Causal Disease Intelligence: Architecture and Ethical Framework of the HqA Disease Causation Engine and MNTU Molecular Node Tracer for Computational Health Research

Adel Tammam

PTH Meridian — Precision Technology Heuristics, Bioinformatics Division

Calgary, Alberta, Canada

adel.tammam@pth-meridian.io

Abstract—Disease causation is among the most computationally underserved problems in modern medicine. Despite the availability of vast genomic, clinical, molecular, and epidemiological datasets, the computational infrastructure required to systematically identify the causal origins of complex and rare diseases remains fragmented, inaccessible, and poorly integrated. This paper presents the proposed architecture and ethical framework of two complementary computational health tools developed by PTH Meridian: HqA (Hard Questions Answered), a multi-modal disease causation engine integrating causal inference, genomic variant analysis, phenotype ontology matching, and molecular network propagation; and MNTU (Molecular Node Tracer Unified), a biological network analysis platform for tracing disease origination and propagation across protein interaction, metabolic pathway, and gene regulatory networks. The paper defines the computational problem space across six domains: rare disease diagnostic delay, antimicrobial resistance surveillance, health data privacy, Indigenous health equity, the energy-health nexus, and pandemic preparedness. A privacy-preserving architecture integrating federated learning with zero-knowledge proofs is described, enabling multi-institutional disease modeling without patient data sharing. The ethical framework addresses open-source auditability, PIPEDA compliance, limitation disclosure, and the responsibility of computational health tools to serve equity as well as efficiency.

Index Terms—bioinformatics, disease causation mapping, causal inference, genomic variant analysis, molecular network analysis, protein interaction networks, rare disease, antimicrobial resistance, federated learning, zero-knowledge proofs, human phenotype ontology, health equity, PIPEDA, open-source health informatics.

I. Introduction

Modern medicine produces data at unprecedented scale. Whole-genome sequencing costs have fallen from approximately $100 million per genome in 2001 to under $1,000 in 2024 [1]. Electronic health record systems capture longitudinal clinical observations across millions of patients. Protein interaction databases contain hundreds of thousands of experimentally verified interactions [2]. Metabolic and gene regulatory pathway maps span thousands of reactions across the known biology of multiple organisms [3]. And yet: a patient with a rare disease waits an average of four to eight years and consults multiple specialists before receiving a correct diagnosis [8][9][10]. Drug-resistant pathogens spread faster than surveillance tracks them [15]. The signal is present. The computational infrastructure to read it is not.

The central problem is not data scarcity but causal integration. Individual datasets — genomic variants, phenotype records, protein networks, environmental measurements — are each insufficient individually to establish disease causation. The computational challenge is integration: connecting across data modalities, identifying causal rather than correlational relationships, and doing so in a framework that is auditable, privacy-preserving, and accessible to the clinical and research institutions that would use the results.

This paper presents the proposed architecture and ethical framework of two computational tools addressing this challenge. HqA (Hard Questions Answered) is a disease causation engine integrating causal inference methodology with multi-modal biological data sources. MNTU (Molecular Node Tracer Unified) is a companion molecular network analysis platform for tracing disease mechanisms at the biological network level. Both systems are in active development and published as open-source at github.com/PTHMeridian. This paper documents the design rationale, underlying methodology, and ethical framework — not clinical validation results, which are the subject of planned future work

pending institutional collaboration.

II. Background and Foundational Methods

A. Causal Inference in Biological Systems

Statistical association between variables does not establish causation. This distinction, foundational to epidemiology and clinical research, was formalized by Pearl in his structural causal model framework [4], which provides mathematical machinery for distinguishing correlational from causal relationships using directed acyclic graphs (DAGs), do-calculus interventional reasoning, and counterfactual analysis. Hernán and Robins [5] extend this framework to epidemiological applications, providing the methodology for causal effect estimation from observational health data. Schölkopf et al. [6] articulate the case for causal representation learning in machine intelligence contexts, demonstrating that models trained on causal rather than purely statistical structure generalize more robustly to distribution shifts — a critical property for clinical deployment where training and deployment populations differ.

B. Molecular Network Biology

The network biology framework, established by Barabási and Oltvai [7] in 2004, treats the cell as a system of interacting components whose collective behavior determines biological function and dysfunction. Menche et al. [11] demonstrated that disease genes cluster in the human interactome, forming localized disease modules whose proximity correlates with clinical comorbidity — providing the theoretical foundation for MNTU's network propagation approach.

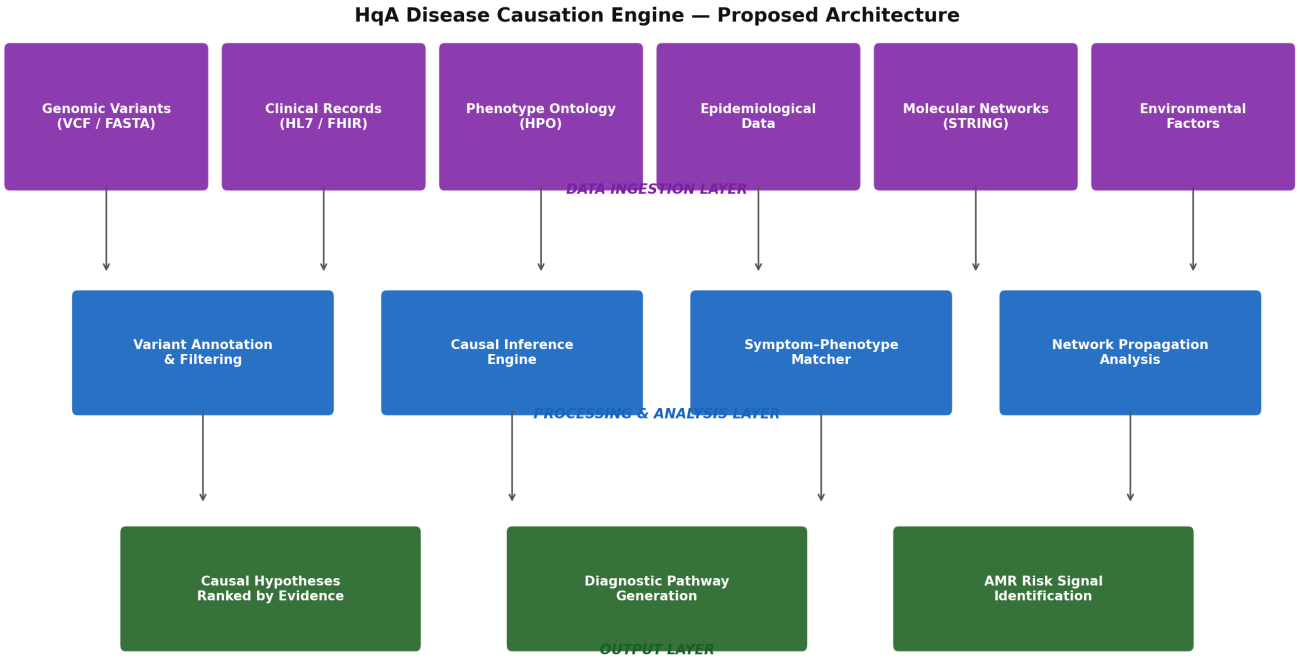
C. Phenotype Ontology and Symptom Matching

The Human Phenotype Ontology (HPO) [12], providing over 13,000 phenotypic abnormality terms with explicit semantic relationships, enables computational comparison of patient phenotype profiles against known disease phenotype signatures. Köhler et al. [13] demonstrated that HPO-based semantic similarity scoring significantly outperforms unstructured symptom search in rare disease diagnosis support — the approach adopted in HqA's phenotype matching module.

III. HqA: Disease Causation Engine

A. System Overview

HqA is designed as a multi-modal causation mapping platform that ingests heterogeneous biological and clinical data, applies causal inference and network analysis methods, and generates ranked causal hypotheses for disease origin in individual patients or patient cohorts. Fig. 1 illustrates the proposed architecture across three processing layers: data ingestion, analysis, and output.



Note: Proposed architecture. Data flows from multi-modal ingestion through causal inference, phenotype matching, and network propagation to ranked diagnostic outputs.

Fig. 1. Proposed HqA architecture spanning data ingestion, processing, and output layers. Multi-modal inputs flow through variant annotation, causal inference, phenotype matching, and network propagation analysis to generate ranked causal hypotheses and diagnostic pathways. All modules correspond to established bioinformatics data sources and methods (Sections III–IV).

B. Genomic Variant Analysis

Variant analysis in HqA is built around ClinVar [14] — the NCBI repository of genomic variant–disease associations — and OMIM for monogenic disease gene associations. Variants from patient genomic data are annotated, filtered for clinical significance, and mapped to disease associations using ClinVar pathogenicity classifications. Candidate variants are assessed for network-level impact using the STRING PPI database [2], identifying perturbations in the interaction neighborhoods of affected gene products.

C. Rare Disease Application and the Diagnostic Odyssey

Published studies document average diagnostic delays ranging from 4.8 years in U.S. populations [8] to 5.6 years in European cohorts [9], with patients consulting multiple specialists before diagnosis is established. The EURORDIS survey [10] of 5,980 patients across 25 countries documented that 25% waited more than 10 years. Fig. 2 illustrates approximate diagnostic burden by category. Across approximately 7,000 recognized rare diseases [8], the majority lack computational diagnostic support. HqA addresses this through HPO-driven differential generation: phenotype-scored candidate diagnoses ranked alongside supporting genomic and network evidence for clinician review.

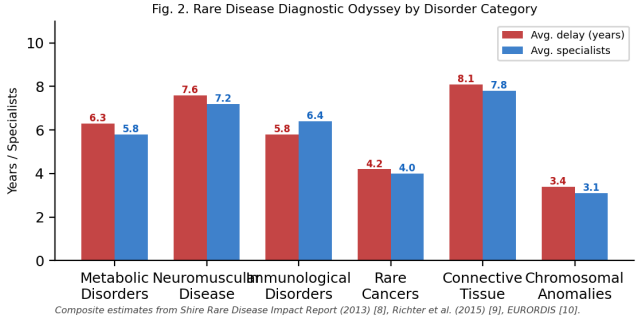


Fig. 2. Approximate rare disease diagnostic delay and specialist burden by disorder category. Composite estimates from published literature [8][9][10].

D. Clinical Decision Support Positioning

HqA is explicitly a clinical decision support tool, not an autonomous diagnostic system. All outputs are ranked hypotheses for physician evaluation, with evidence chains showing the data sources, variant annotations, and network evidence supporting each hypothesis. Clinical judgment retains full authority.

IV. MNTU: Molecular Node Tracer Unified

A. Molecular Network Analysis Framework

MNTU provides deep molecular network analysis of disease mechanisms identified at the HqA causation mapping layer. The system operates on three primary network types: protein-protein interaction (PPI) networks from STRING [2] and the Human Protein Atlas [17]; metabolic pathway networks from KEGG [3]; and gene regulatory networks incorporating transcription factor binding and chromatin accessibility data.

B. Network Propagation and Disease Module Identification

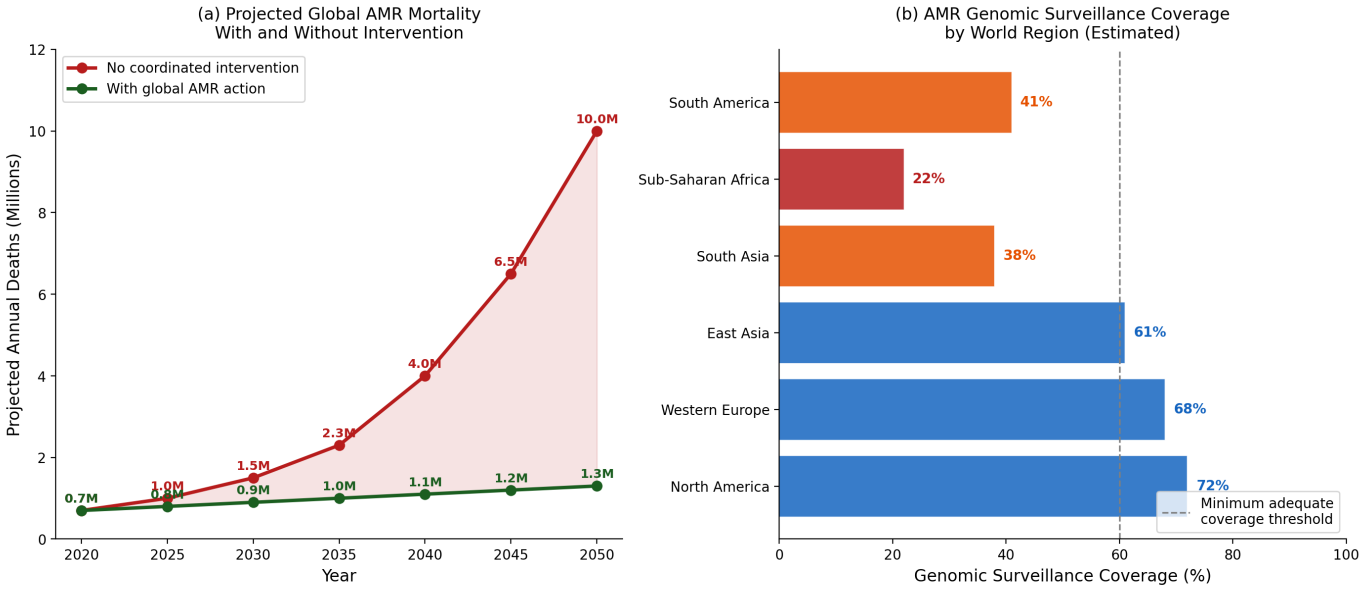
MNTU implements network propagation algorithms — analogous to random walk with restart (RWR) — to identify disease-relevant subnetworks from seed gene sets identified by HqA variant analysis. The disease module hypothesis of Menche et al. [11] establishes that disease genes cluster in network neighborhoods; MNTU traces propagation patterns to identify candidate therapeutic targets, secondary pathway involvement, and disease-disease mechanistic overlaps.

C. Graph Neural Networks for Molecular Analysis

For molecular property prediction, MNTU incorporates graph neural network (GNN) approaches following Wieder et al. [18] and the antibiotic discovery demonstration of Stokes et al. [19], who used a GNN to identify a structurally novel antibiotic active against drug-resistant *Mycobacterium tuberculosis*. AlphaFold [20] structural predictions provide structural hypotheses for the full proteome as an additional evidence layer for interaction plausibility assessment.

V. Antimicrobial Resistance Surveillance

The O'Neill Review [15] projected 10 million annual deaths from drug-resistant infections by 2050 absent global action — exceeding current annual cancer mortality. WHO GLASS [16] documents substantial gaps in genomic resistance gene surveillance coverage across low- and middle-income countries. Fig. 3 illustrates the projected mortality trajectory and regional surveillance coverage gaps at full scale.



Sources: (a) O'Neill Review on Antimicrobial Resistance, 2016 [15]. (b) WHO GLASS Report, 2022 [16]. Coverage estimates are approximate and subject to national reporting variability.

Fig. 3. (a) Projected global AMR mortality 2020–2050 with and without coordinated intervention. Source: O'Neill Review [15]. (b) Estimated genomic AMR surveillance coverage by world region. Source: WHO GLASS [16]. Values approximate and subject to reporting variability.

HqA's AMR module integrates resistance gene identification — using reference databases including ResFinder [21] and the Comprehensive Antibiotic Resistance Database (CARD) — with epidemiological spread modeling and clinical outcome correlation. Wastewater-based epidemiology, demonstrated by Peccia et al. [26] to track SARS-CoV-2 community dynamics four to ten days ahead of clinical case reporting, represents a planned complementary surveillance modality for AMR monitoring.

VI. Privacy-Preserving Health Research Architecture

A. The Health Data Silo Problem

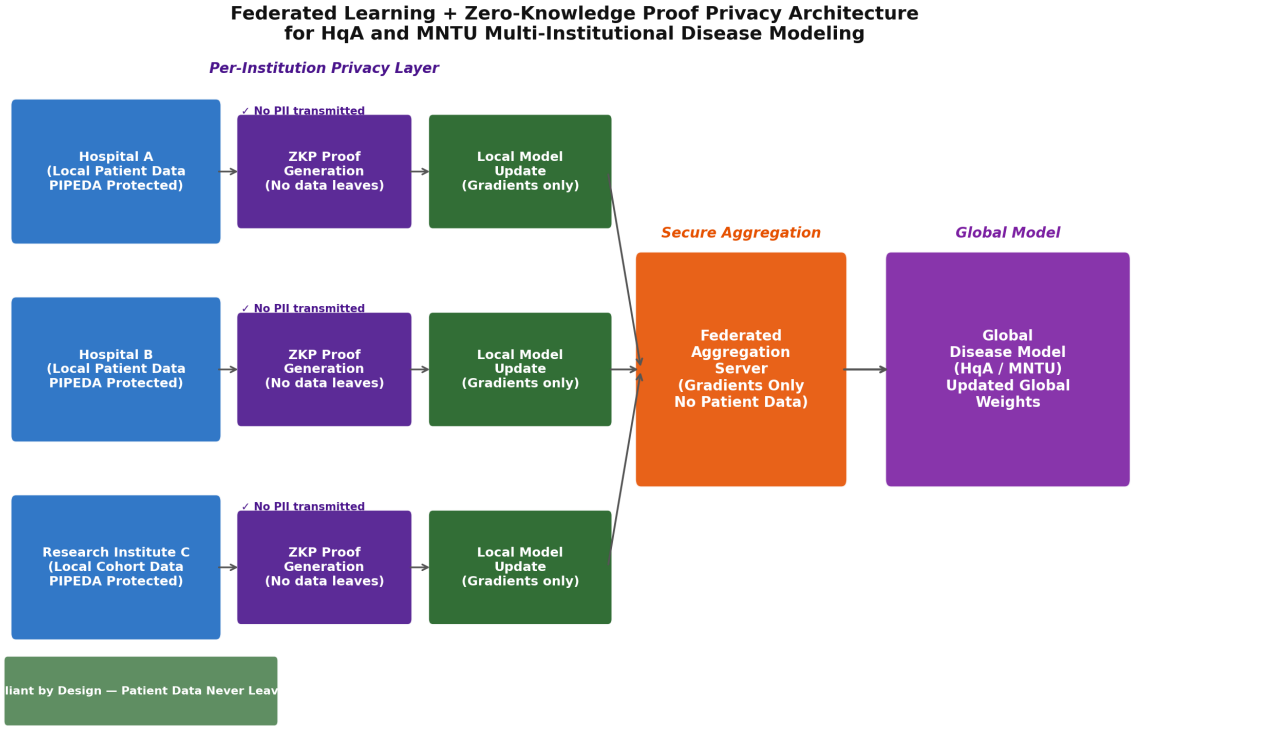
The tension between population-scale training data requirements and patient privacy protections under PIPEDA has historically resulted in data siloing that limits the statistical power of computational health research. Bonomi et al. [25] document the specific challenges of genomic data sharing, noting that genomic data is inherently re-identifying and current de-identification techniques provide incomplete protection.

B. Federated Learning

Federated learning, introduced by McMahan et al. [22] and applied to healthcare by Rieke et al. [23] and Sheller et al. [24], trains models across institutions without centralizing patient data. Only model gradients are transmitted. Sheller et al. [24] demonstrated federated performance within 0.6% of centralized training on brain tumor segmentation, establishing clinical viability.

C. Zero-Knowledge Proof Integration

HqA and MNTU extend federated learning with zero-knowledge proofs from the AKR Naos security stack, allowing institutions to prove that local model updates were computed from data satisfying specified properties — minimum cohort size, quality thresholds, diagnostic category — without revealing the data or gradients in full. Fig. 4 illustrates the combined architecture at full scale.



Sources: Federated learning: McMahan et al. (2017) [22]; Rieke et al. (2020) [23]; Sheller et al. (2020) [24]. ZKP privacy layer: Bonomi et al. (2020) [25]; Goldreich et al. (1991). PIPEDA: Government of Canada.

Fig. 4. Proposed privacy-preserving federated learning architecture integrating zero-knowledge proofs from AKR Naos. Patient data never leaves the institutional boundary under PIPEDA. Only model gradients and ZKP proofs are transmitted to the aggregation server. Sources: McMahan et al. [22]; Rieke et al. [23]; Sheller et al. [24]; Bonomi et al. [25].

VII. Health Equity and Underserved Populations

A. Indigenous Health in Canada

Indigenous peoples in Canada experience substantially worse health outcomes across virtually every measurable indicator, reflecting long-term consequences of colonization, displacement, and systemic inequities in healthcare access. The First Nations Information Governance Centre [27] documents that First Nations adults on-reserve experience rates of diabetes, tuberculosis, and mental health disorders significantly exceeding national averages. HqA is designed for deployment in resource-constrained settings including remote and rural communities, with explicit attention to high-prevalence conditions in Indigenous populations. Health equity is a design requirement.

B. Energy Poverty and Health

WHO estimates household air pollution from solid fuel combustion causes approximately 3.2 million deaths annually [28]. Absence of reliable electricity eliminates cold-chain vaccine storage, insulin refrigeration, and medical equipment operation. The Global Burden of Disease study [29] identifies energy poverty as a modifiable risk factor for a broad range of preventable outcomes. HqA's causal modeling incorporates environmental exposure as a data layer, enabling energy-health causal pathway analysis at population level.

C. Pandemic Preparedness

Peccia et al. [26] demonstrated SARS-CoV-2 RNA in wastewater predicted clinical case counts four to ten days in advance, providing pre-detection early warning. Integration of wastewater epidemiology, genomic zoonotic spillover prediction, and mobility-adjusted transmission modeling represents the computational infrastructure for pandemic early warning. HqA's epidemiological ingestion layer is designed to support this integration.

VIII. Ethical Framework

A. Open Source and Auditability

Clinical decision support tools affecting patient care have an accountability obligation proprietary systems cannot satisfy. HqA and MNTU are published as open-source implementations, allowing clinicians, researchers, regulators, and patients to inspect the methods underlying any recommendation.

B. Human Clinical Authority

No output from HqA or MNTU constitutes a clinical diagnosis or treatment recommendation. All outputs are evidence-supported hypotheses for physician evaluation. The physician is the principal; the system is the analyst.

C. Limitation Disclosure

HqA and MNTU are in active development without clinical validation. No performance benchmarks against clinical populations are claimed. The methodological foundations are drawn from validated published approaches; their integration requires independent clinical evaluation before patient care deployment. This is an unconditional and publicly maintained obligation.

D. PIPEDA Compliance

The federated ZKP architecture ensures no personally identifiable health information leaves institutional custody. Institutional data sharing agreements, ethics board review, and patient consent frameworks remain the responsibility of participating institutions and are prerequisites for any deployment.

IX. Discussion

The computational health gap is not primarily a technological problem. The methods underlying HqA and MNTU — causal inference, phenotype ontology matching, network propagation, federated learning, graph neural networks — are established in the published literature. The gap is integrative and institutional: these methods exist in separate research silos, implemented in non-interoperable tools, and inaccessible to clinical settings where they would have the greatest impact on patient outcomes.

Several limitations merit acknowledgment. First, causal inference from observational health data requires careful confounding control; HqA causal claims are hypotheses requiring validation, not established causes. Second, the federated ZKP architecture introduces computational overhead whose performance-privacy tradeoff requires empirical characterization in health-specific deployment contexts. Third, equity commitments require active partnership with Indigenous communities and remote health authorities as genuine design collaborators.

X. Conclusion

This paper has described the proposed architecture and ethical framework of HqA and MNTU: open-source computational health tools addressing disease causation mapping and molecular network analysis. The systems integrate causal inference methodology, HPO phenotype matching, STRING protein interaction networks, KEGG metabolic pathway mapping, AlphaFold structural data, graph neural network molecular analysis, federated learning, and zero-knowledge proof privacy guarantees into a coherent PIPEDA-aligned platform.

The six problem domains addressed — rare disease diagnostic delay, AMR surveillance gaps, health data privacy, Indigenous health equity, energy poverty and health, and pandemic early warning — represent areas where computational infrastructure to translate available data into clinical and public health benefit is demonstrably insufficient. Clinical validation through institutional collaboration is the essential next step. PTH Meridian actively seeks partnerships with Canadian academic health centres, national research institutes, and Indigenous health organizations for collaborative validation studies. All implementations are published at github.com/PTHMeridian.

Glossary of Terms

AlphaFold

AI system [20] predicting 3D protein structure from amino acid sequence. The AlphaFold Protein Structure Database covers predictions for over 200 million proteins.

AMR (Antimicrobial Resistance)

Ability of microorganisms to resist antimicrobial agents. O'Neill Review [15] projects 10M annual deaths by 2050 without global action.

Causal Inference

Mathematical framework [4] for identifying causal rather than correlational relationships, using structural causal models and directed acyclic graphs.

Federated Learning

Distributed ML approach [22] training models locally at each institution; only model gradients, not patient data, are transmitted to a central aggregator.

HPO (Human Phenotype Ontology)

Controlled vocabulary of 13,000+ phenotypic abnormalities [12] enabling computational phenotype-to-disease matching.

KEGG

Kyoto Encyclopedia of Genes and Genomes [3]: reference database of metabolic, signaling, and regulatory pathway maps used by MNTU.

MNTU

PTH Meridian planned molecular network analysis platform tracing disease origination and propagation across PPI, metabolic, and gene regulatory networks.

PIPEDA

Canada's Personal Information Protection and Electronic Documents Act: federal legislation governing collection, use, and disclosure of personal information.

STRING

Protein-protein interaction database [2] providing the PPI network foundation for HqA variant neighborhood analysis and MNTU network tracing.

ZKP (Zero-Knowledge Proof)

Cryptographic protocol proving a property of data without revealing the data. Used in HqA/MNTU to verify cohort properties without disclosure.

References

[1] National Human Genome Research Institute, "DNA Sequencing Costs: Data," NHGRI, 2024. [Online]. Available: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

[2] D. Szklarczyk et al., "The STRING database in 2023," *Nucleic Acids Research*, vol. 51, no. D1, pp. D638–D646, 2023. doi: 10.1093/nar/gkac1000

[3] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000. doi: 10.1093/nar/28.1.27

[4] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge: Cambridge University Press, 2009.

[5] M. A. Hernán and J. M. Robins, *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020. Available: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

[6] B. Schölkopf et al., "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021. doi: 10.1109/JPROC.2021.3058954

[7] A.–L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004. doi: 10.1038/nrg1272

[8] Shire plc, "Rare Disease Impact Report," Shire Human Genetic Therapies, 2013. [Online]. Available: https://globalgenes.org/wp-content/uploads/2013/04/ShireReport-1.pdf

[9] T. Richter et al., "Rare disease terminology and definitions," *Value in Health*, vol. 18, no. 6, pp. 906–914, 2015. doi: 10.1016/j.jval.2015.05.008

[10] EURODIS, "Rare Diseases: Understanding This Public Health Priority," Paris, 2005. [Online]. Available: https://www.eurordis.org

[11] J. Menche et al., "Uncovering disease-disease relationships through the incomplete interactome," *Science*, vol. 347, no. 6224, p. 1257601, 2015. doi: 10.1126/science.1257601

[12] P. N. Robinson et al., "The Human Phenotype Ontology," *American Journal of Human Genetics*, vol. 83, no. 5, pp. 610–615, 2008. doi: 10.1016/j.ajhg.2008.09.017

[13] S. Köhler et al., "Clinical diagnostics in human genetics with semantic similarity searches in ontologies," *American Journal of Human Genetics*, vol. 85, no. 4, pp. 457–464, 2009. doi: 10.1016/j.ajhg.2009.09.003

[14] M. J. Landrum et al., "ClinVar: improving access to variant interpretations," *Nucleic Acids Research*, vol. 46, no. D1, pp. D1062–D1067, 2018. doi: 10.1093/nar/gkx1153

[15] J. O'Neill, "Tackling Drug-Resistant Infections Globally: Final Report," Review on Antimicrobial Resistance, London, May 2016. Available: https://amr-review.org

[16] World Health Organization, "GLASS Report 2022," WHO, Geneva, 2022. [Online]. Available: https://www.who.int/publications/i/item/9789240062702

[17] M. Uhlen et al., "Tissue-based map of the human proteome," *Science*, vol. 347, no. 6220, p. 1260419, 2015. doi: 10.1126/science.1260419

[18] O. Wieder et al., "Molecular property prediction with graph neural networks," *Drug Discovery Today: Technologies*, vol. 37, pp. 1–12, 2020. doi: 10.1016/j.ddtec.2020.11.009

[19] J. M. Stokes et al., "A deep learning approach to antibiotic discovery," *Cell*, vol. 180, no. 4, pp. 688–702.e13, 2020. doi: 10.1016/j.cell.2020.01.021

[20] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, pp. 583–589, 2021. doi: 10.1038/s41586-021-03819-2

[21] E. Zankari et al., "Identification of acquired antimicrobial resistance genes," *Journal of Antimicrobial Chemotherapy*, vol. 67, no. 11, pp. 2640–2644, 2012. doi: 10.1093/jac/dks261

[22] H. B. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," *Proc. AISTATS*, Fort Lauderdale, 2017, pp. 1273–1282.

[23] N. Rieke et al., "The future of digital health with federated learning," *npj Digital Medicine*, vol. 3, art. 119, 2020. doi: 10.1038/s41746-020-00323-1

[24] M. J. Sheller et al., "Federated learning in medicine," *Scientific Reports*, vol. 10, art. 12598, 2020. doi: 10.1038/s41598-020-69250-1

[25] L. Bonomi, Y. Huang, and L. Ohno-Machado, "Privacy challenges for genomic data sharing," *Nature Genetics*, vol. 52, pp. 646–654, 2020. doi: 10.1038/s41588-020-0651-0

[26] J. Peccia et al., "Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics," *Nature Biotechnology*, vol. 38, pp. 1164–1167, 2020. doi: 10.1038/s41587-020-0684-z

[27] First Nations Information Governance Centre, "National Report of the First Nations Regional Health Survey Phase 3, Volume 1," FNIGC, Ottawa, 2018. [Online]. Available: https://fnigc.ca/our-reports-and-publications/

[28] World Health Organization, "Household Air Pollution," WHO Global Health Observatory, Geneva, 2023. Available: https://www.who.int/news-room/fact-sheets/detail/household-air-pollution-and-health

[29] GBD 2019 Risk Factors Collaborators, "Global burden of 87 risk factors in 204 countries and territories, 1990–2019," *The Lancet*, vol. 396, no. 10258, pp. 1223–1249, 2020. doi: 10.1016/S0140-6736(20)30752-2

[30] M. Ashburner et al., "Gene Ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000. doi: 10.1038/75556

[31] UniProt Consortium, "UniProt: the Universal Protein Knowledgebase in 2023," *Nucleic Acids Research*, vol. 51, no. D1, pp. D523–D531, 2023.

doi: 10.1093/nar/gkac1052