 https://doi.org/10.7494/csci.2025.26.SI.7067

Eda Cakir
Marc-Thorsten Hütt

NETWORK-BASED COMPUTATIONAL PIPELINE FOR STUDYING VARIABILITY OF TRANSCRIPTOME PROFILES FOR HUMAN DISEASES

Abstract *Machine learning applications to high-throughput data in medicine – one of the biggest resources for understanding complex diseases – have been limited thus far. Here, we present a computational approach for assessing the intrinsic variability in the most prominent data type, transcriptomics data for disease cohorts. Our study looks at situations where multiple data sets for the same disease are available. We leverage concepts of network medicine to assess how the match between a biological network and a set of differentially expressed genes varies across different networks and experiments. Our results showed that different biological networks yielded markedly different results; also, the clustering of diseases depended strongly on the choice of the parameters that were contained in the data analysis and network processing.*

Keywords transcriptomics data, network medicine, disease cohorts

Citation Computer Science 26(SI) 2025: 69–91

1. Introduction

Over the last decades, biology and medicine have become data sciences. High-throughput ('omics') data on the levels of gene expression, metabolic activity, epigenetic regulation, and others now serve as a prominent source of systemic information; this makes these fields accessible to data-driven computational methods – particularly, network science and machine learning.

Network science [5,19] employs the formal view of graph theory to understand the design principles of complex systems. Abstracting cellular processes (gene regulation, metabolism, protein interactions) into networks has revolutionized the way we think about biological systems [2, 7].

The two biggest success stories of machine learning (ML)/artificial intelligence (AI) that have been applied to medicine are arguably AlphaFold [1,27] and the diverse variants of medical-image classification (see, e.g., [28,33,55]). With protein-structure prediction and, in the most recent version [1], docking and ligand binding prediction, AlphaFold holds enormous potential for drug-target prediction and drug repurposing. Medical-image classification benefits from reliable network architectures for general image classification and the substantial volumes of reasonably standardized medical images [33, 55].

For the data driving the emerging fields of Systems Biology and Systems Medicine [3, 23, 34, 53], high-throughput data has made this possible due to technological advances in sequencing and the assessments of the DNA structure and its modifications; however, the situation regarding AI applications is a bit different. So far, the promise of applying machine learning to these data sets has been widely acknowledged (see [11, 62] as examples); however, the actual applications are limited and still not performing well (see [13, 18] as examples).

Here, we focus on the most prominent type of high-throughput data – the simultaneous measurements of the activities of (nearly) all of the genes in a cell (or, often rather, average activity levels across whole populations of cells). For human diseases, such gene expression patterns or transcriptome profiles allow for functional characterization of disease phenotype [12]. A typical data set consists of the transcriptome profiles of N_p patients and of N_c healthy controls. By using established statistical approaches [35], these two sets can be compared and differentially expressed genes (up-regulated or down-regulated in the patients as compared to the controls) can be extracted. Mapping these data into given biological networks is a common interpretation strategy of such gene sets (see, e.g., [24,30,44,61,65]). This is also the approach that we will follow here. Our research question is whether this data will produce reliable functional characterizations of a disease or whether the experiment-to-experiment variability is too great (e.g., larger than the differences among different diseases). To address this question, we select publicly available gene expression data sets for a set of diseases in which more than one data set is available. In this way, we can quantify the experiment-to-experiment variability of network indicators for differentially expressed gene sets and compare it with the disease-to-disease variability of the same quantities.

The reasons for the potentially high variability of transcriptomics data (as well as other high-throughput data) for human diseases are diverse and multifaceted: on the one hand, the phenotype space is huge – particularly, when compared to the available cohort sizes. Hence, any inter-individual differences beyond a phenotype under investigation (i.e., a disease at hand) may mask the 'signal' (disease-related expression patterns; differentially expressed genes) that the experiment was designed to detect.

On a more general level, information in biology is organized via an interplay of digital and analog data. This has been studied most clearly in the context of bacterial gene regulation [31, 40, 41, 67, 68]; however, it is present in all aspects of biology and across all levels of organization. In the context of the characterization of diseases via high-throughput data, this is perhaps most clearly seen in the rich universe of novel findings regarding chromosomal organization [32, 57, 63] and the coexistence of network-based (i.e., essentially digital) and chromosome-based (i.e., essentially analog) interpretations of such data (see [26], for example). How does this affect the experiment-to-experiment variability in the signals that are extracted from 'omics' data? If one analyzes a three-dimensional object via a projection onto a two-dimensional plane, then several data sets of the same (only slightly tilted) object will lead to high variability. To a certain degree, this is the situation with the interpretation of 'omics' data: attempts of the functional interpretation of such data (e.g., via concepts of network medicine [4, 6]) often use only one category of information. This restricted perspective can contribute to the perceived variability, as data is only analyzed as projections of higher-dimensional objects. The case of bacterial gene regulation deserves a further comment. Even in this case, the differences in the spatial organization of iModulons (derived via machine learning applied to gene expression data, [60]) as compared to standard regulons (derived from biological knowledge) suggests that analog information could well be responsible for the still-limited power of machine learning approaches to predict expression patterns – even for cases of bacterial cells [10].

Here, we design a network-based computational analysis pipeline with the purpose of quantitatively assessing the variability of transcriptome profiles for human diseases. Section 2 introduces the computational pipeline; in Section 3, information about the gene expression data, the biological networks, and the statistical methods that are contained in our pipeline are summarized. The key results of our investigation are given in Section 4. Finally, Section 5 briefly discusses these results in a broader context. Details about the experimental data that was used (the sizes of the data sets, references, and metadata) are listed in tabular form in Appendices A and B.

2. Computational pipeline

We retrieved 20 RNA-Seq gene expression data sets that were associated with 7 distinct diseases from the Expression Atlas database [48] and identified differentially expressed gene sets by varying the log2 fold-change values that fell between 1 and 5 (see Section 3.1).

In order to thoroughly examine the network coherence values of the differentially expressed genes within the framework of the biological networks, we employed a gene-centric metabolic network known as Recon3D [9] (see Section 3.2) alongside two distinct gene-level protein-protein interaction networks, which were identified as String [66] and Biogrid [46] (see Section 3.3). Our investigation centered on the induced subgraphs that represented the projections of the differentially expressed genes (up- and down-regulated, up-regulated only, or down-regulated only) onto the aforementioned gene-centric metabolic network or the gene-level protein-protein interaction networks that we utilized in our analytical procedures (see Section 3.4). A comprehensive analysis of the connectivity ratios (the nodes that were connected to all of nodes in the induced subgraph) was conducted on these subgraphs; we assessed whether the observed connectivity ratios exhibited values that were statistically larger or smaller than those that would be anticipated under conditions of randomness. These z-scores (called *network coherences*) were the bases for the hierarchical cluster analyses.

The computational pipeline is summarized in Figure 1.

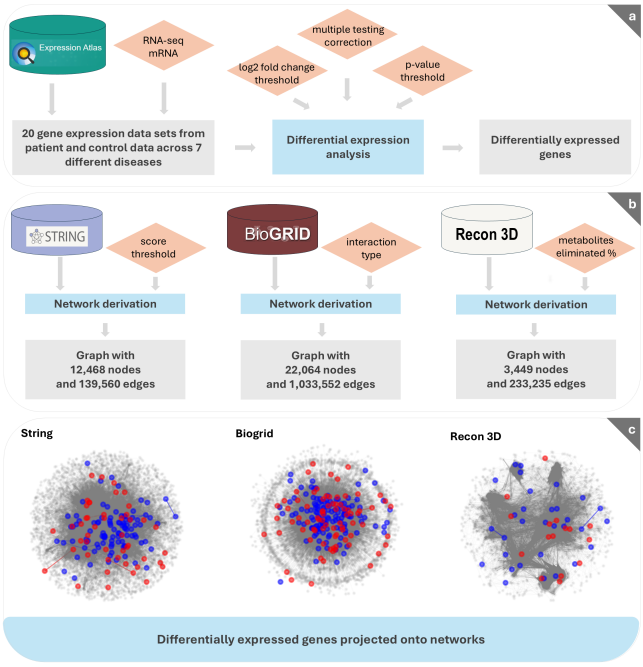

Figure 1. **(a–b)** Investigation was performed by projecting differentially expressed genes from 20 data sets obtained from Expression Atlas database that were associated with 7 distinct diseases onto 3 separate networks – specifically, gene-level protein-protein interaction networks that were sourced from String and Biogrid databases as well as gene-centric metabolic network that was derived from Recon3D human metabolic model; **(c)** three network figures illustrate projections of up-regulated genes (blue dots) and down-regulated genes (red dots) from single data set (E-MTAB-7915 [54]) onto these networks

3. Methods

3.1. Gene-expression data sets

Twenty RNA-Seq gene-expression data sets that were relevant to human diseases were obtained from the Expression Atlas database [48], along with experimental design and analysis data. A compilation of the gene-expression data sets, the associated diseases, and the relevant scientific articles is presented in Appendix A.

The comprehensive metadata that was associated with each data set was systematically obtained from the Expression Atlas and BioStudies databases [59]; these can be found in Appendix B. The metadata included the following: (1) the name of the disease; (2) the total number of assays (which referred to the number of experimental tests that were performed as parts of the studies); (3) the years of the publications or releases to the public; (4) the organism group (which referred to the specific taxonomic classifications of the organism parts or tissues that were studied or used as samples in the studies); (5) indications of whether the biological samples under consideration were derived from blood or were sourced from alternative tissues; (6) the categorization of the disease group based on the International Classification of Diseases (ICD); and (7) an indication of whether the identified disease was classified within the cancer category or otherwise.

3.2. Gene-centric metabolic-network construction

We derived the gene-centric metabolic model from Recon3D [9] human metabolic models while adhering to the methodologies that were detailed in [30, 47, 64, 65]. In this network, a connection between two genes is established when the metabolic reactions that are linked to these genes share a common metabolite.

The primary exchange metabolites (which include ATP, ADP, CO2, H, NAD, NADH, and several others) are recognized as the most highly interconnected metabolic species; their presence tends to obscure the establishments of links among those genes that possess similar metabolic functionalities. This consequently results in an artificially inflated density of the metabolic network [29, 37, 38]. To effectively mitigate this confounding effect, we decided to eliminate a subset of the metabolites that corresponded to the top 2% of the most highly connected metabolites prior to the construction of the network.

The resulting gene-centric metabolic network that was obtained through the utilization of the Recon3D human metabolic network model had 3449 nodes and 233,235 edges.

3.3. Gene-level protein-protein interaction-network construction

For the purpose of our analysis, we conducted our investigations by utilizing two distinct protein-protein interaction networks at the gene level (GPINs): these were derived from the String [66] and Biogrid [46] protein-protein interaction databases.

For the GPIN that was constructed from the String database, we selectively included protein interactions that were relevant to the human species – specifically, those with scores that were greater than 850. We then cross-referenced the protein identifiers with the corresponding gene identifiers by using the Ensembl database [15]. The GPIN was represented as a graph, with the genes acting as nodes and edges that illustrated the relationships among the genes through their protein interactions. This graph contained 12,468 nodes and 139,560 edges.

The GPIN that was derived from the Biogrid database was assembled following the aforementioned methodology but without any filtering of the interaction scores. This graph contained totals of 22,064 nodes and 1,033,552 edges.

3.4. Network analysis

Differential expression-analysis results were retrieved from the Expression Atlas database [48]. The typical approach for identifying differentially expressed genes within RNA-Seq data sets involves comparisons of gene-expression levels across various experimental conditions following the normalization of the raw read counts for the sequencing depth and RNA composition. In principle, the log2 fold change LF_k for gene k can be calculated as follows:

$$LF_k = \log_2\left(\frac{\langle e_k\rangle}{\langle f_k\rangle}\right) = \log_2\left(\frac{\frac{1}{n_p}\sum_i^{n_p} e_k^{(i)}}{\frac{1}{n_c}\sum_j^{n_c} f_k^{(j)}}\right),$$

where n_p is the number of patient samples, n_c is the number of control samples, and $e_k^{(i)}$ and $f_k^{(j)}$ are the normalized expression levels of the k^{th} gene for patient sample i and control sample j, respectively.

The DESeq2 algorithm goes beyond this by fitting a generalized linear model to each gene to test for significant expression changes [36]. The results include log2 fold changes and adjusted p-values, which account for the multiple testing.

The differentially expressed genes were subjected to a filtering process based on log2 fold-change values that ranged from 1 to 5; additionally, we required that the adjusted p-values be less than or equal to 0.05 (see Figure 2a). Then, the filtered differentially expressed genes systematically projected onto the network under consideration (see Figure 2b). To assess the clustering of these genes in the given network, we analyzed the induced subgraph that was spanned by these genes.

The *connectivity ratio* of this subgraph was the ratio of the non-isolated genes to the total number of genes in the subgraph. It is important to note that this analysis was conducted exclusively under the condition that the number of differentially expressed genes was equal to or exceeded the threshold of 5.

In principle, a range of other properties of the induced subgraph could have been used for this analysis. In a previous work, it was shown that the connectivity ratio worked well for comparatively weak signals [45].

To derive the null distribution of the network connectivity ratios, we randomly drew 5000 gene sets; each was equivalent in size to the induced subgraph (from the

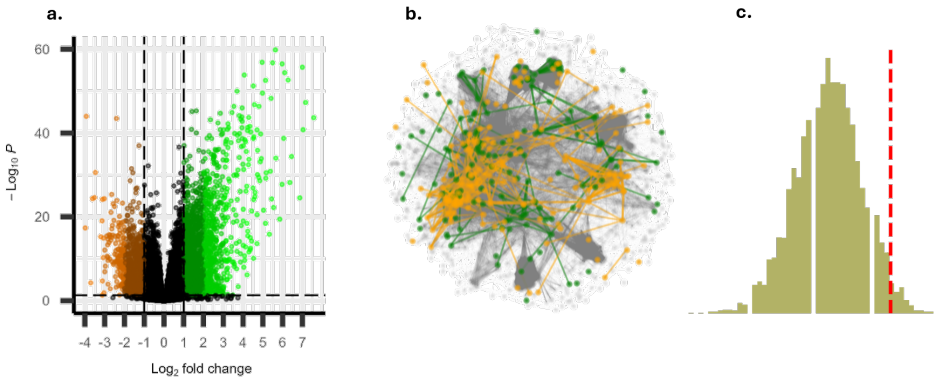

Figure 2. Differential expression and network analysis of single data set (E-GEOD-57945): **(a)** differentially expressed genes were detected by changing log2 fold-change values between 1 and 5 (black dots: log2 fold changes between –1 and 1; dark-green to light-green: up-regulated genes [color becomes lighter as log2 fold-change value increases]; brown to light-orange: down-regulated genes [color becomes lighter as log2 fold-change value decreases]) – adjusted p-value threshold is less than or equal to 0.05; **(b)** network figure illustrates projection of up-regulated genes (green dots) and down-regulated (orange dots) genes onto gene-centric metabolic network that was derived from Recon3D human metabolic model; **(c)** network coherence is z-score of connectivity ratio of induced subgraph (illustrated by vertical dashed red line) in relation to null distribution of connectivity ratios of 5000 randomly selected subnetworks

employed network), and we subsequently calculated the connectivity ratios for these randomly selected subnetworks. The z-score of the connectivity ratio with respect to this null distribution is called *network coherence* [30, 64, 65] and will be at the center of our subsequent analysis (see Figure 2c).

3.5. Hierarchical clustering

Following the calculation of the network coherence values for all of the data sets across the three different networks, these values were then used to perform hierarchical clusterings of the data sets by utilizing average linkage and Euclidean distance metrics. The dendrograms now allowed us to assess whether the same diseases were clustered together (first color bar). We then compared the metadata table with the disease clusters to see if the resulting clusters could be explained by the categorical attributes that were present.

4. Results

In order to investigate the extent to which the different diseases produced analogous network signals and to assess the robustness of these signals across the different gene-expression data sets, we used three different biological networks to compute

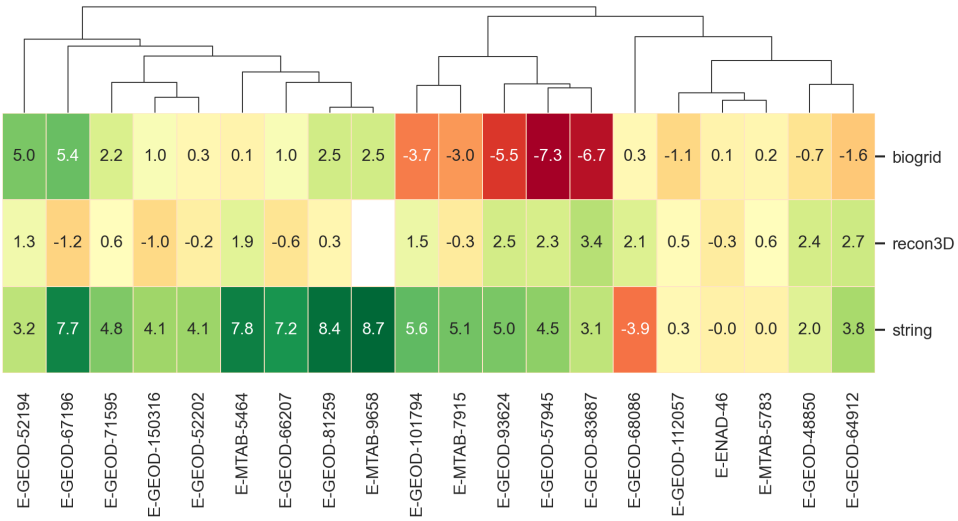

Figure 3. Heatmap representing network coherence values that were calculated by mapping differentially expressed genes (both up-regulated and down-regulated where the log2 fold change was 1) onto three distinct networks: STRING, BioGRID, and Recon3D

the network coherences that were associated with the differentially expressed genes. While setting the log2 fold-change values to within a range of 1 to 5 (denoted as LF1–LF5), this computational analysis was performed in three specific ways: first, by considering both the up- and down-regulated genes; second, by examining only the up-regulated genes; and third, by focusing only on the down-regulated genes. Following this comprehensive analysis, we performed a hierarchical clustering analysis to determine whether those diseases with similar characteristics tended to cluster together based on the premise that their respective network signals exhibited some degrees of similarity.

Figure 3 shows an example of such a result for one choice of a log2 fold-change threshold and for the joint set of up- and down-regulated genes. The clustering tree that arose from the hierarchical clustering that was performed on the three-component vectors of the network coherences for each disease consisted of three quite-pronounced disease clusters. These clusters could be confirmed by visual inspections of the network coherence values (shown in the color coding in Figure 3) for the three networks across the diseases. Figure 4 now shows this clustering tree (top left) and all of the other clustering trees that arose from our analysis, together with a range of disease characterizations and metadata that attempted to explain the disease clusters that were visible in these dendrograms.

The dendrograms now allowed us to assess information on whether the same diseases were clustered together (the first color bar); in none of the cases were the

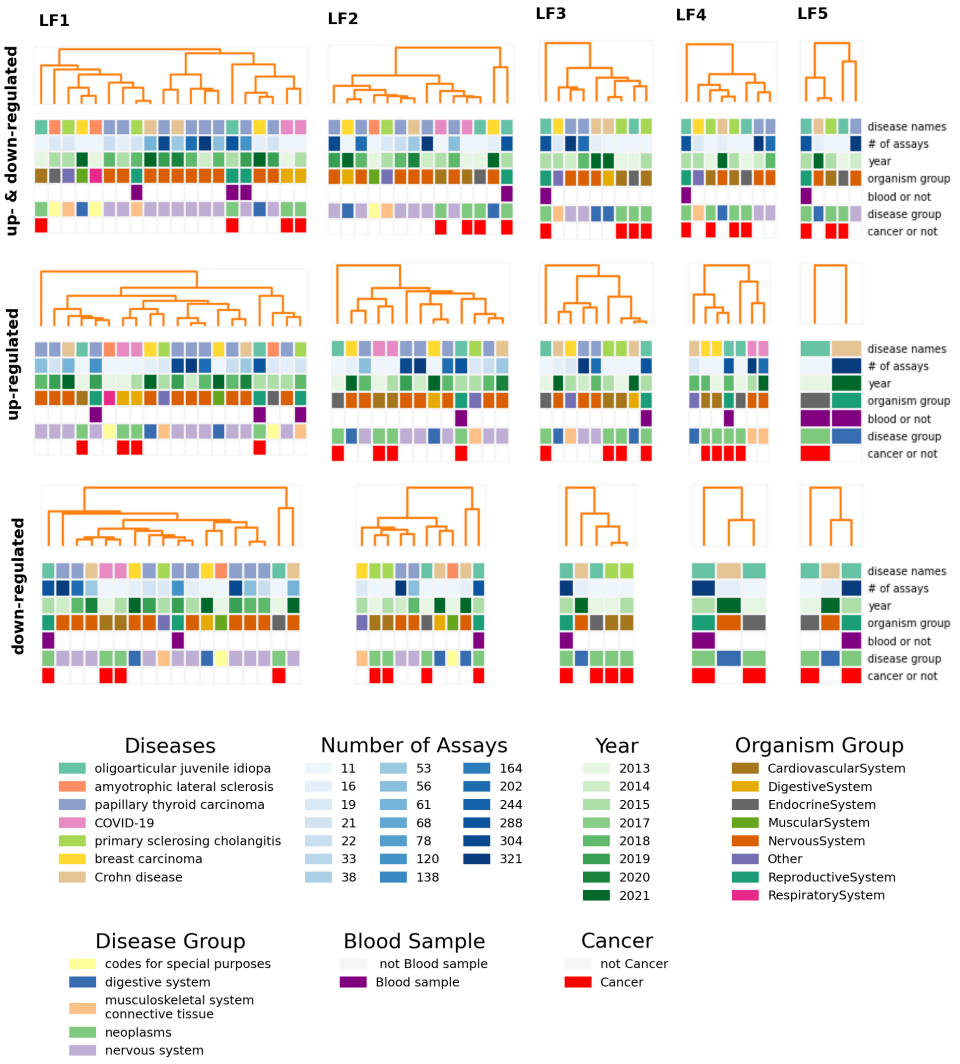

Figure 4. Hierarchical clustering and categorization of diseases; differential expression analyses were performed on each gene-expression data set by adjusting parameters to various configurations (i.e., filtering both up-& down-regulated, only up-regulated, or only down-regulated genes, and varying log2 fold change between 1 and 5 [LF1–LF5]) – network coherence values were computed by projecting differentially expressed genes onto three different networks (STRING, BioGRID, and Recon3D) and comparing these values to connectivity values that were derived from 5000 randomly selected gene sets of equivalent size; disease clusters were formed by taking the correlations of these network coherence values into account (lower section of subplots shows categories of diseases): Row 1: disease names; Row 2: number of assays in associated study; Row 3: year of publication; Row 4: organism group of associated sample; Row 5: specification of whether sample was derived from blood or not; Row 6: relevant disease group; Row 7: indication of whether disease fell within cancer category or not

disease clusters explained by the same diseases. We added other information in a color-coded form to see whether this additional metadata from the studies could explain the clusters in the dendrograms. Visual inspections suggested that none of these explained the clustering in a discernible manner.

Figure 5. Comparison of network-coherence values that were computed by projecting differentially expressed genes (up- & down-regulated, up-regulated, or down-regulated) onto three networks: STRING, BioGRID, and Recon3D (diseases indicated by color coding below graphic)

Table 1 confirms the general impression from Figures 4 and 5 for the example of a log2 fold change of 1 (the same as in Figure 5): the experiment-to-experiment variability for a single disease was typically not smaller than the disease-to-disease differences (represented below by the average network coherence differences across all of the diseases).

Summarizing, we presented the results of the hierarchical clustering and the subsequent categorization of the different diseases under the different parameter settings in Figure 4. Despite the observation of clear clustering patterns among the data sets in almost all of the scenarios, the clusters themselves could not be adequately explained by either the disease pairs nor the available metadata that was associated with the data sets that were included in this study.

Table 1

Average difference in network coherence for LF1 case – rows indicate different networks and differential expression types (with last rows indicating averages across all networks); last column is average difference in network coherence between any two data sets (independent of associated disease)

network	method	breast carcinoma	Crohn's disease	papillary thyroid carcinoma	oligoarticular juvenile idiopa	primary sclerosing cholangitis	amyotrophic lateral sclerosis	COVID-19	avg
string	up&down	7.13	3.46	1.81	3.56	3.56	3.63	4.10	3.68
	up	1.67	3.33	1.13	3.27	8.00	7.39	1.32	3.19
	down	10.60	2.41	0.04		10.58		4.71	3.94
biogrid	up&down	4.65	4.00	0.96	0.30	5.57	5.15	0.89	3.92
	up	7.17	4.67	0.31	0.90	0.66	6.60	4.07	3.80
	down	7.64	2.57	0.69		7.23		5.78	3.77
recon3D	up&down	0.83	1.56	0.23	0.38		1.09	0.76	1.63
	up	0.75	0.94	0.88	0.44			1.79	1.28
	down	0.26	1.23	1.09				0.08	1.60
cross-network	up&down	4.20	3.01	1.00	1.41	4.56	3.29	1.92	3.13
	up	3.20	3.17	0.77	1.54	4.33	6.99	2.39	2.91
	down	6.17	2.22	0.61		8.91		3.52	3.28

In the next phase of our analysis, we directed our attention toward the network-coherence signals that were derived from each of the networks when the log2 fold change was fixed at a value of 1. This approach was implemented in order to facilitate the evaluation of the majority of the signals without imposing a strict threshold that could potentially exclude any relevant data. As is shown in Figure 5, it was clear that the signals showed considerable variability – even when the same disease and/or the same data set were analyzed across the different networks. Focusing on Crohn's disease, we observed that the two different sources of protein-protein interactions (namely, String, and BioGrid) showed contrasting signals across the analyzed data sets for both the up-regulated and down-regulated genes as well as in those scenarios where both types of genes were assessed together. In particular, the String network consistently yielded markedly positive network coherences, whereas the BioGrid network exhibited significantly negative ones. Furthermore, the signals that were obtained from the Recon3D metabolic network predominantly fell within a range of -2 to 2, thus indicating the lack of a statistical significance in the results.

5. Conclusion

Bringing machine learning to 'omics' data requires a substantial volume of reliable data. For the most prominent class of 'omics' data – gene-expression patterns or transcriptomics data – we studied data variability here by comparing different experiments for the same diseases using a network medicine perspective.

We analyzed seven diseases; for each, between two and eight transcriptomics data sets were available, together with the corresponding controls. Our results suggested that the network signatures of transcriptomics data indeed showed a strong variability, with the experiment-to-experiment variation being of a similar size as the disease-to-disease variation.

Network coherence is a simple indicator of the clustering of gene sets (in this case, the differentially expressed genes for a particular disease) in a given biological network. Our computational pipeline assessed this clustering of differentially expressed genes in three biological networks, hence arriving at a multi-network version of network coherence. Disease clusters that were based on this multi-network coherence did not group the same diseases together in a systematic way and were not explained by the metadata (like the organismal subsystem or disease category).

Returning to the origins of the variability in human disease data that was listed in the introduction, we came to the conclusion that, indeed, variability is currently prohibitive of large-scale machine-learning applications to this data based on our results. Larger data sets and joint perspectives using the features of both network and chromosomal organizations may reduce this variability in the future.

These results also showed that any conclusions that are drawn from a single disease data set and a single biological network may not provide a comprehensive picture of a disease.

Acknowledgements

MH is grateful to the organizers of the 2nd International Workshop on Machine Learning and Quantum Computing Applications in Medicine and Physics (WMLQ2024) for their kind invitation as well as the opportunity to speak at this fascinating and highly interdisciplinary event.

A. Gene-expression data sets

Experiment ID	Experiment Title	Number of Assays	Disease ID	Disease Name	Reference
E-GEOD-52194	RNA-seq of 17 breast tumor samples of three different subtypes and normal human breast organoid samples	19	EFO:0000305	breast carcinoma	[17]
E-GEOD-68086	RNA-seq of blood platelets from six tumor types and healthy donors	288	EFO:0000305	breast carcinoma	[8]
E-GEOD-83687	RNA-seq of 134 patients undergoing bowel resection for inflammatory bowel disease and controls	138	EFO:0000384	Crohn's disease	[51]
E-MTAB-5783	RNA-seq of formalin-fixed, paraffin-embedded uninvolved terminal ileal tissue obtained from ileo-colic resection surgeries of Crohn's disease and control patients	68	EFO:0000384	Crohn's disease	[69]
E-GEOD-66207	mRNA and small RNA associated with Crohn's disease behavior [RNA-Seq]	33	EFO:0000384	Crohn's disease	[49]
E-GEOD-112057	Whole-blood transcriptome profiling in juvenile idiopathic arthritis and inflammatory bowel disease	202	EFO:0000384	Crohn's disease	[42]

Continued on next page

Experiment ID	Experiment Title	Number of Assays	Disease ID	Disease Name	Reference
E-GEOD-101794	RNA-seq of ileal biopsies from diagnostic endoscopy of pediatric Crohn's disease patients and non inflammatory bowel disease controls	304	EFO:0000384	Crohn's disease	[20]
E-GEOD-57945	RNA-seq of 359 treatment-naive pediatric patients with Crohn's disease, patients with ulcerative colitis, and control individuals	321	EFO:0000384	Crohn's disease	[21]
E-GEOD-93624	RNA-seq of 210 treatment-naive patients of pediatric Crohn's disease and 35 non-IBD controls from RISK study	244	EFO:0000384	Crohn's disease	[39]
E-MTAB-5464	RNA-sequencing of purified intestinal epithelial cells from paediatric biopsies (including inflammatory bowel disease and healthy controls)	78	EFO:0000384	Crohn's disease	[22]
E-GEOD-48850	Novel kinase fusion oncogenes in post-Chernobyl radiation-induced pediatric thyroid cancers	11	EFO:0000641	papillary thyroid carcinoma	[56]
E-GEOD-64912	RNA-sequencing of human papillary thyroid carcinomas	22	EFO:0000641	papillary thyroid carcinoma	[14]

Continued on next page

Experiment ID	Experiment Title	Number of Assays	Disease ID	Disease Name	Reference
E-GEOD-71595	RNA-sequencing of cells derived from sites of inflammation of juvenile idiopathic arthritis patients	56	EFO:0002609	juvenile idiopathic arthritis	[50]
E-GEOD-81259	Transcriptional profiling revealed monocyte signature associated with JIA patient poor response to methotrexate	61	EFO:0002609	juvenile idiopathic arthritis	[43]
E-MTAB-7915	PSC-IBD mucosal biology	120	EFO:0004268	sclerosing cholangitis	[54]
E-MTAB-9658	Tissue-dependent transcriptional and bacterial associations in primary sclerosing cholangitis-associated inflammatory bowel disease	164	EFO:0004268	sclerosing cholangitis	[25]
E-GEOD-52202	Transcription profiling by high-throughput sequencing of iPSC-derived motor neuron cultures from C9ORF72 carriers	16	MONDO:0004976	amyotrophic lateral sclerosis	[58]
E-GEOD-67196	Transcription profiling by high-throughput sequencing of cerebellum and frontal cortex from patients of amyotrophic lateral sclerosis	53	MONDO:0004976	amyotrophic lateral sclerosis	[52]

Continued on next page

Experiment ID	Experiment Title	Number of Assays	Disease ID	Disease Name	Reference
E-GEOD-150316	Spectrum of viral load and host response seen in autopsies of SARS-CoV-2 infected lungs	21	MONDO:0100096	COVID-19	[16]
E-ENAD-46	Lung and colon transcriptome profiling of fatal COVID-19 cases	38	MONDO:0100096	COVID-19	[70]

B. Metadata

Exp. Name	Disease ID	Disease Name	Num. of Assays	Year	Organism Part	Organism Group	Blood or not	Disease Group	Cancer or not
E-GEOD-52194	EFO: 0000305	breast carcinoma	19	2013	breast	reproductive system	not blood	neoplasms	cancer
E-GEOD-68086	EFO: 0000305	breast carcinoma	288	2015	blood	cardiovascular system	blood	neoplasms	cancer
E-GEOD-66207	EFO: 0000384	Crohn's disease	33	2015	colon	digestive system	not blood	digestive system	not cancer
E-MTAB-5783	EFO: 0000384	Crohn's disease	68	2018	small intestine	digestive system	not blood	digestive system	not cancer
E-GEOD-93624	EFO: 0000384	Crohn's disease	244	2017	ileum	digestive system	not blood	digestive system	not cancer
E-GEOD-101794	EFO: 0000384	Crohn's disease	304	2018	ileum	digestive system	not blood	digestive system	not cancer
E-GEOD-57945	EFO: 0000384	Crohn's disease	321	2014	ileum	digestive system	not blood	digestive system	not cancer
E-MTAB-5464	EFO: 0000384	Crohn's disease	78	2017	ascending colon	digestive system	not blood	digestive system	not cancer
E-GEOD-112057	EFO: 0000384	Crohn's disease	202	2018	blood	cardiovascular system	blood	digestive system	not cancer
E-GEOD-83687	EFO: 0000384	Crohn's disease	138	2017	colon	digestive system	not blood	digestive system	not cancer
E-GEOD-48850	EFO: 0000641	papillary thyroid carcinoma	11	2013	thyroid gland	endocrine system	not blood	neoplasms	cancer
E-GEOD-64912	EFO: 0000641	papillary thyroid carcinoma	22	2015	thyroid	endocrine system	not blood	neoplasms	cancer

Continued on next page

Exp. Name	Disease ID	Disease Name	Num. of Assays	Year	Organism Part	Organism Group	Blood or not	Disease Group	Cancer or not
E-GEOD-71595	EFO: 0002609	juvenile idiopathic arthritis	56	2015	synovial fluid	other	not blood	musculoskeletal system, connective tissue	not cancer
E-GEOD-81259	EFO: 0002609	juvenile idiopathic arthritis	61	2017	blood	cardiovascular system	blood	musculoskeletal system, connective tissue	not cancer
E-MTAB-9658	EFO: 0004268	sclerosing cholangitis	164	2021	caecum	digestive system	not blood	digestive system	not cancer
E-MTAB-7915	EFO: 0004268	sclerosing cholangitis	120	2019	colon	digestive system	not blood	digestive system	not cancer
E-GEOD-67196	MONDO: 0004976	amyotrophic lateral sclerosis	53	2015	cerebellum	nervous system	not blood	nervous system	not cancer
E-GEOD-52202	MONDO: 0004976	amyotrophic lateral sclerosis	16	2013	muscle	muscular system	not blood	nervous system	not cancer
E-GEOD-150316	MONDO: 0100096	COVID-19	21	2020	lung	respiratory system	not blood	codes for special purposes	not cancer
E-ENAD-46	MONDO: 0100096	COVID-19	38	2020	colon	digestive system	not blood	codes for special purposes	not cancer

References

[1] Abramson J., Adler J., Dunger J., Evans R., Green T., Pritzel A., Ronneberger O., *et al.*: Accurate structure prediction of biomolecular interactions with AlphaFold 3, *Nature*, pp. 1–3, 2024. doi: 10.1038/s41586-024-07487-w.

[2] Alon U.: Network motifs: theory and experimental approaches, *Nature Reviews Genetics*, vol. 8(6), pp. 450–461, 2007. doi: 10.1038/nrg2102.

[3] Alon U.: *Systems medicine: physiological circuits and the dynamics of disease*, CRC Press, 2023. doi: 10.1201/9781003356929.

[4] Barabási A.L.: Network medicine—from obesity to the "diseasome", 2007.

[5] Barabási A.L.: *Network Science*, Cambridge University Press, 2016.

[6] Barabási A.L., Gulbahce N., Loscalzo J.: Network medicine: a network-based approach to human disease, *Nature Reviews Genetics*, vol. 12(1), pp. 56–68, 2011.

[7] Barabasi A.L., Oltvai Z.N.: Network biology: understanding the cell's functional organization, *Nature Reviews Genetics*, vol. 5(2), 101, 2004. doi: 10.1038/nrg1272.

[8] Best M.G., Sol N., Kooi I., Tannous J., Westerman B.A., Rustenburg F., Schellen P., *et al.*: RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics, *Cancer Cell*, vol. 28(5), pp. 666–676, 2015. doi: 10.1016/j.ccell.2015.09.018.

[9] Brunk E., Sahoo S., Zielinski D.C., Altunkaya A., Dräger A., Mih N., Gatto F., *et al.*: Recon3D enables a three-dimensional view of gene variation in human metabolism, *Nature Biotechnology*, vol. 36(3), pp. 272–281, 2018. doi: 10.1038/nbt.4072.

[10] Cakir E., Lesne A., Hütt M.T.: The economy of chromosomal distances in bacterial gene regulation, *npj Systems Biology and Applications*, vol. 7(1), p. 49, 2021. doi: 10.1038/s41540-021-00209-2.

[11] Cannarozzi A.L., Latiano A., Massimino L., Bossa F., Giuliani F., Riva M., Ungaro F., *et al.*: Inflammatory bowel disease genomics, transcriptomics, proteomics and metagenomics meet artificial intelligence, *United European Gastroenterology Journal*, 2024. doi: 10.1002/ueg2.12655.

[12] Casamassimi A., Federico A., Rienzo M., Esposito S., Ciccodicola A.: Transcriptome profiling in human diseases: new advances and perspectives, *International Journal of Molecular Sciences*, vol. 18(8), p. 1652, 2017. doi: 10.3390/ijms18081652.

[13] Chen K.A., Nishiyama N.C., Kennedy Ng M.M., Shumway A., Joisa C.U., Schaner M.R., Lian G., *et al.*: Linking gene expression to clinical outcomes in pediatric Crohn's disease using machine learning, *Scientific Reports*, vol. 14(1), 2667, 2024. doi: 10.1038/s41598-024-52678-0.

[14] Costa V., Esposito R., Ziviello C., Sepe R., Bim L.V., Cacciola N.A., Decaussin-Petrucci M., *et al.*: New somatic mutations and WNK1-B4GALNT3 gene fusion in papillary thyroid carcinoma, *Oncotarget*, vol. 6(13), pp. 11242–11251, 2015. doi: 10.18632/oncotarget.3593.

[15] Cunningham F., Allen J.E., Allen J., Alvarez-Jarreta J., Amode M.R., Armean I.M., Austine-Orimoloye O., *et al.*: Ensembl 2022, *Nucleic Acids Research*, vol. 50(D1), pp. D988–D995, 2022. doi: 10.1093/NAR/GKAB1049.

[16] Desai N., Neyaz A., Szabolcs A., Shih A.R., Chen J.H., Thapar V., Nieman L.T., *et al.*: Temporal and spatial heterogeneity of host response to SARS-CoV-2 pulmonary infection, *Nature Communications*, vol. 11(1), p. 6319, 2020. doi: 10.1038/s41467-020-20139-7.

[17] Eswaran J., Cyanam D., Mudvari P., Reddy S.D.N., Pakala S.B., Nair S.S., Florea L., *et al.*: Transcriptomic landscape of breast cancers through mRNA sequencing, *Scientific Reports*, vol. 2(1), p. 264, 2012. doi: 10.1038/srep00264.

[18] Fröhlich H., Balling R., Beerenwinkel N., Kohlbacher O., Kumar S., Lengauer T., Maathuis M.H., *et al.*: From hype to reality: data science enabling personalized medicine, *BMC Medicine*, vol. 16, pp. 1–15, 2018. doi: 10.1186/s12916-018-1122-7.

[19] Gosak M., Markovič R., Dolenšek J., Rupnik M.S., Marhl M., Stožer A., Perc M.: Network science of biological systems at different scales: A review, *Physics of Life Reviews*, vol. 24, pp. 118–135, 2018. doi: 10.1016/j.plrev.2017.11.003.

[20] Haberman Y., Schirmer M., Dexheimer P.J., Karns R., Braun T., Kim M.O., Walters T.D., *et al.*: Age-of-diagnosis dependent ileal immune intensification and reduced alpha-defensin in older versus younger pediatric Crohn Disease patients despite already established dysbiosis, *Mucosal Immunology*, vol. 12(2), pp. 491–502, 2019. doi: 10.1038/s41385-018-0114-4.

[21] Haberman Y., Tickle T.L., Dexheimer P.J., Kim M.O., Tang D., Karns R., Baldassano R.N., *et al.*: Erratum: Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature (Journal of Clinical Investigation (2014) 124: 8 (3617-3633) DOI: 10.1172/JCI75436), *Journal of Clinical Investigation*, vol. 125(3), p. 1363, 2015. doi: 10.1172/JCI79657.

[22] Howell K.J., Kraiczy J., Nayak K.M., Gasparetto M., Ross A., Lee C., Mak T.N., *et al.*: DNA Methylation and Transcription Patterns in Intestinal Epithelial Cells From Pediatric Patients With Inflammatory Bowel Diseases Differentiate Disease Subtypes and Associate With Outcome, *Gastroenterology*, vol. 154(3), pp. 585–598, 2018. doi: 10.1053/j.gastro.2017.10.007.

[23] Hütt M.T.: Understanding genetic variation – the value of systems biology, *British Journal of Clinical Pharmacology*, vol. 77(4), pp. 597–605, 2014. doi: 10.1111/bcp.12266.

[24] Ideker T., Krogan N.J.: Differential network biology, *Molecular Systems Biology*, vol. 8(1), p. 565, 2012. doi: 10.1038/msb.2011.99.

[25] Ilott N.E., Neyazi M., Arancibia-Cárcamo C.V., Powrie F., Geremia A., Investigators O.T.G.U., *et al.*: Tissue-dependent transcriptional and bacterial associations in primary sclerosing cholangitis-associated inflammatory bowel disease, *Wellcome Open Research*, vol. 6, 2021. doi: 10.12688/wellcomeopenres.16901.1.

[26] Jablonski K.P., Carron L., Mozziconacci J., Forné T., Hütt M.T., Lesne A.: Contribution of 3D genome topological domains to genetic risk of cancers: a genome-wide computational study, *Human Genomics*, vol. 16(1), 2, 2022. doi: 10.1186/s40246-022-00375-2.

[27] Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O., Tunyasuvunakool K., *et al.*: Highly accurate protein structure prediction with AlphaFold, *Nature*, vol. 596(7873), pp. 583–589, 2021. doi: 10.1038/s41586-021-03819-2.

[28] Ker J., Wang L., Rao J., Lim T.: Deep learning applications in medical image analysis, *IEEE Access*, vol. 6, pp. 9375–9389, 2017. doi: 10.1109/access.2017.2788044.

[29] Kharchenko P., Church G.M., Vitkup D.: Expression dynamics of a cellular metabolic network, *Molecular Systems Biology*, vol. 1(1), 2005.0016, 2005. doi: 10.1038/msb4100023.

[30] Knecht C., Fretter C., Rosenstiel P., Krawczak M., Hütt M.T.: Distinct metabolic network states manifest in the gene expression profiles of pediatric inflammatory bowel disease patients and controls, *Scientific Reports*, vol. 6(32584), 2016. doi: 10.1038/srep32584.

[31] Kosmidis K., Jablonski K.P., Muskhelishvili G., Hütt M.T.: Chromosomal origin of replication coordinates logically distinct types of bacterial genetic regulation, *npj Systems Biology and Applications*, vol. 6(1), pp. 1–9, 2020. doi: 10.1038/s41540-020-0124-1.

[32] Krijger P.H.L., De Laat W.: Regulation of disease-associated gene expression in the 3D genome, *Nature Reviews Molecular Cell Biology*, vol. 17(12), pp. 771–782, 2016. doi: 10.1038/nrm.2016.138.

[33] Ktena I., Wiles O., Albuquerque I., Rebuffi S.A., Tanno R., Roy A.G., Azizi S., *et al.*: Generative models improve fairness of medical classifiers under distribution shifts, *Nature Medicine*, pp. 1–8, 2024. doi: 10.1038/s41591-024-02838-6.

[34] Loscalzo J., Barabasi A.L.: Systems biology and the future of medicine, *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, vol. 3(6), pp. 619–627, 2011. doi: 10.1002/wsbm.144.

[35] Love M., Anders S., Huber W.: Differential analysis of count data – the DESeq2 package, *Genome Biology*, vol. 15, 550, 2014. doi: 10.1186/s13059-014-0550-8.

[36] Love M.I., Huber W., Anders S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, *Genome Biology*, vol. 15, pp. 1–21, 2014. doi: 10.1186/s13059-014-0550-8.

[37] Ma H., Zeng A.P.: Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms, *Bioinformatics*, vol. 19(2), pp. 270–277, 2003. doi: 10.1093/bioinformatics/19.2.270.

[38] Ma H.W., Zeng A.P.: The connectivity structure, giant strong component and centrality of metabolic networks, *Bioinformatics*, vol. 19(11), pp. 1423–1430, 2003. doi: 10.1093/bioinformatics/btg177.

[39] Marigorta U.M., Denson L.A., Hyams J.S., Mondal K., Prince J., Walters T.D., Griffiths A., *et al.*: Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease, *Nature Genetics*, vol. 49(10), pp. 1517–1521, 2017. doi: 10.1038/ng.3936.

[40] Marr C., Geertz M., Hütt M.T., Muskhelishvili G.: Dissecting the logical types of network control in gene expression profiles, *BMC Systems Biology*, vol. 2(1), pp. 1–9, 2008. doi: 10.1186/1752-0509-2-18.

[41] Meyer S., Reverchon S., Nasser W., Muskhelishvili G.: Chromosomal organization of transcription: in a nutshell, *Current Genetics*, vol. 64, pp. 555–565, 2018.

[42] Mo A., Marigorta U.M., Arafat D., Chan L.H.K., Ponder L., Jang S.R., Prince J., *et al.*: Disease-specific regulation of gene expression in a comparative analysis of juvenile idiopathic arthritis and inflammatory bowel disease, *Genome Medicine*, vol. 10, 48, 2018. doi: 10.1186/s13073-018-0558-x.

[43] Moncrieffe H., Bennett M.F., Tsoras M., Luyrink L.K., Johnson A.L., Xu H., Dare J., *et al.*: Transcriptional profiles of JIA patient blood with subsequent poor response to methotrexate, *Rheumatology (United Kingdom)*, vol. 56(9), pp. 1542–1551, 2017. doi: 10.1093/rheumatology/kex206.

[44] Nyczka P., Hütt M.T.: Generative network model of transcriptome patterns in disease cohorts with tunable signal strength, *Physical Review Research*, vol. 2(3), 033130, 2020. doi: 10.1103/physrevresearch.2.033130.

[45] Nyczka P., Hütt M.T., Lesne A.: Inferring pattern generators on networks, *Physica A*, vol. 566, 125631, 2021. doi: 10.1016/j.physa.2020.125631.

[46] Oughtred R., Rust J., Chang C., Breitkreutz B.J., Stark C., Willems A., Boucher L., *et al.*: The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions, *Protein Science*, vol. 30(1), pp. 187–200, 2021. doi: 10.1002/PRO.3978.

[47] Palsson B.Ø.: *Systems biology: properties of reconstructed networks*, Cambridge university press, 2006. doi: 10.1017/cbo9780511790515.

[48] Papatheodorou I., Fonseca N.A., Keays M., Tang Y.A., Barrera E., Bazant W., Burke M., *et al.*: Expression Atlas: gene and protein expression across multiple studies and organisms, *Nucleic Acids Research*, vol. 46(D1), pp. D246–D251, 2018.

[49] Peck B.C., Weiser M., Lee S.E., Gipson G.R., Iyer V.B., Sartor R.B., Herfarth H.H., *et al.*: MicroRNAs classify different disease behavior phenotypes of Crohn's disease and may have prognostic utility, *Inflammatory Bowel Diseases*, vol. 21(9), pp. 2178–2187, 2015. doi: 10.1097/MIB.0000000000000478.

[50] Peeters J.G., Vervoort S.J., Tan S.C., Mijnheer G., de Roock S., Vastert S.J., Nieuwenhuis E.E., *et al.*: Inhibition of Super-Enhancer Activity in Autoinflammatory Site-Derived T Cells Reduces Disease-Associated Gene Expression, *Cell Reports*, vol. 12(12), pp. 1986–1996, 2015. doi: 10.1016/j.celrep.2015.08.046.

[51] Peters L.A., Perrigoue J., Mortha A., Iuga A., Song W.M., Neiman E.M., Llewellyn S.R., *et al.*: A functional genomics predictive network model identifies regulators of inflammatory bowel disease, *Nature Genetics*, vol. 49(10), pp. 1437–1449, 2017. doi: 10.1038/ng.3947.

[52] Prudencio M., Belzil V.V., Batra R., Ross C.A., Gendron T.F., Pregent L.J., Murray M.E., *et al.*: Distinct brain transcriptome profiles in C9orf72-associated and sporadic ALS, *Nature Neuroscience*, vol. 18(8), pp. 1175–1182, 2015. doi: 10.1038/nn.4065.

[53] Qiao L., Khalilimeybodi A., Linden-Santangeli N.J., Rangamani P.: The evolution of systems biology and systems medicine: From mechanistic models to uncertainty quantification, *arXiv preprint arXiv:240805395*, 2024. doi: 10.1146/annurev-bioeng-102723-065309.

[54] Quraishi M.N., Acharjee A., Beggs A.D., Horniblow R., Tselepis C., Gkoutos G., Ghosh S., *et al.*: A Pilot Integrative Analysis of Colonic Gene Expression, Gut Microbiota, and Immune Infiltration in Primary Sclerosing Cholangitis-Inflammatory Bowel Disease: Association of Disease With Bile Acid Pathways, *Journal of Crohn's and Colitis*, vol. 14(7), pp. 935–947, 2020. doi: 10.1093/ecco-jcc/jjaa021.

[55] Rajpurkar P., Lungren M.P.: The current and future state of AI interpretation of medical images, *New England Journal of Medicine*, vol. 388(21), pp. 1981–1990, 2023. doi: 10.1056/nejmra2301725.

[56] Ricarte-Filho J.C., Li S., Garcia-Rendueles M.E., Montero-Conde C., Voza F., Knauf J.A., Heguy A., *et al.*: Identification of kinase fusion oncogenes in post-Chernobyl radiation-induced thyroid cancers, *Journal of Clinical Investigation*, vol. 123(11), pp. 4935–4944, 2013. doi: 10.1172/JCI69766.

[57] Sabari B.R., Dall'Agnese A., Young R.A.: Biomolecular condensates in the nucleus, *Trends in Biochemical Sciences*, vol. 45(11), pp. 961–977, 2020. doi: 10.1016/j.tibs.2020.06.007.

[58] Sareen D., O'Rourke J.G., Meera P., Muhammad A.K.M.G., Grant S., Simpkinson M., Bell S., *et al.*: Targeting RNA Foci in iPSC-Derived Motor Neurons from ALS Patients with a C9ORF72 Repeat Expansion, *Science Translational Medicine*, vol. 5(208), pp. 208ra149–208ra149, 2013. doi: 10.1126/scitranslmed.3007529.

[59] Sarkans U., Gostev M., Athar A., Behrangi E., Melnichuk O., Ali A., Minguet J., *et al.*: The BioStudies database-one stop shop for all data supporting a life sciences study, *Nucleic Acids Research*, vol. 46(D1), pp. D1266–D1270, 2018.

[60] Sastry A.V., Gao Y., Szubin R., Hefner Y., Xu S., Kim D., Choudhary K.S., *et al.*: The *Escherichia coli* transcriptome mostly consists of independently regulated modules, *Nature Communications*, vol. 10(1), 5536, 2019.

[61] Schlicht K., Nyczka P., Caliebe A., Freitag-Wolf S., Claringbould A., Franke L., Võsa U., *et al.*: The metabolic network coherence of human transcriptomes is associated with genetic variation at the cadherin 18 locus, *Human Genetics*, vol. 138(4), pp. 375–388, 2019. doi: 10.1007/s00439-019-01994-x.

[62] Sharma A., Lysenko A., Jia S., Boroevich K.A., Tsunoda T.: Advances in AI and machine learning for predictive medicine, *Journal of Human Genetics*, pp. 1–11, 2024. doi: 10.1038/s10038-024-01231-y.

[63] Shin Y., Chang Y.C., Lee D.S., Berry J., Sanders D.W., Ronceray P., Wingreen N.S., *et al.*: Liquid nuclear condensates mechanically sense and restructure the genome, *Cell*, vol. 175(6), pp. 1481–1491, 2018. doi: 10.1016/j.cell.2019.02.025.

[64] Sonnenschein N., Geertz M., Muskhelishvili G., Hütt M.T.: Analog regulation of metabolic demand, *BMC Systems Biology*, vol. 5(1), 40, 2011. doi: 10.1186/1752-0509-5-40.

[65] Sonnenschein N., Golib Dzib J.F., Lesne A., Eilebrecht S., Boulkroun S., Zennaro M.C., Benecke A., *et al.*: A network perspective on metabolic inconsistency, *BMC Systems Biology*, vol. 6, pp. 1–13, 2012. doi: 10.1186/1752-0509-6-41.

[66] Szklarczyk D., Gable A.L., Nastou K.C., Lyon D., Kirsch R., Pyysalo S., Doncheva N.T., *et al.*: The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets, *Nucleic Acids Research*, vol. 49(D1), pp. D605–D612, 2021. doi: 10.1093/NAR/GKAA1074.

[67] Travers A., Muskhelishvili G.: DNA supercoiling – a global transcriptional regulator for enterobacterial growth?, *Nature Reviews Microbiology*, vol. 3(2), pp. 157–169, 2005. doi: 10.1038/nrmicro1088.

[68] Travers A., Muskhelishvili G., Thompson J.: DNA information: from digital code to analogue structure, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 370(1969), pp. 2960–2986, 2012. doi: 10.1098/rsta.2011.0231.

[69] VanDussen K.L., Stojmirović A., Li K., Liu T.C., Kimes P.K., Muegge B.D., Simpson K.F., *et al.*: Abnormal Small Intestinal Epithelial Microvilli in Patients With Crohn's Disease, *Gastroenterology*, vol. 155(3), pp. 815–828, 2018. doi: 10.1053/j.gastro.2018.05.028.

[70] Wu M., Chen Y., Xia H., Wang C., Tan C.Y., Cai X., Liu Y., *et al.*: Transcriptional and proteomic insights into the host response in fatal COVID-19 cases, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117(45), pp. 28336–28343, 2020. doi: 10.1073/PNAS.2018030117/SUPPL_FILE/PNAS.2018030117.SD01.XLSX.

Affiliations

Eda Cakir
Constructor University, School of Science, Campus Ring 1 28759 Bremen, Germany, ecakir@constructor.university

Marc-Thorsten Hütt
Constructor University, School of Science, Campus Ring 1 28759 Bremen, Germany, mhuett@constructor.university

Received: 8.03.2025
Revised: 12.03.2025
Accepted: 12.03.2025