Pharmacophore

ISSN-2229-5402

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

THE TRANSPORTABILITY MAP FOR MULTI-OMICS MODELS ACROSS POPULATIONS, PLATFORMS, LABORATORIES, AND BIOLOGICAL CONTEXTS

Lensa Abebe[1]*, Fatuma Hussein[2], Chala Daba[3], Jan van den Bergh[4]

1. ***Department of Multi-Omics Transportability and Model Transfer, Faculty of Pharmacy, Addis Ababa University, Addis Ababa, Ethiopia.***
2. ***Department of Platform and Laboratory Transferability, Faculty of Pharmacy, Haramaya University, Haramaya, Ethiopia.***
3. ***Department of Biological Context and Population Adaptation, Faculty of Pharmacy, Jimma University, Jimma, Ethiopia.***
4. ***Department of Transportability Mapping and Validation, Faculty of Veterinary Medicine, Utrecht University, Utrecht, Netherlands.***

ARTICLE INFO

Received:
05 November 2024
Received in revised form:
02 February 2025
Accepted:
06 February 2025
Available online:
28 February 2025

Keywords: Multi-omics integration, Model transportability, Population shift, Batch effects, Biological-context mismatch, External validation

ABSTRACT

Multi-omics models are increasingly used to integrate genomic, transcriptomic, epigenomic, proteomic, metabolomic, and cellular information for biomarker discovery, disease stratification, target prioritization, and therapeutic-response modeling. Yet strong performance within a development dataset does not establish that a model will retain its meaning or reliability when transferred to a different population, assay platform, laboratory workflow, tissue, disease state, or intervention context. Existing evaluation practices commonly compress transportability into one external performance measure, obscuring whether failure arises from inadequate population representation, incompatible measurement processes, unstable learned representations, altered biological relationships, or combinations of these factors. This article proposes a Multi-Omics Transportability Map as a conceptual and methodological architecture for characterizing source–target differences without collapsing them into a single domain-shift label. The map distinguishes four interacting shift axes—population, platform, laboratory, and biological context—from three diagnostic mismatch classes: representation, measurement, and biological mismatch. It further links these diagnoses to external validation, adaptation, recalibration, uncertainty assessment, restricted use, additional evidence collection, or abstention. A versioned evidence ledger and graded claim structure are proposed to preserve provenance, expose conflicting evidence, and align statements of generalizability with the target settings actually evaluated. The contribution is intended to support research design, model appraisal, and transparent pharmaceutical interpretation rather than to function as a validated prediction system or deployment rule. Its usefulness remains conditional on complete metadata, appropriate target data, domain expertise, experimentally credible biological evidence, and future prospective evaluation. By reframing transportability as a relational and multidimensional property, the proposed map may help prevent benchmark performance from being mistaken for cross-context pharmaceutical usefulness.

To Cite This Article: Abebe L, Hussein F, Daba C, van den Bergh J. The Transportability Map for Multi-Omics Models across Populations, Platforms, Laboratories, and Biological Contexts. Pharmacophore. 2025;16(1):61-70. https://doi.org/10.51847/paWzUTmsvq

Introduction

Multi-omics analysis seeks to relate complementary molecular layers rather than interpret genomic, transcriptomic, proteomic, metabolomic, epigenomic, or cellular measurements in isolation. This integration can reveal coordinated disease processes, molecular subtypes, candidate biomarkers, and potentially actionable biological relationships that may remain invisible within a single data modality. However, every additional molecular layer also introduces distinct sampling procedures, measurement scales, missing-data structures, technical artifacts, and biological dependencies. Consequently, a model trained on integrated molecular evidence inherits not only the biological information contained in its source data but also the conditions under which those data were generated. Multi-omics integration therefore expands the scope of possible inference while simultaneously increasing the number of conditions that must remain sufficiently stable for those inferences to transfer [1].

Corresponding Author: Lensa Abebe; Department of Multi-Omics Transportability and Model Transfer, Faculty of Pharmacy, Addis Ababa University, Addis Ababa, Ethiopia. E-mail: lensa.abebe@aau.edu.et.

Machine-learning approaches intensify this problem because their learned representations are shaped by choices concerning modality integration, feature selection, dimensionality reduction, supervision, optimization, preprocessing, and validation. Early fusion, late fusion, latent-factor modeling, multiblock discrimination, graph-based learning, and deep generative modeling do not merely provide alternative computational routes to the same biological truth. They preserve different structures, emphasize different dependencies, and fail under different forms of source–target mismatch. A model may perform well because it captures a durable molecular relationship, but it may also exploit platform-specific intensity patterns, laboratory-specific preprocessing signatures, population composition, outcome-label practices, or other non-transportable regularities. Reviews of machine learning for multi-omics analysis accordingly show substantial heterogeneity in integration strategies, evaluation designs, interpretability, and suitability for particular scientific tasks [2].

Transportability is often treated as though it were demonstrated once a model is applied to an external dataset and produces an acceptable summary metric. This interpretation is too broad. Generalizability is relational: it concerns whether a model developed under one set of conditions remains scientifically meaningful for a specified target population, measurement environment, biological state, and intended use. A model can discriminate adequately in a new dataset while being miscalibrated, can preserve average performance while failing in underrepresented subgroups, or can align molecular profiles while erasing rare or context-specific biology. Conversely, apparent performance decline may reflect correct detection of a genuinely different biological regime rather than mere technical instability. The broader critique of generalizability in health-related modeling therefore applies directly to multi-omics research: no model is transportable in the abstract, because transportability can only be assessed in relation to a defined target setting [3].

The unresolved problem is not simply the absence of stronger normalization, larger training datasets, or more complex learning architectures. It is the absence of a sufficiently explicit conceptual structure for separating why a multi-omics model may or may not transfer. This article addresses that gap by proposing the Multi-Omics Transportability Map, an evidence-organizing architecture that distinguishes population, platform, laboratory, and biological-context shifts; diagnoses representation, measurement, and biological mismatch; and links those diagnoses to external validation, adaptation, recalibration, uncertainty characterization, restricted use, or abstention. The map does not assign a universal transportability score and does not claim that all forms of shift can be corrected. Instead, it is designed to preserve the provenance, assumptions, uncertainty, and target specificity of transportability evidence so that benchmark performance is not mistaken for mechanistic validity, experimental confirmation, clinical usefulness, or pharmaceutical readiness.

Transportability Failures in Contemporary Multi-Omics Modeling

A first source of transportability failure is invalid evidence produced before genuine source–target transfer is even examined. Information leakage can occur when feature selection, normalization, imputation, latent-representation construction, hyperparameter tuning, or sample grouping incorporates information from data later described as independent evaluation data. In multi-omics studies, leakage may be especially difficult to detect because preprocessing is frequently performed separately for multiple molecular layers before the resulting representations are combined. Related samples, repeated specimens, technical replicates, longitudinal observations, or institution-specific batches can also be divided across development and evaluation partitions in ways that compromise independence. Under such conditions, apparently strong performance may reflect access to target information or duplicated structure rather than a model's ability to generalize beyond the development workflow. Leakage and non-independent evaluation can consequently produce false evidence of reproducibility and transportability even when standard performance metrics appear convincing [4].

A second failure arises from the assumption that one benchmark ranking identifies a generally superior integration method. Comparative evaluation of joint multi-omics dimensionality-reduction approaches shows that method performance depends on the dataset, integration objective, biological task, and assessment criterion. A representation that is useful for recovering known cancer subtypes may not be optimal for survival-associated structure, cross-omics feature relationships, sample clustering, or another downstream application. This dependence matters because the model's transportability is partly determined by what its representation was optimized to preserve. If the source task rewards a feature structure that is absent, reweighted, or biologically altered in the target setting, an apparently successful integration may fail even when all source data were processed correctly. Transportability should therefore be evaluated in relation to the intended scientific use of the representation rather than inferred from an undifferentiated benchmark rank [5].

A third failure occurs when technical harmonization is treated as equivalent to preservation of biological meaning. Batch-correction methods are generally intended to reduce technical variation, but stronger mixing across batches can remove genuine biological differences when batch membership is correlated with population, tissue, disease stage, treatment, or cell-state composition. Conversely, preservation of all observed biological separation may also preserve technical artifacts. Benchmarking of single-cell batch-correction methods demonstrates that batch removal and biological conservation are distinct, potentially competing objectives that require separate assessment [6]. This distinction is especially important for pharmaceutical multi-omics studies, in which treatment exposure, sampling centre, assay date, and disease severity may be structurally linked. A correction procedure cannot recover a clean biological signal merely because the corrected embedding appears visually well mixed.

The same multidimensional problem becomes more evident at atlas scale, where integration methods must balance removal of unwanted technical structure, preservation of meaningful biology, scalability, rare-state retention, and downstream utility. No single metric adequately represents all of these properties, and an integration may appear successful under one criterion while

failing under another [7]. Comparative benchmarks collectively show that method rankings change with dataset composition, batch structure, biological-conservation objective, and evaluation metric [5–7]. The proposed transportability model therefore treats performance values as evidence components rather than final transportability determinations. It asks what the model retained, what it removed, which source conditions shaped the result, which target conditions differ, and what kinds of validity remain untested. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop transportability failures in contemporary multi-omics modeling within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for Transportability failures in contemporary multi-omics modeling

Construct or Evidence Domain	Core Question	Scientific Basis	Role in the Article	Failure or Overclaiming Risk	Boundary Statement
Evaluation Independence	Was the target evidence protected from model development, preprocessing, feature selection, and optimization?	Valid external appraisal requires separation between development decisions and evaluation information.	Establishes whether apparent transportability evidence is admissible before its magnitude is interpreted.	Leakage may produce optimistic performance that is incorrectly described as external generalization.	A technically separate dataset is not independent when information from it influenced the analytical workflow.
Source-Data Composition	Which populations, tissues, disease states, interventions, platforms, and laboratories shaped the learned model?	Learned representations reflect the distribution and provenance of their training evidence.	Defines the empirical support from which later transport claims are made.	Large sample size may be mistaken for broad biological or technical coverage.	More source observations do not compensate for absent target-relevant contexts.
Representation Objective	What structure was the integration method optimized to preserve?	Supervised, unsupervised, factor-based, graph-based, and generative approaches retain different relationships.	Connects computational architecture to the kinds of transfer that can reasonably be assessed.	A representation useful for one task may be treated as universally biologically informative.	Latent structure is model- and task-dependent rather than a direct observation of biological truth.
Batch-Effect Control	Were technical differences reduced without eliminating target-relevant biological variation?	Harmonization can trade technical mixing against biological conservation.	Separates correction quality from biological validity.	Visual batch mixing may conceal overcorrection or removal of rare states.	Technical alignment alone does not establish preserved biology.
Biological-Signal Preservation	Are relevant cell states, pathways, subtype relationships, treatment responses, or molecular gradients retained?	Transportability requires preservation of the biological structure needed for the intended task.	Provides a counterweight to purely technical measures of integration success.	Known annotations may be preserved while unknown or context-specific biology is erased.	Preservation of selected markers does not prove preservation of every relevant mechanism.
Task-Conditioned Performance	Does the model remain useful for the specific target task rather than merely within a generic benchmark?	Method rankings vary with endpoints, datasets, and evaluation criteria.	Prevents benchmark superiority from being generalized beyond the tested use.	One favorable metric may be reported as comprehensive model validity.	Performance is conditional on the evaluated task, outcome definition, and target context.
Subgroup and Rare-State Behavior	Does acceptable average performance conceal failure in underrepresented groups or uncommon biological states?	Aggregate measures can obscure heterogeneous support and error.	Extends transportability assessment beyond population averages.	Minority groups or rare cell states may be forced into dominant source patterns.	Average performance cannot establish equitable or biologically complete transport.
Provenance and Preprocessing Stability	Are assay, laboratory, feature-selection, normalization, and software choices traceable and reproducible?	Preprocessing choices can materially alter integrated representations and model behavior.	Makes the route from raw measurement to transportability claim auditable.	Unreported analytical decisions can be mistaken for inherent model properties.	Reproducibility of a workflow does not by itself demonstrate external validity.
Metric Concordance	Do technical, biological, predictive, calibration, and uncertainty assessments support the same conclusion?	Different evaluation dimensions may disagree because they test different properties.	Preserves conflicting evidence instead of collapsing it into a composite score.	A favorable composite may conceal failure in a consequential dimension.	Disagreement between metrics is diagnostic evidence, not merely statistical inconvenience.
Pharmaceutical Interpretation	Does the transported output support only prediction, or also experimental, mechanistic, translational, or decision-relevant inference?	Predictive association and pharmaceutical usefulness require different evidence.	Establishes the ultimate boundary of the article's central argument.	Cross-dataset prediction may be presented as target validation, mechanism, efficacy, or therapeutic value.	Transported prediction alone does not establish causality, experimental confirmation, safety, developability, or clinical utility.

Population, Platform, Laboratory, and Context Shifts

Population shift refers to differences between the source and target in genetic ancestry, demographic composition, environmental exposure, disease prevalence, comorbidity, treatment history, socioeconomic conditions, sampling pathways,

or other variables that influence molecular measurements and their relationship to outcomes. Human genetic evidence remains concentrated in a restricted subset of global populations, limiting the range of allele frequencies, linkage structures, environmental interactions, and disease presentations represented during discovery and model development [8]. In multi-omics modeling, this underrepresentation can propagate across molecular layers: ancestry-associated genomic structure may interact with gene expression, methylation, protein abundance, metabolism, medication exposure, diet, or disease stage. Population shift should therefore not be reduced to a single demographic covariate. It is a compound source–target difference whose components may be partly biological, partly environmental, partly social, and partly consequences of study recruitment and health-system access.

The consequences of insufficient population representation are illustrated by the reduced transfer of polygenic risk models across ancestry groups. Differences in allele frequencies, linkage disequilibrium, estimated effect sizes, phenotype definitions, and training-data composition can substantially alter predictive performance and may amplify disparities when source-trained scores are applied without appropriate population-specific evaluation [9]. The broader multi-omics implication is not that every population requires an entirely separate model, nor that ancestry categories should be treated as fixed biological entities. Rather, model developers must determine which population-dependent relationships are actually relevant to the target task, which groups were sufficiently represented, whether performance and calibration remain stable within those groups, and whether observed differences originate from genetic structure, environmental context, measurement practice, access to care, or combinations of these factors. Population transportability is thus both a scientific-validity problem and an equity-sensitive boundary on permissible generalization.

Platform shift concerns differences in the instruments, chemistries, molecular capture processes, sequencing depth, dynamic range, detection limits, feature definitions, and data structures through which biological material becomes computational input. Such differences do not simply add random noise. Comparative analysis of single-cell and single-nucleus RNA-sequencing methods shows that sampling and protocol choices systematically alter the molecular signals observed, including the cellular compartments represented and the relative detection of transcript classes [10]. A source model may consequently learn relationships that depend on the accessibility of intact cells, nuclear enrichment, transcript capture, or platform-specific feature availability. Even when two platforms nominally measure the same omics layer, the resulting variables may differ in scale, missingness, sensitivity, and biological interpretation. Cross-platform normalization can reduce numerical discrepancies, but it cannot recreate information that a target assay does not measure or guarantee that apparently corresponding features retain equivalent meaning.

Laboratory shift includes differences in site, operator, reagent lot, specimen handling, storage, processing interval, instrument maintenance, quality-control practices, and local analytical workflow. A multicentre benchmark using shared reference samples demonstrates that site and technology effects can remain detectable even when laboratories examine common biological materials under coordinated protocols [11]. These effects may be nested within platform differences, but they should not be treated as interchangeable: the same platform can behave differently across sites, and a laboratory can introduce variability across multiple platforms. Biological-context shift adds a further layer encompassing tissue, cell state, disease stage, species, developmental condition, treatment exposure, temporal phase, and regulatory environment. Unlike many technical differences, genuine biological-context changes may not be correctable and may indicate that the source-learned relationship is no longer applicable. The proposed map therefore treats population, platform, laboratory, and biological context as interacting but analytically distinct axes. Their purpose is not to classify every discrepancy into one exclusive category, but to prevent a generic “domain shift” label from concealing the particular evidence needed to determine whether transfer is plausible, testable, correctable, or scientifically inappropriate.

The Proposed Multi-Omics Transportability Map

The proposed Multi-Omics Transportability Map conceptualizes transportability as a relationship between a defined source system, a defined target-use system, and the evidence available to connect them. The source specification records populations, specimens, molecular modalities, platforms, laboratories, preprocessing decisions, labels, prediction tasks, and learned representations. The target specification records the corresponding conditions under which reuse is proposed. Between them, four shift axes—population, platform, laboratory, and biological context—describe where relevant differences may arise. Latent-factor approaches demonstrate that integrated representations can contain shared and modality-specific variation, supporting the need to record which dimensions of variation a source model was constructed to preserve [12]. The map nevertheless treats such factors as model-dependent representations rather than direct measurements of biological truth.

A second layer describes the model’s representational purpose. Unsupervised integration may prioritize dominant covariance, shared variation, or sample organization, whereas supervised integration may preferentially retain molecular features associated with a specified phenotype or outcome. DIABLO illustrates how multiblock covariance and feature selection can be explicitly conditioned on an outcome-related task [13]. This task dependence means that a representation transported successfully for classification may not remain suitable for mechanistic interpretation, subgroup discovery, target prioritization, or treatment-response modeling. The proposed map therefore requires every transportability claim to identify the scientific task, the molecular relationships that must remain stable, and the consequences of failure. It does not permit a general statement that a representation is transportable without specifying what it is expected to transport.

A third layer records structural conditions that affect source–target comparison, including missing modalities, heterogeneous data likelihoods, group structure, feature availability, and reference coverage. MOFA+ demonstrates that multimodal and

multi-group integration can be extended to heterogeneous observations and incomplete views [14]. Statistical accommodation of missingness, however, does not establish that an absent target modality can be reconstructed reliably or that groups are biologically equivalent. Reference-atlas transfer further shows that new observations can be mapped into an existing representation while retaining some query-specific variation [15]. Such transfer remains conditional on whether relevant target states are represented by the reference, whether feature definitions are compatible, and whether novel states are allowed to remain uncertain or unassigned. **Figure 1** presents the multi-omics transportability bridge, showing how the article's key components and boundaries are connected within the proposed multi-omics transportability map.

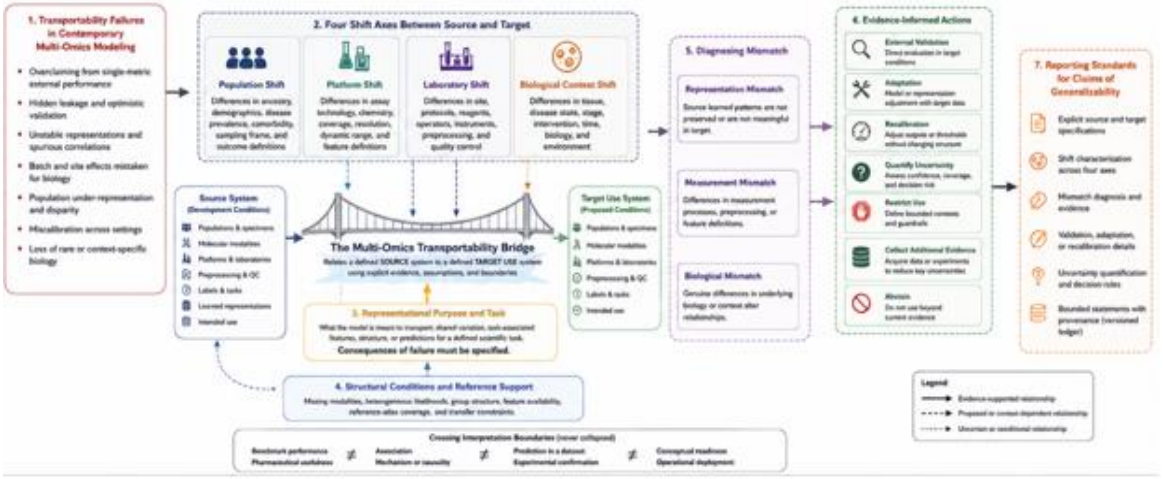

Figure 1. Multi-Omics Transportability Bridge.

The figure is an original conceptual synthesis that organizes the article's central contribution across transportability failures in contemporary multi-omics modeling, population, platform, laboratory, and context shifts, context shifts, proposed multi-omics transportability map, diagnosing representation, measurement, and biological mismatch, diagnosing representation. 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.

The bridge terminates not in one transportability score but in a versioned evidence ledger and a bounded claim grade. The ledger records source and target specifications, observed shifts, diagnostic results, unchanged external performance, adaptations, recalibration, residual uncertainty, subgroup behavior, and unresolved biological questions. Claim grades may range from untested or unsupported to context-bound or conditionally transportable for a precisely defined target use. These proposed grades are not validated thresholds and should not be interpreted as regulatory categories. Their function is to constrain language: evidence obtained for one population, platform, laboratory, or biological context cannot automatically justify use in another. The map may therefore support comparison, study planning, and transparent scientific communication, but its practical value requires future evaluation against real transport failures and independent expert judgment.

Diagnosing Representation, Measurement, and Biological Mismatch

Representation mismatch occurs when the source model's features, latent variables, neighborhoods, or learned correspondences fail to encode target-relevant structure. Mutual-nearest-neighbor correction illustrates this dependence by identifying locally corresponding observations across batches and using them to estimate correction vectors [16]. The approach is informative for transportability because it makes an important assumption visible: sufficiently similar biological populations must exist in both datasets. When the target contains novel cell states, altered disease processes, or different population structure, a mathematically close neighbor may not represent a biologically equivalent state. Representation diagnosis should therefore examine feature support, neighborhood stability, latent-factor consistency, rare-state retention, and uncertainty in correspondence rather than relying only on global visual alignment.

Harmony provides another form of representation-level correction by iteratively adjusting a shared embedding while encouraging locally diverse clustering across datasets [17]. Such methods may efficiently reduce dataset-specific structure, but their success depends on recoverable shared biological organization. Excessive correction may obscure meaningful subgroup or treatment effects, whereas insufficient correction may leave site or platform signatures available to the downstream model. The proposed map therefore distinguishes the disappearance of an identifiable batch pattern from evidence that target biology has been preserved. Diagnostic questions should include whether corrected clusters retain independent molecular support, whether target-specific states remain detectable, and whether improvement in one integration metric is accompanied by deterioration in another.

Anchor-based integration similarly relies on identifiable correspondences between source and target observations. Anchors can support data transformation, label transfer, and cross-dataset comparison when equivalent states are sufficiently represented [18]. They become less reliable when a target state is absent from the source, when corresponding states have

changed under treatment or disease progression, or when platform effects alter the feature space used to establish correspondence. A transportability analysis should consequently report not only transferred labels or aligned embeddings but also anchor coverage, uncertainty, unsupported regions, and evidence for biological equivalence. Forced assignment of every target observation to a source category may conceal the very novelty that external evaluation should detect.

Measurement mismatch concerns how molecular quantities are generated, detected, scaled, and rendered missing. Reference-material-based correction can help estimate technical variation across large transcriptomic, proteomic, or metabolomic studies, including settings in which biological groups and analytical batches are partly confounded [19]. Its validity depends on whether the reference is stable, processed comparably, and commutable to the biological samples. Representation-oriented methods also encode different assumptions about shared states, local neighborhoods, cluster structure, and valid cross-dataset correspondence [16–18]. Biological mismatch is more fundamental: it occurs when the molecular–phenotypic or molecular–mechanistic relationship itself changes across tissue, disease stage, intervention, species, or population. Technical harmonization cannot prove that such relationships remain invariant, and transported associations cannot alone establish causal mechanism or therapeutic relevance.

External Validation, Adaptation, Recalibration, and Uncertainty

External validation should begin with the unchanged source model and a target dataset that was not used to select features, tune preprocessing, choose hyperparameters, or decide whether adaptation was needed. Its purpose is not merely to reproduce an aggregate score but to determine how model behavior changes under the target population, measurement process, laboratory workflow, and biological context. Analyses of cross-site probabilistic model failure show that transport can be impaired by differences in covariates, measurement procedures, labels, workflows, and conditional relationships even when the nominal task remains unchanged [20]. Multi-omics validation should therefore report target-data provenance, subgroup behavior, feature availability, missing modalities, prediction errors, biological-conservation measures, and uncertainty rather than presenting one external metric as a complete transportability determination.

Recalibration is appropriate when the model retains useful ordering or structural information but its predicted probabilities or quantitative outputs are systematically misaligned with the target setting. Calibration must be assessed separately from discrimination because a model can rank observations adequately while consistently overestimating or underestimating the relevant outcome [21]. Recalibration can correct some forms of target-specific misalignment, but it does not repair missing biological states, incompatible assays, invalid labels, or changed mechanisms. Moreover, a recalibrated model is no longer identical to the source model and requires its own external assessment. The map therefore records unchanged validation results before modification so that adaptation does not erase evidence of the original transport failure.

Uncertainty assessment should address predictive, representational, measurement, and structural uncertainty. Probabilistic generative modeling can explicitly represent count-generating processes, latent biological structure, library effects, and posterior uncertainty in single-cell transcriptomics [22]. Such estimates can be valuable when the target differs in sequencing depth, feature availability, or sample composition, but posterior uncertainty remains conditional on the assumed model. A system can remain confidently wrong when the source model omits an important shift mechanism or encounters biology outside its support. Uncertainty should therefore be tested empirically through coverage, subgroup calibration, unsupported-state detection, sensitivity analyses, and selective abstention under realistic source–target differences. **Figure 2** presents the evidence, validation, and translation boundary observatory for the transportability map for multi-omics models across, showing how the article's key components and boundaries are connected within external validation, adaptation, recalibration, and uncertainty.

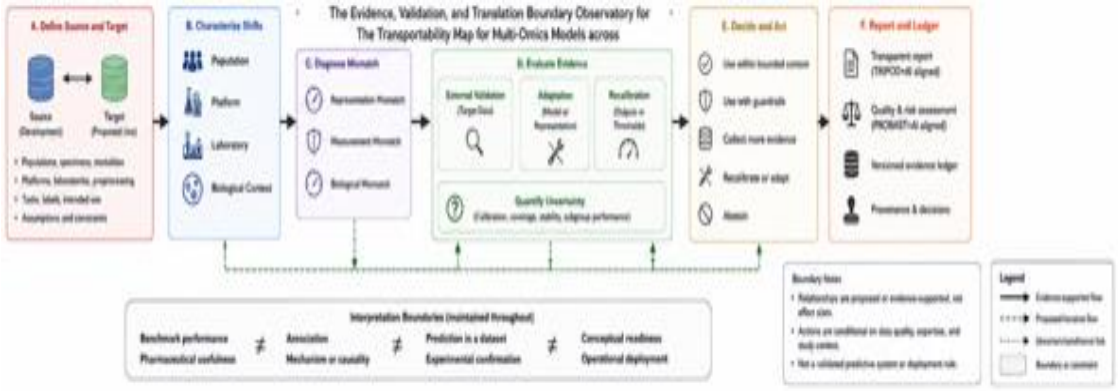

Figure 2. Evidence, Validation, and Translation Boundaries.

The figure is an original conceptual synthesis that shows how claims move from conceptual plausibility through evaluation, uncertainty assessment, and bounded pharmaceutical use. 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.

Adaptation may involve feature remapping, representation updating, domain-informed normalization, transfer learning, target-specific fine-tuning, or restricted retraining. These interventions should be selected in response to a diagnosed mismatch rather than used reflexively after poor external performance. Feature-selection choices alone can materially alter single-cell integration and reference-query performance, demonstrating that preprocessing is part of the transportability specification rather than a neutral preliminary step [23]. Every adapted model should be evaluated on untouched target evidence and compared with the unchanged source model to detect negative transfer. The proposed response layer therefore permits several outcomes: conditional transport after independent validation, recalibration for a specified target, adaptation followed by new validation, restricted use, collection of additional evidence, redevelopment, or abstention. Abstention is treated as a scientifically valid conclusion when the available evidence cannot support transfer.

Reporting Standards for Claims of Generalizability

Claims of generalizability should begin with a precise description of the model's source, target, task, and intended interpretation. Biological machine-learning validation guidance emphasizes separate reporting of data, optimization, model, and evaluation decisions [24]. For multi-omics models, this should include specimen selection, molecular modalities, assay versions, laboratories, preprocessing, feature construction, missing-data handling, representation learning, outcome definition, subgroup composition, and all uses of target data. A claim such as "externally validated" is insufficient unless readers can determine whether the evaluation dataset was independent, whether preprocessing was locked, which target differences were present, and whether performance was examined beyond one aggregate measure.

Computational reproducibility is another necessary but insufficient condition. Reproducibility standards for life-science machine learning emphasize access to data or justified substitutes, executable code, software environments, model configurations, parameters, and analytical decisions [25]. These elements allow others to reconstruct how a result was produced and to identify hidden dependencies on software versions, feature dictionaries, reference assemblies, or preprocessing choices. Reproducing a reported value does not establish transportability, however, because the same workflow may reproduce the same source-specific artifact. Reporting should therefore distinguish computational repeatability from independent target validity, biological preservation, calibration, robustness, and pharmaceutical usefulness.

Generalizability reporting should also distinguish model development, internal evaluation, unchanged external validation, adaptation, and post-adaptation testing. TRIPOD+AI guidance supports explicit reporting of intended use, participants, predictors, outcomes, model development, evaluation, and applicability [26]. Adapted to multi-omics research, such reporting should identify which modalities were present in each dataset, whether feature spaces differed, how target missingness was handled, whether target labels informed model modification, and which claims remain untested. The term "external" should not obscure close similarity between source and target institutions, shared preprocessing, overlapping participants, common reference resources, or retrospective selection of a favorable target dataset.

Applicability and risk of bias must remain distinct from reported performance. PROBAST+AI provides a structured basis for examining participants, predictors, outcomes, analysis, and target applicability rather than accepting predictive results at face value [27]. Existing reporting initiatives converge on the need to document data provenance, model development, evaluation design, external validation, computational reproducibility, and intended use [24–26]. The proposed reporting standard adds transportability-specific requirements: explicit source–target comparison, shift-axis description, mismatch diagnosis, unchanged and adapted results, calibration, subgroup and rare-state behavior, uncertainty, unresolved biology, and a bounded claim grade. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop reporting standards for claims of generalizability within the article's central argument.

Table 2. Evaluation, implementation, and research priorities arising from The Transportability Map for Multi-Omics Models across Populations, Platforms, Laboratories, and Biological

Component or Layer	Inputs	Core Function	Expected Output	Uncertainty or Limitation	Validation Requirement
Source Specification	Population composition, specimens, modalities, assay platforms, laboratories, preprocessing, feature space, labels, task, model version.	Defines the evidence and conditions from which the model was developed.	Versioned source profile.	Metadata may be incomplete or aggregated; recorded variables may not capture all relevant processes.	Independent audit of metadata completeness and analytical provenance.
Target-Use Specification	Intended population, platform, laboratory, biological state, outcome, workflow, and decision context.	Replaces vague generalizability claims with a defined target proposition.	Explicit target-use statement.	Future targets may be only partly observable before application.	Prospective confirmation that the evaluated target corresponds to intended use.
Population-Shift Assessment	Ancestry, demography, environment, disease spectrum, treatment exposure, subgroup support.	Identifies underrepresentation and potentially unstable population-dependent relationships.	Population-support profile and subgroup limitations.	Population categories may combine heterogeneous biological, environmental, and social processes.	Independent subgroup evaluation with appropriate uncertainty and calibration assessment.

Platform-Shift Assessment	Assay chemistry, instrument, resolution, feature definitions, detection limits, paired measurements, reference materials.	Determines whether source and target measurements are analytically comparable.	Platform-compatibility profile.	Numerical agreement may not imply equivalent biological information.	Paired-sample and reference-material studies under target conditions.
Laboratory-Shift Assessment	Site, operator, reagents, handling, storage, processing time, quality-control procedures.	Identifies procedural and site-dependent variation.	Laboratory-effect profile and residual-site analysis.	Laboratory factors may be confounded with biological groups.	Multicentre replication or deliberately balanced site studies.
Biological-Context Assessment	Tissue, cell state, disease stage, intervention, time, species, developmental condition, pathway evidence.	Tests whether source-learned relationships remain biologically relevant.	Context-compatibility statement and unresolved mechanism list.	Biological states may be continuous, incompletely annotated, or causally uncertain.	Independent experimental, perturbational, temporal, or mechanistic evidence.
Representation Diagnosis	Feature overlap, latent-factor stability, neighborhood preservation, anchor coverage, query uncertainty.	Determines whether the source representation can encode target-relevant structure.	Representation-mismatch profile.	Apparent embedding overlap can conceal rare-state loss or overcorrection.	Evaluation with independent biological markers and downstream target tasks.
Measurement Diagnosis	Technical replicates, reference materials, batch indicators, missingness, scaling, detection limits.	Separates measurement incompatibility from biological difference.	Measurement-mismatch profile and correction rationale.	Technical and biological effects may be inseparable in confounded designs.	Validation of correction using independent or deliberately replicated materials.
Biological-Mismatch Diagnosis	Outcome relationships, subgroup effects, perturbations, pathway coherence, temporal stability.	Tests whether the modeled molecular relationship changes in the target context.	Biological-mismatch statement.	Observational stability cannot establish mechanism or causality.	Experimental or prospective evidence appropriate to the claimed interpretation.
Unchanged External Validation	Locked source model, independent target data, predefined metrics and subgroup analyses.	Measures target behavior before modification.	Transparent external-validation record.	One target dataset cannot establish universal applicability.	Replication across independently selected target settings.
Adaptation and Recalibration	Diagnosed mismatch, target features, calibration data, transfer or updating procedure.	Modifies the model only where evidence justifies a target-specific response.	Versioned adapted model and comparison with the unchanged model.	Adaptation may overfit small target datasets or create negative transfer.	Untouched post-adaptation external evaluation.
Residual Uncertainty and Abstention	Predictive distributions, coverage tests, unsupported states, subgroup uncertainty, sensitivity analyses.	Defines where model outputs remain unreliable or should not be issued.	Uncertainty profile, restricted-use conditions, or abstention decision.	Estimated uncertainty may omit structural or mechanistic uncertainty.	Prospective testing under realistic compound shifts.
Evidence Ledger	All source–target metadata, diagnostics, validation results, adaptations, limitations, and version history.	Preserves provenance and conflicting evidence without collapsing it into one score.	Auditable transportability record.	Completeness depends on data access and reporting quality.	Inter-rater evaluation, reproducibility testing, and prospective use studies.
Graded Generalizability Claim	Evidence ledger, applicability assessment, intended consequence, residual uncertainty.	Aligns scientific language with the exact target settings supported.	Untested, unsupported, context-bound, or conditionally transportable claim.	Proposed grades are not validated regulatory or deployment categories.	Empirical assessment of consistency, usefulness, and predictive validity of claim grades.

Limitations and Boundary Conditions

The Multi-Omics Transportability Map is a proposed conceptual architecture rather than an empirically validated instrument. Its components are grounded in established problems involving batch effects, population representation, platform variation, integration assumptions, external validation, calibration, and reporting, but the combined map has not been prospectively tested. The categories may overlap, and some failures will resist clean attribution. A platform change may interact with laboratory handling, a population difference may be confounded with recruitment site, and biological context may alter both measurement quality and outcome definition. The map should therefore support structured investigation rather than imply that every transport failure has one identifiable cause.

The quality of the map's output is conditional on metadata and target evidence. Historical datasets may lack assay-version records, reagent information, sample-processing times, participant context, or raw measurements needed to distinguish technical from biological variation. Small or selectively sampled target datasets may produce unstable subgroup, calibration, or uncertainty estimates. Reference materials may not reproduce the complexity of primary tissues, and known annotations may favor familiar biological states while obscuring novel ones. No diagnostic score can recover information that was never measured, remove complete confounding retrospectively, or establish biological invariance where relevant experimental evidence is absent.

The map also cannot convert transported prediction into pharmaceutical validation. A stable association across datasets does not establish causality, target tractability, synthesizability, developability, pharmacokinetics, safety, therapeutic efficacy, clinical utility, or regulatory acceptability. Likewise, a successfully aligned molecular representation does not prove that the same mechanism operates across tissues, disease stages, treatments, species, or populations. Human oversight remains necessary to interpret mismatch evidence, weigh consequences, determine whether adaptation is scientifically defensible, and decide when abstention or new data collection is preferable. The proposed framework should not be used as an automated authorization tool or as a substitute for experimental, clinical, ethical, or regulatory evaluation.

Research Agenda and Implementation Priorities

A first priority is the creation of factorial transportability benchmarks in which population, platform, laboratory, and biological-context differences are varied deliberately rather than encountered as inseparable historical confounders. Such studies should use shared specimens, technical replicates, multiple laboratories, diverse populations, complementary assays, and defined perturbations where feasible. Evaluation should include unchanged external performance, representation preservation, calibration, subgroup behavior, rare-state recovery, uncertainty, and negative transfer after adaptation. The purpose would not be to rank methods universally but to identify which diagnostic signatures correspond to particular shift mechanisms and which methods fail when several shifts occur simultaneously.

A second priority is development of interoperable provenance and reporting infrastructure. Multi-omics studies require machine-readable descriptions of assay versions, feature definitions, preprocessing, reference resources, model versions, adaptation histories, and source–target relationships. Evidence ledgers should be versioned so that recalibration or fine-tuning cannot overwrite the unchanged external-validation record. Shared schemas should permit linkage among raw-data provenance, model documentation, biological annotations, calibration analyses, and applicability assessments. Their evaluation should examine completeness, inter-rater agreement, reproducibility, usability, and whether they reduce unsupported generalizability claims rather than merely increase reporting burden.

A third priority is prospective evaluation in pharmaceutical research workflows. Candidate settings include biomarker transfer across cohorts, reference-atlas querying, cross-platform pharmacogenomics, disease-subtype modeling, target prioritization, and treatment-response prediction. Studies should prespecify the target use, source–target shifts, permitted adaptations, stopping criteria, abstention rules, and independent evidence required before a transported result influences further experimentation. Particular attention should be given to biological mismatch, because technically successful transfer may still preserve an association that is mechanistically irrelevant in the target context. Implementation should proceed in stages—from retrospective map completion, to blinded external evaluation, to prospective research use—without implying clinical or operational readiness.

Conclusion

The proposed Multi-Omics Transportability Map reframes model transfer as a multidimensional, source–target evidence problem rather than a property demonstrated by one performance measure. It separates population, platform, laboratory, and biological-context shifts from representation, measurement, and biological mismatch; preserves unchanged validation results before adaptation; and links residual uncertainty to bounded claims, restricted use, further evidence collection, or abstention. Its principal contribution is an explicit architecture for asking what is being transported, between which conditions, through which measurements and representations, with what evidence, and for what scientific purpose. The map does not validate a model, establish biological mechanism, demonstrate pharmaceutical usefulness, or authorize deployment. Its value will depend on future empirical evaluation, complete provenance, independent target evidence, and disciplined interpretation of what transported predictions can and cannot establish.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18(1):83. doi:10.1186/s13059-017-1215-1
2. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: A review. Biotechnol Adv. 2021;49:107739. doi:10.1016/j.biotechadv.2021.107739
3. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health. 2020;2(9). doi:10.1016/S2589-7500(20)30186-2

4. 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
5. Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D, Remy E, et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nat Commun. 2021;12(1):124. doi:10.1038/s41467-020-20430-7
6. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020;21(1):12. doi:10.1186/s13059-019-1850-9
7. 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
8. Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. Cell. 2019;177(1):26-31. doi:10.1016/j.cell.2019.02.048
9. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51(4):584-91. doi:10.1038/s41588-019-0379-x
10. Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat Biotechnol. 2020;38(6):737-46. doi:10.1038/s41587-020-0465-8
11. Chen W, Zhao Y, Chen X, Yang Z, Xu X, Bi Y, et al. A multicenter study benchmarking single-cell RNA sequencing technologies using reference samples. Nat Biotechnol. 2021;39(9):1103-14. doi:10.1038/s41587-020-00748-9
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. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics. 2019;35(17):3055-62. doi:10.1093/bioinformatics/bty1054
14. 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
15. Lotfollahi M, Naghipourfar M, Luecken MD, Khajavi M, Büttner M, Wagenstetter M, et al. Mapping single-cell data to reference atlases by transfer learning. Nat Biotechnol. 2022;40(1):121-30. doi:10.1038/s41587-021-01001-7
16. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018;36(5):421-7. doi:10.1038/nbt.4091
17. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16(12):1289-96. doi:10.1038/s41592-019-0619-0
18. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888-902. doi:10.1016/j.cell.2019.05.031
19. Yu Y, Zhang N, Mai Y, Ren L, Chen Q, Cao Z, et al. Correcting batch effects in large-scale multiomics studies using a reference-material-based ratio method. Genome Biol. 2023;24(1):201. doi:10.1186/s13059-023-03047-z
20. Lasko TA, Strobl EV, Stead WW. Why do probabilistic clinical models fail to transport between sites. NPJ Digit Med. 2024;7(1):53. doi:10.1038/s41746-024-01037-4
21. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating Diagnostic Tests and Prediction Models' of the STRATOS Initiative. Calibration: The Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. doi:10.1186/s12916-019-1466-7
22. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12):1053-8. doi:10.1038/s41592-018-0229-2
23. Zappia L, Richter S, Ramírez-Suástegui C, Kfuri-Rubens R, Vornholz L, Wang W, et al. Feature selection methods affect the performance of scRNA-seq data integration and querying. Nat Methods. 2025;22(4):834-44. doi:10.1038/s41592-025-02624-3
24. Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G, ELIXIR Machine Learning Focus Group, et al. DOME: Recommendations for supervised machine learning validation in biology. Nat Methods. 2021;18(10):1122-7. doi:10.1038/s41592-021-01205-4
25. Heil BJ, Hoffman MM, Markowetz F, Lee SI, Greene CS, Hicks SC. Reproducibility standards for machine learning in the life sciences. Nat Methods. 2021;18(10):1132-5. doi:10.1038/s41592-021-01256-7
26. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385. doi:10.1136/bmj-2023-078378
27. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388. doi:10.1136/bmj-2024-082505