



Perspective

Artificial intelligence-driven paradigms in polysaccharide structure-activity relationships: a new revolution

Shi-Kang Chen^{1, 2, 3, 4, 5, 6, †}, Tao Li^{7, †}, Qiang Li⁸, Lei Duan⁹, Chen Chen¹⁰, Zhi Huang¹¹, Yan-Long Cui¹², Zhi-Fei Fu¹³, Xi Chen¹⁴, Xiao-Bo Deng², Ning-Xian Yang¹⁵, Fu-Li Yu¹⁶, Pei-Lin Chen¹⁷, Yan Meng¹⁸, Xing-Fang Zhang¹⁹, Qiang Ren²⁰, Ren-Peng Du²¹, Guang-Ping Lü²², Ying-Yue Zhang²³, Yu-Ting Su²⁴, Hong-Tao Bi²⁵, Li-Li Guan²⁶, Lei Feng²⁷, Tao Hu²⁸, Hazem Golshany^{1, 2, 3, 4, 5, 6}, Jia-Lun Duan²⁹ (✉), Wen-Yi Kang^{1, 2, 3, 4, 5, 6} (✉)

¹ National R & D Center for Edible Fungus Processing Technology, Henan University, Kaifeng 475004, China

² College of Agriculture, Henan University, Kaifeng 475004, China

³ Joint International Research Laboratory of Food & Medicine Resource Function, Kaifeng 475004, China

⁴ Functional Food Engineering Technology Research Center, Kaifeng 475004, China

⁵ Henan Key University-Enterprise R & D Center for Special Foods, Kaifeng 475004, China

⁶ Key Laboratory for Innovation and Manufacturing of Specialty Plant Health Foods in Henan Province, Kaifeng 475004, China

⁷ Qilin Hesheng Network Technology INC, Beijing 100012, China

⁸ Department of Critical Care Medicine, Beijing Hospital, National Center for Gerontology; National Clinical Research Center for Gerontology; The Key Laboratory of Geriatrics of NHC; Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, China

⁹ Department of Civil and Environmental Engineering, Duke University Durham, Durham 27705, USA

¹⁰ Shaanxi Province Key Laboratory of Bio-resources; Qinba Mountain Area Collaborative Innovation Center of Bioresources Comprehensive Development; Qinba State Key Laboratory of Biological Resources and Ecological Environment (Incubation); College of Biological Science and Engineering, Shaanxi University of Technology, Hanzhong 723001, China

¹¹ School of Light Industry and Food Engineering, Guangxi University, Nanning 530004, China

¹² College of Food Science, Henan Institute of Science and Technology, Xinxiang 453003, China

¹³ National Key Laboratory of Chinese Medicine Modernization, State Key Laboratory of Component-based Chinese Medicine, Tianjin University of Traditional Chinese Medicine, Tianjin 301617, China

¹⁴ School of Food Science and Engineering, Wuhan Polytechnic University, Wuhan 430023, China

¹⁵ School of Public Health, State Key Laboratory of Discovery and Utilization of Functional Components in Traditional Chinese Medicine, Guizhou Medical University, Guiyang 561113, China

¹⁶ School of Basic Medical Science, Guizhou Medical University, Guiyang 561113, China

¹⁷ College of Oceanology and Food Sciences, Quanzhou Normal University, Quanzhou 362000, China

¹⁸ School of Pharmacy, Hubei University of Chinese Medicine, Wuhan 430065, China

¹⁹ Department of Pharmacy, Xijing Hospital, Fourth Military Medical University, Xi'an 710032, China

²⁰ Department of Pharmacy, Jining Medical University, Rizhao 276826, China

²¹ College of Heilongjiang Province, School of Life Sciences, Heilongjiang University, Harbin 150000, China

²² School of Food Science and Pharmaceutical Engineering, Nanjing Normal University, Nanjing 210023, China

²³ School of Life Sciences and Chemical Engineering, Jiangsu Second Normal University, Nanjing 211200, China

²⁴ School of Food Science, Guangdong Pharmaceutical University, Zhongshan 510006, China

²⁵ Qinghai Provincial Key Laboratory of Tibetan Medicine Pharmacology and Safety Evaluation, Northwest Institute of Plateau Biology, Chinese Academy of Science, Xining 810001, China

²⁶ College of Life Sciences, Engineering Research Center of Bioreactor and Pharmaceutical Development, Ministry of Education, Jilin Agricultural University, Changchun 130118, China

²⁷ Department of Pharmacy, the First Hospital of China Medical University, Shenyang 110001, China

²⁸ School of Management Science and Engineering, Guizhou University of Finance and Economics, Guiyang 550025, China

²⁹ Jiangsu Collaborative Innovation Center of Chinese Medicinal Resources Industrialization, National and Local Collaborative Engineering Center of Chinese Medicinal Resources Industrialization and Formulae Innovative Medicine, Jiangsu Provincial Key Laboratory of Functional Substances in Traditional Chinese Medicine Formulae and Innovative Drug Discovery, Nanjing University of Chinese Medicine, Nanjing 210023, China

[†] Shi-Kang Chen and Tao Li contributed equally to this work.

Received: 12 July 2026 / **Revised:** 20 July 2026 / **Accepted:** 20 July 2026



1 Introduction and background

Polysaccharides are macromolecular carbohydrates assembled from monosaccharide residues through glycosidic bonds. They may be classified by composition (homopolysaccharides or heteropolysaccharides), source (plant, fungal, algal, microbial or animal), topology (linear or branched), charge and substitution (neutral, acidic, sulfated, phosphorylated, acetylated or otherwise modified), and biological role (storage, structural, recognition or extracellular-matrix functions). Starch, cellulose and glycogen illustrate compositionally simple but architecturally distinct glucans, whereas pectins, hemicelluloses, microbial exopolysaccharides, glycosaminoglycans and marine sulfated polysaccharides occupy far broader compositional and conformational spaces. Molecular-weight distribution, branching density, linkage position and anomeric configuration jointly shape solubility, rheology, supramolecular assembly, enzymatic accessibility, receptor binding and microbiota utilization. Consequently, two samples described by the same source name can exhibit different activities because of cultivar, harvest, extraction, purification, processing or batch history. In recent decades, however, polysaccharide research has progressively evolved into a systematic discipline centered on the “structure–property–function” relationship. Nevertheless, the inherent heterogeneity, hierarchical conformational complexity, and highly nonlinear structure–activity relationships (SAR) of polysaccharides have posed substantial challenges to traditional research paradigms that rely heavily on empirical observations and single-factor analyses, limiting further advances in mechanistic understanding and functional development.

Against this backdrop, artificial intelligence (AI) has emerged as a transformative technological approach for polysaccharide research. The concept of digital intelligent polysaccharides has consequently gained increasing attention, referring to the deep integration of digital intelligence technologies—including big data analytics, cloud computing, machine learning, deep learning, and generative AI—throughout the entire polysaccharide research pipeline. This framework aims to achieve intelligent upgrading across all stages, from raw material acquisition and structural characterization to structure–activity analysis, mechanistic elucidation, and functional product development. To date, AI applications in polysaccharide science have begun to encompass multiple areas, including quality control, extraction optimization, SAR modeling, assisted structural characterization, and the development of functional products.

Looking forward, AI is expected not only to accelerate the translation of knowledge “from structure to function” in polysaccharide science, but also to fundamentally transform the research paradigm—from experience-driven to data-driven discovery, from isolated analyses to end-to-end integration, and from passive characterization to rational design—thereby ushering in a new revolution in the study of polysaccharide structure–activity relationships.

2 Challenges and technical barriers in polysaccharide structure–activity relationship research

2.1 Challenges in polysaccharide acquisition

Polysaccharide acquisition represents the starting point for AI-driven modeling; however, it is constrained by two major

bottlenecks: the difficulty of chemical synthesis and the poor batch-to-batch consistency of natural products. From the perspective of AI modeling, these challenges can be addressed through two complementary strategies. The first is the precise reconstruction of biosynthetic pathways through synthetic biology and synthetic chemistry, enabling the standardized and rational production of polysaccharides. The second is the integration of active learning strategies, whereby AI autonomously recommends experimental samples with the highest expected information gain, thereby maximizing structural diversity coverage at minimal cost and improving the quality and representativeness of training datasets at the source.

2.2 Structural characterization under ambiguity

Polysaccharide structural elucidation currently confronts several formidable bottlenecks: the inherent difficulty in effectively separating and resolving heterogeneous mixtures, the lack of a universal methodological framework to guarantee the accuracy of monosaccharide composition analysis, and the severe crowding and overlapping of nuclear magnetic resonance (NMR) spectral peaks in complex polysaccharides that impedes reliable assignment.

To overcome these challenges, the critical imperative is to establish a systematic elucidation strategy that integrates high-resolution mass spectrometry (HRMS), high-field NMR, chemical and enzymatic degradation, and computational simulations. In this context, AI-driven multi-spectroscopic collaborative elucidation offers a highly viable trajectory. Utilizing multimodal fusion models, this approach maps NMR chemical shifts, MS fragmentation behaviors, and infrared (IR) characteristic peak morphologies into a unified, shared embedding space, thereby realizing reciprocal complementation and cross-validation across distinct spectroscopic modalities.

Concurrently, generative AI architectures can be introduced to execute an inverse spectral validation loop: deriving theoretical spectra from generated candidate structures and benchmark-comparing them against experimental data to close a robust “hypothesis-validation” structural confirmation loop. Such a strategy holds strong potential to bridge traditionally isolated spectroscopic datasets, fundamentally enhancing both the throughput and fidelity of polysaccharide structural elucidation.

2.3 Insights into SAR

The core challenge in deciphering the SAR of polysaccharides lies in identifying the key structural determinants that truly drive biological effects from amidst an overwhelming complexity of structural data. A persistent bottleneck in current research is that while numerous studies have validated correlations between structure and activity, they have generally failed to crystallize clear, universal laws analogous to the sequence–function paradigms established for proteins. Urban *et al.*^[1] pointed out that mainstream approaches conventionally simplify carbohydrate backbones into monosaccharide node graphs. However, this abstraction discards the intrinsic atomic-level structural information within the monosaccharide units, which is precisely what is critical to their functional outcomes.

More critically, statistically significant SAR correlations do not inherently imply causality; relying solely on correlation analysis can easily lead to the misinterpretation of concomitant structural variations as the actual drivers of altered bioactivity. The integration of causal representation learning and counterfactual

inference holds promise in enabling researchers to look beyond superficial correlations and discern which structural modifications truly drive changes in activity. This will provide a more robust logical foundation for precise, function-driven design. Furthermore, the lack of comparability among bioactivity data presents another formidable challenge. For the same biological indicator across different studies, experimental conditions, positive controls, and data reporting formats vary widely. This multi-dimensional heterogeneity severely hinders the integration of cross-literature data, directly impeding the construction of standardized datasets essential for AI modeling.

2.4 A minimum information standard for AI-ready polysaccharide data

Data standardization is the most immediate community bottleneck. The MIRAGE (minimum information required for a glycomics experiment) initiative established minimum reporting requirements for glycomics experiments, including sample preparation and MS workflows. MIAME (minimum information required for a glycomics experiment) and MIAPE (minimum information about a microarray experiment) show how community-agreed metadata can make microarray and proteomics results interpretable and reusable, while the FAIR (Findable, Accessible, Interoperable, and Reusable) principles emphasize findability, accessibility, interoperability and reusability. Polysaccharide AI requires these ideas to be extended beyond analytical method metadata to encompass the full sample-structure-assay chain.

3 Methodological foundations of AI in polysaccharide research

The application of AI to polysaccharide research necessitates addressing foundational challenges at two distinct levels (As

shown in Table 1): first, how to represent the intricate chemical architecture and biological modalities of polysaccharides into computer-interpretable formats; and second, determining the optimal learning strategies to deploy when confronting these complex datasets.

3.1 Digital representation of polysaccharide data

3.1.1 Types and sources of polysaccharide data

The data modalities implicated in polysaccharide research can be broadly categorized into three distinct hierarchical levels. At the foundational layer is the structural characterization data, which encompasses monosaccharide composition, glycosidic linkage patterns, branching topologies, molecular weight distributions, and chemical modifications, along with raw spectral data from NMR, MS, and IR spectroscopy. The intermediate layer comprises physicochemical property data, such as solubility, viscosity, rheological characteristics, and thermal stability. At the apex resides the functional and mechanistic data, which includes cell-level immunomodulatory indices, animal-model efficacy evaluations, and multi-omics information across transcriptomics, proteomics, metabolomics, and microbiomics. Together, these three vertical layers sustain a comprehensive information chain spanning from chemical architecture to biological function.

However, these datasets exhibit profound phenotypic heterogeneity, manifesting as diverse modalities such as structured tables, spectral signals, molecular topological structures, and high-dimensional omics matrices. Concurrently, polysaccharide research is universally confronted with a “high-dimensional, small-sample” dilemma—wherein the structural features of a single sample can span hundreds of dimensions, whereas high-quality annotated samples typically number only in the dozens to hundreds. Furthermore, the selection of assay endpoints and data reporting formats for bioactivity evaluations

Table 1 Proposed minimum information for an AI-ready polysaccharide record.

Domain	Required fields	Why the model needs them	Recommended implementation
Biological source	Species/strain, tissue or matrix, voucher or accession, origin, cultivation/harvest, biological replicate	Prevents source and taxonomic leakage; supports external validity	Controlled taxonomy identifiers; persistent sample ID
Extraction and processing	Pretreatment, solvent, solid/liquid ratio, pH, temperature, time, energy input, enzymes, purification/fractionation, drying and storage	Distinguishes chemistry from process history and batch effects	Structured protocol plus versioned workflow
Composition and purity	Monosaccharide composition with method and uncertainty, protein/phenolic/endotoxin/nucleic-acid contaminants, ash and moisture	Contaminants can drive apparent bioactivity	Values, units, detection limits and raw/processed files
Primary structure	Linkage positions, anomeric configuration, branching, sequence or repeating unit, terminal groups and chemical substitutions	Provides the causal feature space for SAR modeling	IUPAC (international union of pure and applied chemistry)-condensed plus GlycoCT/WURCS where applicable
Distribution and higher order	Number- and weight-average molecular weight, dispersity, conformation, aggregation, particle size and solvent state	Many food and medicinal polysaccharides are ensembles rather than single structures	Distribution data, not only mean values; temperature, ionic strength and concentration
Analytical evidence	Raw and processed NMR, MS, IR, chromatography and scattering data; instrument, acquisition and preprocessing parameters	Enables re-analysis, multimodal learning and uncertainty audit	MIRAGE-aligned metadata; repository accession
Bioactivity assay	System, dose, exposure, vehicle, positive/negative controls, endpoint definition, normalization, replicate type and n	Makes outcomes comparable and exposes assay confounding	Controlled endpoint vocabulary and complete protocol
Model provenance	Train/validation/test split, duplicate and similarity audit, code/model version, hyperparameters, random seeds and compute	Enables reproducibility and detects leakage	Version-controlled workflow and immutable split manifest
Validation and uncertainty	Internal and external performance, calibration, confidence interval, applicability domain, failed cases and experimental follow-up	Separates reliable prediction from extrapolation	Model card linked to sample and assay records

Note: GlyTouCan, GlycoPOST, GlyGen, GlyCosmos and the Carbohydrate Structure Database demonstrate complementary components of this infrastructure. A practical next step is not another disconnected database, but a federated record model that links persistent structure identifiers to raw analytical evidence, process metadata, bioactivity assays and physical reference materials.

varies across laboratories, posing a formidable barrier to cross-study data integration. Consequently, transforming the aforementioned heterogeneous information into a unified, comparable, and computer-interpretable format constitutes a paramount prerequisite for the integration of AI into polysaccharide research.

3.1.2 Digital encoding of structural features

Unlike proteins or nucleic acids, polysaccharide structures lack a natural, sequence-based coding paradigm; consequently, their digital representation necessitates the concurrent integration of multiple dimensions, including composition, linkage, branching, modification, and conformation. Currently, four predominant encoding strategies have emerged: 1) Vectorization of compositional and linkage features. Monosaccharide compositions and their ratios, glycosidic linkage types, branching densities, and degrees of modification (e.g., sulfation, acetylation) are directly encoded into fixed-length numerical vectors. This approach is intuitive and straightforward, making it highly compatible with conventional machine learning models. 2) Graph structure representation. By treating monosaccharide residues as nodes and glycosidic linkages as edges, the branching topology of a polysaccharide is encoded as a graph, which aligns more closely with its intrinsic structural essence. Leveraging this formulation, Graph Neural Networks (GNNs) can directly learn from these graph architectures. Moving a step further, GNN framework based on combinatorial complexes and higher-order message passing, effectively capturing subtle structural nuances that remain indistinguishable *via* traditional methods^[2]. 3) Atom-Level heterogeneous graph encoding. The GlycanAA framework^[1], designates monosaccharide residues as global nodes and the constituent atoms within those residues as local nodes. Through hierarchical message passing, this framework simultaneously captures both the inter-residue topological relationships and intra-residue atomic chemical interactions, thereby overcoming the limitations of previous methods that overlooked atom-level granularity. 4) Direct encoding of spectral signals. This strategy directly utilizes raw signals or preprocessed feature matrices from NMR, MS, and IR spectroscopy as inputs for deep learning models, bypassing the labor-intensive step of manual peak extraction.

At the foundational model level, GlycanGT^[3], warrants significant attention. This is a graph Transformer-based pre-trained model for glycan chains, leveraging millions of unlabeled glycan structures *via* self-supervised pre-training. It has outperformed existing methods in tasks such as immunogenicity prediction, and its learned embedding space successfully clusters structures in a biologically meaningful manner^[3]. The success of the “pre-training and fine-tuning” paradigm in the glycan domain provides a potent new foundational toolkit for polysaccharide structural encoding.

3.1.3 Fusion encoding of multimodal data

The objective of multimodal fusion is to integrate structural attributes, spectral datasets, literature-derived knowledge, and multi-omics data into a unified computational framework. In terms of spectral data, deep learning methodologies have advanced to directly extract structural features from MS, NMR, and IR spectra. The GlycoBERT model, achieves high accuracy in glycan structure prediction from mass spectrometry data; furthermore, its generative counterpart, GlycoBART, goes a step further by demonstrating the capacity to de novo infer unknown

glycan structures from MS profiles^[4]. Concurrently, the synergy between knowledge graphs and large language models (LLMs) significantly enhances domain-specific knowledge retrieval and reasoning capabilities. Regarding multi-omics data, transcriptomic, proteomic, metabolomic, and microbiomic profiles collected post-polysaccharide intervention reflect the host's systemic response across distinct biological dimensions. The UNIQ system, engineered by Song *et al.*^[5], successfully elucidates the correlations between bioactive components of natural products and their targeted biological pathways by integrating genomic, transcriptomic, and metabolomic information. Once these diverse modalities are encoded, they must be aligned and fused within a unified, shared embedding space to establish a robust foundation for subsequent joint learning.

3.2 Machine learning modeling strategies for polysaccharide data

Once data encoding is finalized, the core challenge for modeling strategies shifts to reliably extracting underlying patterns from the “high-dimensional, small-sample” polysaccharide datasets.

3.2.1 Feature dimensionality reduction and critical information extraction

When the feature dimensionality substantially outnumbers the available sample size, direct modeling often leads to severe overfitting. Consequently, feature dimensionality reduction and critical information extraction must be executed prior to or during the modeling process. The choice of these methodologies depends fundamentally on data scale and modality. When annotated samples are limited (typically fewer than 200 cases) and feature dimensions are moderate with explicit chemical interpretations, conventional machine learning models—such as random forest (RF), support vector machines (SVM), gradient boosting decision trees (GBDT), and extreme gradient boosting (XGBoost)—demonstrate robust applicability. These approaches are remarkably resilient to small-sample constraints, exhibit lower computational costs, and can directly identify key structural variables governing bioactivity *via* feature importance rankings.

Conversely, when dealing with unstructured or weakly structured modalities such as raw spectra or multi-omics matrices, and where the sample size is relatively sufficient, deep learning architectures offer distinct advantages. For instance, Convolutional Neural Networks (CNNs) can automatically learn localized peak morphologies from spectral profiles; transformer architectures excel at processing sequential structural information; GNNs can directly capture node relationships within molecular topologies; and autoencoders (AEs) can be deployed for unsupervised dimensionality reduction and anomaly detection.

Therefore, in practical polysaccharide AI modeling, priority should be given to selecting appropriate methods based on data scale and feature modalities, rather than single-mindedly pursuing model complexity. A noteworthy emerging avenue is the deployment of State Space Models (SSMs), such as the Mamba architecture. By processing long-sequence data with linear time complexity, SSMs offer significant efficiency advantages in analyzing large-scale datasets like high-dimensional multi-omics matrices.

3.2.2 Few-Shot learning strategies

The scarcity of annotated data represents a pervasive bottleneck in AI-driven polysaccharide modeling. To alleviate this constraint, transfer learning and meta-learning have emerged as the primary

paradigms. The “pre-training and fine-tuning” framework has recently achieved milestone progress in glycan research. A prime example is GlycanGT, which utilizes massive archives of unlabeled glycan structures during the pre-training phase to master universal structural grammar, and is subsequently fine-tuned on target tasks with limited annotated instances. This paradigm significantly mitigates the heavily entrenched reliance on large-scale labeled datasets^[3]. Similarly, contrastive learning holds strong utility for polysaccharide spectral classification and bioactivity screening. By pre-training a feature extractor on substantial volumes of unlabeled spectral profiles, the architecture can be efficiently fine-tuned on highly constrained annotated cohorts.

Concurrently, meta-learning aims to endow models with the capacity to “learn how to learn,” thereby facilitating rapid adaptation to novel tasks. When confronted with an entirely unprecedented polysaccharide variant characterized by sparse bioactivity data, meta-learning algorithms can rapidly transfer knowledge generalized from previously characterized polysaccharide families. This capacity for cross-type knowledge transfer is of paramount value given the starkly imbalanced data accumulation across heterogeneous subfields of glycobiology.

3.2.3 Model interpretability

In the investigation of polysaccharide SAR, researchers demand more than mere predictive accuracy; they require a definitive understanding of which specific structural determinants truly drive variations in bioactivity. Consequently, model interpretability serves as the pivotal bridge intersecting AI-driven predictions with biological mechanistic elucidations.

Interpretability frameworks, such as SHAP (SHapley Additive exPlanations), systematically quantify the individual contributions of diverse features to the model’s output, thereby enabling researchers to pinpoint critical structural factors, including molecular weight ranges, monosaccharide compositions, glycosidic linkage patterns, and branching densities. Within GNNs, attention mechanisms and graph-explainer algorithms can further spatialize and isolate the structural regions prioritizing the model’s focus, offering actionable hints regarding potential bioactive motifs. Moving a step further, Concept Bottleneck Models (CBMs) offer a paradigm for embedding interpretability natively within the model architecture itself: the network is engineered to first predict human-interpretable intermediate concepts (e.g., discrete molecular weight intervals or branching density echelons), which subsequently drive the final bioactivity prediction. This architecture empowers researchers to directly audit and intervene in the model’s intermediate reasoning, providing a highly transparent toolkit for the interactive exploration of SAR.

3.2.4 Joint inference across multimodal data

Unimodal analysis is inherently constrained to answering “what”—establishing that a specific polysaccharide exhibits a certain bioactivity. In contrast, the objective of cross-modal joint inference is to decipher “why”: how do disparate structural determinants converge to elicit identical phenotypic outcomes? Or, why does the same polysaccharide variant yield fundamentally different metabolic profiles across distinct gut microbiotic backgrounds?

Achieving this objective necessitates constructing robust causal inference chains spanning structural representations, spectral profiles, literature-derived knowledge, and multi-omics data. Currently, relatively independent methodological advancements

have been achieved within individual segments—such as mapping spectra to structures, structures to microbiota, microbiota to metabolisms, and metabolisms to phenotypes; nonetheless, bridging these components into an end-to-end, cross-modal pipeline remains a formidable bottleneck.

From recent developmental trajectories, multi-agent cooperative frameworks offer a highly viable technical pathway to address this challenge. Within this architecture, domain-specific models operate as specialized agents responsible for localized inferences within their respective expertise (e.g., transforming spectra to structures, distilling literature to universal laws, or mapping omics to pathways), while a centralized orchestration layer coordinates cross-modal information alignment and causal chain integration. Rather than relying on a single omnipotent foundation model, this modular architecture achieves complex reasoning through the synergy of distinct expert agents—a design that aligns precisely with the decentralized data infrastructure and highly diverse task scenarios characteristic of contemporary polysaccharide AI research.

3.3 Low-barrier strategies for small datasets

Most experimental groups do not need to pretrain a foundation model. A low-barrier workflow can begin with a curated spreadsheet, explicit data dictionary, regularized regression or XGBoost, repeated nested cross-validation and SHAP-based interpretation. No-code or low-code environments can support exploratory modeling, but split logic and provenance should be documented independently. Glycwork provides reusable glycan processing, motif analysis, trained representations and machine-learning functions for end users. CandyCrunch is available as an open-source package and notebook that can run on a typical laptop or free cloud environment. Web resources such as GlyCosmos, GlyGen, GlyTouCan, GlycoPOST and GLYCAM-Web provide identifiers, curated context, deposited spectra and structure-building services.

Transfer learning is useful when the target task has few labels but shares representation with a larger source task. Pretrained embeddings can be frozen and used with a simple classifier before full fine-tuning is attempted. Data augmentation must preserve chemistry and experimental meaning; arbitrary graph perturbations or synthetic spectra can create invalid examples. Collaborative consortia can also contribute through shared protocols, interlaboratory reference materials, federated analysis and pre-specified external validation rather than by pooling incompatible datasets indiscriminately.

4 Future applications of AI-powered end-to-end polysaccharide research

4.1 Multi-Spectroscopic driven elucidation of complex structure

The conventional workflow for polysaccharide structural elucidation heavily relies on the comprehensive synthesis of multi-spectroscopic evidence, including NMR, MS, and IR spectroscopy. In recent years, deep learning methodologies have been actively driving this workflow toward automation. In NMR data processing, machine learning has been progressively deployed for peak assignment, spectral classification, and computer-assisted structural elucidation. Progress in the MS domain has been particularly striking—transformer-based glycan structure

prediction models can now directly infer glycan architectures from tandem mass spectrometry data. For instance, the GlycoBERT model and its generative counterpart, GlycoBART, have demonstrated high accuracy on benchmarking datasets and have exhibited initial capabilities in *de novo* structural inference from MS profiles. Regarding IR spectroscopy, CNNs coupled with attenuated total reflectance–Fourier transform infrared spectroscopy (ATR-FTIR) and near-infrared (NIR) technologies have enabled rapid prediction of polysaccharide content, source identification, and quality evaluation. A more compelling avenue of research lies in multimodal fusion technologies, which are reshaping the traditionally isolated paradigms of individual spectroscopic analyses. The core strategy entails embedding the complementary information garnered from NMR, MS, and IR spectra—which respectively provide distinct insights into glycosidic linkage modes, molecular weights, and functional groups—into a unified, shared embedding space. Exploiting these inter-spectroscopic complementarities facilitates cross-validation and comprehensive inference, thereby enhancing both the accuracy and coverage of structural characterization. Building upon this foundation, generative architectures can introduce an inverse validation loop: predicting theoretical spectra for generated candidate structures and benchmark-comparing them against experimental data. This establishes a closed “hypothesis-validation” loop that effectively mitigates the risk of structural misinterpretation.

4.2 Prediction of advanced conformations and physical properties

The processing functionalities of polysaccharides are intimately coupled with their multi-scale architecture. At the conformational prediction level, GLYCAM-Web enables the online generation of oligosaccharide 3D structures and supports molecular docking, whereas the polysaccharide builder provides atomistic models specifically tailored for systems like cellulose and chitin that are fully compatible with mainstream molecular dynamics (MD) engines; furthermore, frameworks neural network potentials (NNPs) with conformational sampling have achieved qualitative agreement with experimental spectra during the exploration of monosaccharide and disaccharide conformations.

In terms of macroscopic physical property prediction, researchers have leveraged architectures such as XGBoost and neural networks to successfully predict the viscosity of konjac glucomannan, the rheological characteristics of gellan gum composite gels, and the 3D printing applicability of polysaccharide hydrogels. Currently, constructing an end-to-end predictive workflow spanning from primary sequence structures to macroscopic physical properties still necessitates overcoming major bottlenecks, including data sparsity, time-scale mismatches, and structural diversity. To address these challenges, methodologies such as active learning, coarse-grained (CG) simulations, and enhanced sampling are actively being integrated into current research paradigms.

4.3 Structure-activity relationship modeling and bioactivity screening

The objective of SAR research is to establish quantitative mapping relationships between the structural features of polysaccharides and their biological activities, thereby providing a theoretical foundation for bioactivity screening and structural optimization. The application of machine learning within this

domain has achieved preliminary progress. The distinct advantages of this strategy lie in its capacity to significantly narrow the scope of experimental screening, integrate multi-dimensional structural variables, enhance the efficiency of deciphering complex SAR, and deliver theoretical guidance for structural modification strategies such as targeted purification, enzymatic hydrolysis, sulfation, and carboxymethylation.

However, shared limitations persist across current methodologies. These include the restricted scale of polysaccharide SAR databases, the lack of standardized protocols for structural characterization, and the high variance in assay methods and endpoint indices across different studies. Furthermore, the difficulty in entirely standardizing the description of advanced conformations, aggregation behaviors, and receptor-binding modes poses formidable barriers to cross-study model training and validation.

Looking forward, future endeavors must center on the establishment of standardized, multi-dimensional SAR databases, alongside the seamless integration of machine learning with multi-omics analyses and molecular simulations to bolster the reliability and applicability of predictive models. In terms of methodological selection, SAR investigations characterized by limited sample sizes should prioritize conventional machine learning models to prevent overfitting, rather than blindly pursuing the complexity of deep learning architectures.

4.4 Atom-level interaction prediction: from phenotypic screening to target validation

The biological activities of polysaccharides, such as immunomodulation and anti-inflammation, are typically initiated through direct interactions with specific receptor proteins. Tools like AlphaFold 2, AlphaFold 3, and AF2BIND have enabled high-precision prediction of protein structures and their corresponding binding sites without requiring homology modeling or multiple sequence alignments (MSAs)^[6]. Building upon these advancements, structural prediction of non-covalent protein-polysaccharide complexes is emerging as a major frontier. Furthermore, extending beyond static structural predictions, Molecular Docking and MD simulations provide indispensable toolkits for further deciphering the dynamic interaction mechanisms between polysaccharides and their target receptors.

4.5 Function-driven *de novo* design and optimization

The conventional paradigm of polysaccharide research rigorously adheres to the linear “extraction-characterization-functional validation” pipeline. However, the convergence of generative AI and synthetic biology has introduced a revolutionary framework for the proactive design of functional polysaccharides. The foundational philosophy of *de novo* protein design is progressively expanding into the carbohydrate domain. Generative architectures, spearheaded by diffusion models and flow matching models, possess the unique capacity to generate target molecular blueprints that satisfy specific structural expectations under pre-defined functional constraints. In polysaccharide investigations, researchers can center their objectives on tailored functional demands—such as immunomodulation, prebiotic activity, or targeted drug delivery—and leverage generative models to deep-learn the structural syntax of naturally active polysaccharides. Consequently, this enables the directional design of entirely novel polysaccharide sequences characterized by specific molecular

weight intervals, defined glycosidic linkage combinations, and tailored side-chain modification patterns.

Complementarily, synthetic biology acts as the translational engine to materialize these digital blueprints into physical entities. Utilizing cell-free protein synthesis systems and multi-enzyme cascade reactions, *in vitro* platforms can rapidly reconstruct the polysaccharide biosynthetic pathways recommended by AI algorithms. Furthermore, by engineering the metabolic networks of model strains such as *Streptomyces* and *Lactobacillus*, the polysaccharide biosynthetic process can be precisely regulated, ensuring the highly efficient and stable synthesis of target products.

The deep synergy between AI and synthetic biology has successfully orchestrated the directed evolution and iterative optimization of functional polysaccharides. Currently, *de novo* design is advancing from single-round generation to iterative evolutionary paradigms. The AI-assisted directed glycan evolution workflow operates *via* a three-step closed loop: 1) Generative AI creates an initial library of polysaccharide variants; 2) high-throughput screening assays rapidly evaluate the bioactivity of these variants; and 3) machine learning algorithms analyze the screening outputs to dynamically update the generation strategy.

4.6 AI glycomics in precision nutrition

AI glycomics offers a novel trajectory for precision nutrition, spanning from polysaccharide structural elucidation to functional prediction and personalized intervention (Fig. 1). The current research trend is shifting from the functional evaluation of single polysaccharides or prebiotics toward multi-modal analyses that seamlessly integrate glycomics, gut microbiota, metabolomics, and clinical phenotypes. The foundational workflow entails capturing the structural profiles of dietary carbohydrates alongside

individual microbiotic and metabolic signatures, leveraging machine learning or molecular simulations to decipher the interconnected correlations across the “glycan structure–microbiota–metabolite–health outcome” axis, and subsequently formulating personalized nutritional regimes or tailored prebiotic combinations.

Future endeavors must continuously advance the construction of standardized glycomics databases, the development of interpretable multi-omics fusion models, and the validation of these frameworks through real-world nutritional intervention trials. A particularly compelling avenue of research is the Digital Twin Gut. By integrating an individual's gut microbiota composition, metabolomic profiles, and host genotype, this approach constructs a personalized, computational model of the intestinal ecosystem. Within this virtual framework, AI can simulate the specific intervention outcomes of diverse polysaccharides, enabling the *in silico* optimization of nutritional formulations prior to actual clinical or dietary administration.

5 Translation to food–medicine homology and precision nutrition

5.1 Functional foods

For food–medicine homologous materials, the first translational target is often not a novel molecular entity but a reproducible ingredient. AI can integrate botanical or microbial provenance, extraction parameters, structural descriptors, sensory and processing properties, bioactivity and safety to define a quality-by-design space. The immediate opportunity is batch qualification and experiment prioritization. Models should be trained against verified compositional and functional endpoints, and claims

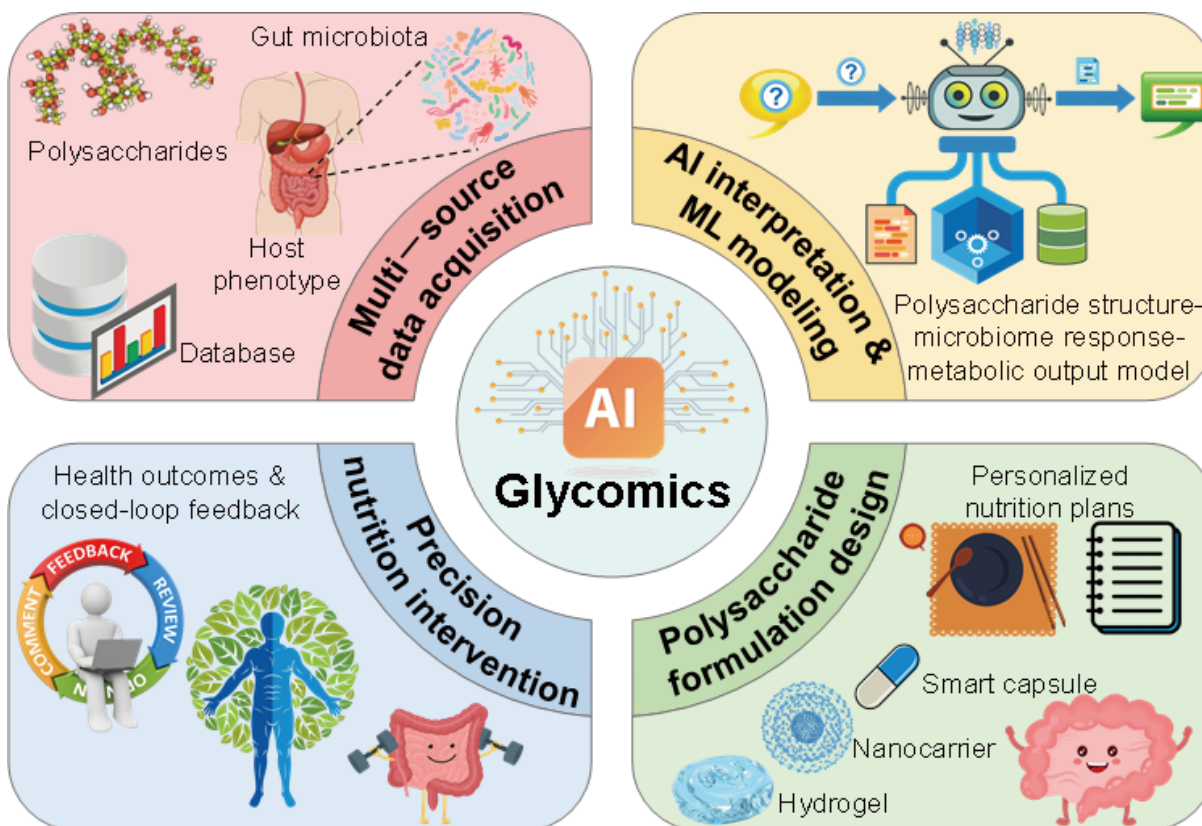


Figure 1 AI glycomics for precision nutrition.

should be limited to the tested preparation. A high-performing model does not justify transferring an activity from a purified fraction to a commercial extract with different molecular-weight distribution or contaminants.

5.2 Prebiotic and gut-targeted products

Prebiotic function emerges from a chain of events: accessibility to microbial enzymes, strain- and community-specific utilization, cross-feeding, metabolite production and host response. The relevant representation must therefore combine polysaccharide linkage and accessibility with carbohydrate-active enzyme repertoires, microbiome composition, diet and host phenotype. Rather than predicting a single universal “prebiotic score,” models should estimate conditional responses and uncertainty. *In vitro* fermentation can screen structural panels, but prospective human studies are needed to confirm responders, dose and durability.

5.3 Medical polysaccharide excipients and delivery systems

Medical excipients require consistent rheology, gelation, degradation, compatibility, sterilization stability and safety. Here, process and distributional data are as important as nominal structure. AI can support formulation optimization, degradation forecasting and risk-based specification setting, but deployment should follow regulatory quality-by-design principles. For hydrogels, nanocarriers or smart capsules, the model should connect formulation variables to critical quality attributes and release behavior, with predefined acceptance criteria and external batch validation.

5.4 Precision nutritional glycomics

Precision nutrition provides a motivating precedent: machine-learning studies have shown substantial interindividual variation in postprandial responses and the value of combining meal, microbiome and host factors. These studies do not yet establish personalized polysaccharide prescriptions, but they define the cohort size, prospective design and external validation that such claims require. The glycan structure–microbiome–metabolite–health axis should therefore be treated as a testable causal chain, not a single correlation.

6 Current challenges and future perspectives

This review systematically contextualizes the application pathways of AI in deciphering polysaccharide structure-activity relationship. However, this domain remains in its embryonic stages; the primary bottleneck at present resides not in algorithmic constraints, but rather in the acute scarcity of high-quality data resources and high-precision structural characterization capabilities. In the next 2 years, the priority is community agreement on minimum metadata, reference materials and leakage-resistant benchmark splits. Within 3–5 years, multicenter datasets should enable external validation across instruments, batches and laboratories, together with prospective active-learning studies. Over 5–10 years, closed-loop platforms may combine constrained design, synthesis or biosynthesis, structural confirmation and functional testing. Movement between phases should depend on evidence rather than calendar time. This transition will primarily manifest across three core dimensions.

First, the establishment of open-access, standardized, and traceable polysaccharide databases coupled with repository

systems for physical reference standards, while leveraging advanced privacy-preserving architectures such as federated learning to facilitate seamless cross-institutional data sharing. Second, the development of explainable models that intrinsically integrate structural chemical priors, multimodal datasets, and causal inference techniques, thereby shifting the research focus from superficial correlation analyses to robust causal mechanism elucidations, which will significantly bolster the extrapolation power and rational design capabilities of predictive models. Third, harnessing the potential of foundational models and AI for Science (AI4S) agents to propel the underlying research modality from an experience-driven paradigm into a data-driven, autonomously iterative phase.

Collectively, these advancements will transition the study of polysaccharide SAR away from its historical reliance on localized expertise and fragmented datasets toward a newly crystallized paradigm centered on standardized data, interpretable architectures, and causal inference. The intrinsic value of AI lies not merely in accelerating research throughput, but more fundamentally in discovering novel laws and biochemical mechanisms that transcend conventional heuristic understandings. Nevertheless, AI remains a tool subordinated to fundamental scientific inquiry; its predictive outputs will invariably necessitate rigorous experimental validation and the professional discernment of researchers. As data repositories, algorithmic architectures, and disciplinary consensus continue to mature, the deep integration of AI and polysaccharide science is poised to reshape the research logic of the field, opening unprecedented trajectories for the rational design and precision application of functional polysaccharides.

Data availability

Additional data related to this paper may be requested from the corresponding author upon request.

Conflicts of interest

The authors declare that there are no conflicts of interest in this work. Yan-Long Cui, Ning-Xian Yang and Jia-Lun Duan are the Editorial Board Members of Food & Medicine Homology, Wen-Yi Kang is an Editors-in-Chief of Food & Medicine Homology, but were not involved in the processing, peer review or decisionmaking of this article.

Acknowledgements

None.

Funding

This work was supported by Central Plains Innovation Leading Talent (264200510055).

Author contribution statement

Shi-Kang Chen: writing manuscript, Project administration. Tao Li: writing manuscript. Qiang Li: writing manuscript. Lei Duan: writing manuscript. Chen Chen: writing manuscript. Zhi Huang: writing manuscript. Yan-Long Cui: writing manuscript. Zhi-Fei Fu: writing manuscript. Xi Chen: writing manuscript. Xiao-Bo Deng: writing manuscript. Ning-Xian Yang: writing manuscript. Fu-Li Yu: writing manuscript. Pei-Lin Chen: writing

manuscript. Yan Meng: writing manuscript. Xing-Fang Zhang: writing manuscript. Qiang Ren: writing manuscript. Ren-Peng Du: writing manuscript. Guang-Ping Lü: writing manuscript. Ying-Yue Zhang: writing manuscript. Yu-Ting Su: writing manuscript. Hong-Tao Bi: writing manuscript. Li-Li Guan: writing manuscript. Lei Feng: writing manuscript. Tao Hu: writing manuscript. Hazem Golshany: writing manuscript. Jia-Lun Duan: writing manuscript; project administration. Wen-Yi Kang: funding acquisition.

Informed consent

Not applicable.

Ethics statement

Not applicable.

Use of AI statement

During the preparation of this work, the authors used ChatGPT (OpenAI) to improve the clarity, grammar, and readability of the manuscript. After using this tool, the authors carefully reviewed and edited the generated content and take full responsibility for

the content of the publication.

References

- [1] Urban, J., Jin, C., Thomsson, K. A., et al. Predicting glycan structure from tandem mass spectrometry via deep learning. *Nature Methods*, **2024**, 21: 1206–1215. <https://doi.org/10.1038/s41592-024-02314-6>
- [2] Yamada, I., Shiota, M., Shinmachi, D., et al. The GlyCosmos Portal: a unified and comprehensive web resource for the glycosciences. *Nature Methods*, **2020**, 17: 649–650. <https://doi.org/10.1038/s41592-020-0879-8>
- [3] Abramson, J., Adler, J., Dunger, J., et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, **2024**, 630: 493–500. <https://doi.org/10.1038/s41586-024-07487-w>
- [4] Bran, A. M., Cox, S., Schilter, O., et al. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, **2024**, 6: 525–535. <https://doi.org/10.1038/s42256-024-00832-8>
- [5] Song, J., Yuan, C., Wang, F., et al. Omics-based profiling and therapeutic potential of natural components in pan-Shennongjia medicinal herbs. *Chinese Medicine*, **2025**, 20: 172. <https://doi.org/10.1186/s13020-025-01208-9>
- [6] Gazizov, A., Lian, A., Goverde, C., et al. AF2BIND: predicting small-molecule binding sites using the pair representation of AlphaFold2. *Nature Methods*, **2026**, 23: 626–635. <https://doi.org/10.1038/s41592-026-03011-2>