Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

MULTIMODAL GENERATIVE ARTIFICIAL INTELLIGENCE BEYOND MOLECULE CREATION: A SCOPING REVIEW OF THERAPEUTIC DESIGN ACROSS DATA MODALITIES

Ali Hassan[1]*, Noor Siddiqui[1], Bilal Khan[2], Sana Malik[3], Hassan Ali[4]

1. ***Department of Multimodal Generative AI for Therapeutic Design, Faculty of Pharmacy, Aga Khan University, Karachi, Pakistan.***
2. ***Department of Beyond-Molecule-Creation AI, Faculty of Pharmacy, Qatar University, Doha, Qatar.***
3. ***Department of Multi-Modality Drug Design, Faculty of Pharmacy, University of Punjab, Lahore, Pakistan.***
4. ***Department of Therapeutic Design across Modalities, Faculty of Pharmacy, University of Agriculture Faisalabad, Faisalabad, Pakistan.***

ARTICLE INFO

Received:
06 April 2026
Received in revised form:
03 August 2026
Accepted:
07 August 2026
Available online:
28 August 2026

Keywords: Multimodal artificial intelligence, Generative models, Therapeutic design, Drug discovery, Protein design, Multi-omics

ABSTRACT

Generative artificial intelligence in pharmaceutical science is commonly evaluated through its capacity to create novel molecular structures, yet therapeutic design depends on a wider evidence environment that includes protein sequence and conformation, cellular and multi-omics states, biomedical images, scientific text, target context, and experimental observations. The resulting literature is heterogeneous in terminology, model architecture, generated output, conditioning strategy, and validation depth, making conventional effect synthesis inappropriate and encouraging potentially misleading comparisons between representation learning, prediction, generation, and therapeutic development. This scoping review maps how multimodal generative approaches are defined, implemented, and evaluated across pharmaceutical data modalities. The review uses transparent eligibility, study-selection, data-charting, and terminology-mapping principles to distinguish modality integration from cross-modal generation and to separate demonstrated model capability from benchmark performance, experimental confirmation, translational value, and routine pharmaceutical readiness. The synthesis identifies a field with comparatively mature methods for molecular generation, emerging experimental evidence for protein design, developing applications in perturbational omics, and primarily enabling roles for biomedical images and scientific text. It further distinguishes conditioning, alignment, fusion, and cross-modal decoding as related but non-equivalent operations. An original review-derived taxonomy and evidence-maturity interpretation are proposed to organize these approaches; neither is presented as empirically validated. Major limitations include narrow reference datasets, inconsistent definitions of validation, limited external testing, weak causal grounding, sparse negative evidence, and insufficient connection between computational outputs and prospective therapeutic decisions. Multimodal generative systems may broaden therapeutic design beyond molecule creation, but their scientific value remains conditional on claim-matched evaluation, experimentally grounded validation, transparent provenance, and explicit boundaries between computational plausibility and pharmaceutical usefulness.

To Cite This Article: Hassan A, Siddiqui N, Khan B, Malik S, Ali H. Multimodal Generative Artificial Intelligence beyond Molecule Creation: A Scoping Review of Therapeutic Design across Data Modalities. Pharmacophore. 2026;17(4):104-14. https://doi.org/10.51847/LZB9kLxouu

Introduction

Artificial intelligence now contributes to numerous stages of pharmaceutical research, including target identification, molecular representation, virtual screening, structure-based analysis, candidate optimization, and development planning. However, the expansion of computational capability has not produced a uniform level of scientific or translational evidence. Drug-design systems operate within workflows in which chemical novelty must be reconciled with biological relevance, experimental tractability, medicinal-chemistry judgment, and downstream development constraints. Computational

Corresponding Author: Ali Hassan; Department of Multimodal Generative AI for Therapeutic Design, Faculty of Pharmacy, Aga Khan University, Karachi, Pakistan. E-mail: ali.hassan@aku.edu.pk

acceleration therefore cannot be interpreted independently of the evidence used to define the target, condition the design process, and validate the resulting intervention [1-3].

The dominant account of generative pharmaceutical artificial intelligence has nevertheless remained centered on the creation of small-molecule structures. This focus is understandable because chemical representations are comparatively standardized, large molecular datasets are available, and generated outputs can be assessed using established syntax, similarity, property, and docking-based procedures. Yet a therapeutically meaningful design problem begins before a candidate structure is generated. It includes the selection of disease-relevant biological states, the interpretation of target evidence, the identification of context-dependent mechanisms, and the specification of the intervention that should alter those states. AI-supported target discovery illustrates why candidate generation must be connected to disease biology, multi-omics evidence, and the uncertainty surrounding target–phenotype relationships [4].

Multimodal foundation models offer a possible route toward this broader conception by learning across images, language, sequences, structured measurements, and other data types. Their importance lies not merely in accommodating several inputs but in potentially connecting complementary forms of evidence that describe a therapeutic problem at different scales. Nevertheless, cross-modal representation, retrieval, or task transfer does not itself establish that a model can generate a therapeutically useful intervention. Generalist medical models demonstrate the feasibility of broad multimodal learning while also showing that capability is shaped by data availability, task formulation, validation setting, and intended use [5]. These distinctions are especially important in pharmaceutical science, where a plausible generated output may remain unsynthesizable, biologically inactive, unsafe, non-developable, or irrelevant to the decision for which it was produced.

The unresolved problem is therefore both technical and evidentiary: the literature lacks a sufficiently clear synthesis of what "multimodal generation" means across molecular, protein, omics, image, and text domains, how modalities condition or interact within generative systems, and what validation is required before outputs can influence therapeutic research. This scoping review maps the breadth, terminology, methods, application contexts, evidence maturity, and unresolved gaps in multimodal generative therapeutic design. Its central contribution is an evaluative taxonomy that separates modality integration, representation alignment, conditional generation, fusion, and cross-modal decoding while distinguishing computational validity from experimental confirmation and pharmaceutical relevance. The taxonomy is proposed as an organizing synthesis rather than a validated ontology or deployment framework.

Scoping-Review Objectives and Eligibility Framework

The review was structured using PRISMA-ScR-compatible reporting logic, with explicit articulation of the review questions, eligibility concepts, information sources, study-selection principles, charting fields, and synthesis approach [6]. The primary objective was to determine how generative artificial intelligence has been used, or credibly proposed for use, in therapeutic design when more than one pharmaceutical data modality contributes to the representation, conditioning, transformation, or evaluation of an output. Secondary objectives were to map the terminology used for multimodality, identify the therapeutic objects being generated or transformed, characterize the validation attached to those outputs, and locate areas in which representation or prediction has been described as generation without sufficient methodological justification. The review was not designed to estimate a pooled treatment, prediction, or model-performance effect.

A scoping design was selected because the available literature spans heterogeneous scientific objects, model families, data structures, application stages, and validation settings. The relevant questions concern the boundaries and distribution of evidence rather than the comparative effectiveness of a single intervention or method. Guidance distinguishing scoping from conventional systematic reviews supports this approach when the purpose is to clarify concepts, map evidence types, examine how research is conducted, and identify knowledge gaps [7]. Accordingly, eligibility was based on conceptual relevance to multimodal therapeutic generation rather than on the availability of a common outcome measure. Formal meta-analysis, pooled benchmark rankings, and numerical aggregation across incompatible tasks were outside the review scope.

Eligibility was operationalized around three linked concepts: multimodality, generation, and therapeutic design. Multimodality required either the use of distinct data forms—such as molecular graphs, protein sequences, three-dimensional structures, omics measurements, images, or text—or clearly differentiated representational channels that contributed non-redundant information. Generation required the creation, reconstruction, editing, transformation, or counterfactual proposal of an output rather than classification or prediction alone. Therapeutic design required a defensible connection to targets, interventions, candidate properties, biological states, experimental hypotheses, or pharmaceutical decisions. Because scoping-review conduct allows iterative refinement when emerging terminology reveals previously unrecognized evidence categories, definitions could be clarified during charting provided that every refinement remained documented and was applied consistently [8].

Study selection was designed to retain empirical investigations, methodological studies, benchmark analyses, and sufficiently detailed critical or conceptual articles when they directly supported the review questions. Prediction-only studies were charted as contextual evidence only when they clarified a necessary enabling modality, validation boundary, or terminology distinction. Studies were not treated as equivalent merely because they used the label "multimodal," and quality appraisal was used to calibrate interpretation rather than to impose a single exclusion threshold across incompatible evidence types. Practical scoping-review guidance supports documented screening roles, iterative charting, and transparent resolution of classification uncertainty [9]. The final synthesis therefore maps the nature and boundaries of the evidence rather than claiming comprehensive effect estimation or uniform methodological quality, consistent with the formal definition of scoping reviews

as evidence-mapping exercises [10]. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop scoping-review objectives and eligibility framework within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for Scoping-review objectives and eligibility framework

Evidence domain	Eligible evidence or method	Reported capability or claim	Strength of support	Reproducibility or bias concern	Unresolved gap
Scoping-review methodology	Reporting guidance, methodological guidance, transparent review procedures, and explicit charting frameworks	Heterogeneous evidence can be mapped without implying pooled-effect certainty	Strong methodological support	Incomplete reporting of eligibility refinement or charting decisions may reduce reproducibility	Limited consensus on appraisal methods for rapidly evolving generative-model evidence
Multimodal representation	Models integrating distinct molecular, biological, imaging, or textual data types	Complementary modalities can be embedded, related, or jointly analyzed	Moderate to strong for representation tasks	Paired-data scarcity, batch effects, modality imbalance, and hidden dependence can distort shared representations	Representation quality is rarely linked prospectively to therapeutic decisions
Generative operation	Creation, reconstruction, editing, transformation, or counterfactual proposal of an output	Models can generate candidates or transformed states under stated conditions	Uneven across output types	Terminology may conflate generation with prediction, interpolation, retrieval, or imputation	No generally accepted cross-domain definition of therapeutic generation
Therapeutic relevance	Target-conditioned design, intervention proposals, perturbational state generation, protein design, or decision-support hypotheses	Generated outputs may contribute to therapeutic research	Strongest for bounded molecular and protein tasks; emerging elsewhere	Proxy objectives may not represent potency, mechanism, safety, developability, or clinical utility	Few studies connect generated outputs to prospective pharmaceutical decisions
Experimental confirmation	Synthesis, structural assessment, binding assays, functional experiments, cellular testing, or other physical validation	Selected generated outputs can exhibit measurable physical or biological properties	Strong when predefined experiments are reported; otherwise limited	Positive-result selection, narrow assay contexts, and absence of independent replication	Sparse experimental validation across image-, text-, and omics-conditioned generation
External validity	Evaluation on new targets, datasets, cohorts, laboratories, temporal settings, or modality combinations	Model capability may generalize beyond the development data	Generally limited and application-specific	Similarity leakage and non-independent splitting can inflate apparent generalization	Few locked-model, cross-site, or prospective evaluations
Evidence maturity	Explicit distinction among conceptual, computational, external, experimental, and prospective evidence	Studies can be compared according to validation depth rather than a single performance metric	Proposed review-level synthesis	Maturity categories may simplify heterogeneous development pathways	The proposed maturity interpretation requires future testing and refinement
Interpretation boundary	Claim-matched distinctions between association, prediction, generation, mechanism, experimental confirmation, and therapeutic utility	Evidence can support bounded conclusions without implying deployment readiness	Strong conceptual necessity	Promotional terminology and inconsistent definitions of validation encourage overstatement	Shared reporting language for multimodal therapeutic generation remains underdeveloped

Search, Study Selection, Data Charting, and Modality Taxonomy

The search strategy was designed to retrieve literature at the intersections of generative modeling, pharmaceutical or therapeutic design, and heterogeneous data modalities. Search concepts covered molecular and structure-based generation, protein design, perturbational and multi-omics modeling, image–language systems, scientific-text conditioning, multimodal alignment, modality fusion, cross-modal generation, evaluation, reproducibility, and translational boundaries. Information

sources, complete search strings, field restrictions, limits, deduplication procedures, and supplementary searching should be reported in sufficient detail for reconstruction, in accordance with PRISMA-S principles [11]. Search sensitivity was prioritized over reliance on the term "multimodal," because relevant studies frequently describe their methods as target-conditioned, structure-guided, protein-aware, text-guided, multi-view, paired-modal, or joint-representation models.

Study selection proceeded through title-and-abstract assessment followed by full-text eligibility review using the operational definitions of multimodality, generation, and therapeutic design. During charting, each study was assigned separate fields for input modality, generated or transformed output, conditioning variable, alignment mechanism, fusion level, model family, therapeutic context, evaluation endpoint, external-validation status, experimental confirmation, and stated limitation. This structure was necessary because multimodal data integration can involve shared latent factors, modality-specific factors, or adaptive cell-level relationships without generating a therapeutic output. Multi-omics factor models and multimodal single-cell integration methods demonstrate several such representational strategies and therefore provide a basis for distinguishing integration from generation [12-14].

The proposed modality taxonomy contains five primary input domains: molecular representations; protein sequences and structures; omics and cellular states; biomedical images; and scientific or biomedical text. Studies can receive more than one input label and more than one output label, preventing systems that combine modalities from being forced into a single category. A second coding axis distinguishes representation, retrieval, prediction, imputation, editing, unconditional generation, conditional generation, and cross-modal decoding. A third axis records whether modalities are connected through early fusion, intermediate representation alignment, cross-attention or related conditional interaction, late decision fusion, or tool-mediated orchestration. These categories are review-derived analytical devices rather than empirically validated architectural classes. Their purpose is to make heterogeneous evidence comparable while preserving the central boundary that access to multiple modalities does not itself establish multimodal generation, and multimodal generation does not itself establish therapeutic usefulness.

Generative Design Using Molecular, Protein, Omics, Image, and Text Inputs

Molecular generation remains the most technically developed component of the evidence base. Continuous latent chemical representations allow molecular structures to be encoded, sampled, reconstructed, and optimized with respect to learned or externally specified properties [15]. This capability demonstrates that chemical design can be formulated as navigation through a learned representation rather than only enumeration from fixed libraries. Its therapeutic interpretation is nevertheless conditional. Molecular validity, reconstruction accuracy, novelty, or predicted property improvement does not establish target engagement, synthetic accessibility, selectivity, metabolic stability, toxicity, formulation feasibility, or clinical value. Molecular generation should therefore be treated as candidate proposal under bounded computational assumptions rather than as completed therapeutic design.

Protein generation broadens the output space from chemical graphs to macromolecular sequence, conformation, and function. Diffusion-based protein design has demonstrated the capacity to generate three-dimensional structures and selected functional configurations that can be evaluated experimentally [16]. Compared with purely in silico molecular benchmarks, physical expression, structural assessment, or functional testing provides a stronger evidentiary bridge between generated representation and biological object. Even so, success remains design- and assay-specific. A generated protein that adopts an intended structure may still face challenges involving stability, immunogenicity, manufacturability, pharmacokinetics, tissue distribution, or context-dependent activity. The evidence supports experimentally testable protein generation, but not universal therapeutic-protein readiness.

Omics-based generation differs because the generated object may be a biological state rather than a discrete therapeutic candidate. Latent generative models have been used to predict perturbation-associated single-cell states by learning transformations across observed cellular conditions [17]. Such approaches may help organize hypotheses about how cells respond to interventions, identify response heterogeneity, or prioritize experiments. Their interpretation requires particular caution because a generated transcriptomic state may reflect interpolation within correlated data rather than a causal simulation of treatment response. Generalization to new cell types, disease states, doses, combinations, temporal trajectories, or measurement platforms must therefore be assessed separately from reconstruction or held-out prediction performance.

Biomedical images and scientific text currently function more often as contextual, conditioning, or alignment modalities than as independently validated therapeutic generators. Visual–language foundation models in computational pathology show that image and text representations can be aligned and transferred across diagnostic or descriptive tasks [18]. This may support therapeutic design indirectly by linking morphology with disease context, biomarker hypotheses, or response phenotypes. Text can also provide descriptions of targets, desired properties, mechanisms, or constraints. However, neither a shared image–text embedding nor a linguistically plausible design explanation establishes that the corresponding therapeutic hypothesis is experimentally valid. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop generative design using molecular, protein, omics, image, and text inputs within the article's central argument.

Table 2. Components, relationships, uncertainties, and validation needs within Generative design using molecular, protein, omics, image, and text inputs

Approach or application	Data and task context	Evaluation practice	Evidence maturity	Translational limitation	Review interpretation

Latent molecular generation	Molecular graphs or strings with property labels or optimization objectives	Chemical validity, novelty, reconstruction, distribution similarity, and predicted properties	Established computational benchmark evidence	Proxy objectives do not establish synthesis, biological activity, safety, or developability	The most mature generative domain, but still predominantly candidate-proposal technology
Structure-conditioned ligand generation	Molecular outputs conditioned by target shape, pocket, or structural features	Structural compatibility, docking-related assessment, chemical filters, and selected assays	Computationally mature with limited experimental extension	Scoring functions and structural assumptions may be exploited or incomplete	More therapeutically contextualized than unconditional generation, but not sufficient evidence of efficacy
Protein structure and function generation	Protein sequence and three-dimensional design tasks	Structural prediction, expression, physical characterization, binding, or functional testing	Computational and experimental evidence in selected settings	Generalization across therapeutic functions and development constraints remains uncertain	Strong evidence that generative design can extend beyond small molecules
Omics-state generation	Single-cell or multi-omics states before and after perturbation	Reconstruction, held-out condition prediction, latent-state comparison, and biological annotation	Emerging retrospective evidence	Confounding, sparse perturbation coverage, batch effects, and weak causal identification	Useful for hypothesis generation and experiment prioritization, not a validated causal simulator
Image-conditioned therapeutic reasoning	Histopathology, microscopy, or phenotype images linked to language or molecular context	Retrieval, classification, representation transfer, and association with downstream labels	Stronger for representation than direct generation	Images often provide indirect evidence and may contain site-specific artefacts	Primarily an enabling modality for context and stratification
Text-guided generation or editing	Scientific descriptions, property instructions, target information, or conversational prompts	Instruction following, molecule retrieval, editing success, and computational property checks	Emerging computational evidence	Semantic compliance may conceal chemical, biological, or evidentiary errors	Text is a useful control interface but should not be treated as an authoritative evidence source
Multimodal therapeutic hypothesis generation	Combined molecular, biological, imaging, omics, and textual evidence	Coherence, traceability, expert review, and experimental follow-up	Predominantly conceptual or early-stage	No common standard links cross-modal coherence to improved pharmaceutical decisions	Promising review-level category requiring prospective validation

Conditioning, Alignment, Fusion, and Cross-Modal Generation

Conditioning refers to the use of an auxiliary signal to steer generation toward a desired target, property, structure, phenotype, or design constraint. In structure-based ligand design, three-dimensional target information can guide molecular generation toward candidates compatible with a specified binding environment [19]. Conditioning therefore differs from unconditional sampling because it imposes a task-relevant context on the output distribution. Its scientific value depends on whether the conditioning variable is reliable, sufficiently informative, and linked to the claimed therapeutic objective. A model may comply with a geometric or textual condition while failing to satisfy unmodeled biological, pharmacokinetic, or safety requirements.

Alignment establishes correspondence between modalities without necessarily combining all information into one representation or producing a new therapeutic object. Molecule–text models can place chemical structures and textual descriptions in related spaces, supporting retrieval and text-guided molecular editing [20]. Alignment may facilitate communication between scientific language and chemical representation, but it does not prove that the textual concept has been translated into the intended mechanism. The underlying descriptions may be incomplete, ambiguous, generated from weak annotations, or correlated with superficial structural features. Alignment should consequently be evaluated through bidirectional retrieval, editing fidelity, perturbation tests, and evidence traceability rather than through semantic similarity alone.

Fusion combines information from more than one modality and may occur before encoding, within intermediate representations, or after modality-specific outputs have been produced. Early fusion can expose a model directly to concatenated or jointly tokenized inputs, whereas intermediate fusion allows cross-attention or shared latent interaction after modality-specific encoding. Late fusion combines predictions, rankings, or decisions rather than raw representations. Conversational molecular systems illustrate an additional tool-mediated form in which language instructions coordinate

molecular generation, retrieval, or editing operations [21]. Such systems may improve accessibility, but apparent conversational fluency must remain separate from scientific grounding, uncertainty communication, and reproducible control over the generated output.

Cross-modal generation is the most demanding category because one or more input modalities are used to create an output in another representational domain. A target-aware chemical language model, for example, can condition molecular generation on protein information and submit selected candidates to experimental assessment [22]. This provides stronger evidence than text-only or benchmark-only generation, while remaining bounded by the targets, datasets, assays, and candidate-selection procedures studied. The review therefore proposes that cross-modal generation be reported with four separate descriptors: the source modality, the generated modality, the conditioning or fusion mechanism, and the validation level. This proposed reporting structure is intended to prevent a shared embedding, retrieval result, or prediction from being mislabeled as therapeutic generation.

Evaluation of Controllability, Validity, Novelty, and Therapeutic Relevance

Evaluation begins with model-level questions but should not end there. Benchmarking frameworks for molecular generation assess dimensions such as chemical validity, uniqueness, novelty, distribution learning, and optimization against specified objectives. They improve comparability across methods but can also reveal that models exploit benchmark definitions, objective functions, or reference distributions in ways that produce favorable metrics without corresponding therapeutic quality [23-25]. Controllability should therefore mean more than movement in a requested numerical direction. It should include reproducible compliance with explicit constraints, preservation of non-target properties, robustness to prompt or condition variation, and evidence that the control signal relates to the intended scientific objective.

Validity is multidimensional. Representational validity concerns whether an output can be parsed or physically instantiated; chemical validity concerns valence and structural rules; biological validity concerns whether the intervention affects the relevant system; and therapeutic validity concerns whether the effect is useful under realistic disease and development conditions. Evaluation guidance for chemical machine learning emphasizes that test design, splitting strategy, baseline selection, applicability analysis, and reported uncertainty should match the scientific claim and intended use [26]. A model evaluated only on internal benchmark samples should not be described using language that implies external generalization or prospective pharmaceutical usefulness.

Novelty is also an incomplete criterion. Structural dissimilarity from a training set may indicate that a generated candidate is not an exact duplicate, but it does not establish that the candidate occupies a useful or previously inaccessible region of therapeutic space. Excessive novelty may reduce synthesizability, assay compatibility, or confidence in learned property estimates, while low apparent novelty may still yield meaningful optimization within a validated chemical series. The appropriate question is therefore whether novelty contributes to a scientifically justified design objective without moving the output beyond the model's reliable domain. Similar reasoning applies to proteins, cellular states, images, and text-derived hypotheses, for which novelty must be interpreted relative to modality-specific constraints.

Therapeutic relevance requires a hierarchy of evidence that extends beyond computational scoring. At minimum, evaluation should identify the intended decision, the comparator, the conditioning evidence, the model's applicability domain, and the consequences of failure. Stronger evidence may include external datasets, temporal or scaffold-aware evaluation, independent experimental confirmation, comparison with established design strategies, and prospective influence on candidate selection. Model confidence should not be treated as calibrated uncertainty, and expert approval of a plausible output should not be treated as validation. **Table 3** organizes the evidence, constructs, and interpretation boundaries needed to develop evaluation of controllability, validity, novelty, and therapeutic relevance within the article's central argument.

Table 3. Components, relationships, uncertainties, and validation needs within Evaluation of controllability, validity, novelty, and therapeutic relevance

Claim category	Supporting evidence type	Consistency across sources	Quality concern	What can be concluded	What remains uncertain
Generated outputs are computationally valid	Syntax, valence, reconstruction, structural checks, or model-based plausibility	Relatively consistent within molecular benchmarks	Validity definitions differ across modalities and may be permissive	Outputs can satisfy specified representational rules	Whether they can be synthesized, expressed, or function biologically
Generation is controllable	Property conditioning, target conditioning, prompt compliance, or constrained optimization	Supported in bounded tasks	Models may exploit proxy objectives or ignore unmeasured constraints	Outputs can be shifted toward selected computational conditions	Whether control persists under new targets, datasets, or experimental conditions
Generated sets are novel	Distance from reference sets, scaffold analysis, or sequence dissimilarity	Commonly reported for molecular generation	Novelty depends on representation and reference-set construction	Outputs may differ from indexed training examples	Whether novelty is useful, safe, synthesizable, or therapeutically meaningful

Benchmark performance is comparable	Shared datasets, metrics, baselines, and task definitions	Improved by benchmark platforms	Dataset overlap, tuning practices, and metric selection may distort comparison	Methods can be compared under stated benchmark conditions	Whether ranking transfers to real discovery programs
Generated candidates are biologically relevant	Binding, activity, perturbation, functional, or phenotype evidence	Uneven across domains	Positive-result selection and narrow assay contexts	Selected outputs may exhibit reported biological effects	Generality, mechanism, selectivity, durability, and in vivo relevance
Multimodal integration improves design	Ablations, retrieval, prediction, editing, or generation with multiple inputs	Emerging and task-dependent	Improvements may arise from additional information rather than meaningful cross-modal reasoning	Multiple modalities can improve selected computational tasks	Whether multimodality improves prospective therapeutic decisions
Models generalize	External, temporal, scaffold-aware, target-held-out, or cross-site evaluation	Limited and inconsistent	Leakage and hidden similarity can inflate apparent transfer	Some systems retain capability outside internal random splits	Reliability under genuine pharmaceutical distribution shift
Systems are therapeutically ready	Prospective decisions, experimental programs, and development outcomes	Sparse	Readiness terminology is inconsistently defined	Selected systems may support bounded research tasks	Routine clinical, regulatory, or operational readiness

Evidence Gaps beyond Small-Molecule Creation

The clearest evidence gap is the imbalance between mature molecular-generation benchmarks and the comparatively limited validation of protein-, omics-, image-, and text-centered generation. Even within chemical modeling, apparent generalization may reflect the similarity structure of established datasets rather than transferable learning. Ligand-based benchmarks can reward memorization when training and test examples remain chemically related [27]. This concern extends to multimodal systems because paired modalities may share hidden identifiers, source-specific artefacts, duplicated entities, or near-neighbor biological contexts. Random splitting is therefore insufficient whenever the intended claim concerns new targets, diseases, scaffolds, cohorts, laboratories, or modality combinations.

A second gap concerns the distance between reported computational capability and actual pharmaceutical decisions. Critical analyses of AI in drug discovery emphasize that retrospective model performance does not establish impact unless outputs are compared with credible alternatives, assessed prospectively, and connected to a decision that matters [28]. Beyond small molecules, this requirement becomes more complex because the output may be a designed protein, predicted cellular state, pathology-grounded hypothesis, or language-mediated intervention proposal. Each object requires a different validation chain. A common readiness label would obscure these differences and encourage false equivalence across modalities.

Data suitability is a third limitation. Chemical and biological datasets vary in assay protocol, measurement quality, missingness, selection bias, temporal relevance, and compatibility with the intended use [29]. Multimodal models may magnify these problems because combining data does not remove uncertainty; it can propagate errors across representations or allow one dominant modality to overwhelm weaker but clinically important evidence. Missing modalities, contradictory signals, and unbalanced pairing are rarely treated as central evaluation conditions. Moreover, published datasets often underrepresent inactive candidates, failed experiments, manufacturing constraints, toxicity findings, and null biological responses, limiting the ability of generative systems to learn realistic failure boundaries.

Leakage and reproducibility failures create a fourth cross-cutting gap. Information can pass improperly between training and evaluation through preprocessing, feature selection, duplicated entities, shared experimental batches, target similarity, or post hoc benchmark construction [30]. These risks are especially difficult to detect when multiple encoders, pretrained models, retrieval systems, or external tools contribute to a generated result. **Figure 1** presents the multimodal therapeutic generation prism, showing how the article's key components and boundaries are connected within evidence gaps beyond small-molecule creation.

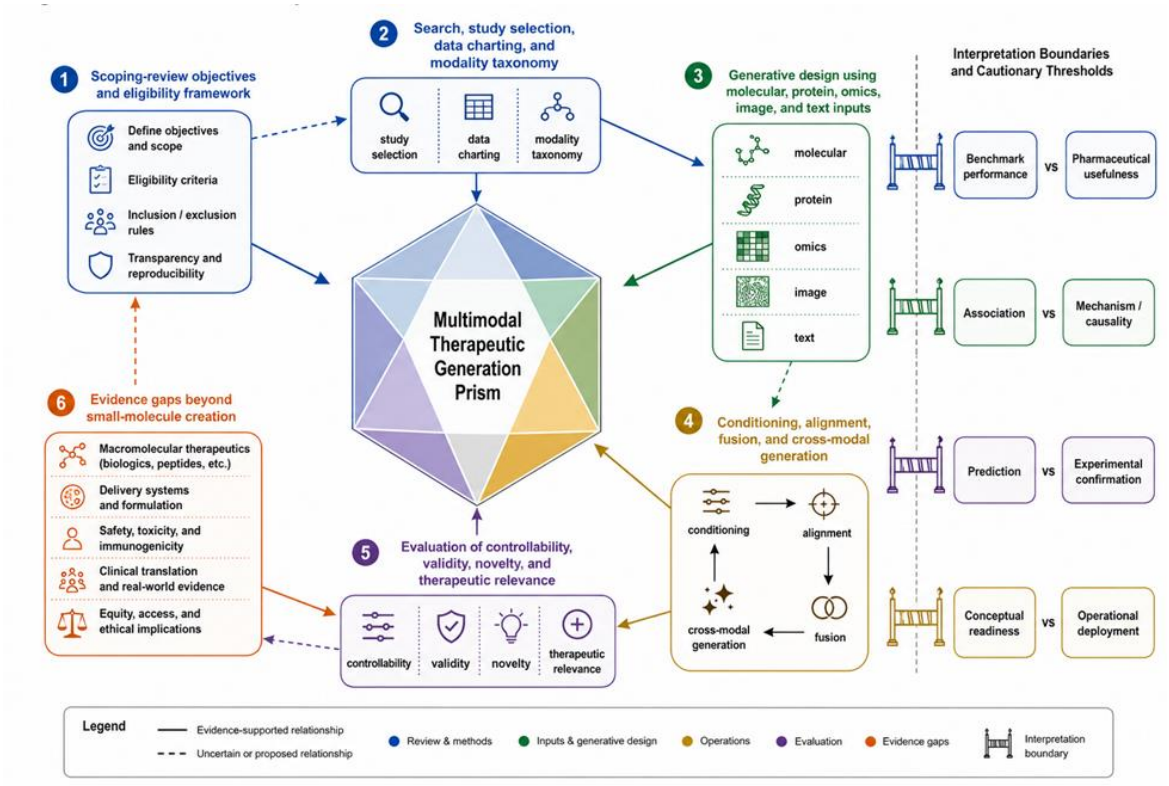

Figure 1. Multimodal Therapeutic Generation Prism.

The figure is an original conceptual synthesis that organizes the article's central contribution across scoping-review objectives and eligibility framework, search, study selection, data charting, and modality taxonomy, study selection, data charting, modality taxonomy, generative design using molecular, protein, omics, image, and text inputs. Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, regulatory determination, clinical recommendation, or deployment-ready system.

Priorities for Multimodal Therapeutic Design

The first priority is validation under biologically meaningful distribution shift. Genomic and omics machine learning is vulnerable to confounding, population structure, technical batch effects, dependence among samples, and leakage across related observations [31]. Multimodal generative studies should therefore move beyond random held-out evaluation toward target-held-out, scaffold-aware, cell-type-held-out, temporal, cross-site, and prospectively collected test settings. The selected design should correspond to the intended claim. A system proposed for generating interventions for previously unseen targets requires a stronger evaluation than a system intended to refine candidates within a known series.

The second priority is minimum reporting that enables reconstruction of the evidence chain. Reporting should specify data provenance, preprocessing, modality pairing, missing-data handling, model architecture, pretrained components, split construction, external tools, candidate-selection procedures, validation endpoints, and intended use. Clinical-AI reporting guidance demonstrates the value of explicit documentation of modeling choices, validation, and application context [32]. For generative therapeutic systems, this should be extended to include generated-output lineage, prompt or condition versioning, filtering rules, failed outputs, human interventions, and the evidence used to advance a candidate from computational generation to experimental testing.

The third priority is to separate technical performance from replicability, ethics, scientific effectiveness, and decision benefit. Critical questions developed for health-related artificial intelligence emphasize that high model performance does not resolve whether a system is transparent, reproducible, fair, useful, or effective in practice [33]. Multimodal therapeutic design additionally requires governance of proprietary datasets, sensitive patient-derived modalities, foundation-model provenance, uncertain biological claims, and human authority over experimental escalation. These requirements do not imply that every research prototype needs a clinical governance framework; rather, governance intensity should increase with the consequence of the decision and the proximity of the output to physical or human testing.

The fourth priority is infrastructure that links multimodal data science with experimental discovery. Early-stage drug-discovery roadmaps emphasize governed datasets, shared metadata, interoperable representations, reusable workflows, and integration across scientific teams [34]. Multimodal generation will require these foundations because no model can reliably reconcile molecular, structural, omics, imaging, and textual evidence when identifiers, assay context, version history, and uncertainty are not preserved. **Table 4** organizes the evidence, constructs, and interpretation boundaries needed to develop priorities for multimodal therapeutic design within the article's central argument.

Table 4. Evaluation, implementation, and research priorities arising from Multimodal Generative Artificial Intelligence beyond Molecule Creation A Scoping Review of Therapeutic

Research priority	Evidence gap	Recommended study or reporting design	Validation requirement	Near-term value	Longer-term significance
Claim-matched external validation	Internal benchmarks do not establish transfer to new targets, scaffolds, cohorts, or laboratories	Predefine intended use and use target-held-out, scaffold-aware, temporal, or cross-site evaluation	Locked-model assessment on genuinely external data	Reduces exaggerated generalization claims	Establishes transportable therapeutic-design evidence
Prospective multimodal datasets	Existing modalities are often weakly paired, retrospectively assembled, or collected under incompatible protocols	Collect molecular, biological, imaging, omics, textual, and experimental metadata under shared identifiers	Cross-site replication and missing-modality stress testing	Improves alignment and fusion quality	Enables robust cross-modal therapeutic modeling
Perturbational and causal grounding	Correlated multimodal representations may not predict intervention effects	Incorporate controlled perturbations, longitudinal measurements, and mechanism-recovery analyses	Independent experiments testing predicted direction and context of response	Distinguishes association from intervention-relevant evidence	Supports causal therapeutic-design models
Experimental validation ladders	Computational validity is frequently treated as sufficient evidence	Predefine staged progression from computational checks to synthesis, binding, function, cellular response, and development assessment	Blinded or independent testing with explicit advancement criteria	Makes evidence maturity visible	Links generative systems to reproducible discovery programs
Uncertainty and abstention	Model scores are often misinterpreted as confidence	Report calibration, ensemble disagreement, applicability domain, modality conflict, and abstention behavior	Evaluation under distribution shift and contradictory inputs	Improves risk-sensitive candidate review	Supports accountable decision integration
Negative and failed-design data	Published datasets overrepresent successful or measurable outcomes	Capture inactive, unsynthesizable, unstable, toxic, and experimentally failed candidates	Versioned release of negative outcomes and failure reasons	Improves failure detection	Produces more realistic generative objectives
Provenance and output lineage	Generated proposals may depend on opaque pretrained models, prompts, retrieval sources, and filters	Maintain versioned records of data, models, prompts, tools, selection rules, and human edits	Independent reproduction from frozen artifacts	Strengthens auditability	Enables controlled updating and organizational learning
Human–AI decision studies	Technical output quality is rarely connected to decision quality	Compare established workflows with multimodal-assisted workflows using predefined decision tasks	Prospective assessment of error detection, scientific justification, and decision consequences	Reveals where models genuinely assist researchers	Determines practical pharmaceutical value
Contradiction-aware fusion	Modalities may disagree because of biology, measurement error, or contextual mismatch	Model modality reliability, discordance, missingness, and context-dependent weighting explicitly	Stress tests with deliberately incomplete or conflicting inputs	Prevents unjustified averaging of evidence	Supports scientifically interpretable multimodal systems
Shared maturity reporting	"Validation" and "readiness" are used inconsistently	Report conceptual, computational, external, experimental, and prospective evidence separately	Independent appraisal against declared maturity criteria	Improves comparison across studies	Supports cumulative and defensible field development

Conclusion

Multimodal generative artificial intelligence expands therapeutic design from the production of chemical structures toward a broader space that includes proteins, cellular states, images, text, targets, and experimentally grounded biological context. The evidence mapped in this scoping review supports distinct capabilities in representation, conditioning, alignment, fusion, editing, and cross-modal generation, but these capabilities are unevenly mature and should not be treated as interchangeable. Molecular generation has the most developed benchmark infrastructure, protein design provides important examples of experimental confirmation, omics generation remains dependent on context and causal assumptions, and image and text modalities currently function primarily as sources of conditioning and evidence. The review's proposed taxonomy, maturity interpretation, and multimodal therapeutic generation prism are conceptual organizing contributions rather than validated frameworks. Their principal purpose is to separate benchmark success from pharmaceutical usefulness, association from mechanism, prediction from experimental confirmation, and computational plausibility from therapeutic value. Progress will depend less on increasing model scale alone than on prospective multimodal datasets, claim-matched external validation, experimental testing, calibrated uncertainty, transparent provenance, negative evidence, and integration with accountable scientific decisions.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77. doi:10.1038/s41573-019-0024-5
2. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA Jr, et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov. 2020;19(5):353-64. doi:10.1038/s41573-019-0050-3
3. Sadybekov AV, Katritch V. Computational approaches streamlining drug discovery. Nature. 2023;616(7958):673-85. doi:10.1038/s41586-023-05905-z
4. Pun FW, Ozerov IV, Zhavoronkov A. AI-powered therapeutic target discovery. Trends Pharmacol Sci. 2023;44(9):561-72. doi:10.1016/j.tips.2023.06.010
5. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259-65. doi:10.1038/s41586-023-05881-4
6. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med. 2018;169(7):467-73. doi:10.7326/M18-0850
7. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol. 2018;18(1):143. doi:10.1186/s12874-018-0611-x
8. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth. 2020;18(10):2119-26. doi:10.11124/JBIES-20-00167
9. Pollock D, Davies EL, Peters MDJ, Tricco AC, Alexander L, McInerney P, et al. Undertaking a scoping review: A practical guide for nursing and midwifery students, clinicians, researchers, and academics. J Adv Nurs. 2021;77(4):2102-13. doi:10.1111/jan.14743
10. Munn Z, Pollock D, Khalil H, Alexander L, McInerney P, Godfrey CM, et al. What are scoping reviews? Providing a formal definition of scoping reviews as a type of evidence synthesis. JBI Evid Synth. 2022;20(4):950-2. doi:10.11124/JBIES-21-00483
11. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev. 2021;10(1):39. doi:10.1186/s13643-020-01542-z
12. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14(6). doi:10.15252/msb.20178124
13. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21(1):111. doi:10.1186/s13059-020-02015-1
14. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-87.e29. doi:10.1016/j.cell.2021.04.048

15. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018;4(2):268-76. doi:10.1021/acscentsci.7b00572
16. Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, et al. De novo design of protein structure and function with RFdiffusion. Nature. 2023;620(7976):1089-100. doi:10.1038/s41586-023-06415-8
17. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. Nat Methods. 2019;16(8):715-21. doi:10.1038/s41592-019-0494-8
18. Lu MY, Chen B, Williamson DFK, Chen RJ, Liang I, Ding T, et al. A visual-language foundation model for computational pathology. Nat Med. 2024;30(3):863-74. doi:10.1038/s41591-024-02856-4
19. Skalic M, Sabbadin D, Sattarov B, Sciabola S, De Fabritiis G. From target to drug: Generative modeling for the multimodal structure-based ligand design. Mol Pharm. 2019;16(10):4282-91. doi:10.1021/acs.molpharmaceut.9b00634
20. Liu S, Nie W, Wang C, Lu J, Qiao Z, Liu L, et al. Multi-modal molecule structure–text model for text-based retrieval and editing. Nat Mach Intell. 2023;5(12):1447-57. doi:10.1038/s42256-023-00759-6
21. Zeng Z, Yin B, Wang S, Liu J, Yang C, Yao H, et al. ChatMol: Interactive molecular discovery with natural language. Bioinformatics. 2024;40(9). doi:10.1093/bioinformatics/btae534
22. Wu K, Xia Y, Deng P, Liu R, Zhang Y, Guo H, et al. TamGen: Drug design with target-aware molecule generation through a chemical language model. Nat Commun. 2024;15(1):9360. doi:10.1038/s41467-024-53632-4
23. Brown N, Fiscato M, Segler MHS, Vaucher AC. GuacaMol: Benchmarking models for de novo molecular design. J Chem Inf Model. 2019;59(3):1096-108. doi:10.1021/acs.jcim.8b00839
24. Polykovskiy D, Zhebrak A, Sánchez-Lengeling B, Golovanov S, Tatanov O, Belyaev S, et al. Molecular Sets (MOSES): A benchmarking platform for molecular generation models. Front Pharmacol. 2020;11:565644. doi:10.3389/fphar.2020.565644
25. Renz P, Van Rompaey D, Wegner JK, Hochreiter S, Klambauer G. On failure modes in molecule generation and optimization. Drug Discov Today Technol. 2019;32-33:55-63. doi:10.1016/j.ddtec.2020.09.003
26. Bender A, Schneider N, Segler M, Walters WP, Engkvist O, Rodrigues T. Evaluation guidelines for machine learning tools in the chemical sciences. Nat Rev Chem. 2022;6(6):428-42. doi:10.1038/s41570-022-00391-9
27. Wallach I, Heifets A. Most ligand-based classification benchmarks reward memorization rather than generalization. J Chem Inf Model. 2018;58(5):916-32. doi:10.1021/acs.jcim.7b00403
28. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. Drug Discov Today. 2021;26(2):511-24. doi:10.1016/j.drudis.2020.12.009
29. Bender A, Cortés-Ciriano I. Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 2: A discussion of chemical and biological data. Drug Discov Today. 2021;26(4):1040-52. doi:10.1016/j.drudis.2020.11.037
30. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y). 2023;4(9):100804. doi:10.1016/j.patter.2023.100804
31. Whalen S, Schreiber J, Noble WS, Pollard KS. Navigating the pitfalls of applying machine learning in genomics. Nat Rev Genet. 2022;23(3):169-81. doi:10.1038/s41576-021-00434-9
32. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. Nat Med. 2020;26(9):1320-4. doi:10.1038/s41591-020-1041-y
33. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ. 2020;368. doi:10.1136/bmj.l6927
34. Edfeldt K, Edwards AM, Engkvist O, Günther J, Hartley M, Hulcoop DG, et al. A data science roadmap for open science organizations engaged in early-stage drug discovery. Nat Commun. 2024;15(1):5640. doi:10.1038/s41467-024-49777-x