Experimental & Molecular Medicine https://doi.org/10.1038/s12276-025-01583-1

Article in Press

A survey on large language models in biology and chemistry

Received: 7 July 2025

Accepted: 27 August 2025

Published online: 15 November 2025

Cite this article as: Islambek Ashyrmamatov, Su Ji Gwak, Su-Young Jin *et al.* A survey on large language models in biology and chemistry *Exp Mol Med.* (2025). https://doi.org/10.1038/s12276-025-01583-1

Islambek Ashyrmamatov, Su Ji Gwak, Su-Young Jin, Ikhyeong Jun, Umit V. Ucak, Jay-Yoon Lee & Juyong Lee

A Survey on Large Language Models in Biology and Chemistry

Islambek Ashyrmamatov[1†], Su Ji Gwak[2†], Su-Young Jin[2], Ikhyeong Jun[3], Umit V. Ucak[1*], Jay-Yoon Lee[2*] and Juyong Lee[1,3*]

[1] Research Institute of Pharmaceutical Science, College of Pharmacy, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

[2] Graduate School of Data Science, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

[3] Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

† These authors have contributed equally to this work.

** Corresponding authors: braket@snu.ac.kr, lee.jayyoon@snu.ac.kr, nicole23@snu.ac.kr*

Abstract

Artificial intelligence (AI) is reshaping biomedical research by providing scalable computational frameworks suited to the complexity of biological systems. Central to this revolution are bio/chemical language models (LMs), including Large Language Models (LLMs), which are re-conceptualizing molecular structures as a form of "language" amenable to advanced computational techniques. This review critically examines the role of these models in biology and chemistry, tracing their evolution from molecular representation to molecular generation and optimization. This review covers key molecular representation strategies for both biological macromolecules and small organic compounds—ranging from protein and nucleotide sequences to single-cell data, string-based chemical formats, graph-based encodings, and 3D point clouds—highlighting their respective advantages and inherent limitations in AI applications. The discussion further explores core model architectures, such as BERT-like encoders, GPT-like decoders, and encoder-decoder transformers, alongside their sophisticated pre-training strategies like self-supervised learning, multi-task learning, and retrieval-augmented generation. Key biomedical applications, spanning protein structure and function prediction, de novo protein design, genomic analysis, molecular property prediction, de novo molecular design, reaction prediction, and retrosynthesis, are explored through representative studies and emerging trends. Finally, the review considers the emerging landscape of agentic and interactive AI systems, showcasing briefly their potential to automate and accelerate scientific discovery while addressing critical technical, ethical, and regulatory considerations that will shape the future trajectory of AI in biomedicine.

1. **Introduction**

Large language models (LLMs), built on deep neural architectures and trained on massive text corpora, have achieved state-of-the-art performance in language understanding, generation, and reasoning. Although originally developed for natural language, their core modeling principles are broadly transferable to symbolic scientific data. This has spurred growing interest in adapting LLMs to scientific domains, particularly in chemistry and biology.[1,2]

Scientific knowledge and un

derstanding critically depend on the construction of formal representations that encode the structure and behavior of physical and biological systems. These representations are designed for fidelity in capturing domain-specific properties, but rarely align with the distributional and syntactic patterns of language models. Thus, various attempts have been suggested for better alignment between LLMs and scientific representations**.**[3,4]

What enables LLMs to perform so effectively is not an understanding of individual tokens, but their ability to model the statistical structure that governs token composition. In scientific domains, a model's ability to infer properties depends on how well the input representation encodes underlying structure. Thus, representational design is not peripheral but fundamental for developing scientific LLMs. It determines what models can learn, generalize, and ultimately, discover. In addition, it is well-known that the scales of model architecture and training data are critical in accuracy and emergent behaviors of LLMs.[5] Thus, the success of scientific LLMs rests on both scale and architecture of the models, and how effectively the representation translates a domain structure into a learnable entity.

Recent progress in using LLMs in biology and chemistry has been accelerated by the growth of curated, domain-specific datasets. Molecular and protein databases, along with scientific literature, now support diverse training strategies, from self-supervised objectives to multimodal integration. However, much of this development remains fragmented, and systematic comparisons across chemical and biological domains are still limited.

In this review, we examine how LLMs are being adapted to the unique demands of chemical and biological topics. We focus on how representations, architectures, and training regimes influence model performance across domains and tasks. The foundational challenge lies in converting complex, multi-dimensional molecular information into formats that language models can process (**Fig. 1**). Our goal is to clarify what has been achieved, what remains challenging, and how these models will better serve scientific understanding.

2. **Biological language models**

The unprecedented success of large language models (LLMs) has opened a new paradigm in data analysis. In the field of biology, the utilization of various biological data such as protein sequences,[6] structures,[7] nucleotides,[8] and species taxonomy[9] has been considered. The application of Transformer architectures to biological problems has led to significant breakthroughs, with AlphaFold2 (AF2)[10] and RoseTTAFold (RF)[11] emerging as landmark models in protein structure prediction. In parallel, ongoing research is being conducted to describe biological complexity more accurately within the models (see **Table 1**).

2.1. **Protein language models**

The sequential nature of protein has enabled the application of language modeling techniques from natural language processing. Early models such as ProtBERT,[12] MSA Transformer,[13] and ProtTrans[14] leveraged core techniques from the deep language models while exploring variations in both input formats, e.g., single sequences, multiple sequence alignments (MSAs), and architectures, e.g., unidirectional and BERT-style bidirectional encoders. ESMFold[2] achieves AlphaFold2-level accuracy in protein structure prediction without relying on MSAs, capturing contextual dependencies solely through language modeling. The scaling of model parameters and faster structure prediction highlight the potential of language models when trained on large-scale biological data. ProtMamba[15] also showed that protein language modeling is feasible without MSAs. The model adopts a Mamba[16] based state space architecture instead of attention-based to handle long-range sequences.

Protein design aims to generate proteins with completely new functions and structures, and generative models can play a key role in the process. ProGen[17] enables controlled protein sequence generation by incorporating conditioning tags into an autoregressive transformer architecture. ProGen2[18] and ProtGPT2[19] further improve upon previous models by leveraging more complex conditioning tags to generate sequences that satisfy both structural and functional constraints. Recently, diffusion architectures, developed for image generation from text prompts, have been adapted for protein structure generation. RFdiffusion[20] incorporates spatial constraints through SE(3) equivariance, enabling more efficient and physically consistent sampling of protein structures. Such structural modeling has facilitated scaffolding tasks, and tools including ProteinMPNN[21] and Foldseek[22] have accelerated advances in protein design.

2.2. **Protein structure models**

Protein structure models predict the tertiary structures of proteins from their primary amino acid sequences. Traditionally, techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM) have been employed to elucidate protein structures. However, these experimental methods are often constrained by high costs, time requirements, and technical limitations, resulting in a considerably slower accumulation of structural data compared to the rapidly expanding number of known protein

sequences.[23] This sequence–structure data imbalance (e.g., between UniProtKB[24] and the PDB[7]) underscores the need for computational prediction approaches to complement experimental efforts.

AlphaFold (AF)[25] and AlphaFold2 (AF2)[10] have demonstrated outstanding performance in the field of protein structure prediction, as evidenced by their success in Critical Assessment of protein Structure Prediction13 (CASP13) and CASP14, respectively. AF2 consists of two primary modules: the Evoformer and the structure module. Unlike AF, which employs a ResNet-based convolutional neural network (CNN), AF2 introduces an attention-based Evoformer, enabling efficient processing of MSAs and pairwise residue interactions. The Evoformer can be interpreted as a biology-specific transformer, where MSAs are treated as sequences in natural language, capturing evolutionary patterns across homologous proteins. This approach has been more fully realized in protein language models (pLMs), which are designed to replace MSAs by implicitly modeling evolutionary information. The structure module allows for end-to-end learning from primary sequence to 3D structural reconstruction, achieving near experimental accuracy.

Several platforms have been developed to extend the applicability and accessibility of protein structure models. ColabFold[26] leverages a metagenomic sequence database (ColabFoldDB) to enhance the diversity and quality of MSAs, and it is implemented to run on web-based GPU resources through Google Colaboratory. This approach improves accessibility to high-accuracy protein structure prediction while effectively reducing computational resource burdens. Phyre2.2[27] is an upgraded platform for protein structure and function prediction that maintains a user-friendly interface while integrating AlphaFold-predicted structures as new templates. It enables large-scale structural analysis by utilizing a broader range of structural templates beyond those available in the PDB. Furthermore, it supports domain-level optimization and batch-mode prediction, thereby serving as a computational alternative that complements experimental studies.

2.3. **Nucleotide language models**

Unlike natural language, DNA does not possess an inherent concept of "words," and its composition is limited to just four nucleotides—adenine (A), thymine (T), guanine (G), and cytosine (C)—as opposed to protein sequences, which are composed of approximately 20 amino acids. This limited alphabet reduces the overall information density, making the development of effective DNA language models more challenging.

Earlier approaches, such as DeepSite,[28] utilized CNNs and recurrent neural networks (RNNs) for modeling DNA sequences. However, CNNs often struggle with capturing long-range dependencies, and RNNs suffer from computational inefficiency and scalability issues. To address these limitations, DNABERT[29] adopted a masked language modeling (MLM) based on bidirectional encoder representations from transformers (BERT) using k-mer tokenization (a.k.a. n-gram in computer science), enabling more effective sequence representation. Subsequent models, including GROVER[30] and DNABERT2,[31] leveraged Byte Pair Encoding (BPE)[32]—tokenization employed by the SentencePiece[33] framework—to flexibly define token units. This helped reduce sequence information loss and improved computational efficiency. As a result,

transformer-based models have been successfully applied to tasks such as identifying promoters and transcription factor binding sites (TFBSs) directly from DNA sequences. Caduceus[34] employs character-level (base-pair) tokenization, which ensures robustness to minor sequence variations. Furthermore, by modeling DNA sequences bidirectionally and incorporating reverse complement (RC) equivariance, Caduceus demonstrates superior performance on tasks such as regulatory site prediction and long-range SNP effect inference. Recently, research has been performed beyond masked language modeling toward generative approaches, such as MegaDNA,[35] a transformer-based DNA sequence generation model.

GenSLM[36] is an RNA language model capable of mutation effect prediction by capturing the differences between original and mutated RNA sequences and predicting their functional effects. The model uses a codon-level vocabulary, which avoids frame shift issues, for tokenizing RNA sequences. The study addresses input lengths that exceed the standard maximum capacity of the standard transformer. This limitation has been identified as a fundamental architectural bottleneck in early foundation models designed for nucleotide sequence analysis. Evo,[37] HyenaDNA,[38] and Caduceus[34] have adopted specialized architectures, such as Hyena[39] and Mamba, to support long-sequence modeling.

2.4. **Single-cell language models**

With the accumulation of high-dimensional gene expression data, single-cell language models have emerged as a new frontier in biology. While proteins and nucleotides are naturally sequential, single-cell gene expression data are not universally sequential. Therefore, a method of ranking genes based on their expression levels has been proposed. Genes within a cell are treated as words in a sentence, and Transformer-based models are applied to capture their underlying dependencies, as in other biological language modeling tasks.

Recent advances in single-cell representation learning have surpassed traditional marker gene-based approaches in capturing cellular heterogeneity.[40] scBERT[41] addresses this limitation by leveraging full gene expression profiles, achieving strong performance in cell type annotation. Geneformer[42] handles the non-sequential nature of gene expression data by ordering genes based on count statistics, also showing effectiveness in classification tasks. Building on this, scGPT[43] takes gene embeddings as input tokens and outputs a cell embedding, jointly learning representations at both levels. It achieves state-of-the-art results across tasks such as cell type classification, perturbation prediction, batch correction, and multi-omics integration. These findings emphasize the value of large-scale single-cell datasets (e.g., the Human Cell Atlas,[44] CellMarker[45]) and the potential of embedding models to capture cellular complexity.

At the same time, approaches have been proposed to leverage general-purpose LLMs for directly incorporating prior biological knowledge, going beyond gene sequence modeling alone. For example, despite being trained on common human languages, GPT-4 has shown the ability to perform automatic cell type annotation based on text prompts describing gene expression levels.[46] Accordingly, GenePT[47] and scELMo[48] have constructed gene- and cell-level embeddings by applying text embedding APIs from a corpus of biomedical literature including the NCBI database. It has been reported to outperform some biological data-driven models such

as Geneformer.[42] In addition, CancerGPT,[49] a GPT-3[50] model fine-tuned on corpora of text, predicts drug response pairs within rare tissue types by aligning textual representations with cellular information. Developing disease-specific models with refined cell embeddings may further advance precision medicine.

2.5. **Biomolecule representations**

Biological macromolecules such as proteins and nucleic acids can be represented through diverse modalities to support machine learning applications. Sequence-based representations use amino acid or nucleotide strings and serve as the foundation for protein and genomic language models such as ESM,[2] ProtBERT,[12] and DNABERT[29,31]. Structural representations capture spatial information using atomic coordinates, contact maps, or distance matrices, which are leveraged in structure models like AF and ESMFold. Graph-based approaches abstract biomolecules into nodes and edges, enabling the use of geometric deep learning models such as SE(3) Transformer.[51] Functional representations include Gene Ontology terms, protein family annotations, and subcellular localization, enriching models with biological context. At the cellular level, omics data like scRNA-seq is encoded as high-dimensional expression vectors.

2.6. **Tokenization strategies**

Tokenization methods have evolved from traditional machine learning techniques, including k-mer approaches,[52] to biomolecule-specialized strategies such as structure- and codon-based tokenization,[53] which are critical for accurate and detailed biomolecular modeling. In protein and nucleotide models, k-mer tokenization (e.g., 3-mer, 6-mer) is used to capture local biochemical context, as seen in DNABERT and ProtBERT. Some models use byte-pair encoding (BPE) or unigram models trained on large corpora of sequences, such as DNABERT2, ESM, and ProGen. Codon-based or codon-preserving tokenization are also adopted to avoid frame-shift artifacts in nucleotide modeling. scBERT employs the gene2vec approach to generate gene embeddings, which facilitates the application of the BERT architecture to single-cell RNA sequencing data. These customized strategies ensure efficient representation of biological syntax and semantics in pretrained language models.

2.7. **Application of BLMs in Biomedicine**

2.7.1. **Integrative modeling for molecular cell biology**

AF2 demonstrated the strength of AI in protein structure prediction and has since inspired a wide range of follow-up studies. Models such as AlphaFold3,[54] RoseTTAFoldNA,[55] and RoseTTAFold All-Atom[56] extend their focus beyond proteins to include other biologically relevant molecules such as RNA, DNA, and ligands. In particular, all-atom structure prediction introduces computational challenges in accurately reconstructing 3D coordinates. This reflects a growing recognition that structural accuracy is essential for understanding biomolecular function

not only in proteins, but also in RNA, where structure plays a critical role in regulatory activity.[57] Concurrently, large language model (LLM)-based methods have begun to incorporate structural information, moving beyond sequence modeling. ESM3[58] jointly embeds sequence, structure, and function marking a transition toward multimodal representation. Specialized models such as ESM-DBP[59] have also been developed to predict DNA-binding proteins, adopting hybrid approaches that leverage both sequence and structure features. In the context of unified modeling in biological language models, foundation models aim to learn comprehensive cellular representations by integrating diverse biological modalities. These include epigenetic marks, spatial transcriptomics, protein expression data, and perturbation signatures, which can be explored to gain a deeper understanding into cellular function.[60] The integration signals a broader shift from modality-specific models toward unified representations that more reasonably reflect the inherent complexity of biological systems.

2.7.2. **Multimodal foundation models**

Multimodal Large Language Models (MLLMs) offer a framework for aligning heterogeneous data types such as clinical notes, protein sequences, and molecular structures.

BiomedGPT[61] aligns natural language with biomedical modalities, particularly visual representations, to enable cross-modal reasoning for visual-language tasks. It focuses on applications such as diagnosis, summarization, clinical decision support through flexible query answering. However**,** such models still exhibit limitations in reasoning across complex clinical scenarios, including the interpretation of radiological images and the resolution of textual conflicts. MediConfusion[62] provides a diagnostic benchmark that systematically evaluates failure modes of multimodal medical LLMs.

Tx-LLM[63] leverages the advantages of large-scale pretraining on diverse biological datasets. Specifically, It is trained on sequence-level information encompassing RNA, DNA, protein sequences, as well as SMILES. This comprehensive approach enables positive transfer performance in end-to-end drug discovery tasks, outperforming models that do not incorporate biological sequence data. Similarly, BioMedGPT-10B[64] contributes to drug discovery by specializing in protein and molecule Question and Answering (QA), having been trained on cell sequences, protein and molecule structures. These advancements highlight the potential of LLMs to serve as unified multimodal platforms in biomedicine. (**Fig. 2**).

3. **Chemical language models**

Chemical Language Models (CLMs) have been suggested to learn the structure-activity relationship of small molecules from large-scale chemical data using various sequential representations of molecules, e.g. Simplified Molecular Input Line Entry System (SMILES).[65]

3.1. Models Types

Similar to pLMs, most CLMs leverage Transformer architectures,[66] akin to those in natural language processing, to understand, generate, and manipulate chemical structures and reactions. These models are broadly categorized based on their architectural design, each optimized for distinct tasks within cheminformatics and drug discovery. The primary model types include encoder-only (BERT-like) models, decoder-only (GPT-like) models, and encoder-decoder architectures, as well as emerging multi-modal LLMs that integrate diverse data formats (**Fig. 3**). These architectural choices dictate how the models process molecular representations and perform tasks ranging from property prediction to *de novo* molecular design and retrosynthesis.

3.1.1. **Chemical encoders**

Encoder-only transformer models, primarily inspired by BERT, are designed to extract contextual representations of molecules and are well-suited for property prediction and molecular understanding. ChemBERTa[67] adapts the RoBERTa[68] framework with MLM and multitask regression, where auxiliary property prediction tasks are defined using molecular features computed by RDKit[69]. Mol-BERT[70] applies MLM to learn chemically informed token-level dependencies and is fine-tuned for tasks such as property classification and activity prediction. MoLFormer[71] extends this approach using linear attention and rotary embeddings, yielding compact representations useful for downstream regression and classification tasks, though it is limited to relatively small molecules. Further encoder variants refine token representations or integrate structural priors. MolRoPE-BERT[72] enhances positional encoding, while MFBERT,[73] SELFormer,[74] and semi-RoBERTa[75]introduce architectural modifications for greater chemical expressiveness. Graph-enhanced encoders like GROVER[76] incorporate topological features directly, bridging the gap between sequence and graph representations.

3.1.2. **Chemical decoders**

Decoder-only transformer models, following the GPT architecture, are optimized for autoregressive generation and have become essential in de novo molecular design. MolGPT[77] prioritizes causality to learn token-wise dependencies and ultimately generates novel molecules. It supports conditional generation strategies to bias outputs toward specific chemical properties. GP-MoLFormer[78] is a decoder-only adaptation of MoLFormer-XL[71] and optimized for tasks such as unconstrained molecule generation, scaffold completion, and conditional property optimization. Other GPT-based chemical models include SMILES-GPT[79] and iupacGPT,[80] both adapted from GPT-2[81] for molecular and nomenclature sequence generation. cMolGPT[82] extends this framework for controllable generation under property or scaffold constraints. Taiga[83] combines GPT modeling with reinforcement learning to guide molecule synthesis toward multi-objective goals.

3.1.3. **Encoder-Decoder Architectures**

Encoder–decoder transformer models are designed for seq-to-seq tasks, making them particularly effective for applications such as retrosynthesis, reaction prediction, and cross-domain molecular translation. Text+ChemT5[84] employs a shared encoder-decoder T5 backbone to support dual-modality tasks across chemical and natural language domains, including text-to-molecule generation and vice versa. SELFIES-TED,[85] built upon a BART-style encoder-decoder structure, is tailored for chemically constrained generation tasks. It consistently performs well across molecular prediction and generative benchmarks, showing strong generalizability.

Beyond these, Chemformer[86] and BARTSmiles[87] adopt the BART architecture for generative and discriminative molecular tasks. MOLGEN[88] introduces self-feedback during pretraining to better align model output with chemically realistic constraints. Models like Molecular Transformer,[1] Retrosynthesis Transformer,[89] and SCROP[90] focus on forward and backward reaction prediction, employing techniques such as snapshot learning, syntax correction, and beam search to enhance accuracy and syntactic validity. Hybrid approaches also emerge: GO-PRO[91] integrates context-free grammars, RetroTRAE[92] tracks atom-level transformations via fragment tokenization, and GCT[93] augments the Transformer with a conditional variational autoencoder for latent sampling. Prompt-driven models such as RetroSynth-Diversity[94] and the Disconnection-Aware Transformer[95] further refine retrosynthesis by guiding outputs based on fragmentation strategies or disconnection heuristics.

3.1.4. **Multi-modal LLMs**

Chemical information is inherently multi-modal, encompassing textual descriptions, molecular graphs, 2D depictions, 3D coordinates, and higher-dimensional properties, such as polarizability. Standard CLMs, designed for handling only with text format, cannot fully capture heterogeneous information. To address this, recent CLMs integrate LLMs with structural encoders to enable cross-modal reasoning. Mol-LLaMA[96] incorporates graph representations into a language model, improving tasks such as functional group identification and retrosynthesis. GIT-Mol[97] processes graphs, images, and text through separate encoders, then fuses their modality-specific tokens via a shared representation layer. Contrastive objectives are used to align modalities, and a joint prediction head enables multi-task learning across modalities. LLM-MPP[98] similarly aligned SMILES, 2D graphs, and text descriptions through cross-attention and contrastive learning to enable coherent molecular representation. Vision-language models such as PRESTO[99] and ChemVLM[100] jointly encode molecular depictions and associated texts to support synthesis planning and reaction condition inference. nach0[101] treats SMILES, images, and text as aligned modalities within a shared representation space for multimodal reasoning. Collectively, these approaches reflect the expanding range of design strategies aimed at achieving effective modality fusion within chemical language models.

3.2. **Pretraining and fine-tuning Strategies**

3.2.1. **Self-Supervised Learning (SSL)**

Self-supervised learning (SSL) is a powerful subset of unsupervised learning where labels are automatically generated from the input data itself. This approach is commonly employed for pretraining models on large, unlabeled datasets, which is crucial for ensuring the generalizability of the learned representations. In this regard, Masked Language Modeling (MLM) is a widely adopted pretraining task for encoder-based language models. In this approach, a certain percentage of tokens (e.g., 15%) in the input sequence are randomly masked, and the model is trained to predict these masked tokens based on the surrounding context.[102] This forces the model to learn deep contextual representations and implicitly understand the underlying chemical syntax and molecular structures, which can then be transferred to various downstream tasks. Denoising objectives are another form of SSL where the model is trained to reconstruct the original, clean input from a corrupted or "noisy" version.[103,104]

3.2.2. **Multi-Task Learning (MTL)**

Multi-task learning (MTL) is a powerful paradigm that utilizes shared information across multiple related learning tasks to improve generalization and overall performance. By training a single model on several tasks simultaneously, it is compelled to learn common patterns and representations that are beneficial across all tasks.[105] This approach can be conceptualized as machines mimicking human learning, where knowledge acquired from one task can effectively benefit and improve performance on other related tasks.[106] MTL is particularly advantageous in molecular prediction tasks, as it helps alleviate data sparsity issues by allowing models to draw strength from diverse but related datasets, leading to improved accuracy. Models like **Text+ChemT5** exemplify this by being multi-domain, multi-task LMs that concurrently handle both chemical and natural language. They achieve this by sharing weights across these distinct domains and tasks, fostering a unified understanding. Similarly, **nach0-pc**[107] is a multi-task LM specifically designed for 3D molecular structures, demonstrating its capability to process complex point cloud data effectively within a multi-task framework

3.2.3. **Retrieval-Augmented Generation (RAG)**

Retrieval-Augmented Generation (RAG) enhances language models by integrating a latent retriever that dynamically accesses external documents during pretraining, fine-tuning, and inference.[108] In chemistry, this modular architecture improves performance on tasks like molecule design, retrosynthesis, and reaction prediction, with reported gains of up to 17.4% over standard inference.[109] It also mitigates hallucinations by grounding predictions in up-to-date domain-specific data.[110] However, conventional RAG often neglects structural dependencies among retrieved documents. Models like **ATLANTIC**[111] address this by building heterogeneous document graphs and using frozen GNNs for context encoding, thus improving retrieval quality while preserving computational efficiency.

3.2.4. **Supervised Fine-tuning**

Supervised fine-tuning adapts pretrained CLMs to specific tasks using labeled datasets. It aligns model outputs with experimental annotations through continued gradient-based optimization, supporting applications such as property prediction, reaction classification, and synthesis planning. While full-model fine-tuning often yields strong performance, it can be computationally intensive and prone to overfitting in low-resource settings. To mitigate these issues, several parameter-efficient alternatives have emerged including adapter tuning, prefix tuning, prompt tuning, and LoRA (Low-Rank Adaptation),[112] which constrain the number of trainable parameters while maintaining task adaptability. These methods offer scalable alternatives that retain the benefits of large-scale pretraining. Regardless of strategy, fine-tuning success depends critically on data quality, which remains central to achieving reliable and interpretable outcomes.

3.3. **Molecular Representations**

Just as natural language processing relies on effective text tokenization and embedding, CLMs depend on robust and informative molecular representations. This section explores the primary representation schemes, detailing their principles, advantages, and the challenges they pose for AI systems.

3.3.1. **String-Based Representations**

SMILES[65] encodes molecular structures as linear ASCII strings, originally developed for efficient data storage. Its brevity, machine-readability, and invertibility have made it widely adopted in cheminformatics and compatible with language models. However, SMILES presents several limitations when used with LLMs. A key issue is non-uniqueness: a single molecule can have multiple valid SMILES representations, which complicates training and reduces model generalization. Canonicalization reduces this ambiguity but often encourages overfitting to syntactic patterns rather than learning underlying chemical rules. SMILES also lacks explicit stereochemistry and 3D spatial information, and many datasets omit annotations for chiral centers and geometric isomers, limiting its utility in structure-sensitive tasks. To address these shortcomings, several extensions to the SMILES grammar have been proposed, including DeepSMILES,[113] SELFIES[3] and Atom-in-SMILES (AIS)[4] with particular attention to improving validity, interpretability, and compatibility with machine learning systems.

3.3.2. **Graph-Based Representations**

Graph-based inputs capture connectivity and topological constraints absent in SMILES, offering richer structural context. Models like GROVER[76] and MG-BERT[114] incorporate GNN-derived embeddings or graph-attention mechanisms to bridge this gap. Despite their promise, graph-based hybrid CLMs face challenges in tokenization, alignment with sequence models, and limited standardized pipelines. Ongoing efforts focus on improving graph serialization,

integrating positional encodings for graphs, and combining GNNs with pre-trained transformers in parameter-efficient ways.

3.3.3. **3D Point Cloud Representations**

Recent advances have extended CLMs beyond linear and 2D representations by incorporating explicit three-dimensional molecular structure, particularly via point cloud-based methodologies. These models exploit geometric deep learning to capture spatial features critical to tasks such as molecular property prediction and drug design. Notably, models range from specialized 3D-aware transformers like Uni-Mol[115] to multimodal architectures such as nach0-pc[107] and 3D-MolT5[116] combine molecular point cloud encoders with language models to learn from atomic spatial arrangements via joint training on multi-task 3D datasets. These approaches represent a shift toward spatially grounded CLMs with improved capacity to model complex molecular geometry and interactions.

3.4. **Tokenization strategies**

Tokenization in CLMs refers to the transformation of molecular strings into discrete, model-readable units. Unlike general NLP, token boundaries in chemistry must respect atomic symbols, charges, and bonding syntax. For instance, character-level tokenization is ill-suited for SMILES, as it often produces tokens that represent unphysical or chemically meaningless characters. Canonical schemes like Byte Pair Encoding (BPE),[32] while effective in NLP, tend to merge chemically unrelated substrings under frequency pressure. To address this, domain-specific approaches such as Atom Pair Encoding (APE)[117] or token frequency regularization have been proposed. Atom-in-SMILES (AIS)[4] tokenization embeds local topological context—such as neighboring atoms or ring membership—into tokens, improving resolution without altering syntax. This yields more balanced token distributions and enhances optimization performance in low-data regimes.

3.5. **Applications CLMs in Biomedicine**

CLMs are being increasingly used across biomedicinal research, particularly in drug discovery. These models predict molecular properties such as solubility, bioavailability, and toxicity directly from string representations, allowing rapid screening of candidate compounds and reducing dependence on experimental assays. CLMs also support de novo molecule generation.[118] Autoregressive and diffusion-based models can design novel compounds with optimized activity, selectivity, or synthetic accessibility. In biomedicine, such tools are applied to generate inhibitors, antibiotics, and CNS-targeted molecules tailored to therapeutic needs.

In chemical synthesis, CLMs predict reaction outcomes and assist in retrosynthetic planning. Sequence-to-sequence models trained on reaction corpora like USPTO[119] suggest plausible routes for synthesizing drug-like molecules, improving efficiency and creativity in medicinal chemistry.[120] CLMs further aid early toxicity assessment by learning structural patterns linked to adverse effects. When trained on toxicology datasets, they support preclinical risk evaluation

more effectively than rule-based methods.[121] These applications reflect how CLMs contribute to faster and more informed decision-making in biomedicinal pipelines.

4. **Datasets for Bio/Chemical Language Models and Benchmarks.**

The effectiveness of language models in biological and chemical domains is closely tied to the diversity, structure, and scale of training data. Numerous datasets have been curated to support tasks such as molecular property prediction, reaction modeling, clinical text understanding, and biomedical question answering.

Chemical structure databases such as ZINC,[122] PubChem,[123] and ChEMBL[124] provide millions of small molecules in SMILES format, supporting the learning of chemical syntax and structure–activity relationships. For modeling reactivity and synthesis, datasets like USPTO,[119] Reaxys,[125] QM9[126] and QMugs[127] offer extensive reaction and quantum property annotations. To evaluate predictive performance across physical and bioactivity tasks, benchmarks such as MoleculeNet[128] (ESOL, FreeSolv, Lipophilicity, Tox21, SIDER, BBBP, and HIV) are widely adopted.

On the biomedical side, large corpora such as PubMed,[129] PubMed Central (PMC),[130] and clinical datasets including MIMIC-III,[131] eICU,[132] and i2b2[133] enable models to learn domain-specific language and clinical reasoning patterns. Complementary benchmarks—such as MedQA,[134] PubMedQA,[135] BioASQ[136] and MultiMedQA[137]—serve for evaluating medical question answering and multi-hop inference capabilities. For therapeutic science and knowledge extraction, specialized datasets like Therapeutics Data Commons (TDC),[138] DisGeNET,[139] DrugBank,[140] PHARMGKB[141] and STRING[142] offer structured annotations across gene–disease, drug–target, and protein–interaction networks.

5. **The use of LLMs in biology and chemistry**

Prompt engineering has emerged as the most accessible way to adapt ChatGPT and other LLMs to scientific problems without additional training. This method relies on carefully crafted textual prompts to direct model outputs, with applications ranging from SMILES translation to molecular property prediction. Techniques like zero-shot, few-shot, and chain-of-thought (CoT) prompting[143] have demonstrated utility across diverse chemical tasks. Liu et al.[144] showed that performance can be surprisingly strong for tasks like retrosynthesis planning and reaction classification. However, prompt sensitivity, limited domain knowledge, and inconsistent outputs remain major limitations.[145,146] Despite these drawbacks, prompt engineering has facilitated rapid, low-resource adaptation of LLMs to domain-specific tasks.

For greater task specificity, fine-tuning offers a more robust route. By continuing pretraining on chemical corpora—including scientific literature and curated datasets—ChatGPT-like models

can internalize domain language and logic. Studies have demonstrated improved performance in chemical property regression, reaction prediction, and literature-based knowledge extraction.[147,148] Domain-specific fine-tuning, even on modest datasets, has shown effectiveness in areas like inorganic chemistry or thermoelectrics.[149,150] Nevertheless, outcomes are highly sensitive to dataset quality and task design.[151,152]

At the frontier, ChatGPT has been combined with agentic systems that allow LLMs to use tools, yielding multi-step workflows and autonomous decision-making. Approaches like ReAct[153] orchestrate LLM reasoning with external tool use. Coscientist,[154] ChemCrow,[155] and ChatMOF[156] exemplify such frameworks, integrating web search, retrosynthesis tools, and lab protocols into interactive agents. In automation contexts, platforms like Chemputer[157] and Organa[158] have demonstrated LLM-guided synthesis planning and lab execution. These systems promise scalable scientific automation but remain dependent on deterministic toolchains and human oversight.

6. **Conclusion**

The convergence of large-scale data and advanced computation has established a new paradigm in the molecular sciences, re-conceptualizing both biological and chemical systems as structured languages amenable to deep learning. At the core of this transformation is the shared challenge of representation: translating complex, multi-dimensional molecular information—from protein sequences and single-cell expression profiles to SMILES strings—into formats that learning architectures can effectively process. As detailed in this review, this has led to advanced models for predicting protein structure, interpreting genomic regulation, and performing de novo molecular design and synthesis planning.

This development reflects a broader shift toward unified, multimodal frameworks that integrate diverse data types to build more comprehensive and robust foundation models. As these architectures mature, future progress will depend critically on aligning model capabilities with fundamental biological and chemical prior knowledge during the learning process. The development of standardized benchmarks for rigorous model evaluation and improving model interpretability to build trust and guide experimental validation are also essential.

Looking ahead, the development of bio/chemical LLMs is heading toward more interactive and agentic systems capable of assisting with hypothesis generation and experimental design. By successfully addressing the aforementioned challenges, these bio/chemical language models are poised to become foundational platforms for creating more predictive, generative, and reliable tools, which will significantly accelerate the design-build-test-learn cycle in molecular science.

Acknowledgments

The authors are grateful for support from the National Research Foundation of Korea (2020M3A9G7103933, 2022M3E5F3081268, RS-2023-00256320, 2022R1C1C1005080 J.L.), the Korea Environment Industry & Technology Institute (KEITI) funded by the Korea Ministry of Environment (MOE) (RS-2023-00219144 J.L.), and Institute of Information & communications Technology Planning &Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2023-00220628 J.L.). This work was also partially supported by New Faculty Startup Fund from Seoul National University.

Conflict of Interest

The authors declare no competing interests.

Reference

1. Schwaller, P. *et al.* Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
2. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
3. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn.: Sci. Technol.* **1**, 045024 (2020).
4. Ucak, U. V., Ashyrmamatov, I. & Lee, J. Improving the quality of chemical language model outcomes with atom-in-SMILES tokenization. *J Cheminform* **15**, 55 (2023).
5. Kaplan, J. *et al.* Scaling Laws for Neural Language Models. Preprint at https://doi.org/10.48550/arXiv.2001.08361 (2020).
6. Bairoch, A. & Boeckmann, B. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res* **20 Suppl**, 2019–2022 (1992).
7. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
8. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365–370 (2003).
9. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Research* **40**, D136–D143 (2012).
10. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
11. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
12. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102–2110 (2022).
13. Rao, R. M. *et al.* MSA Transformer. in *Proceedings of Machine Learning Research* (eds. Meila, M. & Zhang, T.) vol. 139 8844–8856 (PMLR, 2021).
14. Elnaggar, A. *et al.* ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans Pattern Anal Mach Intell* **44**, 7112–7127 (2022).

15. Sgarbossa, D., Malbranke, C. & Bitbol, A.-F. ProtMamba: a homology-aware but alignment-free protein state space model. *Bioinformatics* btaf348 (2025) doi:10.1093/bioinformatics/btaf348.
16. Gu, A. & Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. Preprint at https://doi.org/10.48550/arXiv.2312.00752 (2024).
17. Madani, A. *et al.* ProGen: Language Modeling for Protein Generation. *arXiv [q-bio.BM]* (2020).
18. Nijkamp, E., Ruffolo, J., Weinstein, E. N., Naik, N. & Madani, A. ProGen2: Exploring the Boundaries of Protein Language Models. *arXiv [cs.LG]* (2022).
19. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications* **13**, 4348 (2022).
20. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
21. Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
22. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* **42**, 243–246 (2024).
23. Chen, L. *et al.* AI-Driven Deep Learning Techniques in Protein Structure Prediction. *IJMS* **25**, 8426 (2024).
24. The UniProt Consortium *et al.* UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research* **53**, D609–D617 (2025).
25. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
26. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682 (2022).
27. Powell, H. R., Islam, S. A., David, A. & Sternberg, M. J. E. Phyre2.2: A Community Resource for Template-based Protein Structure Prediction. *Journal of Molecular Biology* **437**, 168960 (2025).
28. Zhang, Y., Qiao, S., Ji, S. & Li, Y. DeepSite: bidirectional LSTM and CNN models for predicting DNA–protein binding. *International Journal of Machine Learning and Cybernetics* **11**, 841–851 (2020).
29. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).
30. Sanabria, M., Hirsch, J. & Poetsch, A. R. The human genome's vocabulary as proposed by the DNA language model GROVER. *bioRxiv* 2023.07.19.549677 (2023) doi:10.1101/2023.07.19.549677.
31. Zhou, Z. *et al.* DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. *arXiv [q-bio.GN]* (2024).
32. Sennrich, R., Haddow, B. & Birch, A. Neural Machine Translation of Rare Words with Subword Units. Preprint at https://doi.org/10.48550/arXiv.1508.07909 (2016).
33. Kudo, T. & Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. Preprint at https://doi.org/10.48550/arXiv.1808.06226 (2018).

34. Schiff, Y. *et al.* Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling. (2024) doi:10.48550/arXiv.2403.03234.
35. Shao, B. A long-context language model for deciphering and generating bacteriophage genomes. *bioRxiv* 2023.12.18.572218 (2024) doi:10.1101/2023.12.18.572218.
36. Zvyagin, M. *et al.* GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. *bioRxiv* 2022.10.10.511571 (2022) doi:10.1101/2022.10.10.511571.
37. Nguyen, E. *et al.* Sequence modeling and design from molecular to genome scale with Evo. *Science* **386**, eado9336 (2024).
38. Nguyen, E. *et al.* HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. *arXiv [cs.LG]* (2023).
39. Poli, M. *et al.* Hyena Hierarchy: Towards Larger Convolutional Language Models. Preprint at https://doi.org/10.48550/arXiv.2302.10866 (2023).
40. Zhang, Z. *et al.* SCINA: A semi-supervised subtyping algorithm of single cells and bulk samples. *Genes (Basel)* **10**, 531 (2019).
41. Yang, F. *et al.* scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence* **4**, 852–866 (2022).
42. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
43. Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods* **21**, 1470–1480 (2024).
44. Rood, J. E. *et al.* The Human Cell Atlas from a cell census to a unified foundation model. *Nature* **637**, 1065–1071 (2025).
45. Hu, C. *et al.* CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Research* **51**, D870–D876 (2023).
46. Hou, W. & Ji, Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nature Methods* **21**, 1462–1465 (2024).
47. Chen, Y. & Zou, J. GenePT: A Simple But Effective Foundation Model for Genes and Cells Built From ChatGPT. *bioRxiv* 2023.10.16.562533 (2024) doi:10.1101/2023.10.16.562533.
48. Liu, T., Chen, T., Zheng, W., Luo, X. & Zhao, H. scELMo: Embeddings from Language Models are Good Learners for Single-cell Data Analysis. *bioRxiv* 2023.12.07.569910 (2023) doi:10.1101/2023.12.07.569910.
49. Li, T. *et al.* CancerGPT for few shot drug pair synergy prediction using large pretrained language models. *npj Digit. Med.* **7**, 40 (2024).
50. Brown, T. B. *et al.* Language Models are Few-Shot Learners. Preprint at https://doi.org/10.48550/arXiv.2005.14165 (2020).
51. Yim, J. *et al.* SE(3) diffusion model with application to protein backbone generation. Preprint at https://doi.org/10.48550/arXiv.2302.02277 (2023).
52. Dotan, E., Jaschek, G., Pupko, T. & Belinkov, Y. Effect of tokenization on transformers for biological sequences. *Bioinformatics* **40**, btae196 (2024).
53. Outeiral, C. & Deane, C. M. Codon language embeddings provide strong signals for use in protein engineering. *Nat Mach Intell* **6**, 170–179 (2024).
54. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

55. Baek, M. *et al.* Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nature Methods* **21**, 117–121 (2024).
56. Krishna, R. *et al.* Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, eadl2528.
57. Vicens, Q. & Kieft, J. S. Thoughts on how to think (and talk) about RNA structure. *Proceedings of the National Academy of Sciences* **119**, e2112677119 (2022).
58. Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).
59. Zeng, W., Dou, Y., Pan, L., Xu, L. & Peng, S. Improving prediction performance of general protein language model by domain-adaptive pretraining on DNA-binding protein. *Nature Communications* **15**, 7838 (2024).
60. Cui, H. *et al.* Towards multimodal foundation models in molecular cell biology. *Nature* **640**, 623–633 (2025).
61. Zhang, K. *et al.* A generalist vision–language foundation model for diverse biomedical tasks. *Nat Med* **30**, 3129–3141 (2024).
62. Sepehri, M. S., Fabian, Z., Soltanolkotabi, M. & Soltanolkotabi, M. MediConfusion: Can you trust your AI radiologist? Probing the reliability of multimodal medical foundation models. in (2024).
63. Chaves, J. M. Z. *et al.* Tx-LLM: A Large Language Model for Therapeutics. Preprint at https://doi.org/10.48550/ARXIV.2406.06316 (2024).
64. Luo, Y. *et al.* BioMedGPT: Open Multimodal Generative Pre-trained Transformer for BioMedicine. (2023) doi:10.48550/arXiv.2308.09442.
65. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
66. Vaswani, A. *et al.* Attention Is All You Need. Preprint at https://doi.org/10.48550/arXiv.1706.03762 (2023).
67. Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv* (2020) doi:10.48550/arxiv.2010.09885.
68. Liu, Y. *et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* (2019) doi:10.48550/arxiv.1907.11692.
69. Landrum, G. The RDKit 2019.09.01 Documentation. (2019).
70. Li, J. & Jiang, X. Mol-BERT: An Effective Molecular Representation with BERT for Molecular Property Prediction. *Wireless Communications and Mobile Computing* **2021**, 1–7 (2021).
71. Ross, J. *et al.* Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence* **4**, 1256–1264 (2022).
72. Liu, Y. *et al.* MolRoPE-BERT: An enhanced molecular representation with Rotary Position Embedding for molecular property prediction. *Journal of Molecular Graphics and Modelling* **118**, 108344 (2023).
73. Abdel-Aty, H. & Gould, I. R. Large-Scale Distributed Training of Transformers for Chemical Fingerprinting. *Journal of Chemical Information and Modeling* **62**, 4852–4862 (2022).
74. Yüksel, A., Ulusoy, E., Ünlü, A. & Doğan, T. SELFormer: molecular representation learning via SELFIES language models. *Machine Learning: Science and Technology* **4**, 025035 (2023).

75. Tran, T. & Ekenna, C. Molecular Descriptors Property Prediction Using Transformer-Based Approach. *International Journal of Molecular Sciences* **24**, 11948 (2023).
76. Rong, Y. *et al.* Self-Supervised Graph Transformer on Large-Scale Molecular Data. *arXiv* (2020) doi:10.48550/arxiv.2007.02835.
77. Bagal, V., Aggarwal, R., Vinod, P. K. & Priyakumar, U. D. MolGPT: Molecular Generation Using a Transformer-Decoder Model. *Journal of Chemical Information and Modeling* **62**, 2064–2076 (2022).
78. Ross, J. *et al.* GP-MoLFormer: A Foundation Model For Molecular Generation. *arXiv* (2024) doi:10.48550/arxiv.2405.04912.
79. Adilov, S. Generative Pre-Training from Molecules. (2021) doi:10.26434/chemrxiv-2021-5fwjd.
80. Cho, K.-H. & No, K. T. iupacGPT: IUPAC-based large-scale molecular pre-trained model for property prediction and molecule generation. (2023) doi:10.26434/chemrxiv-2023-5kjvh.
81. Alec Radford *et al.* Language Models are Unsupervised Multitask Learners. (2019).
82. Wang, Y., Zhao, H., Sciabola, S. & Wang, W. cMolGPT: A Conditional Generative Pre-Trained Transformer for Target-Specific De Novo Molecular Generation. *Molecules* **28**, 4430 (2023).
83. Mazuz, E., Shtar, G., Shapira, B. & Rokach, L. Molecule generation using transformers and policy gradient reinforcement learning. *Scientific Reports* **13**, 8799 (2023).
84. Christofidellis, D. *et al.* Unifying Molecular and Textual Representations via Multi-task Language Modelling. *arXiv* (2023) doi:10.48550/arxiv.2301.12586.
85. Priyadarsini, I. *et al.* SELFIES-TED : A Robust Transformer Model for Molecular Representation using SELFIES. (2025).
86. Irwin, R., Dimitriadis, S., He, J. & Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology* **3**, 015022 (2022).
87. Chilingaryan, G. *et al.* BartSmiles: Generative Masked Language Models for Molecular Representations. *Journal of Chemical Information and Modeling* **64**, 5832–5843 (2024).
88. Fang, Y. *et al.* Domain-Agnostic Molecular Generation with Chemical Feedback. *arXiv* (2023) doi:10.48550/arxiv.2301.11259.
89. Schwaller, P. *et al.* Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical Science* **11**, 3316–3325 (2020).
90. Zheng, S., Rao, J., Zhang, Z., Xu, J. & Yang, Y. Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks. *J. Chem. Inf. Model.* **60**, 47–55 (2020).
91. Mann, V. & Venkatasubramanian, V. Predicting chemical reaction outcomes: A grammar ontology-based transformer framework. *AIChE Journal* **67**, (2021).
92. Ucak, U. V., Ashyrmamatov, I., Ko, J. & Lee, J. Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. *Nat Commun* **13**, 1186 (2022).
93. Kim, H., Na, J. & Lee, W. B. Generative Chemical Transformer: Neural Machine Learning of Molecular Geometric Structures from Chemical Language via Attention. *Journal of Chemical Information and Modeling* **61**, 5804–5814 (2021).
94. Toniato, A., Vaucher, A. C., Schwaller, P. & Laino, T. Enhancing diversity in language based models for single-step retrosynthesis. *Digital Discovery* **2**, 489–501 (2023).
95. Thakkar, A. *et al.* Unbiasing Retrosynthesis Language Models with Disconnection Prompts. *ACS Central Science* **9**, 1488–1498 (2023).

96. Kim, D., Lee, W. & Hwang, S. J. Mol-LLaMA: Towards General Understanding of Molecules in Large Molecular Language Model. Preprint at https://doi.org/10.48550/arXiv.2502.13449 (2025).
97. Liu, P., Ren, Y., Tao, J. & Ren, Z. GIT-Mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in Biology and Medicine* **171**, 108073 (2024).
98. Jin, C., Guo, S., Zhou, S. & Guan, J. Effective and Explainable Molecular Property Prediction by Chain-of-Thought Enabled Large Language Models and Multi-Modal Molecular Information Fusion. *J. Chem. Inf. Model.* **65**, 5438–5455 (2025).
99. Cao, H. *et al.* PRESTO: Progressive Pretraining Enhances Synthetic Chemistry Outcomes. *Findings of the Association for Computational Linguistics: EMNLP 2024* 10197–10224 (2024) doi:10.18653/v1/2024.findings-emnlp.597.
100. Li, J. *et al.* ChemVLM: Exploring the Power of Multimodal Large Language Models in Chemistry Area. *arXiv* (2024) doi:10.48550/arxiv.2408.07246.
101. Livne, M. *et al.* nach0: multimodal natural and chemical languages foundation model. *Chem. Sci.* **15**, 8380–8389 (2024).
102. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* (2018) doi:10.48550/arxiv.1810.04805.
103. Priyadarsini, I. *et al.* SELF-BART : A Transformer-based Molecular Representation Model using SELFIES. *arXiv* (2024) doi:10.48550/arxiv.2410.12348.
104. Ji, X. *et al.* Uni-Mol2: Exploring Molecular Pretraining Model at Scale. *arXiv* (2024) doi:10.48550/arxiv.2406.14969.
105. Deng, Y., Ericksen, S. S. & Gitter, A. Chemical Language Model Linker: blending text and molecules with modular adapters. *arXiv* (2024) doi:10.48550/arxiv.2410.20182.
106. Zhang, Y. & Yang, Q. An overview of multi-task learning. *National Science Review* **5**, 30–43 (2018).
107. Kuznetsov, M. *et al.* nach0-pc: Multi-task Language Model with Molecular Point Cloud Encoder. *arXiv* (2024) doi:10.48550/arxiv.2410.09240.
108. Guu, K., Lee, K., Tung, Z., Pasupat, P. & Chang, M.-W. REALM: Retrieval-Augmented Language Model Pre-Training. *arXiv* (2020) doi:10.48550/arxiv.2002.08909.
109. Zhong, X. *et al.* Benchmarking Retrieval-Augmented Generation for Chemistry. *arXiv* (2025) doi:10.48550/arxiv.2505.07671.
110. Reed, S. M. Augmented and Programmatically Optimized LLM Prompts Reduce Chemical Hallucinations. *Journal of Chemical Information and Modeling* **65**, 4274–4280 (2025).
111. Munikoti, S., Acharya, A., Wagle, S. & Horawalavithana, S. ATLANTIC: Structure-Aware Retrieval-Augmented Language Model for Interdisciplinary Science. *arXiv* (2023) doi:10.48550/arxiv.2311.12289.
112. Hu, E. J. *et al.* LoRA: Low-Rank Adaptation of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2106.09685 (2021).
113. O'Boyle, N. & Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. (2018) doi:10.26434/chemrxiv.7097960.v1.
114. Zhang, X.-C. *et al.* MG-BERT: leveraging unsupervised atomic representation learning for molecular property prediction. *Briefings in Bioinformatics* **22**, bbab152 (2021).

115. Zhou, G. *et al.* Uni-Mol: A Universal 3D Molecular Representation Learning Framework. (2023) doi:10.26434/chemrxiv-2022-jjm0j-v4.
116. Pei, Q., Wu, L., Gao, K., Zhu, J. & Yan, R. 3D-MolT5: Towards Unified 3D Molecule-Text Modeling with 3D Molecular Tokenization. *arXiv* (2024) doi:10.48550/arxiv.2406.05797.
117. Leon, M., Perezhohin, Y., Peres, F., Popovič, A. & Castelli, M. Comparing SMILES and SELFIES tokenization for enhanced chemical language modeling. *Scientific Reports* **14**, 25016 (2024).
118. Grisoni, F. Chemical language models for de novo drug design: Challenges and opportunities. *Current Opinion in Structural Biology* **79**, 102527 (2023).
119. Lowe, D. M. Extraction of chemical structures and reactions from the literature. (Apollo - University of Cambridge Repository, 2012). doi:10.17863/CAM.16293.
120. Zhong, W., Yang, Z. & Chen, C. Y.-C. Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing. *Nat Commun* **14**, (2023).
121. Tran, T. T. V., Surya Wibowo, A., Tayara, H. & Chong, K. T. Artificial Intelligence in Drug Toxicity Prediction: Recent Advances, Challenges, and Future Perspectives. *J. Chem. Inf. Model.* **63**, 2628–2643 (2023).
122. Tingle, B. I. *et al.* ZINC-22─A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery. *J. Chem. Inf. Model.* **63**, 1166–1176 (2023).
123. Kim, S. *et al.* PubChem 2025 update. *Nucleic Acids Research* **53**, D1516–D1525 (2024).
124. Davies, M. *et al.* ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Research* **43**, W612–W620 (2015).
125. Reaxys. Reaxys Reaction Database. https://www.reaxys.com/
126. Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci Data* **1**, (2014).
127. Isert, C., Atz, K., Jiménez-Luna, J. & Schneider, G. QMugs, quantum mechanical properties of drug-like molecules. *Scientific Data* **9**, 273 (2022).
128. Wu, Z. *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
129. National Institues of Health (NIH). PubMed. https://pubmed.ncbi.nlm.nih.gov/
130. National Institues of Health (NIH). PubMed Central. https://pmc.ncbi.nlm.nih.gov/
131. Johnson, A. E. W. *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* **3**, (2016).
132. Pollard, T. J. *et al.* The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci Data* **5**, (2018).
133. Uzuner, Ö., South, B. R., Shen, S. & DuVall, S. L. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* **18**, 552–556 (2011).
134. Jin, D. *et al.* What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. Preprint at https://doi.org/10.48550/arXiv.2009.13081 (2020).
135. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. & Lu, X. PubMedQA: A Dataset for Biomedical Research Question Answering. in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Association for Computational Linguistics, Hong Kong, China, 2019). doi:10.18653/v1/d19-1259.

136. Tsatsaronis, G. *et al.* An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics* **16**, (2015).
137. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
138. Huang, K. *et al.* Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. Preprint at https://doi.org/10.48550/arXiv.2102.09548 (2021).
139. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* (2019) doi:10.1093/nar/gkz1021.
140. Wishart, D. S. *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research* **36**, D901–D906 (2008).
141. Whirl-Carrillo, M. *et al.* An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine. *Clin Pharma and Therapeutics* **110**, 563–572 (2021).
142. Szklarczyk, D. *et al.* The STRING database in 2025: protein networks with directionality of regulation. *Nucleic Acids Research* **53**, D730–D737 (2025).
143. Wei, J. *et al.* Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2201.11903 (2023).
144. Liu, H., Yin, H., Luo, Z. & Wang, X. Integrating chemistry knowledge in large language models via prompt engineering. *Synthetic and Systems Biotechnology* **10**, 23–38 (2025).
145. Vidhani, D. V. & Mariappan, M. Optimizing Human–AI Collaboration in Chemistry: A Case Study on Enhancing Generative AI Responses through Prompt Engineering. *Chemistry* **6**, 723–737 (2024).
146. Hatakeyama-Sato, K., Yamane, N., Igarashi, Y., Nabae, Y. & Hayakawa, T. Prompt engineering of GPT-4 for chemical research: what can/cannot be done? *Science and Technology of Advanced Materials: Methods* **3**, 2260300 (2023).
147. Jacobs, R. *et al.* Regression with Large Language Models for Materials and Molecular Property Prediction. Preprint at https://doi.org/10.48550/ARXIV.2409.06080 (2024).
148. Zhang, W. *et al.* Fine-tuning large language models for chemical text mining. *Chem. Sci.* **15**, 10600–10611 (2024).
149. Thway, M. *et al.* Harnessing GPT-3.5 for text parsing in solid-state synthesis – case study of ternary chalcogenides. *Digital Discovery* **3**, 328–336 (2024).
150. Kim, S., Jung, Y. & Schrier, J. Large Language Models for Inorganic Synthesis Predictions. *J. Am. Chem. Soc.* **146**, 19654–19659 (2024).
151. Xie, Z. *et al.* Fine-tuning GPT-3 for machine learning electronic and functional properties of organic molecules. *Chem. Sci.* **15**, 500–510 (2024).
152. Van Herck, J. *et al.* Assessment of fine-tuned large language models for real-world chemistry and material science applications. *Chem. Sci.* **16**, 670–684 (2025).
153. Yao, S. *et al.* ReAct: Synergizing Reasoning and Acting in Language Models. Preprint at https://doi.org/10.48550/arXiv.2210.03629 (2023).
154. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).
155. M. Bran, A. *et al.* Augmenting large language models with chemistry tools. *Nat Mach Intell* **6**, 525–535 (2024).

156. Kang, Y. & Kim, J. ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nat Commun* **15**, 4705 (2024).
157. Steiner, S. *et al.* Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* **363**, (2019).
158. Darvish, K. *et al.* ORGANA: A robotic assistant for automated chemistry experimentation and characterization. *Matter* **8**, 101897 (2025).

Figure legends

Figure 1. A Comparative Overview of Molecular Representation Modalities. Molecular information can be encoded at multiple levels of abstraction. In the chemical domain (top), a molecule is represented as a 1D SMILES string, a 2D topological graph, or a 3D spatial point cloud. Analogously, in the biological domain (bottom), a protein is represented by its 1D amino acid sequence, a 2D contact map of residue proximities, or its folded 3D structure. The choice of representation dictates the complexity and type of information available to a language model.

Figure 2. Downstream Tasks of Biological Language Models. The diagram organizes downstream tasks of biological language models according to data modalities (DNA/RNA, protein, and single-cell RNA) and illustrates their practical applications in drug discovery, medical QA, and scientific research.

Figure 3. Representative architectures of chemical language models (CLMs). (a) *Encoder-only models* are trained to learn well-representing molecular embeddings, supporting following downstream tasks. (b) *Decoder-only models* generate molecules autoregressively, enabling de novo design or conditional generation. (c) *Encoder-decoder models* handle structured Seq-2-Seq mappings, effective for retrosynthesis and reaction prediction. (d) *Multi-modal models* combine SMILES, graphs, and various types of data to enable integrated molecular reasoning across modalities.

Tables

Category	Modality		Name	Architecture
Modality-specific	ProteinQ	Protein sequence	ProtBERT, MSA Transformer, ESMFold	Transformer (Encoder-only)
			ProtGPT2	Transformer (Decoder-only)
			ProGen, ProGen2	Transformer (Encoder-Decoder)
			ProteinMPNN**	Message-passing neural network*
			ProtMamba	Mamba*
		Protein structure**	AlphaFold2, RoseTTAFold	Biology-specific modules*
			RFdiffusion	Diffusion model (RF architecture)*
	Nucleotide	DNA sequence	DeepSite	CNN, LSTM
			DNABERT, GROVER, DNABERT2	Transformer (Encoder-only)
			MegaDNA	Transformer (Decoder-only)
			HyenaDNA	Hyena*
			Caduceus	Mamba*
		RNA sequence	GenSLM	Transformer
	Single-cell	scRNA sequence	scBERT, Geneformer	Transformer (Encoder-only)
			scGPT	Transformer (Decoder-only)
			scELMo, GenePT	Pre-trained LLM
			CancerGPT	Fine-tuned LLM
Multi-modality	Protein sequence, Protein structure, Protein function		ESM3	Transformer (Encoder-only)
	DNA sequence, RNA sequence		Evo	Hyena* (StripedHyena)
	Protein, Nucleic acid, Ligand		AlphaFold3	Biology-specific modules*
	Image, Text		BiomedGPT	Transformer (Encoder-Decoder)

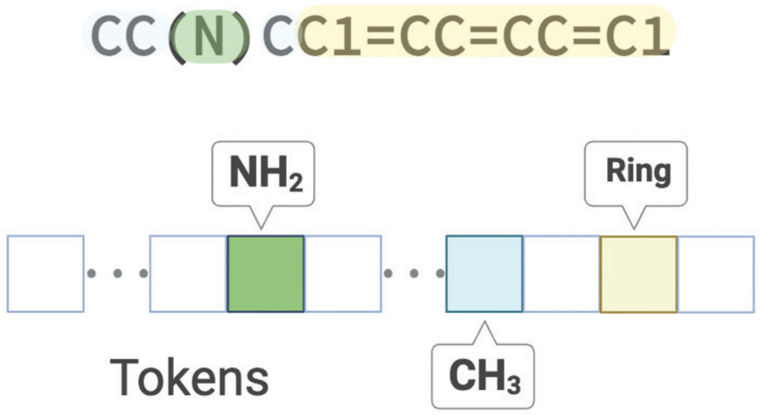
	Protein, Molecule, Text	BioMedGPT-10B	Transformer (Decoder-only)
	Molecule, Protein, Nucleic acid, Cell, Disease, Text	Tx-LLM	Transformer (Decoder-only)

Table 1. Categorization of biological language, structure, and multimodal models

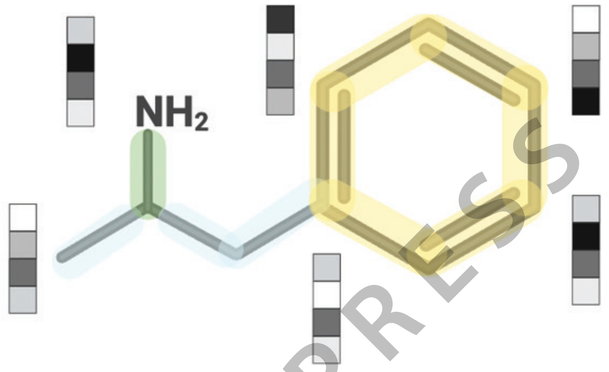
Biological language, structure, and multimodal models can be categorized based on data modality, architecture, and biological specificity. Single-modality data were applied to the modality-specific model, while several modalities of data were used in the multimodal model. The modality includes protein, nucleotide, single-cell, and it includes biological data such as a single sequence and 3D structure, and general data including text and image. * indicates models being customized to reflect biological characteristics using existing architecture. ** The modality is classified as based on the characteristics of the output.

Chemical domain

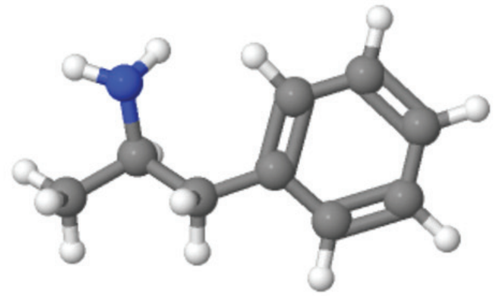
SMILES



Mol Graph



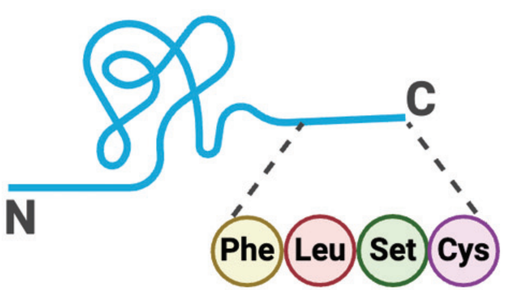
3D Point-Cloud



Sequence → Graph → 3D Structure

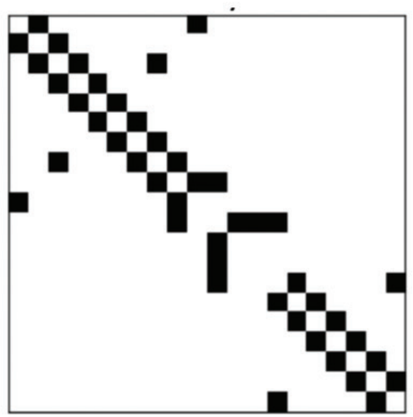
Biological domain

Protein sequence



MLALGGTAHTLDVIAFSQGPGLG...

Contact map



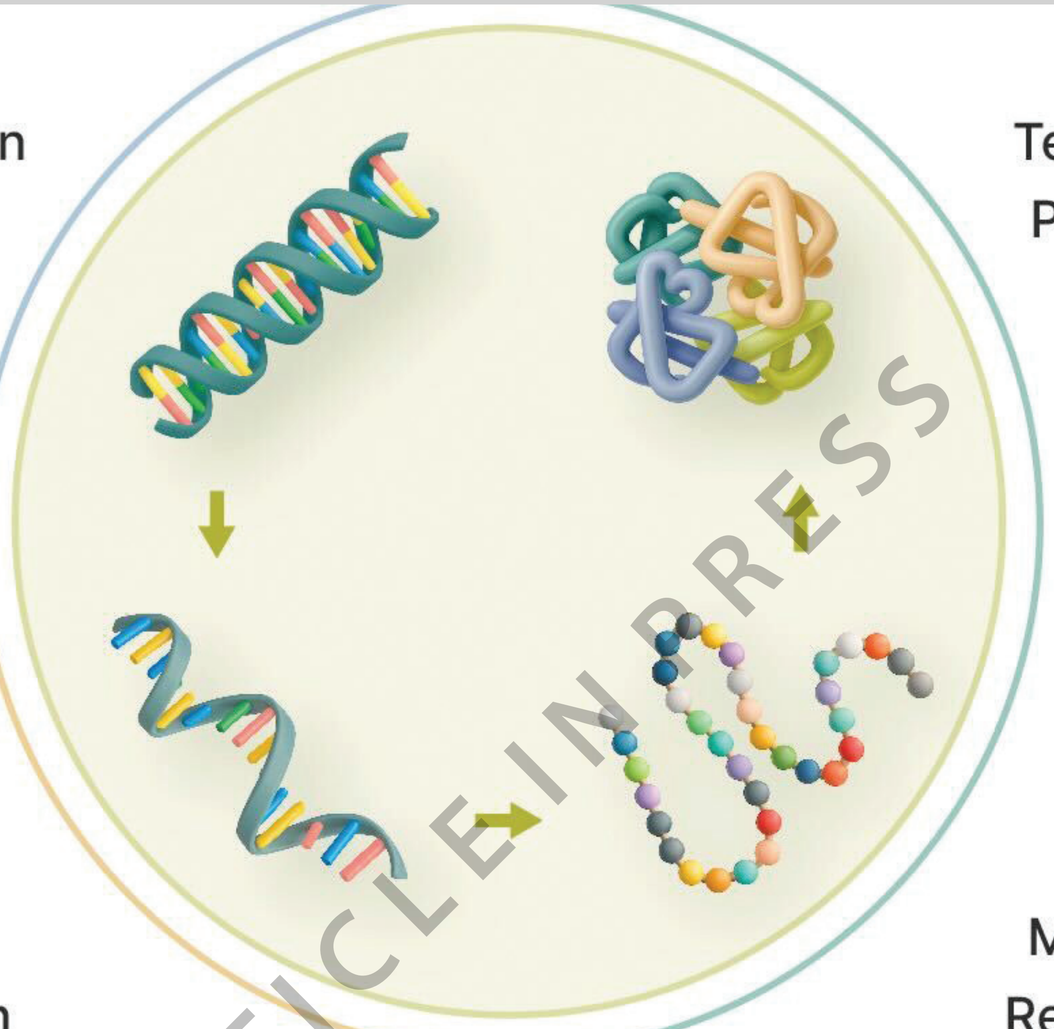
3D Structure

DNA / RNA

Regulatory Region Prediction
Mutation Effect Prediction
Binding Site Prediction
Sequence generation

Protein

Tertiary Structure Prediction
PPI / Binding Site Prediction
De Novo Protein Design
Proteome Prediction



scRNA

Batch Correction
In Silico Treatment
Cell Type Annotation
Gene Function Classification

Subcellular Localization
Conditional Generation
Mutational Effect Prediction
Remote Homology Detection

Application

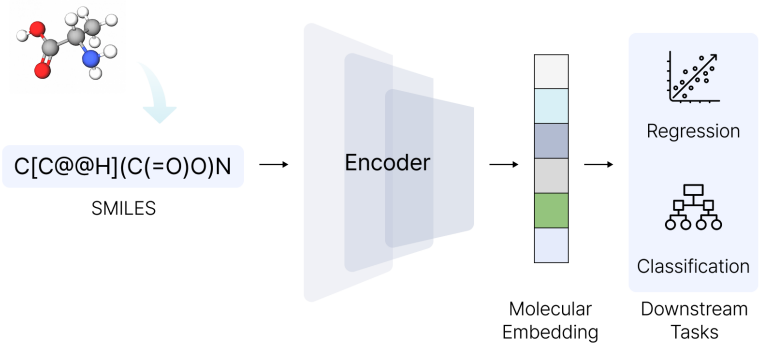
Drug Discovery

Medical QA

Research QA

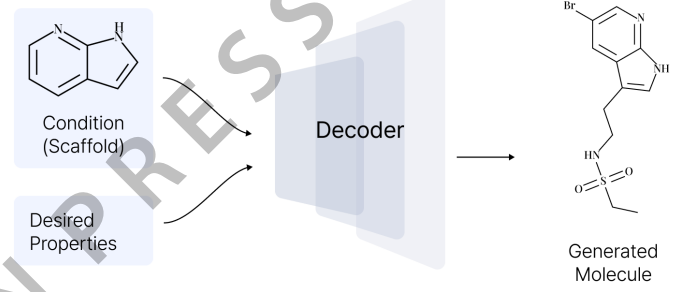
a. BERT-like (Encoder only)

- Molecular Embedding
- Property Prediction



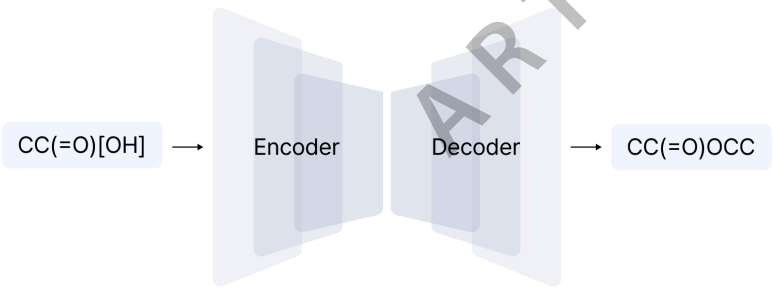
b. GPT-like (Decoder only)

- Molecule Generation
- Scaffold Completion



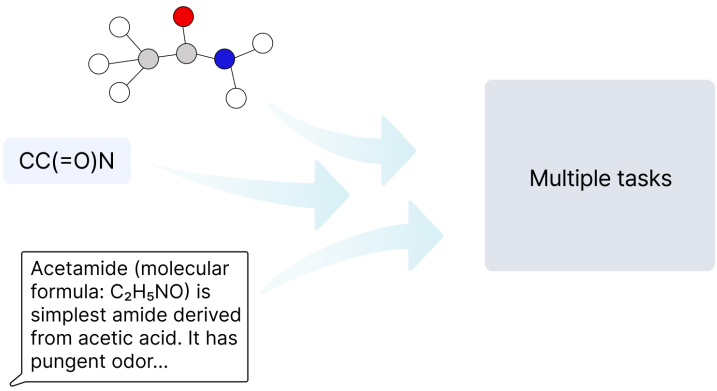
c. Encoder-Decoder (Seq-2-Seq)

- Retrosynthesis
- Reaction Prediction



d. Multi-modal

- Modality Fusion
- Cross-modal Reasoning



Experimental & Molecular Medicine

Language Models Drive Innovation in Biomedical Discovery

Large language models (LLMs) are artificitial intelligence models that understand and generate human language. Scientists want to use LLMs to better understand complex scientific data, but there are challenges because scientific data are different from human language. Researchers reviewed how LLMs are being adapted for tasks in chemistry and biology by training them using large datasets, such as protein sequences and chemical structures. The study highlights the importance of designing effective representations that LLMs can understand, which is crucial for their success in scientific applications. The results show that LLMs can predict protein structures and design new molecules, potentially revolutionizing drug discovery and related areas. The researchers conclude that while progress has been made, more work is needed to align natural-language LLMs to fully address scientific needs.

Related Article Manuscript number: EMM20251396
Article Title: A Survey on Large Language Models in Biology and Chemistry
Corresponding Author and affiliation/s: Prof. Juyong Lee

About your Research Summary — please read

This **Research Summary** is based on your manuscript that was recently accepted for publication in *Experimental & Molecular Medicine* (EMM). It provides a non-specialist audience with a synopsis of your key research outcomes and conclusions. This value-added service provided by NPG is designed to raise interest in your research across the broader community.

NPG will publish the summary on the journal's website, and it will be freely available under a Creative Commons "by-nd-nc 4.0 unported" license (see the journal website for details). We encourage you to re-use the summary to bring attention to your research; for example, you can host it on your own website and share it via social-networking platforms. Please attribute the summary to *EMM* and your article (e.g. by providing a link to your article) and do not make derivatives.

Please note that to maximise the usefulness of these summaries they must follow several stringent guidelines:

-- Spelling, punctuation and style are set according to *Nature* editorial guidelines. As this summary is aimed at non-expert readers, some concepts and technical terms will be simplified.

-- Total length must be no more than 135 words. It is likely that not all points in the paper will be

covered.

-- The first sentence must be no more than 280 characters, including spaces, to allow use on microblogging sites.

-- The headline must consist of a brief generic subject identifier followed by a short description. No more than 10 words in total.

Please contact the editorial office (ksbmb3@ksbmb.or.kr) immediately with corrections should you find any factual errors in this Research Summary.