Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

REVIEW

View Article Online
View Journal

Cite this: DOI: 10.1039/d5dd00504c

AI-driven natural product-based antiviral drug development: a technical overview

Junxi Song,†[a] Kunhuan Yang,†[a] Yingcai Xiong,†[b] Keyu Tao,†[a] Liangyu Cai,*[c] Peng Cao*[d] and Jianjian Ji *[ac]

The emergence of viral pandemics and rapid pathogen evolution present formidable challenges for conventional antiviral development, including prolonged timelines, high costs, and susceptibility to resistance mechanisms. Natural products (NPs) offer promising antiviral potential through structural diversity and multi-target synergism, while their development faces critical bottlenecks in structural characterization, target identification, and synthetic optimization. Given the current situation, artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), is revolutionizing drug development by transforming data analysis and predictive modeling. This review explores AI applications across the antiviral NP drug development continuum, providing insights for AI-driven pharmaceutical research.

Received 13th November 2025
Accepted 3rd April 2026

DOI: 10.1039/d5dd00504c

rsc.li/digitaldiscovery

1. Introduction

In recent years, many new and re-emerging viral pathogens—such as severe acute respiratory syndrome coronavirus (SARS-CoV), Middle East respiratory syndrome coronavirus (MERS-CoV), and SARS-CoV-2—have continued to pose a risk to global public health.[1] Antiviral medicine development is associated with high costs and extended timescales.[2] Moreover, the current pace of pharmaceutical research and development critically lags behind the exponentially growing need for rapid therapeutic interventions during emerging pandemic scenarios (Fig. 1). The remdesivir development timeline exemplifies both the opportunities and persistent limitations in this race against time. Initially explored in 2009 for hepatitis C and later Ebola, it gained emergency use authorization in May 2020 for COVID-19—only half a year after the SARS-CoV-2 outbreak, showcasing a smooth translation from scientific discovery to emergency response.[3] Yet, as depicted in Fig. 1, even this rapid repurposing occurred amid ongoing viral evolution, with variants like Alpha and Beta emerging before full clinical implementation, highlighting that such success depends on pre-existing scaffolds and may not suffice for *de novo* threats. For novel viral threats where no such prior knowledge exists, the *de novo* drug development process remains substantially slower than the pace of outbreaks.[4] Due to the virus's unique genetic system (lack of a complex genetic information synthesis proofreading system), the virus can rapidly mutate and evade drug treatment. Consider the H1N1 influenza variants: ingenious mutations in the Sb region of the HA protein serve like a molecular camouflage kit, allowing them to evade vaccine-educated antibodies and leaving vaccination campaigns in the lurch.[5] Such complex pressures require more sophisticated therapeutic strategies, which demand transformed drug development paradigms.

Natural products—complex metabolites produced by plants, fungi, animals, and microorganisms—exhibit the highest chemical diversity in nature,[6] and are an important source of antiviral medicines. Many licensed treatments and prospects stem directly or indirectly from plants, microbes, and marine animals. For instance, artemisinin's antimalarial potency signified a milestone for natural-product-based anti-infective discovery;[7] diammonium glycyrrhizinate from licorice root is authorized in China and Japan as an adjuvant for chronic hepatitis B;[8] and numerous antivirals (*e.g.*, acyclovir, ganciclovir, vidarabine, and zidovudine) emerged from natural leads *via* structural optimization.[9] Despite this promise, development confronts obstacles: limited access to bioactive substances (only ~1% of microbial species are culturable[10]); complex metabolite isolation and characterization (bioassay-guided fractionation and structural elucidation are often essential); uncertain pharmacological mechanisms (*e.g.*, the multi-component synergy of Lianhua Qingwen capsules remains incompletely defined[11]);

[a]Jiangsu Key Laboratory of Children's Health and Chinese Medicine, The First Clinical College, Jiangsu Provincial Research Institute of Chinese Medicine Schools, Nanjing University of Chinese Medicine, Nanjing, 210028, China. E-mail: 039222215@njucm.edu.cn; tky615322123@163.com

[b]The State Key Laboratory of Pharmaceutical Biotechnology, Chemistry and Biomedicine Innovation Center (ChemBIC), Division of Immunology, Medical School, Nanjing University, Nanjing, 210093, China. E-mail: yingcai_x@163.com

[c]Wuxi Affiliated Hospital of Nanjing University of Chinese Medicine, 138 Xianlin Avenue, Wuxi, 210023, China. E-mail: jijj@njucm.edu.cn; wxzy018@njucm.edu.cn; 039222333@njucm.edu.cn

[d]State Key Laboratory on Technologies for Chinese Medicine Pharmaceutical Process Control and Intelligent Manufacture, Nanjing University of Chinese Medicine, Nanjing, China. E-mail: cao_peng@njucm.edu.cn

† These authors contributed equally

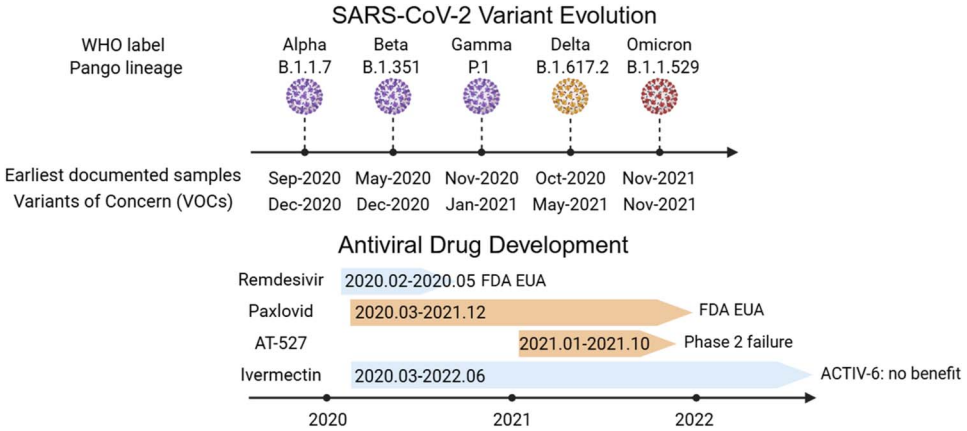

Fig. 1 Asynchrony between SARS-CoV-2 variant evolution and antiviral drug development. This figure illustrates the evolutionary dynamics of major SARS-CoV-2 variants during the COVID-19 pandemic, aligned with the development and clinical implementation timelines of four representative small-molecule antivirals. The upper axis delineates SARS-CoV-2 evolutionary dynamics using a generational color scheme: purple signifies the initial waves of variants of concern (Alpha, Beta, and Gamma) marked by early increases in transmissibility; yellow identifies the Delta variant as a pivotal transition toward significantly higher viral loads and pathogenicity; and red represents the Omicron lineage, which constitutes a fundamental paradigm shift in immune evasion and mutation density. Each variant is annotated with its earliest documented detection date and official WHO Variant of Concern (VOC) designation, with earliest documented detections referring to retrospectively identified sequences rather than real-time discovery, to highlight the inherent delay in global surveillance and response. Antivirals plotted on the lower axis were selected according to two defining dimensions: the development pathway (repurposing of established agents *versus de novo* structural design) and the clinical resolution (regulatory authorization *versus* termination due to futility). Blue bands (remdesivir and ivermectin) represent the drug repurposing strategy, aimed at immediate deployment based on known safety profiles. In contrast, orange bands (Paxlovid and AT-527) signify *de novo* discovery programs. Despite such technological acceleration, vertical alignment across the axes reveals that by the time these high-potency agents reached their respective clinical endpoints, the viral landscape had already transitioned through multiple generational cycles, illustrating the persistent structural lag between therapeutic intervention and emergent pandemic needs. This figure highlights the asynchrony between viral evolution and therapeutic development, underscoring the challenges in maintaining efficacy against phylogenetically divergent lineages. Figure created with https://www.BioRender.com.

and complex synthesis (*e.g.*, up to 30 enzymatic cascade steps to generate vinblastine[12]).

AI provides transformative solutions for these challenges. Advanced algorithms—including Transformer architectures and graph neural networks (GNNs)—have made great strides in accuracy for predicting drug–target interaction.[13] In addition, biomedical data are exploding (*e.g.* the ChEMBL compound library contains >20.3 million bioactivity measurements and >2.4 million unique compounds).[14] Furthermore, the Traditional Chinese Medicine Systems Pharmacology Database (TCMSP) includes 29 384 components, 3311 with targets, and 837 linked to diseases,[15] thus simplifying model training. On the other hand, GPU clusters speed up computational tasks, improving, by several orders of magnitude, the speed of training large-scale AI models, such as DL architectures for drug–target interaction prediction, when compared with CPUs.[16,17]

The recent emergence of AI-driven platforms has already yielded significant breakthroughs in antiviral discovery. For instance, a sophisticated AI pipeline recently identified established antiretrovirals, such as bictegravir and etravirine, as potent broad-spectrum inhibitors against monkeypox virus and related poxviruses.[18] While such success underscores the transformative potential of AI in accelerating drug repurposing and novel application discovery, there remains a need for a more granular technical overview focused specifically on the end-to-end integration of AI across the entire natural product-based antiviral drug development continuum.

In this review we summarize some of the critical applications of AI in upstream (resource mining and target identification), midstream (drug candidate screening and optimization), and downstream (preclinical and clinical stages). We subsequently explore the current technological limitations and emerging possibilities. Finally, we discuss the future directions for the field. We hope to highlight a new era of technology, efficiency and precision in drug development that is expected to speed delivery of new and improved medicines to patients.

2. AI applications in natural product-based antiviral drug development

AI technologies are increasingly integrated into various stages of natural product-based antiviral drug development, forming a systematic and intelligent pipeline from resource mining and target identification to preclinical and clinical applications (Fig. 2).

2.1 Upstream: resource mining and target identification

2.1.1 AI for genome mining and biosynthetic gene clusters (BGCs). Multiple antiviral NPs are the products of secondary metabolites prescribed in microbial[19,20] and plant genomes.[21–23] AI is enhancing the search for these biosynthetic gene clusters. Standard genome-mining tools, such as anti-SMASH, employ rule-based pattern matching to identify known classes of BGCs,

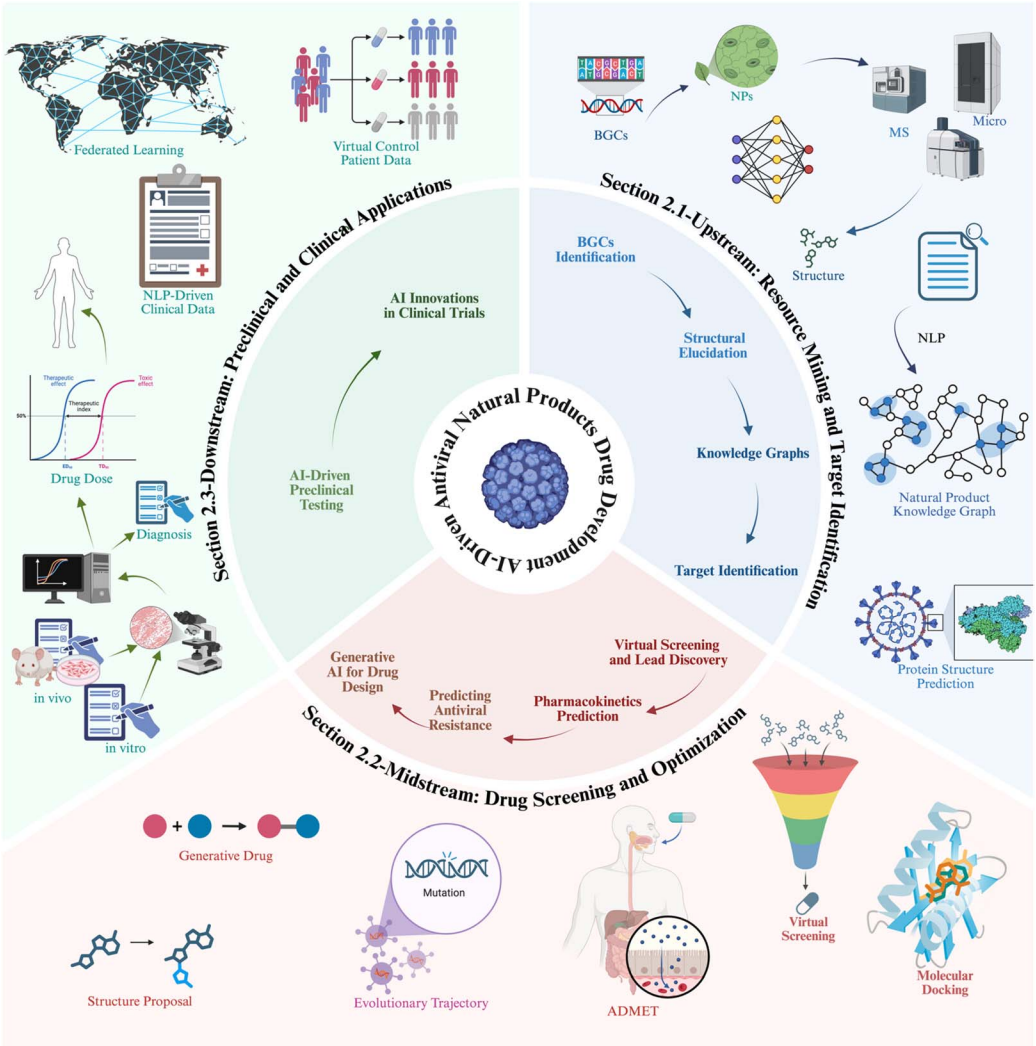

Fig. 2 AI-Driven strategies for natural product-based antiviral drug development across the full pipeline. This figure illustrates the comprehensive application of AI throughout the full pipeline of natural product-based antiviral drug development. It is divided into three main stages: upstream (resource mining and target identification), midstream (drug screening and optimization), and downstream (preclinical and clinical applications). In the upstream stage, AI facilitates biosynthetic gene cluster (BGC) identification, structural elucidation of NPs, construction of knowledge graphs, protein structure prediction, and target identification. In the midstream stage, AI supports virtual screening, molecular docking, ADMET prediction, evolutionary trajectory analysis, antiviral resistance prediction, and generative drug design. In the downstream stage, AI enhances preclinical testing, dose optimization, diagnostic assistance, NLP-driven clinical data mining, federated learning, virtual control cohort construction, and innovations in clinical trials. Different color-coded sections visualize key processes across stages, highlighting the integrative role of AI in advancing natural product-based antiviral drug discovery. Figure created with https://www.BioRender.com.

which results in under-representative identification of clusters that are "atypical" clusters and in significant false-negative rates.[24] Recognizing small sequence motifs beyond human-defined elements, DL approaches can be used to discover novel BGCs. For example, the RNN-based DeepBGC found novel BGCs in Streptomyces genomes.[25] Such AI models leverage gene context, conserved domains, and amino acid property generation to illuminate potentially hidden BGCs that may yield new antiviral drugs. By bridging genetic and metabolomic data, these growing algorithms can correlate putative gene clusters with actual molecules, completing the link from genes to the NPs they produce. Genome-to-metabolite prediction is essential for discovering new antiviral chemicals that are stored in nature's genomic libraries.

2.1.2 AI-assisted structural elucidation. The unambiguous structure determination of NPs is typically a labor-intensive, time-consuming process.[26] AI enhances understanding of complex spectrometric data, including mass spectrometry (MS) and liquid chromatography–mass spectrometry (LC-MS).[27,28] Methods for AI-assisted structural elucidation can be broadly classified into three categories:[1] ML/DL-based MS/MS spectrum annotation and substructure prediction;[2] predictive modeling for LC-MS retention time, peak feature extraction, and spectrum grouping; and[3] integrated pipelines for advanced techniques like microcrystal electron diffraction (MicroED). This

classification reflects the shift from rule-based manual analysis to data-driven automation.[29]

Recent advancements further strengthen this capability: MZmine 3, a scalable open-source platform, supports integrative processing of multimodal MS data (including LC-MS and ion mobility), enabling efficient feature detection, visualization, and annotation tailored to natural product workflows.[30] Similarly, studies employing LC-MS/MS and molecular networking have identified marine-derived secondary metabolites with anti-SARS-CoV-2 activity, such as homofascaplysin A and aureol, although structural confirmation remains labor-intensive due to stereochemical complexity.[31] MicroED with streamlined AI pipelines has enabled 3D structure determination of macrocyclic NPs with antiviral potential, overcoming the need for large single crystals by utilizing microcrystals—yet sample preparation is challenging, as NPs often form amorphous or poorly diffracting microcrystals, limiting resolution and throughput.[32] ML-based retention-time prediction further enhances confidence by avoiding re-isolation of known compounds, but struggles with NPs' high chemical diversity and batch variability.[33–35] In summary, AI-facilitated spectrometric data analysis has begun to streamline structural elucidation in natural product research, thereby facilitating the identification of potential antiviral candidates. Nevertheless, overcoming NP-inherent hurdles—such as mixture complexity, data scarcity for rare scaffolds, and the substantial validation gap—will be essential to translate these computational advances into robust, clinically relevant antiviral natural products.

2.1.3 Knowledge graphs and natural language processing (NLP) for natural product data. New tools, including AI, are also increasingly utilized in organizing and mining this vast knowledge repository of NPs and traditional medicine. Knowledge graphs are structured networks that integrate heterogeneous data sources, such as chemical structures, biological targets, biosynthetic pathways, and literature references, enabling cross-domain analysis.[36] NLP methods, on the other hand, extract structured information from unstructured texts, particularly historical herb manuals and medical literature. Knowledge graphs offer high connectivity and query efficiency but face challenges in entity resolution and data integration due to the heterogeneity of NP sources (*e.g.*, varying nomenclature and incomplete annotations). NLP excels at text mining but struggles with ancient language ambiguity, OCR errors, and domain-specific terminology in traditional medicine texts.

In applications, knowledge graphs have demonstrated value. For example, a knowledge graph on natural products connects segments of tandem MS to predicted metabolites and those predicted metabolites to potential generating genes, as implemented in frameworks like the Experimental Natural Products Knowledge Graph (ENPKG), which integrates multimodal data for plant-derived compounds.[37] Recent efforts further demonstrate this by leveraging AI to associate MS fragmentation patterns with biosynthetic gene clusters through substructure discovery and BGC–metabolite mapping.[38] This graph can emulate an expert chemist's intuition, mining correlations that yield novel antivirals. In the textual side, NLP methods are being applied to the vast literature on medicinal plants and traditional therapies. One of them constructed TCMBank, one of the largest integrative databases associating TCM with multi-omics data, based on text-mining historical herb manuals and medical texts. The system, employing advanced NLP techniques initially based on bidirectional LSTM networks and conditional random fields with subsequent enhancements incorporating Transformer-based models, amassed structured data on herbs, chemicals and known effects from dozens of ancient manuscripts.[39] This "knowledge reconstruction" re-works past empirical material into a machine-readable database enabling the AI to rapidly sift through potential antiviral medicines in conventional literature and then match them against prevailing biomedical information.

By transforming fragmented knowledge into connected, machine-readable data through knowledge graphs and curated databases, AI enhances upstream discovery of natural antivirals. Researchers can now query these systems to identify promising compounds and targets far more efficiently than manual curation. Nevertheless, the field must address NP-specific hurdles—such as textual ambiguity in ancient sources, data heterogeneity, and the persistent validation gap—to ensure reliable, translational impacts in antiviral drug development.

2.1.4 AI-driven target identification. Another essential upstream step is targeting molecular targets (viral or host) for natural antivirals. AI accelerates this process through a combination of data-driven and structure-based approaches, enabling more efficient prediction and validation.[40] To achieve comprehensive coverage, we organize AI-driven target identification into a trinity framework: chemical-centric, system-centric, and physics-centric. This structure addresses ligand-based, network-based, and structure-based paradigms in pharmacology, while incorporating emerging techniques to fill critical gaps such as phenotypic profiling, metabolomic interference, and dynamic simulations.

(i) Chemical-centric strategies focus on ligand-chemical space using deep transfer learning and matrix factorization to associate "orphan ligands" with known targets, including phenotypic/image-based AI (*e.g.*, cell painting assays) that enables forward pharmacology by inferring pathways from cellular morphological fingerprints without prior ligand–target data.[41] For instance, the STarFish platform, a stacked ensemble model, identifies potential targets for NPs by leveraging known ligand–target data, achieving high accuracy in multi-target predictions on benchmark datasets.[42] Similarly, DeepPurpose employs DL for drug–target interaction prediction, facilitating virtual screening of NPs with improved hit rates.[43]

(ii) System-centric strategies integrate knowledge graphs and multimodal GNNs to reveal hidden nodes in virus–host interaction networks, integrated with AI-driven metabolomic analysis (*e.g.*, flux balance analysis in integrated metabolic models) to predict how NPs alter host metabolic environments, uncovering indirect antiviral targets.[44] For example, graph neural networks have been applied to construct and analyze SARS-CoV-2 knowledge graphs based on virus–host interactions, pathways, and drug associations, identifying potential host genes and biological processes for antiviral drug repurposing.[45] Tools

like TCMBank leverage NLP and knowledge graphs for traditional medicine data mining, enabling synergistic target discovery.[39] This dimension also extends to genomic surveillance for pathogen evolution, enhancing target relevance in dynamic viral contexts.[46]

(iii) Physics-centric strategies employ generative AI for biophysical predictions and dynamic simulations. AlphaFold 3 exemplifies this by providing ~50% improved accuracy in predicting protein–ligand and nucleic acid interactions compared to physics-based docking[47] and RoseTTAFold All-Atom for all-atom modeling of protein-ligand dynamics.[48]

Overall, these AI technologies—from knowledge graphs to predictive models—hold significant potential to streamline upstream discovery by connecting chemical leads with biological targets, paving the way for next-generation antiviral drugs. Notably, polypharmacology represents a core advantage of NPs: their multi-target effects enable synergistic therapeutic outcomes, and AI shows emerging potential to actively design and quantify these synergistic or antagonistic effects for optimized efficacy.[49] However, challenges persist, including data bias in training sets (*e.g.*, underrepresentation of rare interactions leading to high false-positive rates in a data-starved setting) and the experimental–computational gap, necessitating rigorous validation and diverse datasets to mitigate biases and improve generalization.

2.2 Midstream: drug screening and optimization

2.2.1 Virtual screening and lead discovery. In the screening and lead identification phase, AI has significantly enhanced processes by enabling ultra-efficient virtual screening of enormous chemical libraries. While traditional high-throughput wet lab screens are often resource-intensive, AI *in silico* methods can evaluate billions of compounds more rapidly. However, the effectiveness of these *in silico* screenings heavily relies on the quality and maturity of the underlying machine learning models, such as the accuracy of training data and model generalization to avoid common pitfalls like overfitting or data biases.[50,51] Quantitative structure–activity relationship (QSAR) models play a central role in this, with performance that is nuanced and contingent upon the available data regime. In high-data regimes, modern approaches utilizing GNNs and transformers excel at learning complex molecular features from structural information, eliminating the need for manual descriptor selection and improving prediction accuracy for large datasets.[52] For example, DL-QSAR models integrating molecular fingerprints and GNN-derived features can accelerate antiviral activity prediction and prioritize natural product analogs.[53] In contrast, in low-data regimes—common in natural product antiviral discovery—classical techniques such as tree-based models (*e.g.*, random forests) combined with circular fingerprints often perform comparably or better, offering robustness, simplicity, and superior generalization with fewer samples.[51,54,55] Hybrid strategies blending classical and advanced methods may yield optimal results across diverse scenarios, balancing computational efficiency and reliability.

In structure-based virtual screening, traditional physics-based molecular docking faces prohibitive computational costs when traversing billion-scale chemical spaces, necessitating AI-driven strategies that optimize both accuracy and throughput while addressing limitations in generalization, especially for structurally complex NPs. Key facets of this paradigm shift encompass enhancing evaluation precision through DL-driven affinity estimation—exemplified by curvature-based GNNs such as CurvAGN, which capture intricate 3D molecular geometries and multi-scale interactions to mitigate systematic biases in classical empirical scoring functions.[56] To surmount scalability constraints of exhaustive docking, Deep Docking-type paradigms deploy DL surrogate models trained on representative subsets to forecast docking scores for the remainder of the library, thereby excluding over 99% of non-binders without explicit physics-based computations—a capability indispensable for trillion-scale ultra-large chemical library screening.[57] This acceleration is further augmented by Active Learning (AL) strategies, which reconfigure virtual screening into a dynamic 'sampling–docking–training–prediction' iterative loop.[58] Through iterative selection of the most informative compounds, AL frameworks identify top-tier leads while requiring docking of less than 1% of the library, drastically mitigating resource demands. Integrating these AI accelerators with robust error-control mechanisms, such as the Conformal Prediction (CP) methodology, assures high sensitivity and reliability in ligand discovery for challenging targets like G protein-coupled receptors (GPCRs).[59] Although these innovations markedly expedite antiviral natural product screening, persistent challenges include limited model generalization to novel scaffolds and substantial computational infrastructure requirements.

AI also permits multi-target screening, interrogating interactions between compounds and multiple viral proteins (*i.e.*, polymerase and protease), and enabling broad-spectrum antivirals with tuned polypharmacology.[60] The synergistic use of generative models, CP-guided docking, and multi objective optimization allows AI-driven virtual screening to accelerate hit discovery while increasing chemical diversity when compared to brute-force approaches in terms of both speed and lead quality.

2.2.2 AI-enhanced pharmacokinetics and toxicity prediction. The optimization of pharmacokinetic (PK) and toxicity profiles has traditionally represented a high-attrition, late-stage bottleneck in drug development; however, AI is increasingly transforming this process into an early-stage, parallel Multi-Parameter Optimization (MPO) paradigm. Rather than treating ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) as a sequential experimental filter, multi-task learning algorithms now enable the simultaneous prediction of diverse drug-likeness endpoints from a single molecular representation (Fig. 3).[61,62] This holistic evaluation facilitates the navigation of complex trade-offs—such as balancing oral bioavailability against cardiotoxicity risks—during the hit-to-lead transition.[63,64] Furthermore, the integration of transfer learning addresses data sparsity challenges in natural product research by adapting animal-derived datasets to human-specific predictions, while interpretability tools like SHAP (SHapley

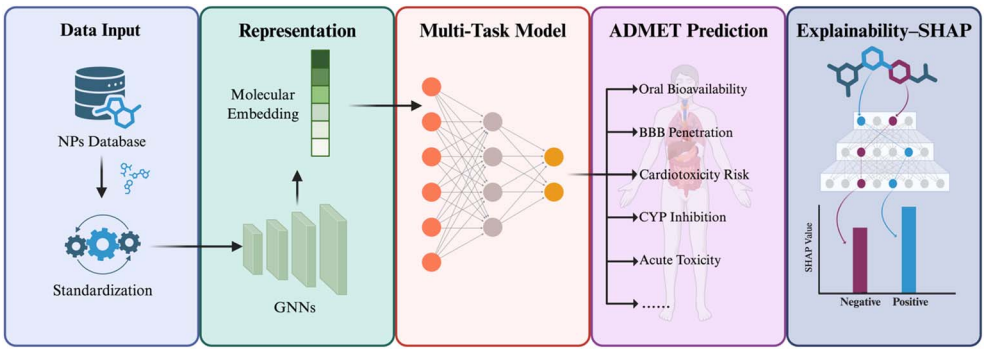

Fig. 3 Schematic workflow of AI-driven multi-task learning for ADMET prediction in natural product drug development. The workflow starts with data input from NP databases. Data standardization follows, normalizing formats (*e.g.*, SMILES) and curating biases. GNNs then generate molecular embeddings by encoding compounds as graphs, capturing NP-specific features like macrocycles. A multi-task model predicts interrelated endpoints simultaneously, such as oral bioavailability, blood–brain barrier (BBB) penetration (*e.g.*, *via* permeability coefficients), cardiotoxicity risk (*e.g.*, hERG channel inhibition), cytochrome P450 (CYP) enzyme inhibition (*e.g.*, CYP3A4 isoform specificity), and acute toxicity (*e.g.*, LD_{50} estimates), using shared representations for efficiency. SHAP, a game-theoretic interpretability framework based on Shapley values from cooperative game theory, quantifies feature contributions (positive or negative) to each prediction, visualizing substructure impacts (*e.g.*, highlighting aromatic rings contributing to hepatotoxicity) to guide targeted structural modifications in lead optimization. Figure created with https://www.BioRender.com.

Additive exPlanations) offer mechanistic insights by identifying toxicophoric substructures for targeted medicinal chemistry modifications.[65,66]

The industrial applicability of these AI-driven workflows is supported by documented improvements in throughput and predictive accuracy. For instance, platforms such as ADMETLab 3.0, which employ directed message-passing neural networks, have shown the ability to evaluate over 119 endpoints with AUROC values up to 0.94, contributing to enhanced efficiency in computational screening compared to traditional empirical models.[67,68] Regulatory horizon-scanning reports, such as those from the European Medicines Agency (EMA), underscore the deployment of tools like Toxometris.ai and the Deep-PK framework in toxicity and pharmacokinetic predictions and preclinical study designs within pharmaceutical pipelines.[59,69] These implementations suggest that AI can enhance predictive capabilities and potentially de-risk the development of natural antivirals by supporting a more streamlined transition from hit identification to clinical candidate nomination. Nonetheless, ongoing challenges in model generalization to novel scaffolds highlight the need for rigorous validation and hybrid approaches to ensure translational reliability.[59,61]

2.2.3 Predicting and mitigating antiviral resistance. Viruses evolve rapidly, posing a persistent challenge to antiviral drug efficacy as resistance mechanisms emerge, often rendering treatments obsolete within months or years. Traditional approaches rely on reactive surveillance and empirical testing, but AI introduces a proactive paradigm by forecasting evolutionary trajectories, identifying resistance signatures, and guiding resilient inhibitor design.

AI models leverage sequence data and structural predictions to anticipate viral mutations at binding sites, enabling the development of inhibitors that maintain potency against future variants. For example, EVEscape computationally produced multi-mutant SARS-CoV-2 spikes to replicate immune escape, with experimental validation.[70] Predicting mutations at binding sites can guide the design of inhibitors resilient to future variants. AI may also search sequence databases for resistance signatures and offer chemical modifications to bypass common mechanisms.[71] These methods also facilitate chemical modifications, using generative AI to suggest scaffold alterations that bypass common resistance pathways, shifting from post-resistance response to preemptive antiviral engineering.

Despite these advances, significant dilemmas arise from the interplay of viral biology and technological limitations. Viruses' high mutation rates create out-of-distribution challenges for AI models, where training on historical variants may fail to generalize to novel, phylogenetically divergent strains, leading to inaccurate predictions and false confidence in drug resilience.[46] Computationally, scaling simulations for multi-mutant landscapes demands immense resources, often exceeding available infrastructure and introducing biases from incomplete datasets. NPs' structural diversity offers a vast pool for discovering new scaffolds less prone to resistance, but integrating this with AI remains underdeveloped due to data scarcity in NP-specific resistance profiles.

2.2.4 Generative AI for novel drug design and synthesis planning. Generative AI serves as an impressive midstream approach in antiviral NP drug development, supporting the *de novo* generation of molecular structures inspired by NPs' structural diversity and multi-target synergies.[72] These models process chemical patterns from large datasets, such as general chemical databases like ChEMBL (which includes NPs) or NP-focused libraries like TCMSP and COCONUT, to suggest candidates that could address viral evolution and resistance issues.[73] Generative frameworks commonly include variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion models, often implemented with sequence-based backbones (*e.g.*, recurrent neural networks (RNNs) or Transformers) or graph-based structures, allowing exploration of

chemical spaces beyond conventional screening and potentially improving timeline efficiency.[74] These major generative AI paradigms can be summarized as follows (Fig. 4). They efficiently generate molecular structures inspired by natural products through mechanisms such as sequence modeling, latent space exploration, diffusion processes, or evolutionary optimization. Notably, generative AI generally produces *de novo* molecules that are NP analogs or mimics, often struggling to replicate the intricate structural features of authentic NPs, such as high chirality centers or complex ring systems, yet it can integrate NP-like motifs (*e.g.*, macrocycles or polyketide scaffolds) to approximate bioactivity benefits like polypharmacology, as discussed in recent AI–NP literature.[75] This method aims to align natural product-inspired designs with synthetic feasibility, facilitating multi-objective considerations for aspects like binding affinity, ADMET properties, and accessibility in antiviral settings.

Examining individual architectures, VAE project molecular structures into latent spaces for sampling NP-inspired analogs, supporting the development of structurally complex antiviral candidates, although accurately mirroring full NP polypharmacology can be difficult.[76] GANs utilize adversarial training to yield plausible NP-like molecules, such as analogs of artemisinin for possible broad-spectrum antiviral exploration, *via* generator-discriminator refinement.[75] Diffusion models construct full 3D molecular conformations by progressively denoising atomic coordinates and atom types in an equivariant manner, offering advantages in generating geometrically accurate NP-mimetic structures that might counter viral mutations.[77] RNNs and Transformers, as sequence-based backbones, manage formats like SMILES strings to produce varied analogs, enabling investigation of NP-adjacent chemical spaces for antiviral purposes.[78] GAs, meanwhile, simulate evolution *via* selection and mutation of known structures, embedding synthetic accessibility measures to optimize NP analogs.[74] By linking these architectures with retrosynthesis planning tools like AiZynthFinder, generative AI can propose candidate structures alongside practical synthesis routes, encouraging an iterative design-make-test process that broadens chemical diversity in line with NP bioactivity concepts, although validation against diverse viral variants remains a key limitation.[79]

2.3 Downstream: preclinical and clinical applications

2.3.1 AI-driven preclinical testing. Before advancing to clinical trials, natural product-based antiviral drug must undergo rigorous preclinical evaluation, including *in vitro* assays, animal studies, and toxicity assessments. AI has been integrated into these stages to enhance efficiency, but its application in NPs—characterized by structural complexity and multi-target interactions—presents unique challenges. This subsection provides a systematic overview of AI methods in preclinical testing for antiviral NPs, focusing on experimental optimization, automated data analysis, and predictive modeling, while critically discussing limitations and real-world impacts.

First, AI facilitates experimental design through optimization algorithms, such as Bayesian optimization (BO) and active

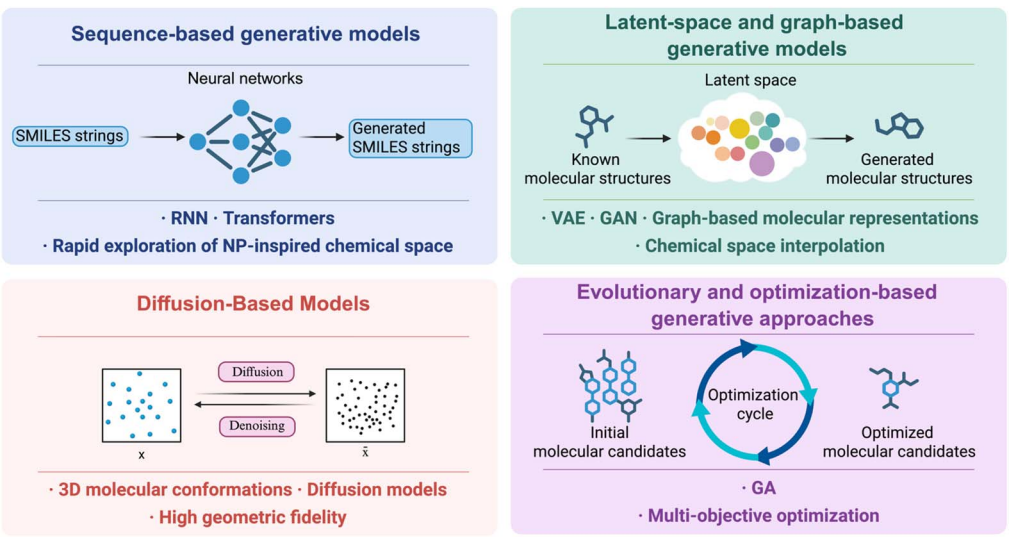

Fig. 4 Overview of major generative AI paradigms applied to natural product–inspired molecular design. Sequence-based generative models treat molecular generation as a language modeling problem, in which molecular structures are encoded as SMILES strings and processed by sequence-based neural networks to rapidly explore NP-inspired chemical space. Latent-variable and graph-based generative models embed known molecular structures into a continuous latent space using latent representations and graph-based molecular encodings, enabling interpolation and sampling to generate novel molecular structures. Diffusion-based generative models generate molecular structures by progressively denoising stochastic representations, allowing the reconstruction of high-fidelity three-dimensional molecular conformations. Evolutionary and optimization-based generative approaches iteratively refine molecular candidates through mutation and selection under multi-objective optimization criteria, explicitly incorporating considerations such as chemical properties and synthetic feasibility. Figure created with https://www.BioRender.com.

learning. BO iteratively selects experimental parameters (*e.g.*, compound dosages or combinations) based on prior data to maximize information gain with minimal trials. In the context of antiviral NPs, BO has been applied to refine *in vitro* assays for plant-derived compounds like artemisinin analogs against broad-spectrum antivirals including coronaviruses, achieving comparable enrichment factors while reducing the experimental footprint by approximately 40–50% compared to traditional methods.[80] Similarly, active learning frameworks combine machine learning with high-throughput screening to prioritize NPs from microbial sources, such as polyketides discovered through genome mining with anti-viral potential.[81] These approaches accelerate preclinical workflows by intelligently navigating the vast chemical space of NPs.

Second, AI enhances automated analysis of preclinical data, particularly in imaging and histopathology. DL models, such as CNNs, automate the scoring of pathology slides from animal models. For instance, the CSGO (Cell Segmentation with Globally Optimized boundaries) pipeline has recently been validated for high-throughput, whole-cell segmentation in Hematoxylin-and-Eosin (H&E)-stained tissues, enabling precise quantification of inflammatory cell infiltration in lung injury models.[82] Such methodologies provide objective, high-resolution readouts for evaluating the therapeutic efficacy of NP candidates, such as quercetin, in alleviating virus-induced pulmonary inflammation. This automation significantly reduces inter-observer variability and ensures the reproducibility of toxicity and efficacy assessments for natural antivirals in preclinical animal trials.

Third, AI-driven predictive modeling, including physiologically based PBPK models augmented with ML, bridges *in vitro* data to *in vivo* outcomes. These models predict pharmacokinetics (*e.g.*, clearance and half-life) for NPs by integrating physicochemical properties and multi-omics data. A notable application involves forecasting human systemic exposure for plant-derived antivirals like berberine. For instance, mechanistic PBPK models have been refined to capture complex interactions between berberine and multiple transporters (*e.g.*, P-gp and OCTs), providing a quantitative framework to evaluate its therapeutic potential against viral infections, particularly in the context of drug–drug interactions (DDI).[83] Furthermore, ML-enhanced PBPK has demonstrated significant superiority, reducing prediction errors (*e.g.*, RMSE) by 20–30% compared to empirical scaling methods, as shown in small-molecule validations.[84] Although primarily validated in targeted therapeutics like oligonucleotides,[85] the transfer learning strategies utilized for cross-species translation are increasingly being adapted to NPs, facilitating more accurate animal-to-human extrapolations for complex natural compounds and offering methodological insights for antiviral drug development.

However, despite these advancements, AI in preclinical testing for antiviral NPs faces significant failure modes. Recent high-profile blind challenges, such as the CACHE series and the ASAP-Polaris initiatives, have exposed a stark reality: the majority of AI-prioritized compounds fail to exhibit reproducible activity in wet-lab assays, with failure rates exceeding 90% in many cases.[86,87] During the COVID-19 pandemic, the "rush to screen" led to an influx of low-quality *in silico* studies where herbal compounds like quercetin were frequently identified as hits. Retrospective analyses now confirm that these were largely false positives—often due to the molecules being Pan-Assay Interference Compounds (PAINS) or AI overestimating binding affinities by neglecting complex solvation effects and structural plasticity.[87] These failures underscore a systemic translational gap, necessitating a shift toward more rigorous, physics-informed AI models for natural products.

2.3.2 AI innovations in clinical trials. Natural product-based antiviral drugs, like other therapeutics, undergo expensive and lengthy clinical trial phases. Artificial intelligence is beginning to offer targeted improvements in efficiency and informativeness, although many applications remain in early stages or face substantial limitations in real-world translation. Several AI-based methodologies are under exploration in Phase I–III studies for antiviral agents:

(i) Federated learning (FL) for multi-center data integration—FL enables collaborative model training across institutions without centralizing sensitive patient data, thereby preserving privacy while leveraging diverse cohorts. A landmark example is its application to predict clinical outcomes in COVID-19 patients from multiple hospitals, demonstrating improved generalizability across heterogeneous populations.[88] Although this approach has not yet been widely reported for antiviral NP trials, it holds methodological promise for modeling disease trajectories or treatment responses in multi-ethnic or multi-risk-group settings, provided data harmonization and model robustness challenges are addressed.

(ii) NLP for unstructured clinical data—NLP algorithms can extract relevant outcomes, adverse events, or symptom patterns from electronic health records, physician notes, and free-text entries in near real-time. Systematic reviews indicate that NLP enhances signal detection in clinical decision support and could support more responsive monitoring in trials.[89] In the context of antiviral NPs, such tools may facilitate earlier identification of efficacy or safety signals in adaptive trial designs, although current implementations are largely limited to general medical contexts rather than NP-specific endpoints.

(iii) Synthetic control arms and generative models—Generative AI methodologies, particularly GANs, are being explored to create synthetic patient data (SPD) to augment or partially replace traditional control arms. Early benchmarks using tabular clinical datasets demonstrate that GAN-based frameworks, such as GANerAid, can synthesize patient-level records that preserve the complex statistical correlations and longitudinal trajectories of actual trial participants.[90] This approach is particularly promising for establishing synthetic control arms in rare disease or oncology trials, where simulating survival data (*e.g.*, progression-free survival) can reduce the reliance on large placebo cohorts while maintaining statistical power.[91] However, applications to antiviral NPs remain nascent, with significant challenges in ensuring biological plausibility—the risk that GANs may generate pharmacologically impossible trajectories for multi-component natural extracts. Furthermore, regulatory acceptance requires rigorous validation against "hallucinated"

correlations and the avoidance of bias amplification in complex patient profiles.

3. Bottlenecks in AI-enabled antiviral NP development

3.1 Data constraints: NP complexity and viral dynamics

Data scarcity and heterogeneity are the key obstacles. NP data are multi-modal, complex, and unevenly standardized across sources and formats.[92] High-quality, empirically validated annotations for structures, biosynthetic processes, antiviral activity (including negatives), and toxicity are limited. Multi-target/network pharmacology mechanisms are tough to capture adequately, hindering mechanistic modeling. Knowledge integrated in traditional medicine literature is valuable; however, it is unstructured, archaic, and philosophically complex, challenging NLP and knowledge-graph development.[93]

Viral evolution data are limited in sample size and dynamic. For new strains or unexpected resistance mutations, establishing robust models is challenging.[94] Rapid genomic change needs frequent model updating.[46] Many viruses replicate within liquid–liquid phase-separated (LLPS) condensates, which lack well-defined structural targets, complicating design; changing sequences may affect condensate physicochemical features and dynamics.[95] Host-target inhibition may generate compensating responses that current AI can rarely foresee.[49] Addressing such "moving targets" requires models that spot dynamic or allosteric sites or even intervene in phase behavior, potentially *via* system-level "digital twins"—well beyond existing capabilities.

3.2 Model limitations: from pattern recognition to biological understanding

AI's advancements primarily rely on pattern recognition, but deep biological and chemical understanding for intricate NP structures and polypharmacology remains inadequate. NPs feature intricate ring structures, multiple chiral centers, and flexible conformations; capturing fine-grained conformational dynamics and flexible protein–ligand interactions is difficult.[76] Even when employing AlphaFold 3, difficulties exist in modeling dynamics, allostery, and binding of unusual ligands, which can affect docking accuracy.[96] NPs frequently act *via* several viral and host targets, forming complex networks; current AI struggles to interpret synergy or antagonism or predict system-level ramifications. Multi-component traditional formulations confront additional "black-box" issues that limit optimization and sensible combination design.[97] For swiftly evolving viruses, generalization to out-of-distribution variations is limited; even single-point mutations might have structural ramifications that models fail to capture.[98]

3.3 Synthetic accessibility

Bridging computational hits to physical molecules remains a formidable hurdle for NPs. While NPs benefit from innate bioactivity and initial accessibility through extraction or fermentation, scaling their production to meet clinical or industrial demands reveals significant bottlenecks. Direct large-scale extraction from natural sources is often unsustainable due to resource scarcity, seasonal variability, environmental degradation (*e.g.*, overharvesting), and extremely low yields (frequently below 0.01% dry weight).[6] Similarly, native microbial fermentation is hampered by the low productivity of wild-type strains and difficult-to-control fermentation conditions, although modern strain improvement and optimization techniques (*e.g.*, fed-batch or continuous fermentation) can enhance productivity.[99]

Heterologous biosynthesis has emerged as a transformative strategy, but metabolic engineering still faces a "scale-up gap," where titer optimization often stalls at the mg L^{-1} level, failing to reach the g L^{-1} threshold required for commercial viability.[100] For molecules with extreme structural complexity, semi-synthesis offers a middle ground, yet it remains constrained by limited reagent-accessible space and high operation costs.[101]

4. Outlook and future directions

4.1 High-quality, multi-modal benchmark datasets

To overcome data scarcity and fragmented evaluation, the field needs FAIR (Findable, Accessible, Interoperable, and Reusable), high-quality, multi-modal benchmark datasets. These should integrate NP structures, biosynthetic pathways, multi-dimensional bioactivity profiles (including negative results), and ADMET properties, along with standardized metrics for fair model comparison and industrial translation.

4.2 Model and algorithmic innovation

Beyond correlational prediction, future AI should prioritize interpretability (*e.g.*, explainable AI, XAI) and robust out-of-distribution generalization to novel biological systems (*e.g.*, new variants). Incorporating causal inference and neuro-symbolic techniques may facilitate a move from pattern recognition to scientific discovery, yielding better mechanistic understanding.

4.3 Stronger experimental closed loops and high-throughput validation

To match AI's high-throughput hypothesis development, experimental validation must accelerate. Automated platforms (self-driving laboratories) and active learning can connect *in silico* forecasts tightly with wet-lab feedback during DBTL cycles, eliminating the computational–experimental gap.

4.4 Cross-disciplinary collaboration and ecosystem building

Antiviral NP discovery needs fundamental integration across chemistry, biology, medicine, and data science. Training cross-domain competence and promoting academia–industry–international partnerships will enable secure data exchange (*e.g.*, privacy-preserving federated learning). Building an automated and intelligent end-to-end pipeline—from data collection and modeling to candidate generation and validation—will require continual investment but is essential to realize AI's full promise in antiviral NP discovery.

5. Conclusion

Antiviral drug research today faces significant difficulties: it is expensive, time-consuming, and limited in efficacy against rapidly evolving viruses. NPs are still hindered in their analysis, target detection, and synthesis because of their structural diversity and multi-target synergy. AI is a powerful new paradigm that is changing how the drug discovery process is conducted through the integration of data and algorithmic learning.

In the upstream phase, AI helps to refine genome mining and the modeling of biosynthetic gene clusters, while knowledge graphs and natural language processing aid in extracting insight from conventional medical databases. AI-driven virtual screening and generative design during the midstream phase significantly expand chemical space exploration, yielding multi-target antiviral candidates with optimal pharmacokinetics and resilience against resistance. In the downstream phase, AI optimizes preclinical models and clinical trial efficiency using techniques such as federated learning and synthetic control arms, accelerating translational outcomes.

There are many critical issues, including data scarcity, limited model interpretability, and synthetic accessibility challenges, necessitating integrated solutions to advance the field. The complexity of NPs and viral dynamics demands FAIR, multi-modal benchmark datasets that unify structural, biosynthetic, and bioactivity data, while interpretable AI and causal inference are critical to move beyond pattern recognition toward mechanistic understanding of polypharmacology and dynamic targets like liquid–liquid phase-separated condensates. Automated platforms, such as self-driving laboratories, must bridge the computational–experimental gap through high-throughput design–build–test–learn cycles, embedding synthetic feasibility early in the design process. By fostering cross-disciplinary collaboration and privacy-preserving data-sharing ecosystems, the field can build an intelligent, end-to-end pipeline, transforming reactive antiviral NP discovery into a proactive, resilient strategy against evolving viral threats.

Although not a cure-all, AI represents the best catalyst for converting NP-based antiviral discovery from empirical serendipity to predictive engineering. If adopted more widely—as long as there's solid validation and international data-sharing agreements—the virus's arsenal of antiviral responses may be unleashed. We are on the cusp of the fourth drug discovery revolution enabled by symbiotic human–AI collaboration, where the speed of machines meets the smarts of chemical evolution to move beyond the capacity for viral adaptation.

Author contributions

Conceptualization: Jianjian Ji, Peng Cao, Liangyu Cai, Junxi Song, Kunhuan Yang, Yingcai Xiong, and Keyu Tao. Methodology: Junxi Song, Kunhuan Yang, Yingcai Xiong, and Keyu Tao. Investigation: Junxi Song, Kunhuan Yang, Yingcai Xiong, and Keyu Tao. Writing—original draft preparation: Jianjian Ji, Peng Cao, Liangyu Cai, Junxi Song, Kunhuan Yang, Yingcai Xiong, and Keyu Tao. Visualization: Junxi Song, Kunhuan Yang, Yingcai Xiong, and Keyu Tao. Writing—review & editing: Jianjian Ji, Peng Cao, and Liangyu Cai. Funding acquisition: Jianjian Ji, Peng Cao, and Liangyu Cai. Supervision: Jianjian Ji, Peng Cao, and Liangyu Cai. Project administration: Jianjian Ji, Peng Cao, and Liangyu Cai.

Conflicts of interest

There are no conflicts to declare.

Data availability

This article is a review, and as such, does not report any new primary data. All data and information discussed are available in the original publications and sources cited within the manuscript's references section.

Acknowledgements

We would like to express our profound gratitude to all participants for their invaluable contributions to this research. Additionally, we appreciate the assistance of https://www.BioRender.com in creating the figures for this study. This work was supported by The Key Research Project of Jiangsu Provincial Academy of Chinese Medicine Schools (LPZD2025012).

References

1 Z. Abdelrahman, M. Li and X. Wang, Comparative Review of SARS-CoV-2, SARS-CoV, MERS-CoV, and Influenza A Respiratory Viruses, *Front Immunol*, 2020, **11**, 552909.
2 J. A. DiMasi, H. G. Grabowski and R. W. Hansen, Innovation in the pharmaceutical industry: New estimates of R&D costs, *J Health Econ*, 2016, **47**, 20–33.
3 R. T. Eastman, J. S. Roth, K. R. Brimacombe, A. Simeonov, M. Shen, S. Patnaik, *et al.*, Remdesivir: A Review of Its Discovery and Development Leading to Emergency Use Authorization for Treatment of COVID-19, *ACS Cent. Sci.*, 2020, **6**(5), 672–683.
4 T. Cihlar and R. L. Mackman, Journey of remdesivir from the inhibition of hepatitis C virus to the treatment of COVID-19, *Antivir. Ther.*, 2022, **27**(2), 13596535221082773.
5 T. Guarnaccia, L. A. Carolan, S. Maurer-Stroh, R. T. Lee, E. Job, P. C. Reading, *et al.*, Antigenic drift of the pandemic 2009 A(H1N1) influenza virus in A ferret model, *PLoS Pathog.*, 2013, **9**(5), e1003354.
6 D. J. Newman and G. M. Cragg, Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019, *J. Nat. Prod.*, 2020, **83**(3), 770–803.
7 L. H. Miller and X. Su, Artemisinin: discovery from the Chinese herbal garden, *Cell*, 2011, **146**(6), 855–858.
8 J. Y. Li, H. Y. Cao, P. Liu, G. H. Cheng and M. Y. Sun, Glycyrrhizic acid in the treatment of liver diseases: literature review, *Biomed Res Int*, 2014, **2014**, 872139.
9 E. De Clercq, Antiviral drugs in current clinical use, *J Clin Virol*, 2004, **30**(2), 115–133.

10 M. S. Rappé and S. J. Giovannoni, The uncultured microbial majority, *Annu. Rev. Microbiol.*, 2003, **57**, 369–394.
11 T. Liu and S. Lin, Comprehensive characterization of the chemical constituents of Lianhua Qingwen capsule by ultra high performance liquid chromatography coupled with Fourier transform ion cyclotron resonance mass spectrometry, *Heliyon*, 2024, **10**(6), e27352.
12 J. Zhang, L. G. Hansen, O. Gudich, K. Viehrig, L. M. M. Lassen, L. Schruebbers, *et al.*, A microbial supply chain for production of the anti-cancer drug vinblastine, *Nature*, 2022, **609**(7926), 341–347.
13 Z. Q. Zhu, X. Zheng, G. Q. Qi, Y. F. Gong, Y. Y. Li, N. Mazur, *et al.*, Drug-target binding affinity prediction model based on multi-scale diffusion and interactive learning, *Expert Syst. Appl.*, 2024, **255**, 124647.
14 B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, *et al.*, The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods, *Nucleic Acids Res.*, 2023, **52**(D1), D1180–D1192.
15 J. Ru, P. Li, J. Wang, W. Zhou, B. Li, C. Huang, *et al.*, TCMSP: a database of systems pharmacology for drug discovery from herbal medicines, *J. Cheminf.*, 2014, **6**(1), 13.
16 R. D. Darmawan, W. A. Kusuma and H. Rahmawan, Deep learning optimization for drug-target interaction prediction in COVID-19 using graphic processing unit, *Int. J. Electr. Comput. Eng.*, 2023, **13**(3), 3111–3123.
17 Y. Chen, D. Luo and W. Xue, Deep Learning for Drug–Target Interaction Prediction: A Comprehensive Review, *Chem. Biol. Drug Des.*, 2025, **106**(4), e70183.
18 Y. Wang, A. Ünlü, X. Wang, E. Çevrim, D. M. Offermans, M. P. Flesseman, *et al.*, AI-driven discovery of antiretroviral drug bictegravir and etravirine as inhibitors against monkeypox and related poxviruses, *Commun. Biol.*, 2025, **8**(1), 1734.
19 A. K. Mishra, N. Sudalaimuthuasari, K. M. Hazzouri, E. E. Saeed, I. Shah and K. M. A. Amiri, Tapping into Plant-Microbiome Interactions through the Lens of Multi-O mics Techniques, *Cells*, 2022, **11**(20), 3254.
20 J. Zhang, B. Li, Y. Qin, L. Karthik, G. Zhu, C. Hou, *et al.*, A new abyssomicin polyketide with anti-influenza A virus activity from a marine-derived Verrucosispora sp. MS100137, *Appl. Microbiol. Biotechnol.*, 2020, **104**(4), 1533–1543.
21 H. Li, L. Yang, F.-F. Liu, X.-N. Ma, P.-L. He, W. Tang, *et al.*, Overview of therapeutic drug research for COVID-19 in China, *Acta Pharmacol. Sin.*, 2020, **41**(9), 1133–1140.
22 F. Ge, Y. Yang, Z. Bai, L. Si, X. Wang, J. Yu, *et al.*, The role of Traditional Chinese medicine in anti-HBV: background, prog ress, and challenges, *Chin Med*, 2023, **18**(1), 159.
23 L. Ma, L. Ji, T. Wang, Z. Zhai, P. Su, Y. Zhang, *et al.*, Research progress on the mechanism of traditional Chinese medicine reg ulating intestinal microbiota to combat influenza a virus infection, *Virol J*, 2023, **20**(1), 260.
24 M. H. Medema, K. Blin, P. Cimermancic, V. de Jager, P. Zakrzewski, M. A. Fischbach, *et al.*, antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome s equences, *Nucleic Acids Res.*, 2011, **39**(Web Server issue), W339–W346.
25 G. D. Hannigan, D. Prihoda, A. Palicka, J. Soukup, O. Klempir, L. Rampula, *et al.*, A deep learning genome-mining strategy for biosynthetic gene cluster prediction, *Nucleic Acids Res.*, 2019, **47**(18), e110.
26 Y. Zhang, J. Zhang, M. Li, Y. Qiao, W. Wang, L. Ma, *et al.*, Target discovery of bioactive natural products with native-compound-co upled CNBr-activated Sepharose 4B beads (NCCB): Applications, mechanisms and outlooks, *Bioorg. Med. Chem.*, 2023, **96**, 117483.
27 Y. Xu, L. Cao, Y. Chen, Z. Zhang, W. Liu, H. Li, *et al.*, Integrating Machine Learning in Metabolomics: A Path to Enhanced Diagnostics and Data Interpretation, *Small Methods*, 2024, **8**(12), e2400305.
28 Z. Zhang, H. Yang, Y. Wang, L. Zhang and S. H. Lin, QuanFormer: A Transformer-Based Precise Peak Detection and Quantification Tool in LC-MS-Based Metabolomics, *Anal. Chem.*, 2025, **97**(5), 2698–2706.
29 G. Hu and M. Qiu, Machine learning-assisted structure annotation of natural products based on MS and NMR data, *Nat. Prod. Rep.*, 2023, **40**(11), 1735–1753.
30 R. Schmid, S. Heuckeroth, A. Korf, A. Smirnov, O. Myers, T. S. Dyrlund, *et al.*, Integrative analysis of multimodal mass spectrometry data in MZmine 3, *Nature biotechnology*, 2023, **41**(4), 447–449.
31 B. K. Chhetri, P. R. Tedbury, A. M. Sweeney-Jones, L. Mani, K. Soapi, C. Manfredi, *et al.*, Marine natural products as leads against SARS-CoV-2 infection, *J. Nat. Prod.*, 2022, **85**(3), 657–665.
32 J. G. Song, W. C. Ye and Y. Wang, Advanced crystallography for structure determination of natural products, *Nat. Prod. Rep.*, 2025, **42**(3), 429–442.
33 S. A. Dymura, O. O. Viniichuk, K. P. Melnykov, D. S. Radchenko and O. O. Grygorenko, Machine Learning-Based Retention Time Prediction Tool for Routine LC-MS Data Analysis, *J. Chem. Inf. Model.*, 2025, **65**(14), 7415–7425.
34 Y. Liu, A. C. Yoshizawa, Y. Ling and S. Okuda, Insights into predicting small molecule retention times in liquid chromatography using deep learning, *J Cheminform*, 2024, **16**(1), 113.
35 D. Song, T. Tang, R. Wang, H. Liu, D. Xie, B. Zhao, *et al.*, Enhancing compound confidence in suspect and non-target screening through machine learning-based retention time prediction, *Environ. Pollut.*, 2024, **347**, 123763.
36 C. Peng, F. Xia, M. Naseriparsa and F. Osborne, Knowledge Graphs: Opportunities and Challenges, *Artif Intell Rev*, 2023, **56**(11), 13071–13102.
37 A. Gaudry, M. Pagni, F. Mehl, S. Moretti, L.-M. Quiros-Guerrero and L. Cappelletti, *et al. A sample-centric and knowledge-driven computational framework for natural products drug discovery*. ACS Publications; 2024.
38 T. F. Leão, M. Wang, R. da Silva, A. Gurevich, A. Bauermeister, P. W. P. Gomes, *et al.*, NPOmix:

a machine learning classifier to connect mass spectrometry fragmentation data to biosynthetic gene clusters, *PNAS Nexus*, 2022, **1**(5), pgac257.

39 Q. Lv, G. Chen, H. He, Z. Yang, L. Zhao, H.-Y. Chen, *et al.*, TCMBank: bridges between the largest herbal medicines, chemical ingredients, target proteins, and associated diseases with intelligence text mining, *Chem. Sci.*, 2023, **14**(39), 10684–10701.

40 S. Galati, M. Di Stefano, E. Martinelli, G. Poli and T. Tuccinardi, Recent Advances in *In Silico* Target Fishing, *Molecules*, 2021, **26**(17), 5124.

41 S. Sivanandan, B. Leitmann, E. Lubeck, M. M. Sultan, P. Stanitsas, N. Ranu, *et al.*, A pooled cell painting CRISPR screening platform enables *de novo* inference of gene function by self-supervised deep learning, *Nat. Commun.*, 2026, **17**, 77.

42 N. T. Cockroft, X. Cheng and J. R. Fuchs, STarFish: A Stacked Ensemble Target Fishing Approach and its Application to Natural Products, *J. Chem. Inf. Model.*, 2019, **59**(11), 4906–4920.

43 K. Huang, T. Fu, L. M. Glass, M. Zitnik, C. Xiao and J. Sun, DeepPurpose: a deep learning library for drug–target interaction prediction, *Bioinformatics*, 2020, **36**(22–23), 5545–5547.

44 P. Kundu, S. Beura, S. Mondal, A. K. Das and A. Ghosh, Machine learning for the advancement of genome-scale metabolic modeling, *Biotechnol. Adv.*, 2024, **74**, 108400.

45 K. Hsieh, Y. Wang, L. Chen, Z. Zhao, S. Savitz, X. Jiang, *et al.*, Drug repurposing for COVID-19 using graph neural network with genetic, mechanistic, and epidemiological validation, *Research square*, 2020, (3), 114758.

46 S. Lytras, K. D. Lamb, J. Ito, J. Grove, K. Yuan, K. Sato, *et al.*, Pathogen genomic surveillance and the AI revolution, *J. Virol.*, 2025, **99**(2), e0160124.

47 J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3, *Nature*, 2024, **630**(8016), 493–500.

48 R. Krishna, J. Wang, W. Ahern, P. Sturmfels, P. Venkatesh, I. Kalvet, *et al.*, Generalized biomolecular modeling and design with RoseTTAFold All-Atom, *Science*, 2024, **384**(6693), eadl2528.

49 M. Abdelsayed, AI-Driven Polypharmacology in Small-Molecule Drug Discovery, *Int. J. Mol. Sci.*, 2025, **26**(14), 6996.

50 A. Bender and I. Cortes-Ciriano, Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data, *Drug Discov Today*, 2021, **26**(4), 1040–1052.

51 J. Deng, Z. Yang, H. Wang, I. Ojima, D. Samaras and F. Wang, A systematic study of key elements underlying molecular property prediction, *Nat. Commun.*, 2023, **14**(1), 6395.

52 T. Wen, X. Cai and J. Li, Graph Neural Networks *vs.* Traditional QSAR: A Comprehensive Comparison for Multi-Label Molecular Odor Prediction, *Molecules*, 2025, **30**(23), 4605.

53 H. Cai, H. Zhang, D. Zhao, J. Wu and L. Wang, FP-GNN: a versatile deep learning architecture for enhanced molecular property prediction, *Brief Bioinform*, 2022, **23**(6), bbac408.

54 M. Praski, J. Adamczyk and W. Czech. Benchmarking pretrained molecular embedding models for molecular representation learning. *arXiv:250806199*. 2025.

55 M. R. Dobbelaere, I. Lengyel, C. V. Stevens and K. M. Van Geem, Geometric deep learning for molecular property predictions with chemical accuracy across chemical space, *J. Cheminf.*, 2024, **16**(1), 99.

56 J. Wu, H. Chen, M. Cheng and H. Xiong, CurvAGN: Curvature-based Adaptive Graph Neural Networks for Predicting Protein-Ligand Binding Affinity, *BMC Bioinf.*, 2023, **24**(1), 378.

57 F. Gentile, V. Agrawal, M. Hsing, A.-T. Ton, F. Ban, U. Norinder, *et al.*, Deep docking: a deep learning platform for augmentation of structure based drug discovery, *ACS central science*, 2020, **6**(6), 939–949.

58 D. E. Graff, E. I. Shakhnovich and C. W. Coley, Accelerating high-throughput virtual screening through molecular pool-based active learning, *Chem. Sci.*, 2021, **12**(22), 7866–7881.

59 A. Luttens, I. Cabeza de Vaca, L. Sparring, J. Brea, A. L. Martínez, N. A. Kahlous, *et al.*, Rapid traversal of vast chemical space using machine learning-guided docking screens, *Nat. Comput. Sci.*, 2025, 1–12.

60 L.-C. Zhang, H.-L. Zhao, J. Liu, L. He, R.-L. Yu and C.-M. Kang, Design of SARS-CoV-2 Mpro, PLpro dual-target inhibitors based on deep reinforcement learning and virtual screening, *Future Med. Chem.*, 2022, **14**(6), 393–405.

61 European Medicines Agency, in *Review of artificial intelligence and machine learning applications in medicines lifecycle (2024): Horizon Scanning Short Report*, ed. Agency E. M., Amsterdam, Netherlands, European Medicines Agency, 2025.

62 J. Zhang, H. Li, Y. Zhang, J. Huang, L. Ren, C. Zhang, *et al.*, Computational toxicology in drug discovery: applications of artificial intelligence in ADMET and toxicity prediction, *Briefings Bioinf.*, 2025, **26**(5), bbaf533.

63 H. Lee, J. Kim, J.-W. Kim and Y. Lee, Recent advances in AI-based toxicity prediction for drug discovery, *Front. Chem.*, 2025, **13**, 1632046.

64 A. Lavecchia, Explainable artificial intelligence in drug discovery: bridging predictive power and mechanistic insight, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2025, **15**(5), e70049.

65 I. Pathan, A. Raza, A. Sahu, M. Joshi, Y. Sahu, Y. Patil, *et al.*, Revolutionizing pharmacology: AI-powered approaches in molecular modeling and ADMET prediction, *Med. Drug Discovery*, 2025, 100223.

66 Q. Wang, B. Sun, Y. Yi, T. Velkov, J. Shen, C. Dai, *et al.*, Progress of AI-driven drug–target interaction prediction and lead optimization, *Int. J. Mol. Sci.*, 2025, **26**(20), 10037.

67 F. J. Ferreira and A. S. Carneiro, AI-Driven Drug Discovery: A Comprehensive Review, *ACS Omega*, 2025, **10**(23), 23889–23903.

68 L. Fu, S. Shi, J. Yi, N. Wang, Y. He, Z. Wu, *et al.*, ADMETlab 3.0: an updated comprehensive online ADMET prediction platform enhanced with broader coverage, improved performance, API functionality and decision support, *Nucleic Acids Res.*, 2024, **52**(W1), W422–W431.
69 Y. Myung, A. G. de Sá and D. B. Ascher, Deep-PK: deep learning for small molecule pharmacokinetic and toxicity prediction, *Nucleic Acids Res.*, 2024, **52**(W1), W469–W475.
70 N. Youssef, S. Gurev, F. Ghantous, K. P. Brock, J. A. Jaimes, N. N. Thadani, *et al.*, Computationally designed proteins mimic antibody immune evasion in viral evolution, *Immunity*, 2025, **58**(6), 1411–21.e6.
71 K. I. Sahibzada, S. Shahid, M. Akhter, R. Abid, M. Azhar, Y. Hu, *et al.*, HIV OctaScanner: A Machine Learning Approach to Unveil Proteolytic Cleavage Dynamics in HIV-1 Protease Substrates, *J. Chem. Inf. Model.*, 2025, **65**(2), 640–648.
72 C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay and K. F. Jensen, Generative models for molecular discovery: Recent advances and challenges, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2022, **12**(5), e1608.
73 A. G. Atanasov, S. B. Zotchev, V. M. Dirsch and C. T. Supuran, Natural products in drug discovery: advances and opportunities, *Nat. Rev. Drug Discovery*, 2021, **20**(3), 200–216.
74 A. Nigam, R. Pollice and A. Aspuru-Guzik, Parallel tempered genetic algorithm guided by deep neural networks for inverse molecular design, *Digital Discovery*, 2022, **1**(4), 390–404.
75 J. Meyers, B. Fabian and N. Brown, De novo molecular design and generative models, *Drug discovery today*, 2021, **26**(11), 2707–2715.
76 T. Ochiai, T. Inukai, M. Akiyama, K. Furui, M. Ohue, N. Matsumori, *et al.*, Variational autoencoder-based chemical latent space for large molecular structures with 3D complexity, *Commun. Chem.*, 2023, **6**(1), 249.
77 Hoogeboom E., Satorras V. G., Vignac C. and Welling M., *Equivariant diffusion for molecule generation in 3d. International conference on machine learning*, PMLR, 2022.
78 M. Parrot, H. Tajmouati, V. B. R. da Silva, B. R. Atwood, R. Fourcade, Y. Gaston-Mathé, *et al.*, Integrating synthetic accessibility with AI-based generative drug design, *J. Cheminf.*, 2023, **15**(1), 83.
79 A. M. Westerlund, L. Saigiridharan and S. Genheden, Human-guided synthesis planning *via* prompting, *Chem. Sci.*, 2025, **16**(32), 14655–14667.
80 T.-Z. Long, D.-J. Jiang, S.-H. Shi, Y.-C. Deng, W.-X. Wang and D.-S. Cao, Enhancing Multi-species Liver Microsomal Stability Prediction through Artificial Intelligence, *J. Chem. Inf. Model.*, 2024, **64**(8), 3222–3236.
81 D. Yan, M. Zhou, A. Adduri, Y. Zhuang, M. Guler, S. Liu, *et al.*, Discovering type I cis-AT polyketides through computational mass spectrometry and genome mining with Seq2PKS, *Nat. Commun.*, 2024, **15**(1), 5356.
82 Z. Gu, S. Wang, R. Rong, Z. Zhao, F. Wu, Q. Zhou, *et al.*, Cell segmentation with globally optimized boundaries (csgo): A deep learning pipeline for whole-cell segmentation in hematoxylin-and-eosin–stained tissues, *Lab. Invest.*, 2025, **105**(2), 102184.
83 J. Adiwidjaja, A. V. Boddy and A. J. McLachlan, Physiologically based pharmacokinetic model predictions of natural product-drug interactions between goldenseal, berberine, imatinib and bosutinib, *Eur. J. Clin. Pharmacol.*, 2022, **78**(4), 597–611.
84 Y. Li, Z. Wang, Y. Li, J. Du, X. Gao, Y. Li, *et al.*, A combination of machine learning and PBPK modeling approach for pharmacokinetics prediction of small molecules in humans, *Pharm. Res.*, 2024, **41**(7), 1369–1379.
85 A. Derbalah, F. Stader, C. Liu, A. Zyla, T. Abdulla, Q. Wu, *et al.*, Cross-species translational modelling of targeted therapeutic oligonucleotides using physiologically based pharmacokinetics, *J. Pharmacokinet. Pharmacodyn.*, 2025, **52**(4), 35.
86 S. Ackloo, R. Al-Awar and R. E. Amaro, CACHE (Critical Assessment of Computational Hit-finding Experiments): A public–private partnership benchmarking initiative to enable the development of computational methods for hit-finding, *Nat. Rev. Chem.*, 2022, **6**, 287–295.
87 H. MacDermott-Opeskin, J. Scheen and C. Wognum, A computational community blind challenge on pan-coronavirus drug discovery data, *J. Chem. Inf. Model.*, 2026, **66**(6), 3129–3149.
88 I. Dayan, H. R. Roth, A. Zhong, A. Harouni, A. Gentili, A. Z. Abidin, *et al.*, Federated learning for predicting clinical outcomes in patients with COVID-19, *Nat. Med.*, 2021, **27**(10), 1735–1743.
89 H. Eguia, C. L. Sánchez-Bocanegra, F. Vinciarelli, F. Alvarez-Lopez and F. Saigí-Rubió, Clinical decision support and natural language processing in medicine: systematic literature review, *J. Med. Internet Res.*, 2024, **26**, e55315.
90 L. Krenmayr, R. Frank, C. Drobig, M. Braungart, J. Seidel, D. Schaudt, *et al.*, GANerAid: realistic synthetic patient data for clinical trials, *Inform. Med. Unlocked*, 2022, **35**, 101118.
91 I. Akiya, T. Ishihara and K. Yamamoto, Comparison of synthetic data generation techniques for control group survival data in oncology clinical trials: simulation study, *JMIR Med. Inform*, 2024, **12**(1), e55118.
92 M. W. Mullowney, K. R. Duncan, S. S. Elsayed, N. Garg, J. J. J. van der Hooft, N. I. Martin, *et al.*, Artificial intelligence for natural product drug discovery, *Nat Rev Drug Discov*, 2023, **22**(11), 895–916.
93 L. Zhou, S. Liu, C. Li, Y. Sun, Y. Zhang, Y. Li, *et al.*, Natural Language Processing Algorithms for Normalizing Expressions of Synonymous Symptoms in Traditional Chinese Medicine, *Evid.-based Complement. Altern. Med.*, 2021, **2021**, 6676607.
94 S. Mallapaty, What will viruses do next? AI is helping scientists predict their evolution, *Nature*, 2025, **637**(8046), 527–528.
95 M. Galloux and S. Longhi, Unraveling Liquid-Liquid Phase Separation (LLPS) in Viral Infections to Understand and Treat Viral Diseases, *Int. J. Mol. Sci.*, 2024, **25**(13), 6981.

96 S. Y. Shen, J. R. Li, Y. S. Wang, S. N. Li, H. E. Xu and X. H. He, An update for AlphaFold3 *versus* experimental structures: assessing the precision of small molecule binding in GPCRs, *Acta Pharmacol. Sin.*, 2025, 1–10.
97 L. Yang, H. Wang, Z. Zhu, Y. Yang, Y. Xiong, X. Cui, *et al.*, Network Pharmacology-Driven Sustainability: AI and Multi-Omics Synergy for Drug Discovery in Traditional Chinese Medicine, *Pharmaceuticals*, 2025, **18**(7), 1074.
98 J. Wee and G. W. Wei, Rapid response to fast viral evolution using AlphaFold 3-assisted topological deep learning, *Virus Evol*, 2025, **11**(1), veaf026.
99 X. Liu, W. Ding and H. Jiang, Engineering microbial cell factories for the production of plant natural products: from design principles to industrial-scale production, *Microb. Cell Fact.*, 2017, **16**(1), 125.
100 H. Zhou, H. Eun and S. Y. Lee, Systems metabolic engineering for the production of pharmaceutical natural products, *Curr. Opin. Syst. Biol.*, 2024, **37**, 100491.
101 R. A. Shenvi, Natural product synthesis in the 21st century: Beyond the mountain top, *ACS Cent. Sci.*, 2024, **10**(3), 519–528.