihiro Nomura, Naoto Hayashi, Osamu Abe, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2020. Kart: Privacy leakage framework of language models pre-trained with clinical records. arXiv:2101.00036. Retrieved from https://arxiv.org/abs/2101.00036

[198] Usman Naseem, Matloob Khushi, Vinay Reddy, Sakthivel Rajendran, Imran Razzak, and Jinman Kim. 2020. Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. arXiv:2009.09223. Retrieved from https://arxiv.org/abs/2009.09223

[199] Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. 2019. Results of the seventh edition of the bioasq challenge. In *ECML PKDD*. Springer, 553–568.

[200] Oded Nov, Nina Singh, and Devin Mann. 2023. Putting ChatGPT's Medical Advice to the (Turing) Test. arXiv.2301.10035. Retrieved from https://arxiv.org/abs/2301.10035

[201] Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J. Marshall, Ani Nenkova, and Byron C. Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *ACL*, Vol. 2018. NIH Public Access, 197.

[202] Ibrahim Burak Ozyurt. 2020. On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining. In *Workshop on Scholarly Document Processing*. 104–112.

[203] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M. Friedrich. 2018. Radiology objects in COntext (ROCO): A multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer, 180–189.

[204] Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *BioNLP Workshop and Shared Task*. 58–65.

[205] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.

[206] Bethany Percha. 2021. Modern clinical text mining: A guide and review. *Annu. Rev. Biomed. Data Sci.* 4 (2021), 165–187.

[207] Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. arXiv:2105.11447. Retrieved from https://arxiv.org/abs/2105.11447

[208] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*. 2227–2237.

[209] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *EMNLP-IJCNLP*. 2463–2473.

[210] Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: A text-to-text transformer model for biomedical literature. arXiv:cs.CL/2106.03598. Retrieved from https://arxiv.org/abs/2106.03598

[211] Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. Inexpensive domain adaptation of pretrained language models: Case studies on biomedical NER and Covid-19 QA. In *EMNLP*. 1482–1490.

[212] Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Andrew Rowley, Hong-Woo Chun, Sung-Jae Jung, Sung-Pil Choi, Jun'ichi Tsujii, and Sophia Ananiadou. 2015. Overview of the cancer genetics and pathway curation tasks of bionlp shared task 2013. *BMC Bioinf.* 16, 10 (2015), 1–19.

[213] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Sci. Chin. Technol. Sci.* (2020), 1–26.

[214] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[215] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. arXiv:2103.00020. Retrieved from https://arxiv.org/abs/2103.00020

[216] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. *Improving Language Understanding by Generative Pre-training*. OpenAI Technical Report. (2018).

[217] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv:cs.LG/1910.10683. Retrieved from https://arxiv.org/abs/1910.10683

[218] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. arXiv:1606.05250. Retrieved from https://arxiv.org/abs/1606.05250

[219] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. arXiv:2204.06125. Retrieved from https://arxiv.org/abs/2204.06125

[220] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. arXiv:cs.CV/2102.12092. Retrieved from https://arxiv.org/abs/2102.12092

[221] Alan Ramponi, Rob van der Goot, Rosario Lombardo, and Barbara Plank. 2020. Biomedical event extraction as sequence labeling. In *EMNLP*. 5357–5367.

[222] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. 2019. Evaluating protein transfer learning with tape. *NeuIPS* 32 (2019), 9689.

[223] Roshan Rao, Jason Liu, Robert Verkuil, Joshua Meier, John F. Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. 2021. Msa transformer (unpublished).

[224] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2020. Med-bert: Pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. arXiv:2005.12833. Retrieved from https://arxiv.org/abs/2005.12833

[225] Bhanu Pratap Singh Rawat, Wei-Hung Weng, So Yeon Min, Preethi Raghavan, and Peter Szolovits. 2020. Entity-enriched neural models for clinical question answering. In *SIGBioMed Workshop on Biomedical Language Processing*. 112–122.

[226] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci. U.S.A.* 118, 15 (2021).

[227] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*. 10684–10695.

[228] Subendhu Rongali, Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. 2021. Continual domain-tuning for pretrained language models. arXiv:cs.CL/2004.02288. Retrieved from https://arxiv.org/abs/2004.02288

[229] Frank Rudzicz and Raeid Saqur. 2020. Ethics of artificial intelligence in surgery. arXiv:2007.14302. Retrieved from https://arxiv.org/abs/2007.14302

[230] Devendra Singh Sachan, Pengtao Xie, Mrinmaya Sachan, and Eric P. Xing. 2018. Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition. In *Machine Learning for Healthcare Conference*. PMLR, 383–402.

[231] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. arXiv:2205.11487. Retrieved from https://arxiv.org/abs/2205.11487

[232] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. arXiv:cs.CL/1910.01108. Retrieved from https://arxiv.org/abs/1910.01108

[233] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. arXiv:2211.05100. Retrieved from https://arxiv.org/abs/2211.05100

[234] Elisa Terumi Rubel Schneider, Joao Vitor Andrioli de Souza, Yohan Bonescki Gumiel, Claudia Moro, and Emerson Cabrera Paraiso. 2021. A gpt-2 language model for biomedical texts in portuguese. In *CBMS*. IEEE, 474–479.

[235] Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt - A Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, 65–72. https://doi.org/10.18653/v1/2020.clinicalnlp-1.7

[236] Constantin Seibold, Simon Reiß, M. Saquib Sarfraz, Rainer Stiefelhagen, and Jens Kleesiek. 2022. Breaking with fixed set pathology recognition through report-guided contrastive training. arXiv:2205.07139. Retrieved from https://arxiv.org/abs/2205.07139

[237] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*. Association for Computational Linguistics, 1715–1725.
[238] Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019. Pre-training of graph augmented transformers for medication recommendation. arXiv:1906.00346. Retrieved from https://arxiv.org/abs/1906.00346
[239] Shreyas Sharma and Ron Daniel Jr au2. 2019. Bioflair: Pretrained pooled contextualized embeddings for biomedical sequence labeling tasks. arXiv:cs.CL/1908.05760. Retrieved from https://arxiv.org/abs/1908.05760.
[240] Soumya Sharma, Bishal Santra, Abhik Jana, Santosh Tokala, Niloy Ganguly, and Pawan Goyal. 2019. Incorporating domain knowledge into medical NLI using knowledge graphs. In *EMNLP-IJCNLP*. 6092–6097.
[241] Golnar Sheikhshab, Inanc Birol, and Anoop Sarkar. 2018. In-domain context-aware token embeddings improve biomedical named entity recognition. In *Workshop on Health Text Mining and Information Analysis*. 160–164.
[242] Xiaoming Shi, Haifeng Hu, Wanxiang Che, Zhongqian Sun, Ting Liu, and Junzhou Huang. 2020. Understanding medical conversations with scattered keyword attention and weak supervision from responses. In *AAAI*, Vol. 34. 8838–8845.
[243] Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. Biomegatron: Larger biomedical domain language model. In *EMNLP*. 4700–4706.
[244] Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embeddings. *J. Am. Med. Inf. Assoc.* 26, 11 (2019), 1297–1304.
[245] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. arXiv:2212.13138. Retrieved from https://arxiv.org/abs/2212.13138
[246] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. arXiv:2305.09617. Retrieved from https://arxiv.org/abs/2305.09617
[247] Larry Smith, Lorraine K. Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M. Friedrich, Kuzman Ganchev, et al. 2008. Overview of biocreative ii gene mention recognition. *Genome Biol.* 9, 2 (2008), 1–19.
[248] Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. 2017. Biosses: A semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics* 33, 14 (2017), i49–i58.
[249] Jose Roberto Ayala Solares, Francesca Elisa Diletta Raimondi, Yajie Zhu, Fatemeh Rahimian, Dexter Canoy, Jenny Tran, Ana Catarina Pinho Gomes, Amir H. Payberah, Mariagrazia Zottoli, Milad Nazarzadeh, et al. 2020. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *J. Biomed. Inf.* 101 (2020), 103337.
[250] Sarvesh Soni and Kirk Roberts. 2020. Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering. In *LREC*. 5532–5538.
[251] Peter Spyns. 1996. Natural language processing in medicine: An overview. *Methods Inf. Med.* 35, 4-5 (1996), 285–301.
[252] Pascal Sturmfels, Jesse Vig, Ali Madani, and Nazneen Fatema Rajani. 2020. Profile Prediction: An alignment-based pre-training task for protein sequence models. arXiv:2012.00195. Retrieved from https://arxiv.org/abs/2012.00195
[253] Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham Barezi, and Pascale Fung. 2020. CAiRE-COVID: A question answering and query-focused multi-document summarization system for COVID-19 scholarly information management. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.
[254] Peng Su, Yifan Peng, and K. Vijay-Shanker. 2021. Improving BERT model using contrastive learning for biomedical relation extraction. In *Proceedings of the 20th Workshop on Biomedical Language Processing*. 1–10.
[255] Peng Su and K. Vijay-Shanker. 2020. Investigation of bert model on biomedical relation extraction based on revised fine-tuning mechanism. In *BIBM*. IEEE, 2522–2529.
[256] Sanjay Subramanian, Lucy Lu Wang, Ben Bogin, Sachin Mehta, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. 2020. MedICaT: A dataset of medical images, captions, and textual references. In *EMNLP*. 2112–2120.
[257] Dianbo Sui, Yubo Chen, Jun Zhao, Yantao Jia, Yuantao Xie, and Weijian Sun. 2020. Feded: Federated learning via ensemble distillation for medical relation extraction. In *EMNLP*. 2118–2128.
[258] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*. Springer, 194–206.
[259] Cong Sun, Zhihao Yang, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. 2020. Biomedical named entity recognition using bert in the machine reading comprehension framework. arXiv:2009.01560. Retrieved from https://arxiv.org/abs/2009.01560
[260] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. arXiv:1906.08976. Retrieved from https://arxiv.org/abs/1906.08976

[261] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*. 5100–5111.
[262] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020. Efficient transformers: A survey. arXiv:2009.06732. Retrieved from https://arxiv.org/abs/2009.06732
[263] Ashok Thillaisundaram and Theodosia Togia. 2019. Biomedical relation extraction with pre-trained language representations and minimal task-specific architecture. In *Workshop on BioNLP Open Shared Tasks*. 84–89.
[264] Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Fine-tuning large neural language models for biomedical natural language processing. arXiv:2112.07869. Retrieved from https://arxiv.org/abs/2112.07869
[265] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P. Langlotz, Andrew Y. Ng, and Pranav Rajpurkar. 2022. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng.* (2022), 1–8.
[266] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNll-2003 shared task: language-independent named entity recognition. In *NAACL-HLT*. 142–147.
[267] Yiqi Tong, Yidong Chen, and Xiaodong Shi. 2021. A multi-task approach for improving biomedical named entity recognition by incorporating multi-granularity information. In *ACL-IJCNLP*. 4804–4813.
[268] Hai-Long Trieu, Thy Thy Tran, Khoa N. A. Duong, Anh Nguyen, Makoto Miwa, and Sophia Ananiadou. 2020. Deepeventmine: End-to-end neural nested event extraction from biomedical texts. *Bioinformatics* 36, 19 (2020), 4910–4917.
[269] Ryan Turner, David Eriksson, Michael McCourt, Juha Kiili, Eero Laaksonen, Zhen Xu, and Isabelle Guyon. 2021. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020. arXiv:2104.10201. Retrieved from https://arxiv.org/abs/2104.10201
[270] Elena Tutubalina, Ilseyar Alimova, Zulfat Miftahutdinov, Andrey Sakhovskiy, Valentin Malykh, and Sergey Nikolenko. 2021. The russian drug reaction corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics* 37, 2 (2021), 243–249.
[271] Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J. Am. Med. Inf. Assoc.* 18, 5 (2011), 552–556.
[272] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*. 5998–6008.
[273] Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. Bertology meets biology: Interpreting attention in protein language models. arXiv:cs.CL/2006.15222. Retrieved from https://arxiv.org/abs/2006.15222.
[274] Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. arXiv:2010.05731. Retrieved from https://arxiv.org/abs/2010.05731
[275] Shoya Wada, Toshihiro Takeda, Shiro Manabe, Shozo Konishi, Jun Kamohara, and Yasushi Matsumura. 2020. Pre-training technique to localize medical bert and enhance biomedical bert. arXiv:2005.07202. Retrieved from https://arxiv.org/abs/2005.07202
[276] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *EMNLP-IJCNLP*. 5788–5793.
[277] Byron C. Wallace, Sayantan Saha, Frank Soboczenski, and Iain J. Marshall. 2020. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. arXiv:2008.11293. Retrieved from https://arxiv.org/abs/2008.11293
[278] Zhongwei Wan, Che Liu, Mi Zhang, Jie Fu, Benyou Wang, Sibo Cheng, Lei Ma, César Quilodrán-Casas, and Rossella Arcucci. 2023. Med-UniC: Unifying cross-lingual medical vision-language pre-training by diminishing bias. arXiv:2305.19894. Retrieved from https://arxiv.org/abs/2305.19894
[279] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv:1804.07461. Retrieved from https://arxiv.org/abs/1804.07461
[280] Benyou Wang, Lifeng Shang, Christina Lioma, Xin Jiang, Hao Yang, Qun Liu, and Jakob Grue Simonsen. 2021. On position embeddings in bert. In *ICLR*, Vol. 2. 12–13.
[281] Benyou Wang, Donghao Zhao, Christina Lioma, Qiuchi Li, Peng Zhang, and Jakob Grue Simonsen. 2019. Encoding word order in complex embeddings. In *ICLR 2020 Spotlight*.
[282] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. 2022. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *Advances in Neural Information Processing Systems*.
[283] Xiaosong Wang, Ziyue Xu, Leo Tam, Dong Yang, and Daguang Xu. 2021. Self-supervised image-text pre-training with mixed data in chest x-rays. arXiv:2103.16022. Retrieved from https://arxiv.org/abs/2103.16022

[284] Xing David Wang, Leon Weber, and Ulf Leser. 2020. Biomedical event extraction as multi-turn question answering. In *ACL Workshop on Health Text Mining and Information Analysis*. 88–96.
[285] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2018. Clinical information extraction applications: A literature review. *J. Biomed. Inf.* 77 (2018), 34–49.
[286] Yu-An Wang and Yun-Nung Chen. 2020. What do position embeddings learn? an empirical study of pre-trained language model positional encoding. In *EMNLP*. 6840–6849.
[287] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. MedCLIP: Contrastive learning from unpaired medical images and text. arXiv:2210.10163. Retrieved from https://arxiv.org/abs/2210.10163
[288] Neha Warikoo, Yung-Chun Chang, and Wen-Lian Hsu. 2021. Lbert: Lexically aware transformer-based bidirectional encoder representation model for learning universal bio-entity relations. *Bioinformatics* 37, 3 (2021), 404–412.
[289] Qiang Wei, Zongcheng Ji, Yuqi Si, Jingcheng Du, Jingqi Wang, Firat Tiryaki, Stephen Wu, Cui Tao, Kirk Roberts, and Hua Xu. 2019. Relation extraction from clinical narratives using pre-trained language models. In *AMIA Annual Symposium Proceedings*, Vol. 2019. American Medical Informatics Association, 1236.
[290] Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *ACL*. 201–207.
[291] Wei-Hung Weng and Peter Szolovits. 2019. Representation learning for electronic health records. arXiv:1909.09248. Retrieved from https://arxiv.org/abs/1909.09248
[292] Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and Heuiseok Lim. 2020. An effective domain adaptive post-training method for bert in response selection. In *INTERSPEECH*. 1585–1589.
[293] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*. Association for Computational Linguistics, 1112–1122.
[294] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. arXiv:cs.CL/1910.03771. Retrieved from https://arxiv.org/abs/1910.03771
[295] Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, et al. 2020. Deep learning in clinical natural language processing: A methodical review. *J. Am. Med. Inf. Assoc.* 27, 3 (2020), 457–470.
[296] Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting bert. In *ACL*. 4166–4176.
[297] Zhaofeng Wu, Yan Song, Sicong Huang, Yuanhe Tian, and Fei Xia. 2019. WTMED at MEDIQA 2019: A hybrid approach to biomedical natural language inference. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. 415–426.
[298] Yuan Xia, Jingbo Zhou, Zhenhui Shi, Chao Lu, and Haifeng Huang. 2020. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In *AAAI*, Vol. 34. 1062–1069.
[299] Yijia Xiao, Jiezhong Qiu, Ziang Li, Chang-Yu Hsieh, and Jie Tang. 2021. Modeling protein using large-scale pretrain language model. arXiv:2108.07435. Retrieved from https://arxiv.org/abs/2108.07435
[300] Qianqian Xie, Jennifer Amy Bishop, Prayag Tiwari, and Sophia Ananiadou. 2022. Pre-trained language models with domain knowledge for biomedical extractive summarization. *Knowl.-Bas. Syst.* 252 (2022), 109460.
[301] Qianqian Xie, Jinpeng Hu, Jiayu Zhou, Yifan Peng, and Fei Wang. 2023. Factreranker: Fact-guided reranker for faithful radiology report summarization. arXiv:2303.08335. Retrieved from https://arxiv.org/abs/2303.08335
[302] Qianqian Xie, Jimin Huang, Tulika Saha, and Sophia Ananiadou. 2022. GRETEL: Graph contrastive topic enhanced language model for long document extractive summarization. In *COLING*. 6259–6269.
[303] Qianqian Xie, Zheheng Luo, Benyou Wang, and Sophia Ananiadou. 2023. A survey on biomedical text summarization with pre-trained language model. arXiv:2304.08763. Retrieved from https://arxiv.org/abs/2304.08763
[304] Qianqian Xie and Fei Wang. 2023. Faithful AI in healthcare and medicine (unpublished).
[305] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. Simmim: A simple framework for masked image modeling. In *CVPR*. 9653–9663.
[306] Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *AAAI*, Vol. 33. 7346–7353.
[307] Song Xu, Haoran Li, Peng Yuan, Yujia Wang, Youzheng Wu, Xiaodong He, Ying Liu, and Bowen Zhou. 2021. K-plug: Knowledge-injected pre-trained language model for natural language understanding and generation in e-commerce. arXiv:2104.06960. Retrieved from https://arxiv.org/abs/2104.06960
[308] Kui Xue, Yangming Zhou, Zhiyuan Ma, Tong Ruan, Huanhuan Zhang, and Ping He. 2019. Fine-tuning bert for joint entity and relation extraction in chinese medical text. In *BIBM*. IEEE, 892–897.

[309] Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. 2018. Mining electronic health records (ehrs) a survey. *ACM Comput. Surv.* 50, 6 (2018), 1–40.
[310] Shweta Yadav, Vishal Pallagani, and Amit Sheth. 2020. Medical knowledge-enriched textual entailment framework. In *ICCL*. 1795–1801.
[311] Keisuke Yamada and Michiaki Hamada. 2022. Prediction of RNA–protein interactions using a nucleotide language model. *Bioinformatics Advances* 2, 1 (2022), vbac023.
[312] Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhaochun Ren, Xin Xin, Huasheng Liang, Maarten de Rijke, and Zhumin Chen. 2022. ReMeDi: Resources for multi-domain, multi-service, medical dialogues. In *ACM SIGIR*. 3013–3024.
[313] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.* 10, 2 (2019), 1–19.
[314] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.* 10, 2, Article 12 (Jan 2019), 19 pages. https://doi.org/10.1145/3298981
[315] Xi Yang, Jiang Bian, William R. Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *J. Am. Med. Inf. Assoc.* (10 2020). https://doi.org/10.1093/jamia/ocaa189
[316] Xi Yang, Xing He, Hansi Zhang, Yinghan Ma, Jiang Bian, and Yonghui Wu. 2020. Measurement of semantic textual similarity in clinical texts: Comparison of transformer-based models. *JMIR Med. Inf.* 8, 11 (2020), e19735.
[317] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding. arXiv:cs.CL/1906.08237. Retrieved from https://arxiv.org/abs/1906.08237
[318] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *ACL (Volume 1: Long Papers)*. 8003–8016.
[319] Wonjin Yoon, Richard Jackson, Jaewoo Kang, and Aron Lagerberg. 2021. Sequence tagging for biomedical extractive question answering. arXiv:2104.07535. Retrieved from https://arxiv.org/abs/2104.07535
[320] Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, and Jaewoo Kang. 2019. Pre-trained language model for biomedical question answering. In *ECML PKDD*. Springer, 727–740.
[321] Fei Yu, Hongbo Zhang, and Benyou Wang. 2023. Nature language reasoning, a survey. arXiv:2303.14725 (2023).
[322] Xin Yu, Wenshen Hu, Sha Lu, Xiaoyan Sun, and Zhenming Yuan. 2019. Biobert based named entity recognition in electronic medical record. In *ITME*. IEEE, 49–52.
[323] Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. *BioNLP 2022@ ACL 2022*, 97.
[324] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Adv. Neural Inf. Process. Syst.* 34 (2021), 27263–27277.
[325] Zheng Yuan, Zhengyun Zhao, and Sheng Yu. 2020. Coder: Knowledge infused cross-lingual medical term embedding for term normalization. arXiv:2011.02947. Retrieved from https://arxiv.org/abs/2011.02947
[326] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. arXiv:2007.14062. Retrieved from https://arxiv.org/abs/2007.14062
[327] Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. Meddialog: A large-scale medical dialogue dataset. In *EMNLP*. 9241–9250.
[328] Zhiqiang Zeng, Hua Shi, Yun Wu, and Zhiling Hong. 2015. Survey of natural language processing techniques in bioinformatics. *Comput. Math. Methods Med.* 2015 (2015).
[329] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. HuatuoGPT, towards taming language models to be a doctor. arXiv:2305.15075. Retrieved from https://arxiv.org/abs/2305.15075
[330] Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: Quantifying biases in clinical contextual word embeddings. In *CHIL*. 110–120.
[331] Hongbo Zhang, Xiang Wan, and Benyou Wang. 2023. Injecting knowledge into biomedical pre-trained models via polymorphism and synonymous substitution. arXiv:2305.15010. Retrieved from https://arxiv.org/abs/2305.15010
[332] Ningyu Zhang, Zhen Bi, Xiaozhuan Liang, Lei Li, Xiang Chen, Shumin Deng, Luoqiu Li, Xin Xie, Hongbin Ye, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Mosha Chen, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Huajun Chen, Buzhou Tang, and Qingcai Chen. 2021. Cblue: A chinese biomedical language understanding evaluation benchmark. arXiv:cs.CL/2106.08087. Retrieved from https://arxiv.org/abs/2106.08087
[333] Ningyu Zhang, Qianghuai Jia, Kangping Yin, Liang Dong, Feng Gao, and Nengwei Hua. 2020. Conceptualized representation learning for chinese biomedical text mining. arXiv:2008.10813. Retrieved from https://arxiv.org/abs/2008.10813

[334] Rong Zhang, Revanth Gangi Reddy, Md Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avirup Sil, and Todd Ward. 2020. Multi-stage pre-training for low-resource domain adaptation. arXiv:cs.CL/2010.05904. Retrieved from https://arxiv.org/abs/2010.05904

[335] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *ICLR*.

[336] Wei Zhang, Lu Hou, Yichun Yin, Lifeng Shang, Xiao Chen, Xin Jiang, and Qun Liu. 2020. Ternarybert: Distillation-aware ultra-low bit bert. arXiv:cs.CL/2009.12812. Retrieved from https://arxiv.org/abs/2009.12812

[337] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. 2020. Contrastive learning of medical visual representations from paired images and text. arXiv:2010.00747. Retrieved from https://arxiv.org/abs/2010.00747

[338] Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao. 2020. Mie: A medical information extractor towards medical dialogues. In *ACL*. 6460–6469.

[339] Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *ACL*. 5108–5120.

[340] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B. Dolan. 2020. DIALOGPT: Large-scale generative pre-training for conversational response generation. In *ACL: System Demonstrations*. 270–278.

[341] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv:1707.09457. Retrieved from https://arxiv.org/abs/1707.09457

[342] Huiwei Zhou, Xuefei Li, Weihong Yao, Chengkun Lang, and Shixian Ning. 2019. Dut-nlp at mediqa 2019: An adversarial multi-task network to jointly model recognizing question entailment and question answering. In *BioNLP Workshop and Shared Task*. 437–445.

[343] Henghui Zhu, Ioannis C. Paschalidis, and Amir M. Tahmasebi. 2018. Clinical concept extraction with contextual word embedding. In *NeuIPS Workshop on Machine Learning for Health*.

[344] Wei Zhu, Yuan Ni, Xiaoling Wang, and Guotong Xie. 2021. Discovering better model architectures for medical query understanding. In *NAACL-HLT*. 230–237.

[345] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*. 19–27.

[346] Konstantin Weißenow, Michael Heinzinger, and Burkhard Rost. 2021. Protein language model embeddings for fast, accurate, alignment-free protein structure prediction. *bioRxiv* (2021).

Received 28 October 2021; revised 15 June 2023; accepted 22 June 2023