Natural Product Reports

ROYAL SOCIETY OF CHEMISTRY

REVIEW

View Article Online
View Journal

Cite this: DOI: 10.1039/d6np00002a

Chemical language models for natural product discovery

Koh Sakano, Kairi Furui, Apakorn Kengkanna, Yuta Kikuchi and Masahito Ohue *

Covering: up to 2026

Natural products are an important source of medicines, yet their discovery can be a slow and laborious process. The recent development of chemical language models (CLMs), which process string-based molecular representations, is reshaping the field of natural product science. This review provides an overview of the role of CLMs in natural product drug discovery, tracing their evolution from early neural networks to modern large-scale Transformers. We describe how these models accelerate discovery timelines by predicting bioactivity, biosynthetic pathways, and spectral data. Furthermore, we cover their use in proposing novel, natural-product-like scaffolds that expand the computationally explored chemical space. The review also addresses persistent challenges, including the limited availability of natural product data and the need for model interpretability. Finally, we discuss future directions, outlining the current status and prospects for CLM-enabled natural product science.

Received 12th January 2026

DOI: 10.1039/d6np00002a

rsc.li/npr

Department of Computer Science, School of Computing, Institute of Science, Tokyo. E-mail: ohue@comp.isct.ac.jp

1 Introduction

Natural products, owing to their unique diversity of chemical architectures and distinctive biological activities, have long underpinned medicinal and agrochemical discovery as one of the most important sources for screening. Indeed, among the small-molecule drugs approved between 1981 and 2019, 32% are natural products or their derivatives, and when restricted to anti-infectives, the proportion exceeds one half.[1] Many breakthrough therapeutics, exemplified by penicillins, originate from natural products that bear complex ring systems and numerous stereocenters, which are difficult to access by synthetic chemistry. Such distinctive architectures continue to serve as fertile sources of novel drug seeds.[2] However, conventional natural

product discovery still faces many challenges. The sequential processes from microbial isolation and cultivation, isolation of bioactive compounds, to complex structure determination require considerable time and effort.[3] Consequently, there is a growing need for new technologies that can reshape traditional approaches in natural products chemistry, efficiently explore uncharted chemical space, and accelerate the drug discovery pipeline.

As a promising response to these challenges, machine learning (ML) and deep learning (DL) have recently come to the fore and are transforming pharmaceutical research. While traditionally experts would select molecular descriptors and analyze their relationships with activity, deep learning has enabled the automatic discovery of features important for activity directly from molecular structure data itself. Among these methods, chemical language models (CLMs), which treat molecular structures as a form of language and apply techniques from natural language processing (NLP), open new avenues for exploring chemical space. CLMs learn from one-dimensional molecular representations such as SMILES strings by interpreting them analogously to sentences composed of tokens. In doing so, they can infer the grammar of chemistry (rules governing atomic connectivity) and the context (properties arising from combinations of functional groups) directly from data, and they are capable of predicting biological activities and physicochemical properties with high accuracy.[4] Architecturally, CLMs have evolved from early sequence models such as recurrent neural networks (RNNs)[5] to Transformers equipped with self-attention mechanisms that capture global

Koh Sakano

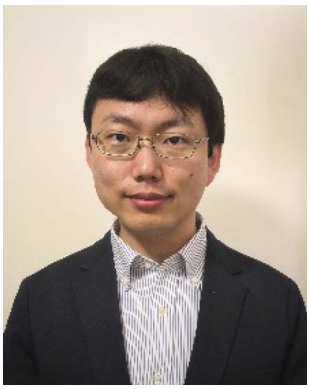
Koh Sakano is a PhD student in the Ohue Lab at the Department of Computer Science, School of Computing, Institute of Science Tokyo, and a JSPS Research Fellow (DC1). His research focuses on chemical language models and generative modeling for molecular design, with particular interest in natural product-inspired compound generation and data-efficient learning. He works on representation learning for molecular strings to support exploration of natural-product-like chemical space.

Kairi Furui

Kairi Furui is a PhD student in the Ohue Lab at the Department of Computer Science, School of Computing, Institute of Science Tokyo, and a JSPS Research Fellow (DC1). He is interested in AI and computer simulation techniques for designing antibodies and small-molecule drugs. His research explores how language models and molecular simulations can be combined to support optimization and prioritization in drug discovery.

Apakorn Kengkanna

Apakorn Kengkanna is a PhD student in the Ohue Lab at the Department of Computer Science, School of Computing, Institute of Science Tokyo, supported by a Japanese Government (MEXT) Scholarship. He is interested in AI-driven prediction of molecular properties and catalyst design for chemical reactions. His work focuses on machine learning models for structure–property relationships and reaction-related tasks.



Yuta Kikuchi

Yuta Kikuchi received his PhD in Life Science from Kitasato University in 2024 and was appointed as a Specially Appointed Assistant Professor at the Graduate School of Infection Control Science, Kitasato University, where he conducted research on microbial drug discovery at the Ōmura Satoshi Memorial Institute. Since 2025, he has been a Specially Appointed Assistant Professor at the School of Computing, Institute of Science Tokyo. His research focuses on natural product drug discovery by integrating cheminformatics and bioinformatics with experimental natural product chemistry.

Masahito Ohue

Masahito Ohue received his PhD in Computer Science from Tokyo Institute of Technology in 2014 and is now an Associate Professor at the School of Computing, Institute of Science Tokyo. His research integrates bioinformatics and supercomputing, particularly computational methods for predicting protein–protein/ligand/peptide/antibody interactions. He has been a JST FOREST researcher since 2022 and has received awards including the JSPS Ikushi Prize (2014), the MEXT Young Scientist Award (2019), and the IPSJ Microsoft Faculty Award (2024).

context.[6] This has led to the construction of foundation models[7] that have learned chemical knowledge from large-scale known compound datasets. They can achieve high performance on specific chemical tasks through fine-tuning that adapts the model to particular downstream tasks.

In this review, we focus specifically on language models among the many AI techniques and provide a comprehensive overview from their theoretical underpinnings to applications in natural product discovery. Whereas existing reviews of AI in natural products research often survey AI/ML methods broadly covering QSAR, docking, virtual screening, omics integration, and synthetic-biology workflows and tend to treat language models as just one among many tools, our article is differentiated by making language models the central theme. We provide an in-depth technical exploration of their principles, architectures, molecular representations, and tokenization methods, while systematically discussing language models' applications specifically tailored to natural products. Specifically, we first explain how language models operate and how CLMs learn molecular structures to function as predictors of biological activity, target proteins, and ADMET properties. We then synthesize cross-disciplinary applications, including predicting the structures and functions of compounds produced from biosynthetic gene cluster (BGC) sequences and annotating structures directly from spectral data such as mass spectrometry (MS) and nuclear magnetic resonance (NMR). In parallel, we discuss the potential of CLMs as generative models for *de novo* compound design, assessing both their strengths and limitations. Finally, we highlight practical challenges, including the quantity and quality of available natural product datasets and the interpretability of model predictions. Through these analyses, we outline future directions to guide the next generation of natural products science empowered by AI.

Fig. 1 provides a visual overview of how CLMs integrate into the natural product discovery pipeline across various stages. These applications, detailed in subsequent sections, demonstrate how CLMs accelerate each stage of the discovery process while maintaining the chemical validity essential for natural product research.

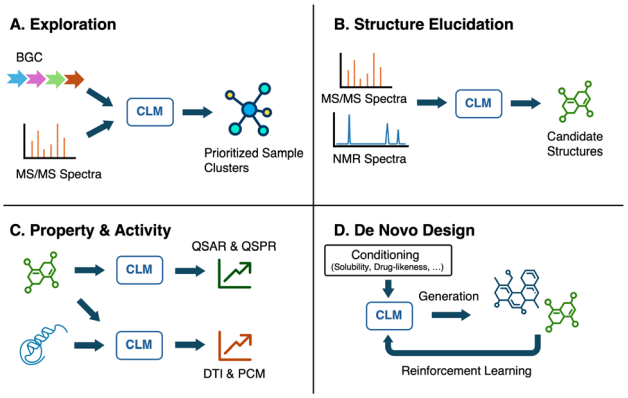


Fig. 1 Overview of how CLMs interface with natural product discovery. (A) Genome- and metabolome-based prioritization, in which CLMs take biosynthetic gene cluster sequences and/or MS-based profiles as input to suggest plausible metabolite structures. (B) Structure elucidation from spectral data, where CLMs map MS/MS or NMR spectra to candidate molecular structures that can be validated experimentally. (C) Activity and property prediction, with CLMs encoding natural product structures alone or jointly with protein sequences to predict bioactivity, target profiles, and ADMET-relevant properties. (D) *De novo* design and optimization, where CLMs generate novel, natural-product-like molecules conditioned on desired properties and iteratively refine them under multi-objective constraints.

2 Fundamentals of chemical language models

To understand how CLMs operate, it is first necessary to grasp the evolution of language models and their position within deep learning. Language models originally developed as a core technology in NLP, aiming to describe sequences of words or sentences probabilistically and to predict the next word or sentence.[8] Early language models were based on *n*-grams and related statistical techniques, leveraging word co-occurrence frequencies to capture sentence structure; however, they struggled to model long-range contextual dependencies in extended texts.[9] From the mid-2000s onward, neural language models emerged, and distributed word representations such as word2vec[10] and GloVe[11] became widespread, embedding semantic distances between words into low-dimensional vector spaces.[8,12]

With the advancement of deep learning, language model architectures have improved their accuracy by incorporating structures suitable for sequential data, including recurrent neural networks (RNNs),[12] long short-term memory (LSTM),[13,14] and gated recurrent units (GRUs).[15,16] However, these architectures still suffered from vanishing/exploding gradients and difficulty capturing long-distance dependencies.[17,18] The Transformer, proposed in 2017, introduced self-attention to learn dependencies across entire sequences in parallel, accelerating the development of large language models such as BERT[19] and GPT.[20] The Transformer's simultaneous improvements in both accuracy and training speed, combined with its adherence to scaling laws,[21] whereby performance increases with more training data and parameters, have ushered in the era of foundation models. These models undergo pretraining on massive text corpora and can be fine-tuned for diverse downstream tasks.

In chemistry, language model techniques have been adapted by representing molecules as text sequences, most prominently *via* SMILES.[22] Casting molecules as strings makes the learning problem formally analogous to sentence modeling, enabling direct transfer of NLP methods. This idea spread rapidly in the late 2010s and early work employed RNNs/LSTMs for molecular generation and prediction.[5,23] Subsequently, Transformer-based CLMs were developed, leveraging self-attention for long-range structural dependencies and large-scale pretraining, with applications expanding to molecular design, property prediction, and reaction pathway prediction.[4,24–27]

CLMs do more than memorize strings. They internalize statistical regularities and grammatical constraints of chemical

space, enabling high-validity prediction and generation for unseen structures. Recently released large-scale CLMs trained on extensive datasets[26–28] are rapidly increasing their utility in drug discovery and natural products research. Moreover, the scope of CLMs has expanded beyond molecular strings to encompass diverse chemical and biological data modalities, including BGCs, protein sequences, mass spectrometry data, NMR spectra, and textual metadata, all of which can be tokenized and processed using language modeling techniques. This section reviews the molecular string representations that underpin CLMs, explores how language modeling approaches are being extended to these broader data types, and examines the neural network architectures used to learn from them.

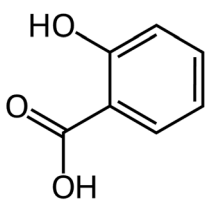
2.1 Molecular string representations

With CLM applications in mind, we introduce major string-based molecular representations and summarize their advantages and challenges for CLM usage. Fig. 2 illustrates how the same molecular structure can be encoded using different string representations, each with distinct advantages and limitations for CLM applications.

2.1.1 IUPAC nomenclature. IUPAC nomenclature is a human-readable, systematic naming scheme defined by IUPAC and widely used in research articles and patents. Some reports have shown better performance than SMILES for certain generative and property-prediction tasks,[29] but IUPAC names are generally long and complex, making them cumbersome to learn and generate directly as token sequences.

2.1.2 SMILES. Simplified Molecular Input Line Entry System (SMILES)[22] is a concise representation developed in the 1980s. It is relatively interpretable to humans and has been the de facto standard for CLMs. However, long-range syntactic dependencies (*e.g.*, branching and ring closures) can cause grammatical errors and learning difficulties for sequence models. Non-uniqueness and variation in stereochemical annotation also introduce representation instability.

2.1.3 DeepSMILES. DeepSMILES aims to mitigate some syntactic shortcomings of SMILES by, for example, representing ring closures with a single symbol and using only right parentheses for branching, thereby reducing the incidence of syntax errors in generation.[30] It often suppresses syntactically invalid outputs in generative models, though at some cost to human readability and compatibility with existing tooling.



IUPAC
2-hydroxybenzoic acid
SMILES
O=C(O)c1ccccc1O
DeepSMILES
O=CO)ccccccO
SELFIES
[O][=C][Branch1][C][O][C][=C][C][=C][C][=C][Ring1][=Branch1][O]

Fig. 2 Comparison of molecular string representations for salicylic acid. The same molecular structure is shown with its IUPAC nomenclature, SMILES, DeepSMILES with simplified branching and ring notation, and SELFIES with guaranteed chemical validity.

2.1.4 SELFIES. SELF-referencIng Embedded Strings (SELFIES), proposed by Krenn and colleagues, guarantees that any string decodes to a chemically valid molecule, *i.e.*, 100% validity.[31] Whereas SMILES can violate grammar or chemical constraints such as valence, SELFIES enforces both syntactic and semantic validity *via* a state-transition scheme. Its robustness benefits VAEs and CLM-based generation, although some studies note limitations in reproducing certain aspects of chemical diversity, including aromatic systems, potentially making it harder to reflect training data distributions faithfully.[32,33]

In practice, SMILES remains the dominant representation across CLM use cases such as generation, property prediction, and reaction prediction. DeepSMILES is used in scenarios prioritizing reduced syntax errors during generation, with reports of benefits in some settings. SELFIES is particularly attractive where generation robustness and chemical validity are crucial. IUPAC names are highly readable for humans but are not commonly adopted as the primary learning representation for CLMs.

2.2 Tokenizing diverse chemical modalities

While most chemical language modeling research focuses on SMILES or SELFIES representations, various data types essential to natural product discovery can be tokenized and processed using language modeling techniques. This approach treats each modality as a token sequence and uses standard pretraining objectives, including masked modeling, denoising, and contrastive alignment to develop useful representations. Fig. 3 illustrates how heterogeneous data types relevant to natural

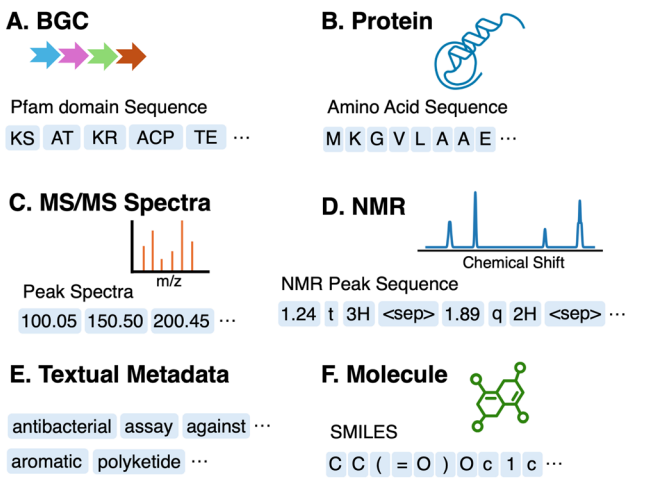


Fig. 3 Multimodal tokenization for chemical language models. (A) Biosynthetic gene clusters tokenized at the Pfam protein-domain level, where each domain is treated as a single token and the sequence preserves cluster organization. (B) Protein sequences tokenized at the single-amino-acid level. (C) MS/MS spectra serialized as ordered peak tokens, here shown as rounded or binned *m/z* values. (D) One-dimensional ^{1}H NMR spectra summarized as sequences of peak descriptors combining chemical shift, multiplicity, and proton count. (E) Textual metadata such as assay descriptions processed with wordpiece tokenization. (F) Molecules represented as SMILES strings.

product discovery can be converted into token sequences for chemical language modeling.

For BGC, these structures are represented as sequences of functional domain tokens arranged according to their cluster organization, with the option of including module boundaries and tailoring enzyme annotations.[34,35] Applying masked language modeling techniques to these domain sequences produces encoders that capture long-range biosynthetic dependencies, while taxonomic or environmental context can be added through additional token representations. Protein sequences use standard amino acid tokenization, and their embeddings are aligned with molecular representations through contrastive objectives or cross-attention mechanisms to enable cross-modality reasoning.[36]

Mass spectrometry data, specifically MS/MS spectra, are encoded as ordered sequences of peaks indexed along the mass-to-charge axis, with intensity values represented either as categorical bins or as scalar features.[37,38] The pretraining method typically combines masked peak recovery with denoising procedures applied to modified or perturbed spectra, along with contrastive objectives designed to associate spectra from the same compound measured under different conditions. NMR data are serialized differently based on dimensionality, with one-dimensional experiments represented as binned chemical shift sequences and two-dimensional experiments encoded as sequences of peak coordinate pairs, including solvent and temperature information as conditioning tokens.[39]

Additionally, textual metadata, including organism names, occurrence information, assay descriptions, and bibliographic context, are processed using standard natural language processing tokenization methods and then integrated with molecular encoders.[28,40]

Across these modalities, a consistent design pattern appears that first defines a chemically or biologically meaningful tokenization scheme with explicit context tokens, followed by pretraining procedures using masking, denoising, and contrastive losses to develop robust encoder architectures capable of learning useful representations from the data.

2.3 Neural network architectures

The neural architectures powering CLMs have evolved substantially. Here we outline the main families, including RNNs and Transformers, their characteristics, and typical use cases.

2.3.1 RNN. As noted above, RNN-based architectures were widely used in early CLMs, ingesting molecular strings token by token while maintaining a latent state for context. Vanilla RNNs work well on shorter sequences but suffer from vanishing/exploding gradients on long sequences.[17,18]

LSTM and GRU introduce gating mechanisms to mitigate these issues, dynamically deciding how much information to retain or forget and thereby stabilizing long-range dependencies. For SMILES, models must track relations between distant positions (*e.g.*, the correspondence of branch delimiters or ring indices across tens of characters), and LSTMs/GRUs can capture such dependencies to a degree.

During the formative years of CLMs, many LSTM/GRU-based generators were proposed, valued for producing syntactically valid molecules even with limited data.[5,41,42] However, their inherently sequential computation hinders parallelization, making large-scale training less efficient than with Transformers, and very long contexts remain challenging.

2.3.2 Transformer. The Transformer[6] centers on self-attention, enabling parallel learning of relationships among all tokens in a sequence. Unlike RNNs, which process inputs stepwise in one direction, Transformers compute pairwise token affinities in a single pass, facilitating capture of long-range dependencies and efficient large-scale pretraining.

For CLMs, molecular strings are tokenized and augmented with positional encodings, then processed by stacked layers of self-attention and feedforward networks.[4,24–27] This architecture excels at modeling complex, molecule-spanning dependencies such as branching and ring systems. However, it is important to note that self-attention has computational and memory complexity of $O(n^2)$ with respect to sequence length n, which can be a bottleneck for long SMILES typical of natural products with multiple rings or glycosylations.

Transformer architectures can be broadly classified into three categories: encoder-only, decoder-only, and encoder-decoder models. While all fundamentally utilize attention mechanisms, they differ in their training methods and applications. Below, we provide a detailed description of each architecture.

2.3.2.1 Encoder-only. Encoder-only models, exemplified by BERT,[19] process entire input sequences jointly, allowing each token to attend bidirectionally to every position. Pretraining typically uses masked language modeling (MLM), predicting randomly masked tokens from their surrounding context to learn statistical and syntactic features. The resulting representations transfer effectively to classification, regression, and sequence labeling.

In chemistry, applying MLM to SMILES enables models to learn grammatical constraints and chemical knowledge. Encoder-only CLMs are adept at capturing correspondences such as ring closures and branch boundaries. Models such as ChemBERTa leverage such pretrained representations for downstream tasks, including molecular classification, property prediction, and similarity search, often surpassing descriptor-based baselines.[25,26]

2.3.2.2 Decoder-only. Decoder-only models are autoregressive language models that generate entire sequences by repeatedly predicting the next token while referring only to preceding context through causal masking at each step. GPT[20] is a representative example, achieving strong generative coherence and promptability after large-scale pretraining. At inference, decoding strategies include greedy and beam search as well as stochastic schemes such as top-k and nucleus (top-p) sampling, which trade off diversity and quality.[43] Lightweight adaptation methods, such as prompt design, prefix/prompt tuning, and adapters, further ease task transfer.

In CLMs, SMILES/SELFIES are treated as token sequences for *de novo* molecular generation. Conditional generation supplies targets such as desired properties (*e.g.*, logP, QED, synthetic

Table 1 Comparison of CLM architectures and their suitability for natural product discovery tasks

Architecture	Training objective	Typical NP tasks
Encoder-only	Masked language modeling	Property prediction, BGC classification
Decoder-only	Autoregressive LM	*De novo* generation, scaffold exploration
Encoder-decoder	Denoising/infilling	Spectrum → structure, conditional molecule generation, retrosynthesis

accessibility), functional groups, or scaffolds to bias outputs.[44,45] Reinforcement learning can also guide generators toward objectives like predicted activity or docking scores.[46–48]

2.3.2.3 Encoder–decoder. Encoder–decoder models pair a bidirectional encoder that builds a sequence representation with an autoregressive decoder that generates outputs conditioned on that representation. They suit text-to-text mappings such as translation and summarization. Pretraining often employs self-supervised tasks like text infilling or span corruption (*e.g.*, BART, T5).[49,50] While they typically require sizable, high-quality paired data, they tend to be stable and straightforward to train for text-to-text generation tasks.

In chemistry, SMILES-to-SMILES mappings are widely used for forward/reaction prediction, retrosynthesis-like molecular optimization, and related tasks.[24,27,51] Encoder-decoder models also support cross-modality translation (*e.g.*, generating candidate structures directly from MS/MS or NMR spectral features).[38,39,52–54]

As each architecture has distinct strengths, Table 1 summarizes which types of natural product discovery tasks each is best suited for.

3 Applications of language models in natural product discovery

This chapter organizes how language models are integrated into practical workflows in natural product drug discovery and how they concretely shorten bottlenecks on the path to discovery. We detail which CLM types and applications suit each stage of the research flow (exploration, structure elucidation, activity/property assessment, and design). Table 2 summarizes the models covered in this chapter by domain and model family. While this chapter primarily focuses on language model-based approaches, we briefly reference some non-language model-based methods, such as graph-based models, where relevant. Although these fall outside our main scope, they are mentioned sparingly to provide context for understanding language model applications.

3.1 Exploration in genome mining and screening

At the outset of exploration, one must decide which genomes or samples to prioritize given limited experimental resources. We first discuss genome-based approaches that identify and learn representations of secondary-metabolite BGCs to estimate product classes and novelty for targeting. We then describe extract-based approaches that select samples or fractions with high novelty based on MS/MS or activity data.

3.1.1 BGC identification and representation learning. For genome-based exploration, approaches that treat BGCs as token sequences and learn context with sequence models are effective. DeepBGC maps Pfam domain sequences to word2vec-derived distributed representations (pfam2vec), detects BGC boundaries with a bidirectional LSTM (BiLSTM), and further assists product-class classification with a random forest, improving both recall of known classes and recovery of putative novel classes.[55] BiGCARP introduces BERT-style self-supervised pretraining. It tokenizes Pfam domains, learns representations *via* masked language modeling, and transfers them to BGC detection and product classification, boosting accuracy and robustness.[34] BGC-Prophet, centered on a Transformer encoder, leverages self-attention to capture long-range positional dependencies and enables efficient scanning at whole-genome and metagenome scales, surpassing prior methods in both sensitivity and throughput.[35] Predicted novelty scores and product classes from these models guide target selection, *i.e.*, which BGCs to route to the bench first.

3.1.2 Extract-based prioritization. Complementing genome-based methods, extract-level approaches leverage physicochemical and spectral signatures to prioritize samples. In extract-based exploration, properties exhibited by culture broths or extracts (activity, physicochemical signatures, spectral data) can be used to avoid rediscovering known compounds while prioritizing samples or fractions with high novelty. When leveraging MS/MS, one can achieve more reliable structure prediction and library search than with MS alone. Representation learning that captures semantic similarity between spectra is well-suited for this prioritization. Spec2Vec learns word2vec-style embeddings from large spectral corpora using peak co-occurrence patterns and shows better alignment with structural similarity than cosine scores that have long been used.[37] Complementarily, MS2DeepScore uses a Siamese neural network to directly predict the Tanimoto similarity between molecular fingerprints from a pair of spectra, bridging the gap between spectral similarity and structural similarity.[57] Both integrate well with molecular networking,[72] enabling one to distinguish clusters enriched for known *versus* unknown chemistry and to allocate isolation and structure-elucidation resources preferentially to the latter.

3.2 Structure elucidation

Correct structural assignment is essential for understanding the properties and functions of isolated compounds. While our focus is on language models, in this section, we briefly reference graph-based Transformers that currently lead forward

Table 2 Models available for natural product drug discovery

Name	Domain	Model family
DeepBGC[55]	Genome mining	LSTM + random forest
BiGCARP[34]	Genome mining	Transformer (encoder)
BGC-Prophet[35]	Genome mining	Transformer (encoder)
BioNavi-NP[56]	Retrobiosynthesis	Transformer (single-step) + AND-OR search
Spec2Vec[37]	Spectral similarity	Other (word2vec embedding)
MS2DeepScore[57]	Spectral similarity	Other (Siamese DNN)
Spec2Mol[38]	Structure elucidation (MS/MS → SMILES)	GRU
MassGenie[53]	Structure elucidation (fragments → SMILES)	Transformer (decoder)
MSNovelist[52,58]	Structure elucidation (FP-guided)	RNN + constraints
CFM-ID 4.0[59]	MS forward model	Other (learnable fragmentation model)
MassFormer[60]	MS forward model	Graph transformer
GT-NMR[61]	NMR forward model	Graph transformer
Alberts *et al.*[39]	NMR *de novo*	Transformer (encoder-decoder)
ChemBERTa[25]	Activity/ADMET	Transformer (encoder)
ChemBERTa-2 (ref. 26)	Activity/ADMET	Transformer (encoder)
SMILES-Mamba[62]	Activity/ADMET	Other (Mamba)
TransformerCPI[63]	DTI/CPI	Transformer (encoder)
T-ALPHA[64]	DTI/CPI	Transformer (encoder, encoder-decoder)
MolTrans[65]	DTI/CPI	Transformer (encoder)
DrugCLIP[66]	Virtual screening	Dual encoder
Graphormer[67]	QSAR/physchem	Graph transformer
MolBART[24]	*De novo* design	Transformer (encoder-decoder)
Chemformer[27]	*De novo* design	Transformer (encoder-decoder)
BARTSmiles[68]	*De novo* design	Transformer (encoder-decoder)
MolT5 (ref. 28)	Text-molecule	Transformer (encoder-decoder)
REINVENT (v1-3)[46–48]	RL optimization	LSTM
Tay *et al.*[69]	NP-like molecule generation	LSTM
NPGPT[70]	NP-like molecule generation	Transformer (decoder)
NP-VAE[71]	NP-like molecule generation	VAE

prediction of spectrum (MassFormer, GT-NMR).[60,61] In structure elucidation from extracts, databases often lack the molecules encountered, making pure library matching insufficient. A common integrated workflow uses tandem mass spectrometry (MS/MS) to broadly propose candidates and verifies them by nuclear magnetic resonance (NMR). Natural products are often structurally complex, and elucidation can be challenging, so, in recent years, modeling through machine learning and deep learning has advanced for both MS/MS and NMR. Fig. 4 shows three different approaches for structure elucidation using MS/MS data.

In methods that generate structures directly from MS/MS, the core is an encoder-decoder that encodes spectra, either as continuous sequences or sets of peaks, and autoregressively decodes SMILES. Spec2Mol embeds multi-collision-energy MS/MS spectra with an encoder and generates SMILES using a GRU-based decoder in an end-to-end scheme, proposing candidate structures even when not present in databases.[38] Spec2Mol was not trained on a natural-product-specific dataset, but the general chemistry databases used for training (PubChem, ZINC-12) and the NIST20 MS/MS library are likely to contain natural-product-related molecules indirectly, such as metabolites and plant-derived compounds. MassGenie formulates the problem as translation from fragment information to SMILES, pretraining a Transformer largely on synthetic spectra and showing that scaling to millions of structure-fragment pairs is effective.[53] More recently, BART-style pretraining has been adopted in generators such as MS2Mol, prompting discussion of tokenization designs for treating spectra as language, including discretization of encodings and choices of bin width.[73]

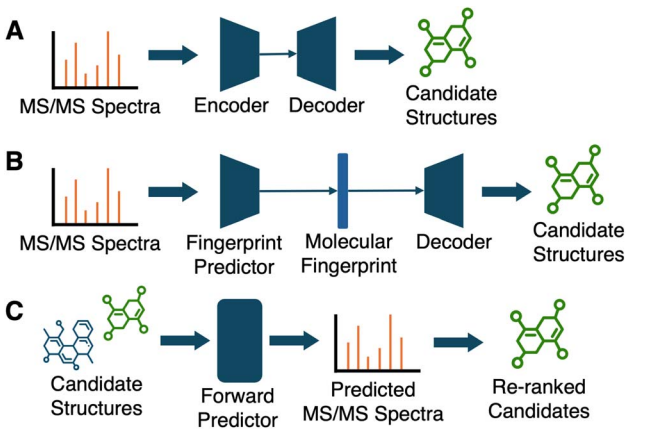


Fig. 4 Structure elucidation workflows combining MS/MS with machine learning models. (A) Direct MS/MS to SMILES generation, where an encoder-decoder model maps tandem mass spectra to candidate molecular structures. (B) Fingerprint-based two-stage approach that first predicts molecular fingerprints from MS/MS and then uses a conditional generator to sample structures consistent with the predicted fingerprint. (C) Forward-prediction-based re-scoring, in which candidate structures are passed to MS/MS forward predictors and ranked according to the agreement between predicted and experimental spectra.

Two-stage methods that first infer molecular fingerprints from MS/MS spectra and then generate structures consistent with those fingerprints have also been proposed. MSNovelist stacks a language-model-based generator atop CSI:FingerID, which predicts fingerprints from spectra, achieving high top-*k* identification rates for complex isomers while enabling *de novo* proposals beyond the database.[52,58]

For candidate re-scoring, forward predictors of spectra from structures play a key role. Predicted spectra are compared with observed spectra to re-rank candidates by physical plausibility. CFM-ID 4.0 refines the competitive fragmentation model by introducing learnable, topology–aware parameters, achieving accurate MS/MS spectra prediction.[59] MassFormer uses a Graph Transformer with self-attention to capture long-range correlations on molecular graphs and reports state-of-the-art accuracy for collision-energy-dependent spectra.[60] Such forward models stabilize top-ranked results by re-evaluating candidate sets proposed by direct generation or fingerprint-guided models against the observed spectra.

For NMR, both high-accuracy chemical-shift prediction and structure prediction from spectra have advanced. GT-NMR uses a Graph Transformer that inputs molecular graphs to predict chemical shifts, achieving state-of-the-art mean absolute errors on nmrshiftdb2 and on natural product benchmarks.[61] In parallel, models that predict SMILES directly from NMR spectra are being trained: Alberts *et al.* pretrain an encoder-decoder Transformer on synthetic spectra and, using combined ^{1}H and ^{13}C data, achieve 67% top-1 structural accuracy.[39] Multimodal Transformers that jointly ingest MS and NMR (and in some cases IR or 2D NMR) are also under investigation, showing advantages in difficult cases such as distinguishing structural isomers, especially regioisomers, by fusing complementary information sources.[74]

3.3 Activity and property prediction

Once structures are determined, it is crucial to estimate biological activities and ADMET properties rapidly and consistently to concentrate limited experimental resources on promising candidates. Encoder-style CLMs that take SMILES as input are widely used at this stage. ChemBERTa demonstrated that Transformer-based molecular representation learning is effective for QSAR/ADMET by showing competitive performance under low-label regimes *via* self-supervised pretraining on approximately 10 million SMILES from PubChem, which covers a broad range of compounds including both synthetic and natural-origin entries.[25] ChemBERTa-2 redesigns the training scheme and data scale, improving across multiple benchmarks, including MoleculeNet, and clarifying the direction of foundation-model-based representation learning in chemistry.[26] In parallel, BERT-family encoder models such as SMILES-BERT and Mol-BERT have been proposed, refining contextual representations of functional groups and substructures *via* masked language modeling and tasks such as canonical-equivalence prediction across different SMILES for the same molecule.[4,75] Recently, SMILES-Mamba applies a selective state-space model (computationally lighter than Transformers) to SMILES sequences, showing strengths on several ADMET tasks under large-scale pretraining.[62]

When target proteins must be considered, multimodal models that integrate protein sequences with molecular representations are effective. TransformerCPI applies self-attention to both protein sequences and molecular structure information, improving target–compound interaction prediction and offering interpretability *via* attention weight visualization of interaction regions.[63] Although its training data from ChEMBL and BindingDB may include some natural products, these databases are not natural-product-focused. MolTrans extracts substructure patterns from SMILES and from protein sequences, then encodes them using Transformer encoders to predict drug–target interactions (DTIs) with high accuracy.[65] Its training data similarly cover broad drug-like chemical space. It also supports semi-supervised learning with unlabeled molecules. More recently, retrieval-style virtual screening aligns binding-pocket and molecular representations *via* contrastive learning. For example, DrugCLIP is pretrained on large sets of compound–protein–pocket pairs without explicit affinity labels and achieves fast, accurate screening.[66] Recent advances include T-ALPHA, which employs a hierarchical Transformer framework integrating multimodal representations through three distinct channels (protein, ligand, and protein-ligand complex), demonstrating state-of-the-art performance even when using predicted structures instead of crystal structures.[64] For chemical spaces like natural products (often out of distribution relative to standard drug-like corpora), combining sequence-language-model representations on the target side with CLM representations on the ligand side is advantageous for target identification and matching to known targets.

Care must be taken with activity cliffs, in which small substructural changes produce large activity differences. Such pairs (compounds with similar structures yet markedly different activities) often degrade the performance of traditional ML and DL models. MoleculeACE provides a benchmark platform to quantify activity cliffs and systematically evaluate model performance.[76] Robust activity prediction is especially important for natural products, which often feature diverse functional groups and high stereochemical complexity.

CLMs that take SMILES as input often compete closely with graph neural networks (GNNs). For example, while ChemBERTa-2 outperforms Chemprop (a model based on the D-MPNN variant of GNNs) on multiple MoleculeNet tasks,[26] D-MPNN and Graphormer remain strong for quantum and physicochemical properties.[67,77] Hybridization is an active direction. MolPROP integrates CLM embeddings with GNNs, and UniMAP fuses SMILES and graph embeddings, both reporting improvements.[78,79]

Across Sections 2.2 and 3.3, most of the models introduced have been developed and validated on general chemical datasets rather than those focused specifically on natural products. However, these approaches are highly flexible, and with domain-specific fine-tuning or adaptation using natural product datasets, they hold strong potential to capture the unique structural complexity, molecular size ranges, and stereochemical diversity characteristic of natural products.

3.4 *De novo* design

This section surveys how *de novo* molecular design with CLMs can help open previously unexplored regions of natural product chemistry. Although slightly apart from in-the-trenches CLM usage in discovery workflows, we summarize efforts to generate novel compounds.

Within the encoder-decoder models, BARTSmiles is a representative example. After pretraining with masked language modeling, it undergoes fine-tuning on toxicity and chemical reaction datasets, demonstrating state-of-the-art performance across many tasks.[68] Such models have the advantage that encoder embeddings can power prediction tasks while the decoder enables generation. In practice, BARTSmiles handles property prediction on MoleculeNet and supports generation tasks such as retrosynthesis and reaction prediction. MolT5 enables bidirectional translation between text and molecular representations, opening the door to natural-language-driven molecular design *via* captioning and conditional text-to-molecule generation.[28]

To bias generation toward the natural product space, CLMs trained on natural product data have been explored. Tay *et al.* trained an LSTM on natural product structures and generated approximately 67 million natural-product-like compounds.[69] NPGPT fine-tunes a pretrained GPT on natural product data, bringing the generation distribution closer to natural products and yielding candidates with promising drug-like potential.[70] While this review primarily focuses on language models, it is worth noting that other deep learning approaches have also been developed for natural product generation. For instance, NP-VAE employs graph-based VAE architectures to handle large molecular structures with 3D complexity, including chirality.[71]

For controllability, conditional Transformers have been widely studied for multi-objective optimization. By conditioning on targets such as protein sequences or pocket descriptors, property vectors, or scaffolds, autoregressive generation can be steered.[80–82] Such conditional generation is useful for obtaining compounds that satisfy desired properties while retaining natural-product-like characteristics and scaffold diversity.

To explore chemical space under explicit objectives, reinforcement learning (RL) is commonly combined with generators. REINVENT updates the policy of an RNN-based generator using rewards derived from objective functions (predicted activity, docking scores, *etc.*).[46–48] For example, by supplying a natural-product-likeness objective, one can train a model to generate natural-product-like compounds. REINVENT has seen continual refinement and widespread application.

A critical question for *de novo* design in the natural product space is whether a generated compound is synthetically accessible. BioNavi-NP addresses this by combining a Transformer-based single-step retrobiosynthesis model with AND-OR tree search (Retro*), trained on datasets including general organic and biosynthetic reactions, to propose biologically plausible multi-step biosynthetic routes.[56] In internal tests, it proposed routes for 90.2% of 368 cases, with a building-block hit rate of 56.0% (72.8% when user-defined building blocks were specified). Coupling such retrobiosynthesis tools with CLM-based generators offers a promising workflow. *De novo* generated candidates can be filtered or ranked by the feasibility of their predicted biosynthetic routes, thereby focusing experimental effort on structures for which plausible pathways exist.

Another important consideration is whether CLMs generate structurally diverse compounds rather than close analogs of training-set members. Encouragingly, recent studies suggest that CLMs biased toward natural products can produce diverse outputs. It has been reported that natural-product-like compounds generated by CLMs trained on the structures of natural products are distributed in a broader physicochemical space than the training data.[69] Nevertheless, generative models can exhibit mode collapse, and practitioners should routinely report diversity and novelty metrics such as the Fréchet ChemNet Distance (FCD), internal diversity, and scaffold diversity, as standardized in benchmarks like MOSES and GuacaMol.[83,84] Systematic use of these metrics will help the community assess whether CLMs are truly expanding the explored chemical space.

3.5 Case studies

The practical impact of CLMs is perhaps best illustrated by studies in which language-model predictions were followed by experimental validation. A notable example on the discovery side is a recent Transformer-based platform for BGC prediction and design.[85] In this work, BGCs were represented as sequences of biosynthetic domains, allowing a RoBERTa-based model to learn domain co-occurrence and positional relationships in a language-like manner. Beyond benchmarking on known BGCs, the authors used the cyclooctatin pathway as a case study for model-guided BGC engineering. A model trained on whole genomes proposed several domains that were absent from the original cluster and from the predictions of a BGC-only model. To test one such prediction experimentally, the authors located a gene encoding the predicted domain in a related Streptomyces genome and co-expressed it with the cyclooctatin biosynthetic genes in Streptomyces albus, which led to the production of an unknown cyclooctatin derivative.

A complementary example comes from *de novo* design. In a study on retinoid X receptor (RXR) modulation, an RNN-based CLM was pretrained on bioactive compounds and then fine-tuned using a small set of natural product templates with known RXR activity.[86] The resulting model generated natural-product-inspired mimetics that were prioritized, synthesized, and experimentally tested. Two of the four designed compounds showed RXR modulatory activity that was isofunctional with that of the natural product templates, illustrating how CLMs can translate natural product bioactivity patterns into experimentally validated design hypotheses.

4 Data resources and curation for CLMs

The performance of CLMs depends critically on which data are selected and how they are preprocessed. In this chapter, we first organize the available natural product databases. We then

describe the procedures required to prepare retrieved records as model-ready training data, outlining the end-to-end pipeline from preprocessing through dataset splitting.

4.1 Major natural product databases

When using natural product databases as training resources for CLMs, it is essential to understand, for each resource, whether structures are rigorously determined, whether annotations are consistent, and how accessible the data are. Here, we focus on four representative databases (COCONUT 2.0, The Natural Products Atlas 3.0, LOTUS, and SuperNatural 3.0) and summarize their characteristics and recommended uses for CLM pretraining and evaluation.[87–90]

COCONUT 2.0 aggregates publicly available natural product data and supports community-driven curation.[87] The data can be obtained *via* bulk download or API, facilitating ML applications and pretraining. It integrates diverse sources and contains roughly 700 000 unique structures, although some non-natural entries are known to be present and warrant caution.

The Natural Products Atlas 3.0 focuses on microbially derived compounds and employs literature-based curation with human verification in the workflow.[88] The latest release incorporates structural corrections, improving overall accuracy. With 36 545 natural products, it is especially valuable as a high-confidence set for evaluation.

LOTUS publishes occurrence information (compound, organism, and literature) as a knowledge graph and provides rich, reusable metadata such as classifications and synonyms.[89] Although led by the same group as COCONUT, LOTUS offers more comprehensive annotations for organismal and bibliographic context. For CLMs, this supports the design of hard negatives, assessment of phylogenetic or reporting biases, and integration with textual information.

SuperNatural 3.0 covers natural products and derivatives broadly, and alongside structural data it includes predicted properties, toxicity-related information, and supplier metadata.[90] With approximately 450 000 compounds, it is convenient for training, but annotations are tiered into two confidence levels, so reliability is not uniform. Compared with COCONUT, the explicit confidence indicators in SuperNatural 3.0 make it easier to exclude dubious entries, which is useful when balancing data volume against annotation quality.

Despite these valuable resources, the scale of available natural product data remains considerably smaller than synthetic or drug-like compound collections. ZINC 22 (ref. 91) contains several billion commercially available compounds, and ChEMBL[92] includes approximately 2.5 million bioactive molecules. Recent CLMs like MolFormer[93] have been pretrained on over 1 billion compounds from these large-scale databases. In contrast, the largest natural product database, COCONUT 2.0, contains roughly 700 000 structures. The limited dataset size constrains model training, and the structural complexity of natural products means that existing databases provide sparse coverage of the chemical space. Addressing this data gap through continued curation and development of data-efficient learning methods will be crucial for realizing the full potential of CLMs in natural product discovery.

Beyond scale, data quality and consistency present additional challenges. Among these, incomplete and inconsistent stereochemical annotations stand out as a critical issue for natural product databases. Natural products frequently possess numerous stereocenters, making stereochemical information important, yet database entries vary widely. Some include fully determined configurations, while others report stereochemistry with undetermined centers. Community-driven curation initiatives and standardized natural-product-focused benchmarks, similar to MoleculeNet[94] for general chemistry, would help address these issues by providing shared, high-quality evaluation resources and encouraging consistent data practices across the field.

4.2 Data processing and splitting pipeline

This subsection explains the sequence of steps from acquisition to CLM-ready datasets, covering standardization of string representations, data augmentation, tokenization, and dataset splitting.

The first step is to standardize string representations such as SMILES. RDKit and MolVS are commonly used tools for this purpose.[95,96] Standard procedures include normalization, reionization, cleavage of metal coordination, and exclusion of mixtures, which are applied in a prescribed order. Harmonizing representations across heterogeneous sources is crucial for smooth CLM training.

Quality control should proceed in parallel with standardization. For natural product datasets, in addition to detecting obvious valence violations and ring inconsistencies, it is important to compute diagnostic metrics (*e.g.*, the NP-likeness score[97]) to identify outliers and characterize distributional biases.

A particularly important quality-control decision concerns the treatment of incomplete stereochemical assignments, which are common in natural product databases. Three main approaches are available. The first option is (1) to exclude molecules with incomplete stereochemistry, yielding a cleaner but smaller dataset suited for tasks where stereochemical correctness is essential. The second option is (2) to retain molecules as-is, because SMILES can represent undefined stereocenters alongside defined ones. This preserves data volume while accepting partial stereochemical information. The third option is (3) to strip all stereochemical annotations entirely to reduce token complexity and sequence length, which can simplify training at the cost of discarding stereochemical knowledge. In practice, option (3) is often chosen for large-scale pretraining where reducing training cost is a priority, whereas option (1) is preferred when the task explicitly requires stereochemically complete structures. Reporting which strategy was adopted and its effect on dataset size is recommended to facilitate reproducibility.

For data augmentation, enumerated SMILES are commonly employed.[98] Because a single molecule admits multiple valid SMILES, this approach increases the effective dataset size.

Generating multiple SMILES per molecule during training generally improves model robustness.

Several tokenization strategies are available. Subword segmentation *via* byte pair encoding (BPE), widely used in NLP, tends to reduce sequence length but may not align tokens with chemically meaningful units, potentially making substructure recognition by CLMs more difficult. Given that chemistry has a relatively small vocabulary of atomic symbols compared with natural language words, atom-level tokenization is also common. In addition, chemistry-specific tokenizers such as SMILES Pair Encoding and atom-in-SMILES have been proposed.[99,100]

Dataset splitting has a direct impact on the assessment of generalization. A widely adopted principle is scaffold splitting based on Bemis-Murcko frameworks, which ensures that the same scaffold does not appear across training and evaluation sets.[94,101] By requiring that evaluation sets contain novel scaffolds, this strategy tests whether the model generalizes to unseen chemistry. When feasible, scaffold splits can be combined with temporal splits based on publication or deposition year to create a more rigorous learning setup.

In practical use, performance on unseen scaffolds is important. Although scaffold splits are stricter than random splits, they can still overestimate performance because closely related but scaffold-distinct molecules may be mixed.[102] Splitting based on chemical similarity (*e.g.*, using UMAP or related approaches to partition the chemical space) has been recommended to obtain more realistic estimates.

5 Future perspectives

The use of CLMs is expanding and transitioning into technologies that can be embedded into day-to-day natural product discovery. Nevertheless, the natural product chemical space differs from drug-like small molecules in its statistical properties and stereochemical complexity, and conventional model designs and evaluation protocols may not always be appropriate. This chapter discusses prospects for CLMs in natural product discovery under these domain-specific conditions.

5.1 Natural-product-specialized CLMs and 3D representations

We anticipate increasing demand for CLMs specialized in natural products. The natural product chemical space differs from typical drugs defined by the "Rule of Five",[103] yet it yields scaffolds of importance for drug discovery and contains many structures with affinity for biological targets. Historically, it has provided numerous leads with novel mechanisms of action and activity against difficult targets.[1] Models trained on general chemical corpora tend to undervalue stereochemically rich and highly functionalized scaffolds typical of natural products, which can misalign prioritization and generation with real-world discovery value. Accordingly, to fully leverage CLMs for natural product discovery, representations and priors adapted to this domain are essential.

A practical training strategy is to pretrain on large, diverse general chemistry datasets to acquire chemical grammar and broad knowledge, followed by continued pretraining on comparatively smaller natural product datasets. This downstream domain adaptation shifts the exploration bias toward the natural product space by imprinting domain-specific structural regularities. In turn, one can expect improvements in activity and property prediction for natural products, better re-ranking of candidates with unfamiliar scaffolds, and richer scaffold diversity in *de novo* generation.

However, transferring representations pretrained on general chemistry can suffer from distribution shift, leading to negative transfer or catastrophic forgetting.[104] Parameter-efficient fine-tuning (PEFT) has therefore attracted attention. A representative approach, LoRA, adjusts only a small subset of parameters atop a frozen backbone, enabling effective adaptation under limited data.[105]

Insufficient stereochemical awareness in CLMs is another key challenge. Compared with typical screening libraries, natural products often have numerous stereocenters, high 3D complexity and conformational flexibility. Natural products span a wide range of flexibility, from rigid macrocyclic peptides and polyketides to highly flexible glycosylated or acyclic terpenoid structures. While this enhances complementarity to binding pockets, string-based CLMs do not always capture stereochemistry or conformational preferences reliably. Indeed, in Transformer training on SMILES, substructural patterns are learned early, whereas the acquisition of chirality lags until later stages.[106] Such tendencies are particularly problematic for sets dense in stereocenters, as in natural products. Further advances in stereochemical representation and learning are therefore crucial to extract more value from CLMs in this domain.

Beyond optimization of training recipes and fine-tuning strategies, a complementary direction is to refine the underlying tokenization and representation schemes so that three-dimensional and conformational information can be expressed directly at the sequence level. Most current CLMs operate on SMILES/SELFIES-based sequences and, at best, incorporate static stereochemical annotations.[22,26,31] This design implicitly treats conformation as something to be recovered downstream by separate 3D models rather than as part of the primary representation.

Recent work on conformation-aware line notations and 3D-aware tokenization suggests one pathway forward. Conformation-specific SMILES variants such as CSMILES encode a particular conformer by decorating SMILES atoms or bonds with additional chiral, dihedral, or distance labels.[107] TokenMol instead represents a molecule as an ordered sequence of discretized torsion angles along its backbone, so that a CLM can model conformational preferences explicitly as a sequence prediction problem.[108] These methods make it possible for CLMs to treat conformations as tokens rather than as information reconstructed solely by a separate three-dimensional module.

For natural products, such conformation-aware tokenization is particularly attractive. Natural product activity often depends on local conformational preferences, stereochemical

orientation of pharmacophores, or macrocycle rigidity, features that are only indirectly encoded in standard SMILES.[109–111] CLMs trained on conformation-augmented sequences can learn distributions over three-dimensional motifs rather than only over two-dimensional connectivity patterns.[108,112] Moreover, conformational tokens offer a natural interface between string-based CLMs and SE(3)-equivariant models, enabling hybrid architectures in which language models capture global scaffold and substitution patterns while equivariant networks refine local geometries.

Realizing these possibilities will likely depend on data and evaluation. For natural products, curated three-dimensional information remains fragmentary. Most public resources provide only 2D structures or a single low-energy conformer, and even 3D-focused databases cover at most tens of thousands of metabolites.[113] As research into conformation-aware representations advances, the demand for corresponding 3D structural data is expected to increase, highlighting the need for greater availability of such datasets in the future. In parallel, benchmarks for conformation-aware CLMs should move beyond 2D property prediction to tasks that evaluate the recovery of conformational ensembles and property prediction from 3D structures, so that competing representations and training objectives can be compared on questions that matter for natural product discovery.

5.2 Multimodal CLMs integrating BGC sequences, MS/NMR spectra, textual metadata

Natural product discovery spans multiple steps (structure elucidation, prioritization, and evaluation) and involves multiple modalities. A single modality is often insufficient to reduce uncertainty. Recent tools like MultiT2 demonstrate the value of connecting fragmented multimodal data in natural products, specifically for bacterial aromatic polyketides, through causal inference approaches.[114] Rather than treating structure, BGC sequence, MS/NMR spectra, and literature/database text as independent features to be fused post hoc, multimodal CLMs map them into a shared latent space and, when needed, train encoder-decoder models to translate between modalities. Such models can accelerate both the identification of unknown compounds and the focusing of resources on candidates with high discovery value. As noted above, self-supervised representation learning for BGCs,[34,35,55] structure prediction from spectra,[38,53,73] and text-structure translation[28,115] are already feasible. Recently, this text-structure multimodality has been further advanced by models like mCLM, which tokenizes molecules as functional building blocks, enabling direct and bilingual translation between natural language functional descriptions and makeable molecular structures.[116] We expect the development of CLMs that jointly handle these diverse modalities to accelerate.

A persistent obstacle to higher-performing multimodal models is data scarcity. Large, high-confidence datasets that systematically link BGCs, metabolites, and text do not yet exist. When relying on low-confidence associations, learning may become dependent on inferred links or weak labels and degrade in performance. Benchmarking for multimodal models is also underdeveloped. Standardizing datasets and evaluation metrics will make model comparison possible and should catalyze progress.

5.3 Explainable and uncertainty-aware CLMs

For CLMs in natural product discovery, high average accuracy alone is insufficient. Decisions about which samples to deepen, which candidates to synthesize, and which to scrutinize are made under limited experimental budgets, blending expert judgment with occasional ML predictions. If models output not only predictions but also scientific rationales and uncertainty estimates, the evidential basis becomes clearer and decision-making becomes easier. In particular, when activity cliffs or measurement noise contribute to uncertainty, numerical confidence is essential to trust predictions. Explainability and uncertainty estimation for CLMs remain nascent and warrant further development.

For explainability, methods such as XSMILES, which visualize token attributions and attention weights from SMILES-based models onto molecular graphs, provide foundations for chemists' decisions.[117] However, it should be noted that attention weights do not constitute direct explanations. Interpretation consistency should be enhanced by combining multiple explanation methods and providing counterexamples where slight modifications change activity.

For uncertainty, ensemble modeling can quantify predictive dispersion. For activity prediction and related tasks, conformal prediction can endow outputs with top-k candidate sets and guarantees on confidence. Such frameworks are particularly beneficial for compounds with natural-product-specific features that are underrepresented in generic small-molecule corpora.

5.4 Interdisciplinary collaboration

A practical challenge in applying CLMs to natural product discovery is the gap between computational and experimental communities. For computational scientists, obtaining high-quality experimental data, such as bioassay results, annotated spectra, and verified structures, remains difficult because these data are often generated in small batches and dispersed across laboratories. Conversely, for natural product chemists, the overhead required to train, validate, and deploy CLMs can be prohibitive without dedicated computational support. Advancing interdisciplinary collaboration, where both communities develop a shared vocabulary and mutual understanding of each other's needs and constraints, will be essential for CLMs to become routine tools in natural product discovery. Through such collaboration, the learning and evaluation cycle of CLMs can be accelerated, thereby increasing the maturity and utility of these models as practical tools for researchers.

6 Conclusions

This review provides a comprehensive overview of CLMs for natural product design and discovery, ranging from fundamental concepts and model architectures to practical

applications and future perspectives. The number of tools readily applicable to natural product drug discovery research is increasing, and their introduction at appropriate times can contribute to resolving bottlenecks. Recent CLM development has embraced large-scale pretraining centered on Transformers, shifting emphasis from single-task models to more general foundation models. Large chemical foundation models are now available, and continued progress in CLM research is expected.

We summarize four priority directions. The first is the establishment of natural-product-specialized CLMs. By biasing models toward the natural product space, domain-specific tasks can be optimized. The second is the development of multimodal models that treat BGC sequences, MS/NMR, structures, and textual literature within a shared latent space. The third is the integration of explainability and uncertainty estimation as standard features, so that model predictions are accompanied by rationale and confidence to better support expert decision-making. The fourth is the reduction of barriers to adoption through interdisciplinary collaboration between computational and experimental groups.

Underpinning all four directions, foundational work on data quality and reproducibility is indispensable. Addressing inconsistent stereochemical annotations, incomplete metadata, and the absence of standardized natural product benchmarks will strengthen the datasets on which CLMs depend. Publishing pre-processing pipelines, covering structure normalization, quality control, data augmentation, and dataset splitting, together with training and evaluation code, will enable comparison and portability of results. As such shared infrastructure matures, CLMs are poised to become a practical, consistent decision-support technology throughout natural product discovery.

7. Author contributions

KS, KF, AK, YK, and MO designed the study and wrote the manuscript. All authors read and approved the final manuscript.

8. Conflicts of interest

There are no conflicts to declare.

9. Data availability

No primary research results, software or code have been included and no new data were generated or analysed as part of this review.

10. Acknowledgements

This work was financially supported by the Japan Science and Technology Agency FOREST (Grant No. JPMJFR216J), Japan Science and Technology Agency ACT-X (Grant No. JPMJAX25LB), Japan Society for the Promotion of Science KAKENHI (Grant No. JP23H04880, JP23H04887, JP23K28186, JP24KJ1091), and Japan Agency for Medical Research and Development Basis for Supporting Innovative Drug Discovery and Life Science Research (Grant No. JP25ama121026).

References

1 D. J. Newman and G. M. Cragg, *J. Nat. Prod.*, 2020, **83**, 770–803.
2 A. G. Atanasov, S. B. Zotchev, V. M. Dirsch, International Natural Product Sciences Taskforce and C. T. Supuran, *Nat. Rev. Drug Discovery*, 2021, **20**, 200–216.
3 K. L. Kurita and R. G. Linington, *J. Nat. Prod.*, 2015, **78**, 587–596.
4 S. Wang, Y. Guo, Y. Wang, H. Sun and J. Huang, *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2019, pp. 429–436.
5 M. H. S. Segler, T. Kogej, C. Tyrchan and M. P. Waller, *ACS Cent. Sci.*, 2018, **4**, 120–131.
6 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser and I. Polosukhin, *Adv. Neural Inf. Process. Syst.*, 2017.
7 R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou and P. Liang, *arXiv*, 2021, preprint, arXiv:2108.07258, DOI: 10.48550/arXiv.2108.07258.
8 Y. Bengio, R. Ducharme, P. Vincent and C. Janvin, *J. Mach. Learn. Res.*, 2003, **3**, 932–938.
9 R. Rosenfeld, *Proc. IEEE Inst. Electr. Electron. Eng.*, 2000, **88**, 1270–1278.
10 T. Mikolov, K. Chen, G. Corrado and J. Dean, *International Conference on Learning Representations (ICLR 2013)*, 2013.
11 J. Pennington, R. Socher and others, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

12 T. Mikolov, M. Karafiát, L. Burget, J. Černocký and S. Khudanpur, *Conf Int Speech Commun Assoc*, 2010, 1045–1048.
13 S. Hochreiter and J. Schmidhuber, *Neural Comput.*, 1997, **9**, 1735–1780.
14 F. A. Gers, J. Schmidhuber and F. Cummins, *Neural Comput.*, 2000, **12**, 2451–2471.
15 K. Cho, B. Van Merriënboer, Ç. Gulçehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, *EMNLP*, 2014, pp. 1724–1734.
16 J. Chung, C. Gulcehre, K. Cho and Y. Bengio, *arXiv*, 2014, preprint, arXiv:1412.3555, DOI: 10.48550/arXiv.1412.3555.
17 Y. Bengio, P. Simard and P. Frasconi, *IEEE Trans. Neural Netw.*, 1994, **5**, 157–166.
18 R. Pascanu, T. Mikolov and Y. Bengio, *ICML*, 2012, 28, 1310–1318.
19 J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, *NAACL*, Stroudsburg, PA, USA, 2019, pp. 4171–4186.
20 A. Radford, J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever, *OpenAI Blog*, 2019.
21 J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu and D. Amodei, *arXiv*, 2020, preprint, arXiv:2001.08361, DOI: 10.48550/arXiv.2001.08361.
22 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.
23 A. Gupta, A. T. Müller, B. J. H. Huisman, J. A. Fuchs, P. Schneider and G. Schneider, *Mol. Inform.*, 2018, **37**, 1700111.
24 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.
25 S. Chithrananda, G. Grand and B. Ramsundar, *arXiv*, 2020, preprint, arXiv:2010.09885, DOI: 10.48550/arXiv.2010.09885.
26 W. Ahmad, E. Simon, S. Chithrananda, G. Grand and B. Ramsundar, *arXiv*, 2022, arXiv:2209.01712, DOI: 10.48550/arXiv.2209.01712.
27 R. Irwin, S. Dimitriadis, J. He and E. J. Bjerrum, *Mach. Learn.: Sci. Technol.*, 2022, **3**, 015022.
28 C. Edwards, T. Lai, K. Ros, G. Honke and others, *EMNLP*, 2022, pp. 375–413.
29 J. Mao, T. Sui, K.-H. Cho, K. T. No, J. Wang and D. Shan, *Mol. Divers.*, 2026, **30**(2), 1913–1921.
30 N. O'Boyle and A. Dalke, *ChemRxiv*, 2018, DOI: 10.26434/chemrxiv.7097960.v1.
31 M. Krenn, F. Häse, A. Nigam, P. Friederich and A. Aspuru-Guzik, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 045024.
32 A. H. Cheng, A. Cai, S. Miret, G. Malkomes, M. Phielipp and A. Aspuru-Guzik, *Digit. Discov.*, 2023, **2**, 748–758.
33 S. Piao, J. Choi, S. Seo and S. Park, *Appl. Intell.*, 2023, **53**, 25868–25880.
34 C. Rios-Martinez, N. Bhattacharya, A. P. Amini, L. Crawford and K. K. Yang, *PLoS Comput. Biol.*, 2023, **19**, e1011162.
35 Q. Lai, S. Yao, Y. Zha, H. Zhang, H. Zhang, Y. Ye, Y. Zhang, H. Bai and K. Ning, *Nucleic Acids Res.*, 2025, **53**, gkaf305.
36 R. Singh, S. Sledzieski, B. Bryson, L. Cowen and B. Berger, *Proc. Natl. Acad. Sci. U. S. A.*, 2023, **120**, e2220778120.
37 F. Huber, L. Ridder, S. Verhoeven, J. H. Spaaks, F. Diblen, S. Rogers and J. J. J. van der Hooft, *PLoS Comput. Biol.*, 2021, **17**, e1008724.
38 E. Litsa, V. Chenthamarakshan, P. Das and L. Kavraki, *Commun. Chem.*, 2023, **6**, 132.
39 M. Alberts, F. Zipoli and A. C. Vaucher, *AI for Accelerated Materials Design-NeurIPS 2023 Workshop*, 2023.
40 S. Liu, W. Nie, C. Wang, J. Lu, Z. Qiao, L. Liu, J. Tang, C. Xiao and A. Anandkumar, *Nat. Mach. Intell.*, 2023, **5**, 1447–1457.
41 E. J. Bjerrum and R. Threlfall, *arXiv*, 2017, preprint, arXiv:1705.04612, DOI: 10.48550/arXiv.1705.04612.
42 J. Arús-Pous, S. V. Johansson, O. Prykhodko, E. J. Bjerrum, C. Tyrchan, J.-L. Reymond, H. Chen and O. Engkvist, *J. Cheminform.*, 2019, **11**, 71.
43 A. Holtzman, J. Buys, L. Du, M. Forbes and Y. Choi, *arXiv*, 2019, preprint, arXiv:1904.09751, DOI: 10.48550/arXiv.1904.09751.
44 Y. Yang, S. Zheng, S. Su, C. Zhao, J. Xu and H. Chen, *Chem. Sci.*, 2020, **11**, 8312–8322.
45 V. Fialková, J. Zhao, K. Papadopoulos, O. Engkvist, E. J. Bjerrum, T. Kogej and A. Patronov, *J. Chem. Inf. Model.*, 2022, **62**, 2046–2063.
46 M. Olivecrona, T. Blaschke, O. Engkvist and H. Chen, *J. Cheminform.*, 2017, **9**, 48.
47 T. Blaschke, J. Arús-Pous, H. Chen, C. Margreitter, C. Tyrchan, O. Engkvist, K. Papadopoulos and A. Patronov, *J. Chem. Inf. Model.*, 2020, **60**, 5918–5922.
48 H. H. Loeffler, J. He, A. Tibo, J. P. Janet, A. Voronov, L. H. Mervin and O. Engkvist, *J. Cheminform.*, 2024, **16**, 20.
49 M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov and L. Zettlemoyer, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2020, pp. 7871–7880.
50 C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu, *J. Mach. Learn. Res.*, 2020, **21**, 5485–5551.
51 J. He, E. Nittinger, C. Tyrchan, W. Czechtizky, A. Patronov, E. J. Bjerrum and O. Engkvist, *J. Cheminform.*, 2022, **14**, 18.
52 M. A. Stravs, K. Dührkop, S. Böcker and N. Zamboni, *Nat. Methods*, 2021, **19**, 865–870.
53 A. D. Shrivastava, N. Swainston, S. Samanta, I. Roberts, M. Wright Muelas and D. B. Kell, *Biomolecules*, 2021, **11**, 1793.
54 F. Hu, M. S. Chen, G. M. Rotskoff, M. W. Kanan and T. E. Markland, *ACS Cent. Sci.*, 2024, **10**, 2162–2170.
55 G. D. Hannigan, D. Prihoda, A. Palicka, J. Soukup, O. Klempir, L. Rampula, J. Durcak, M. Wurst, J. Kotowski, D. Chang, R. Wang, G. Piizzi, G. Temesi, D. J. Hazuda, C. H. Woelk and D. A. Bitton, *Nucleic Acids Res.*, 2019, **47**, e110.
56 S. Zheng, T. Zeng, C. Li, B. Chen, C. W. Coley, Y. Yang and R. Wu, *Nat. Commun.*, 2022, **13**, 3342.
57 F. Huber, S. van der Burg, J. J. J. van der Hooft and L. Ridder, *J. Cheminform.*, 2021, **13**, 84.
58 K. Dührkop, H. Shen, M. Meusel, J. Rousu and S. Böcker, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 12580–12585.

59 F. Wang, J. Liigand, S. Tian, D. Arndt, R. Greiner and D. S. Wishart, *Anal. Chem.*, 2021, **93**, 11692–11700.
60 A. Young, H. L. Röst and B. Wang, *Nat. Mach. Intell.*, 2024, **6**, 404–416.
61 H. Chen, T. Liang, K. Tan, A. Wu and X. Lu, *J. Cheminform.*, 2024, **16**, 132.
62 B. Xu, Y. Lu, C. Li, L. Yue, X. Wang, N. Hao, T. Fu and J. Chen, *arXiv*, 2024, preprint, arXiv:2408.05696, DOI: 10.48550/arXiv.2408.05696.
63 L. Chen, X. Tan, D. Wang, F. Zhong, X. Liu, T. Yang, X. Luo, K. Chen, H. Jiang and M. Zheng, *Bioinformatics*, 2020, **36**, 4406–4414.
64 G. W. Kyro, A. M. Smaldone, Y. Shee, C. Xu and V. S. Batista, *J. Chem. Inf. Model.*, 2025, **65**, 2395–2415.
65 K. Huang, C. Xiao, L. Glass and J. Sun, *Bioinformatics*, 2020, **37**, 830–836.
66 B. Gao, B. Qiang, H. Tan, M. Ren, Y. Jia, M. Lu, J. Liu, W.-Y. Ma and Y. Lan, *Adv. Neural Inf. Process. Syst.*, 2023, 44595–44614.
67 C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen and T.-Y. Liu, *Adv. Neural Inf. Process. Syst.*, 2021, 28877–28888.
68 G. Chilingaryan, H. Tamoyan, A. Tevosyan, N. Babayan, K. Hambardzumyan, Z. Navoyan, A. Aghajanyan, H. Khachatrian and L. Khondkaryan, *J. Chem. Inf. Model.*, 2024, **64**, 5832–5843.
69 D. W. P. Tay, N. Z. X. Yeo, K. Adaikkappan, Y. H. Lim and S. J. Ang, *Sci. Data*, 2023, **10**, 296.
70 K. Sakano, K. Furui and M. Ohue, *J. Supercomput.*, 2025, **81**, 352.
71 T. Ochiai, T. Inukai, M. Akiyama, K. Furui, M. Ohue, N. Matsumori, S. Inuki, M. Uesugi, T. Sunazuka, K. Kikuchi, H. Kakeya and Y. Sakakibara, *Commun. Chem.*, 2023, **6**, 249.
72 M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W.-T. Liu, M. Crüsemann, P. D. Boudreau, E. Esquenazi, M. Sandoval-Calderón, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C.-C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, E. E. Carlson, B. Aigle, C. F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B. T. Murphy, L. Gerwick, C.-C. Liaw, Y.-L. Yang, H.-U. Humpf, M. Maansson, R. A. Keyzers, A. C. Sims, A. R. Johnson, A. M. Sidebottom, B. E. Sedio, A. Klitgaard, C. B. Larson, C. A. B. P, D. Torres-Mendoza, D. J. Gonzalez, D. B. Silva, L. M. Marques, D. P. Demarque, E. Pociute, E. C. O'Neill, E. Briand, E. J. N. Helfrich, E. A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J. J. Kharbush, Y. Zeng, J. A. Vorholt, K. L. Kurita, P. Charusanti, K. L. McPhail, K. F. Nielsen, L. Vuong, M. Elfeki, M. F. Traxler, N. Engene, N. Koyama, O. B. Vining, R. Baric, R. R. Silva, S. J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P. G. Williams, J. Dai, R. Neupane, J. Gurr, A. M. C. Rodríguez, A. Lamsa, C. Zhang, K. Dorrestein, B. M. Duggan, J. Almaliti, P.-M. Allard, P. Phapale, L.-F. Nothias, T. Alexandrov, M. Litaudon, J.-L. Wolfender, J. E. Kyle, T. O. Metz, T. Peryea, D.-T. Nguyen, D. VanLeer, P. Shinn, A. Jadhav, R. Müller, K. M. Waters, W. Shi, X. Liu, L. Zhang, R. Knight, P. R. Jensen, B. O. Palsson, K. Pogliano, R. G. Linington, M. Gutiérrez, N. P. Lopes, W. H. Gerwick, B. S. Moore, P. C. Dorrestein and N. Bandeira, *Nat. Biotechnol.*, 2016, **34**, 828–837.
73 T. Butler, A. Frandsen, R. Lightheart, B. Bargh, B. Kerby, K. West, J. Davison, J. Taylor, C. Krettler, T. J. Bollerman, G. Voronov, K. Moon, T. Kind, P. Dorrestein, A. Allen, V. Colluru and D. Healey, *ChemRxiv*, 2023, DOI: 10.26434/chemrxiv-2023-vsmpx-v4.
74 M. Priessner, R. Lewis, J. P. Janet, I. Lemurell, M. Johansson, J. Goodman and A. Tomberg, *ChemRxiv*, 2024, DOI: 10.26434/chemrxiv-2024-zmmnw-v2.
75 J. Li and X. Jiang, *Wirel. Commun. Mob. Comput.*, 2021, **181815**, 1–7.
76 D. van Tilborg, A. Alenicheva and F. Grisoni, *J. Chem. Inf. Model.*, 2022, **62**, 5938–5951.
77 K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen and R. Barzilay, *J. Chem. Inf. Model.*, 2019, **59**, 3370–3388.
78 Z. A. Rollins, A. C. Cheng and E. Metwally, *J. Cheminform.*, 2024, **16**, 56.
79 S. Feng, L. Yang, Y. Huang, Y. Ni, W. Ma and Y. Lan, *arXiv*, 2023, preprint, arXiv:2310.14216, DOI: 10.48550/arXiv.2310.14216.
80 M. Yang, H. Sun, X. Liu, X. Xue, Y. Deng and X. Wang, *Brief. Bioinform.*, 2023, **24**, bbad185.
81 Y. Wang, H. Zhao, S. Sciabola and W. Wang, *Molecules*, 2023, **28**, 4430.
82 L. Yang, C. Jin, G. Yang, Z. Bing, L. Huang, Y. Niu and L. Yang, *Phys. Chem. Chem. Phys.*, 2023, **25**, 2377–2385.
83 D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, A. Kadurin, S. Johansson, H. Chen, S. Nikolenko, A. Aspuru-Guzik and A. Zhavoronkov, *Front. Pharmacol*, 2020, **11**, 565644.
84 N. Brown, M. Fiscato, M. H. S. Segler and A. C. Vaucher, *J. Chem. Inf. Model.*, 2019, **59**, 1096–1108.
85 T. Kawano, T. Shiraishi, T. Kuzuyama and M. Umemura, *PLoS Comput. Biol.*, 2026, **22**, e1013181.
86 D. Merk, F. Grisoni, L. Friedrich and G. Schneider, *Commun. Chem.*, 2018, **1**, 68.
87 V. Chandrasekhar, K. Rajan, S. R. S. Kanakam, N. Sharma, V. Weißenborn, J. Schaub and C. Steinbeck, *Nucleic Acids Res.*, 2025, **53**, D634–D643.
88 E. F. Poynton, J. A. van Santen, M. Pin, M. M. Contreras, E. McMann, J. Parra, B. Showalter, L. Zaroubi, K. R. Duncan and R. G. Linington, *Nucleic Acids Res.*, 2025, **53**, D691–D699.
89 A. Rutz, M. Sorokina, J. Galgonek, D. Mietchen, E. Willighagen, A. Gaudry, J. G. Graham, R. Stephan, R. Page, J. Vondrášek, C. Steinbeck, G. F. Pauli,

J.-L. Wolfender, J. Bisson and P.-M. Allard, *eLife*, 2022, **11**, e70780.
90 K. Gallo, E. Kemmler, A. Goede, F. Becker, M. Dunkel, R. Preissner and P. Banerjee, *Nucleic Acids Res.*, 2023, **51**, D654–D659.
91 B. I. Tingle, K. G. Tang, M. Castanon, J. J. Gutierrez, M. Khurelbaatar, C. Dandarchuluun, Y. S. Moroz and J. J. Irwin, *J. Chem. Inf. Model.*, 2023, **63**, 1166–1176.
92 B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, M. de Veij, H. Ioannidis, D. M. Lopez, J. F. Mosquera, M. P. Magarinos, N. Bosc, R. Arcila, T. Kizilören, A. Gaulton, A. P. Bento, M. F. Adasme, P. Monecke, G. A. Landrum and A. R. Leach, *Nucleic Acids Res.*, 2024, **52**, D1180–D1192.
93 J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh and P. Das, *arXiv*, 2021, preprint, arXiv:2106.09553, DOI: 10.48550/arXiv.2106.09553.
94 Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu and V. Pande, *Chem. Sci.*, 2018, **9**, 513–530.
95 RDKit: Open-source cheminformatics, https://www.rdkit.org.
96 M. Swain, MolVS: Molecule Validation and Standardization, https://github.com/mcs07/MolVS, 2018, Accessed: 2024-1-15.
97 P. Ertl, S. Roggo and A. Schuffenhauer, *J. Chem. Inf. Model.*, 2008, **48**, 68–74.
98 E. J. Bjerrum, *arXiv*, 2017, preprint, arXiv:1703.07076, DOI: 10.48550/arXiv.1703.07076.
99 X. Li and D. Fourches, *J. Chem. Inf. Model.*, 2021, **61**, 1560–1569.
100 U. V. Ucak, I. Ashyrmamatov and J. Lee, *J. Cheminform.*, 2023, **15**, 55.
101 G. W. Bemis and M. A. Murcko, *J. Med. Chem.*, 1996, **39**, 2887–2893.
102 Q. Guo, S. Hernandez-Hernandez and P. J. Ballester, *33rd International Conference on Artificial Neural Networks (ICANN 2024), Lecture Notes in Computer Science*, 2024, vol. 15025, pp. 58–72.
103 C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, 1997, **23**, 3–25.
104 X. Chen, S. Wang, B. Fu, M. Long and J. Wang, *Adv. Neural Inf. Process. Syst.*, 2019, pp. 1906–1916.
105 E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang and W. Chen, *International Conference on Learning Representations (ICLR 2022)*, 2022.
106 Y. Yoshikai, T. Mizuno, S. Nemoto and H. Kusuhara, *Nat. Commun.*, 2024, **15**, 1197.
107 J. W. Furness, K. B. Moore 3rd and A. Bochevarov, *J. Chem. Inf. Model.*, 2025, **65**, 10289–10310.
108 J. Wang, R. Qin, M. Wang, M. Fang, Y. Zhang, Y. Zhu, Q. Su, Q. Gou, C. Shen, O. Zhang, Z. Wu, D. Jiang, X. Zhang, H. Zhao, J. Ge, Z. Wu, Y. Kang, C.-Y. Hsieh and T. Hou, *Nat. Commun.*, 2025, **16**, 4416.
109 R. E. Taylor, Y. Chen, G. M. Galvin and P. K. Pabba, *Org. Biomol. Chem.*, 2004, **2**, 127–132.
110 M. Reese, V. M. Sánchez-Pedregal, K. Kubicek, J. Meiler, M. J. J. Blommers, C. Griesinger and T. Carlomagno, *Angew Chem. Int. Ed. Engl.*, 2007, **46**, 1864–1868.
111 D. W. Carney, K. R. Schmitz, J. V. Truong, R. T. Sauer and J. K. Sello, *J. Am. Chem. Soc.*, 2014, **136**, 1922–1929.
112 M. Bedrosian and H. Khachatrian, *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025.
113 M. H. Maeda and K. Kondo, *J. Chem. Inf. Model.*, 2013, **53**, 527–533.
114 L. Ge, Q. Gao, J. He, X. Wang, J. Huang, H. Zhang and Z. Qin, *ACS Omega*, 2025, **10**, 5105–5110.
115 Z. Liu, W. Zhang, Y. Xia, L. Wu, S. Xie, T. Qin, M. Zhang and T.-Y. Liu, *Proceedings of the 61st Annual Meeting of the*, Association for Computational Linguistics, 2023, pp. 1606–1616.
116 C. Edwards, C. Han, G. Lee, T. Nguyen, S. Szymkuć, C. K. Prasad, B. Jin, J. Han, Y. Diao, G. Liu, H. Peng, B. A. Grzybowski, M. D. Burke and H. Ji, *arXiv*, 2026, preprint, arXiv:2505.12565, DOI: 10.48550/arXiv.2505.12565.
117 H. Heberle, L. Zhao, S. Schmidt, T. Wolf and J. Heinrich, *J. Cheminform.*, 2023, **15**, 2.