Pharmacophore

ISSN-2229-5402

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

WHAT TRANSFERS FROM MOLECULES TO MEDICINES? A SYSTEMATIC REVIEW OF FOUNDATION MODELS ACROSS PHARMACEUTICAL SCALES

James Walker[1]*, Olivia Harris[1], George Wilson[2]

1. ***Department of Foundation Model Transfer from Molecules to Medicines, Faculty of Pharmacy, University of Edinburgh, Edinburgh, United Kingdom.***
2. ***Department of Multi-Scale Foundation Model Evaluation, Faculty of Pharmacy, University of Leeds, Leeds, United Kingdom.***

ARTICLE INFO

Received:
28 April 2026
Received in revised form:
14 August 2026
Accepted:
16 August 2026
Available online:
28 August 2026

Keywords: Foundation models, Pharmaceutical scales, Transfer learning, Molecular representation, Protein language models, Multi-omics

ABSTRACT

Foundation models are increasingly used to encode chemical structures, protein sequences, three-dimensional conformations, cellular states, biomedical knowledge, and clinical narratives. Their shared premise is that large-scale pretraining can produce reusable representations that reduce task-specific data requirements and support adaptation to multiple pharmaceutical problems. The unresolved issue is not whether such representations can improve selected benchmarks, but what information transfers across pharmaceutical scales and which evidence is required before transfer can influence medicine-development decisions. This systematic review evaluates foundation-model evidence using two complementary dimensions: scale distance, extending from molecules through proteins, structures, cells, omics, disease contexts, and development decisions; and evidence distance, extending from internal benchmark capability through external evaluation, experimental confirmation, prospective decision impact, and bounded operational readiness. Protocol-defined searching, eligibility assessment, qualitative risk-of-bias appraisal, structured evidence extraction, and narrative synthesis were used to distinguish demonstrated capability from plausible utility and translational value. The synthesis indicates that transfer is most convincingly demonstrated within individual modalities or between closely adjacent biological scales. Cross-scale claims become less secure when representations must preserve mechanistic relevance across changing assays, biological contexts, populations, or decision environments. Adaptation, grounding, calibration, leakage control, external validation, and experimental confirmation therefore function as evidence gates rather than optional performance refinements. A scale-aware interpretation of foundation models is proposed in which transfer is treated as a validated relationship among a source representation, target task, changed context, and decision consequence. The available evidence remains heterogeneous, frequently benchmark-centred, and insufficient to establish routine pharmaceutical readiness. Progress will require evaluation designs that connect representation reuse to reproducible, externally valid, and experimentally consequential decisions.

To Cite This Article: Walker J, Harris O, Wilson G. What Transfers from Molecules to Medicines? A Systematic Review of Foundation Models across Pharmaceutical Scales. Pharmacophore. 2026;17(4):125-34. https://doi.org/10.51847/44vwgWGyvz

Introduction

Foundation models are pretrained on broad data collections to produce representations that can be reused, adapted, or combined across downstream tasks. In pharmaceutical science, the relevant source material may include molecular strings and graphs, protein sequences, three-dimensional structures, gene-expression profiles, biomedical literature, electronic health records, or heterogeneous knowledge networks. This breadth has encouraged claims that a common computational paradigm might connect early molecular design with target biology, disease interpretation, drug repurposing, and clinical reasoning. Yet the reuse of a representation does not itself demonstrate that scientifically relevant information has survived a change in task, biological scale, population, assay, or decision context. Reviews of generalist medical artificial intelligence, machine learning across drug discovery and development, and AI-enabled drug design collectively show that broad computational capability remains conditional on suitable data, meaningful adaptation, rigorous validation, and alignment with pharmaceutical objectives [1-3].

Corresponding Author: James Walker; Department of Foundation Model Transfer from Molecules to Medicines, Faculty of Pharmacy, University of Edinburgh, Edinburgh, United Kingdom. E-mail: james.walker@ed.ac.uk

The central problem is therefore one of transfer rather than model size alone. A model may transfer statistically when pretrained weights improve performance on a downstream dataset, but pharmaceutical transfer demands a stronger account of what has moved, between which contexts, and with what validation. Molecular similarity may support a property prediction without establishing target engagement; sequence embeddings may predict protein annotations without demonstrating functional consequences in a disease system; structural prediction may improve access to plausible conformations without establishing ligandability, dynamics, selectivity, or therapeutic value. Biological machine-learning guidance similarly emphasizes that data partitioning, metric selection, interpretation, and independent validation must be aligned with the scientific question rather than chosen only to maximize apparent predictive performance [4]. Consequently, cross-scale transfer should not be inferred merely because one pretrained architecture can accept multiple data types or be fine-tuned for several tasks.

The distinction is particularly important for generative and representation-learning systems positioned near medicinal chemistry. Novelty, validity, or high benchmark scores can be useful technical properties, but they do not determine whether generated or prioritized compounds are synthesizable, experimentally active, selective, developable, or superior to existing decision processes. Expert appraisal of generative chemistry has therefore argued that impact should be evaluated through consequences for medicinal-chemistry practice and experimental programs rather than through computational novelty alone [5]. The same boundary applies beyond molecular generation: predictive confidence is not equivalent to calibrated uncertainty, textual fluency is not factual grounding, association is not mechanism, and retrospective prediction is not prospective decision value.

This review asks what transfers from molecules to medicines and how such transfer should be evaluated across heterogeneous pharmaceutical scales. Rather than presenting a chronology of individual models, the review organizes the evidence along two axes. The first is scale distance, distinguishing within-scale adaptation from transfer between molecules, proteins, structures, cells, omics, disease contexts, and medicine-development decisions. The second is evidence distance, distinguishing internal benchmark capability from domain-shifted evaluation, laboratory confirmation, prospective decision impact, and bounded readiness. The original contribution is a proposed scale-aware synthesis in which transfer is treated as an evidence-bearing relationship among a source representation, a target task, a changed context, and an appropriate validation gate. This synthesis is intended to clarify the evidence; it is not a validated maturity instrument, causal model, regulatory classification, or claim that pharmaceutical development follows a single linear progression.

Systematic-Review Questions and Protocol

The review was structured using protocol-defined questions, eligibility criteria, appraisal domains, extraction fields, and synthesis rules. PRISMA-compatible logic was used to distinguish identification, eligibility assessment, evidence extraction, and interpretation, while avoiding claims about screening quantities that could not be verified from auditable database exports [6]. Seven questions guided the review: how protocol transparency should be operationalized; how searching, screening, appraisal, and extraction influence the conclusions; what capabilities are demonstrated for molecular, sequence, structural, omics, and textual foundation models; what evidence supports within-scale and cross-scale transfer; how adaptation, grounding, calibration, and external validity have been evaluated; what evidence moves beyond benchmarks toward medicine development; and which research priorities are required for scale-aware models. These questions were defined before narrative synthesis so that references were selected for explicit evidentiary roles rather than incorporated retrospectively to support a predetermined technological narrative.

The unit of evidence was a peer-reviewed journal article reporting a model, benchmark, experiment, methodological framework, reporting guideline, or critical analysis relevant to pharmaceutical-scale capability or transfer. Eligible empirical articles required sufficient detail to identify the input modality, representation or pretraining strategy, adaptation procedure, target task, evaluation context, and validation level. Methodological and review articles were eligible when they directly informed search reproducibility, quality appraisal, leakage assessment, external validity, calibration, reporting, or translational interpretation. Complete database-specific strategies were developed for MEDLINE/PubMed, Embase, Scopus, Web of Science Core Collection, and IEEE Xplore with journal filters, using combinations of terms for foundation models, pretrained models, language models, self-supervised learning, transfer learning, pharmaceutical science, drug discovery, molecules, proteins, structures, single-cell data, omics, and biomedical text. Search sources, syntax adaptations, execution dates, limits, deduplication rules, and citation-chaining procedures were documented because reproducibility requires more than naming the databases searched [7].

The appraisal protocol separated reporting completeness from confidence in the methods and conclusions. A well-described study could still have high risk of bias, weak applicability, unsuitable validation, or limited pharmaceutical relevance. Conversely, incomplete reporting was treated as uncertainty rather than automatically interpreted as methodological failure. Review-level appraisal principles were adapted to examine whether objectives, eligibility decisions, search coverage, study overlap, extraction procedures, risk-of-bias judgments, and synthesis claims were sufficiently transparent and supported [8]. For empirical model studies, additional domains included data provenance, train–validation–test dependency, leakage, baseline suitability, task–metric alignment, adaptation transparency, grounding, uncertainty assessment, external testing, experimental confirmation, reproducibility, and applicability to the claimed decision. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop systematic-review questions and protocol within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for Systematic-review questions and protocol

Evidence domain	Eligible evidence or method	Reported capability or claim	Strength of support	Reproducibility or bias concern	Unresolved gap
Protocol transparency	Prespecified review questions, eligibility criteria, databases, search strategies, screening stages, extraction fields, and synthesis rules	The review process can be traced from question formulation to final interpretation	Strong methodological support when all decisions are documented before synthesis	Retrospective modification of questions or eligibility criteria may favour a preferred conclusion	Prospective registration and independent reproduction of the complete review process
Search coverage	Database searching, journal filtering, backward and forward citation chasing, DOI confirmation, and updated searching	The selected literature captures relevant model, validation, and translational evidence	Conditional on database coverage, terminology, indexing, and transparent syntax	Rapidly changing terminology may omit relevant studies; citation chasing may reinforce influential research clusters	Validated search filters for foundation models across pharmaceutical modalities
Eligibility and screening	Dual-stage title–abstract and full-text assessment with explicit inclusion and exclusion reasons	Included studies directly address scale-specific capability, transfer, reliability, or translation	Stronger when independent reviewers apply protocol-defined criteria	Subjective judgments about "foundation model," pharmaceutical relevance, or evidence maturity may affect inclusion	Greater consensus on operational definitions and reproducible classification rules
Model capability	Empirical studies reporting pretraining, downstream adaptation, task evaluation, and comparator methods	A pretrained representation supports one or more specified downstream tasks	Usually strongest for internal benchmark capability and adjacent-task transfer	Shared datasets, dependent splits, unrepresentative baselines, and task-specific tuning may inflate apparent transfer	Independent replication under changed datasets, assays, scaffolds, populations, or institutions
Cross-scale transfer	Studies connecting molecular, sequence, structural, cellular, omics, textual, disease, or decision contexts	Information learned at one scale remains useful at another scale	Variable and generally weaker as biological and decision distance increases	Improvements may arise from correlated datasets rather than preserved biological meaning	Direct tests of what information transfers, what is lost, and which adaptation steps are essential
Reliability and validity	Leakage analysis, calibration assessment, external validation, robustness testing, and applicability appraisal	Model outputs remain interpretable and useful beyond the development dataset	Strong only when evaluation matches the intended use and changed context	Confidence scores may be uncalibrated; external datasets may not be genuinely independent	Prospective, multi-context evaluation with predefined failure criteria and uncertainty reporting
Pharmaceutical translation	Synthesis, biochemical or cellular assays, perturbational studies, preclinical evaluation, or prospective workflow testing	Selected outputs survive an experimental or decision-relevant evidence gate	Stronger than benchmark evidence but still bounded by the tested system	Selective validation, missing comparators, and unsuccessful candidates may distort perceived utility	Comparative evidence that model use improves decisions, resource allocation, or development outcomes
Review synthesis	Structured extraction, qualitative appraisal, scale-distance classification, and narrative comparison	Heterogeneous evidence can be organized without treating unlike benchmarks as equivalent	Appropriate for explanatory synthesis when statistical pooling is unjustified	Reviewer interpretation may impose artificial order on diverse tasks and evidence types	Independent testing of the proposed scale-aware taxonomy and maturity distinctions

Table note: The categories are original review constructs used to organize heterogeneous evidence. They do not constitute a validated appraisal instrument, formal readiness classification, or quantitative ranking of models.

Search, Screening, Quality Assessment, and Evidence Extraction

Searches were designed to retrieve journal articles addressing foundation or broadly pretrained models in molecular science, protein sequence analysis, structural biology, single-cell and multi-omic analysis, biomedical text, drug repurposing, and

pharmaceutical decision support. A complementary reliability search targeted calibration, external validation, transportability, reproducibility, leakage, grounding, and benchmark design. Exact DOI matching was specified as the primary deduplication method, followed by normalized title, first-author, and publication-year comparison. Conference versions and subsequent journal articles were linked, with only eligible peer-reviewed journal versions considered for final inclusion. Screening proceeded through title and abstract relevance, full-text and metadata eligibility, mapping to a predefined manuscript claim, redundancy assessment, and an integrity check requiring each retained article to have a clear evidentiary role. Because prediction studies may appear technically strong while remaining biased or poorly applicable, quality assessment examined the data or participants represented, predictors and inputs, outcomes or target labels, analytical procedures, and correspondence between the evaluated setting and the claimed use [9].

Evidence extraction was designed to preserve the distinction between model description and decision-relevant evidence. For each empirical study, extraction fields included the data and application context, input modality, pretraining objective, model family, adaptation method, target task, comparator, splitting strategy, evaluation metric, uncertainty treatment, external validation, experimental confirmation, intended use, and stated limitations. For studies making prospective or clinically adjacent claims, the review additionally examined whether the protocol specified the intended users, model inputs and outputs, integration with human judgment, handling of unavailable or poor-quality inputs, and procedures for detecting or responding to model failures [10]. Reports of evaluated AI systems were separately assessed for disclosure of model versioning, human–AI interaction, error analysis, exclusions, and analytical procedures, because reproducible interpretation depends on how an intervention was actually used rather than on architecture descriptions alone [11]. These reporting frameworks were applied proportionately: they informed extraction and appraisal but were not treated as proof that pharmaceutical foundation-model systems had achieved clinical effectiveness.

Risk-of-bias coding used the qualitative categories low concern, some concern, high concern, and not clearly reported. The domains comprised data provenance and suitability; dependency among training, validation, and test sets; preprocessing or target leakage; relevance of baselines; alignment among task, metric, and intended use; adaptation transparency; biological or textual grounding; calibration or uncertainty; external validation; experimental confirmation; reproducibility; conflicts of interest; and applicability to the claimed pharmaceutical decision. Leakage received specific attention because information shared across preprocessing, feature construction, related biological entities, temporal boundaries, or dataset partitions can create apparently strong results without valid generalization [12]. The extracted evidence was then organized narratively by modality, scale distance, and validation level rather than pooled statistically across incomparable tasks. Translational interpretation required a meaningful pharmaceutical problem, representative data, independent evaluation, workflow compatibility, and evidence that the model's output could affect a real scientific or development decision [13]. Where such information was absent, the review records "not clearly reported" and restricts the conclusion rather than imputing missing methods, outcomes, or readiness.

Foundation Models For Molecular, Sequence, Structural, Omics, and Textual Tasks

At the molecular, protein-sequence, and protein-structure scales, large-scale representation learning has demonstrated reusable capability, but the meaning of transfer differs across domains. Chemical language models can encode structural and property regularities for downstream molecular tasks; protein language models can recover signals associated with evolutionary, structural, and functional organization; and structure-prediction systems can infer highly informative three-dimensional models in benchmarked settings [14-16]. These results establish that pretraining can compress domain regularities into adaptable representations. They do not establish that the same information remains sufficient when the target changes from property prediction to synthesis, from sequence annotation to therapeutic function, or from static structure to binding, selectivity, dynamics, and developability.

Single-cell and multi-omic foundation models extend pretraining from individual biological entities to heterogeneous cellular states. Generative modeling across transcriptomic and multi-omic profiles can support annotation, integration, perturbation analysis, and other downstream tasks [17]. The evidence is important because cell-state representations provide a possible bridge between molecular intervention and biological response. That bridge is nevertheless conditional on tissue coverage, assay technology, preprocessing, batch structure, perturbational diversity, and the correspondence between training cells and the biological context of use. A model that reconstructs or classifies cellular profiles may learn stable biological structure, technical regularities, or both; downstream success alone does not identify which information is responsible.

Biomedical language models address a different transfer problem: adapting statistical representations of scientific language to entity recognition, relation extraction, question answering, evidence retrieval, and other text-mining tasks. Domain-adaptive pretraining improves performance across several biomedical-language benchmarks, but the resulting capability remains dependent on source corpora, annotation conventions, task definitions, and evaluation design [18]. Textual transfer is especially vulnerable to category errors in pharmaceutical interpretation. Accurate extraction of a reported association does not verify the association, fluent synthesis does not establish factual completeness, and retrieval of mechanistic language does not demonstrate that a model is grounded in experimental biology.

Across these domains, "foundation model" therefore names a reusable training strategy rather than a uniform level of pharmaceutical evidence. Molecular, sequence, structural, omics, and textual systems differ in their observables, missing information, biological assumptions, and downstream error costs. Their benchmarks are not directly commensurable, and apparent generality may reflect repeated adaptation to related datasets rather than stable transfer across scales. The review

synthesis consequently classifies capability by the source representation, target task, adaptation burden, context change, and validation gate. This prevents a strong result in one modality from being interpreted as evidence that the model can support decisions at more distant pharmaceutical scales.

Transfer within and Across Pharmaceutical Scales

Within-scale transfer occurs when pretraining supports a new task without crossing a major biological level, whereas cross-scale transfer requires a representation to remain informative after the target moves from sequence to function, structure to activity, cell state to network behavior, or heterogeneous evidence to a development decision. Protein language models that generate sequences subsequently shown to possess experimentally tested function provide stronger evidence than sequence plausibility or computational scoring alone [19]. Even here, the validated claim remains bounded: functional activity in selected protein families does not establish therapeutic efficacy, manufacturability, immunogenicity, pharmacology, or general transfer to unrelated sequence families.

Transfer from cell-state representation to network-level prediction illustrates a second pathway. Pretraining across large collections of perturbational transcriptomes can support predictions in data-sparse contexts and can help identify biologically relevant network relationships [20]. This is a meaningful advance over training a separate model for every narrowly sampled task. However, prediction in a new cell or perturbation context still depends on whether the pretrained data cover the relevant regulatory programs and whether the downstream labels correspond to the intended biological claim. Network prediction may organize hypotheses, but it does not by itself establish causal control, treatment response, or validity in a disease population.

Adaptation strategy determines how much of a pretrained representation is genuinely reusable. Systematic comparisons of protein language models indicate that task-specific fine-tuning can improve predictions across diverse tasks, while the magnitude and direction of benefit vary with the task, model, data regime, and adaptation method [21]. Transfer should therefore not be treated as an intrinsic scalar property of a model. Frozen embeddings, full fine-tuning, parameter-efficient adaptation, prompting, retrieval, and supervised heads impose different data requirements and may alter previously learned features. Evidence for transfer should report both the baseline from which improvement is measured and the adaptation resources needed to obtain that improvement.

The most distant transfer claims arise when heterogeneous biomedical representations are used to support drug repurposing or clinician-facing prioritization. A pretrained model can integrate drugs, diseases, biomedical knowledge, and patient-related evidence for repurposing tasks, indicating that cross-domain representations may help organize candidate hypotheses [22]. Yet this is not equivalent to transferring molecular knowledge intact into treatment benefit. Each added scale introduces new confounding, missingness, temporal structure, and decision consequences. The review therefore distinguishes representational transfer, predictive transfer, experimentally grounded transfer, and decision transfer. Evidence at one level may justify progression to the next gate, but it cannot substitute for that gate.

Evaluation of Adaptation, Grounding, Calibration, and External Validity

Grounding asks whether model representations correspond to independently meaningful scientific phenomena rather than merely to dataset regularities. In viral-sequence modeling, learned representations were related to experimentally relevant patterns of evolution and immune escape, showing that language-model structure can align with biological constraints in a defined system [23]. Such alignment is stronger than success on an arbitrary proxy task, but it remains system-specific. Grounding should be tested through orthogonal assays, perturbations, structural or mechanistic consistency, temporal prediction, and failure analysis. A representation may be biologically informative without being causal, and a model may capture evolutionary constraints without supporting therapeutic intervention.

Calibration addresses a separate reliability question: whether uncertainty estimates or predicted probabilities correspond to observed frequencies or other decision-relevant uncertainty targets. High discrimination, ranking quality, or retrieval accuracy does not guarantee calibrated uncertainty [24]. This distinction matters because foundation models are often evaluated by aggregate accuracy while downstream users must decide which outputs warrant synthesis, experimentation, escalation, or rejection. Calibration should be examined after adaptation and under relevant domain shifts, not assumed from pretraining scale or confidence scores. For generative and nonprobabilistic outputs, evaluation may require task-specific measures such as selective prediction, abstention behavior, conformal coverage, stability, or empirical success rates.

External validity asks whether performance persists when the institution, assay, scaffold, sequence family, population, time period, or data-generation process changes. Multi-site imaging evidence demonstrates that models can exploit institution-specific signals and show variable performance across settings despite apparently successful internal evaluation [25]. Although the modality differs from molecular and omics modeling, the methodological lesson transfers directly: random internal splits may preserve hidden dependencies and underestimate deployment-relevant shift. Pharmaceutical foundation-model studies should therefore define the intended target population or chemical-biological domain and test changes that challenge the particular invariances the model is claimed to have learned.

Evaluation must also be multidimensional. Benchmarking of single-cell integration methods shows that rankings depend on dataset characteristics and on how biological conservation is balanced against removal of technical variation [26]. A single aggregate score can therefore conceal incompatible strengths and weaknesses. For foundation models, adaptation quality, grounding, calibration, robustness, computational burden, data efficiency, and external validity should be reported separately. The review does not infer superiority across unrelated tasks from a common metric. Instead, evidence is judged against the

intended use: a model suitable for exploratory representation may remain unsuitable for candidate exclusion, resource allocation, or safety-sensitive progression.

Evidence for Translation from Benchmark Performance to Medicine Development

Selected model-guided programs have progressed beyond virtual benchmarks to compounds that were synthesized and experimentally tested, including target-focused inhibitor generation, antibacterial discovery, and structure-enabled hit identification [27-29]. These studies demonstrate that computational prioritization can survive at least one laboratory evidence gate and can contribute to integrated discovery workflows. Their importance lies in the connection between model output and physical experiment, not in proving that any one architecture generally accelerates development. Case studies differ in targets, assays, comparator strategies, chemical starting spaces, and follow-up depth, so they cannot establish a universal efficiency advantage or routine medicine-development readiness.

Further evidence comes from model-guided identification of a structurally distinct antibiotic class supported by interpretable substructure analysis and extensive experimental evaluation [30]. This type of work strengthens translational plausibility because the computational model is linked to chemical selection, biological activity, mechanistic investigation, and preclinical testing. Nevertheless, preclinical confirmation remains bounded by the compounds, organisms, assays, and models studied. Explanation methods may help identify features associated with predictions, but they do not automatically provide causal mechanisms. Nor does successful validation of prioritized candidates establish that the model outperforms alternative screening, medicinal-chemistry, or experimental strategies under prospective comparison.

The evidence base therefore supports a graded interpretation rather than a binary judgment of success. Internal benchmarks establish capability within a sampled task. Domain-shifted evaluation tests conditional transportability. Synthesis, biochemical assays, cellular assays, perturbational studies, or animal experiments provide experimental confirmation for selected outputs. Prospective decision impact would require predeclared use of the model in an active workflow, appropriate comparators, documentation of human interaction, and measurement of consequential decisions or outcomes. Routine readiness would require repeated independent validation, monitoring, governance, and evidence that benefits persist without unacceptable scientific or operational harms.

Across the reviewed literature, the main translational gap is not the absence of impressive models but the scarcity of designs that isolate incremental decision value. Many studies show that a model can rank, generate, retrieve, or predict; fewer determine whether its use improves which experiment is performed, which candidate advances, how uncertainty is managed, or how resources are allocated compared with credible alternatives. Negative predictions, failed syntheses, inactive candidates, and unsuccessful replication are also less visible than successful examples. A defensible medicine-development claim therefore requires transparent denominators, prespecified progression rules, appropriate comparators, and reporting of failures as well as successes.

Research Agenda for Scale-Aware Foundation Models

The first priority is to design cross-scale models around explicit scientific interfaces rather than indiscriminate modality accumulation. Integrating molecular, genomic, phenotypic, and clinical data can reveal complementary information, but heterogeneity, missingness, differing resolutions, and incompatible sampling processes must be represented rather than hidden [31]. Scale-aware architectures should specify what each modality observes, how entities are linked, which temporal and causal assumptions are imposed, and where information may be lost. Shared latent spaces are useful only when alignment preserves task-relevant meaning and contradictions between modalities remain visible.

The second priority is evaluation under deliberate challenge. Genomic machine-learning experience shows how confounding, leakage, dependent samples, weak baselines, and inadequate biological validation can overstate performance and interpretation [32]. Future studies should predefine scale-specific domain shifts, maintain independent test resources, separate model selection from final evaluation, and include negative controls and ablations that identify what the pretrained representation contributes. External validation should be chosen to test the claimed transfer pathway—for example, new scaffolds for chemical generalization, remote families for protein transfer, new tissues or perturbations for omics, and later time periods or new institutions for decision support.

The third priority is evidence-bearing data governance. Dataset documentation should describe why data were collected, what they contain, how they were processed, their intended and discouraged uses, known limitations, maintenance, and relevant social or legal considerations [33]. Pharmaceutical foundation models additionally require lineage for assay protocols, molecular standardization, sequence and structure versions, cell and tissue provenance, ontology mappings, temporal cut-offs, and links between training examples and downstream evaluation sets. Documentation does not remove bias, but it makes hidden dependencies and applicability limits more inspectable and supports reproducible updating.

The fourth priority is a minimum reporting and validation standard tied to intended consequence. Biomedical AI studies should disclose essential information about data, model development, evaluation, reproducibility, and limitations [34]. For scale-aware foundation models, this minimum should also include the source and target scales, adaptation resources, grounding evidence, calibration method, domain-shift design, failure criteria, and the strongest validation gate actually crossed. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop research agenda for scale-aware foundation models within the article's central argument.

Pharmacophore, 17(4) 2026, Pages 125-134

Table 2. Evaluation, implementation, and research priorities arising from What Transfers from Molecules to Medicines? A Systematic Review of Foundation Models

Approach or application	Data and task context	Evaluation practice	Evidence maturity	Translational limitation	Review interpretation
Molecular chemical-language modeling	Molecular strings, graphs, conformations, and property tasks	Dependency-aware splits, scaffold or temporal tests, relevant baselines, and synthesis-aware assessment	Predominantly benchmark capability, with selected experimental extensions	Property prediction or generation does not establish synthesizability, selectivity, exposure, safety, or therapeutic value	Useful for representation and prioritization when chemical-domain limits and downstream evidence gates are explicit
Protein sequence modeling and generation	Evolutionary sequence corpora, function prediction, mutation effects, and conditional generation	Remote-family testing, task-specific adaptation comparisons, orthogonal functional assays, and uncertainty analysis	Strong within-scale evidence; selected sequence-to-function confirmation	Functional assays in selected families do not establish therapeutic developability or broad family transfer	Transfer is credible when task adaptation and experimental function are separately demonstrated
Structure prediction and structure-enabled design	Protein sequences, alignments, structural templates, target structures, docking, and design	Confidence assessment, structural comparison, binding assays, and prospective hit evaluation	Strong structural-prediction evidence; early development-facing proof of concept	Static structural accuracy does not establish dynamics, ligandability, selectivity, or development success	Structure is an enabling intermediate, not a surrogate for pharmacological validation
Single-cell and multi-omic foundation modeling	Transcriptomes, perturbations, cell atlases, tissues, and linked omics	Cross-dataset and cross-tissue tests, biological-conservation metrics, perturbational validation, and batch-sensitivity analysis	Multi-task benchmark and retrospective biological evidence	Technical variation, incomplete cell-state coverage, and causal ambiguity constrain transfer	Cellular representations may bridge scales when biological context and shift are explicitly tested
Biomedical text and knowledge integration	Scientific literature, ontologies, relations, drug–disease evidence, and clinical narratives	Corpus-shift tests, retrieval attribution, factuality checks, evidence tracing, and task-specific external evaluation	Mature text-mining benchmarks; emerging decision-support evidence	Linguistic accuracy does not establish factual, mechanistic, or clinical grounding	Text models can organize evidence but should not replace source verification or experimental confirmation
Cross-scale repurposing and candidate prioritization	Drugs, targets, diseases, patient-related evidence, and heterogeneous knowledge graphs	External validation, comparator workflows, uncertainty-based abstention, and expert review	Retrospective cross-domain evidence with limited prospective testing	Correlated evidence can mimic biological transfer; treatment benefit remains unproven	Appropriate for hypothesis prioritization when claims stop at the validation gate actually crossed
Prospective pharmaceutical workflow use	Live target selection, experiment choice, candidate advancement, or study-design decisions	Predeclared intended use, human–model interaction logging, credible comparator, failure reporting, and monitored outcomes	Insufficient evidence for routine readiness	Implementation may alter behavior, error propagation, resource allocation, and accountability	Highest-priority validation target; conceptual readiness must not be presented as operational deployment

Table note: Evidence maturity is classified qualitatively for review synthesis. The categories do not constitute validated readiness levels, regulatory determinations, or proof of clinical utility.

Figure 1 presents the molecule-to-medicine foundation-model transfer bridge, showing how the article's key components and boundaries are connected within research agenda for scale-aware foundation models.

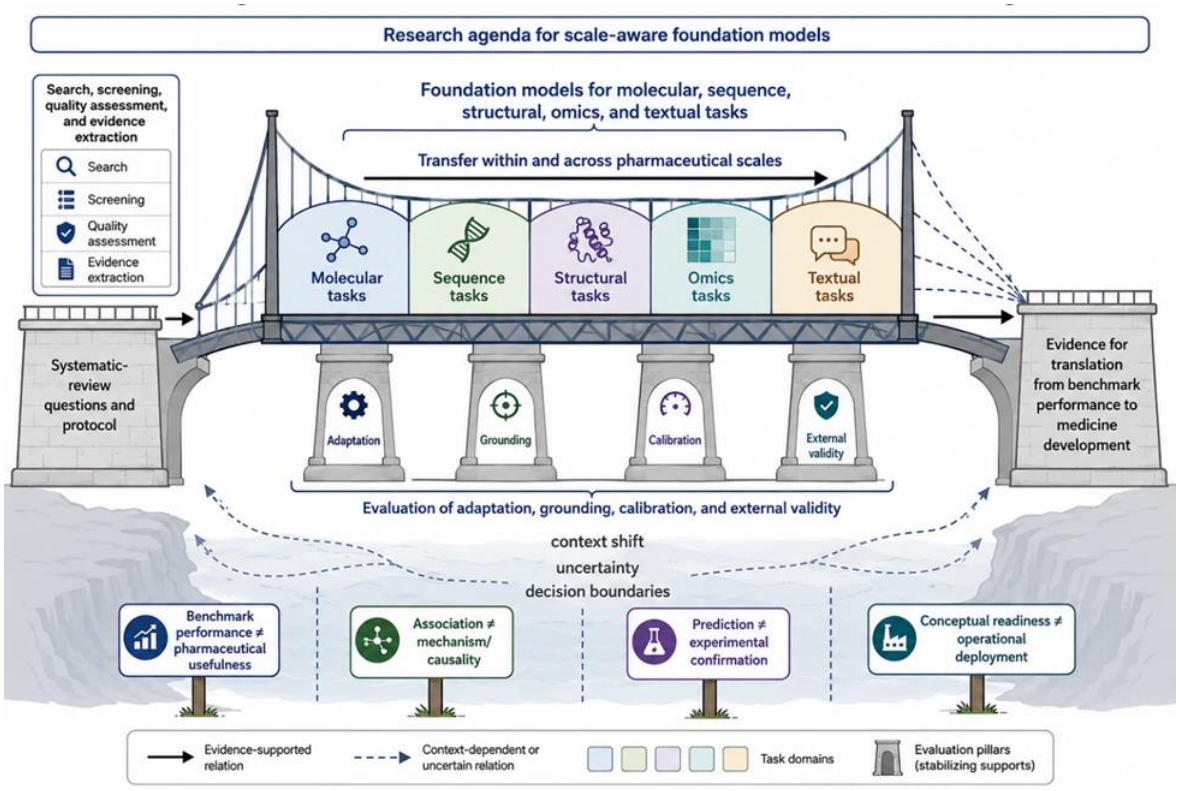

Figure 1. Molecule-to-Medicine Foundation-Model Transfer Bridge

The figure is an original conceptual synthesis that organizes the article's central contribution across systematic-review questions and protocol, search, screening, quality assessment, and evidence extraction, quality assessment, evidence extraction, foundation models for molecular, sequence, structural, omics, and textual tasks, foundation models for molecular. 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.

Conclusion

Foundation models have established reusable statistical representations across molecular, protein, structural, cellular, omic, and textual data, but the evidence does not support an unqualified claim that representations transfer seamlessly from molecules to medicines. Transfer is strongest within a modality or between adjacent scales and becomes progressively more conditional as biological context, assay conditions, populations, and decision consequences change. This review proposes that every transfer claim be specified as a relationship among a source representation, target task, changed context, adaptation procedure, and validation gate. Benchmark performance can justify further investigation; it cannot replace grounding, calibrated uncertainty, external validation, experimental confirmation, or prospective decision evaluation. Scale-aware development should therefore prioritize explicit interfaces, challenge-based testing, data lineage, failure reporting, and consequence-matched evidence. The proposed bridge and maturity distinctions organize the current literature but are not validated readiness instruments. Their value lies in preventing capability, utility, experimental success, and routine pharmaceutical use from being treated as equivalent claims.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Moor M, Banerjee O, Shakeri Hossein Abad Z, 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

2. 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
3. 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
4. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022;23(1):40-55. doi:10.1038/s41580-021-00407-0
5. Walters WP, Murcko MA. Assessing the impact of generative AI on medicinal chemistry. Nat Biotechnol. 2020;38(2):143-5. doi:10.1038/s41587-020-0418-2
6. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ. 2021;372. doi:10.1136/bmj.n71
7. 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
8. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358. doi:10.1136/bmj.j4008
9. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51-8. doi:10.7326/M18-1376
10. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. BMJ. 2020;370. doi:10.1136/bmj.m3210
11. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. BMJ. 2020;370. doi:10.1136/bmj.m3164
12. 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
13. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. doi:10.1186/s12916-019-1426-2
14. Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das P. Large-scale chemical language representations capture molecular structure and properties. Nat Mach Intell. 2022;4(12):1256-64. doi:10.1038/s42256-022-00580-7
15. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A. 2021;118(15). doi:10.1073/pnas.2016239118
16. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9. doi:10.1038/s41586-021-03819-2
17. Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods. 2024;21(8):1470-80. doi:10.1038/s41592-024-02201-0
18. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40. doi:10.1093/bioinformatics/btz682
19. Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, et al. Large language models generate functional protein sequences across diverse families. Nat Biotechnol. 2023;41(8):1099-106. doi:10.1038/s41587-022-01618-2
20. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. Nature. 2023;618(7965):616-24. doi:10.1038/s41586-023-06139-9
21. Schmirler R, Heinzinger M, Rost B. Fine-tuning protein language models boosts predictions across diverse tasks. Nat Commun. 2024;15(1):7407. doi:10.1038/s41467-024-51844-2
22. Huang K, Chandak P, Wang Q, Havaldar S, Vaid A, Leskovec J, et al. A foundation model for clinician-centered drug repurposing. Nat Med. 2024;30(12):3601-13. doi:10.1038/s41591-024-03233-x
23. Hie B, Zhong ED, Berger B, Bryson B. Learning the language of viral evolution and escape. Science. 2021;371(6526):284-8. doi:10.1126/science.abd7331
24. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt PM, et al. Calibration: The Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. doi:10.1186/s12916-019-1466-7
25. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. 2018;15(11). doi:10.1371/journal.pmed.1002683
26. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Müller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022;19(1):41-50. doi:10.1038/s41592-021-01336-8
27. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019;37(9):1038-40. doi:10.1038/s41587-019-0224-x
28. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. Cell. 2020;180(4):688-702. doi:10.1016/j.cell.2020.01.021

29. Ren F, Ding X, Zheng M, Korzinkin M, Cai X, Zhu W, et al. AlphaFold accelerates artificial intelligence powered drug discovery: Efficient discovery of a novel CDK20 small molecule inhibitor. Chem Sci. 2023;14(6):1443-52. doi:10.1039/d2sc05709c
30. Wong F, Zheng EJ, Valeri JA, Donghia NM, Anahtar MN, Omori S, et al. Discovery of a structural class of antibiotics with explainable deep learning. Nature. 2024;626(7997):177-85. doi:10.1038/s41586-023-06887-8
31. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Inf Fusion. 2019;50:71-91. doi:10.1016/j.inffus.2018.09.012
32. 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
33. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H III, et al. Datasheets for datasets. Commun ACM. 2021;64(12):86-92. doi:10.1145/3458723
34. 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