

# Large Language Models in Drug Discovery and Development: From Disease Mechanisms to Clinical Trials

**Yizhen Zheng<sup>1</sup> \* Huan Yee Koh<sup>1,2</sup> \* Maddie Yang<sup>4</sup> Li Li<sup>4,5</sup> Lauren T. May<sup>2</sup> Geoffrey I. Webb<sup>1</sup>**  
**Shirui Pan<sup>3+</sup> George Church<sup>4,5+</sup>**

1. Department of Data Science and AI, Monash University
2. Drug Discovery Biology, Monash Institute of Pharmaceutical Sciences, Monash University
3. School of Information and Communication Technology, Griffith University
4. Harvard Medical School, Harvard University
5. Wyss Institute for Biologically Inspired Engineering, Harvard University

\* indicates equal contribution and + indicates corresponding authors:

George Church(george\_church@hms.harvard.edu), Shirui Pan(s.pan@griffith.edu.au)

## Abstract

The integration of Large Language Models (LLMs) into the drug discovery and development field marks a significant paradigm shift, offering novel methodologies for understanding disease mechanisms, facilitating drug discovery, and optimizing clinical trial processes. This review highlights the expanding role of LLMs in revolutionizing various stages of the drug development pipeline. We investigate how these advanced computational models can uncover target-disease linkage, interpret complex biomedical data, enhance drug molecule design, predict drug efficacy and safety profiles, and facilitate clinical trial processes. Our paper aims to provide a comprehensive overview for researchers and practitioners in computational biology, pharmacology, and AI4Science by offering insights into the potential transformative impact of LLMs on drug discovery and development.

## 1. Introduction

*“Language is only the instrument of science, and words are but the signs of ideas.”*

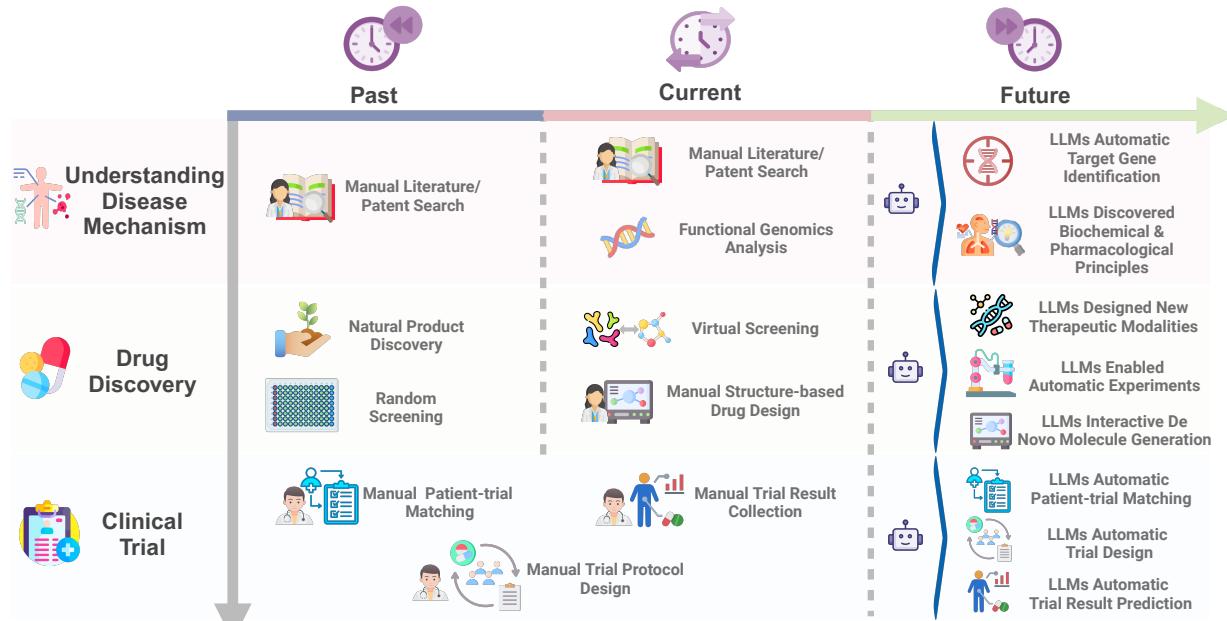
— Samuel Johnson

The pursuit of new drugs to research and develop is a long-term commitment that typically takes 10-15 years and costs over \$2 billion in order to bring a new drug to a patient (Berdigaliyev & Aljofan, 2020). This complex procedure is traditionally divided into three stages: the first stage is to understand the disease and to choose the target of treatment; the second stage is to develop a focused approach

to developing treatments towards the target; and the third stage is to test the treatments in clinical trials for their effectiveness. Each phase of the process is both time-consuming and resource-intensive, this is because of the complexity of biological systems and the extensive nature of the review required of each phase in the research and validation process. The slow and protracted nature of the process often prevents the introduction of new therapies that would improve and extend human life. Consequently, there are extraordinary dividends to be reaped by introducing efficiencies and expanding the capabilities of current practices.

Artificial intelligence (AI) tools have emerged as preeminent innovation in the quest to accelerate drug discovery and development. Among these tools, large language models (LLMs)<sup>1</sup> have distinguished themselves through their capabilities in understanding scientific language and executing various downstream tasks essential in drug discovery and development. Recent LLM breakthroughs, Geneformer (Theodoris et al., 2023), pretrained on 30 million single-cell transcriptomes, can help in disease modeling and successfully identified candidate therapeutic targets for cardiomyopathy via in silico deletion. Notable LLMs for facilitating chemistry experiments, Boiko et al. (2023) and Chemcrow (Bran et al., 2023), have highlighted the potential of LLMs in automating chemistry experiments related to drug discovery, specifically in the fields of directed synthesis and chemical reaction prediction. Other works, such as LLM4SD (Zheng et al., 2023), showed that LLMs can perform scientific synthesis, inference, and explanation directly from raw experimental data and formulate hypotheses that resonate with human experts’ analysis. Med-PaLM (Singhal et al., 2023), a mega size LLM encoding clinical knowledge, was the first to reach human expert in USMLE-styled ques-

<sup>1</sup>Large Language Models (LLMs) are also known as large pre-trained language models.



**Figure 1. Large Language Models Shaping the Future Landscape of Drug Discovery and Development.** In the past, each stage of drug discovery involved numerous manual tasks, which requires significant human effort and substantial resources. Nowadays, advancements in biotechnology, alongside the integration of AI and computer-aided *in silico* computation tools, have reduced the need of human labor and resources. However, we have yet to have a highly automated drug discovery pipeline, especially in the clinical trial phase, where trial design and matching are still mainly done by clinical practitioners. In the future, it is anticipated that the continued development of LLMs and their application in drug discovery will enable a highly automated drug discovery process.

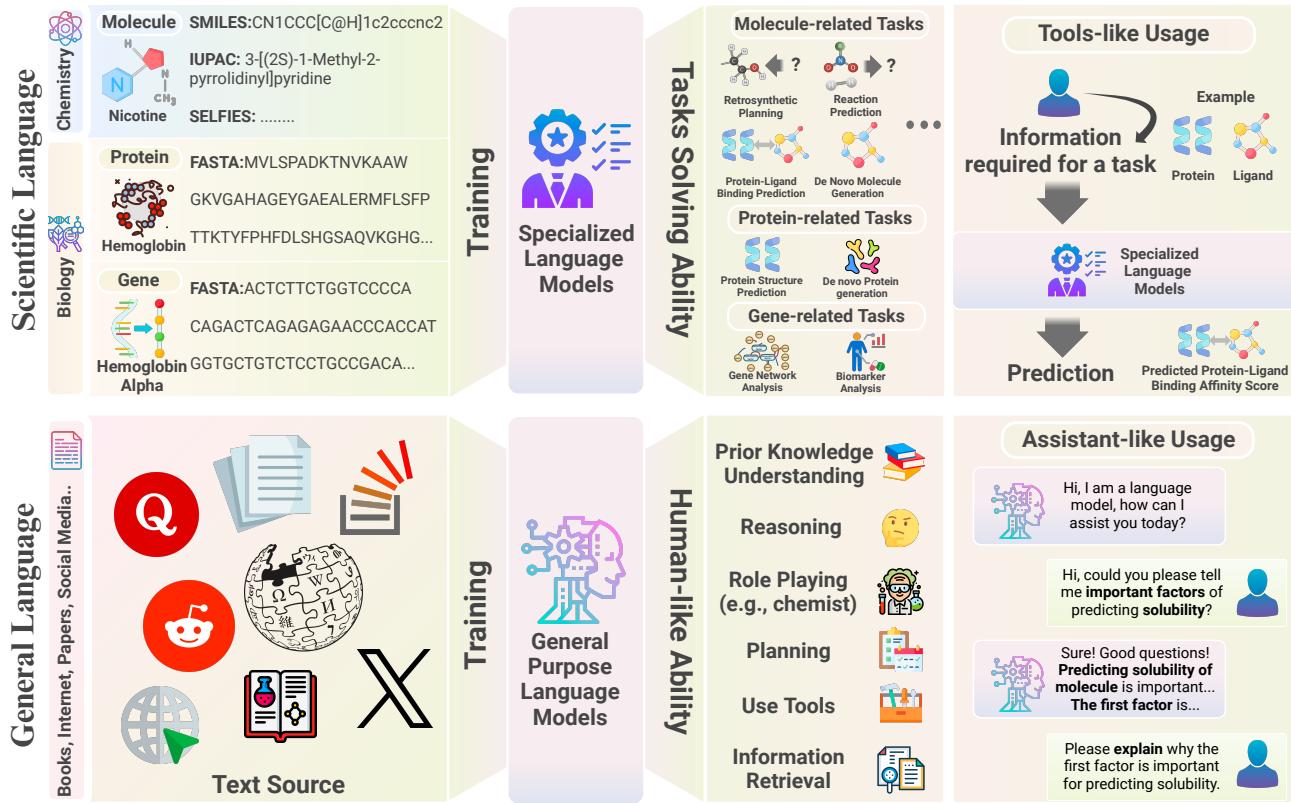
tions, a medical licensing examination. This advancement highlights the potential of LLMs to liberate clinical practitioners from the laborious activities associated with clinical trials.

With advancements in LLMs, these technologies have the potential to revolutionize the drug discovery pipeline, with future drug discovery including highly automated LLM applications across the three stages of drug discovery (Figure 1). To understand disease mechanisms and aid in target identification, LLMs can perform comprehensive literature reviews and patent analyses to explore the biological pathways involved in diseases. Additionally, they can conduct functional genomics analysis to pinpoint target genes. By analyzing gene-related literature, including results from *in vivo* or *in vitro* experiments, LLMs can compare data on various genes and recommend those with favorable characteristics, such as a desirable mechanism of action or strong potential as drug targets. Furthermore, through analysis and review of literature, LLMs may infer new insights and uncover principles of biochemistry and pharmacology. In the drug discovery and development phase, LLMs have the potential to automate related chemistry experiments by understanding chemical reaction and controlling robotic equipments. In addition, LLMs can offer an interactive platform aiding experts in discovering novel and effective compounds through suggestions for molecule editing and generation. LLMs could also assist in the design of new therapeutic ap-

proaches, such as gene therapy. For instance, LLMs could help in the design of Adeno-associated virus (AAV) vectors by quickly summarizing scientific literature to identify novel strategies and by analyzing genomic sequences to predict the most effective vector sequences for safe and efficient gene delivery. During the clinical trial phase, LLMs could streamline the tedious tasks of matching patients with trials and designing trials by interpreting patient profiles and trial requirements. Additionally, early research has shown that LLMs might be capable of predicting trial outcomes by examining historical clinical data.

**In this survey, we aim to comprehensively address three questions** for researchers and practitioners seeking to harness the power of LLMs to improve the drug discovery and development pipeline:

**1) How can LLMs be effectively integrated into various drug discovery and development stages?** First, the types of LLMs under consideration must be defined (Figure 2). Then, we categorize the whole drug discovery and development pipeline into three linear stages: “Understanding Disease Mechanisms” (Figure 3), “Drug Discovery” (Figure 4), and “Clinical Trials” (Figure 5) to illustrate the blueprint of integrating LLMs into these processes, respectively. The left column of each figure describes the specific processes involved in these stages, while the right column covers the tasks that LLMs can perform to facilitate these stages. The



**Figure 2. The two main paradigms of language models.** Specialized language models are trained on specific scientific languages and are typically tailored for specific or a few science-related tasks. These models are used as tools to perform a specific task, in which users provide the information required for a task, and the model outputs the prediction. General-purpose language models are trained on diverse textual information sourced from various materials, including scientific papers and textbooks. These models are used like an assistant that allows users to use plain language to interact with the model.

visual representation seeks to illustrate how LLMs can optimize various aspects of drug development.

**2) How advanced are LLMs in facilitating downstream tasks across various drug discovery and development stages?** In order to determine the level of advancement of LLMs in supporting downstream tasks throughout different stages of drug discovery and development, we have evaluated current applications of LLMs and classified each one into one of four categories: not applicable, nascent, advanced, and mature. These indicators provide an overview of the current state in the field and indicate promising future directions (Figure 6).

**3) What are the future directions of LLMs in drug discovery and development?** We explore the evolving landscape of LLM development, promoting LLMs in more biological use cases, and addressing ethical, privacy, fairness, and bias concerns in future development. These concerns are increasingly apparent as LLMs are applied to handle sensitive health data and make critical medical decisions. We

also discuss the need to overcome certain technical limitations associated with LLMs, such as the occurrence of hallucinations, the constraints of context window limit, and the need for better model interpretability and scientific understanding. Solving these challenges can enable LLMs to become trusted and efficient tools in the applications of drug discovery as well as in patient care. This is discussed in Section 5.

In this paper, we first elucidate the two paradigms of LLMs for drug discovery and development: specialized language models trained on specific scientific languages and general-purpose language models trained on general textual language. We then delve into how LLMs can be helpful throughout each drug discovery and development stage, from understanding disease mechanisms to facilitating drug discovery and optimizing clinical trials. After covering each stage, we analyze the maturity of these LLM applications. Lastly, we discuss the future directions for LLMs in drug discovery and development.

## 2. Main Paradigms of Language Models

In drug discovery and development, the intricate text-based scientific languages used to describe chemicals and proteins, such as SMILES strings for encoding molecular structures (Weininger, 1988), and FASTA format for encoding protein, DNA and RNA sequences (fas, 1995), represent a unique form of structured language crafted by humans to encode domain-specific knowledge. To effectively interpret these languages, two main language model paradigms, including specialized language models (specialized LLMs) and general-purpose language models (general LLMs) emerged (Figure 2).

### 2.1. Specialised Large Language Models

The first paradigm of language models in drug discovery and development is specialized LLMs trained in specific scientific languages. These LLMs aim to decode the statistical patterns of scientific language, thereby enabling the interpretation of scientific data in its raw form (Figure 2).

**Understanding Disease Mechanisms.** Specialized LLMs can be used in many ways to explore diseases. For instance, LLMs can extract genomic information from single-cell RNA transcriptomic data and DNA sequences (Consens et al., 2023), enabling practitioners to determine epigenetic marks, transcription factor binding sites, functional genetic variants, and gene network analysis, all of which contribute to understanding the genetic basis of disease.

Additionally, protein LLMs, such as ESM (Rives et al., 2021), can be trained to predict parts of the amino acid sequence that have been intentionally hidden or ‘masked’ during training, such as “MVL<MASK>PAD” (Figure 2). Despite a simple training procedure, these specialized LLMs have been proven helpful in annotating the functions (Matic et al., 2022) and predicting the structures of proteins directly from protein sequences (Lin et al., 2023), significantly advancing our understanding of protein structures, and informing downstream drug discovery efforts.

**Drug Discovery.** In drug discovery, specialized LLMs are particularly helpful for accelerating various chemistry experiments (Bran et al., 2023; Park et al., 2023). A specialized LLM trained in the molecular SMILES language can help in retrosynthetic planning and predicting reaction outcomes; at the same time, it can help chemists in de novo molecules guided by some specific molecular properties, such as increasing binding affinity towards targets. Moreover, these models can also play a role in ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) prediction, a critical step in assessing molecule properties and filtering out those with undesirable characteristics.

**Usage.** Specialized LLMs are tool-like, where the user inputs information needed for a given task and receives a

model prediction in return (Figure 2). For instance, when we use a specialized LLM to query protein-ligand binding affinity, both protein sequences and ligand SMILES strings must be provided to the model, which will subsequently output the predicted binding affinity score.

### 2.2. General-purpose Language Models

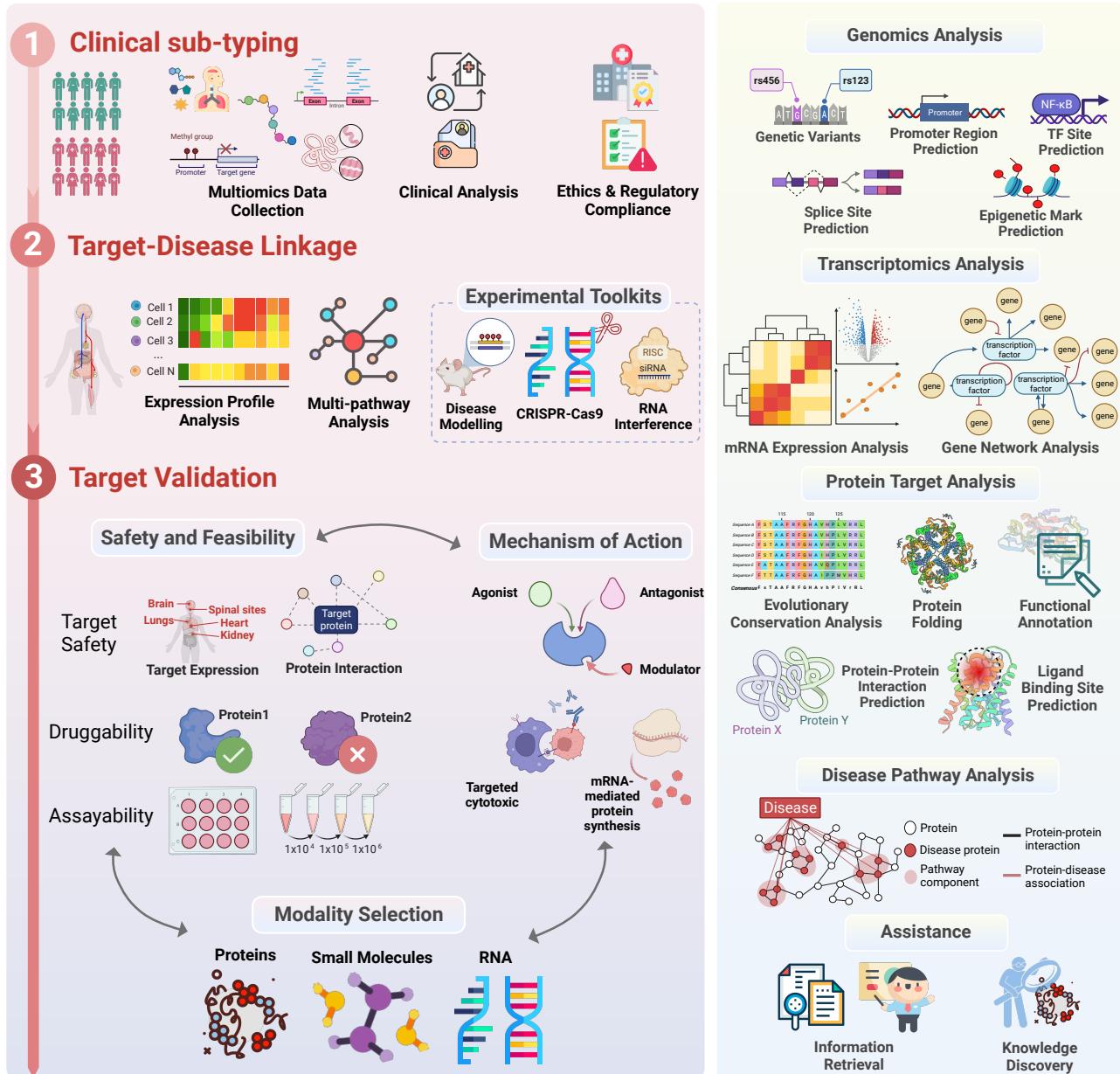
The second paradigm encompasses general LLMs, which are trained on a diverse array of textual information sourced from various materials, including but not limited to scientific papers, textbooks, and general literature. Such breadth in training allows them to achieve a broad understanding of human language, which includes a significant grasp of scientific contexts. Models like GPT-4 (OpenAI, 2023; AI4Science & Quantum, 2023) and Galactica (Taylor et al., 2022) have been noted for their proficiency in also mastering complex formal scientific description languages, including SMILES strings and FASTA format. Using this capacity, general LLMs can work on tasks that would typically require the participation of domain professionals, such as making inferences, doing reasoning and analysis, and applying field-specific knowledge across different scientific domains.

**Understanding Disease Mechanisms.** General LLMs can traverse a large volume of literature, extract data, and summarize for users. Furthermore, it can also synthesize the extracted data into a knowledge graph, revealing how genes and diseases are connected, helping scientists uncover the basis behind diseases (Savage, 2023). Furthermore, these models can explain technical terminologies in layperson’s language, making understanding complex concepts and principles easy, significantly aiding in education and communication.

**Drug Discovery.** In drug discovery, general LLMs have great potential to accelerate experimental practices. Recently, general LLMs have been applied in chemistry robotics for automated experiments. General LLMs also exhibited expert-level capabilities in retrosynthetic planning and reaction prediction (Bran et al., 2023; Boiko et al., 2023), while only costing a fraction of human experts.

Some preliminary attempts are underway to train general language models to perfect tasks currently more suited for specialized LLMs, such as de novo molecule and protein generation and editing (Liu et al., 2023c). The primary motivation is that, unlike specialized LLMs that can only learn data patterns from specific scientific languages, these LLMs can reason and apply the domain knowledge learned from extensive literature. However, research in this direction is still in its infancy.

**Clinical Trials.** General LLMs provide significant advantages in analyzing electronic health records and clinical



**Figure 3. Understanding Disease Mechanisms.** The left part of the figure illustrates the processes involved in understanding disease mechanisms. This process involves clinical sub-typing, target-disease linkage analysis, and target validation. Clinical subtyping refers to identifying subgroups of patients with similar clinical characteristics during which data can be collected from multi-omics. Target-disease linkage analysis refers to identifying the relationship between targets and diseases. Target validation typically involves three steps: safety and feasibility, mechanisms of action, and modality selection. The right part of the figure highlights the tasks that LLMs can perform to facilitate these processes, including genomics analysis, RNA analysis, pathway analysis, target profiling, strategic profiling, and assistance.

protocols (Singhal et al., 2023; Jin et al., 2023c; Huang et al., 2020). They can facilitate patient-trial matching, assist in trial planning, help predict trial outcomes, and assist in document writing. The user-friendly chat interfaces of general LLMs also make it easier for practitioners to interact with them.

**Usage.** A natural language-based AI Assistant is a strong use case for general LLMs (Figure 2). For example, if a user wanted to know what some key features are in predicting the solubility of molecules, they could ask a general LLM, and it would retrieve and summarize relevant information from the literature.

### 3. LLMs in Drug Discovery and Development

This section discusses how LLMs can be applied in three drug discovery and development pipeline stages: understanding disease mechanisms, drug discovery, and clinical trials.

#### 3.1. Understanding Disease Mechanisms

Understanding disease mechanisms is the initial and crucial stage of the drug discovery and development pipeline. The primary aim of this stage (Figure 3) is to identify a suitable protein target for a potential drug to act upon (Lindsay, 2003). This process involves three key steps: clinical sub-typing, target-disease linkage analysis, and target validation.

Clinical sub-typing in drug discovery involves categorizing patients into subgroups to collect clinical and multiomics data and aids in understanding disease variations and identifying potential differences in disease mechanisms across patient groups (Cortés-Cros et al., 2013; Pun et al., 2023).

The target-disease linkage analysis phase in drug discovery involves establishing connections between potential protein targets and specific diseases. This phase encompasses pathway analysis to investigate biological pathways involved in the disease and expression profile analysis to study disease-related gene expression patterns (Plenge et al., 2013). Additionally, practitioners leverage experimental techniques to establish causal links between a target and the disease, including CRISPR-Cas9 (Lin et al., 2017), in-vivo disease modeling (Lindsay, 2003), and interference RNA (siRNA) (Cortés-Cros et al., 2013).

After identifying a target in the drug discovery process, target validation is a crucial, non-linear step that follows target identification, involving a continuous validation cycle with no fixed starting point (Figure 3). This cycle includes assessing the necessary actions to be performed on the target for disease treatment (mechanism of action), choosing the most appropriate therapeutic intervention (modality selection), and conducting a comprehensive safety and feasibility as-

essment (Emmerich et al., 2021). The safety and feasibility assessment evaluates both the potential organismal impact (safety) and the target's druggability (Floris et al., 2018), as well as the practicality of assays for feasibility (Vincent et al., 2015). This flexible, iterative approach ensures thorough evaluation of the target's viability and safety at any stage before advancing in drug development, ensuring the selected targets are both theoretically promising and practical for further development.

#### 3.1.1. GENOMICS ANALYSIS

Decades of genome-wide association studies (GWAS) have identified critical genomic regions linked to various diseases (Michailidou et al., 2015; 2017; Nelson et al., 2017; Zengini et al., 2018) that have significantly advanced genomic-based analysis for disease understanding and target discovery. Notably, integrating genetic associations in drug discovery, could significantly improve the success rate of clinical targets (Nelson et al., 2015).

Recently, there has been significant interest in adapting advancement in LLMs used for human languages to genomic analysis, such as DNA-BERT (Ji et al., 2021), due to the structural similarities between DNA and human language. Through specialized training on vast amounts of nucleotide sequences, these LLMs are adept at decoding the language of genetics. As a result, there has been an explosion in the field of specialized nucleotide LLMs (Ji et al., 2021) that are increasingly capable of understanding the cryptic “language” used by genomes more efficiently, enabling various downstream tasks in understanding genetic mechanisms of diseases.

**Genetic variant analysis.** The application of nucleotide LLMs in genetic variant analysis hinged on the fact that genetic sequences follow specific language patterns and rules (Yanofsky et al., 1964; Altschuh et al., 1988). Variations in these sequences—be it single nucleotide polymorphisms (SNPs), insertions, deletions, or more complex rearrangements—can significantly impact gene function (Brendel & Busse, 1984; Searls, 2002). Hence, they employ masked language modeling when nucleotide LLMs are trained on extensive genomic data. In this approach, the model learns to predict parts of the nucleotide sequence that have been intentionally hidden or ‘masked’ during training. This learning process enables the LLMs to decode the intricate, often hidden patterns and rules that govern the language of genes (Ji et al., 2021; Dalla-Torre et al., 2023).

Post-training, nucleotide LLMs have demonstrated the ability to detect significant functional genetic variants directly from DNA sequences. For example, DNA-BERT (Ji et al., 2021) showed that nucleotide LLMs selectively concentrate on the most relevant genomic regions. This enables the extraction of motif patterns that are evolutionarily conserved

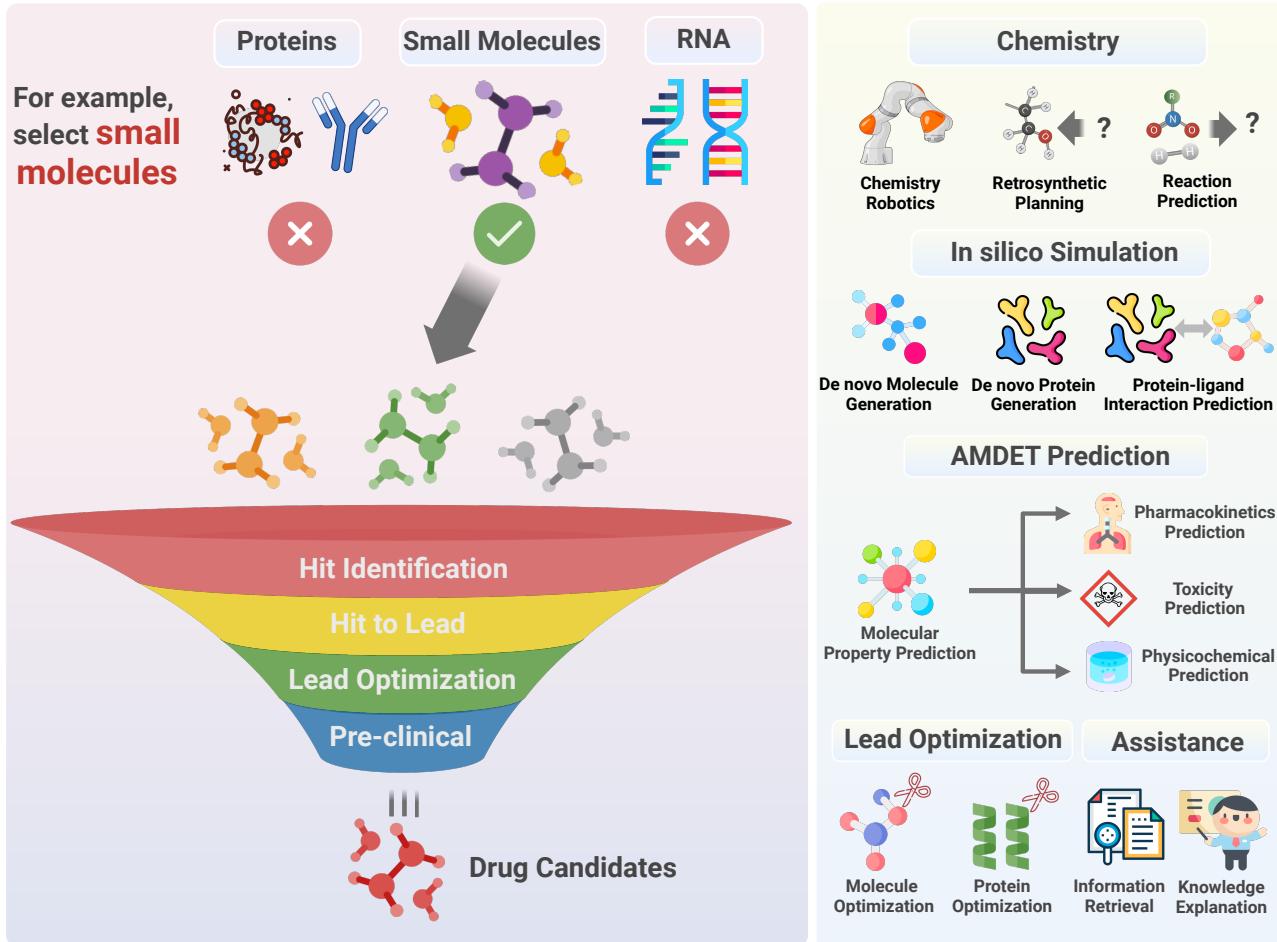


Figure 4. **Drug Discovery.** The left part of the figure illustrates the processes involved in drug discovery. The right part of the figure highlights the tasks that LLMs can perform to facilitate these processes.

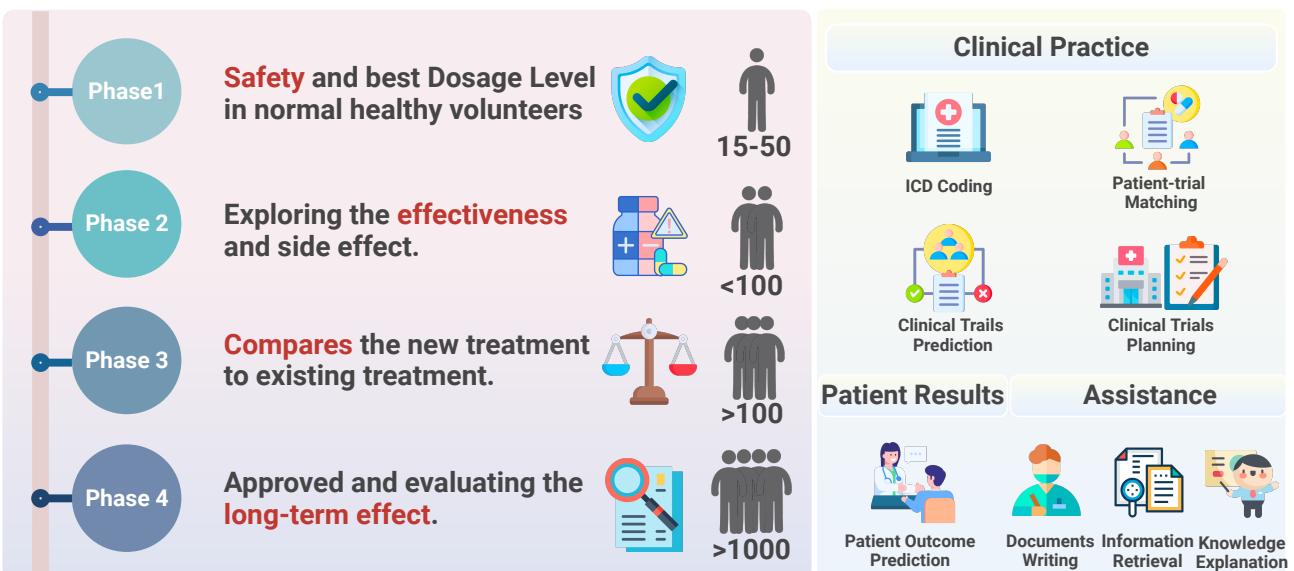


Figure 5. **Clinical Trials.** The left part of the figure illustrates the processes involved in clinical trials. Clinical trials consist of four phases: Phase 1, Phase 2, Phase 3, and Phase 4. The right part of the figure highlights the tasks that LLMs can perform to facilitate these processes.

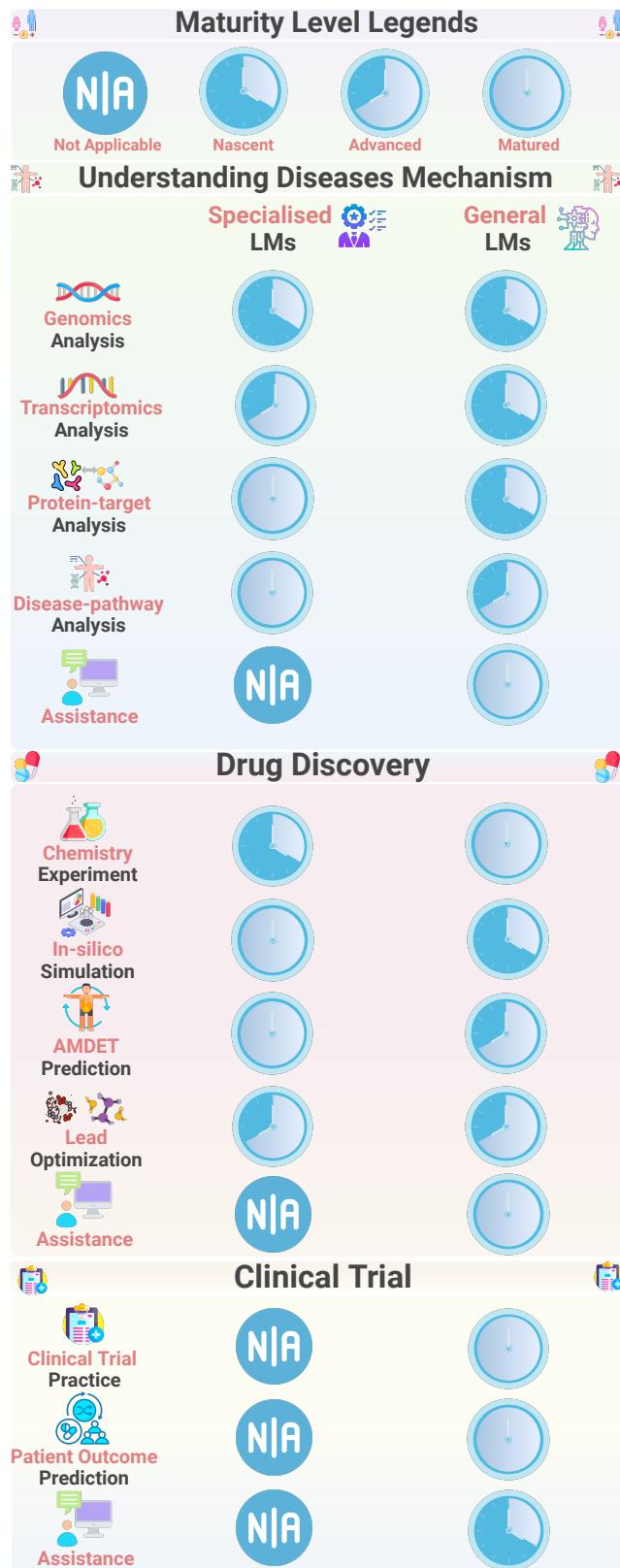
and aid in identifying functional variants of significance. Similarly, Nucleotide Transformer (Dalla-Torre et al., 2023), another specialized nucleotide LLM, has also demonstrated the ability to prioritize functional genetic variants. Moreover, it can be further trained for specific variant identification tasks, such as classifying SARS-CoV-2 variants (Zhou et al., 2023) and understanding SARS-CoV-2 evolutionary dynamics (Zvyagin et al., 2023).

More recently, HyenaDNA, built on the Hyena LLM framework, has pushed the boundaries of genetic variant analysis by enabling the modeling of extremely long genomic sequences—up to 1 million tokens—at the single nucleotide level (Nguyen et al., 2024b). This is a significant leap from previous models that were constrained by the quadratic scaling of attention and limited to much shorter sequences. HyenaDNA’s ability to process such extensive context lengths allows it to capture long-range interactions in DNA, which are crucial for understanding complex genetic variations. It has achieved state-of-the-art performance on multiple benchmarks with a much smaller model and less pretraining data, marking a substantial advance in the field of genomic sequence analysis. This long-range capability, coupled with the precision of single nucleotide resolution, positions HyenaDNA as a powerful tool in detecting and prioritizing functional genetic variants, further enhancing our understanding of genomic data.

**Genomic regions-of-interest predictions.** Promoter regions, transcription factor (TF) binding sites, and splice sites are all crucial elements in regulating gene expression. Despite their varied roles, they all contribute to the complex regulation of when, where, and how genes are activated or silenced. Alterations or mutations in these regions may bring about overexpression or underexpression of a gene, potentially causing diseases. However, despite the importance, predicting these regions remains challenging due to the need for more understanding in the language of DNA (Ji et al., 2021).

To address these challenges, specialized nucleotide LLMs are being fine-tuned to predict these regions of interest, with results showing an outperformance against previous state-of-the-art methods (Dalla-Torre et al., 2023; Zhou et al., 2023). Fine-tuning these LLMs for genomic applications involves two key steps. First, an LLM is pre-trained on extensive datasets of nucleotide sequences, allowing the LLM to grasp the nuances of genetic language. Then, in the second stage, the LLM is further fine-tuned to incorporate domain-specific knowledge, which is the locations, biological functions, and additional biochemical and biophysical insights of different gene regulatory sites.

Epigenetic marks refer to chemical modifications, such as DNA methylation and histone modifications, that affect gene expression without altering the underlying DNA sequence.



**Figure 6. Maturity Assessment of LLMs in Downstream Tasks.** This figure is segmented into Maturity Level Legends, Understanding Disease Mechanism, Drug Discovery, and Clinical Trials, detailing LLMs maturity across various tasks and phases.

These marks play a crucial role in regulating gene activity, influencing both disease development and therapeutic targeting (Atlasi & Stunnenberg, 2017). Accurately predicting these marks is essential for understanding how epigenetic changes impact gene expression and their implications in various diseases (Miranda Furtado et al., 2019). However, the complexity and variability of epigenetic marks present a significant challenge in making accurate predictions. Similar to the discussion above, pre-trained nucleotide LLMs are further fine-tuned to predict specific histone modifications, such as H3K14ac, H3K36me3, and H3K4me1 (Zhou et al., 2023).

### 3.1.2. TRANSCRIPTOMICS ANALYSIS.

Transcriptomics, a field that investigates the entirety of RNA transcripts that an organism or cell system generates under certain conditions, has experienced a surge in the volume of transcriptomic data derived from a wide range of human tissues due to the development of high-throughput technologies and single-cell technologies. However, the data is often sparse for specific disease states, particularly for rare diseases and diseases affecting clinically inaccessible tissues (Shao et al., 2021), so relying solely on these data for specific diseases would likely not suffice to develop robust and accurate models. To address existing limitations, specialized gene LLMs have been proposed to obtain a comprehensive understanding of transcriptomic data while offering the goods to adapt to scenarios with sparse data samples.

The primary technological development in this subfield is the specialized transcriptomic LLM, Geneformer (Theodoris et al., 2023), which developed an innovative method for mapping each single-cell transcriptome into a sequence of genes ranked by their expression levels. This approach, known as “rank value encoding”, represents the transcriptome of each cell as a sequence of genes ordered based on their expression levels. These levels are then normalized against the overall expression observed across all human tissues. Through this technique, rank value encoding offers a distinct representation of gene activity within individual cells and facilitates a comprehensive comparison of gene expression across a diverse array of data (Theodoris et al., 2023). This approach is akin to learning the language of transcriptomics through specialized LLMs, enhancing our understanding of cellular behaviors and interactions at the molecular level.

By transforming single-cell transcriptomic data into gene sequences, Geneformer (Theodoris et al., 2023), scGPT (Cui et al., 2023), and other models like scMulan (Bian et al., 2024) and scFoundation (Hao et al., 2024) have demonstrated a remarkable ability to analyze transcriptomic data effectively as foundational LLM models. Furthermore, spe-

cialized transcriptomic LLMs can be adapted to scenarios with sparse data through fine-tuning to model gene networks accurately to comprehend complex dynamics, including network interactions, extending beyond simple cell-level annotations (Ma et al., 2024).

In parallel, efforts have also been made to leverage biomedical literature for predicting future therapeutic targets by training LLMs on historical text corpora. A study used Word2Vec models on abstracts published between 1995 and 2022, allowing these LLMs to prioritize gene-disease associations and protein-protein interactions likely to be validated in future research (Narganes-Carlón et al., 2023). This approach, termed Publication-Wide Association Study (PWAS), encodes biomedical knowledge as word embeddings without human supervision, effectively capturing drug discovery concepts and prioritizing hypotheses years before experimental confirmation. PWAS demonstrates the potential of LLMs as a scalable system for early-stage target ranking, enhancing the ability to mine literature for under-explored therapeutic opportunities.

**mRNA expression analysis.** mRNA expression analysis can be challenging due to the need to derive meaningful insights from limited data scenarios, a vital aspect in enhancing our understanding of diseases. Traditional machine learning approaches, such as XGBoost (Chen & Guestrin, 2016) and standard deep neural networks, usually start from scratch for each specific task. This methodology can be ineffective, especially if the data is limited, which is often the case in the fields of rare disease research or when working with tissues that are not accessible in a clinical setting.

The specialized transcriptomic LLM, Geneformer (Theodoris et al., 2023), leverages the general knowledge acquired from pretraining on transcriptomic data to adapt efficiently to a specific disease use case and demonstrates remarkable efficiency in gene network analysis with minimal data. A study successfully distinguished key factors in the NOTCH1-dependent network by fine-tuning on just 884 endothelial cells from healthy versus dilated aortas, outperforming other methods that used a much larger dataset of about 30,000 cells (Theodoris et al., 2023).

scGPT (Cui et al., 2023), another generated pre-trained transformer model for single-cell multi-omics data analysis, on the other hand, also showed the ability to generate meaningful cell-type clusters directly from the pre-trained model in a zero-shot manner (i.e., without additional fine-tuning).

**Gene network analysis.** Gene network analysis typically begins by mapping the gene regulatory networks and tracing the critical genes in a disease’s progression to identify potential therapeutic targets. Specifically, the gene network analysis strives to uncover vital regulatory elements that can alter or modulate these networks in a desired manner (Theodoris

et al., 2015; 2021). Due to the lack of data, these regulatory elements are challenging to find, especially for rare diseases or conditions affecting clinically inaccessible tissues (Cui et al., 2023; Theodoris et al., 2023).

To address the challenge of gene network analysis, Geneformer (Theodoris et al., 2023) proposed leveraging the self-attention mechanism within the transformer (Vaswani et al., 2017) backbone of the LLM to address the challenge of gene network analysis. This self-attention is crucial, as the trained attention weights of the model for each gene reveal two essential aspects: (1) the genes to which a particular gene is paying attention and (2) the genes that focus on it. This process inherently constructs a gene network, intricately mapping the web of gene interactions. Geneformer (Theodoris et al., 2023) also employs “in silico deletion”, a technique similar to perturbation analysis for virtually removing specific genes to study their impact on gene networks. scGPT (Cui et al., 2023), on the other hand, uses embedding computing to construct a gene network via gene-gene similarities.

### 3.1.3. PROTEIN TARGET ANALYSIS

A protein sequence is often the most accessible data about a target that can help explain its potential to play a role in disease mechanisms, with drug discovery scientists often targeting it as a starting point. In this application, specialized LLMs can be particularly valuable as these models have shown the ability to provide extensive analyses, including evolutionary conservation, functional annotation, protein folding, and binding site prediction. The unique ability of the LLMs to extract relevant information from sequence data alone has provided a means of characterization of target biological traits and functions even without experimental data like experimentally determined 3D protein structures.

**Evolutionary conservation.** The use of specialized LLMs in protein analysis, as explained in the representative work of ESM (Rives et al., 2021), is based on a fundamental idea: the statistical patterns of protein sequences contain valuable information about their biological function and structure, which have been shaped by evolutionary processes (Yanofsky et al., 1964; Altschuh et al., 1988). This idea proposes that mutations that improve an organism’s fitness are more likely to be selected by evolutionary forces among the multitude of possible mutations a sequence can undergo (Göbel et al., 1994), resulting in unique signatures in protein sequence patterns.

Specialized protein-based LLMs can effectively predict likely mutations within protein sequences that contain masked amino acids. This proficiency not only enables them to understand evolutionary conservation but also allows them to grasp the selection processes driving the evolution of these sequences. Follow-up research has demon-

strated the efficacy of this approach using the ESM language model, which can make accurate predictions of mutational effects across a variety of proteins with different functions without any additional training (Meier et al., 2021). MSA-Transformer takes this approach further by analyzing multiple sequences using a multiple sequence alignment (MSA) approach (Rao et al., 2021). With the added information from the MSA as input, MSA-Transformers enhance their ability to interpret complex relationships within protein sequences and outperform single-sequence LLMs (Meier et al., 2021).

Understanding evolutionary conservation using specialized LLMs is significant because it provides information not only on the functional landscape of proteins according to amino acid conservation patterns but also enables the establishment of the role and significance of individual residues in binding and activity (Altschuh et al., 1987). This is important as such conserved sites typically contribute to protein function and structure, while covarying mutations are similarly associated with these features, including contact surface, structure, and binding (Levitt, 1978; Yanofsky et al., 1964; Altschuh et al., 1988). Such findings are crucial for developing specialized LLMs for proteomics and form the basis for using these models to provide insight into predicting protein folding, binding sites, and functional annotation.

**Protein folding.** The sequence patterns of a protein are shaped by its hidden structure, which is linked to evolutionary conservation and mutation. Specific structures and sequences are conserved due to their functional importance, while mutations occur in response to evolutionary pressures. As a result, LLMs that learn from protein sequence data can indirectly capture these evolutionary trends. This is exemplified by the groundbreaking work in ESM (Rives et al., 2021) and MSA-Transformer (Rao et al., 2021), which showed that LLMs can accurately decode the structural nuances of proteins from sequence data alone. Specifically, when these LLMs create pairwise interaction maps (attention matrices) between all amino acid positions in a sequence, they demonstrate an ability to infer which pairs of amino acids should be in contact with unparalleled accuracy (Rao et al., 2020; Fung et al., 2022). This remarkable ability strongly suggests that a significant amount of structural information can be directly inferred from the LLM model using only sequence data, in line with Anfinsen’s dogma (Anfinsen, 1973).

Building on foundational research, AlphaFold2 (Jumper et al., 2021) and RosettaFold (Baek et al., 2021) have revolutionized the field of protein structure prediction. These models can now produce atom-level accuracy even in cases where similar structures are not known, due to the Evoformer component in AlphaFold2, a specialized protein-based LLM. In this way, AlphaFold2’s training objective mirrors that of MSA-Transformer, where residues

from MSA undergo random masking before being reconstructed by the Evoformer (Jumper et al., 2021; Hu et al., 2022). Through this process, AlphaFold2 can reason over MSAs and incorporate valuable evolutionary information to achieve near-experimental accuracy at the whole protein’s structure level (Mirdita et al., 2022; Ahdritz et al., 2024). Furthermore, this same approach has been further extended to understanding biomolecular interaction with RosettaFold All-Atom, enabling the modeling of complex biomolecular assemblies, such as protein-protein complexes, protein-DNA/RNA interactions, and protein-small molecule interactions (Krishna et al., 2024).

In a similar timeframe, RGN2 (Chowdhury et al., 2022) developed ProtBERT to encode protein sequence data and predict structure directly from a single sequence, without requiring evolutionary information from an MSA. RGN2 has demonstrated the ability to match or even surpass AlphaFold2 (Jumper et al., 2021) in predicting the structure of orphan proteins that lack sequence homologs (Chowdhury et al., 2022).

**Functional annotation.** As elucidated by the early works (Rao et al., 2019; Rives et al., 2021; Brandes et al., 2022; Lin et al., 2023), specialized protein-based LLMs can encode rich structural and functional information (Bepler & Berger, 2021). Naturally, the rich information encoded within LLMs is now being harnessed in advanced applications like NetGO 3.0 (Wang et al., 2023a) to advance automated function prediction of proteins without costly experiments. The same approach was used by Matic et al. (2022) for GPCR sequence analysis using ESM (Rives et al., 2021), aiming to predict their signaling and functional repertoire. The research showed that the interaction mechanism between different protein variants is due to alternative splicing of genes, and GPCRs can be clearly defined with the help of specialized protein-based LLMs.

With advancements in general LLMs, ProteinChat (Guo et al., 2023) was proposed to provide an interactive platform where users can upload protein sequences and structures and pose questions about a brief description of a protein’s functionalities. A recent work (AI4Science & Quantum, 2023) also demonstrates that GPT-4 demonstrates considerable expertise in understanding proteins.

Despite these advances, the study of protein language models has remained relatively limited in scope. However, ESM2 (Lin et al., 2023) and ESM3 (Hayes et al., 2024) have been addressing this gap by scaling up to an impressive 15 billion and 98 billion parameters respectively, making them some of the largest protein language models evaluated to date. This substantial increase in scale has enabled ESM models to better learn the sequence-structure-function relationships of proteins. Using ESM as the foundational LLM, ESMFold has demonstrated an unprecedented ability

to closely match the performance of AlphaFold2 in predicting protein structures (Lin et al., 2023). Notably, ESMFold achieves this high level of precision by analyzing a single sequence, and its ability to operate without querying a database significantly enhances its speed and ease of use.

**Protein-ligand interaction and binding site prediction.** Protein-ligand interactions and understanding protein binding sites play a pivotal role in understanding protein function and interactions and serves as an essential foundation for rational-based drug discovery. Gaining insight into these interactions is crucial in identifying potential safety issues related to target proteins and acts as guidance for the design of therapies. Protein-based specialized LLMs have demonstrated notable successes leveraging their extensive understanding of protein language. These include the prediction of metal ion binding sites (Yuan et al., 2022), protein-protein binding sites (Fang et al., 2023), and small molecule binding sites (Zhang & Xie, 2023). Understanding the interactions between proteins is thus fundamentally important in deciding a valid target for treating diseases. Furthermore, this knowledge can contribute to the design of biologics-based drugs, as discussed in sections 3.2.2 and 3.2.4.

To this end, protein-based LLMs have advanced with the introduction of AlphaFold-Multimer (Evans et al., 2021). While its initial application is in predicting multimeric complex structures, it turns out that AlphaFold-Multimer is capable of predicting protein-protein interaction directly from protein sequences as accurately as unique protein-protein docking methods that use experimentally-determined structures (Ketata et al., 2023).

Another significant development in protein-based LLMs for protein-protein interaction is DockGPT (McPartlon & Xu, 2023), an innovative approach in protein docking. This end-to-end deep learning method stands out for its flexible and site-specific protein docking capability, effectively accommodating conformational flexibility and utilizing binding site information compared to AlphaFold-Multimer. Its strength lies in its ability to process unbound and predicted monomer structures. Notably, DockGPT (McPartlon & Xu, 2023) showed that the protein-based LLM can effectively deal with antibody-antigen complexes, achieving high accuracy in predicting binding poses as well as co-design the sequence and structure of antibody regions targeting specific epitopes.

When it comes to assessing ligand interaction sites, Phosformer has proven to be a significant improvement. Unlike previous methods such as MusiteDeep, DeepPhos, and Ember (Wang et al., 2017; Luo et al., 2019; Kirchoff & Gomez, 2022), which relied on multiple models specific to different protein families or groups, Phosformer uses its comprehensive understanding of protein language to make accurate predictions with just one model. This means that virtually

any kinase interacting with any peptide can be used to predict phosphorylation sites. Meanwhile, ProtT5 (Elnaggar et al., 2021), a specialized protein-based machine learning model, has been used to successfully predict binding sites for metal ions, nucleic acids, and small molecules (Littmann et al., 2021). This approach has even been advanced to predict genome-wide annotations for these binding sites (Yuan et al., 2023).

Finally, recent advancements in protein-ligand interaction modeling using LLMs have been significantly enhanced by RosettaFold All-Atom (Krishna et al., 2024). Unlike earlier tools that focused primarily on polypeptide chains, RosettaFold All-Atom incorporates a wide range of ligands, including small molecules, metal ions, and nucleic acids, into its predictions. This comprehensive approach not only enables highly accurate modeling of protein-ligand complexes, but provides analysis of key cellular processes for protein target analysis, offering deep insights into disease mechanisms and providing a powerful tool for drug discovery (Krishna et al., 2024).

#### 3.1.4. PATHWAY ANALYSIS

In pathway analysis, gene regulatory network analysis can be a powerful tool for researchers seeking to decipher complex disease pathways. In this subsection, we will be focusing on the use of general LLMs, which can provide all-around assistance for pathway analysis.

Unlike their specialized counterparts, general LLMs are innately equipped with a wealth of prior knowledge gleaned from vast scientific literature and datasets (Taylor et al., 2022; OpenAI, 2023). This extensive background enables them to approach pathway analysis with a broad, informed perspective rather than specializing in a single scientific language. Furthermore, general-purpose LLMs have the distinct advantage of being able to interactively and conversationally engage with complex scientific data (OpenAI, 2023; Jeblick et al., 2023), providing researchers with a powerful tool for understanding and exploring their findings.

A recent study showcased the capacity of general-purpose LLMs, such as GPT-4, in analyzing blood transcriptional modules related to erythroid cells, demonstrating that these models are efficient in knowledge-driven pathway analysis (Toufiq et al., 2023). This research uses general LLMs to automatically generate codes for gene networks, summarize candidate genes ranked based on association tests, generate reports for users, and fact-check the report against the literature. In each task, the rich prior knowledge and interactive capabilities of general LLMs are exploited to analyze scientific data. By leveraging the rich prior knowledge and interactive capabilities of general LLMs, this study highlights how they can enhance disease mechanisms and

target identification, allowing for a better understanding of complex gene networks. Ultimately, this transforms pathway analysis from static approaches to more dynamic and interpretable methods.

#### 3.1.5. ASSISTANCE

The exploration of disease mechanisms is a complex task, requiring the contribution of experts from fields across health and pharmaceutical industries. In this environment, general-purpose LLMs that can perform tasks related to interactive and conversational skills, including information retrieval and knowledge explanation, can play a crucial role (Taylor et al., 2022).

General-purpose LLMs offer fast and accurate information retrieval, clear explanations tailored to user needs, and the ability to organize and categorize large datasets, enhancing workflow and productivity (Jeblick et al., 2023). By integrating with search engines, recent LLM iterations provide real-time access to scientific data, improving hypothesis generation and validation in disease research. Additionally, they aid in scientific communication by simplifying complex ideas for laypeople, fostering better collaboration among specialists with different expertise.

### 3.2. Drug Discovery

The drug discovery process is a crucial phase in the drug development pipeline, encompassing several critical steps as depicted in Figure 4. These steps include hit identification, hit to lead, lead optimization, and preclinical development.

The process starts with “hit identification”, where professionals find compounds with potential therapeutic effects. Next, “hit to lead” involves a more refined selection from these hits, identifying those most promising for further development. The third step, “lead optimization”, is a critical process of enhancing a lead compound’s efficacy, stability, and safety via editing. Finally, “preclinical development” entails rigorous testing of the optimized lead compound in animal models to assess its suitability for human trials.

Our survey will begin by outlining the specific downstream tasks associated with each operation. Following this, we will explore how LLMs can be incorporated into these tasks to advance the drug discovery process.

#### 3.2.1. CHEMISTRY

Medicinal chemistry is essential to drug discovery and development through independent laboratory work and compound synthesis. Autonomous lab operations use robotic manipulators, controlled and programmed to execute complex chemistry and synthetic reactions. In addition, high-throughput screening requires compounds to be precisely and efficiently synthesized as part of the hit identification

phase. After synthesis, the compound will be evaluated for activity and selectivity using pharmacological assays.

LLMs have proven to be highly valuable in these fields. LLMs can help generate codes that program the chemistry robotics based on user requirements (Bran et al., 2023; Boiko et al., 2023). In particular, LLMs can translate user requirements into complex experimental protocols and convert them into specific, understandable robot instructions. Additionally, LLMs are successful in retrosynthetic planning and reaction prediction, offering to recommend feasible synthetic routes and to predict possible chemical reactions. Using LLMs in this manner is beneficial as it brings efficiencies in compound synthesis and accelerates the drug discovery process.

**Chemistry Robotics.** Nowadays, chemistry robotics is an integral part of conducting chemistry experiments using autonomous laboratory operations. This technique involves converting instructions written in natural language into robot-executable plans, usually described using a fixed, well-defined language that resembles coding.

General LLMs like GPT-4 (OpenAI, 2023) and CodeLlama (Roziere et al., 2023) have shown the ability to generate effective code. Therefore, it is logical to use general LLMs to generate robot-executable plans, as they have been trained on a vast amount of code. One notable application is CLARify (Yoshikawa et al., 2023), which utilizes GPT-3 (Brown et al., 2020) to generate task plans in a specific Chemistry Description Language (XDL) based on descriptive user instructions in natural languages. Constrained task and motion planning problems are then solved using PDDLStream solvers. This approach aims to facilitate the autonomous and safe execution of chemistry experiments using general-purpose robot manipulators. Notably, these plans have shown much higher accuracy than baseline systems like SynthReader (Mehr et al., 2020).

Furthermore, there are preliminary attempts using GPT-4 to generate compatible scripts in Python to control OT-2 (Inagaki et al., 2023), a computer-controlled liquid handling robot, achieving 95% success within five iterations.

An emerging branch of AI research involves utilizing large language models as agents that will autonomously create, perform, and program scientific experiments. Boiko et al. (Boiko et al., 2023) presented one such method, showing how these models could use web search engines for information about molecule synthesis or employ vector search to find relevant documentation on chemical reactions. They have also developed multi-instrument systems code generation agents that can successfully implement complex experiments like Suzuki and Sonogashira cross-coupling reactions.

#### Retrosynthetic Planning & Reaction Prediction.

Retrosynthetic planning involves breaking down complex compounds into simpler precursor compounds, while reaction prediction entails forecasting the outcome of chemical reactions. These two tasks are pivotal for understanding how to synthesize complex molecules from more basic starting materials, which is an essential step for preparing experiments like high-throughput screening.

An early attempt at LLMs for retrosynthetic planning is the Molecular Transformer (Schwaller et al., 2019), which utilizes a simple encoder-decoder transformer framework. The model is trained to take reactants and reagents as input and predict the chemical product that can be synthesized from a reaction. It has demonstrated higher accuracy in reaction prediction than human chemists. Subsequently, Schwaller et al. (Schwaller et al., 2020) combined the Molecular Transformer with a hyper-graph exploration strategy to develop an automated retrosynthetic route planning system. The dynamically constructed hypergraph represents a generic reaction where each molecule is a node, and the hyper-arc symbolizes the reaction arrow. This optimal synthetic route is identified using beam search over a beam search across the hyper-graph of possible disconnection strategies.

The capabilities of LLMs were further extended by Chemformer (Irwin et al., 2022), which is based on the BART (Lewis et al., 2020) architecture. It includes an additional pretraining process that involves reconstructing masked SMILES strings and using an autoencoder to convert original SMILES to embeddings and back. The model is then fine-tuned for various downstream tasks, including reaction and retrosynthesis predictions.

Recently, general-purpose LLMs have emerged in this field, such as Chemcrow (Bran et al., 2023) and Boiko et al (Boiko et al., 2023). Boiko's system uses web search and simple calculations, while Chemcrow adopted a more sophisticated approach. Chemcrow has developed and utilized a more comprehensive range of customized molecule and reaction tools. These include functionalities like converting queries to SMILES, obtaining molecule prices, patent checking, and reaction classification. Additionally, Chemcrow adopts a four-step framework to improve LLMs' ability: think about the necessary steps, take action using tools, provide inputs to these tools, analyze observations, and then deliver the final answer. This approach has been shown to perform better than GPT-4 in most tasks evaluated by humans in synthesis planning. Similarly, a recent study (Jablonka et al., 2024) demonstrated that a fine-tuned GPT-3 model was shown to outperform traditional machine learning models on several chemistry tasks, especially in low-data scenarios. The findings highlighted that LLMs, even those not initially trained on chemical data, could adapt to various predictive chemistry tasks with minimal fine-tuning, showcasing the potential of LLMs in advancing chemical research.

### 3.2.2. IN-SILICO SIMULATION

In-silico simulation leverages computer models to simulate complex biological processes. These simulations are pivotal in understanding and predicting how drugs interact at the molecular level, leading to more efficient and targeted drug development. The three main tasks involved in in-silico simulations are de novo molecule generation, de novo protein generation, and protein-ligand interaction prediction.

**De novo Molecule Generation.** De novo molecule generation is a complex task in in-silico simulations that involves creating new molecular structures with the potential to be effective drugs. This process is categorized into two types: unconstrained molecule generation, which seeks to populate the chemical space of the training set, and constrained molecule generation, where molecules are synthesized to meet specific desired properties (Brown et al., 2019). Constrained generation requires a model to consider various constraints such as affinity to targets, selectivity against off-targets, appropriate physicochemical properties, ADME characteristics, pharmacokinetics/pharmacodynamics, toxicology, and synthesizability (Loeffler et al., 2023).

To benchmark the performance of de novo molecule generation methods, GuacaMol (Brown et al., 2019) was introduced to provide a benchmarking framework considering aspects such as validity, uniqueness, and novelty. Specialized language models have demonstrated remarkable proficiency. For instance, it has been shown that even simple RNN-based models can perform exceedingly well on challenging generative modeling tasks (Flam-Shepherd et al., 2022). These models effectively learn the complex distribution of molecules, such as the highest-scoring penalized logP molecules in ZINC15 or the most significant molecules in PubChem. To explore wide and novel chemical spaces, LLMs such as SMILES-LSTM and ORGAN (Guimaraes et al., 2017) have been assessed with the unconstrained generation ability to directly generate (Brown et al., 2019), and ORGAN. When it comes to constrained molecule generation, several notable approaches have been developed. Previous studies utilized reinforcement learning (Olivecrona et al., 2017) and pharmacophoric features (Skalic et al., 2019) to improve RNN-based models toward generating molecules with desired properties and binding ligands to protein pockets.

Moreover, MolGPT (Bagal et al., 2021b), with its GPT architecture, can handle multiple constraints. It is trained by recovering a molecule with its scaffold and properties. The REINVENT series (Blaschke et al., 2020; Loeffler et al., 2023) represents a more advanced approach in this category. It is sweeping and capable of meeting up to 10 different objectives, including synthesizability, selectivity, etc. This method narrows the chemical search space through a three-staged training process: the prior model, a transfer learning

agent, and staged learning towards generating high-scoring sequences. This sophisticated approach involves pretraining, transfer learning, reinforcement learning, and curriculum learning.

On the other hand, general LLMs usually focus on the constrained molecule generation task. In this context, MoIT5 (Edwards et al., 2022) uses a self-supervised learning framework to pretrain T5 (Raffel et al., 2020), a general-purpose LLM that is trained on a large corpus of text coupled with molecular data pairs. The pretraining methodology comprises unsupervised SMILES recovery and associated chemical texts. More recently, GPT-4 (OpenAI, 2023) has shown its ability to produce a novel molecule guided by textual instructions. However, the effectiveness generated by these two models is inferior to specialized language models. In addition, multimodal methods such as Momu (Su et al., 2022) and GIT-Mol (Liu et al., 2023a) enhance general LLMs' capabilities in molecule generation. Momu (Su et al., 2022) improves upon MoIT5 (Edwards et al., 2022) by adopting CLIP (Radford et al., 2021) to align molecule graphs with related text. At the same time, GIT-Mol (Liu et al., 2023a), inspired by BLIP2's Q-FORMER strategy (Li et al., 2023a), integrates graph, image and text information, using cross-attention and a variety of pretraining tasks. This multimodal approach significantly improves the effectiveness of MoIT5 (Edwards et al., 2022) in constrained molecule generation tasks.

**De novo Protein Generation.** Similar to molecule generation, this task focuses on designing new proteins in an unconditional manner (Hesslow et al., 2022; Ferruz et al., 2022; Nijkamp et al., 2023) or conditional manner, where proteins generated should fit user constraints (Wang et al., 2022a; Ram & Bepler, 2022; Watson et al., 2023). Using these LLMs, scientists can design new artificial proteins that could serve specific functions, such as binding to a particular receptor or acting as enzymes.

Unconstrained generation aims to delve into and map the extensive protein space with specialized protein-based LLMs, such as those using autoregressive models for amino acid sequence generation, demonstrating remarkable effectiveness in this realm (Madani et al., 2020; Hesslow et al., 2022; Nijkamp et al., 2023). A prime example of these LLMs is ProtGPT2, detailed in Ferruz et al. (2022). Trained on a wide range of protein sequences, ProtGPT2 excels in creating de novo protein sequences that mirror natural patterns. Its outputs, marked by ordinary amino acid propensities, are predominantly globular, resembling natural proteins. Intriguingly, comparative analysis with protein databases indicates that ProtGPT2's generated sequences distantly related to existing proteins can form valid structures based on AlphaFold2. This indicates ProtGPT2's ability to explore novel and valid protein space areas.

Constrained protein generation seeks to achieve the controllable design of novel proteins with specified cellular compartments or functions (Ferruz & Höcker, 2022). A typical practical constraint is generating protein sequences within the same family. This method ensures that new proteins retain the critical characteristics of a given group. Specialized LLMs like ProGen (Madani et al., 2023) and PoET (Watson et al., 2023) are utilized in this context. ProGen specifically generates sequences for a particular family, using a prefix like “Protein Family: Pfam ID”, and has produced proteins with efficiencies close to natural counterparts, even with low sequence identity. PoET, on the other hand, compiles multiple sequences from the same family to create new sequences, akin to forming a paragraph from multiple sentences, thereby preserving the family’s structural integrity.

Building on this, a more challenging yet direct path to drug development in constrained protein generation is the design of protein binders (McPartlon & Xu, 2023). This task requires intricately designing proteins for specific binding functions, necessitating a deep understanding of how the entire protein folds into a desired structure with a few functional residues underpinned by the protein’s overall structure. A more streamlined approach involves starting with a desired structure and, conditional on this structure, using a specialized LLM for inverse folding (Hsu et al., 2022). This inverse folding method starts with a complex protein structure and only seeks to leverage LLMs to convert the desired structure into protein sequences.

Watson et al. (2023) modified RoseTTAFold, a specialized LLM already equipped with profound structural knowledge and an understanding of protein folding, with diffusion modeling to create RFDiffusion. This innovative approach allows RFDiffusion to begin with randomly distributed residue frames and systematically denoise them toward a valid protein structure. More recently, the development of RFdiffusion All-Atom (RFdiffusionAA) (Krishna et al., 2024) extends this capability by integrating atomic-level detail to generate folded protein structures that specifically bind to small molecules, metals, and nucleic acids. By initializing the model with random distributions of residues around these small molecules, RFdiffusionAA can design highly specific binding pockets that are validated both computationally and experimentally. Subsequently, RFDiffusion can restructure the sequence in an inverse folding manner for greater accuracy, facilitating the design of diverse functional proteins based on simple molecular specifications. This method complements and enhances the potential of RoseTTAFold by enabling not only constrained protein design but also the creation of entirely new protein structures around various molecular targets. Notably, most of these constrained-based LLMs can also generate unconstrained protein structures, either by iteratively exploring the constrained space or by initializing the model via a random

process.

Specialized protein-based LLMs above have shown considerable promise; however, despite evidence that they can design protein sequences that align with user requirements, it remains to be seen whether these LLMs, trained on natural protein sequences, can generalize beyond these to unnatural sequences. To address this question, Verkuil et al. (2022) leveraged the ESM protein language model (Rives et al., 2021) to design novel protein structures in unconstrained and constrained scenarios. By experimentally validating the generated proteins, it achieved a 67% success rate in creating functional proteins, some of which bear minimal similarity to known proteins. The impact of this validation on drug discovery and development is significant: it substantiates innovative LLM approaches for creating biologics and therapeutic proteins, potentially speeding up the creation of new protein-based treatments. Notably, with the advent of ESM3 (Hayes et al., 2024), the newer version of ESM, improvements in the ability to predict and design complex protein structures have been demonstrated.

More recently, general-purpose LLMs represent a newly emerging de novo protein generation technology. ProteinDT (Liu et al., 2023d), a multi-modal framework, incorporates textual descriptions for protein design. To train ProteinDT, a dataset of 441K text and protein pairs was constructed. By training on these pairs, ProteinDT was able to achieve over 90% accuracy for text-guided de novo protein generation. Furthermore, ProteinDT can be used to perform text-guided protein optimization tasks, as discussed in section 3.2.4.

**Protein-ligand Interaction Prediction.** Understanding the mechanisms underlying the interaction between a drug (ligand) and its protein target is fundamental to drug discovery and development. In recent years, virtual screening via in silico methods such as molecular docking or classic predictive machine learning models has been at the forefront of streamlining drug development processes. These tools form the cornerstone for accelerating the early-stage identification of potential therapeutic agents. In the early stages of drug discovery, specialized large language models (LLMs) are utilized in two primary directions: (i) to use LLM directly as a backbone and (ii) as a standalone yet essential part of a more comprehensive predictive system. Both LLM use cases are highly instrumental in improving the efficiency and effectiveness of the drug development process.

Tools such as AlphaFold-multimer (Evans et al., 2021), used explicitly for protein-protein docking (as detailed in section 3.1.3 on protein-protein interaction), highlight the direct application of specialized LLMs as a backbone. Likewise, Singh et al. (2023) have employed protein LLM alongside molecular fingerprints for virtual screening. This technique has successfully identified binders with sub-nanomolar affin-

ity, underscoring LLMs' increasing impact and potential in refining the drug discovery and development process.

More commonly, incorporating protein or ligand embeddings and structural information from specialized LLMs as part of complex system architectures has become a standard in assisting accurate prediction and docking (Wang et al., 2022b; Lu et al., 2022; Jiang et al., 2022; Corso et al., 2023; Ketata et al., 2023; Koh et al., 2024). PSICHIC (Koh et al., 2024) demonstrated that learning from sequence data alone (protein sequence and ligand SMILES string) can surpass methods that rely on experimental 3D structures or protein-ligand complexes. More significantly, when learning from sequence data alone, PSICHIC demonstrated emergent capabilities in deciphering the mechanisms underlying protein-ligand interactions. It successfully identifies protein residues in the binding site and ligand atoms involved in these interactions, a capability achieved through training PSICHIC exclusively on sequence data without any information on specific binding sites or residue-atom interactions. Such achievements suggest a highly promising avenue for LLMs in processing extensive biochemical data. If successful, this approach can enable prediction and foster an understanding of how various chemical structures interact with different proteins, potentially revealing many hitherto unknown aspects of the protein-ligand interactions without costly experimental endeavors.

In general-purpose Large LLMs, Galactica is equipped with in-depth general scientific knowledge and has also been jointly trained to predict the docking scores of protein-ligand sequences (Taylor et al., 2022). It has demonstrated reasonable correlation with experimental results in certain instances. It is intriguing to consider the potential of developing an LLM that can directly understand, interpret, and predict protein-ligand interactions at a molecular level while also encompassing general knowledge.

### 3.2.3. ADMET PREDICTION

The prediction of absorption, distribution, metabolism, excretion, and toxicity (ADMET) attributes of compounds is a critical phase during the "Hit to Lead" and "Lead Optimization" stages of drug development. This task is essential to distinguish compounds with favorable pharmacokinetic profiles from those with negative characteristics, ensuring the progression of only the most promising drug candidates. Predicting molecular properties for multiple scientific areas, including physiology, physical chemistry, biophysics, and quantum mechanics, is the core of molecular property prediction in drug discovery. In recent years, both specialized and general-purpose large language models (LLMs) have shown remarkable predictive capabilities in the field of ADMET prediction, leveraging their ability to learn large amounts of data.

Specialized LLMs are usually trained on large datasets of SMILES strings and then fine-tuned for specific downstream property prediction tasks. For instance, ChemBERTa, based on the extensive PubChem 77M dataset, has shown comparable performance with traditional machine learning approaches like random forests and support vector machines. Similarly, SMILES Transformer, which employs an encoder-decoder architecture trained on over 861,000 unlabeled SMILES strings, has reached state-of-the-art performance. Recently, large-scale transformer-based models such as Molformer and BARTSMILES have demonstrated new state-of-the-art performance, though they require vast training time and resources.

General LLMs, on the other hand, either find knowledge to augment traditional machine learning models or are fine-tuned for specific tasks. LLM4SD can synthesize rules from literature and data sources that allow even random forest models to outperform all state-of-the-art methods for most tasks. Galactica and other models pre-trained on a significant amount of scientific literature have demonstrated capabilities in molecular property prediction with simple text instructions. Additionally, previous studies have highlighted the potential of other models such as GIT-Mol and MolT5, which could also yield satisfactory results after being fine-tuned on downstream tasks. GPT-4's extensive training on diverse text data enables it to provide valuable insights in the field, although adapting it for specific molecular property predictions might be challenging due to its proprietary nature.

### 3.2.4. LEAD OPTIMIZATION

Lead optimization is a significant step that aims to modify the drug candidate molecular structure or protein sequence to enhance its potency, safety, and stability. This process is usually carried out by chemists or biologists who modify according to their knowledge and experience. However, this process is time-consuming and requires much effort because it may take several attempts before arriving at the desired outcome. LLMs can assist with this task by utilizing statistical analysis on large data sets to predict how altering the structure of a compound would affect its properties. This feature supports more effective decision-making for chemists, minimizing the number of trials required for optimizing compounds.

**Molecular Optimization.** Molecular optimization is a complex task that involves modifying a molecular compound's structure to enhance its efficacy, stability, and safety. This process can be divided into two major categories: uncontrolled and controlled. In uncontrolled optimization, the core scaffold will be preserved but the LLMs will randomly modify the surrounding functional groups with external guidance to improve property values. In contrast, controlled

optimization means users can specify parts of a molecule to be optimized for a given property. Through editing, the final compounds are expected to have enhanced properties. Compared with uncontrolled approaches, these methods allow chemists to have a greater degree of specification.

Specialized LLMs can aid in molecular optimization through both uncontrollable and controllable strategies. For uncontrollable optimization, models like the Reinvent series (Blaschke et al., 2020; Loeffler et al., 2023) and MERMAID (Erikawa et al., 2021) use reinforcement learning to ensure that the synthesized molecules retain the desired structural scaffolds while enhancing properties such as potency, stability, or drug-likeness, guided by external models like drug-likeness filters or predictive algorithms. Specifically, MERMAID (Erikawa et al., 2021) incorporates a Monte Carlo Tree Search (MCTS) strategy, adeptly navigating potential molecular modifications to find the best ones. Apart from reinforcement learning, alternative methods such as fine-tuning and pretraining are also employed to enable specialized LMs with molecular optimization capabilities. An example is LigGPT (Bagal et al., 2021a), which focuses on generating molecules with specific scaffolds and desired properties. This model uses a trained transformer decoder architecture to reconstruct the original SMILES strings with the given scaffold and property information. For controllable optimization, Transformer-R (He et al., 2021a) and He et al. (He et al., 2021b) demonstrate its molecular optimization capability based on matched molecular pair (MMP) analysis. This model successfully considers property constraints and crucial components of SMILES strings while producing the required source R-GROUP for targeted molecular optimization. Improving these two approaches, C5T5 (Rothchild et al., 2021) does not rely on molecular pair data for training. This method is trained to recover masked IUPAC names by using specific tokens denoting the property range of the molecule. It shows successful applications for logP, logD, PSA, and Refractivity among other properties.

Recently, general LLMs have emerged, which present a novel avenue for amalgamating human expertise with LLM capability. The first one to consider is MoleculeSTM (Liu et al., 2023b), an application of multimodal learning for learning molecular structures from textual descriptions through contrastive learning. A dual-phase strategy is employed to enable a generative model to perform molecule editing. The first step involves training an adapter that aligns a molecule generative model with a joint representation space. In contrast, the second step focuses on fine-tuning the latent space to minimize differences between the generated molecule and given molecule-text instructions. An approach called ChatDrug (Liu et al., 2023c) significantly outperforms MoleculeSTM (Liu et al., 2023b) attributed to its integration with ChatGPT, which provides conver-

sational capabilities. This agent-driven technique applies an iterative refinement procedure combining information retrieval and domain-specific feedback, improving drug optimization workflow. ChatDrug incorporates modules like the Prompt Design for Domain-Specific (PDDS) module, which uses extensive prompt engineering from language models like LLMs. Later on, the Retrieval and Domain Feedback (ReDF) module will help find molecules based on specific requirements.

Additionally, GPT-4 has demonstrated basic abilities to optimize molecular structures, even though it was not explicitly developed for this purpose (OpenAI, 2023). However, there are certain limitations; GPT-4 can innovate upon existing compounds without incremental feedback and continuous checks from humans. It can easily lead to inaccurate responses.

**Protein Optimization.** Like molecular optimization, protein optimization involves modifying the structure to enhance protein functionality and safety. In this field, biochemists and molecular biologists meticulously adjust proteins, a process that can be laborious and iterative. LLMs can contribute by offering predictions on how structural changes impact protein properties. Specifically, in the development of antibody drugs, language models can be used to consider multiple essential factors, including improving antigen binding, reducing immunogenicity, enhancing stability, and preventing high viscosity or polyspecificity (Beck et al., 2017; Nichols et al., 2015; Raybould et al., 2019).

In uncontrolled optimization, ESM (Rives et al., 2021), which is a protein-based LLM trained on diverse protein sequences with general evolutionary information, has been used to suggest evolutionarily viable mutations that could help enhance fitness across protein families (Hie et al., 2023). This work relies on evolutionary plausible mutation that generally improve fitness across proteins rather than specific properties. As a proof-of-concept, this work found notable improvements in matured IgG antibodies' affinity against different viral antigens achieved with minimal testing of the variants through only two rounds of evolution.

In controlled optimization, protein hallucination and inpainting have been proposed (Wang et al., 2022a) that resembles constrained optimization have been proposed to optimize proteins using a section of the protein masked off, and a protein LLM (Dauparas et al., 2022) is used to sample and refine the protein sequence while maintaining the original backbone structure.

Among the general-purpose LLMs, ProteinDT (Liu et al., 2023d) is a novel method that optimizes protein sequences using prompts encapsulating specific properties information. ProteinDT can understand both natural texts and protein sequences. Leveraging this ability, the method employs

two techniques: latent interpolation and latent optimization. Latent interpolation merges text prompt and protein sequence representations, while latent optimization aligns which aligns a token-level latent code with both text and protein representations. Experiments showcase that LLMs can conduct protein optimization, including structure, stability, and peptide binding optimization. Adopting such a strategy indicates a novel way in protein engineering for achieving text-guided, exact transformations on protein structures aiming at their required features.

### 3.3. Assistance

In the same way that researchers in the "Understanding Disease Mechanisms" phase require a diverse array of information sources, those involved in drug discovery also need access to various relevant resources. These resources can include, but are not limited to, comprehensive compound libraries, up-to-date research publications, and extensive patent landscapes.

To gather this information, a procedure known as information retrieval is used, which involves using General LLMs with web searching and knowledge retrieval abilities to source information from various sources, such as research articles, compound databases, and patent documents. Additionally, tools like Galactica (Taylor et al., 2022) and GPT4 (AI4Science & Quantum, 2023) can assist in clarifying scientific concepts and data, helping researchers achieve a deeper understanding.

### 3.4. Clinical Trials

Clinical trials represent the last stage of the drug development pipeline and are essential to evaluate a drug candidate's safety and efficiency. These trials take place in four phases, each serving a specific purpose. Phase 1 involves a small number of healthy volunteers who are administered the compound to evaluate its safety and tolerability. Phase 2 involves a larger group of patients to evaluate efficacy and side effects. Phase 3 is the last testing stage performed on many patients after an optimal treatment selection. In this phase, the new treatment is compared to existing treatments to understand its differences. The final phase is postmarketing surveillance, during which the drug is monitored for adverse effects.

Using LLMs, one of the most significant areas where they can be used is clinical practice, result analysis, and clinical assistance. The subsequent paragraphs will delineate the specific downstream tasks for each category of these tasks and how LLMs can be incorporated into them to enhance drug discovery further.

#### 3.4.1. CLINICAL PRACTICE

In the realm of clinical trials, practitioners are typically tasked with four core responsibilities: coding ICD, matching patients with trials, predicting outcomes, and planning the trials themselves. These responsibilities span various areas and have traditionally been fulfilled by experienced practitioners who rely on their knowledge and expertise. Hence, clinical trial practitioners face the challenge of reading and understanding large amounts of information such as electronic health records (EHRs), trial eligibility criteria (ECs), trial protocols, and outcome reports. Fortunately, general LLMs have emerged as a promising solution for accelerating these processes, as they excel at extracting and handling information from large volumes of text data.

**ICD Coding.** ICD coding is an essential practice in clinical practice, which deals with assigning ICD codes to patient records. This is usually a time-consuming and laborious process. By analyzing vast amounts of EHR data, LLMs can predict the most suitable codes for Electronic Health Records (EHRs), thus enabling clinical practitioners to make more enlightened decisions and streamline the process.

Shi et al. created a groundbreaking system (Shi et al., 2017) that utilizes a character-aware LSTM-based network to process diagnosis descriptions from hospital admission records more efficiently. Similarly, Xie and Xin (Xie & Xing, 2018) developed a tree-of-sequences LSTM network to encode diagnosis descriptions and ICD codes. Inspired by Shi et al.'s work (Shi et al., 2017), they added adversarial learning to align various writing styles using an attentional matching module with isotonic constraints. These enhancements improve code assignments and prioritize more significant codes.

Some recent studies have used more contemporary and up-to-date frameworks for their language models. For example, BERT-XML (Zhang et al., 2020b) incorporates BERT pre-training with multi-label attention to encode EHRs and later uses a multi-label classification model to predict ICD codes from EHRs. This approach takes advantage of the capabilities of the BERT framework to increase the accuracy of code assignment. Another example, PLM-ICD (Huang et al., 2022) adapts domain-specific pretrained language models such as BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2021), and RobBERTa-PM (Liu et al., 2019b), for ICD coding by fine-tuning them. It also uses segment pooling and label attention to increase efficiency and accuracy in clinical coding contexts.

**Patient-Trial Matching.** When it comes to matching patients with clinical trials, the process relies on the use of electronic health records (EHRs) to identify viable options based on the patient's medical history. Historically, this task was performed manually by physicians and data an-

alysts who would sift through patient demographics and pre-screening eligibility factors to pinpoint the most suitable trial. However, this approach can be time-consuming and fraught with errors due to the complexity and diversity of trial criteria.

To overcome these challenges, preliminary works typically encode EHRs and eligibility criteria into an embedding pair and then calculate the match score through similarity. An example of this is the cross-modal framework called Deep-Enroll (Zhang et al., 2020a), which utilizes BERT to capture eligibility criteria from the text-based patient records, using a hierarchical structure of latent representation. This framework finally observed a notable 12.4%

COMPOSE (Gao et al., 2020a) significantly improves Deep-Enroll by utilizing a dual pathway encoding framework and a composite loss function. This approach effectively separates inclusion and exclusion criteria. The ECs pathway employs BERT and convolutional neural networks (CNNs) for robust encoding, while in the EHR pathway, hierarchical memory networks are deployed to organize medical concept hierarchies systematically. COMPOSE (Gao et al., 2020a) then dynamically interacts with these memories using EC embeddings, which help it to choose the most precise matches.

Recently, there have been several methods harnessing general-purpose LLMs to facilitate patient-trial matching based on LLMs reasoning ability. Med-monoT5 (Pradeep et al., 2022) is a T5-based system fine-tuned on medical passage ranking tasks that follows a zero-shot approach. It evaluates clinical trial documents' relevance to patient descriptions utilizing specifically designed templates. It employs a two-stage fine-tuning process on general and medical datasets, leveraging a sliding-window approach to handle lengthy text fields for matching patients with appropriate clinical trials. Hamer et al. (Hamer et al., 2023) use InstructGPT (Ouyang et al., 2022) to assist physicians in determining patient eligibility for clinical trials. Employ prompting strategies such as one-shot, selection-inference, and chain-of-thought—to parse and analyze the criteria. While this automation has been shown to potentially reduce up to 90% of the workload, achieving about 72% accuracy in screenability, it is not without issues. Overconfidence in interpreting ambiguous criteria and the risk of generating inaccurate content necessitate continued supervision by medical professionals to ensure reliability. Another pioneering work, TrialGPT (Jin et al., 2023c), uses an architecture that predicts criterion-level eligibility and provides detailed explanations. These explanations are aggregated to rank and exclude candidate clinical trials based on free-text patient notes. Although TrialGPT (Jin et al., 2023c) correlates well with expert annotations, its occasional errors highlight the limited medical knowledge of GPT 3.5 and the need for their

careful integration into clinical trial matching processes.

**Clinical Trial Planning and Prediction.** Planning a clinical trial involves several labor-intensive activities, including searching for historical trials, designing trial criteria, and selecting suitable sites. To accelerate this process, general LLMs have been introduced in the domain. These models utilize historical data on trials, patient outcomes, and demographic trends to provide informed suggestions.

The first step in preparing a clinical trial is the search for historical trials, simplified by embedding trials into documents using document embedding methods. Although some traditional statistical models like BM25 (Trotman et al., 2014) and TF-IDF (Ramos et al., 2003) can produce such embeddings, their performance often needs to catch up to that of deep learning. Deep learning models such as Doc2Vec (Le & Mikolov, 2014) and BERT (Devlin et al., 2019) provide more sophisticated embeddings but still face challenges in medical-specific retrieval tasks. These gaps are narrowed by medically tailored adaptations of BERT (Devlin et al., 2019), including clinical BERT (Alsentzer et al., 2019), clinical bioBERT (Alsentzer et al., 2019), TrialBERT (Wang & Sun, 2022), and Med-monoT5 (Pradeep et al., 2022), which have been fine-tuned on substantial medical literature to provide highly accurate retrieval results. Going beyond these developments, the Trial2Vec (Wang & Sun, 2022) framework extends from TrialBERT (Wang & Sun, 2022) and utilizes hierarchical contrastive learning to generate global and local embeddings that incorporate semantic meaning based on document meta-structure. Additionally, cliniDigest (White et al., 2023) employs the summarization capabilities of GPT-3.5 to provide up-to-date, concise summaries of clinical trials, thereby facilitating quick decision-making for upcoming trials.

AutoTrial utilizes a two-stage training method with GPT-2 as its backbone to automate trial criteria design and clinical trial planning. This model employs the decoder architecture, learns from vast documents regarding previous trials during pretraining, and then gets its task-specific fine-tuning to generate exact trial criteria from given specifications. Its combination of a hybrid prompting strategy and multi-stage training makes it easy to adapt without retraining or performance loss.

Trial site matching involves finding a suitable trial site for a clinical trial. Modern algorithms integrate multimodal data containing unstructured and structured information about a trial to rank sites. One such algorithm is PG-Entropy (Srinivasan et al., 2022), which exploits ClinicalBERT (Huang et al., 2020)-based encoding of trial criteria text in combination with structured clinical trial features via a list-wise policy learning approach. It also gives equal importance to ensuring sensitive attributes do not adversely impact any bias-related outcome. In contrast, FRAMM (Theodorou

et al., 2023) uses a deep reinforcement learning mechanism.

In this way, it is an effective remedy for the absence of data and for creating site representations using the masked cross-attention mechanism. In addition, it balances diversity and enrollment when choosing sites via a Deep Q-Value Network that helps in decision-making by defining the reward function that focuses on both these aspects.

To predict the success of a clinical trial, trial outcome prediction analyzes various trial-related variables, such as information about the disease, the drug, trial criteria, and trial protocol. LLMs can provide insightful predictions to help trial designers enhance the trial plan. This can mitigate the risk of adverse outcomes from suboptimal trial configurations.

Existing methods mainly use LLMs to encode clinical trial information, which is then used to predict clinical trial outcomes. These LLM-based methods performed better than traditional machine learning methods such as logistic regression, MLP, and XGBoost (Chen & Guestrin, 2016). For instance, given drug, disease, and clinical protocol information, HINT (Fu et al., 2022) amalgamates comprehensive web data such as drug properties, disease characteristics, and clinical trial information to assist the model in predicting both phase-level and trial-level clinical trial outcomes. Specifically, it separately encodes the drug molecules, disease characteristics, and clinical protocols using pre-trained models with web knowledge. Then, it builds an interaction graph using these components to generate the final prediction.

SPOT (Wang et al., 2023c) and HINT (Fu et al., 2022) are both methods that aggregate clinical trials of the same topic into a sequence based on timestamps instead of making predictions for each trial individually. SPOT uses a sequenced meta-learning approach that begins with topic discovery and clustering, followed by an RNN structure for sequential predictive modeling. It then uses a meta-learning approach to optimize models for different topics. On the other hand, MediTab (Wang et al., 2023b) can be adapted to perform clinical trial prediction and achieves better performance than both HINT and SPOT, with the added advantage of using larger-scale tabular data from different sources. While DeepEnroll (Zhang et al., 2020a) and COMPOSE (Gao et al., 2020a) were initially designed for patient-trial matching, they can also be adapted for clinical trial prediction. However, they perform less effectively than previous LLM methods.

**Documents Writing.** The generation of various clinical documents, such as discharge summaries, clinical notes, and radiology reports, has traditionally been a time-consuming and laborious task performed by clinical practitioners. However, language models (LLMs) are increasingly being employed

to automate this process with their powerful text-generation ability.

For instance, in discharge summaries generation, Patel et al. (Patel & Lam, 2023) present a prompt to generate discharge using chatGPT automatically, while Shing et al. (Shing et al., 2021) have implemented an extractive-abstractive summarization pipeline. This method consists of a two-stage process that begins by extracting relevant sentences from clinical notes, followed by an abstractive summarization technique to transform these extracts into coherent summaries.

LLMs have also been shown to be highly effective in maintaining regular check-ins with patients, as demonstrated by (Webster, 2023). These models can assist in writing notes in patients' records and summarizing their issues, bridging the gap between conventional encounters with caregivers. In addition, Seppo et al. (Enarvi et al., 2020) generated medical reports from patient-doctor conversation transcripts using RNN and transformer models. To achieve this, they trained train RNN and transformer models using a dataset of around 800,000 orthopedic encounters. When both RNN and Transformer models were compared, the latter showed superior accuracy and training efficiency performance.

For generating reports from randomized controlled trials (RCTs), the RobotReviewer (Marshall et al., 2017) system has shown its capability to automatically generate reports summarizing critical information from RCTs.

Finally, multi-modal LLMs have been widely used for the task of radiology report generation. In the early days, specific language models were employed as encoders. An example is TieNet (Wang et al., 2018), an end-to-end CNN-RNN architecture with multi-level attention models. It involves merging image features and text embeddings from associated reports to improve disease classification accuracy and report quality (Wang et al., 2018). Taking a step further, Liu et al. (Liu et al., 2019a) presented a hierarchical generation strategy via CNN-RNN-RNN architecture with reinforcement learning. The approach focuses on balancing readability and clinical accuracy by considering Clinically Coherent Reward (CCR) to maintain the clinical relevance of the reports. For methods using more modern LLMs, MedViLL (Moon et al., 2022), which also uses the BERT (Devlin et al., 2019) architecture, concatenates visual and language feature embeddings and trains them on tasks like masked language modeling and image report matching to achieve an effective alignment of visual and language features. Another example, RadBERT (Yan et al., 2022), is a BERT-like system pre-trained on millions of radiology reports that can generate concise reports highlighting essential observations and conclusions. More recently, during human evaluation, Med-PaLM M (Tu et al., 2024), a multimodal medical language model, generated chest X-ray

reports that clinicians preferred over those by radiologists in up to 40.50% of cases.

### 3.4.2. PATIENT RESULTS

LLMs can help in predicting patient outcomes using patient visits data. There are two categories of downstream tasks in this domain: hospital-related and disease-related predictions. While the former type includes hospital readmission, length of stay, and mortality, the latter type focuses on disease onset, diagnosis, and morbidity.

**Patient Outcome Prediction.** Patient outcome prediction aims to predict a patient’s current or future health status. Language models help in this area by encoding vast amounts of electronic health records (EHRs) to predict future health outcomes, thereby enabling clinicians to make more informed decisions and potentially saving a substantial amount of time and resources. These LLMs-based methods generally achieved better performance than traditional machine learning methods such as logistic regression, MLP, and XGBoost (Chen & Guestrin, 2016).

There are LLMs focusing on hospital-related predictions. For example, ClinicalBERT (Huang et al., 2020) leverages the BERT (Devlin et al., 2019) architecture to understand clinical records and make readmission predictions. Similarly, NYUTron (Jiang et al., 2023), a BERT-like language model, is pre-trained on a comprehensive collection of clinical notes and fine-tuned for various tasks such as predicting mortality, comorbidity, hospital readmission, insurance denial, and length of stay. Another model, RAIM (Xu et al., 2018), processes multimodal EHR data—including clinical records, electrocardiograph waveforms, and vital signs—to predict outcomes like decompensation and length of stay. It considers both historical and current data points, focusing on the most relevant information for accurate predictions. StageNet (Gao et al., 2020b), which can predict decompensation and mortality, uses a stage-aware LSTM module that captures changes in patients’ health conditions over time. It integrates time information between visits and employs a stage-adaptive convolutional module to recalibrate understanding of disease progression, highlighting the most informative patterns for precise outcome prediction.

The second category of methods focuses on disease-related predictions, including disease onset, diagnosis, and morbidity. For instance, RETAIN (Choi et al., 2016) employs a two-level neural attention mechanism with recurrent neural networks (RNNs) to focus on crucial visits and variables for heart failure predictions. Building upon RETAIN, Dipole (Ma et al., 2017) uses a bidirectional RNN and three types of attention mechanisms to enhance diagnosis predictions, determining future medical codes for patient visits. With more recent LLMs, MediTab (Wang et al., 2023b) consolidate and align different types of medical data, predicting

patient outcomes such as morbidity.

### 3.4.3. ASSISTANCE

General-purpose language models (LLMs) can play a significant role in clinical trial assistance by helping patients understand trial-related information, assisting clinicians in retrieving and understanding relevant literature and patient data, and supporting pharmacovigilance efforts by identifying and reporting adverse events.

General-purpose LLMs, such as GPT-4, and Med-Palm2 are capable of understanding medical knowledge and explaining it in simple, accessible language (Kung et al., 2023; Thirunavukarasu et al., 2023). This ability can help patients better comprehend and participate in clinical trial opportunities. Clinicians can also utilize LLMs to efficiently retrieve relevant clinical trial literature and assess patient eligibility using advanced information retrieval capabilities. In pharmacovigilance, LLMs contribute to understanding drug-drug interactions (Luo et al., 2022; Taylor et al., 2022), providing deeper insights into drug safety. Their code generation capabilities streamline data analysis (OpenAI, 2023), enhancing the efficiency and speed of data interpretation.

## 4. LLMs Maturity Assessment

In this analysis, we evaluate the progress of two LLMs paradigms across 14 downstream task categories within three stages of the drug discovery and development pipeline: understanding disease mechanisms, drug discovery, and clinical trials. We categorize the maturity of these tasks using a four-tiered system ranging from “Not Applicable” to “Matured”. We begin by outlining the specifics of the criteria before detailing the maturity levels of the various tasks within each phase.

### 4.1. Maturity Criteria

The maturity of LLMs in the drug discovery and development pipeline is categorized into four distinct levels (Figure 6):

1. **Not Applicable:** The LLM paradigm is not suitable or relevant for the given downstream task.
2. **Nascent:** The LLM paradigm has been applied to the task in a preliminary, in silico setting only and lacks validation through real-world experiments.
3. **Advanced:** The LLM paradigm has moved beyond theoretical application, with its effectiveness validated through real-world experiments in relevant scenarios.
4. **Matured:** The LLM paradigm application has been integrated and deployed in practical, real-world envi-

ronments such as hospitals or pharmaceutical companies, with evidence demonstrating its effectiveness and utility.

## 4.2. Maturity Assessment of Downstream Tasks

For the maturity assessment, we consider 14 different downstream tasks to be carried out in three significant drug discovery and development stages. In understanding disease mechanisms, we focus on genomics analysis, transcriptomic analysis, protein target analysis, disease pathway analysis, and assistance. At the stage of drug discovery, we analyze chemistry experiments, in silico simulations, ADMET prediction, lead optimization, and assistance. Finally, clinical trials deal with clinical applications in clinical practices, patient results, and assistance.

### 4.2.1. UNDERSTANDING DISEASE MECHANISM

**Genomics Analysis.** Specialized LLMs have been recently created to encode information in the nucleotide sequences (Ji et al., 2021; Dalla-Torre et al., 2023; Li et al., 2023b) with any practical applications, such as genetic variant analysis (Le et al., 2022; Zhou et al., 2023). However, they are still nascent, and further experiments are required to validate their effectiveness.

Recently, general LLMs have emerged and are also in the nascent stage. There is still room for improvement in explaining the evolutionary processes of genomic data or designing DNA sequences, as indicated in recent studies (AI4Science & Quantum, 2023). This underscores the ever-developing field with its enormous potential for future breakthroughs.

**Transcriptomics Analysis.** There have been significant advancements in the real-world applications of specialized LLMs, such as the Geneformer LLM (Theodoris et al., 2023). This particular LLM has played an essential role in gene network analysis. It was able to distinguish between normal and cardiomyopathic cardiomyocytes. This process identified the genes responsible for network perturbations associated with hypertrophic and dilated cardiomyopathy, providing potential therapeutic targets like ADCY5 and SRPK3. The real-world effectiveness of these targets was confirmed through experimental validation using iPSC-derived cardiac microtissues with Titin truncating mutations (Theodoris et al., 2023). Hence, it is evident that specialized LLMs are in an advanced stage of development for analyzing transcriptomic data and deciphering disease mechanisms.

In contrast, general LLMs are still in the early nascent stage of development for transcriptomic analysis. Research is underway to explore auxiliary tasks in this field, such as automating cell type analysis (Hou & Ji, 2023) and analyzing data through code generation (AI4Science & Quantum,

2023).

**Protein Target Analysis.** Specialized LLMs for protein target analysis have significantly matured following the breakthrough of AlphaFold2 (Jumper et al., 2021). AlphaFold2 is now a comprehensive and readily accessible database (Varadi et al., 2022), with various applications in structure-based drug discovery and vaccine development (Varadi & Velankar, 2023). One noteworthy accomplishment is the rapid development of a first-in-class hit molecule for a novel target, CDK20, without an experimental structure, achieved within 30 days from target selection and requiring the synthesis of only seven compounds (Ren et al., 2023). Additionally, ESM (Rives et al., 2021), a prominent protein language model, has been developed into a web server focusing on GPCR proteins. This tool analyzes their signaling and functional repertoire (Matic et al., 2022) and identifies compounds with subnanomolar affinity (Singhal et al., 2023).

While general LLMs like GPT-4 (OpenAI, 2023) have demonstrated potential in analyzing complex scientific data, including protein analysis, they remain in a nascent stage of development within protein target analysis. ProteinChat (Guo et al., 2023) is another example demonstrating how these models can label protein structures based on user prompts. However, these developments primarily focus on generating informative answers based on embeddings without extensive real-world validation. Thus, this indicates that the field is still evolving.

**Disease-pathway analysis.** We are witnessing specialized LLMs reach an advanced stage in disease pathway analysis. These specialized LLMs have made significant contributions, particularly in genomics, transcriptomics, and protein target analysis. Notably, these fields play a critical role in the comprehensive analysis of disease pathways. A notable recent breakthrough is the use of transcriptomic LLM, Geneformer (Theodoris et al., 2023), for gene network analysis, which has undergone laboratory validations and illustrates the capabilities of these models in dissecting disease pathways (Theodoris et al., 2023).

General LLMs have also reached an advanced stage in disease-pathway analysis. For instance, Insilico Medicine, a biotechnology company, already offers ChatGPT integration with its PandaOmics target discovery platform to analyze disease pathways (Savage, 2023). While the PandaOmics target discovery already incorporated general LLMs for this purpose, these tools' widespread adoption and broad usage still need to be fully realized.

**Assistance.** General LLMs have attained a mature development stage, greatly aiding researchers in information retrieval and knowledge discovery for the understanding of disease mechanisms (AI4Science & Quantum, 2023; Sav-

age, 2023; Toufiq et al., 2023). These models possess exceptional capabilities in mining and synthesizing extensive scientific and medical literature, allowing us to understand these mechanisms better. Additionally, their skill in creating and interpreting knowledge graphs (Savage, 2023) is crucial in mapping gene networks and elucidating gene-disease relationships (Luo et al., 2022). Additionally, general LLMs have proven effective in simplifying complex medical and genetic concepts (Jeblick et al., 2023), thus making technical knowledge more accessible and enhancing education and communication in the medical field.

#### 4.2.2. DRUG DISCOVERY

**Chemistry Experiments.** For chemistry experiments, while the utilization of specialized LLMs is still in its nascent stages, general LLMs have advanced considerably. They are now used in sophisticated chemistry experiments (Figure 6). Specialized LLMs, typically conducted in silico, are considered inferior to their general counterparts based on their performance in retrosynthetic planning and reaction prediction (Boiko et al., 2023; Bran et al., 2023). This is mainly due to the tool use capabilities of general LLMs, which include using tools, reading scientific literature, and searching online to assist in molecular synthesis.

In real-world laboratory settings, general LLMs have demonstrated their effectiveness in synthesizing molecules and controlling robotic arms (Boiko et al., 2023; Bran et al., 2023; Yoshikawa et al., 2023). These successes underscore the potential for general LLMs to impact the field of chemistry significantly, presenting new ways of conducting experiments and facilitating discoveries. However, despite these promising developments, there needs to be more evidence of general LLMs being deployed in the industry, e.g., pharmaceutical companies. This gap highlights the need for further research and development to extend the use of general LLMs in chemistry experiments.

**In-silico Simulation.** The use of specialized LLMs in industry is becoming increasingly common, with tools like AlphaFold Multimer (Evans et al., 2021) being used for protein-protein complex prediction. More recently, these tools have expanded to include protein-ligand complexes, broadening their scope to small molecules and nucleic acids. Additionally, In Silico Medicine, a biotechnology company has developed GENTRL (Zhavoronkov et al., 2019) and Chemistry42 (Ivanenkov et al., 2023), which utilized specialized LLMs to identify potent DDR1 kinase inhibitors and generate novel molecular structures, respectively. These structures exhibit optimized properties and have been validated through extensive in vitro and in vivo studies. Similarly, Molformer (Ross et al., 2022), developed by IBM, demonstrates promising results in generating candidate molecules aimed at inhibiting the SARS-CoV-2 virus and

developing antimicrobial peptides (Das et al., 2021).

Conversely, applying general language models remains primarily confined to in-silico environments. These models face substantial challenges in scientific understanding and quantitative analysis. For instance, GPT-4 (AI4Science & Quantum, 2023) struggles with interpreting and generating SMILES strings. Additionally, general LLMs often lack the precision required for quantitative tasks, leading to suboptimal performance in simulations, such as predicting binding affinity (Razeghi et al., 2022).

**ADMET Prediction.** For specialized LLMs, IBM's Molformer (Ross et al., 2022) has built a cloud-based platform that allows chemists to conduct real-time molecular screening and efficient molecular properties prediction.

In the advanced stage, LLM4SD (Zheng et al., 2023) uses general-purpose LLMs such as Falcon (Penedo et al., 2023) and Galactica (Taylor et al., 2022) as backbones to extract meaningful hypotheses from ADMET data. These assumptions have been shown to outperform state-of-the-art professional baselines when applied to traditional methods such as random forests. Pharmacologists validate the majority of these rules, ensuring their relevance and efficacy.

**Lead Optimization.** Specialized LLMs, for lead optimization have been validated via real-world experiments. For instance, in molecular optimization, Moret et al. (Moret et al., 2023) developed a chemical language model that facilitated the discovery of a new PI3K $\gamma$  ligand with sub-micromolar activity. Similarly, in protein optimization, Hie et al. (Hie et al., 2023) employed a language-model-guided process to enhance seven antibodies' affinity and effectiveness against Ebola and SARS-CoV-2 through minimal variant screening and lab evolution.

However, general LLMs are still in the early stages of development and have only undergone in-silico testing. One of the main challenges in adapting general LLMs to this application lies in the profound understanding of scientific language, which is essential for lead optimization.

**Assistance.** General LLMs have reached an advanced stage in information retrieval and explanation for drug discovery. BenevolentAI, a company specializing in AI-enabled drug discovery and development, is investigating ChatGPT retrieval plug-ins that can search through personal or company documents to provide medical answers (Savage, 2023). Furthermore, GPT-4 showcases substantial coding capabilities; it aids in various coding tasks related to drug discovery, such as data downloading and data preprocessing (AI4Science & Quantum, 2023).

#### 4.2.3. CLINICAL TRIAL

The clinical trial phase mainly involves general text data, including electronic health records and trial protocol documents. Therefore, a specialized LLM is generally not suitable at this stage.

**Clinical Trial Practice.** In clinical trial practice, there are tasks including ICD coding, patient-trial matching, and clinical trial planning. Although general LLMs in this field are still in the early stages of implementation, their potential is promising. Despite the lack of extensive real-world testing and application evidence, the rapid development and improvement of the models, particularly in their ability to understand and process medical knowledge (Singhal et al., 2023), signals a promising future.

**Patient Outcome Prediction.** Predicting patient outcomes is one area where the LLMs show promise, helping doctors diagnose and predict patient outcomes. General LLMs specialize in handling large amounts of unstructured data in electronic medical records. For example, Jiang et al. (Jiang et al., 2023) developed NYUTriton, an advanced platform that interfaces with the electronic health record system at NYU Langone Health System. Deployed across a network of hospitals and outpatient facilities in New York, the system performs tasks such as predicting in-hospital mortality, estimating a comprehensive comorbidity index, and predicting 30-day all-cause readmissions. Similarly, Google's MedPalm2 (Singhal et al., 2023) was introduced for medical question-answering tasks and achieved an accuracy of 86.5%, much higher than the approximate medical passing score. This advanced technology is being tested in real-world settings with select client groups, including VHC Health VA, affiliated with Mayo Clinic.

**Assistance.** General LLMs have matured in clinical assistance. These assistants can aid physicians and administrative staff in tasks such as document writing. For example, Webster & Paul (Webster, 2023) have demonstrated the effectiveness of these models in generating clinical notes, maintaining regular check-ins for patients with chronic conditions, and summarizing patient issues. Recently, Oracle unveiled a Clinical Digital Assistant that can handle administrative tasks through voice commands. Additionally, Google's MedPalm2 has been implemented in real-world scenarios for information retrieval and knowledge explanation (Singhal et al., 2023).

## 5. Future Direction

This section explores the future direction of LLM applications in drug discovery and development. We discuss nine areas that require enhancement: integrating biological insights, addressing ethical and privacy concerns and preventing misuse, addressing fairness and mitigating bias,

addressing hallucination, improving multi-modality, improving context window, improving spatio-temporal understanding, and integrating specialized LLMs with general LLMs.

### 5.1. Integrating Biological Insights

Improving the scientific understanding of language models (LLMs) is crucial for their successful application in drug discovery and related downstream tasks. To be practical, specialized and general LLMs require a deep understanding of scientific concepts, such as terminologies, and languages, such as SMILES and IUPAC nomenclature.

A case is molecular generation and editing that requires accurate interpretation and manipulation of these specific, highly scientific languages. Microsoft performed one such evaluation using GPT-4, showing that the model could not understand SMILES strings well (AI4Science & Quantum, 2023). Similarly, in clinical trials, LLMs must be familiar with medical vocabulary from Electronic Health Records (EHRs) and patient profiles regarding diagnoses and clinical criteria. Such knowledge is necessary for reliable patient clinical trial matching, a complicated multistage process.

Benchmarking the scientific understanding of LLMs is a pivotal step in their development and deployment. Though medical QA datasets are available, they may need to be more comprehensive and practical as it is hard to discern whether the model uses scientific knowledge or reproduces answers based on memorization. Hence, more rigorous and customized methods should be used to estimate and develop LLM's scientific understanding.

Arguably, the most applicable developments come in developing explanatory capabilities at a big-data scale in biochemistry and biophysics. These technologies primarily include high-throughput secondary structure prediction technological developments and increasingly robust statistical mechanics predictions.

Recently, high-throughput secondary structure generation for DNA and RNA represents a significant advancement in molecular biology and bioinformatics. These technologies combine in vitro biophysical probing techniques with statistical methods and technologies to predict the secondary structures of nucleic acids at a large scale, which is crucial for understanding their function, interaction, and role in various biological processes. Developing high-throughput sequencing-based technologies has provided a wealth of genomic and transcriptomic data. Future research will increasingly integrate computational predictions with experimental data, like SHAPE (Loughrey et al., 2014) (Selective 2'-Hydroxyl Acylation analyzed by Primer Extension) or DMS (Zubradt et al., 2016) (dimethyl sulfate) mapping, to refine secondary structure models.

Recent developments in statistical mechanics have also sig-

nificantly enhanced the robustness of its predictions, which have important implications for LLMs in drug discovery. In particular, enhanced sampling techniques and multiscale modeling are advanced computational strategies used to overcome limitations in traditional molecular simulations, providing more accurate and comprehensive insights into biomolecular processes. Metadynamics (Bussi & Laio, 2020) is a popular enhanced sampling technique that facilitates the exploration of a molecule's energy landscape more efficiently than conventional molecular dynamics simulations. Metadynamics can be applied to study the binding mechanisms and conformational changes of ligands and proteins in drug discovery. The QM/MM (Bösel et al., 2021) (quantum mechanics/molecular mechanics) approach is a prime example of multiscale modeling. In QM/MM simulations, the region of interest (e.g., a drug interacting with its binding site) is treated using quantum mechanics to model the electronic interactions accurately. In contrast, the surrounding environment (e.g., the rest of the protein and solvent) is treated using classical molecular mechanics, balancing accuracy and computational efficiency. QM/MM simulations can be particularly valuable for studying enzyme-catalyzed reactions, which are often targets in drug design.

The integration of advanced computational techniques developed in fields like statistical mechanics or molecular dynamics into large language models (LLMs) for drug discovery has been gradual due to several factors. Despite the significant progress in both domains over the last decade, the following reasons explain why some advanced techniques have not yet been widely adopted. This is potentially due to interdisciplinary gaps, validation and standardization protocols, and data compatibility. As interdisciplinary collaboration grows and computational resources become more accessible, we will likely see more advanced techniques integrated into LLMs, enhancing their effectiveness and impact in drug discovery.

## 5.2. Addressing Ethical, Privacy Concerns, & Preventing Misuse

The ethical issues in using LLMs for drug discovery are diverse and involve responsibility, fairness, and the potential for unintended consequences. One primary ethical issue revolves around accountability for decisions made or influenced by LLMs. As these models play an increasingly important role in drug development, the question arises of who is responsible for the outcomes, whether positive breakthroughs or negative results. This is particularly challenging given the often opaque nature of the LLM decision-making process. The rapid pace of innovation in this field has raised concerns that regulations and ethical guidelines must be updated.

Privacy issues related to LLMs are essential because they

can memorize training data. For example, when it comes to sensitive multi-omic data collected during patient typing, it is critical to ensure the data is anonymised and therefore cannot be directly related to the patient. Recently, a study showed that adversaries could extract large amounts of training data from LLM (Nasr et al., 2023). This is achieved through extractable storage, where the model accidentally leaks training data in response to specific queries. For example, a model can be reverted to its original language modeling behavior by prompting ChatGPT with a sequence that causes it to break from Chatbot-style generation. This difference can cause the model to output fragments of its training data. Researchers developed a new divergence attack for ChatGPT that exposed training data 150 times higher than usual. These results indicate that current alignment techniques do not entirely prevent memory leakage, which raises serious ethical and safety issues.

Potential misuse of LLMs in areas such as drug discovery necessitates a carefully balanced approach that prioritizes safeguarding against risks without impeding technological advancement. These concerns are raised as LLMs can be used intentionally for malicious purposes as described in MegaSyn2 model (Urbina et al., 2022). They show the MegaSyn2 model (Urbina et al., 2022), initially developed to discover therapeutic inhibitors and exploit them to create highly toxic substances or chemical warfare agents. However, it is crucial to acknowledge that while LLMs democratize the knowledge of compound synthesis, this does not automatically lead to more accessible access to materials for synthesizing dangerous substances due to existing strict regulations. Given the LLMs' potential benefits in areas like medical science, they need more relaxed regulations to ensure their development. This also requires a balanced view considering system safety while avoiding slowing down technology development.

## 5.3. Addressing Hallucination

The use of language large models (LLMs) in drug discovery is growing. However, their tendency to "hallucinate"—generating irrelevant or incoherent responses—presents a major challenge. Researchers and clinicians must be cautious, as these errors can lead them in the wrong direction with false information. These errors can be propagated, which leads to serious consequences.

For example, hallucinations can potentially result in the identification of incorrect biological targets or relationships, driving research in unproductive directions and wasting valuable resources. Some biotech companies are now using LLMs to interact with knowledge graphs of biological entities, such as genes, proteins, and diseases, to identify potential targets for drug development (Savage, 2023). This issue can lead to inaccurate optimization or modifications

in molecule and protein design. For instance, hallucinated molecular structures could be chemically invalid or impractical for synthesis. In clinical settings, where the stakes of inaccuracy include diagnosis and data interpretation, it can lead to serious, even life-threatening consequences.

To address hallucinations in LLMs for drug discovery, mitigation strategies can be used to guide the model toward generating more accurate and relevant answers. Knowledge editing is an approach that fills gaps in the understanding of the model by modifying some parameters or integrating plugins from other sources (Ji et al., 2023). This entails grounding LLMs in retrieval-augmented generation (RAG) with external documents for increased factuality and relevance in their outputs (Ji et al., 2023). Similarly, fine-tuning LLMs on debiased datasets helps remove knowledge shortcuts and spurious correlations that might stem from biased sampling (Ji et al., 2023). Additionally, many techniques can be applied to improve knowledge recall. Chain of Thoughts Prompting and similar techniques help generate outputs based on factuality and relevance (Ji et al., 2023). Finally, the effort of perfecting the decoding algorithms, such as Factuality Enhanced Decoding and Faithfulness Enhanced Decoding, ensures that the output generated is entirely in line with actual data and customer requests, significantly enhancing both the accuracy and reliability of LLMs.

#### 5.4. Addressing Fairness & Bias

Fairness and bias should be among the top priorities when creating and using LLMs in drug development. Biases' effects are evident in medical contexts, which can lead to possible inaccuracies, discrimination between different groups of patients, or even potentially harmful consequences. Biases inherent within data collection, model training, and application channels may prolong disparities, thus negatively impacting the integrity and efficacy of medical treatments.

Bias can originate from various sources, including the need for more data on rare diseases or specific populations, leading to underrepresentation. For example, biases are particularly evident in clinical trials through disparities in participation rates among different populations, influenced by geographic, economic, and cultural factors. Skewing predictive models developed based on these datasets can result in less effective or inappropriate solutions for underrepresented groups. Furthermore, this problem is accentuated when the models turn more toward common or widely examined points, like particular protein targets and demographic groups.

Enhancing the transparency and interpretability of LLMs is essential to address their biases. This can be achieved by employing various data sources, utilizing inclusive data collection and analysis approaches, and conducting rigorous ethical assessments. Furthermore, creating mechanisms

that allow for error rectification, promoting interdisciplinary cooperation, and initiating conversations concerning the responsible implementation of LLMs in drug development would help ensure equity in healthcare access.

#### 5.5. Improving Quantitative Analysis

The role of LLMs in the field of drug discovery has been expanding. One significant skill they need is to analyze vast amounts of numerical datasets. For instance, this applies to transcriptomic expression data interpretation for learning about disease mechanisms or predicting molecular properties while making drugs (Theodoris et al., 2023; Zheng et al., 2023). These examples indicate a growing trend towards relying on LLMs to manage particular problems within data-intensive pharmaceutical research.

LLMs, in general, while proficient in text generation and analysis, have shown limited success with data predominantly comprising numerical values, a critical aspect in drug research. For example, LLMs have historically faced challenges executing straightforward arithmetic operations like multi-digit multiplication (Dziri et al., 2023). They often resort to fabricating answers (OpenAI, 2023; Frieder et al., 2023). As some have argued (Testolin, 2023; Golkar et al., 2023), this can be due to the standard tokenization methods in LLMs that fail to accurately reflect the unique quantitative characteristics of numerical data, separating it from typical language inputs.

Recent explorations propose various methodologies to improve the encoding of numerical information for these LLMs (Thawani et al., 2021; Golkar et al., 2023). Approaches include digit-by-digit encoding (Gruver et al., 2023), base-10 formatting, and alignment between embedding distance against real numerical distance (Sundararaman et al., 2020). However, as LLMs are prone to relying on shortcuts and non-representative correlations in data (Dziri et al., 2023), they still struggle with interpolation and extrapolation in mathematical contexts within scientific fields (Grosse et al., 2023; Anil et al., 2022). Addressing this fundamental issue requires imposing an appropriate inductive bias that acknowledges the continuous nature of numbers, a critical step for advancing LLMs in drug discovery.

#### 5.6. Improving Multi-Modality

Multimodal Large Language Models (MLLMs) are advanced LLMs equipped to receive and process multimodal information (Yin et al., 2023). The application of MLLMs to drug discovery is promising as they can process diverse types of data, including videos, images, and experimental data. This property aligns with the nature of drug discovery, requiring diversified data sources like chemical structures, biological datasets, and scientific literature (Taylor et al.,

2022).

MLLMs are beneficial as they enable scientists to interact intuitively and flexibly with them. They can be accommodating when it comes to complex tasks in chemistry or biology, such as molecular modeling or clinical data analysis. These tasks involve multiple types of data, e.g., 3D protein and molecule structure and 2D medical images. Implementing MLLMs can open up an exciting frontier for research and may significantly optimize the effectiveness of laboratory investigations.

### 5.7. Improving Context Window

In drug discovery, LLMs usually must deal with vast amounts of biological data like sequences, which can easily exceed 2048 tokens (Koh et al., 2022a;b). However, many existing models have a restricted window of 2048 or 4096 tokens, e.g., Galactica (Taylor et al., 2022), and Falcon. Thus, these models are ineffective in processing multiple proteins and gene sequences. Protein sequences typically have 200-300 amino acids, while gene sequences can have over 26,000 nucleotides. Even LLMs with large window sizes of up to 128k tokens fail to thoroughly analyze such enormous volumes of input data. Typically, these models are good at interpreting the input's beginning and end but usually perform poorly for middle sections. This “forgetting” issue can lead to significant gaps in the analysis and interpretation of data.

Several potential solutions exist to address this challenge. One approach is to segment the input into smaller chunks and process them separately, then combine the outputs to generate a final result. To ensure no critical information is lost, intelligent segmentation strategies should be developed to understand the biological significance of different parts of the sequence. Another solution is implementing more sophisticated memory and attention mechanisms that help models better manage and utilize longer context windows. However, this research direction requires intensive computational resources.

### 5.8. Improving Spatio-temporal Understanding

An essential prerequisite for developing rational drug design and discovery is the improvement of spatial-temporal capabilities in LLMs, given that these techniques rely heavily on processing large datasets comprising complex, multi-dimensional data. At present, LLMs can process and interpret textual information reasonably well but have their weaknesses exposed when it comes to spatio-temporal data (Pan et al., 2023; Jin et al., 2023a;b), which plays an essential role in the drug discovery field. For example, this limitation leaves aside areas where a physical change concerning time and space understanding is crucial, e.g., dynamic interactions between molecules. Improving LLMs in this aspect

would provide new opportunities to understand more deeply in fields like spatial-temporal transcriptomics and molecular dynamics simulations (Nguyen et al., 2024a).

Furthermore, LLMs with enhanced spatial-temporal and multi-modal understanding enable a highly autonomous and efficient process. An illustration is in the analysis of molecular dynamics simulations. These models can automatically investigate, document, and even describe data that elaborate on molecule dynamics. This advancement and multi-modal capabilities are pivotal in unearthing potential drug candidates and uncovering molecular pathways commonly hidden beneath vast data. This exciting development has the potential to revolutionize drug discovery and significantly reduce time and resource expenditure.

### 5.9. Integrating Specialised LLMs & General LLMs

Combining a specialized language model and a general-purpose LLM gives an edge in drug discovery. Specialized LLMs perform admirably in precision tasks like understanding biological information, estimating molecule interactions, or examining protein configurations using their niche training datasets. On the other hand, general LLMs provide versatility and a broad knowledge base that can be applied to various subjects and tasks. They are considered essential instruments for many users since they are user-friendly to individuals with different levels of professionalism, such as researchers and medical workers, enabling them to access scientific knowledge and reasoning easily.

Specifically, general LLMs can act as the front-end system responsible for user interaction with conversational interfaces that provide detailed descriptions of case characteristics, assist in problem identification, and facilitate discussion of decision alternatives. They might deliver context, background knowledge, and reasoning ability to aid in understanding the problem situation. On the other hand, specialized LLMs with quantitative analytical ability can further be utilized to accomplish specific downstream tasks. For instance, they can conduct QSAR analysis for molecule compounds, protein folding simulation, or molecular structure optimization. Furthermore, these LLMs can then provide the results back to general-purpose LLMs to synthesize and interpret these results and provide insightful information for users.

## References

- Fasta, 1995. URL <http://ftp.virginia.edu/pub/fasta/>.
- Ahdritz, G., Bouatta, N., Floristean, C., Kadyan, S., Xia, Q., Gerecke, W., O'Donnell, T. J., Berenberg, D., Fisk, I., Zanichelli, N., et al. Openfold: Retraining alphafold2 yields new insights into its learning mechanisms and ca-

- pacity for generalization. *Nature Methods*, pp. 1–11, 2024.
- AI4Science, M. R. and Quantum, M. A. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv preprint arXiv:2311.07361*, 2023.
- Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.
- Altschuh, D., Lesk, A., Bloomer, A., and Klug, A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of molecular biology*, 193(4):693–707, 1987.
- Altschuh, D., Vernet, T., Berti, P., Moras, D., and Nagai, K. Coordinated amino acid changes in homologous protein families. *Protein Engineering, Design and Selection*, 2(3):193–199, 1988.
- Anfinsen, C. B. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.
- Anil, C., Wu, Y., Andreassen, A., Lewkowycz, A., Misra, V., Ramasesh, V., Slone, A., Gur-Ari, G., Dyer, E., and Neyshabur, B. Exploring length generalization in large language models. *Advances in Neural Information Processing Systems*, 35:38546–38556, 2022.
- Atlasi, Y. and Stunnenberg, H. G. The interplay of epigenetic marks during stem cell differentiation and development. *Nature Reviews Genetics*, 18(11):643–658, 2017.
- Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.
- Bagal, V., Aggarwal, R., Vinod, P., and Priyakumar, U. D. Liggpt: Molecular generation using a transformer-decoder model. 2021a.
- Bagal, V., Aggarwal, R., Vinod, P., and Priyakumar, U. D. Molgpt: molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076, 2021b.
- Beck, A., Goetsch, L., Dumontet, C., and Corvaia, N. Strategies and challenges for the next generation of antibody-drug conjugates. *Nature reviews Drug discovery*, 16(5):315–337, 2017.
- Bepler, T. and Berger, B. Learning the protein language: Evolution, structure, and function. *Cell systems*, 12(6):654–669, 2021.
- Berdigaliyev, N. and Aljofan, M. An overview of drug discovery and development. *Future medicinal chemistry*, 12(10):939–947, 2020.
- Bian, H., Chen, Y., Dong, X., Li, C., Hao, M., Chen, S., Hu, J., Sun, M., Wei, L., and Zhang, X. scmluan: a multitask generative pre-trained language model for single-cell analysis. In *International Conference on Research in Computational Molecular Biology*, pp. 479–482. Springer, 2024.
- Blaschke, T., Arús-Pous, J., Chen, H., Margreitter, C., Tyrchan, C., Engkvist, O., Papadopoulos, K., and Patronov, A. Reinvent 2.0: an ai tool for de novo drug design. *Journal of chemical information and modeling*, 60(12):5918–5922, 2020.
- Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.
- Bran, A. M., Cox, S., White, A. D., and Schwaller, P. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.
- Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.
- Brendel, V. and Busse, H. Genome structure described by formal languages. *Nucleic Acids Research*, 12(5):2561–2568, 1984.
- Brown, N., Fiscato, M., Segler, M. H., and Vaucher, A. C. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3):1096–1108, 2019.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Bussi, G. and Laio, A. Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics*, 2(4):200–212, Mar 2020. doi: 10.1038/s42254-020-0153-0.
- Bösel, L., Thürlemann, M., and Riniker, S. Machine learning in qm/mm molecular dynamics simulations of condensed-phase systems. *Journal of Chemical Theory and Computation*, 17(5):2641–2658, Apr 2021. doi: 10.1021/acs.jctc.0c01112.
- Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

- Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., and Stewart, W. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.
- Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., Rochereau, C., Ahdritz, G., Zhang, J., Church, G. M., et al. Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, 40(11):1617–1623, 2022.
- Consens, M. E., Dufault, C., Wainberg, M., Forster, D., Karimzadeh, M., Goodarzi, H., Theis, F. J., Moses, A., and Wang, B. To transformers and beyond: Large language models for the genome. *arXiv preprint arXiv:2311.07621*, 2023.
- Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. S. Diffdock: Diffusion steps, twists, and turns for molecular docking. In *The Eleventh International Conference on Learning Representations*, 2023.
- Cortés-Cros, M., Schmelzle, T., Stucke, V. M., and Hofmann, F. The path to oncology drug target validation: an industry perspective. *Target Identification and Validation in Drug Discovery: Methods and Protocols*, pp. 3–13, 2013.
- Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., and Wang, B. scgpt: towards building a foundation model for single-cell multi-omics using generative ai. *bioRxiv*, pp. 2023–04, 2023.
- Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Caranza, N. L., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., et al. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv*, pp. 2023–01, 2023.
- Das, P., Sercu, T., Wadhawan, K., Padhi, I., Gehrmann, S., Cipcigan, F., Chenthamarakshan, V., Strobel, H., dos Santos, C., Chen, P.-Y., Yang, Y. Y., Tan, J. P. K., Hedrick, J., Crain, J., and Mojsilovic, A. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nature Biomedical Engineering*, 5(6):613–623, 2021. doi: 10.1038/s41551-021-00689-x. URL <https://doi.org/10.1038/s41551-021-00689-x>.
- Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I., Courbet, A., de Haas, R. J., Bethel, N., et al. Robust deep learning-based protein sequence design using proteinmpnn. *Science*, 378 (6615):49–56, 2022.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423>.
- Dziri, N., Lu, X., Sclar, M., Li, X. L., Jian, L., Lin, B. Y., West, P., Bhagavatula, C., Bras, R. L., Hwang, J. D., et al. Faith and fate: Limits of transformers on compositionality. *arXiv preprint arXiv:2305.18654*, 2023.
- Edwards, C., Lai, T., Ros, K., Honke, G., Cho, K., and Ji, H. Translation between molecules and natural language. *EMNLP*, 2022.
- Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.
- Emmerich, C. H., Gamboa, L. M., Hofmann, M. C., Bonin-Andresen, M., Arbach, O., Schendel, P., Gerlach, B., Hempel, K., Bespalov, A., Dirnagl, U., et al. Improving target assessment in biomedical research: the got-it recommendations. *Nature reviews Drug discovery*, 20(1): 64–81, 2021.
- Enarvi, S., Amoia, M., Del-Agua Teba, M., Delaney, B., Diehl, F., Hahn, S., Harris, K., McGrath, L., Pan, Y., Pinto, J., Rubini, L., Ruiz, M., Singh, G., Stemmer, F., Sun, W., Vozila, P., Lin, T., and Ramamurthy, R. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pp. 22–30, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpmc-1.4. URL <https://aclanthology.org/2020.nlpmc-1.4>.
- Erikawa, D., Yasuo, N., and Sekijima, M. Mermaid: an open source automated hit-to-lead method based on deep reinforcement learning. *Journal of Cheminformatics*, 13: 1–10, 2021.
- Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pp. 2021–10, 2021.

- Fang, Y., Jiang, Y., Wei, L., Ma, Q., Ren, Z., Yuan, Q., and Wei, D.-Q. Deepprosite: Structure-aware protein binding site prediction using esmfold and pretrained language model. *Bioinformatics*, pp. btad718, 2023.
- Ferruz, N. and Höcker, B. Controllable protein design with language models. *Nature Machine Intelligence*, 4(6): 521–532, 2022.
- Ferruz, N., Schmidt, S., and Höcker, B. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.
- Flam-Shepherd, D., Zhu, K., and Aspuru-Guzik, A. Language models can learn complex molecular distributions. *Nature Communications*, 13(1):3293, 2022.
- Floris, M., Olla, S., Schlessinger, D., and Cucca, F. Genetic-driven druggable target identification and validation. *Trends in Genetics*, 34(7):558–570, 2018.
- Frieder, S., Pinchetti, L., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., Chevalier, A., and Berner, J. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*, 2023.
- Fu, T., Huang, K., Xiao, C., Glass, L. M., and Sun, J. Hint: Hierarchical interaction network for clinical-trial-outcome predictions. *Patterns*, 3(4), 2022.
- Fung, A., Koehl, A., Jagota, M., and Song, Y. S. The impact of protein dynamics on residue-residue coevolution and contact prediction. *bioRxiv*, pp. 2022–10, 2022.
- Gao, J., Xiao, C., Glass, L. M., and Sun, J. Compose: Cross-modal pseudo-siamese network for patient trial matching. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 803–812, 2020a.
- Gao, J., Xiao, C., Wang, Y., Tang, W., Glass, L. M., and Sun, J. Stagenet: Stage-aware neural networks for health risk prediction. In *Proceedings of The Web Conference 2020*, pp. 530–540, 2020b.
- Göbel, U., Sander, C., Schneider, R., and Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 18(4):309–317, 1994.
- Golkar, S., Pettee, M., Eickenberg, M., Bietti, A., Cranmer, M., Krawezik, G., Lanusse, F., McCabe, M., Ohana, R., Parker, L., et al. xval: A continuous number encoding for large language models. *arXiv preprint arXiv:2310.02989*, 2023.
- Grosse, R., Bae, J., Anil, C., Elhage, N., Tamkin, A., Tajdini, A., Steiner, B., Li, D., Durmus, E., Perez, E., et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023.
- Gruver, N., Finzi, M., Qiu, S., and Wilson, A. G. Large language models are zero-shot time series forecasters. *arXiv preprint arXiv:2310.07820*, 2023.
- Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.
- Guimaraes, G. L., Sanchez-Lengeling, B., Outeiral, C., Farias, P. L. C., and Aspuru-Guzik, A. Objective-reinforced generative adversarial networks (organ) for sequence generation models. *arXiv preprint arXiv:1705.10843*, 2017.
- Guo, H., Huo, M., Zhang, R., and Xie, P. Proteinchat: Towards achieving chatgpt-like functionalities on protein 3d structures. 2023.
- Hamer, D. M. d., Schoor, P., Polak, T. B., and Kapitan, D. Improving patient pre-screening for clinical trials: Assisting physicians with large language models. *arXiv preprint arXiv:2304.07396*, 2023.
- Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., Wang, T., Ma, J., Zhang, X., and Song, L. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, pp. 1–11, 2024.
- Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, pp. 2024–07, 2024.
- He, J., Mattsson, F., Forsberg, M., Bjerrum, E. J., Engkvist, O., Tyrchan, C., Czechtizky, W., et al. Transformer neural network for structure constrained molecular optimization. *Theoretical and Computational Chemistry*, 2021a.
- He, J., You, H., Sandström, E., Nittinger, E., Bjerrum, E. J., Tyrchan, C., Czechtizky, W., and Engkvist, O. Molecular optimization by capturing chemist's intuition using deep neural networks. *Journal of cheminformatics*, 13(1):1–17, 2021b.
- Hesslow, D., Zanichelli, N., Notin, P., Poli, I., and Marks, D. Rita: a study on scaling up generative protein sequence models. *arXiv preprint arXiv:2205.05789*, 2022.
- Hie, B. L., Shanker, V. R., Xu, D., Bruun, T. U., Weidenbacher, P. A., Tang, S., Wu, W., Pak, J. E., and Kim, P. S. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, 2023.
- Hou, W. and Ji, Z. Reference-free and cost-effective automated cell type annotation with gpt-4 in single-cell rna-seq analysis. *Research Square*, 2023.

- Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. In *International Conference on Machine Learning*, pp. 8946–8970. PMLR, 2022.
- Hu, M., Yuan, F., Yang, K., Ju, F., Su, J., Wang, H., Yang, F., and Ding, Q. Exploring evolution-aware &-free protein language models as protein function predictors. *Advances in Neural Information Processing Systems*, 35:38873–38884, 2022.
- Huang, C.-W., Tsai, S.-C., and Chen, Y.-N. PLM-ICD: Automatic ICD coding with pretrained language models. In Naumann, T., Bethard, S., Roberts, K., and Rumshisky, A. (eds.), *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pp. 10–20, Seattle, WA, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.clinicalnlp-1.2. URL <https://aclanthology.org/2022.clinicalnlp-1.2>.
- Huang, K., Altosaar, J., and Ranganath, R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *Workshops at the Conference on Health, Inference, and Learning*, 2020.
- Inagaki, T., Kato, A., Takahashi, K., Ozaki, H., and Kanda, G. N. Llms can generate robotic scripts from goal-oriented instructions in biological laboratory automation. *arXiv preprint arXiv:2304.10267*, 2023.
- Irwin, R., Dimitriadis, S., He, J., and Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, 2022.
- Ivanenkov, Y. A., Polykovskiy, D., Bezrukov, D., Zagribelnyy, B., Aladinskiy, V., Kamya, P., Aliper, A., Ren, F., and Zhavoronkov, A. Chemistry42: an ai-driven platform for molecular design and optimization. *Journal of Chemical Information and Modeling*, 63(3):695–701, 2023.
- Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A., and Smit, B. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169, 2024.
- Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Stüber, A. T., Topalis, J., Weber, T., Wesp, P., Sabel, B. O., Ricke, J., et al. Chatgpt makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *European radiology*, pp. 1–9, 2023.
- Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.
- Jiang, L. Y., Liu, X. C., Nejatian, N. P., Nasir-Moin, M., Wang, D., Abidin, A., Eaton, K., Riina, H. A., Laufer, I., Punjabi, P., et al. Health system-scale language models are all-purpose prediction engines. *Nature*, pp. 1–6, 2023.
- Jiang, M., Wang, S., Zhang, S., Zhou, W., Zhang, Y., and Li, Z. Sequence-based drug-target affinity prediction using weighted graph neural networks. *BMC Genomics*, 23, 2022.
- Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., Shi, X., Chen, P.-Y., Liang, Y., Li, Y.-F., Pan, S., et al. Time-lilm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*, 2023a.
- Jin, M., Wen, Q., Liang, Y., Zhang, C., Xue, S., Wang, X., Zhang, J., Wang, Y., Chen, H., Li, X., et al. Large models for time series and spatio-temporal data: A survey and outlook. *arXiv preprint arXiv:2310.10196*, 2023b.
- Jin, Q., Wang, Z., Floudas, C. S., Sun, J., and Lu, Z. Matching patients to clinical trials with large language models. *ArXiv*, 2023c.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.
- Ketata, M. A., Laue, C., Mammadov, R., Stark, H., Wu, M., Corso, G., Marquet, C., Barzilay, R., and Jaakkola, T. S. Diffdock-PP: Rigid protein-protein docking with diffusion models. In *ICLR 2023 - Machine Learning for Drug Discovery workshop*, 2023.
- Kirchoff, K. E. and Gomez, S. M. Ember: multi-label prediction of kinase-substrate phosphorylation events through deep learning. *Bioinformatics*, 38(8):2119–2126, 2022.
- Koh, H. Y., Ju, J., Liu, M., and Pan, S. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM computing surveys*, 55(8):1–35, 2022a.
- Koh, H. Y., Ju, J., Zhang, H., Liu, M., and Pan, S. How far are we from robust long abstractive summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2682–2698, 2022b.
- Koh, H. Y., Nguyen, A. T. N., Pan, S., May, L. T., and Webb, G. I. Physicochemical graph neural network for

- learning protein–ligand interaction fingerprints from sequence data. *Nature Machine Intelligence*, 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00847-1.
- Krishna, R., Wang, J., Ahern, W., Sturmels, P., Venkatesh, P., Kalvet, I., Lee, G. R., Morey-Burrows, F. S., Anishchenko, I., Humphreys, I. R., et al. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384(6693):eabd12528, 2024.
- Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., et al. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198, 2023.
- Le, N. Q. K., Ho, Q.-T., Nguyen, V.-N., and Chang, J.-S. Bert-promoter: An improved sequence-based predictor of dna promoter using bert pre-trained model and shap feature selection. *Computational Biology and Chemistry*, 99:107732, 2022.
- Le, Q. and Mikolov, T. Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196. PMLR, 2014.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
- Levitt, M. Conformational preferences of amino acids in globular proteins. *Biochemistry*, 17(20):4277–4285, 1978.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL <https://aclanthology.org/2020.acl-main.703>.
- Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *CVPR*, 2023a.
- Li, Z., Gao, E., Zhou, J., Han, W., Xu, X., and Gao, X. Applications of deep learning in understanding gene regulation. *Cell Reports Methods*, 2023b.
- Lin, A., Giuliano, C. J., Sayles, N. M., and Sheltzer, J. M. Crispr/cas9 mutagenesis invalidates a putative cancer dependency targeted in on-going clinical trials. *Elife*, 6: e24179, 2017.
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.
- Lindsay, M. A. Target discovery. *Nature Reviews Drug Discovery*, 2(10):831–838, 2003.
- Littmann, M., Heinzinger, M., Dallago, C., Weissenow, K., and Rost, B. Protein embeddings and deep learning predict binding residues for various ligand classes. *Scientific Reports*, 11(1):23916, 2021.
- Liu, G., Hsu, T.-M. H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., and Ghassemi, M. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pp. 249–269. PMLR, 2019a.
- Liu, P., Ren, Y., and Ren, Z. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *arXiv preprint arXiv:2308.06911*, 2023a.
- Liu, S., Nie, W., Wang, C., Lu, J., Qiao, Z., Liu, L., Tang, J., Xiao, C., and Anandkumar, A. Multi-modal molecule structure-text model for text-based retrieval and editing. *Nature Machine Intelligence*, 2023b.
- Liu, S., Wang, J., Yang, Y., Wang, C., Liu, L., Guo, H., and Xiao, C. Chatgpt-powered conversational drug editing using retrieval and domain feedback. *arXiv preprint arXiv:2305.18090*, 2023c.
- Liu, S., Zhu, Y., Lu, J., Xu, Z., Nie, W., Gitter, A., Xiao, C., Tang, J., Guo, H., and Anandkumar, A. A text-guided protein design framework. *arXiv preprint arXiv:2302.04611*, 2023d.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019b.
- Loeffler, H., He, J., Tibo, A., Janet, J. P., Voronov, A., Mervin, L., and Engkvist, O. Reinvent4: Modern ai-driven generative molecule design. 2023.
- Loughrey, D., Watters, K. E., Settle, A. H., and Lucks, J. B. Shape-seq 2.0: Systematic optimization and extension of high-throughput chemical probing of rna secondary structure with next generation sequencing. *Nucleic Acids Research*, 42(21), Oct 2014. doi: 10.1093/nar/gku909.

- Lu, W., Wu, Q., Zhang, J., Rao, J., Li, C., and Zheng, S. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *Advances in Neural Information Processing Systems*, 35:7236–7249, 2022.
- Luo, F., Wang, M., Liu, Y., Zhao, X.-M., and Li, A. Deepphos: prediction of protein phosphorylation sites with deep learning. *Bioinformatics*, 35(16):2766–2773, 2019.
- Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., and Liu, T.-Y. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 2022.
- Ma, F., Chitta, R., Zhou, J., You, Q., Sun, T., and Gao, J. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1903–1911, 2017.
- Ma, Q., Jiang, Y., Cheng, H., and Xu, D. Harnessing the deep learning power of foundation models in single-cell omics. *Nature Reviews Molecular Cell Biology*, pp. 1–2, 2024.
- Madani, A., McCann, B., Naik, N., Keskar, N. S., Anand, N., Eguchi, R. R., Huang, P.-S., and Socher, R. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.
- Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos Jr, J. L., Xiong, C., Sun, Z. Z., Socher, R., et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pp. 1–8, 2023.
- Marshall, I. J., Kuiper, J., Banner, E., and Wallace, B. C. Automating biomedical evidence synthesis: Roboterreviewer. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2017, pp. 7. NIH Public Access, 2017.
- Matic, M., Singh, G., Carli, F., De Oliveira Rosa, N., Miglionico, P., Magni, L., Gutkind, J. S., Russell, R. B., Inoue, A., and Raimondi, F. Precogx: exploring gpcr signaling mechanisms with deep protein representations. *Nucleic acids research*, 50(W1):W598–W610, 2022.
- McPartlon, M. and Xu, J. Deep learning for flexible and site-specific protein docking and design. *bioRxiv*, pp. 2023–04, 2023.
- Mehr, S. H. M., Craven, M., Leonov, A. I., Keenan, G., and Cronin, L. A universal system for digitization and automatic execution of the chemical synthesis literature. *Science*, 370(6512):101–108, 2020.
- Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.
- Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M. J., Maranian, M. J., Bolla, M. K., Wang, Q., Shah, M., et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nature genetics*, 47(4):373–380, 2015.
- Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., et al. Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551(7678):92–94, 2017.
- Miranda Furtado, C. L., Dos Santos Luciano, M. C., Silva Santos, R. D., Furtado, G. P., Moraes, M. O., and Pessoa, C. Epidrugs: targeting epigenetic marks in cancer treatment. *Epigenetics*, 14(12):1164–1176, 2019.
- Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.
- Moon, J. H., Lee, H., Shin, W., Kim, Y.-H., and Choi, E. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6070–6080, 2022.
- Moret, M., Pachon Angona, I., Cotos, L., Yan, S., Atz, K., Brunner, C., Baumgartner, M., Grisoni, F., and Schneider, G. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nature Communications*, 14(1):114, 2023.
- Narganes-Carlón, D., Crowther, D. J., and Pearson, E. R. A publication-wide association study (pwas), historical language models to prioritise novel therapeutic drug targets. *Scientific Reports*, 13(1):8366, 2023.
- Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.
- Nelson, C. P., Goel, A., Butterworth, A. S., Kanoni, S., Webb, T. R., Marouli, E., Zeng, L., Ntalla, I., Lai, F. Y., Hopewell, J. C., et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nature genetics*, 49(9):1385–1391, 2017.

- Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P. C., Li, M. J., Wang, J., et al. The support of human genetic evidence for approved drug indications. *Nature genetics*, 47(8):856–860, 2015.
- Nguyen, A. T., Nguyen, D. T., Koh, H. Y., Toskov, J., MacLean, W., Xu, A., Zhang, D., Webb, G. I., May, L. T., and Halls, M. L. The application of artificial intelligence to accelerate g protein-coupled receptor drug discovery. *British Journal of Pharmacology*, 181(14):2371–2384, 2024a.
- Nguyen, E., Poli, M., Faizi, M., Thomas, A., Wornow, M., Birch-Sykes, C., Massaroli, S., Patel, A., Rabideau, C., Bengio, Y., et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36, 2024b.
- Nichols, P., Li, L., Kumar, S., Buck, P. M., Singh, S. K., Goswami, S., Balthazar, B., Conley, T. R., Sek, D., and Allen, M. J. Rational design of viscosity reducing mutants of a monoclonal antibody: hydrophobic versus electrostatic inter-molecular interactions. In *MAbs*, volume 7, pp. 212–230. Taylor & Francis, 2015.
- Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N., and Madani, A. Progen2: exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978, 2023.
- Olivecrona, M., Blaschke, T., Engkvist, O., and Chen, H. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9(1):1–14, 2017.
- OpenAI. Gpt-4: A technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., and Wu, X. Unifying large language models and knowledge graphs: A roadmap. *arXiv preprint arXiv:2306.08302*, 2023.
- Park, G., Yoon, B.-J., Luo, X., Lpez-Marrero, V., Johnstone, P., Yoo, S., and Alexander, F. Automated extraction of molecular interactions and pathway knowledge using large language model, galactica: Opportunities and challenges. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pp. 255–264, 2023.
- Patel, S. B. and Lam, K. Chatgpt: the future of discharge summaries? *The Lancet Digital Health*, 5(3):e107–e108, 2023.
- Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Capelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.
- Plenge, R. M., Scolnick, E. M., and Altshuler, D. Validating therapeutic targets through human genetics. *Nature reviews Drug discovery*, 12(8):581–594, 2013.
- Pradeep, R., Li, Y., Wang, Y., and Lin, J. Neural query synthesis and domain-specific ranking templates for multi-stage clinical trial matching. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2325–2330, 2022.
- Pun, F. W., Ozerov, I. V., and Zhavoronkov, A. Ai-powered therapeutic target discovery. *Trends in Pharmacological Sciences*, 2023.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- Ram, S. and Bepler, T. Few shot protein generation. *arXiv preprint arXiv:2204.01168*, 2022.
- Ramos, J. et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pp. 29–48. Citeseer, 2003.
- Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.
- Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., and Rives, A. Transformer protein language models are unsupervised structure learners. *Biorxiv*, pp. 2020–12, 2020.
- Rao, R. M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., and Rives, A. Msa transformer. In *International Conference on Machine Learning*, pp. 8844–8856. PMLR, 2021.
- Raybould, M. I., Marks, C., Krawczyk, K., Taddese, B., Nowak, J., Lewis, A. P., Bujotzek, A., Shi, J., and Deane, C. M. Five computational developability guidelines for

- therapeutic antibody profiling. *Proceedings of the National Academy of Sciences*, 116(10):4025–4030, 2019.
- Razeghi, Y., Logan IV, R. L., Gardner, M., and Singh, S. Impact of pretraining term frequencies on few-shot numerical reasoning. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 840–854, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.59. URL <https://aclanthology.org/2022.findings-emnlp.59>.
- Ren, F., Ding, X., Zheng, M., Korzinkin, M., Cai, X., Zhu, W., Mantsyzov, A., Aliper, A., Aladinskiy, V., Cao, Z., et al. AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel cdk20 small molecule inhibitor. *Chemical Science*, 14(6):1443–1452, 2023.
- Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.
- Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., and Das, P. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.
- Rothchild, D., Tamkin, A., Yu, J., Misra, U., and Gonzalez, J. C5t5: Controllable generation of organic molecules with transformers. *arXiv preprint arXiv:2108.10307*, 2021.
- Roziere, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., Adi, Y., Liu, J., Remez, T., Rapin, J., et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.
- Savage, N. Drug discovery companies are customizing chatgpt: here's how. *Nature Biotechnology*, 2023.
- Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.
- Schwaller, P., Petraglia, R., Zullo, V., Nair, V. H., Haeuselmann, R. A., Pisoni, R., Bekas, C., Iuliano, A., and Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical science*, 11(12):3316–3325, 2020.
- Searls, D. B. The language of genes. *Nature*, 420(6912): 211–217, 2002.
- Shao, X., Yang, H., Zhuang, X., Liao, J., Yang, P., Cheng, J., Lu, X., Chen, H., and Fan, X. scdeepsort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network. *Nucleic acids research*, 49(21):e122–e122, 2021.
- Shi, H., Xie, P., Hu, Z., Zhang, M., and Xing, E. P. Towards automated icd coding using deep learning. *arXiv preprint arXiv:1711.04075*, 2017.
- Shing, H.-C., Shivade, C., Pourdamghani, N., Nan, F., Resnik, P., Oard, D., and Bhatia, P. Towards clinical encounter summarization: Learning to compose discharge summaries from prior notes. *arXiv preprint arXiv:2104.13498*, 2021.
- Singh, R., Sledzieski, S., Bryson, B., Cowen, L., and Berger, B. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proceedings of the National Academy of Sciences*, 120(24): e2220778120, 2023.
- Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- Skalic, M., Sabbadin, D., Sattarov, B., Sciabola, S., and De Fabritiis, G. From target to drug: generative modeling for the multimodal structure-based ligand design. *Molecular pharmaceutics*, 16(10):4282–4291, 2019.
- Srinivasa, R. S., Qian, C., Theodorou, B., Spaeder, J., Xiao, C., Glass, L., and Sun, J. Clinical trial site matching with improved diversity using fair policy learning. *arXiv preprint arXiv:2204.06501*, 2022.
- Su, B., Du, D., Yang, Z., Zhou, Y., Li, J., Rao, A., Sun, H., Lu, Z., and Wen, J.-R. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*, 2022.
- Sundararaman, D., Si, S., Subramanian, V., Wang, G., Hazarika, D., and Carin, L. Methods for numeracy-preserving word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4742–4753, 2020.
- Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.
- Testolin, A. Can neural networks do arithmetic? a survey on the elementary numerical skills of state-of-the-art deep learning models. *arXiv preprint arXiv:2303.07735*, 2023.

- Thawani, A., Pujara, J., Szekely, P. A., and Ilievski, F. Representing numbers in nlp: a survey and a vision. *arXiv preprint arXiv:2103.13136*, 2021.
- Theodoris, C. V., Li, M., White, M. P., Liu, L., He, D., Pollard, K. S., Bruneau, B. G., and Srivastava, D. Human disease modeling reveals integrated transcriptional and epigenetic mechanisms of notch1 haploinsufficiency. *Cell*, 160(6):1072–1086, 2015.
- Theodoris, C. V., Zhou, P., Liu, L., Zhang, Y., Nishino, T., Huang, Y., Kostina, A., Ranade, S. S., Gifford, C. A., Uspenkiy, V., et al. Network-based screen in ipsc-derived cells reveals therapeutic candidate for heart valve disease. *Science*, 371(6530):eabd0724, 2021.
- Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., et al. Transfer learning enables predictions in network biology. *Nature*, pp. 1–9, 2023.
- Theodorou, B., Glass, L., Xiao, C., and Sun, J. Framm: Fair ranking with missing modalities for clinical trial site selection. *arXiv preprint arXiv:2305.19407*, 2023.
- Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.
- Toufiq, M., Rinchai, D., Bettacchioli, E., Kabeer, B. S. A., Khan, T., Subba, B., White, O., Yurieva, M., George, J., Jourde-Chiche, N., et al. Harnessing large language models (llms) for candidate gene prioritization and selection. *Journal of Translational Medicine*, 21(1):728, 2023.
- Trotman, A., Puurula, A., and Burgess, B. Improvements to bm25 and language models examined. In *Proceedings of the 19th Australasian Document Computing Symposium*, pp. 58–65, 2014.
- Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):A1oa2300138, 2024.
- Urbina, F., Lentzos, F., Invernizzi, C., and Ekins, S. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191, 2022.
- Varadi, M. and Velankar, S. The impact of alphafold protein structure database on the fields of life sciences. *Proteomics*, 23(17):2200128, 2023.
- Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Verkuil, R., Kabeli, O., Du, Y., Wicky, B. I., Milles, L. F., Dauparas, J., Baker, D., Ovchinnikov, S., Sercu, T., and Rives, A. Language models generalize beyond natural proteins. *bioRxiv*, pp. 2022–12, 2022.
- Vincent, F., Loria, P., Pregel, M., Stanton, R., Kitching, L., Nocka, K., Doyonnas, R., Steppan, C., Gilbert, A., Schroeter, T., et al. Developing predictive assays: the phenotypic screening “rule of 3”. *Science translational medicine*, 7(293):293ps15–293ps15, 2015.
- Wang, D., Zeng, S., Xu, C., Qiu, W., Liang, Y., Joshi, T., and Xu, D. Musitedeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics*, 33(24):3909–3916, 2017.
- Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J. L., Castro, K. M., Ragotte, R., Saragovi, A., Milles, L. F., Baek, M., et al. Scaffolding protein functional sites using deep learning. *Science*, 377(6604):387–394, 2022a.
- Wang, P., Zheng, S., Jiang, Y., Li, C., Liu, J., Wen, C., Patronov, A., Qian, D., Chen, H., and Yang, Y. Structure-aware multimodal deep learning for drug–protein interaction prediction. *Journal of Chemical Information and Modeling*, 62(5):1308–1317, 2022b.
- Wang, S., You, R., Liu, Y., Xiong, Y., and Zhu, S. Netgo 3.0: Protein language model improves large-scale functional annotations. *Genomics, Proteomics & Bioinformatics*, 2023a.
- Wang, X., Peng, Y., Lu, L., Lu, Z., and Summers, R. M. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9049–9058, 2018.
- Wang, Z. and Sun, J. Trial2vec: Zero-shot clinical trial document similarity search using self-supervision. *EMNLP Findings*, 2022.
- Wang, Z., Gao, C., Xiao, C., and Sun, J. Meditab: Scaling medical tabular data predictors via data consolidation, enrichment, and refinement. *arXiv:2305.12081*, 2023b.
- Wang, Z., Xiao, C., and Sun, J. Spot: Sequential predictive modeling of clinical trial outcome with meta-learning. *arXiv preprint arXiv:2304.05352*, 2023c.

- Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976): 1089–1100, 2023.
- Webster, P. Six ways large language models are changing healthcare. *Nature Medicine*, pp. 1–3, 2023.
- Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.
- White, R., Peng, T., Sripitak, P., Rosenberg Johansen, A., and Snyder, M. Clinidigest: a case study in large language model based large-scale summarization of clinical trial descriptions. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, pp. 396–402, 2023.
- Xie, P. and Xing, E. A neural architecture for automated icd coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1066–1076, 2018.
- Xu, Y., Biswal, S., Deshpande, S. R., Maher, K. O., and Sun, J. Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In *Proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*, pp. 2565–2573, 2018.
- Yan, A., McAuley, J., Lu, X., Du, J., Chang, E. Y., Gentili, A., and Hsu, C.-N. Radbert: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, 2022.
- Yanofsky, C., Horn, V., and Thorpe, D. Protein structure relationships revealed by mutational analysis. *Science*, 146(3651):1593–1594, 1964.
- Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., and Chen, E. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.
- Yoshikawa, N., Skreta, M., Darvish, K., Arellano-Rubach, S., Ji, Z., Bjørn Kristensen, L., Li, A. Z., Zhao, Y., Xu, H., Kuramshin, A., et al. Large language models for chemistry robotics. *Autonomous Robots*, pp. 1–30, 2023.
- Yuan, Q., Chen, S., Wang, Y., Zhao, H., and Yang, Y. Alignment-free metal ion-binding site prediction from protein sequence through pretrained language model and multi-task learning. *Briefings in Bioinformatics*, 23(6): bbac444, 2022.
- Yuan, Q., Tian, C., and Yang, Y. Genome-scale annotation of protein binding sites via language model and geometric deep learning. *bioRxiv*, pp. 2023–11, 2023.
- Zengini, E., Hatzikotoulas, K., Tachmazidou, I., Steinberg, J., Hartwig, F. P., Southam, L., Hackinger, S., Boer, C. G., Styrkarsdottir, U., Gilly, A., et al. Genome-wide analyses using uk biobank data provide insights into the genetic architecture of osteoarthritis. *Nature genetics*, 50(4):549–558, 2018.
- Zhang, S. and Xie, L. Protein language model-powered 3d ligand binding site prediction from protein sequence. *arXiv preprint arXiv:2312.03016*, 2023.
- Zhang, X., Xiao, C., Glass, L. M., and Sun, J. Deepenroll: patient-trial matching with deep embedding and entailment prediction. In *Proceedings of the web conference 2020*, pp. 1029–1037, 2020a.
- Zhang, Z., Liu, J., and Razavian, N. BERT-XML: Large scale automated ICD coding using BERT pretraining. In Rumshisky, A., Roberts, K., Bethard, S., and Naumann, T. (eds.), *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, November 2020b.
- Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., Asadulaev, A., et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature biotechnology*, 37(9):1038–1040, 2019.
- Zheng, Y., Koh, H. Y., Ju, J., Nguyen, A. T., May, L. T., Webb, G. I., and Pan, S. Large language models for scientific synthesis, inference and explanation. *arXiv preprint arXiv:2310.07984*, 2023.
- Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., and Liu, H. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv preprint arXiv:2306.15006*, 2023.
- Zubradt, M., Gupta, P., Persad, S., Lambowitz, A. M., Weissman, J. S., and Rouskin, S. Dms-mapseq for genome-wide or targeted rna structure probing in vivo. *Nature Methods*, 14(1):75–82, Nov 2016. doi: 10.1038/nmeth.4057.
- Zvyagin, M., Brace, A., Hippe, K., Deng, Y., Zhang, B., Bohorquez, C. O., Clyde, A., Kale, B., Perez-Rivera, D., Ma, H., et al. Genslms: Genome-scale language models reveal sars-cov-2 evolutionary dynamics. *The International Journal of High Performance Computing Applications*, 37(6):683–705, 2023.