



Phenotypic Screening Meets Machine Learning in Small-Molecule Discovery Through Cellular Signatures, Representation Learning, Mechanism Inference, and Compound Prioritization

Mikko Lahtinen^{1*}, Elina Salo¹, Juhani Virtanen²

¹Department of Phenotypic Screening and Machine Learning, Faculty of Pharmacy, University of Helsinki, Helsinki, Finland.

²Department of Cellular Signatures and Mechanism Inference, Faculty of Pharmacy, University of Eastern Finland, Kuopio, Finland.

ABSTRACT

Phenotypic screening measures compound-induced biological change without requiring a predefined molecular target, yet modern assays increasingly generate images, transcriptional states, and other high-dimensional readouts whose interpretation depends on computational representation. This integrative review synthesizes peer-reviewed evidence published from 2017 through 2025 across image-based profiling, perturbational transcriptomics, representation learning, mechanism inference, chemical-phenotypic modeling, compound prioritization, and experimental confounding, interpreting studies according to task definition, validation design, mechanistic evidence, and transferability rather than pooling across non-equivalent benchmarks. The literature supports a progression from measured cellular responses to reusable signatures and learned representations that can support perturbation retrieval, activity prediction, response modeling, mechanism-oriented hypothesis generation, and prioritization of compounds for testing, with morphological, transcriptional, and chemical representations providing overlapping but non-equivalent information. This review proposes an integrated phenotype-to-decision framework in which mechanism inference and compound prioritization branch from learned biological-response spaces while remaining conditioned by cell context, technical structure, applicability domain, and experimental validation. Strong internal prediction does not itself establish target identity, causal mechanism, external generalization, or prospective discovery value. Phenotypic machine learning is most defensible as an evidence-organizing and hypothesis-prioritizing strategy, and its discovery value depends on separating representational usefulness, mechanistic attribution, and experimentally validated decision use while treating discordance, confounding, and domain shift as scientific information rather than preprocessing nuisances.

Key Words: Phenotypic screening, Machine learning, Cellular signatures, Morphological profiling, Perturbational transcriptomics, Compound prioritization

eIJPPR 2025; 15(3):99-107

HOW TO CITE THIS ARTICLE: Lahtinen M, Salo E, Virtanen J. Phenotypic Screening Meets Machine Learning in Small-Molecule Discovery Through Cellular Signatures, Representation Learning, Mechanism Inference, and Compound Prioritization. Int J Pharm Phytopharmacol Res. 2025;15(3):99-107. <https://doi.org/10.51847/WhSV4kOqyu>

INTRODUCTION

Phenotypic drug discovery begins with a measurable biological response rather than requiring that a compound first satisfy a predefined molecular-target hypothesis. This can expose biologically useful responses that are not

readily reducible to one molecular target, but it also leaves target deconvolution and mechanistic interpretation as substantial downstream challenges [1]. The contemporary computational problem is therefore broader than distinguishing active from inactive wells. Modern phenotypic experiments can produce multidimensional

Corresponding author: Mikko Lahtinen

Address: Department of Phenotypic Screening and Machine Learning, Faculty of Pharmacy, University of Helsinki, Helsinki, Finland.

E-mail: ✉ mikko.lahtinen@helsinki.fi

Received: 20 February 2025; **Revised:** 30 May 2025; **Accepted:** 01 June 2025

This is an **open access** journal, and articles are distributed under the terms of the Creative Commons Attribution-Non Commercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.



descriptions of how cells or organisms respond to perturbation, making representation itself part of the discovery problem.

Deep learning and increasingly mature high-content profiling strategies, including Cell Painting, have expanded the information that can be extracted from cellular assays and connected to perturbational identity or downstream prediction [2, 3]. A screen can consequently generate reusable biological-response signatures rather than only scalar endpoints. Such signatures may support retrieval of related perturbations, prediction of downstream activity, prioritization of compounds, or hypotheses concerning pathways and targets.

This expansion creates a critical evidential distinction. A high-dimensional representation can be predictive without being mechanistically specific, and a phenotypic neighborhood can be biologically informative without identifying a unique molecular cause. Technical variation may additionally become encoded alongside treatment biology. Phenotypic machine learning is therefore considered here not as an automated path from image or transcript to mechanism, but as a set of transformations through which measured biology acquires different forms of computational and decision value.

Review purpose and integrative questions

This review examines how cellular responses become computational representations and how those representations are subsequently used in small-molecule discovery. Four linked questions organize the synthesis. First, what biological information is retained when high-dimensional images, expression states, or other

perturbational measurements are converted into cellular signatures? Second, how does representation learning alter the similarities that become visible between compounds, cells, and perturbations? Third, under what conditions can those similarities support mechanism-oriented inference without assuming that every phenotype maps uniquely to one target? Fourth, what additional evidence is required before representation-derived information becomes defensible compound-prioritization evidence?

The analysis therefore connects morphological profiling, perturbational transcriptomics, learned cellular representations, chemical–phenotypic integration, mechanism inference, and compound ranking while keeping their validation requirements distinct. Particular attention is given to cell context, technical batch structure, experimental shortcuts, applicability domain, and transfer between datasets or laboratories.

The central synthesis proposed here is that phenotypic machine learning operates across several evidential levels rather than one linear prediction pipeline. A model may improve representation without improving mechanism attribution; mechanism-oriented information may assist prioritization without becoming definitive target evidence; and compounds may sometimes be prioritized directly from reproducible biological-response patterns while molecular mechanism remains unresolved.

The search and eligibility structure supporting this synthesis is summarized in **Table 1**, distinguishing candidate capture, screening, eligibility assessment, and inclusion without implying exhaustive systematic-review coverage.

Table 1. Search, eligibility, and evidence-selection architecture of the integrative review

Review stage	Verified review record	Interpretive role
Evidence scope	Peer-reviewed journal evidence published from 2017 through 2025 with verifiable DOI and direct relevance to the review questions	Defined the eligible computational, biological, and validation evidence space
Information sources	PubMed/MEDLINE-targeted searching supplemented by publisher or journal searching and citation-chain discovery	Captured evidence distributed across imaging, transcriptomics, cheminformatics, and drug-discovery literatures
Candidate capture	49 records retained during targeted searching	Established the auditable candidate pool without implying an exhaustive database denominator
Duplicate/version control	5 duplicate, preprint, or alternative-version records removed	Prevented duplicate evidence attribution
Screening	44 unique records screened; 8 excluded before detailed article-level assessment	Removed bibliographic and substantive mismatches
Detailed eligibility	36 records assessed; 7 excluded for redundancy or supersession, insufficient incremental evidence, or ineligible bibliographic year	Favored direct, nonredundant evidence with claim-level utility
Included literature	29 articles retained across phenotypic profiling, perturbational signatures, representation learning, mechanism inference, multimodal modeling, prioritization, and confounding	Supplied the qualitative integrative evidence base

Note: Article-level eligibility assessment is used because the preserved review record does not justify describing every assessed item as a conventionally retrieved full-text article.

The mature argument developed across the review is integrated in **Figure 1**, which shows why phenotypic data domains, learned representations, mechanism inference,

and compound prioritization must remain distinct while being jointly conditioned by context, technical structure, external transfer, and experimental validation.

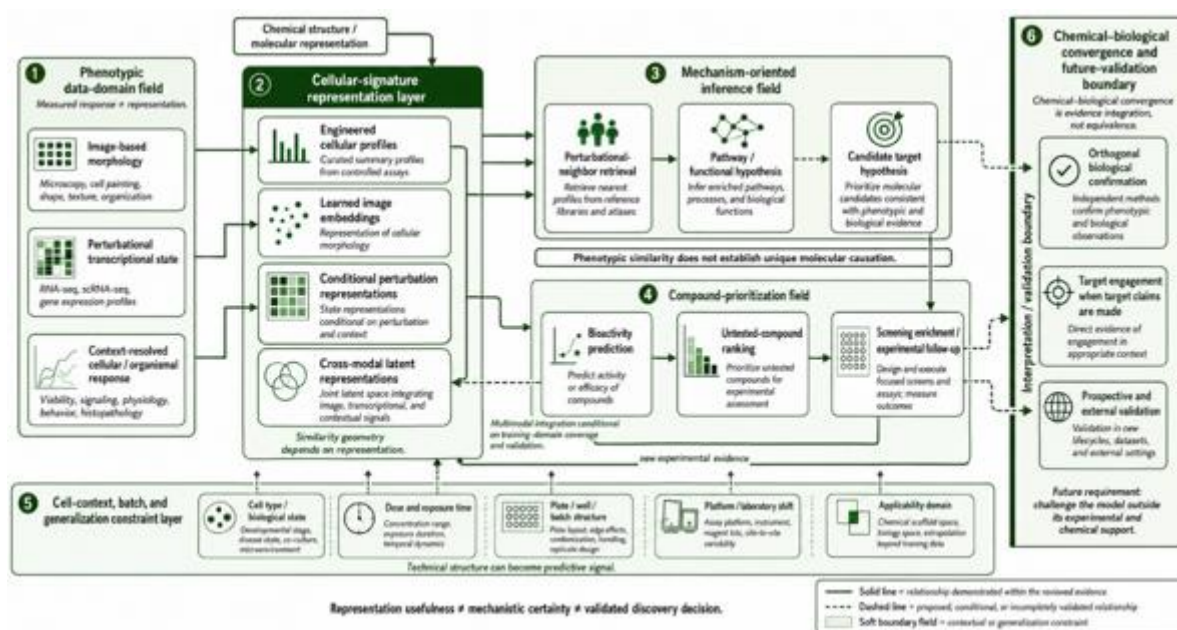


Figure 1. Evidence-qualified integration of phenotypic signatures, mechanism inference, and compound prioritization

Integrative review methodology

This integrative review was designed for an evidence base in which imaging studies, perturbational transcriptomic resources, predictive models, representation benchmarks, mechanism-oriented analyses, and experimental screening studies answer related but non-equivalent questions. The unit of synthesis was therefore the claim supported by a study design, not a pooled model-performance statistic. Computational studies were interpreted according to input modality, prediction task, representation, comparator, holdout design, and applicability boundary. Experimental studies were additionally assessed according to whether observed associations were followed by biological, mechanistic, or prospective testing.

Integration proceeded across the relationship between measurement, representation, inference, and decision use. A study predicting an assay endpoint from morphology was consequently not treated as equivalent to one retrieving mechanistically related perturbations, predicting an unseen transcriptional response, or experimentally enriching a screening campaign. Reviews and large resources supplied methodological or data-space context, whereas empirical and benchmarking studies anchored narrower claims.

This design permits comparison across heterogeneous evidence while avoiding artificial aggregation of accuracy, retrieval, enrichment, or response-prediction metrics that represent different scientific tasks. It also allows

contradictory or incomplete evidence to remain visible rather than forcing every result into a single performance hierarchy.

Eligibility criteria

Eligible sources were peer-reviewed journal articles published from 2017 through 2025 with a verifiable DOI and direct relevance to phenotypic screening, cellular-signature construction, image or perturbational representation learning, mechanism inference, chemical-phenotypic integration, compound prioritization, or the technical and biological factors governing those tasks. A source required sufficient methodological or experimental specificity to support a defined manuscript claim. Articles were excluded when they fell outside the publication interval, were not peer-reviewed journal articles, lacked a verifiable DOI, duplicated another publication version, or addressed an adjacent topic without a direct synthesis role. Primary experimental or computational studies were preferred where they directly addressed a review question. Earlier methodological, platform, or conceptual papers were retained when they supplied foundational evidence not adequately replaced by later work.

Information sources and search strategy

Evidence discovery used PubMed/MEDLINE-targeted searching supplemented by publisher and journal searches

and citation-chain exploration. Search concepts covered phenotypic screening, Cell Painting, high-content and morphological profiling, perturbational transcriptomics, single-cell response prediction, representation learning, mechanism inference, chemical-phenotypic modeling, compound prioritization, batch correction, experimental shortcuts, applicability domains, and generalization.

Search variants also targeted opposing or limiting concepts because failure modes were considered part of the review question rather than ancillary methodological detail. The captured search record contained 49 candidate records. After removal of five duplicate, preprint, or alternative-version records, 44 unique records underwent screening. These counts describe the preserved integrative-review search record and are not presented as an exhaustive denominator of all potentially discoverable publications.

Evidence selection and critical appraisal

Eight screened records were excluded before detailed article-level assessment. Thirty-six records progressed to detailed eligibility assessment, after which seven were excluded: three because their substantive contribution was redundant or superseded within the selected evidence base, three because they provided insufficient incremental evidence relative to stronger direct studies, and one because final bibliographic verification placed it outside the eligible publication interval.

Critical appraisal was claim-specific. Journal prominence, algorithmic complexity, or high benchmark performance was not treated as evidence of mechanistic validity. Greater inferential weight was assigned when validation challenged the specific claim being made—for example through meaningful holdouts, relevant comparators, experimental confirmation, explicit confounder testing, or a demonstrated transfer setting. Mechanism-oriented predictions were kept distinct from biochemical binding, target engagement, and causal attribution unless the study directly established those additional evidential levels.

Data extraction and integrative synthesis

Extraction recorded study purpose, design, perturbational or chemical space, biological system, assay or data source, molecular or cellular representation, computational method, comparator, validation scheme, reported outcome, mechanistic evidence, uncertainty source, and transferability boundary. Unreported information was retained as unreported rather than inferred. This was important because apparently similar models can answer materially different questions depending on whether validation uses random observations, unseen compounds, unseen biological contexts, separate batches, or genuinely external experiments.

Synthesis proceeded across three linked dimensions. The first was the response space being represented, including morphology, transcriptional state, single-cell distributions, and organism-level phenotypes. The second was the computational inference made from that space, such as similarity retrieval, activity prediction, perturbation-response prediction, mechanism nomination, or compound ranking. The third was the evidential distance to a discovery decision.

This final dimension prevented model output from being interpreted as stronger evidence than its validation allowed. A nearby compound in phenotypic space may justify a similarity hypothesis without demonstrating pharmacological equivalence. A predicted target may motivate an experiment without establishing binding or cellular engagement. Likewise, a model that discriminates labels internally may not remain informative when cell state, laboratory conditions, chemical domain, or experimental layout changes. The synthesis therefore preserves uncertainty at the transition between prediction and interpretation rather than attaching it only after candidate selection.

Characteristics of the included literature

The included literature spans several methodological generations of phenotype-based computation. Foundational work establishes phenotype-led discovery and the analytical requirements of high-dimensional cell profiling. More recent studies include large chemical and genetic perturbation resources, learned image representations, perturbational transcriptional and single-cell models, comparative chemical-phenotypic analyses, cross-modal representation learning, mechanism-oriented retrieval or target prediction, experimentally evaluated compound prioritization, and explicit studies of batch effects or learned experimental shortcuts.

The biological unit represented is similarly heterogeneous. Some methods summarize populations as morphological or transcriptional profiles; others represent single-cell distributions or organism-level behavioral responses. Their prediction endpoints also differ. Retrieval of a related perturbation, prediction of an assay label, prediction of a transcriptional response, nomination of a pathway or target, and enrichment of compounds selected for experimentation are distinct validation problems rather than interchangeable measures of model quality.

Accordingly, the literature is best interpreted as an interconnected set of representation and inference problems. Its diversity reveals where biological-response information can complement chemical structure, but also where apparent convergence depends on assay context, reference data, training-domain coverage, and the experimental evidence used to test a computational claim.

How phenotypic screening became a high-dimensional data problem

High-content phenotypic screening becomes a data problem before it becomes a machine-learning problem. Image-derived profiles depend on acquisition quality, feature extraction, normalization, quality control, and statistical treatment of replicate and batch structure [4]. These decisions determine which biological and technical differences remain visible to subsequent models. A profile is therefore not a neutral transcription of cellular state; it is an analytical construction produced from the assay.

Scale increases both opportunity and ambiguity. Large collections containing matched chemical and genetic perturbations make it possible to test whether phenotypic representations recover known biological relationships, yet incomplete or discordant matches show that morphological resemblance is not equivalent to identical mechanism [5]. The relevant discovery object is consequently a context-indexed high-dimensional response rather than a self-interpreting phenotype. Machine learning can organize that response space, but biological meaning still depends on what was measured, how representation was formed, and what reference perturbations are available for comparison.

Cellular signatures as learnable representations

Representation learning changes the geometry through which perturbations are compared. Learned image representations can improve perturbation profiling and retrieval relative to alternative feature spaces in evaluated datasets [6]. The important consequence is not merely dimensionality reduction. Different representations can emphasize different morphological regularities, making “similarity” dependent on the encoder, training objective, normalization strategy, and assay context.

Contemporary morphological profiling therefore spans handcrafted measurements, deep image embeddings, and hybrid strategies rather than one canonical signature format [7]. This review treats a cellular signature as a computationally usable description of response, not as a biological ontology. A signature may preserve information useful for activity prediction while obscuring another feature required for mechanism inference. Representation quality must accordingly be judged against the downstream scientific task, not against compression efficiency alone.

Image-based phenotypes and morphological profiling

Image-derived phenotypes can support compound-level prediction beyond the assay in which images were originally collected. Repurposed high-throughput image assays have been used to predict biological activities in drug-discovery settings [8], and convolutional neural networks have extracted microscopy representations

capable of predicting evaluated assay outcomes [9]. These studies establish that morphology contains reusable predictive information; they do not establish that an image representation uniquely resolves molecular mechanism.

More directly relevant to decision use, Cell Painting-based prediction has improved hit enrichment and compound diversity in experimentally evaluated screening campaigns [10]. This moves the evidence beyond retrospective classification toward tested prioritization, while remaining campaign- and assay-dependent. Morphological profiling is therefore valuable because it can compress multidimensional cellular consequences into a searchable or predictive representation, but its utility depends on the biological coverage of the assay and the experimental regime in which the model is subsequently applied.

Transcriptomic and multi-omic perturbation signatures

Perturbational transcriptomics provides a complementary response space in which compounds and genetic perturbations are compared through induced molecular states. The L1000 Connectivity Map established large-scale systematic comparison of perturbational gene-expression signatures across compounds, genes, doses, and cellular contexts [11]. More recent harmonized single-cell resources extend this logic while making heterogeneity in cell type, protocol, and perturbation design more explicit [12].

Machine learning increasingly attempts to predict responses not directly observed. scGen demonstrated latent-variable prediction of single-cell perturbation responses in evaluated settings [13], while conditional perturbation modeling extended this principle to more complex combinations of perturbation and covariate information [14]. Chemical-conditioned generative modeling has further addressed transcriptional responses to previously unseen compounds within defined benchmark regimes [15]. TranSiGen similarly connects chemical structure with induced transcriptional representations for phenotype-oriented discovery tasks [16].

These approaches broaden the concept of a phenotypic signature from a static vector to a conditional response model. Their principal limitation is correspondingly important: extrapolation remains bounded by chemical, cellular, dosing, and experimental support represented during training.

Representation learning across compound and cell spaces

Not every useful response is adequately represented by an average profile. Neural optimal transport has been used to predict perturbational changes between distributions of single-cell states, retaining information that can be lost when heterogeneous populations are collapsed into one mean vector [17]. This creates a second axis of

representation choice: models differ not only in modality but also in biological resolution.

The implication for small-molecule discovery is that compound representations and cell-state representations need not meet at one fixed abstraction level. Structure can describe what a molecule is; a cellular signature describes what it did under specified conditions; a distributional representation describes how heterogeneous cells moved through state space. These objects can interact computationally without being treated as equivalent measurements.

Inferring mechanism without assuming a single target

Phenotypic machine learning can support mechanism inference without assuming that every informative response must collapse immediately to one molecular target. Functional representations of perturbational gene signatures have been used to predict compound targets, demonstrating that response patterns can nominate target

hypotheses [18]. Image-profile matching has likewise been used to identify small-molecule pathway regulators for experimental follow-up [19].

The appropriate evidential interpretation is graded. A phenotypic neighbor can suggest a shared functional consequence; correspondence with a genetic perturbation can strengthen a mechanistic hypothesis; and a target-prediction model can prioritize a molecular explanation. None of these, by itself, establishes biochemical binding, cellular target engagement, or unique causation. Mechanism inference is therefore best treated as a competitive hypothesis space in which multiple molecular explanations can coexist until discriminating evidence is introduced.

The distinctions between cellular-signature evidence and progressively more specific mechanistic interpretations are organized in **Table 2**. The table deliberately separates what a data domain can support from the mechanistic claim that remains to be tested.

Table 2. Evidence domains linking cellular signatures to mechanism-oriented interpretation

Evidence domain	Typical analytical object	Discovery inference supported	Boundary that remains
Morphological profiling	Engineered or learned cellular image profile	Perturbation similarity and activity-oriented prediction	Similar morphology need not imply identical molecular cause
Chemical-genetic morphology matching	Matched perturbational profiles	Functional or pathway relationship	Incomplete matching and phenotypic degeneracy remain possible
Perturbational transcriptomics	Induced expression signature	Functional-state comparison and perturbation connectivity	Shared expression response need not identify a unique target
Conditional response modeling	Predicted cellular or transcriptional state	Response hypothesis for unmeasured conditions	Validity remains conditional on chemical and biological support
Functional target prediction	Learned functional response representation	Candidate target nomination	Prediction does not establish molecular engagement
Image-profile virtual screening	Phenotypic-reference matching	Candidate pathway-regulator prioritization	Orthogonal experimental confirmation remains necessary

Note: The distinctions constitute an integrative organization of the reviewed evidence, not a validated hierarchy of mechanistic certainty.

The practical consequence of this graded interpretation is illustrated in **Figure 2**: phenotypic resemblance can supply both mechanism hypotheses and compound-ranking

evidence, but these routes require different validation gates and need not occur in a fixed sequence.

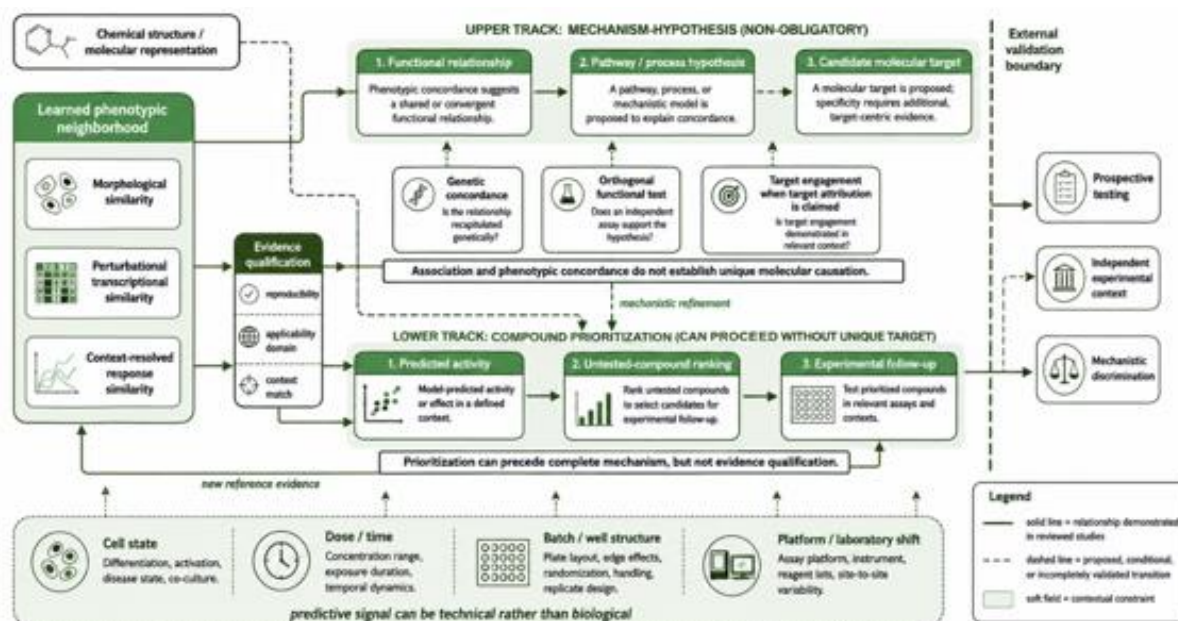


Figure 2. From phenotypic resemblance to falsifiable mechanism and prioritization decisions

Predicting compound similarity from biological response

Chemical resemblance and biological-response resemblance answer different questions. Direct comparisons show that chemical structure and cell morphology contain partly complementary information for bioactivity prediction, while morphology and gene expression capture overlapping but non-identical aspects of cellular state. Joint evaluation of phenotypic profiles and chemical structures likewise shows endpoint-dependent predictive value rather than one universally superior representation [20-22].

Compound similarity should therefore be regarded as representation-specific. Two structurally dissimilar molecules may converge phenotypically through related functional consequences, whereas close analogues can diverge because stereochemistry, permeability, metabolism, target engagement, or cellular context changes the response. Biological similarity is useful precisely because it can expose relationships not encoded by conventional structural resemblance, but it should not be promoted to pharmacological equivalence without additional evidence.

Prioritizing untested molecules from learned phenotypic space

A practical use of learned phenotypic space is deciding which untested compounds deserve experimental resources. Similarity to training data can inform how predictions from chemical fingerprints and morphology should be combined, making applicability domain part of model fusion rather than an afterthought [23]. Contrastive learning can also align molecular structures with cellular images so that chemical queries retrieve relevant

phenotypic information [24]. Graph-image contrastive models extend this principle by jointly learning from molecular and cellular representations for downstream prediction [25].

These approaches support a shift from screening every available compound toward evidence-weighted prioritization. Their value nevertheless remains conditional on coverage. A compound far from both the chemical and phenotypic support of training data should not acquire confidence merely because a model returns a score. Prioritization is therefore strongest when prediction is accompanied by explicit information about representation similarity, context match, and the kind of experimental confirmation required next.

Cell context, batch effects, and biological confounding

High-dimensional models can learn experimental organization as readily as intended biology. A deep phenotyping study in zebrafish identified a shortcut associated with experimental structure and reran the experiment with well-wise randomization; perturbative-map benchmarking and Cell Painting batch-correction studies likewise show that technical organization materially affects learned biological relationships [26-28]. This problem cannot be solved simply by maximizing batch mixing, because aggressive correction can remove genuine biological variation. Cell type, physiological state, dose, exposure time, plate layout, imaging conditions, and processing history all influence the signature available to a model. Confounding is therefore part of the biological interpretation problem: before a learned neighborhood is treated as meaningful, the analysis must ask what

experimental variables could have generated the same geometry.

Cross-platform and cross-laboratory generalization

Generalization has at least two separable dimensions. Biological transfer asks whether a learned relationship persists across cell states, perturbation regimes, or response distributions; technical transfer asks whether it persists across batches, platforms, and laboratories. Large matched perturbation resources expose the first challenge [5], while harmonized single-cell collections demonstrate how heterogeneous experimental contexts remain even after data integration [12].

Distribution-aware prediction provides one strategy for modeling cell-state shifts [17], but it does not remove laboratory or platform variation. Perturbative-map benchmarking shows that representation choices influence cross-dataset comparability [27]. Batch-correction benchmarking additionally demonstrates the trade-off between suppressing technical variation and retaining biological signal [28]. External validity should therefore be tested as a property of a defined transfer scenario, not inferred from strong internal performance.

Convergence of chemical and phenotypic representations

Chemical and phenotypic representations converge most productively when their differences are preserved. Morphology can provide information not captured by structural fingerprints [20]. Transcriptional state contributes complementary information to morphology [21]. Combined phenotype–structure models also show that the relative value of each modality varies by endpoint and chemical neighborhood [22].

Multimodal methods make this complementarity computationally actionable. Similarity-aware fusion can alter how strongly each representation contributes to a prediction [23]. Contrastive alignment allows chemical structures and images to occupy interoperable latent spaces [24]. Graph–image representation learning can further support transfer to downstream tasks [25]. Convergence should therefore mean conditional evidence integration, not conversion of chemistry and phenotype into interchangeable variables.

The resulting evidence pattern is summarized in **Table 3**, where each representation is distinguished by the information it contributes, the discovery use it can support, and the boundary that prevents overinterpretation.

Table 3. Evidence synthesis across chemical and phenotypic representation spaces

Representation space	What it primarily encodes	Discovery contribution	Principal failure boundary
Chemical structure	Molecular constitution and learned structural features	Analogue-oriented prediction and chemical-domain coverage	Structural similarity can miss divergent cellular response
Morphological phenotype	Integrated cellular consequences visible by imaging	Bioactivity-oriented prediction and response-based similarity	Context or technical structure can dominate morphology
Transcriptional phenotype	Perturbation-induced molecular state	Functional-state and response comparison	Shared expression states need not imply one target
Chemical–phenotypic fusion	Complementary structural and response information	Endpoint-dependent multimodal prediction	Fusion is not guaranteed to improve every task
Cross-modal latent space	Learned correspondence between molecular and image representations	Retrieval and transfer across modalities	Latent proximity is not pharmacological equivalence
Graph–image pretraining	Joint molecular and cellular representation	Downstream representation transfer	Generalization remains dependent on pretraining domain

Note: The table expresses evidence-supported distinctions and a proposed synthesis of their discovery roles; it does not rank representation modalities universally.

Emerging discovery architectures

Generative modeling extends phenotypic machine learning from interpreting measured responses toward predicting responses that have not yet been experimentally observed. Generative models have been used to predict cell-morphological responses under held-out perturbational conditions [29]. This is conceptually important because the model output becomes a proposed phenotype rather than merely a label attached to an observed profile.

The emerging architecture is therefore likely to combine conditional response prediction, cross-modal alignment, chemical structure, and experimentally acquired reference phenotypes. Its central scientific challenge is falsifiability.

A realistic-looking or internally consistent generated response is not evidence that the biology occurred. Synthetic phenotypes should be treated as hypotheses about future measurements, with value determined by whether they improve experimental selection or survive prospective testing.

Evidence gaps and review limitations

Several gaps constrain the field. External validation across laboratories, platforms, biological contexts, and chemical domains remains less common than internal benchmarking. Harmonized perturbation resources expose substantial contextual heterogeneity [12]. Perturbative-

map studies show that preprocessing and representation choices can alter apparent biological relationships [27]. Explicit shortcut detection is especially important because predictive performance can remain high while the learned signal is scientifically misleading [26].

This review is also limited by deliberate integration of heterogeneous tasks. Retrieval, activity prediction, response generation, target nomination, and screening enrichment cannot be meaningfully pooled into one performance estimate. The proposed synthesis consequently organizes evidential transitions rather than defining validated thresholds. Prospective scaffold holdouts, independent laboratories, deliberate perturbation of nuisance structure, orthogonal mechanism tests, and experimentally adjudicated model disagreements are needed to determine where integrated phenotypic–chemical representations genuinely improve discovery decisions.

CONCLUSION

Phenotypic screening and machine learning intersect most productively when cellular response is treated as a structured source of evidence rather than a surrogate for mechanism. Images, transcriptional states, and other perturbational signatures can support representation learning, functional comparison, mechanism hypotheses, and compound prioritization, but each transition changes the evidential claim being made.

The proposed synthesis separates representation quality, mechanistic attribution, and decision utility while placing biological context, technical confounding, applicability domain, and external validation across the entire discovery process. Chemical and phenotypic information can converge without becoming equivalent, and useful prioritization need not await complete target resolution. The principal opportunity is therefore not automatic mechanism discovery, but disciplined use of learned cellular signatures to generate better and more falsifiable experimental choices.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

REFERENCES

[1] Moffat JG, Vincent F, Lee JA, Eder J, Prunotto M. Opportunities and challenges in phenotypic drug

discovery: An industry perspective. *Nat Rev Drug Discov.* 2017;16(8):531-43. doi:10.1038/nrd.2017.111

[2] Krentzel D, Shorte SL, Zimmer C. Deep learning in image-based phenotypic drug discovery. *Trends Cell Biol.* 2023;33(7):538-54. doi:10.1016/j.tcb.2022.11.011

[3] Seal S, Trapotsi MA, Spjuth O, Singh S, Carreras-Puigvert J, Greene N, et al. Cell Painting: A decade of discovery and innovation in cellular imaging. *Nat Methods.* 2025;22(2):254-68. doi:10.1038/s41592-024-02528-8

[4] Caicedo JC, Cooper S, Heigwer F, Warchal S, Qiu P, Molnar C, et al. Data-analysis strategies for image-based cell profiling. *Nat Methods.* 2017;14(9):849-63. doi:10.1038/nmeth.4397

[5] Chandrasekaran SN, Cimini BA, Goodale A, Miller L, Kost-Alimova M, Jamali N, et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nat Methods.* 2024;21(6):1114-21. doi:10.1038/s41592-024-02241-6

[6] Moshkov N, Bornholdt M, Benoit S, Smith M, McQuin C, Goodman A, et al. Learning representations for image-based profiling of perturbations. *Nat Commun.* 2024;15(1):1594. doi:10.1038/s41467-024-45999-1

[7] Tang Q, Ratnayake R, Seabra G, Jiang Z, Fang R, Cui L, et al. Morphological profiling for drug discovery in the era of deep learning. *Brief Bioinform.* 2024;25(4):bbae284. doi:10.1093/bib/bbae284

[8] Simm J, Klambauer G, Arany A, Steijaert M, Wegner JK, Gustin E, et al. Repurposing high-throughput image assays enables biological activity prediction for drug discovery. *Cell Chem Biol.* 2018;25(5):611-8.e3. doi:10.1016/j.chembiol.2018.01.015

[9] Hofmarcher M, Rumetshofer E, Hochreiter S, Clevert DA, Klambauer G. Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks. *J Chem Inf Model.* 2019;59(3):1163-71. doi:10.1021/acs.jcim.8b00670

[10] Fredin Haslum J, Lardeau CH, Karlsson J, Turkki R, Leuchowius KJ, Smith K, et al. Cell Painting-based bioactivity prediction boosts high-throughput screening hit-rates and compound diversity. *Nat Commun.* 2024;15(1):3470. doi:10.1038/s41467-024-47171-1

[11] Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A next generation Connectivity Map: L1000 platform and the first 1,000,000 profiles. *Cell.* 2017;171(6):1437-52.e17. doi:10.1016/j.cell.2017.10.049

- [12] Peidli S, Green TD, Shen C, Gross T, Min J, Garda S, et al. scPerturb: Harmonized single-cell perturbation data. *Nat Methods*. 2024;21(3):531-40. doi:10.1038/s41592-023-02144-y
- [13] Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. *Nat Methods*. 2019;16(8):715-21. doi:10.1038/s41592-019-0494-8
- [14] Lotfollahi M, Klimovskaia Susmelj A, De Donno C, Hetzel L, Ji Y, Ibarra IL, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Mol Syst Biol*. 2023;19(6):e11517. doi:10.15252/msb.202211517
- [15] Qi X, Zhao L, Tian C, Li Y, Chen ZL, Huo P, et al. Predicting transcriptional responses to novel chemical perturbations using deep generative model for drug discovery. *Nat Commun*. 2024;15(1):9256. doi:10.1038/s41467-024-53457-1
- [16] Tong X, Qu N, Kong X, Ni S, Zhou J, Wang K, et al. Deep representation learning of chemical-induced transcriptional profile for phenotype-based drug discovery. *Nat Commun*. 2024;15(1):5378. doi:10.1038/s41467-024-49620-3
- [17] Bunne C, Stark SG, Gut G, Sarabia del Castillo J, Levesque M, Lehmann KV, et al. Learning single-cell perturbation responses using neural optimal transport. *Nat Methods*. 2023;20(11):1759-68. doi:10.1038/s41592-023-01969-x
- [18] Chen H, King FJ, Zhou B, Wang Y, Canedy CJ, Hayashi J, et al. Drug target prediction through deep learning functional representation of gene signatures. *Nat Commun*. 2024;15(1):1853. doi:10.1038/s41467-024-46089-y
- [19] Rohban MH, Fuller AM, Tan C, Goldstein JT, Syangtan D, Gutnick A, et al. Virtual screening for small-molecule pathway regulators by image-profile matching. *Cell Syst*. 2022;13(9):724-36.e9. doi:10.1016/j.cels.2022.08.003
- [20] Trapotsi MA, Mervin LH, Afzal AM, Sturm N, Engkvist O, Barrett IP, et al. Comparison of chemical structure and cell morphology information for multitask bioactivity predictions. *J Chem Inf Model*. 2021;61(3):1444-56. doi:10.1021/acs.jcim.0c00864
- [21] Way GP, Natoli T, Adeboye A, Litichevskiy L, Yang A, Lu X, et al. Morphology and gene expression profiling provide complementary information for mapping cell state. *Cell Syst*. 2022;13(11):911-23.e9. doi:10.1016/j.cels.2022.10.001
- [22] Moshkov N, Becker T, Yang K, Horvath P, Dancik V, Wagner BK, et al. Predicting compound activity from phenotypic profiles and chemical structures. *Nat Commun*. 2023;14(1):1967. doi:10.1038/s41467-023-37570-1
- [23] Seal S, Yang H, Trapotsi MA, Singh S, Carreras-Puigvert J, Spjuth O, et al. Merging bioactivity predictions from cell morphology and chemical fingerprint models using similarity to training data. *J Cheminform*. 2023;15(1):56. doi:10.1186/s13321-023-00723-x
- [24] Sanchez-Fernandez A, Rumetshofer E, Hochreiter S, Klambauer G. CLOOME: Contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nat Commun*. 2023;14(1):7339. doi:10.1038/s41467-023-42328-w
- [25] Zheng S, Rao J, Zhang J, Zhou L, Xie J, Cohen E, et al. Cross-modal graph contrastive learning with cellular images. *Adv Sci (Weinh)*. 2024;11(32):e2404845. doi:10.1002/adv.202404845
- [26] Gendelev L, Taylor J, Myers-Turnbull D, Chen S, McCarroll MN, Arkin MR, et al. Deep phenotypic profiling of neuroactive drugs in larval zebrafish. *Nat Commun*. 2024;15(1):9955. doi:10.1038/s41467-024-54375-y
- [27] Celik S, Hütter JC, Carlos SM, Lazar NH, Mohan R, Tillinghast C, et al. Building, benchmarking, and exploring perturbative maps of transcriptional and morphological data. *PLoS Comput Biol*. 2024;20(10):e1012463. doi:10.1371/journal.pcbi.1012463
- [28] Arevalo J, Su E, Ewald JD, van Dijk R, Carpenter AE, Singh S. Evaluating batch correction methods for image-based cell profiling. *Nat Commun*. 2024;15(1):6516. doi:10.1038/s41467-024-50613-5
- [29] Palma A, Theis FJ, Lotfollahi M. Predicting cell morphological responses to perturbations using generative modeling. *Nat Commun*. 2025;16(1):505. doi:10.1038/s41467-024-55707-8