Review

Application of LLMs/Transformer-Based Models for Metabolite Annotation in Metabolomics

Yijiang Liu [1,†], Feifan Zhang [2,†], Yifei Ge [2], Qiao Liu [3], Siyu He [4], and Xiaotao Shen [1,2,5,*]

[1] School of Chemistry, Chemical Engineering and Biotechnology, Nanyang Technological University, Singapore 637459, Singapore
[2] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore 308232, Singapore
[3] Department of Statistics, Stanford University School of Medicine, Palo Alto, CA 94304, USA
[4] Department of Biomedical Data Science, Stanford University School of Medicine, Palo Alto, CA 94304, USA
[5] Singapore Phenome Center, Nanyang Technological University, Singapore 636921, Singapore
* Correspondence: xiaotao.shen@ntu.edu.sg
† These authors contributed equally to this work.
Received: 20 December 2024; Revised: 6 January 2025; Accepted: 3 March 2025; Published: 15 April 2025

Abstract: Liquid Chromatography-Mass Spectrometry (LC-MS) untargeted metabolomics has become a cornerstone of modern biomedical research, enabling the analysis of complex metabolite profiles in biological systems. However, metabolite annotation, a key step in LC-MS untargeted metabolomics, remains a major challenge due to the limited coverage of existing reference libraries and the vast diversity of natural metabolites. Recent advancements in large language models (LLMs) powered by Transformer architecture have shown significant promise in addressing challenges in data-intensive fields, including metabolomics. LLMs, which when fine-tuned with domain-specific datasets such as mass spectrometry (MS) spectra and chemical property databases, together with other Transformer-based models, excel at capturing complex relationships and processing large-scale data and significantly enhance metabolite annotation. Various metabolomics tasks include retention time prediction, chemical property prediction, and theoretical MS^2 spectra generation. For example, methods such as LipiDetective and MS2Mol have shown the potential of machine learning in lipid species prediction and de novo molecular structure annotation directly from MS^2 spectra. These tools leverage transformer principles and their integration with LLM frameworks could further expand their utility in metabolomics. Moreover, the ability of LLMs to integrate multi-modal datasets—spanning genomics, transcriptomics, and metabolomics—positions them as powerful tools for systems-level biological analysis. This review highlights the application and future perspectives of Transformer-based LLMs for metabolite annotation of LC-MS metabolomics incorporating with multiomics. Such transformative potential paves the way for enhanced annotation accuracy, expanded metabolite coverage, and deeper insights into metabolic processes, ultimately driving advancements in precision medicine and systems biology.

Keywords: large language models; LC-MS; metabolite annotation

1. Introduction

Metabolomics is the comprehensive analysis of metabolites within cells, biological fluids, tissues, and organisms, emphasizing their composition, dynamic variations, and interactions [1]. Key analytical techniques in metabolomics include Mass Spectrometry (MS) and Nuclear Magnetic Resonance (NMR) spectroscopy [2,3]. Among these, Liquid Chromatography-Mass Spectrometry (LC-MS) has emerged as a cornerstone technology in metabolomics research [4], which is widely applied in diverse areas, including disease biomarker discovery, drug development, toxicity evaluation, and integrative multi-omics analyses [5,6].

The standard workflow of LC-MS metabolomics typically comprises sample collection and preparation, data acquisition and processing, data cleaning, metabolite annotation, and biological interpretation [7]. Recent technological advancements have made these workflows increasingly sophisticated and efficient. However, metabolite annotation remains a significant challenge in LC-MS untargeted metabolomics [8]. Despite the development of various tools and methods like GNPS [9], SIRIUS [10], MetFrag [11], MetDNA [12], metID [13], and massdatabase [14], metabolite annotation predominantly relies on comparison with reference databases. This

reliance on publicly available databases is a major limitation, as these databases cover only a fraction of the vast diversity of natural metabolites [15]. At the same time, improvements in MS have enhanced sensitivity and increased the detection of metabolic features. However, only a small proportion of these features can be reliably annotated [16]. Given that metabolite annotation is a critical step for translating chemical data into meaningful biological insights, the development of innovative methods to expand the scope of metabolite annotation continues to be a focal area of research.

The recent success of large language models (LLMs) has garnered significant public attention [17,18]. These models, primarily built on Transformer architecture, leverage the self-attention mechanism to capture relationships across different positions in a sequence [19]. Compared to traditional machine learning models, LLMs exhibit superior generalization capabilities and have demonstrated exceptional performance across various tasks, including natural language processing and representation learning [20]. Additionally, the Transformer architecture enables these models to model complex relationships and process datasets with high-dimensional and non-linear features, such as chemical structures and spectral data, making them particularly suitable for applications in biomedical research [21]. When fine-tuned with domain-specific representations like SMILES strings, molecular graphs, or curated MS spectra, LLMs hold promise for addressing key challenges in analyzing LC-MS metabolomics [22], including retention time prediction, molecular property inference, theoretical MS^2 spectra generation, and metabolite annotation workflows. This review provides an overview of LLMs, explores their current applications in LC-MS untargeted metabolomics, and discusses future perspectives.

2. Overview of Large Language Models

Large language models (LLMs) represent a significant breakthrough in artificial intelligence (AI), transforming how complex data is processed and interpreted. The Transformer architecture, an effective deep learning model widely adopted across various research areas, serves as the fundamental backbone of LLMs [19,23,24]. At its core, the Transformer's self-attention mechanism functions like a dynamic information filtering system, enabling the model to identify and weigh relationships between different elements across vast amounts of sequential data. Through this architectural innovation, LLMs can understand and generate human-like language with remarkable precision [25]. These models are pre-trained on massive datasets, often spanning from gigabytes to terabytes in size, consisting of billions to hundreds of billions in parameters, to learn statistical relationships within text. This pretraining allows LLMs to perform a wide variety of tasks, including text summarization, code generation, and knowledge retrieval [26,27].

Traditional deep learning models, including recurrent neural networks (RNNs) [28], long short-term memory networks (LSTMs) [29], gated recurrent units (GRUs) [30], convolutional neural networks (CNNs) [31], and sequence-to-sequence (Seq2Seq) [32] models, have been widely used in metabolomic studies [33–36]. However, these models primarily rely on supervised learning tailored to specific tasks [37,38]. Although they are effective for well-defined applications, they often struggle to generalize and perform poorly when applied to different tasks [26,27]. In contrast, LLMs leverage a massive number of parameters and extensive pretraining on vast datasets, providing significant advantages over traditional deep learning models. Furthermore, self-supervised pre-training, which automatically derives supervision signals from unlabeled data (e.g., by predicting masked tokens in a sentence), enables these models to learn from large-scale unlabeled text corpora, thereby reducing the need for manual data labeling compared to conventional methods.

One of the most transformative features of LLMs is their ability to solve a wide variety of tasks through prompting alone, often without requiring parameter fine-tuning. However, fine-tuning or instruction tuning remains beneficial for specialized applications.

LLMs can generally be classified into two categories: open-source and closed-source models (Figure 1a). Examples of open-source models include Grok-1 (https://github.com/xai-org/grok-1, accessed on 30 November 2024), Mistral (https://mistral.ai/news/announcing-mistral-7b/, accessed on 30 November 2024), LLaMA 3 (https://github.com/meta-llama/llama3, accessed on 30 November 2024), and Qwen 2.5 (https://github.com/QwenLM/Qwen2.5, accessed on 30 November 2024), while closed-source models include Gemini 1.5 (https://deepmind.google/technologies/gemini/, accessed on 30 November 2024), Grok-2 (https://x.ai/blog/grok-2, accessed on 30 November 2024), OpenAI o1 (https://openai.com/o1/, accessed on 30 November 2024), and Claude 3.5 (https://www.anthropic.com/news/claude-3-5-sonnet, accessed on 30 November 2024). Open-source models typically share their weights and often portions of their training code, allowing researchers and developers to freely download, fine-tune, and deploy them locally. This makes them particularly suitable for applications with strict data security requirements or specialized domain needs. In contrast, closed-source models maintain proprietary control over their core technologies, generally command higher commercial

value, and often benefit from more substantial training investments. Consequently, they tend to demonstrate superior performance compared to open-source models.

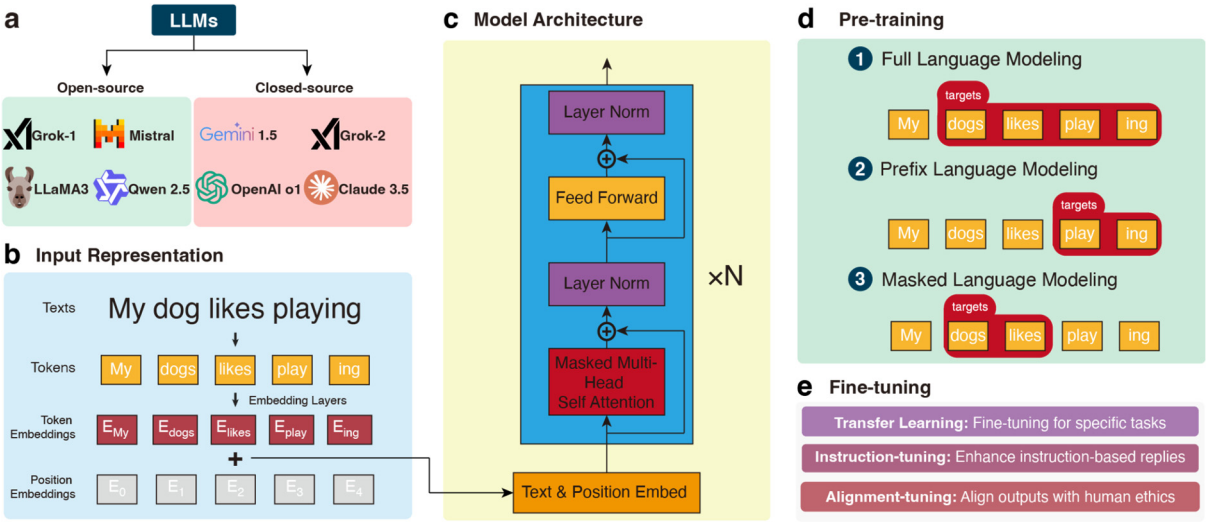

Figure 1. Overview of Large Language Models (LLMs) and their core principles. (**a**) Commonly used LLMs are categorized as open-source (e.g., Grok-1, Mistral, LLaMA3, Qwen 2.5) and closed-source (e.g., Gemini 1.5, OpenAI GPT models, Claude 3.5). (**b**) Input representation for LLMs: Text is tokenized into individual tokens ("My," "dogs," "likes," "play," "ing"), which are then transformed into token embeddings and combined with position embeddings to represent input sequences. (**c**) The model architecture of LLMs: The architecture includes multiple layers (N layers) consisting of Masked Multi-Head Self-Attention, Feed Forward networks (simple fully connected layers), and Layer Normalization (which normalizes the outputs within each layer), with text and positional embeddings as input. (**d**) Pre-training strategies for LLMs. (**e**) Fine-tuning of LLMs: Techniques include Transfer Learning, Instruction-tuning, and Alignment-tuning.

3. Main Principles of LLMs

This section introduces the fundamental principles underlying LLMs.

3.1. Input Representation

Input representation involves transforming natural language inputs into numerical formats that deep learning models can process, commonly referred to as encoding (Figure 1b). Tokenization is the first step, breaking input text into units like characters, subwords, symbols, or words. Each token is then mapped into a fixed-dimensional vector through an embedding layer, optimized during training to capture linguistic and contextual information. Positional embeddings are added to these vectors to encode token order, addressing the position-agnostic nature of the Transformer's attention mechanism [19].

3.2. Model Architecture

The Transformer [19] is the foundational architecture for most modern LLMs [23,39,40]. Introduced in 2017 for machine translation, the Transformer architecture has since become widely adpoted in fields such as computer vision, audio processing, and robotics due to its outstanding performance [41–43], which played a pivotal role in the development of LLMs [26,27].

The success of Transformer is largely attributed to its attention mechanism, which processes sequences in parallel and captures long-range contextual relationships [19]. In this mechanism, the model maps inputs into query, key, and value representations. The attention score is calculated as the scaled dot product of the query and key, normalized through a softmax function, and then used to weigh the value representations. This approach allows the model to focus on the most relevant parts of the input sequence. Typically, Transformers employ a multi-head attention mechanism, where multiple attention heads operate in parallel. This design enables the model to attend to information from different representation subspaces simultaneously (Figure 1c). Unlike traditional language models such as RNNs, which process inputs sequentially, the Transformer processes entire sequences in parallel, making it highly efficient for large-scale data. This parallel processing capability enables LLMs to handle long sequences and large datasets, fostering their emergence.

3.3. Pre-Training

During pre-training, LLMs utilize self-supervised learning to analyze massive unlabeled text corpora, eliminating the need for manually annotated data [44]. Pre-training objectives vary depending on the LLM architecture (Figure 1d):

(1) **Full language modeling:** The model predicts future tokens based on preceding tokens, as seen in autoregressive models like GPT [45].
(2) **Prefix language modeling:** A random prefix is chosen, and the subsequent tokens are used to calculate the loss.
(3) **Masked language modeling:** Tokens are randomly masked in the input, and the model predicts these masked tokens using context from both preceding and following tokens, as in bidirectional models like BERT [23].

These diverse training objectives enable LLMs to learn general language patterns and relationships, forming the foundation for downstream applications.

3.4. Fine-Tuning

Although pre-trained LLMs demonstrate strong generalization capabilities across various tasks, fine-tuning can further optimize them for specific downstream applications [45] (Figure 1e). Key fine-tuning approaches include:

(1) **Transfer learning:** Fine-tuning the pre-trained model on datasets specific to a target task [40,46], enhancing performance for domain-specific applications.
(2) **Instruction-tuning:** The model is fine-tuned on datasets formatted as natural language instructions to enhance the model's ability to generate accurate and context-appropriate responses. These datasets typically span multiple tasks to improve generalization [47].
(3) **Alignment-tuning:** To mitigate issues such as the generation of biased, false, or harmful content, LLMs undergo alignment-tuning, where feedback is used to adjust the models in accordance with human preferences. This ensures models are better calibrated to avoid producing inappropriate or unexpected outputs [48].

These fine-tuning strategies have significantly enhanced the utility of LLMs across diverse applications, improving their adaptability and robustness in various fields [45].

4. The Applications of LLMs in Biomedical Research

The applications of LLMs in biomedical research have grown rapidly, currently primarily across five domains: bioinformatics analysis, AI-powered intelligent agents, text mining and knowledge extraction, molecular design and property prediction, and multi-omics data integration. These advancements transform how complex biomedical challenges are addressed, offering enhanced precision, scalability, and interactivity.

4.1. Bioinformatics Analysis Assistants

LLMs have demonstrated immense potential in supporting a variety of bioinformatics tasks. By leveraging their ability to process and analyze high-dimensional data, LLMs contribute to several key areas:

(1) **Protein structure analysis:** LLMs have been employed to predict and understand protein structures and interactions, significantly advancing efforts to understand molecular biology and disease mechanisms. Tools like ProteinGPT have set new benchmarks for accuracy in this field [49,50].
(2) **Sequence analysis:** LLMs facilitate the analysis of nucleotide and amino acid sequences, enabling efficient pattern recognition, motif detection, and annotation of genomic data [51,52].
(3) **Gene expression and regulation analysis:** By interpreting complex functional genomic and epigenomic datasets, LLMs assist in uncovering gene regulation patterns, identifying biomarkers, and elucidating the molecular basis of diseases [53–55].
(4) **Drug discovery:** The integration of LLMs into drug discovery pipelines accelerates processes such as virtual screening, compound optimization, and predicting drug-target interactions. These models also support the generation of novel molecular structures with desired pharmacological properties [56–58].

In these applications, LLMs excel due to their ability to analyze large datasets, extract meaningful patterns, and provide interpretable outputs that aid decision-making.

4.2. AI Agents in Biomedical Research

LLM-powered AI agents represent a more advanced application of these models, enabling dynamic problem-solving and operational assistance in biomedical research. These agents address complex challenges by leveraging advanced capabilities, including:

(1) **Database retrieval and integration:** AI agents can retrieve, integrate, and analyze information from diverse biomedical databases, such as PubMed (https://pubmed.ncbi.nlm.nih.gov/, accessed on 30 November 2024), UniProt (https://www.uniprot.org/, accessed on 30 November 2024), and KEGG (https://www.genome.jp/kegg/, accessed on 30 November 2024). They streamline data queries, generate comprehensive summaries, and facilitate cross-database analyses [59–61].
(2) **Experimental platform management:** These agents manage experimental workflows by coordinating software tools and laboratory equipment through APIs, and automating processes such as data acquisition, analysis, and reporting [62,63].
(3) **Conversational systems with reflective learning:** Equipped with natural language interfaces, AI agents interact seamlessly with researchers, providing explanations, answering questions, and adapting their knowledge base through reflective learning [64].
(4) **Reasoning and decision-making:** LLMs enable these agents to perform reasoning tasks, such as forming hypotheses, prioritizing experiments, or evaluating alternative strategies [65,66].

By combining these capabilities, LLM-based AI agents act as versatile collaborators, empowering researchers to focus on higher-level problem-solving while automating routine or complex tasks.

4.3. Text Mining and Knowledge Extraction

LLMs are widely used in biomedical text mining to extract meaningful insights from large corpora of scientific literature:

(1) **Named entity recognition:** These models help identify and categorize biomedical entities such as genes, proteins, diseases, and drugs within text, facilitating downstream analysis [67,68].
(2) **Drug repurposing insights:** By analyzing relationships and co-occurrence patterns in the literature, LLMs contribute to drug repurposing efforts by identifying potential therapeutic uses for existing compounds [69].

4.4. Molecular Design and Property Prediction

LLMs are widely used in advancing computational chemistry and molecular design:

(1) **SMILES generation:** LLMs generate novel SMILES strings for molecules, supporting the discovery of compounds with desired properties [70].
(2) **Property prediction:** These models predict molecular properties such as solubility, toxicity, and binding affinity, enabling efficient prioritization of candidate compounds [71].
(3) **Reaction prediction:** LLMs assist in predicting the outcomes of chemical reactions, reducing experimental trial-and-error efforts [72].

4.5. Multi-Omics Data Integration

LLMs enhance the integration of diverse omics datasets, enabling systems-level insights:

(1) **Cross-Modal analysis:** These models facilitate the integration of genomics, transcriptomics, proteomics, and metabolomics data, identifying correlations and causative relationships across modalities [73].
(2) **Biomarker discovery:** Through multi-omics integration, LLMs contribute to identifying biomarkers for disease diagnosis, prognosis, and therapeutic response [74].

5. Applications of LLMs for Metabolite Annotation in LC-MS Metabolomics Data

Compared to other omics fields such as genomics, epigenomics, transcriptomics, and proteomics, the application of LLMs in metabolomics remains relatively limited and less explored. However, an increasing number of studies are now leveraging LLMs and other Transformer-based models for LC-MS data processing and analysis, with a particular focus on metabolite annotation [75–77]. Different from traditional machine learning models and other machine learning methods, LLMs and other Transformer-based models have illustrated numerous advances in the field of metabolite annotaion. For example, LLMs and other Transformer-based models have better generalizability than traditional machine learning models and have a good performance for untargeted metabolomics [20,77]. Also, for LLMs that are fine-tuned with domain-specific datasets and Transformer-based

models that are specially designed for certain tasks, they have been found to have better prediction performances than other methods on multiple applications such as retention time prediction [78], MS^2 spectra prediction [79] and lipid species annotation [76]. These emerging approaches can be broadly categorized into three distinct methods (Figure 2a), as outlined below.

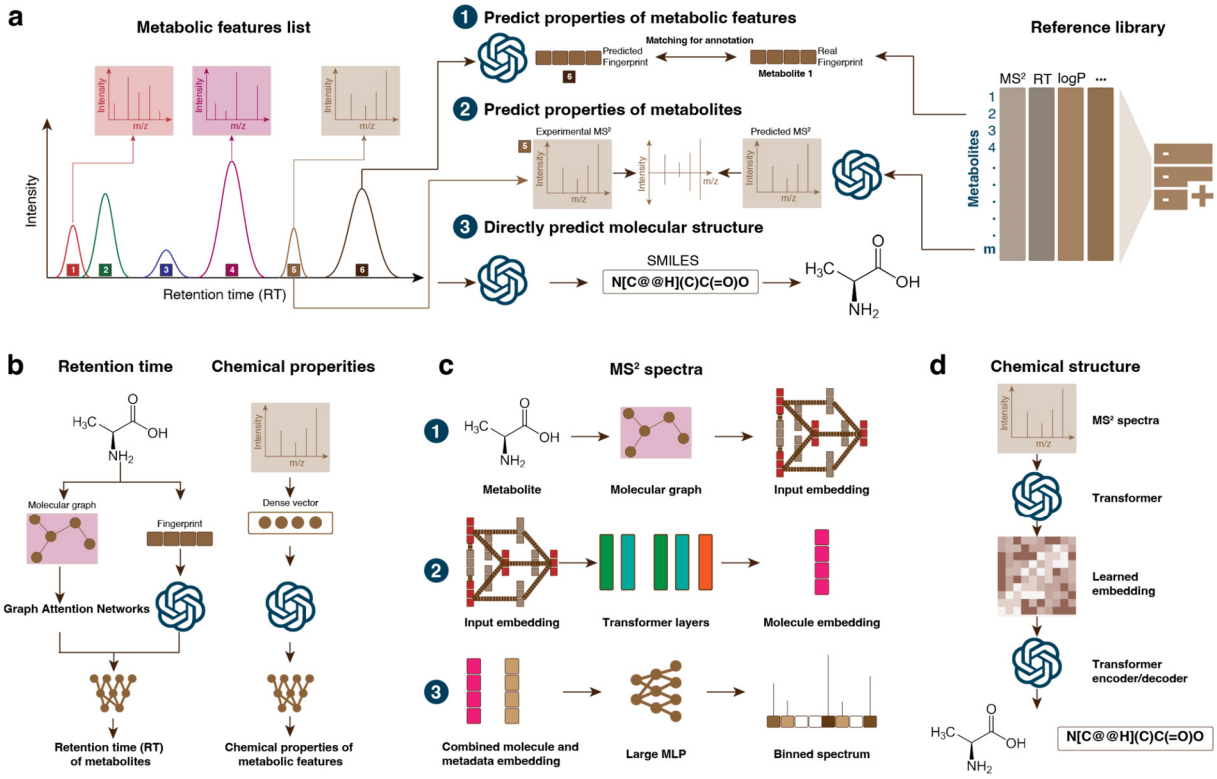

Figure 2. Applications of LLMs for metabolite annotation in LC-MS metabolomics data. (**a**) Overview of three major classes of LLM-based methods for metabolite annotation. (**b**) Retention Time and chemical property prediction: The RT-Transformer predicts retention time (**left panel**) and chemical properties (**right panel**) using MS^2 spectra. (**c**) MS^2 spectral prediction: MassFormer employs a graph Transformer architecture to predict MS^2 spectra by learning relationships between molecular fragments and experimental parameters. (**d**) Molecular structure prediction: MS2Mol performs de novo molecular structure prediction directly from MS^2 spectra, integrating spectral features and fragmentation trees to infer potential molecular structures.

5.1. Retention Time Prediction

Retention time (RT) is a crucial parameter in LC-MS metabolomics, representing the time each compound spends within the chromatographic column during analysis [80]. RT is influenced by various molecular characteristics, including size, polarity, and functional groups, providing an essential basis for the separation and identification of metabolites [81]. However, experimental determination of RT is often costly and time-consuming, highlighting the need for innovative computational methods to improve prediction efficiency and accuracy [78].

Significant progress has been made in RT prediction with the development of computational models. Among these, graph-based neural networks such as GNN-RT [82] and MPNN [83], deep neural networks like DNNpwa [84], and convolutional neural networks such as 1D-CNN [85] have shown promising results. Despite these advancements, the potential of Transformer architectures in RT prediction remains underexplored.

Recently, Xue et al. introduced a novel model, RT-Transformer, which integrates 1D-Transformer and graph attention networks (GAT) to improve RT prediction [78] (Figure 2b). Unlike traditional methods that primarily rely on molecular fingerprints or descriptors, RT-Transformer incorporates detailed molecular structure information, including node and edge attributes from molecular graphs [78,80]. This model uses molecular graphs alongside molecular fingerprints as inputs. The 1D-Transformer generates high-dimensional feature vectors from the fingerprints, which are then combined with feature vectors derived from molecular graphs to predict RT [78]. To evaluate its performance and transfer learning capabilities, RT-Transformer was tested on the SMRT dataset and 41 independent datasets. Results demonstrated superior performance compared to other models like GNN-RT [82] and MPNN [83] across several key metrics, including mean absolute error (MAE), mean relative error (MRE) and

coefficient of determination (R^2) [78]. These findings highlight the potential of Transformer-based architectures for advancing RT prediction in LC-MS metabolomics.

5.2. Chemical Properties Prediction

Chemical properties represent the intrinsic characteristics of a metabolite that dictate its behavior in chemical reactions and interactions, such as chemical stability, solubility, and reactivity [86]. In the context of metabolite annotation, these properties hold significant potential for a variety of applications [87]. For instance, after predicting the chemical properties of metabolic features, these predictions can be matched against the chemical properties of standard metabolites in a reference library, effectively narrowing down the pool of candidate metabolites. Additionally, predicted properties can be integrated with other annotation methods to enable multi-dimensional validation, thereby reducing false positives and improving annotation accuracy [88].

MS2Prop is based on Transformer architecture, which can predict chemical properties of metabolic features directly from their tandem mass spectra (MS^2) spectra [89] (Figure 2b). Thanks to the capabilities of the Transformer architecture, MS2Prop is well-suited for processing disordered MS^2 spectra. Briefly, the MS^2 spectra are first embedded into high-dimensional feature vectors using the Transformer module and then passed through a feedforward neural network to predict chemical properties [89]. A notable advantage of MS2Prop is its exceptional efficiency, generating predictions for an MS^2 spectrum in under 2 milliseconds. This rapid processing capability enables the analysis of large-scale metabolomics datasets in a fraction of the time typically required. Moreover, MS2Prop exhibits robust performance and generalization ability. When applied to novel compounds, the model achieved an average R^2 of 70% across 10 chemical properties, including logP, synthetic accessibility, and polar surface area. These results highlight MS2Prop's potential as a powerful and reliable tool for chemical property prediction and metabolite annotation.

5.3. MS^2 Spectra Prediction

Tandem mass spectra (MS^2 or MS/MS spectra) provide crucial structural information about metabolites and play a pivotal role in metabolite annotation [90]. However, acquiring MS^2 spectra for all standard metabolites is highly challenging due to the time, cost, and practical difficulties in obtaining pure standards for every possible metabolite [14]. This limitation leads to gaps in reference libraries, making theoretical MS^2 spectra prediction an essential solution to address the scarcity of MS^2 spectra for metabolite annotation [91].

Recently, several methods leveraging advanced machine learning models, particularly those based on LLMs and transformer architectures, have been developed for MS^2 spectra prediction [92,93]. These methods aim to efficiently generate accurate theoretical MS^2 spectra for a wide range of metabolites, significantly expanding the coverage of reference libraries.

MassFormer is one such method that utilizes a graph Transformer to predict the MS^2 spectra of metabolites [93] (Figure 2c). This model employs the Transformer attention mechanism to capture global relationships between nodes (atoms) and edges (chemical bonds) in the molecular graph, which serves as the input representation of the metabolite. In addition to structural features, the input includes experimental parameters such as collision energy and precursor adducts, which are critical for accurately modeling MS^2 spectra prediction. The graph Transformer extracts global molecular features, combines them with experimental metadata, and uses multi-layer perceptrons (MLPs) to predict the MS^2 spectra. Comparisons with traditional methods on datasets such as the MassBank of North America (MoNA) and the National Institute of Standards and Technology Database (NIST) showed that MassFormer outperformed traditional approaches like competitive fragmentation modelling (CFM) [94,95], fingerprint neural network model (FP) [96] and Weisfeiler–Lehman Network (WLN) [97], achieving significantly higher scores in cosine similarity and top-5 accuracy metrics [93].

These LLM-based models represent a significant step forward in MS^2 spectra prediction, enabling researchers to expand reference libraries with high-quality theoretical MS^2 spectra [92,93]. By leveraging these advanced computational approaches, the metabolomics community can overcome the limitations of incomplete MS^2 spectral libraries and achieve more comprehensive and accurate metabolite annotation.

5.4. Lipid Species Annotation

Lipids are a diverse group of organic compounds that are widely distributed across animals, plants, and microorganisms [98]. They play essential roles in biological processes, including energy storage, cell membrane structure, and signaling pathways [99]. Unlike other small metabolites, lipids have relatively standardized structures, allowing them to be classified into various categories, such as fatty acids, glycerides, and

glycerophospholipids [100]. Lipids within the same category often share similar structures and functions, making their accurate annotation crucial for addressing biological research questions [101].

To address the challenges in lipid species annotation, LipiDetective was developed as a Transformer-based model designed for lipid species prediction [76]. The model was trained on a dataset comprising MS^2 spectra from eight lipid classes, allowing it to learn the characteristic fragmentation patterns of lipids [76]. Using this knowledge, LipiDetective can predict the identities of novel lipid species based on their MS^2 spectra [76]. A major advantage of LipiDetective over traditional methods lies in its ability to generalize, making it capable of identifying unknown lipid species with higher accuracy. This generalization is particularly important given the structural diversity of lipids and the limitations of existing reference libraries. As the first model to directly utilize deep learning technology for lipid species prediction from MS^2 spectra, LipiDetective represents a significant milestone in lipidomics.

Despite its groundbreaking potential, LipiDetective's performance still has room for improvement, particularly in fine-tuning its prediction accuracy across diverse lipid classes. Nonetheless, its introduction highlights the transformative power of Transformer models in lipidomics, paving the way for future innovations in this field. As researchers continue to optimize these models and expand their training datasets, the potential for highly accurate, high-throughput lipid annotation will be further realized, enhancing our understanding of lipid biology and its implications for human health.

5.5. Fingerprints Prediction

A molecular fingerprint is an abstract representation of a metabolite that encodes its chemical structure into a fixed-length, sparse binary vector [102]. Each bit in the vector typically represents the presence or absence of specific chemical substructures, such as hydroxyl groups or aromatic rings [103]. Molecular fingerprints capture the existence characteristics of a molecule's chemical structure, enabling efficient screening of potential compound candidates by comparing them to the molecular fingerprints of metabolites in a reference library [104].

Some studies have attempted to predict the metabolites' molecular fingerprints directly from MS^2 spectra based on the Transformer model [75,105]. Baygi et al. introduced IDSL_MINT, a model that utilizes Transformer architecture to convert MS^2 spectra into molecular fingerprints [75]. By employing position encoding and attention mechanisms, IDSL_MINT effectively captures the characteristics of spectral segments and achieves high accuracy in both positive and negative ionization modes.

Similarly, Goldman et al. developed MIST (Mass Spectrum to Fingerprint Transformer) to predict molecular fingerprints from MS^2 spectra and then annotate metabolites [105]. The key innovation in MIST is the representation of fragments as chemical formula vectors, which are then processed using a set of transformers to learn the relationships between peaks. This allows MIST to generate high-resolution molecular fingerprints as its output. The model demonstrated superior performance in fingerprint prediction tasks, outperforming existing methods on over 70% of test datasets, according to the authors' evaluation [105].

These advancements highlight the potential of Transformer-based models in molecular fingerprint prediction, paving the way for more accurate and high-throughput metabolite annotation workflows in LC-MS untargeted metabolomics.

5.6. Molecular Structure Prediction for Direct Metabolite Annotation

Unlike the above prediction methods that estimate one or more properties of metabolites and annotate them by matching experimental features to a reference library, some approaches aim to predict the molecular structure directly from MS^2 spectra [106,107]. These end-to-end methods bypass the need for intermediate metrics or database matching, offering a direct route to metabolite annotation.

Shrivastava et al. introduced MassGenie, a model combining a Transformer network with a variational autoencoder (VAE) to generate potential molecular structures from MS^2 spectra [108]. MassGenie features a Transformer architecture comprising a 12-layer encoder and a 12-layer decoder with approximately 400 million parameters. This design enables the model to learn molecular fragments and structural characteristics from MS^2 spectra effectively. While its application is limited to small metabolites with molecular weights below 500 Da and depends heavily on high-quality MS data, MassGenie demonstrates the potential of Transformer-based architectures for annotating unknown metabolites [108].

Similarly, Mass2SMILES leverages a hybrid Transformer and time-convolutional network (TCN) architecture to annotate metabolites directly from high-resolution MS^2 spectra [109]. This model successfully predicted seven completely correct structures out of 744 validation spectra, showcasing its utility in molecular structure prediction.

Another noteworthy method is MS2Mol, which represents a de novo approach to structure prediction [77] (Figure 2d). MS2Mol directly predicts molecular structures from MS^2 spectra without relying on reference compound databases.

A complementary approach incorporates fragmentation trees, a graph-based representation of molecular fragmentation [110]. In a fragmentation tree, nodes represent fragmentations in the MS^2 spectra, and edges represent fragmentation reaction relationships between these fragmentations. This structure provides insight into both global and local dependencies, aiding in the inference of functional groups and molecular substructures. Building on this concept, Zhang et al. developed MS2-Transformer, a Transformer-based model that integrates fragmentation trees derived from spectral data as part of its input [111]. By incorporating information about chemical bonds, this approach significantly improved model performance. Meanwhile, Yang et al. introduced TeFT, a lightweight Transformer model designed to generate SMILES representations of molecular structures [112]. TeFT integrates fragmentation tree information into its predictions, comparing the generated SMILES with the fragmentation tree to refine the inferred molecular structure and generate candidate metabolites.

In summary, these Transformer-based approaches highlight the growing capability of modern LLMs to directly predict molecular structures from MS^2 spectra, reducing reliance on reference libraries and traditional tools [77,108]. As these models continue to evolve, they promise to improve the efficiency, accuracy, and coverage of metabolite annotation in untargeted metabolomics.

6. Future Perspectives

The annotation of single metabolites provides essential insights into their specific structures and properties [15]. However, the vast number of metabolites and the intricate interactions within biological metabolic processes necessitate a broader approach that explores the relationships and correlations between metabolites [113]. One promising strategy is the construction of metabolite networks, which use graph theory to represent metabolites as nodes and define edges based on biochemical reactions, structural similarities, co-occurrence relationships, or functional connections [114]. LLMs can capture complex dependencies and relationships in the networks by integrating with knowledge graphs and using algorithms to generate subgraphs from them, making them very suitable for analyzing these networks [115,116]. These models hold the potential to uncover novel metabolic reactions, metabolites, and pathways/modules, providing a more holistic understanding of metabolic processes (Figure 3a).

A significant bottleneck in metabolomics research is the limited availability of MS^2 spectra for metabolites in reference libraries [117]. While LLMs have shown great success in predicting accurate MS^2 spectra or directly inferring molecular structures from MS^2 spectra, many metabolic features lack corresponding MS^2 spectra [109,118]. This limitation prevents these models from annotating such metabolic features using current approaches [22]. To overcome this challenge, innovative computational methods are needed. By integrating metabolic features (including *m*/*z*, RT, adducts, and isotope patterns) with MS^2 spectral data and embedding using GPT, we can annotate metabolic features using multi-modal and multi-dimensional information rather than relying solely on MS^2 spectra. Their strengths in text learning, logical reasoning, and content generation make them ideal for this task, offering a pathway to extend annotation coverage significantly (Figure 3b).

Simultaneously, multi-omics data integration is becoming a cornerstone of biomedical research [119]. Combining metabolomics with transcriptomics, proteomics, and microbiomes enables a more comprehensive understanding of biological systems [120,121]. Transformer-based models, with their ability to process heterogeneous data types and extract interrelated information, are particularly well-suited for integrating multi-omics datasets. By formatting multi-omics data into structured sentences, LLMs can comprehend complex biological mechanisms and provide deeper insights into system-wide biological processes through training or fine-tuning on domain-specific datasets [73]. This integration could reveal complex interactions among microbiotas, proteins, and metabolites, paving the way for breakthroughs in disease prediction, biomarker discovery, and therapeutic target identification (Figure 3c).

In summary, the incorporation of advanced Transformer-based models into metabolomics research holds immense potential. These models are poised to revolutionize the field by unlocking deeper insights into metabolic networks, improving annotation accuracy, and facilitating systems-level understanding in both basic and applied scientific research.

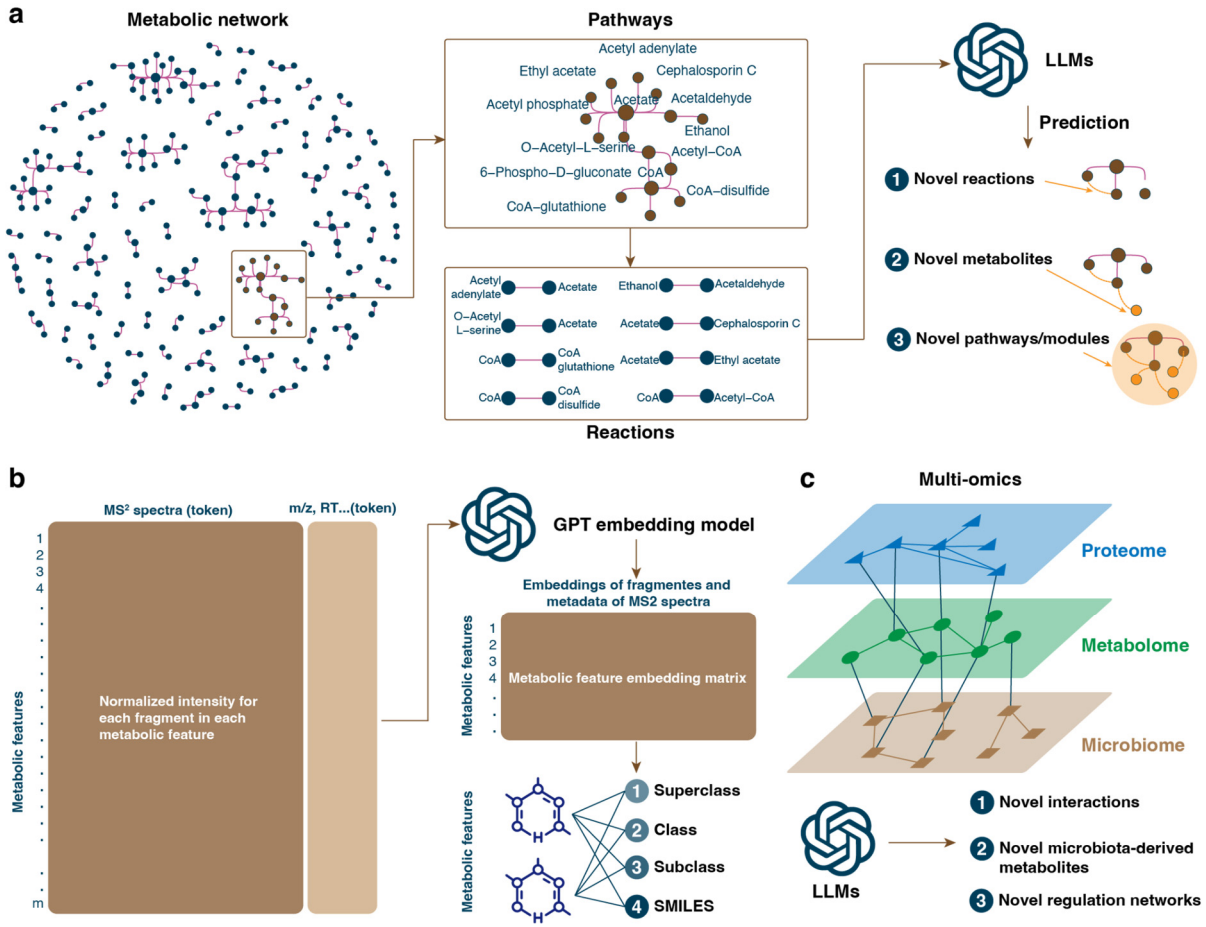

Figure 3. Future perspectives. (**a**) LLMs can utilize known reaction networks between metabolites to predict novel biochemical reactions, previously unidentified metabolites formed through these reactions, and potential novel metabolic pathways or functional modules. (**b**) LLMs can integrate multiple types of input information, such as MS^2 spectra, RT, *m/z*, and other metadata, to predict metabolites in an end-to-end manner. These predictions can be stratified into different levels of detail, including superclass, class, subclass, and molecular structures (e.g., SMILES or fingerprints), with varying levels of confidence. (**c**) LLMs can incorporate multi-omics information, such as proteomics, metabolomics, and microbiomics, to uncover new interactions between different omics layers. This includes identifying novel microbiota-derived metabolites and discovering new regulatory networks that connect genes, proteins, and metabolites.

Author Contributions: Concept development: X.S. and Y.L.; Figures: X.S., Y.L. and F.Z.; Writing and Editing: Y.L., F.Z. and X.S. with input from all the other authors. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the X.S.'s Start-Up Package from Lee Kong Qian School of Medicine (LKCMedicine) and the School of Chemistry, Chemical Engineering and Biotechnology (CCEB), Nanyang Technological University (NTU), Singapore.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Klassen, A.; Faccio, A.T.; Canuto GA, B.; da Cruz PL, R.; Ribeiro, H.C.; Tavares MF, M.; Sussulini, A. Metabolomics: Definitions and Significance in Systems Biology. In *Metabolomics: From Fundamentals to Clinical Applications*; Sussulini, A., Ed.; Springer International Publishing: Cham, Switzerland, 2017; pp. 3–17. https://doi.org/10.1007/978-3-319-47656-8_1.
2. Pan, Z.; Raftery, D. Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics. *Anal. Bioanal. Chem.* **2007**, *387*, 525–527.
3. Emwas, A.-H.M. The Strengths and Weaknesses of NMR Spectroscopy and Mass Spectrometry with Particular Focus

on Metabolomics Research. In *Metabonomics: Methods and Protocols*; Bjerrum, J.T., Ed.; Springer: New York, NY, USA, 2015; pp. 161–193. https://doi.org/10.1007/978-1-4939-2377-9_13.

4. Glish, G.L.; Vachet, R.W. The basics of mass spectrometry in the twenty-first century. *Nat. Rev. Drug Discov.* **2003**, *2*, 140–150.
5. Son, A.; Kim, W.; Park, J.; Park, Y.; Lee, W.; Lee, S.; Kim, H. Mass Spectrometry Advancements and Applications for Biomarker Discovery, Diagnostic Innovations, and Personalized Medicine. *Int. J. Mol. Sci.* **2024**, *25*, 9880.
6. Qin, X.; Hakenjos, J.M.; Li, F. LC-MS-Based Metabolomics in the Identification of Biomarkers Pertaining to Drug Toxicity: A New Narrative. In *Biomarkers in Toxicology*; Patel, V.B., Preedy, V.R., Rajendram, R., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 1–25. https://doi.org/10.1007/978-3-030-87225-0_34-1.
7. Shen, X.; Yan, H.; Wang, C.; Gao, P.; Johnson, C.H.; Snyder, M.P. TidyMass an object-oriented reproducible analysis framework for LC–MS data. *Nat. Commun.* **2022**, *13*, 4365.
8. Chaleckis, R.; Meister, I.; Zhang, P.; Wheelock, C.E. Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. *Anal. Biotechnol.* **2019**, *55*, 44–50.
9. Wang, M.; Carver, J.J.; Phelan, V.V.; Sanchez, L.M.; Garg, N.; Peng, Y.; Nguyen, D.D.; Watrous, J.; Kapono, C.A.; Luzzatto-Knaan, T.; et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **2016**, *34*, 828–837.
10. Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A.A.; Melnik, A.V.; Meusel, M.; Dorrestein, P.C.; Rousu, J. SIRIUS 4: A rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **2019**, *16*, 299–302.
11. Ruttkies, C.; Schymanski, E.L.; Wolf, S.; Hollender, J.; Neumann, S. MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. *J. Cheminform.* **2016**, *8*, 3.
12. Shen, X.; Wang, R.; Xiong, X.; Yin, Y.; Cai, Y.; Ma, Z.; Liu, N.; Zhu, Z.-J. Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. *Nat. Commun.* **2019**, *10*, 1516.
13. Shen, X.; Wu, S.; Liang, L.; Chen, S.; Contrepois, K.; Zhu, Z.J.; Snyder, M. metID: An R package for automatable compound annotation for LC−MS-based data. *Bioinformatics* **2022**, *38*, 568–569.
14. Shen, X.; Wang, C.; Snyder, M.P. massDatabase: Utilities for the operation of the public compound and pathway database. *Bioinformatics* **2022**, *38*, 4650–4651.
15. Nguyen, Q.-H.; Nguyen, H.; Oh, E.C.; Nguyen, T. Current approaches and outstanding challenges of functional annotation of metabolites: A comprehensive review. *Brief. Bioinform.* **2024**, *25*, bbae498.
16. Guo, J.; Yu, H.; Xing, S.; Huan, T. Addressing big data challenges in mass spectrometry-based metabolomics. *Chem. Commun.* **2022**, *58*, 9979–9990.
17. Huang, K.; Mo, F.; Zhang, X.; Li, H.; Li, Y.; Zhang, Y.; Yi, W.; Mao, Y.; Liu, J.; Xu, Y.; et al. A Survey on Large Language Models with Multilingualism: Recent Advances and New Frontiers. *arXiv* **2024**, arXiv:2405.10936.
18. Bharathi Mohan, G.; Prasanna Kumar, R.; Vishal Krishh, P.; Keerthinathan, A.; Lavanya, G.; Meghana, M.K.U.; Sulthana, S. An analysis of large language models: Their impact and potential applications. *Knowl. Inf. Syst.* **2024**, *66*, 5047–5070.
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023.
20. Xu, B.; Poo, M. Large language models and brain-inspired general intelligence. *Natl. Sci. Rev.* **2023**, *10*, nwad267.
21. Zhang, S.; Fan, R.; Liu, Y.; Chen, S.; Liu, Q.; Zeng, W. Applications of transformer-based language models in bioinformatics: A survey. *Bioinforma. Adv.* **2023**, *3*, vbad001.
22. Chi, J.; Shu, J.; Li, M.; Mudappathi, R.; Jin, Y.; Lewis, F.; Boon, A.; Qin, X.; Liu, L.; Gu, H. Artificial intelligence in metabolomics: A current review. *TrAC Trends Anal. Chem.* **2024**, *178*, 117852.
23. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019.
24. Lin, T.; Wang, Y.; Liu, X.; Qiu, X. A Survey of Transformers. *AI Open* **2021**, *3*, 111–132.
25. Rashid, M.M.; Atilgan, N.; Dobres, J.; Day, S.; Penkova, V.; Küçük, M.; Steven; Clapp, R.; Sawyer, B.D. Humanizing AI in Education: A Readability Comparison of LLM and Human-Created Educational Content. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2024**, *68*, 596–603. https://doi.org/10.1177/10711813241261689.
26. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A Comprehensive Overview of Large Language Models. *arXiv* **2024**, arXiv:2307.06435.
27. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* **2019**, *1*, 9.
28. Elman, J.L. Finding structure in time. *Cogn. Sci.* **1990**, *14*, 179–211.
29. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780.
30. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555.

31. O'Shea, K.; Nash, R. An Introduction to Convolutional Neural Networks. *arXiv* **2015**, arXiv:1511.08458.
32. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. *Adv. Neural Inf. Process. Syst.* **2014**, *27*.
33. Li, M.; Wang, X.R. Peak alignment of gas chromatography–mass spectrometry data with deep learning. *J. Chromatogr. A* **2019**, *1604*, 460476.
34. Seddiki, K.; Precioso, F.; Sanabria, M.; Salzet, M.; Fournier, I.; Droit, A. Early Diagnosis: End-to-End CNN–LSTM Models for Mass Spectrometry Data Classification. *Anal. Chem.* **2023**, *95*, 13431–13437.
35. Jain, S.; Safo, S.E. DeepIDA-GRU: A deep learning pipeline for integrative discriminant analysis of cross-sectional and longitudinal multiview data with applications to inflammatory bowel disease classification. *Brief. Bioinform.* **2024**, *25*, bbae339.
36. Kim, H.W.; Zhang, C.; Cottrell, G.W.; Gerwick, W.H. SMART-Miner: A convolutional neural network-based metabolite identification from 1H-13C HSQC spectra. *Magn. Reson. Chem.* **2022**, *60*, 1070–1075.
37. Tang, D.; Qin, B.; Liu, T. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 1422–1432. https://doi.org/10.18653/v1/D15-1167.
38. Huang, X.; Tan, H.; Lin, G.; Tian, Y. A LSTM-based bidirectional translation model for optimizing rare words and terminologies. In Proceedings of the 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 26–28 May 2018; pp. 185–189. https://doi.org/10.1109/ICAIBD.2018.8396191.
39. Radford, A.; Narasimhan, K. Improving Language Understanding by Generative Pre-Training. *OpenAI* **2018**, *12*.
40. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
41. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the Ninth International Conference on Learning Representations, ICLR 2021, Virtual, 3–7 May 2021.
42. Verma, P.; Berger, J. Audio Transformers:Transformer Architectures For Large Scale Audio Understanding. Adieu Convolutions. *arXiv* **2021**, arXiv:2105.00335.
43. Ma, Y.; Chi, D.; Wu, S.; Liu, Y.; Zhuang, Y.; Hao, J.; King, I. Actra: Optimized Transformer Architecture for Vision-Language-Action Models in Robot Learning. *arXiv* **2024**, arXiv:2408.01147.
44. Wang, T.; Roberts, A.; Hesslow, D.; Le Scao, T.; Chung, H.W.; Beltagy, I.; Launay, J.; Raffel, C. What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization? In Proceedings of the 39th International Conference on Machine Learning 2022, Baltimore, MD, USA on 17–23 July 2022.
45. 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. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
46. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; Raffel, C. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv* **2020**, arXiv:2010.11934.
47. Chung, H.W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. Scaling Instruction-Finetuned Language Models. *J. Mach. Learn. Res.* **2022**, *25*, 1–53.
48. 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. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
49. Xiao, Y.; Sun, E.; Jin, Y.; Wang, Q.; Wang, W. ProteinGPT: Multimodal LLM for Protein Property Prediction and Structure Understanding. *arXiv* **2024**, arXiv:2408.11363.
50. Wang, C.; Fan, H.; Quan, R.; Yang, Y. ProtChatGPT: Towards Understanding Proteins with Large Language Models. *arXiv* **2024**, arXiv:2402.09649.
51. Jin, Q.; Yang, Y.; Chen, Q.; Lu, Z. GeneGPT: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics* **2024**, *40*, btae075.
52. Yang, S.; Xu, P. LLM4THP: A computing tool to identify tumor homing peptides by molecular and sequence representation of large language model based on two-layer ensemble model strategy. *Amino Acids* **2024**, *56*, 62.
53. Liu, H.; Wang, H. GenoTEX: A Benchmark for Evaluating LLM-Based Exploration of Gene Expression Data in Alignment with Bioinformaticians. *arXiv* **2024**, arXiv:2406.15341.
54. Lin, X.; Deng, G.; Li, Y.; Ge, J.; Ho JW, K.; Liu, Y. GeneRAG: Enhancing Large Language Models with Gene-Related Task by Retrieval-Augmented Generation. *bioRxiv* **2024**. https://doi.org/10.1101/2024.06.24.600176.
55. Gao, Z.; Liu, Q.; Zeng, W.; Jiang, R.; Wong, W.H. EpiGePT: A pretrained transformer-based language model for context-specific human epigenomics. *Genome Biol.* **2024**, *25*, 310.
56. Chakraborty, C.; Bhattacharya, M.; Lee, S.-S. Artificial intelligence enabled ChatGPT and large language models in drug target discovery, drug discovery, and development. *Mol. Ther. Nucleic Acids* **2023**, *33*, 866–868.

57. Ma, T.; Lin, X.; Li, T.; Li, C.; Chen, L.; Zhou, P.; Cai, X.; Yang, X.; Zeng, D.; Cao, D.; et al. Y-Mol: A Multiscale Biomedical Knowledge-Guided Large Language Model for Drug Development. *arXiv* **2024**, arXiv:2410.11550.
58. Sheikholeslami, M.; Mazrouei, N.; Gheisari, Y.; Fasihi, A.; Irajpour, M.; Motahharynia, A. DrugGen: Advancing Drug Discovery with Large Language Models and Reinforcement Learning Feedback. *arXiv* **2024**, arXiv:2411.14157.
59. Hu, M.; Alkhairy, S.; Lee, I.; Pillich, R.T.; Fong, D.; Smith, K.; Bachelder, R.; Ideker, T. Evaluation of large language models for discovery of gene set function. *Nat. Methods* **2024**, *22*, 82–91. https://doi.org/10.1038/s41592-024-02525-x.
60. Liu, Y.; Chen, Z.; Wang, Y.G.; Shen, Y. TourSynbio-Search: A Large Language Model Driven Agent Framework for Unified Search Method for Protein Engineering. In Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Lisbon, Portugal, 3–6 December 2024.
61. Van Kempen, M.; Kim, S.S.; Tumescheit, C.; Mirdita, M.; Lee, J.; Gilchrist, C.L.M.; Söding, J.; Steinegger, M. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **2024**, *42*, 243–246.
62. Bran, A.M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A.D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **2024**, *6*, 525–535.
63. Boiko, D.A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous chemical research with large language models. *Nature* **2023**, *624*, 570–578.
64. Yang, E.-W.; Velazquez-Villarreal, E. AI-driven conversational agent enhances clinical and genomic data integration for precision medicine research. *medRxiv* **2024**. https://doi.org/10.1101/2024.11.27.24318113.
65. Chen, Q.; Deng, C. Bioinfo-Bench: A Simple Benchmark Framework for LLM Bioinformatics Skills Evaluation. *bioRxiv* **2023**. https://doi.org/10.1101/2023.10.18.563023.
66. Zhou, J.; Zhang, B.; Chen, X.; Li, H.; Xu, X.; Chen, S.; Gao, X. Automated Bioinformatics Analysis via AutoBA. *arXiv* **2023**, arXiv:2309.03242.
67. Biana, J.; Zhai, W.; Huang, X.; Zheng, J.; Zhu, S. VANER: Leveraging Large Language Model for Versatile and Adaptive Biomedical Named Entity Recognition. *arXiv* **2024**, arXiv:2404.17835.
68. Bian, J.; Zheng, J.; Zhang, Y.; Zhou, H.; Zhu, S. One-shot Biomedical Named Entity Recognition via Knowledge-Inspired Large Language Model. In Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Shenzhen China, 22–25 November 2024; Association for Computing Machinery: New York, NY, USA, 2024. https://doi.org/10.1145/3698587.3701356.
69. Wei, J.; Zhuo, L.; Fu, X.; Zeng, X.; Wang, L.; Zou, Q.; Cao, D. DrugReAlign: A multisource prompt framework for drug repurposing based on large language models. *BMC Biol.* **2024**, *22*, 226.
70. Li, Y.; Gao, C.; Song, X.; Wang, X.; Xu, Y.; Han, S. DrugGPT: A GPT-based Strategy for Designing Potential Ligands Targeting Specific Proteins. *bioRxiv* **2023**. https://doi.org/10.1101/2023.06.29.543848.
71. Liu, Y.; Ding, S.; Zhou, S.; Fan, W.; Tan, Q. MolecularGPT: Open Large Language Model (LLM) for Few-Shot Molecular Property Prediction. *arXiv* **2024**, arXiv:2406.12950.
72. Zhang, D.; Liu, W.; Tan, Q.; Chen, J.; Yan, H.; Yan, Y.; Li, J.; Huang, W.; Yue, X.; Ouyang, W.; et al. ChemLLM: A Chemical Large Language Model. *arXiv* **2024**, arXiv:2402.06852.
73. Galkin, F.; Naumov, V.; Pushkov, S.; Sidorenko, D.; Urban, A.; Zagirova, D.; Alawi, K.M.; Aliper, A.; Gumerov, R.; Kalashnikov, A.; et al. Precious3GPT: Multimodal Multi-Species Multi-Omics Multi-Tissue Transformer for Aging Research and Drug Discovery. *bioRxiv* **2024**. https://doi.org/10.1101/2024.07.25.605062.
74. Elsborg, J.; Salvatore, M. Using LLMs and Explainable ML to Analyze Biomarkers at Single-Cell Level for Improved Understanding of Diseases. *Biomolecules* **2023**, *13*, 1516.
75. Baygi, S.F.; Barupal, D.K. IDSL_MINT: A deep learning framework to predict molecular fingerprints from mass spectra. *J. Cheminform.* **2024**, *16*, 8.
76. Würf, V.; Köhler, N.; Molnar, F.; Hahnefeld, L.; Gurke, R.; Witting, M.; Pauling, J.K. LipiDetective—A deep learning model for the identification of molecular lipid species in tandem mass spectra. *bioRxiv* **2024**. https://doi.org/10.1101/2024.10.07.617094 2024.
77. Butler, T.; Frandsen, A.; Lightheart, R.; Bargh, B.; Taylor, J.; Bollerman, T.J.; Kerby, T.; West, K.; Voronov, G.; Moon, K.; et al. MS2Mol: A transformer model for illuminating dark chemical space from mass spectra. *ChemRxiv* **2023**. https://doi.org/10.26434/chemrxiv-2023-vsmpx-v3 2023.
78. Xue, J.; Wang, B.; Ji, H.; Li, W. RT-Transformer: Retention time prediction for metabolite annotation to assist in metabolite identification. *Bioinformatics* **2024**, *40*, btae084.
79. Young, A.; Wang, B.; Röst, H. MassFormer: Tandem Mass Spectrum Prediction for Small Molecules using Graph Transformers. *Nat. Mach. Intell.* **2023**, *6*, 404–416.
80. Liu, Y.; Yoshizawa, A.C.; Ling, Y.; Okuda, S. Insights into predicting small molecule retention times in liquid chromatography using deep learning. *J. Cheminformatics* **2024**, *16*, 113.
81. Cao, M.; Fraser, K.; Huege, J.; Featonby, T.; Rasmussen, S.; Jones, C. Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics. *Metabolomics* **2015**, *11*, 696–706.

82. Yang, Q.; Ji, H.; Lu, H.; Zhang, Z. Prediction of Liquid Chromatographic Retention Time with Graph Neural Networks to Assist in Small Molecule Identification. *Anal. Chem.* **2021**, *93*, 2200–2206.
83. Osipenko, S.; Nikolaev, E.; Kostyukevich, Y. Retention Time Prediction with Message-Passing Neural Networks. *Separations* **2022**, *9*, 291.
84. Ju, R.; Liu, X.; Zheng, F.; Lu, X.; Xu, G.; Lin, X. Deep Neural Network Pretrained by Weighted Autoencoders and Transfer Learning for Retention Time Prediction of Small Molecules. *Anal. Chem.* **2021**, *93*, 15651–15658.
85. Fedorova, E.S.; Matyushin, D.D.; Plyushchenko, I.V.; Stavrianidi, A.N.; Buryak, A.K. Deep learning for retention time prediction in reversed-phase liquid chromatography. *J. Chromatogr. A* **2022**, *1664*, 462792.
86. Zhao, S.; Li, L. Chemical derivatization in LC-MS-based metabolomics study. *TrAC Trends Anal. Chem.* **2020**, *131*, 115988.
87. Domingo-Almenara, X.; Montenegro-Burke, J.R.; Benton, H.P.; Siuzdak, G. Annotation: A Computational Solution for Streamlining Metabolomics Analysis. *Anal. Chem.* **2018**, *90*, 480–489.
88. Lenski, M.; Maallem, S.; Zarcone, G.; Garçon, G.; Lo-Guidice, J.M.; Anthérieu, S.; Allorge, D. Prediction of a Large-Scale Database of Collision Cross-Section and Retention Time Using Machine Learning to Reduce False Positive Annotations in Untargeted Metabolomics. *Metabolites* **2023**, *13*, 282.
89. Voronov, G.; Frandsen, A.; Bargh, B.; Healey, D.; Lightheart, R.; Kind, T.; Dorrestein, P.; Colluru, V.; Butler, T. MS2Prop: A machine learning model that directly predicts chemical properties from mass spectrometry data for novel compounds. *bioRxiv* **2022**. https://doi.org/10.1101/2022.10.09.511482.
90. Heiles, S. Advanced tandem mass spectrometry in metabolomics and lipidomics—Methods and applications. *Anal. Bioanal. Chem.* **2021**, *413*, 5927–5948.
91. Chen, B.; Li, H.; Huang, R.; Tang, Y.; Li, F. Deep learning prediction of electrospray ionization tandem mass spectra of chemically derived molecules. *Nat. Commun.* **2024**, *15*, 8396.
92. Ekvall, M.; Truong, P.; Gabriel, W.; Wilhelm, M.; Käll, L. Prosit Transformer: A transformer for Prediction of MS2 Spectrum Intensities. *J. Proteome Res.* **2022**, *21*, 1359–1364.
93. Young, A.; Röst, H.; Wang, B. Tandem mass spectrum prediction for small molecules using graph transformers. *Nat. Mach. Intell.* **2024**, *6*, 404–416.
94. Allen, F.; Greiner, R.; Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* **2015**, *11*, 98–110.
95. Wang, F.; Liigand, J.; Tian, S.; Arndt, D.; Greiner, R.; Wishart, D.S. CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. *Anal. Chem.* **2021**, *93*, 11692–11700.
96. Wei, J.N.; Belanger, D.; Adams, R.P.; Sculley, D. Rapid Prediction of Electron–Ionization Mass Spectrometry Using Neural Networks. *ACS Cent. Sci.* **2019**, *5*, 700–708.
97. Jin, W.; Coley, C.W.; Barzilay, R.; Jaakkola, T. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. In Proceedings of the 2017 Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
98. Wenk, M. R. Lipidomics: New Tools and Applications. *Cell* **2010**, *143*, 888–895.
99. Kloska, A.; Węsierska, M.; Malinowska, M.; Gabig-Cimińska, M.; Jakóbkiewicz-Banecka, J. Lipophagy and Lipolysis Status in Lipid Storage and Lipid Metabolism Diseases. *Int. J. Mol. Sci.* **2020**, *21*, 6113.
100. Fahy, E.; Cotter, D.; Sud, M.; Subramaniam, S. Lipid classification, structures and tools. *Lipidomics Imaging Mass Spectrom.* **2011**, *1811*, 637–647.
101. Gerhardtova, I.; Jankech, T.; Majerova, P.; Piestansky, J.; Olesova, D.; Kovac, A.; Jampilek, J. Recent Analytical Methodologies in Lipid Analysis. *Int. J. Mol. Sci.* **2024**, *25*, 2249.
102. Yang, J.; Cai, Y.; Zhao, K.; Xie, H.; Chen, X. Concepts and applications of chemical fingerprint for hit and lead screening. *Drug Discov. Today* **2022**, *27*, 103356.
103. Muegge, I.; Mukherjee, P. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opin. Drug Discov.* **2015**, *11*, 137–148.
104. Cereto-Massagué; A; Ojeda, M.J.; Valls, C.; Mulero, M.; Garcia-Vallvé; S; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58–63.
105. Goldman, S.; Wohlwend, J.; Stražar, M.; Haroush, G.; Xavier, R.J.; Coley, C.W. Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nat. Mach. Intell.* **2023**, *5*, 965–979.
106. Stravs, M.A.; Dührkop, K.; Böcker, S.; Zamboni, N. MSNovelist: De novo structure generation from mass spectra. *Nat. Methods* **2022**, *19*, 865–870.
107. Litsa, E.E.; Chenthamarakshan, V.; Das, P.; Kavraki, L.E. An end-to-end deep learning framework for translating mass spectra to de-novo molecules. *Commun. Chem.* **2023**, *6*, 132.
108. Shrivastava, A.D.; Swainston, N.; Samanta, S.; Roberts, I.; Wright Muelas, M.; Kell, D.B. MassGenie: A Transformer-Based Deep Learning Method for Identifying Small Molecules from Their Mass Spectra. *Biomolecules* **2021**, *11*, 1793.
109. Elser, D.; Huber, F.; Gaquerel, E. Mass2SMILES: Deep learning based fast prediction of structures and functional groups

directly from high-resolution MS/MS spectra. *bioRxiv* **2023**. https://doi.org/10.1101/2023.07.06.547963.

110. Vaniya, A.; Fiehn, O. Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics. *TrAC Trends Anal. Chem.* **2015**, *69*, 52–61.
111. Zhang, M.; Xia, Y.; Wu, N.; Qian, K.; Zeng, J. MS^2-Transformer: An End-to-End Model for MS/MS-assisted Molecule Identification. 2022.
112. Yang, Y.; Sun, S.; Yang, S.; Yang, Q.; Lu, X.; Wang, X.; Yu, Q.; Huo, X.; Qian, X. Structural annotation of unknown molecules in a miniaturized mass spectrometer based on a transformer enabled fragment tree method. *Commun. Chem.* **2024**, *7*, 109.
113. Meng, W.; Pan, H.; Sha, Y.; Zhai, X.; Xing, A.; Lingampelly, S.S.; Sripathi, S.R.; Wang, Y.; Li, K. Metabolic Connectome and Its Role in the Prediction, Diagnosis, and Treatment of Complex Diseases. *Metabolites* **2024**, *14*, 93.
114. Amara, A.; Frainay, C.; Jourdan, F.; Naake, T.; Neumann, S.; Novoa-Del-Toro, E.M.; Salek, R.M.; Salzer, L.; Scharfenberg, S.; Witting, M. Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation. *Front. Mol. Biosci.* **2022**, *9*, 841373.
115. Matsumoto, N.; Moran, J.; Choi, H.; Hernandez, M.E.; Venkatesan, M.; Wang, P.; Moore, J.H. KRAGEN: A knowledge graph-enhanced RAG framework for biomedical problem solving using large language models. *Bioinformatics* **2024**, *40*, btae353.
116. Wen, Y.; Wang, Z.; Sun, J. MindMap: Knowledge Graph Prompting Sparks Graph of Thoughts in Large Language Models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024; Association for Computational Linguistics, Bangkok, Thailand, 2024; pp. 10370–10388. https://doi.org/10.18653/v1/2024.acl-long.558.
117. Beck, A.G.; Muhoberac, M.; Randolph, C.E.; Beveridge, C.H.; Wijewardhane, P.R.; Kenttamaa, H.I.; Chopra, G. Recent Developments in Machine Learning for Mass Spectrometry. *ACS Meas. Sci. Au* **2024**, *4*, 233–246.
118. Pinto, R.C.; Karaman, I.; Lewis, M.R.; Hällqvist, J.; Kaluarachchi, M.; Graça, G.; Chekmeneva, E.; Durainayagam, B.; Ghanbari, M.; Ikram, M.A.; et al. Finding Correspondence between Metabolomic Features in Untargeted Liquid Chromatography–Mass Spectrometry Metabolomics Datasets. *Anal. Chem.* **2022**, *94*, 5493–5503.
119. Wörheide, M.A.; Krumsiek, J.; Kastenmüller, G.; Arnold, M. Multi-omics integration in biomedical research—A metabolomics-centric review. *Anal. Chim. Acta* **2021**, *1141*, 144–162.
120. Sanches PH, G.; de Melo, N.C.; Porcari, A.M.; de Carvalho, L.M. Integrating Molecular Perspectives: Strategies for Comprehensive Multi-Omics Integrative Data Analysis and Machine Learning Applications in Transcriptomics, Proteomics, and Metabolomics. *Biology* **2024**, *13*, 848.
121. Maan, K.; Baghel, R.; Dhariwal, S.; Sharma, A.; Bakhshi, R.; Rana, P. Metabolomics and transcriptomics based multi-omics integration reveals radiation-induced altered pathway networking and underlying mechanism. *NPJ Syst. Biol. Appl.* **2023**, *9*, 42.