PLOS Computational Biology

EDUCATION

Ten quick tips for causal analysis of biomedical omics data

Gleb Svinin[1☯], Rebecca Ting Jiin Loo[1☯], Nikhilesh Vasantha Kumar[1☯], Varsha Venkatesha Murthy[1☯], Dilara Uzuner Odongo[1☯], Ramón Díaz-Uriarte[2,3], Ana Conesa[4], Gianluca Bontempi[5], Susana Vinga[6], Ilaria Granata[7], Daniel Domingo-Fernández[8], Paola Lecca[9], Marieke L. Kuijjer[10], Jesse H. Krijthe[11], Ina Koch[12], Laurence Calzone[13], Simona Ester Rombo[14], Fátima Sánchez-Cabo[15], Enrico Glaab[1*]

1 Biomedical Data Science Group, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-sur-Alzette, Luxembourg, **2** Instituto de Investigaciones Biomédicas Sols-Morreale (IIBM), UAM-CSIC, Madrid, Spain, **3** Department of Biochemistry, Universidad Autónoma de Madrid, Madrid, Spain, **4** Institute for Integrative Systems Biology, Spanish National Research Council, Valencia, Spain, **5** Machine Learning Group, Université Libre de Bruxelles, Bruxelles, Belgique and WEL Research Institute, Wavre, Belgique, **6** INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal, **7** Institute for High Performance Computing and Networking, National Research Council, Naples, Italy, **8** Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin, Germany, **9** Faculty of Engineering, Free University of Bozen-Bolzano, Bolzano, Italy, **10** Faculty of Medicine, University of Helsinki, Helsinki, Finland; Norwegian Centre for Molecular Biosciences and Medicine (NCMBM), University of Oslo, Oslo, Norway, **11** Pattern Recognition & Bioinformatics, Delft University of Technology, Delft, The Netherlands, **12** Institute of Computer Science, Goethe University Frankfurt, Frankfurt, Germany, **13** Institut Curie, Inserm U1331, PSL, Mines ParisTech, Paris, France, **14** Department of Mathematics and Computer Science, University of Palermo, Palermo, Italy, **15** Centro Nacional de Investigaciones Cardiovasculares (CNIC), Madrid, Spain

☯ These authors contributed equally to this work and share first authorship.
* enrico.glaab@uni.lu

Citation: Svinin G, Loo RTJ, Kumar NV, Murthy VV, Odongo DU, Díaz-Uriarte R, et al. (2026) Ten quick tips for causal analysis of biomedical omics data. PLoS Comput Biol 22(8): e1014668. https://doi.org/10.1371/journal.pcbi.1014668

Editor: Jana Wolf, Max Delbruck Centre for Molecular Medicine: Max-Delbruck-Centrum fur Molekulare Medizin in der Helmholtz-Gemeinschaft, GERMANY

Published: August 20, 2026

Funding: EG acknowledges funding support by the Luxembourg National Research Fund (FNR) for our lecture series as part of the RESCOM program and for the projects RECAST (INTER/22/17104370/RECAST), AsynIntact (C24/BM/18865990/AsynIntact), PreDYT (INTER/EJP RD22/17027921/PreDYT), AD-PLCG2 (INTER/JPND23/17999421/

Introduction

Omics data generation has become routine in biomedical research. Transcriptomics, proteomics, and metabolomics now produce datasets of considerable scale and resolution. Integrating these diverse modalities requires carefully addressing common analytical pitfalls [1]. A standard analysis involves statistical comparisons between conditions (e.g., disease versus control), providing ranked lists of differentially abundant molecules. While informative for biomarker discovery, these lists describe associations rather than mechanisms: they indicate what changes but not what causes those changes or how interventions might alter the system's behavior [2–4].

Causal inference methods aim to go beyond correlations by modeling directed regulatory influences, identifying candidate upstream drivers of disease phenotypes, and predicting the effects of molecular interventions. They range from counterfactual frameworks and graphical causal models to logic-based representations of regulatory circuits [5,6]. These frameworks address the following distinct tasks: causal discovery (learning structure from data), identification (expressing a causal quantity in terms of the observed data distribution under stated assumptions), estimation (fitting the relevant statistical relationships), and inference (quantifying uncertainty of estimated effects), so that the relevant references and tools differ by task. While some omics studies use controlled lab experiments, many rely on observational data from

ADPLCG2), and EPI_T-ALL (INTER/TRANSCAN23/18333087/EPI_T-ALL). FSC acknowledges the Spanish Agency of Research and FEDER Una manera de hacer Europa under grant PID2022-141527OB-I00. RDU acknowledges partial support by grants PID2024-156888OB-I00 and PID2019-111256RB-I00 funded by MCIN/AEI/10.13039/501100011033/FEDER, EU. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

clinical cohorts or existing biobanks. In these observational settings, researchers do not assign treatments or alter the biology themselves, meaning causal reasoning is necessary to clarify what the data can actually answer and to help manage the limits of the study design. The degree of experimental control, ranging from observational cohorts to designed perturbation screens, further determines how the tips below should be read. Applying these methods to biomedical data requires careful attention to study design, data quality, prior biological knowledge, and the assumptions inherent in each framework, as well as rigorous initial data analysis [7].

The ten quick tips presented here were compiled from the insights shared during an international lecture series on 'Causal Analysis of Biomedical Data' held at the University of Luxembourg in 2024/2025, given by researchers with complementary expertise spanning causal inference, network biology, Boolean modeling, and clinical translation. While no single set of suggestions can address every scenario, this guide provides a practical roadmap for moving from molecular associations to causal, mechanistic understanding (Fig 1). The tips are organized into three phases: causal discovery (Tips 1–3), causal model analysis (Tips 4–8), and validation and translation (Tips 9–10).

Tip 1: Define the causal question and assess assumptions before collecting data

Before collecting data, researchers should formulate the specific causal question the study is intended to address. Whether the aim is to estimate the effect of a molecular intervention, identify an upstream cause of a disease phenotype, or address a counterfactual question (see [6,10] for the distinction between interventions and counterfactuals), the question helps to determine which data and methods are appropriate. Prediction and causal inference should not be conflated: a gene expression signature that accurately predicts tumor recurrence may reflect downstream effects rather than causal drivers [4], and a model predicting cardiovascular risk does not identify the main risk factor for a specific individual [11]. Formally, causal questions concern the consequences of interventions rather than conditional associations [6,12].

To support the intended causal inference, researchers should make their assumptions about the causal structure explicit, including potential confounders. These assumptions can be represented in a directed acyclic graph (DAG), which helps assess whether the available data can support the intended causal claim and reveals potential sources of bias [2,12,13]. If relationships between variables are not carefully considered, paradoxical conclusions may arise, for example, when a treatment appears to increase risk because treated patients already had worse prognosis [5,14]. It is important to separate two kinds of assumptions: identification assumptions, which determine whether a causal estimand can be written as a function of the observed data distribution given a DAG, and statistical or estimation assumptions, which concern the models used to fit the resulting relationships, including assumptions about nuisance functions [8,15]. Three identification assumptions warrant particular attention: exchangeability (absence of unmeasured confounding), positivity (non-zero probability of each intervention

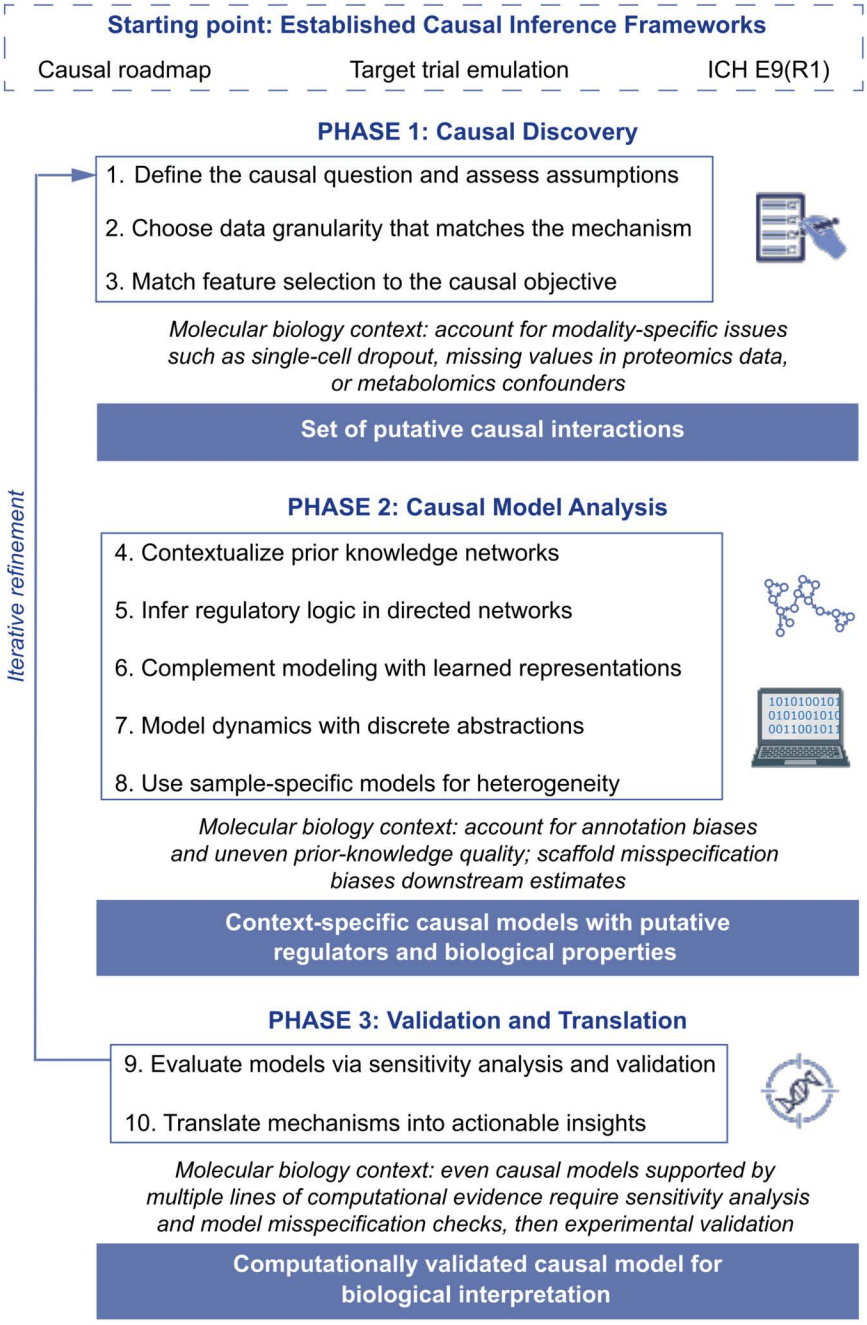

Fig 1. Overview of the ten tips organized by analysis phase. The tips are grouped into three phases: Phase 1 (Causal Discovery, Tips 1–3), Phase 2 (Causal Model Analysis, Tips 4–8), and Phase 3 (Validation and Translation, Tips 9–10). Solid arrows indicate the primary workflow; the arrow from validation back to Phase 1 reflects the iterative nature of causal analysis. The workflow takes the causal roadmap of statistical and causal inference [8] and target trial emulation [9], together with the ICH E9(R1) estimand addendum, as its starting point, and then marks the steps that require field-specific adaptation in molecular biomedicine, such as adapting the estimand to the molecular entities under study, choosing data granularity, contextualizing prior-knowledge networks, and validating predictions experimentally.

https://doi.org/10.1371/journal.pcbi.1014668.g001

level in all subgroups), and consistency (well-defined interventions). These three conditions are sufficient for estimating an average treatment effect under a point intervention but are not universal: other designs require additional or alternative assumptions. For example, Mendelian randomization in population genetics relies on instrumental-variable assumptions [16], whereas g-methods for longitudinal and survival settings with time-varying exposures require sequential exchangeability [12]. Identification assumptions differ in the extent to which they can be examined empirically. Exchangeability and consistency cannot be verified from the observed data, whereas positivity, that is, sufficient overlap between the intervention groups, can be diagnosed in practice, for example, by inspecting propensity score distributions [17]. When the aim is to estimate an effect rather than to recover structure, doubly robust and semiparametric efficient estimators, including augmented inverse-probability weighting, one-step estimators, and targeted maximum likelihood estimation, reduce sensitivity to misspecification of any single nuisance model [15,18,19]. In turn, defining the estimand before analysis helps guide subsequent analytical decisions [20]. This step is formalized by established frameworks, in particular target trial emulation [9] and the causal roadmap for generating real-world evidence [8], which builds on the statistical causal modeling framework [15] and has been taken up by health technology assessment and regulatory bodies, including the ICH E9(R1) estimand addendum [21]. These frameworks provide a structured starting point for molecular studies, with the field-specific challenges addressed in the following tips separately.

Tip 2: Choose data granularity that matches the mechanism

Causal signals can be obscured by inappropriate aggregation. Coarse gene-level quantification may overlook isoform-switching events with distinct functional consequences [22]. Inference of evolutionary accumulation models from bulk data may reconstruct composite genotypes that do not correspond to any real genotype, because mutations from different cell populations are merged [23]. More generally, bulk averaging can mask signals present only in minority cell populations.

Whether aggregation is a problem depends on the target estimand defined in Tip 1. If the question concerns a population-average effect, heterogeneous effects that average out across cells or subgroups are the intended quantity rather than an error; granularity matters when the question instead concerns cell-type-specific or subgroup-specific mechanisms that aggregation would obscure. Data granularity should therefore match the hypothesized causal mechanism. Long-read sequencing can reveal disease-associated isoform switches missed by standard short-read RNA-seq [24]. If the causal hypothesis involves cell-type-specific regulation, single-cell or spatial transcriptomics may be required; some methods, such as TENET [25], infer causal relationships directly from single-cell measurements. Others, including the methods discussed in Tip 5, can take pseudobulk-derived fold changes as input, which improves robustness in settings where single-cell resolution is not needed. If protein function is central, measuring post-translational modifications may be more informative than aggregate gene expression. The greater cost and complexity of higher-resolution assays must be balanced against the risk of losing causal signals through aggregation. Practical challenges and plausibility of assumptions also differ across omics modalities, even when the underlying causal framework is unchanged: single-cell transcriptomics is affected by dropout and sparsity, proteomics often presents values missing not at random and typically achieves lower proteome coverage, and metabolomics is sensitive to environmental and dietary confounders. These differences mainly affect preprocessing choices and which confounders must be considered, and should be weighed alongside the granularity considerations above.

Tip 3: Match feature selection to the causal objective

High-dimensional, multiscale biomedical data involve statistical associations that do not reflect causal relationships. For mechanistic understanding and intervention-oriented analysis, feature selection should distinguish causes from effects and indirect associations [26]. A differentially expressed gene at the bottom of a regulatory cascade is less informative for intervention design than an upstream regulator with a smaller effect size.

Causal feature selection methods address this challenge using strategies distinct from standard correlation-based filtering. Information-theoretic approaches use asymmetric descriptors of inter-variable relationships, or exploit auxiliary variables to clarify the relationship among variables of interest. One such method [26] prioritizes putative direct causes of a target variable by filtering out features whose association with the outcome is fully explained by other features. Subsequent extensions [27] handle multivariate outcomes, while further work [28] infers pairwise causal relationships using asymmetric descriptors, moving closer to network reconstruction. Complementary to this, structured sparsity and network-based penalty terms in regularized regression can incorporate prior knowledge of regulatory relationships, highlighting upstream features within signaling pathways [29,30]. These causal feature selection methods rely on assumptions such as faithfulness and valid conditional-independence testing, but these can be difficult to justify in high-dimensional omics data. Faithfulness can fail, for example, when effects along distinct causal paths cancel, making causally connected variables appear statistically independent and causing true edges to be missed. Under model misspecification or violated assumptions, both false positives and false negatives can arise, and explicit error-rate control is still an open problem (a benchmark survey covers available methods and their failure modes [31]). Such outputs are therefore best treated as hypotheses for downstream validation rather than definitive causal claims.

Tip 4: Contextualize prior knowledge networks

Purely data-driven approaches for learning causal network structure are often underpowered in biomedical settings, where sample sizes are small relative to the number of variables [32]. Prior Knowledge Networks (PKNs) from curated databases provide a biologically informed starting point, and different PKN resources offer complementary strengths: SIGNOR [33] provides directed, signed causal relationships between bioentities, Reactome [34] organizes molecular interactions into pathway maps, and OmniPath [35] integrates over 100 resources into a single compilation. However, these databases contain interactions observed across many conditions, producing dense, non-context-specific networks.

A practical strategy for addressing this issue is to contextualize PKNs using experimental data [36,38]. Nodes not expressed in the tissue or condition of interest are removed along with their associated edges, while recognizing that absence of detection does not always imply biological irrelevance. Tools such as PathMe [36] and NeKo [38] enable integration of mechanistic pathway knowledge from multiple resources, automating the construction of context-specific subnetworks. A complementary strategy is penalized regression, which uses graph-aware penalties derived from PKNs to identify condition-associated subnetworks by encouraging similar regression coefficients across connected genes [29,30]. However, careful attention should be paid to quality, coverage, and potential biases of the underlying databases, as missing or incorrect interactions may propagate into the model. More broadly, when a contextualized network is used as the scaffold for causal or regulatory inference, misspecification of this scaffold relative to the unknown ground truth biases downstream estimates, which is one reason the robustness checks in Tip 9 are needed.

Tip 5: Infer regulatory logic in directed networks

In correlation networks, edges indicate that two variables are associated but not which one influences the other. Because no direction of effect is encoded, such networks cannot propagate signals from putative causes to downstream effects. Directed interaction networks, in which edges are oriented from source to target, overcome this limitation. As discussed in Tip 4, a variety of directed prior-knowledge networks are available, providing a basis for tracing regulatory effects through the system.

These directed networks can be combined with experimental data to infer regulatory activity and formulate signaling hypotheses. VIPER [39], for example, estimates regulator activity by aggregating expression changes across a regulon, i.e., the set of a regulator's known target genes. CARNIVAL [40] identifies consistent signaling paths within a signed, directed PKN by integrating perturbation-response footprints. By combining these data-driven approaches with curated databases for transcription factor–target relationships, such as DoRothEA [41], researchers can maximize the coverage

of directed, signed edges in their networks. The resulting models make explicit predictions about which regulators are active or inactive under a given condition, directly informing causal hypothesis generation. The reliability of predictions depends on the quality and completeness of the underlying regulons and prior knowledge networks, and footprint-based activity estimates assume that the measured downstream changes reflect the activity of the proposed regulator. Where these assumptions are only partially met, independent benchmarking and experimental validation are important before the inferred activities are interpreted causally.

Tip 6: Complement causal modeling with learned representations

Biological interaction networks are structured objects whose topology encodes functional organization [42]. While embedding such networks into metric spaces is not a causal inference technique per se, it generates topology-aware representations that support downstream analyses informing causal modeling. At the node level, mapping gene networks onto a metric space can uncover latent structural properties. Evaluations have shown that hyperbolic space is particularly well suited for biological networks, as it preserves distances of the original graph more faithfully than Euclidean alternatives [43]. In these embeddings, standard clustering techniques can identify centers of information spreading, providing a mathematically grounded approach to nominating candidate drivers of network dynamics.

At the whole-graph level, frameworks like Netpro2vec [37] embed each graph within a collection as a fixed-length vector by summarizing topological features through node distance distributions and transition matrices. These graph-level representations enable direct comparison, clustering, and visualization of multiple networks, for instance to contrast disease and control networks or to identify network subtypes across a patient cohort. Combining node- and graph-level embeddings helps distill complex topological information into features that can guide and refine mechanistic and causal models. When embedding-derived features are used as covariates in downstream causal analyses, for example, in the exposure-assignment model or the outcome model, errors in these representations can propagate into effect estimates. The doubly robust estimators described in Tips 1 and 9 can reduce sensitivity to misspecification of one of these nuisance models, although they do not improve the embedding itself. A related and rapidly developing direction is causal representation learning, which aims to recover latent causal variables and their relations from high-dimensional measurements rather than from predefined features [44]. Steps in this direction for omics include methods that identify regulatory edges which change between conditions such as disease and control states [45], though a general implementation of causal representation learning for high-dimensional omics remains an active area of development. These methods address a distinct but related challenge: rather than representing known topology in metric spaces, they aim to recover latent causal structure directly from high-dimensional data, with their assumptions and identifiability conditions still being actively clarified.

Tip 7: Model system dynamics using discrete abstractions

Frameworks for modeling biological dynamics can be broadly divided into kinetic and non-kinetic approaches. Kinetic models provide quantitative predictions but require extensive parametrization, with reaction rate constants often unavailable at the quality needed for reliable inference in larger systems [46]. Non-kinetic approaches offer an alternative by capturing qualitative system behavior without detailed kinetic parameters.

One non-kinetic formalism is Boolean modeling, in which each biological entity is represented by a binary state (active or inactive) and its state at the next step is determined by logical rules encoding regulatory dependencies. These models can recapitulate stable states (attractors) of signaling pathways and predict qualitative effects of perturbations. The field includes many variants, including frameworks such as MaBoSS [47,48] that use stochastic continuous-time simulations to produce probabilistic predictions of attractor state distributions, without requiring explicit kinetic rate parameters. The CoLoMoTo consortium provides a unified environment for Boolean network analysis [49]. Another discrete formalism is Petri nets, which represent networks as bipartite graphs to distinguish between passive components (places, e.g.,

molecules) and active components (transitions, e.g., reactions), and describe system dynamics using movable tokens [50]. For Petri net modeling and analysis, the software MonaLisa is available [51]. Overall, these discrete approaches are particularly useful for studying knockout or overexpression effects, assessing state reachability, and identifying steady states corresponding to known disease phenotypes.

Tip 8: Account for biological heterogeneity via sample-specific models

A single consensus network inferred from a disease cohort averages over patients and can mask differences in underlying mechanisms. However, causal drivers may differ across individuals due to genetic background, environmental exposures, disease stage, and comorbidities. This matters in therapeutic settings, where a drug target may be relevant only for a subset of patients [52].

Methods for constructing sample-specific networks address this limitation. LIONESS [53,54] derives a network for each sample by computing the change in edge weights between the aggregate network inferred from all samples and the network obtained when that sample is removed. By contrast, PatientProfiler [55] generates sample-specific networks by integrating prior-knowledge interactions with individual sample data. The resulting single-sample networks support topology-aware downstream analyses: patients can be grouped based on similarity between their inferred networks and the resulting groups can then be tested for differences in clinical outcomes such as survival [55]. Collections of individual weighted networks can also be used to identify condition-specific network patterns [56]. More recent work has applied large-scale causal discovery using interventional data to shed light on gene network structure [57]. When full patient-specific modeling is not feasible, stratifying the cohort by known clinical variables before network construction can still help capture mechanistic heterogeneity.

Tip 9: Evaluate causal models through sensitivity analysis and biological validation

Causal analysis depends on assumptions and prior knowledge. Its conclusions should therefore be subjected to sensitivity analyses and checked for robustness under reasonable alternatives, such as different confounder sets, hyperparameters, or prior-knowledge inputs, to assess whether the findings depend strongly on model configuration and assumptions. When possible, diagnostic checks should be used to assess whether the model was applied appropriately; such checks can, for example, reveal systematic biases in estimated treatment effects. In computational settings, credibility is further strengthened when similar conclusions can be reproduced in an independent dataset, ideally from a different modality. Robustness checks should also probe model misspecification directly, for example, by comparing estimators that rely on different working models, or, where suitable, by using doubly robust estimators whose validity holds if at least one of the nuisance models is correctly specified [2,12,15].

Beyond computational checks, causal hypotheses should be assessed against existing biological knowledge and, ultimately, tested experimentally. For gene regulatory networks, CRISPR-based perturbation screens can test whether predicted regulators induce expected changes in their downstream targets. Discrepancies between predicted and observed effects highlight areas where the model, its assumptions, or its inputs need revision, making the process explicitly iterative (Fig 1). Databases such as the Connectivity Map [58] provide perturbation response profiles for systematic *in silico* validation before generating new predictions. Taken together, sensitivity analysis, biological assessment, and perturbation experiments provide stronger support for causal claims.

Tip 10: Translate causal mechanisms into actionable insights

The goal of causal analysis in biomedicine is to produce findings that inform clinical decisions, guide therapeutic development, or advance biological understanding. There are encouraging examples of how the methodological toolkit discussed in this article has been applied in practice. For instance, an approach combining the ideas of tips 4 and 5 was used to study FLT3-ITD-positive acute myeloid leukemia, where subtype-specific roles of the kinase WEE1 were identified and

subsequent pharmacological inhibition restored therapy sensitivity [59]. In another application, a precision oncology framework successfully prioritized compounds by their ability to invert tumor-checkpoint activity in gastroenteropancreatic neuroendocrine tumors [60].

Looking ahead, *in silico* clinical trials [61] offer a further perspective on translation, though computational frameworks remain constrained by the knowledge and data used to build them. Causal models can provide an advantage for clinical adoption because their structure encodes mechanistic hypotheses that can be communicated as interpretable explanations. However, researchers should present results as prioritized intervention points with predicted mechanistic effects and acknowledge the model's assumptions and limitations. The above examples illustrate both the translational potential of causal analyses and the need to remain realistic about what such methods can currently deliver.

Conclusion

Moving from correlation to causation in molecular biology is not only a statistical exercise; it represents a fundamental shift in how biological understanding and therapeutic development are approached. Our ten quick tips provide a workflow for extracting mechanistic insights from high-dimensional molecular data, emphasizing that successful causal analysis requires attention at every stage: formulating causal questions, selecting data at an appropriate granularity, contextualizing prior biological knowledge, accounting for network structure and dynamics, and translating inferred mechanisms into testable hypotheses.

No single approach is universally optimal; the choice of strategy must be guided by the causal question, data quality, and intended application. The iterative nature of causal analysis, where validation informs model refinement and new experiments, should be viewed as a strength rather than a limitation. These quick tips provide a starting point for researchers entering this field and a reference for those already engaged in causal analysis of biomedical data. The applicability of individual tips also depends on the study design: in purely observational cohorts causal reasoning mainly clarifies what can be inferred, whereas designed perturbation experiments can support stronger causal claims.

Acknowledgments

The authors wish to thank all participants and attendees of the lecture series on Causal Analysis of Biomedical Data organized at the Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg.

Author contributions

Conceptualization: Enrico Glaab.

Funding acquisition: Enrico Glaab.

Methodology: Enrico Glaab.

Project administration: Enrico Glaab.

Supervision: Enrico Glaab.

Validation: Enrico Glaab.

Visualization: Enrico Glaab.

Writing – original draft: Gleb Svinin, Rebecca Ting Jiin Loo, Nikhilesh Vasantha Kumar, Varsha Venkatesha Murthy, Dilara Uzuner Odongo, Enrico Glaab.

Writing – review & editing: Gleb Svinin, Rebecca Ting Jiin Loo, Nikhilesh Vasantha Kumar, Varsha Venkatesha Murthy, Dilara Uzuner Odongo, Ramón Díaz-Uriarte, Ana Conesa, Gianluca Bontempi, Susana Vinga, Ilaria Granata, Daniel Domingo-Fernández, Paola Lecca, Marieke L. Kuijjer, Jesse H. Krijthe, Ina Koch, Laurence Calzone, Simona Ester Rombo, Fátima Sánchez-Cab, Enrico Glaab.

References

1. Chicco D, Cumbo F, Angione C. Ten quick tips for avoiding pitfalls in multi-omics data integration analyses. PLoS Comput Biol. 2023;19(7):e1011224. https://doi.org/10.1371/journal.pcbi.1011224 PMID: 37410704
2. Morgan SL, Winship C. Counterfactuals and causal inference: methods and principles for social research. 2nd ed. New York, NY: Cambridge University Press; 2015.
3. Rosenbaum PR. Observation and experiment: an introduction to causal inference. Cambridge, Massachusetts: Harvard University Press; 2017.
4. Pearl J, Mackenzie D. The book of why: the new science of cause and effect. First trade paperback edition ed. New York: Basic Books; 2018.
5. Peters J, Janzing D, Schölkopf B. Elements of causal inference: foundations and learning algorithms. Cambridge, Massachusetts: The MIT Press; 2017.
6. Pearl J. Causality: models, reasoning, and inference. Cambridge, U.K.: Cambridge University Press; 2009.
7. Baillie M, Le Cessie S, Schmidt CO, Lusa L, Huebner M. Ten simple rules for initial data analysis. PLoS Comput Biol. 2022;18:e1009819. https://doi.org/10.1371/journal.pcbi.1009819
8. Dang LE, Gruber S, Lee H, Dahabreh IJ, Stuart EA, Williamson BD, et al. A causal roadmap for generating high-quality real-world evidence. J Clin Transl Sci. 2023;7(1):e212. https://doi.org/10.1017/cts.2023.635 PMID: 37900353
9. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol. 2016;183(8):758–64. https://doi.org/10.1093/aje/kwv254 PMID: 26994063
10. Pearl J. The seven tools of causal inference, with reflections on machine learning. Commun ACM. 2019;62(3):54–60. https://doi.org/10.1145/3241036
11. van Geloven N, Keogh RH, van Amsterdam W, Cinà G, Krijthe JH, Peek N, et al. The risks of risk assessment: causal blind spots when using prediction models for treatment decisions. Ann Intern Med. 2025;178(9):1326–33. https://doi.org/10.7326/ANNALS-24-00279 PMID: 40720833
12. Hernán MA, Robins JM. Causal inference: what if. CRC Press; 2020.
13. Feuerriegel S, Frauen D, Melnychuk V, Schweisthal J, Hess K, Curth A, et al. Causal machine learning for predicting treatment outcomes. Nat Med. 2024;30(4):958–68. https://doi.org/10.1038/s41591-024-02902-1 PMID: 38641741
14. Arah OA. The role of causal reasoning in understanding Simpson's paradox, Lord's paradox, and the suppression effect: covariate selection in the analysis of observational studies. Emerg Themes Epidemiol. 2008;5:5. https://doi.org/10.1186/1742-7622-5-5 PMID: 18302750
15. Petersen ML, van der Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. Epidemiology. 2014;25(3):418–26. https://doi.org/10.1097/EDE.0000000000000078 PMID: 24713881
16. Sanderson E, Glymour MM, Holmes MV, Kang H, Morrison J, Munafò MR, et al. Mendelian randomization. Nat Rev Methods Primers. 2022;2:6. https://doi.org/10.1038/s43586-021-00092-5 PMID: 37325194
17. Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. Stat Methods Med Res. 2012;21(1):31–54. https://doi.org/10.1177/0962280210386207 PMID: 21030422
18. Hines O, Dukes O, Diaz-Ordaz K, Vansteelandt S. Demystifying statistical learning based on efficient influence functions. Am Stat. 2022;76:292–304. doi:10.1080/00031305.2021.2021984
19. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc. 1994;89(427):846–66. https://doi.org/10.1080/01621459.1994.10476818
20. Luijken K, Morzywołek P, van Amsterdam W, Cinà G, Hoogland J, Keogh R, et al. Risk-based decision making: estimands for sequential prediction under interventions. Biom J. 2024;66(8):e70011. https://doi.org/10.1002/bimj.70011 PMID: 39607308
21. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. International Council for Harmonisation (ICH); 2019. Report No.: EMA/CHMP/ICH/436221/2017.
22. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13. https://doi.org/10.1186/s13059-016-0881-8 PMID: 26813401
23. Diaz-Uriarte R, Johnston IG. A picture guide to cancer progression and evolutionary accumulation models: systematic critique, plausible interpretations, and alternative uses. IEEE Access. 2025;13:62306–40. https://doi.org/10.1109/access.2025.3558392
24. Glinos DA, Garborcauskas G, Hoffman P, Ehsan N, Jiang L, Gokden A, et al. Transcriptome variation in human tissues revealed by long-read sequencing. Nature. 2022;608(7922):353–9. https://doi.org/10.1038/s41586-022-05035-y PMID: 35922509
25. Kim J, T Jakobsen S, Natarajan KN, Won K-J. TENET: gene network reconstruction using transfer entropy reveals key regulatory factors from single cell transcriptomic data. Nucleic Acids Res. 2021;49(1):e1. https://doi.org/10.1093/nar/gkaa1014 PMID: 33170214
26. Bontempi G, Meyer PE. Causal filter selection in microarray data. Proceedings of the 27th International Conference on Machine Learning (ICML). 2010. p. 95–102.
27. Bontempi G, Haibe-Kains B, Desmedt C, Sotiriou C, Quackenbush J. Multiple-input multiple-output causal strategies for gene selection. BMC Bioinformatics. 2011;12:458. https://doi.org/10.1186/1471-2105-12-458 PMID: 22118187
28. Bontempi G, Flauder M. From dependency to causality: a machine learning approach. J Mach Learn Res. 2015;16:2437–57.

29. Vinga S. Structured sparsity regularization for analyzing high-dimensional omics data. Brief Bioinform. 2021;22(1):77–87. https://doi.org/10.1093/bib/bbaa122 PMID: 32597465
30. Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics. 2008;24(9):1175–82. https://doi.org/10.1093/bioinformatics/btn081 PMID: 18310618
31. Yu K, Guo X, Liu L, Li J, Wang H, Ling Z. Causality-based feature selection: methods and evaluations. ACM Comput Surv. 2020;53:1–36. https://doi.org/10.1145/3409382
32. Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. Front Genet. 2019;10:524. https://doi.org/10.3389/fgene.2019.00524
33. Lo Surdo P, Iannuccelli M, Contino S, Castagnoli L, Licata L, Cesareni G, et al. SIGNOR 3.0, the SIGnaling network open resource 3.0: 2022 update. Nucleic Acids Res. 2023;51(D1):D631–7. https://doi.org/10.1093/nar/gkac883 PMID: 36243968
34. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A. The reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022;50:D687–92. https://doi.org/10.1093/nar/gkab1028
35. Türei D, Valdeolivas A, Gul L, Palacio-Escat N, Klein M, Ivanova O, et al. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. Mol Syst Biol. 2021;17(3):e9923. https://doi.org/10.15252/msb.20209923 PMID: 33749993
36. Domingo-Fernández D, Mubeen S, Marín-Llaó J, Hoyt CT, Hofmann-Apitius M. PathMe: merging and exploring mechanistic pathway knowledge. BMC Bioinformatics. 2019;20(1):243. https://doi.org/10.1186/s12859-019-2863-9 PMID: 31092193
37. Manipur I, Manzo M, Granata I, Giordano M, Maddalena L, Guarracino MR. Netpro2vec: a graph embedding framework for biomedical applications. IEEE/ACM Trans Comput Biol Bioinform. 2022;19(2):729–40. https://doi.org/10.1109/TCBB.2021.3078089 PMID: 33961560
38. Ruscone M, Tsirvouli E, Checcoli A, Turei D, Barillot E, Saez-Rodriguez J, et al. NeKo: A tool for automatic network construction from prior knowledge. PLoS Comput Biol. 2025;21(9):e1013300. https://doi.org/10.1371/journal.pcbi.1013300 PMID: 40956863
39. Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat Genet. 2016;48(8):838–47. https://doi.org/10.1038/ng.3593 PMID: 27322546
40. Liu A, Trairatphisan P, Gjerga E, Didangelos A, Barratt J, Saez-Rodriguez J. From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. NPJ Syst Biol Appl. 2019;5:40. https://doi.org/10.1038/s41540-019-0118-z PMID: 31728204
41. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 2019;29(8):1363–75. https://doi.org/10.1101/gr.240663.118 PMID: 31340985
42. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5(2):101–13. https://doi.org/10.1038/nrg1272 PMID: 14735121
43. Lecca P, Lombardi G, Latorre RV, Sorio C. How the latent geometry of a biological network provides information on its dynamics: the case of the gene network of chronic myeloid leukaemia. Front Cell Dev Biol. 2023;11:1235116. https://doi.org/10.3389/fcell.2023.1235116 PMID: 38078013
44. Scholkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, et al. Toward causal representation learning. Proc IEEE. 2021;109(5):612–34. https://doi.org/10.1109/jproc.2021.3058954
45. Belyaeva A, Squires C, Uhler C. DCI: learning causal differences between gene regulatory networks. Bioinformatics. 2021;37(18):3067–9. https://doi.org/10.1093/bioinformatics/btab167 PMID: 33704425
46. Calzone L, Barillot E, Zinovyev A. Logical versus kinetic modeling of biological networks: applications in cancer research. Curr Opin Chem Eng. 2018;21:22–31. https://doi.org/10.1016/j.coche.2018.02.005
47. Stoll G, Caron B, Viara E, Dugourd A, Zinovyev A, Naldi A, et al. MaBoSS 2.0: an environment for stochastic Boolean modeling. Bioinformatics. 2017;33(14):2226–8. https://doi.org/10.1093/bioinformatics/btx123 PMID: 28881959
48. Stoll G, Viara E, Barillot E, Calzone L. Continuous time Boolean modeling for biological signaling: application of Gillespie algorithm. BMC Syst Biol. 2012;6:116. https://doi.org/10.1186/1752-0509-6-116 PMID: 22932419
49. Naldi A, Hernandez C, Levy N, Stoll G, Monteiro PT, Chaouiya C, et al. The CoLoMoTo interactive notebook: accessible and reproducible computational analyses for qualitative biological networks. Front Physiol. 2018;9:680. https://doi.org/10.3389/fphys.2018.00680 PMID: 29971009
50. Koch I, Büttner B. Computational modeling of signal transduction networks without kinetic parameters: petri net approaches. Am J Physiol Cell Physiol. 2023;324(5):C1126–40. https://doi.org/10.1152/ajpcell.00487.2022 PMID: 36878844
51. Einloft J, Ackermann J, Nöthen J, Koch I. MonaLisa--visualization and analysis of functional modules in biochemical networks. Bioinformatics. 2013;29(11):1469–70. https://doi.org/10.1093/bioinformatics/btt165 PMID: 23564846
52. Califano A, Alvarez MJ. The recurrent architecture of tumour initiation, progression and drug sensitivity. Nat Rev Cancer. 2017;17(2):116–30. https://doi.org/10.1038/nrc.2016.124 PMID: 27977008
53. Kuijjer ML, Tung MG, Yuan G, Quackenbush J, Glass K. Estimating sample-specific regulatory networks. iScience. 2019;14:226–40. https://doi.org/10.1016/j.isci.2019.03.021 PMID: 30981959
54. Kuijjer ML, Hsieh P-H, Quackenbush J, Glass K. lionessR: single sample network inference in R. BMC Cancer. 2019;19(1):1003. https://doi.org/10.1186/s12885-019-6235-7 PMID: 31653243
55. Lombardi V, Di Rocco L, Meo E, Venafra V, Di Nisio E, Perticaroli V, et al. PatientProfiler: building patient-specific signaling models from proteogenomic data. Mol Syst Biol. 2025;21(12):1845–65. https://doi.org/10.1038/s44320-025-00160-y PMID: 41073799

56. Fassetti F, Rombo SE, Serrao C. Discriminative pattern discovery for the characterization of different network populations. Bioinformatics. 2023;39(4):btad168. https://doi.org/10.1093/bioinformatics/btad168 PMID: 37021928

57. Brown BC, Tokolyi A, Morris JA, Lappalainen T, Knowles DA. Large-scale causal discovery using interventional data sheds light on gene network structure in k562 cells. Nat Commun. 2025;16(1):9628. https://doi.org/10.1038/s41467-025-64353-7 PMID: 41173850

58. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell. 2017;171:1437-1452.e17. https://doi.org/10.1016/j.cell.2017.10.049

59. Massacci G, Venafra V, Latini S, Bica V, Pugliese GM, Graziosi S, et al. A key role of the WEE1-CDK1 axis in mediating TKI-therapy resistance in FLT3-ITD positive acute myeloid leukemia patients. Leukemia. 2023;37(2):288–97. https://doi.org/10.1038/s41375-022-01785-w PMID: 36509894

60. Alvarez MJ, Subramaniam PS, Tang LH, Grunn A, Aburi M, Rieckhof G, et al. A precision oncology approach to the pharmacological targeting of mechanistic dependencies in neuroendocrine tumors. Nat Genet. 2018;50(7):979–89. https://doi.org/10.1038/s41588-018-0138-4 PMID: 29915428

61. Pappalardo F, Russo G, Tshinanu FM, Viceconti M. In silico clinical trials: concepts and early adoptions. Brief Bioinform. 2019;20(5):1699–708. https://doi.org/10.1093/bib/bby043 PMID: 29868882