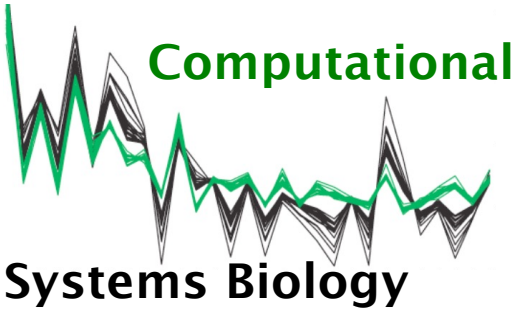
CONSTRUCTOR UNIVERSITY

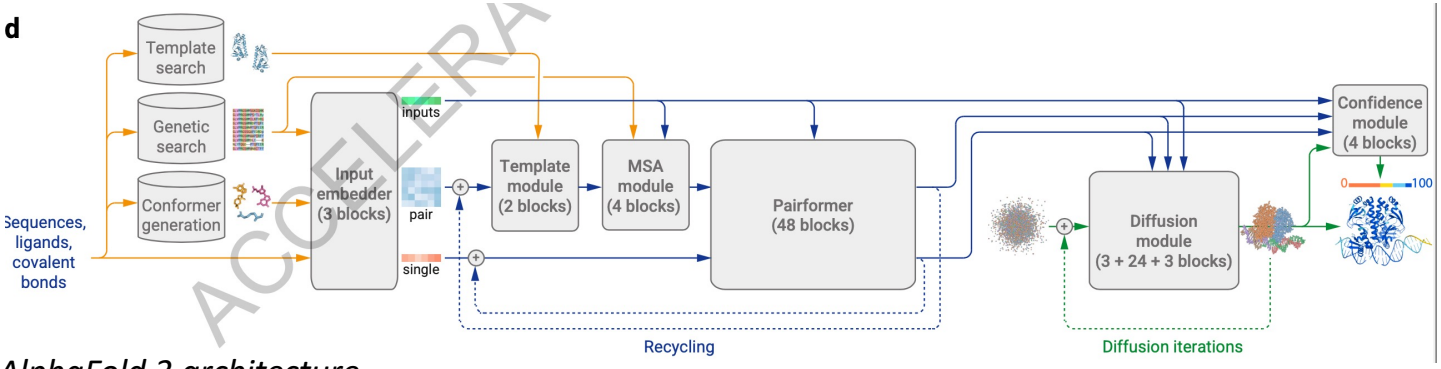
Machine learning vs. network science: A comparison of two paradigms for the interpretation of high-throughput data in biology and medicine

Marc-Thorsten Hütt
Constructor University Bremen
previously Jacobs University

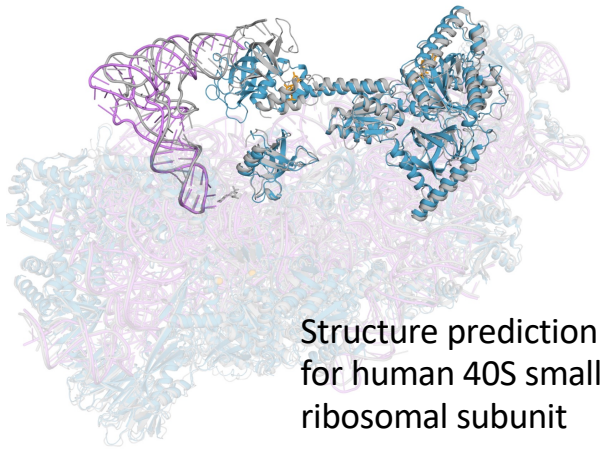
Bremen, Germany



Machine learning in Medicine: two famous examples

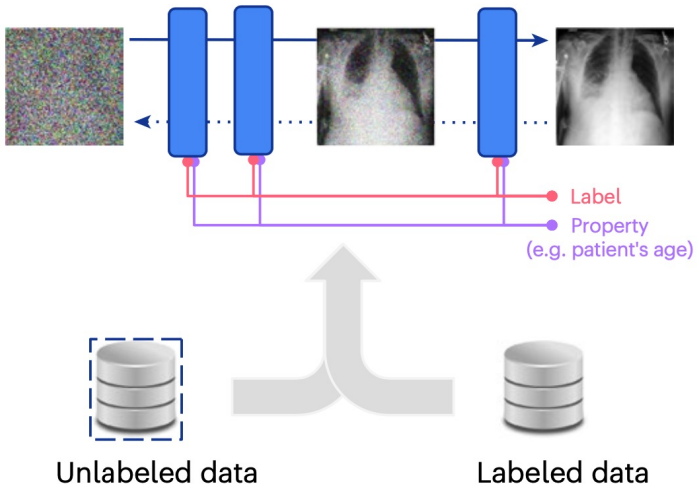


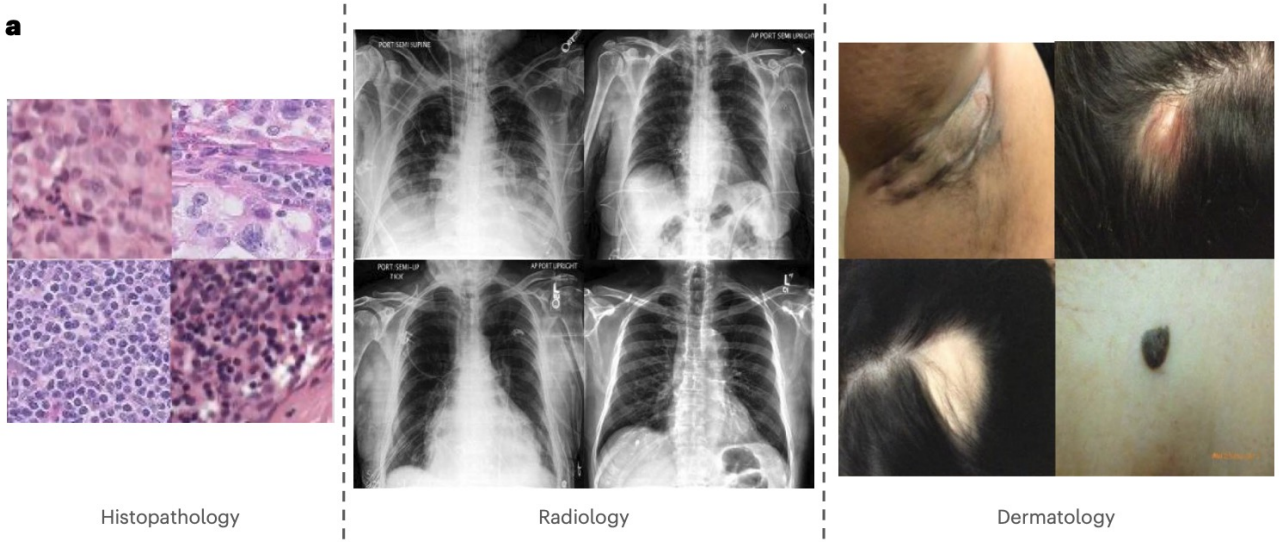
AlphaFold 3 architecture



Structure prediction for human 40S small ribosomal subunit

Taken from: Abramson,... & Jumper (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 1-3.





Taken from: Ktena, ... & Gowal (2024). Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, 1-8.

Machine learning in Medicine beyond AlphaFold and image classifiers

'Omics data' (or high-throughput data)

Ó. Rolfsson and B. O. Palsson Thoughts & Opinion

Decoding the jargon of bottom-up metabolic systems biology

Óttar Rolfsson[1)2)] and Bernhard O. Palsson[1)]* Bioessays 37: 588–591, © 2015

"High throughput data have changed biochemical research such that new skillsets are now required of the modern 21st century life scientist.

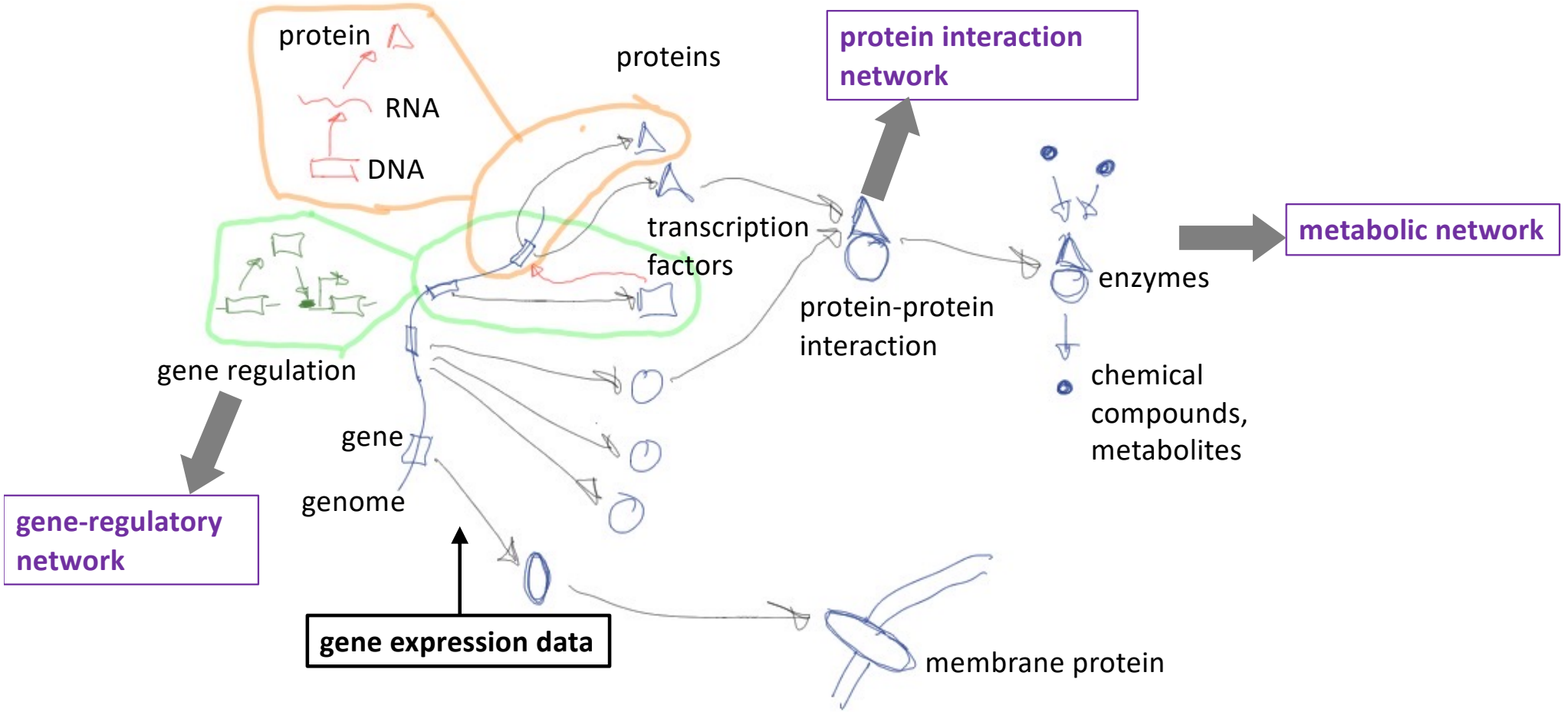
Biochemical research can now be carried out in a manner that is outside the scope of traditional biochemistry and biology methods, and more to the taste of computer scientists and statisticians.

Indeed, bioinformatics has emerged to specifically deal with the manipulation and analysis of these data."

Machine learning in Medicine beyond AlphaFold and image classifiers

Digression: What are **'omics' data** and in particular, **transcriptomics** (gene expression) data?

Gene activity, how it is created and how it fuels the diverse aspects of a cell is at the core of understanding a biological cell



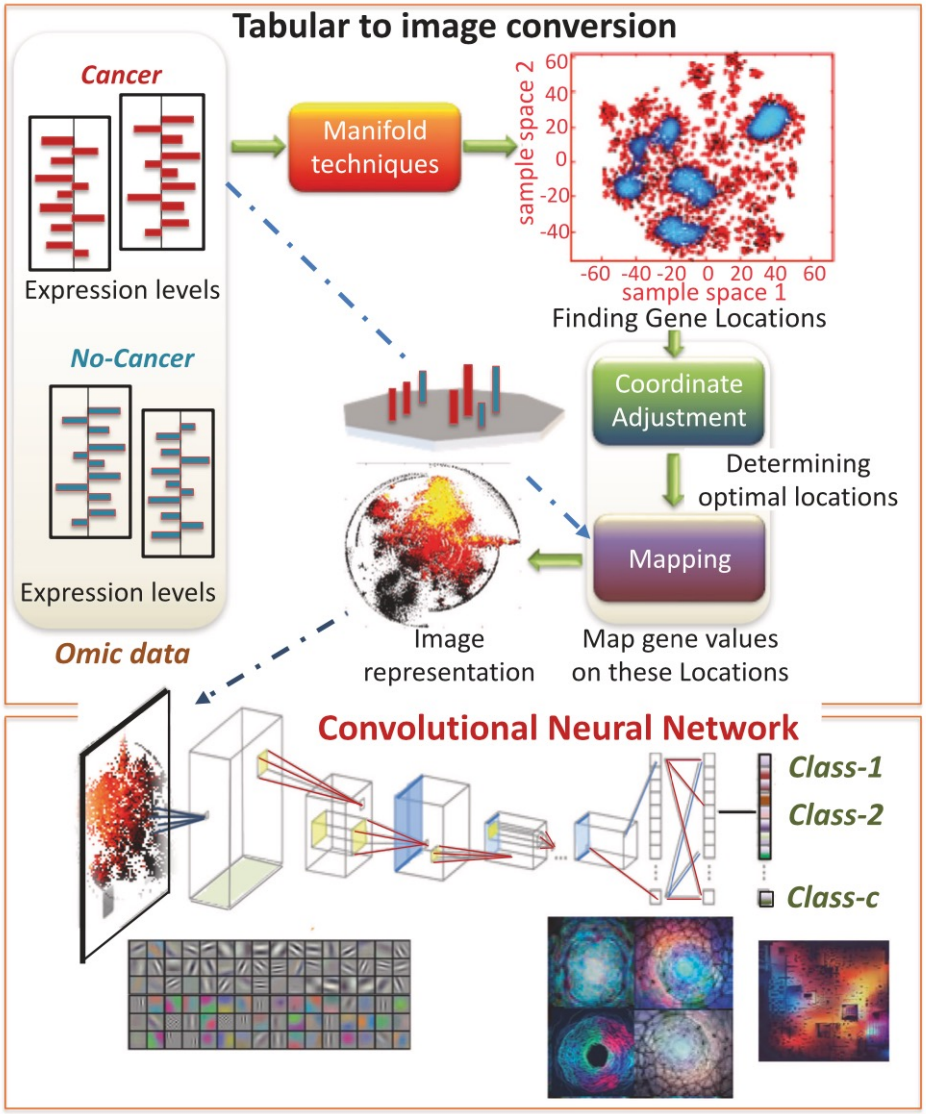
Machine learning in Medicine beyond AlphaFold and image classifiers

'Omics data' (or high-throughput data)

... one strategy: adapting image classifiers for analyzing gene expression data

- hard to interpret the results (features are not linked to biology)
- not (yet) very successful

> In conclusion, as we stand on the cusp of this analytical revolution in genomics, it is imperative to embrace these novel methodologies. Their potential to revolutionize our comprehension of biology, combined with profound clinical implications, cements their role as indispensable instruments in our endeavor to decode the intricacies of life.



Taken from: Sharma, ... & Tsunoda (2024). Advances in AI and machine learning for predictive medicine. *Journal of Human Genetics*, 1-11.

Machine learning in Medicine beyond AlphaFold and image classifiers

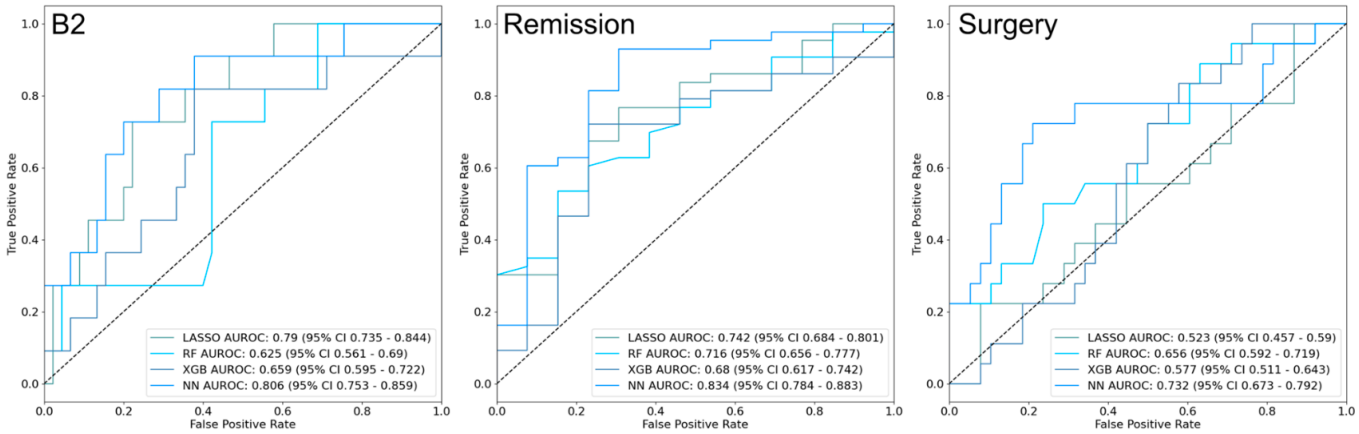
'Omics data' (or high-throughput data)

.... another strategy: dedicated ML devices extracting disease-related genes

Machine learning-based models were able to incorporate colonic gene expression and clinical characteristics to predict outcomes with high accuracy. Models showed an area under the receiver operating characteristic curve (AUROC) of 0.84 for strictures, 0.83 for remission, and 0.75 for surgery. Genes with potential prognostic importance for strictures (*REG1A*, *MMP3*, and *DUOX2)* were not identified in single gene differential analysis but were found to have strong contributions to predictive models.

- substantial variation with method choices

Clinical variables with gene expression



- typically very small datasets

After applying quality control, 56 CD patients with colon samples and 56 CD patients with ileum samples were included in the study cohort, while 46 non-IBD patients with colon samples and 46 non-IBD patients with ileum samples were used as controls.

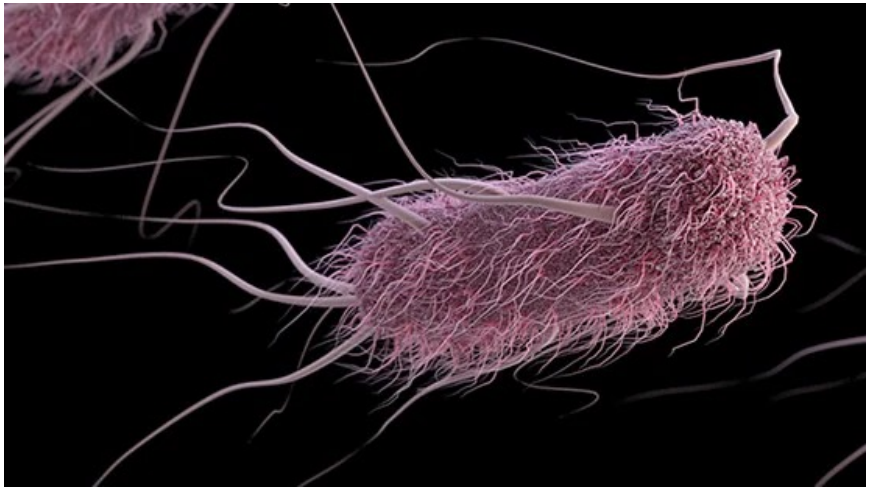
Taken from: Chen, ... & Sheikh (2024). Linking gene expression to clinical outcomes in pediatric Crohn's disease using machine learning. *Scientific Reports*, 14(1), 2667.

Machine learning in Medicine beyond AlphaFold and image classifiers

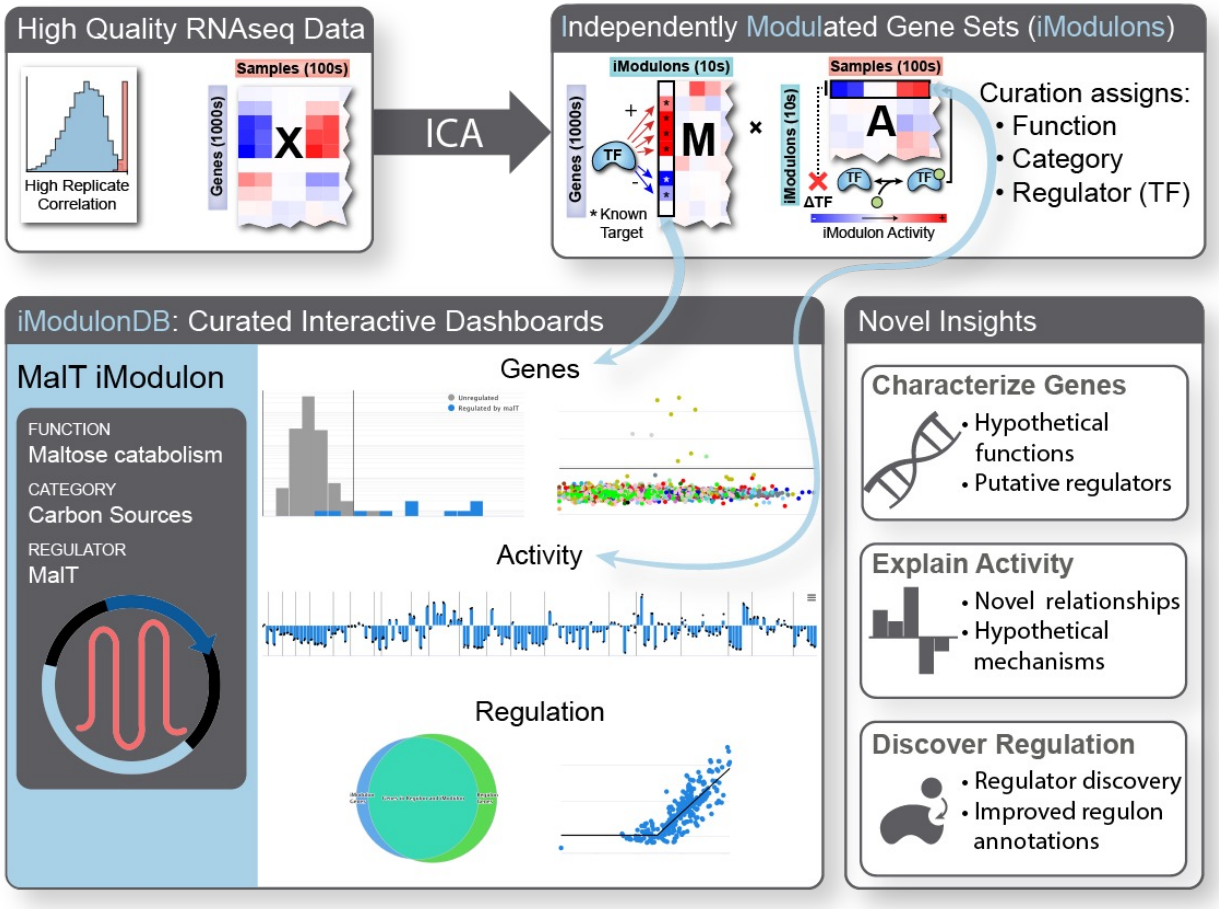
'Omics data' (or high-throughput data)

.... what about simpler organisms?

Gene expression data from a bacterium, *Escherichia coli*



Taken from: www.cdc.gov/ecoli/



Taken from:

Sastry,... & Palsson (2019). The *Escherichia coli* transcriptome mostly consists of independently regulated modules. *Nat. Commun.* 10, 1–14

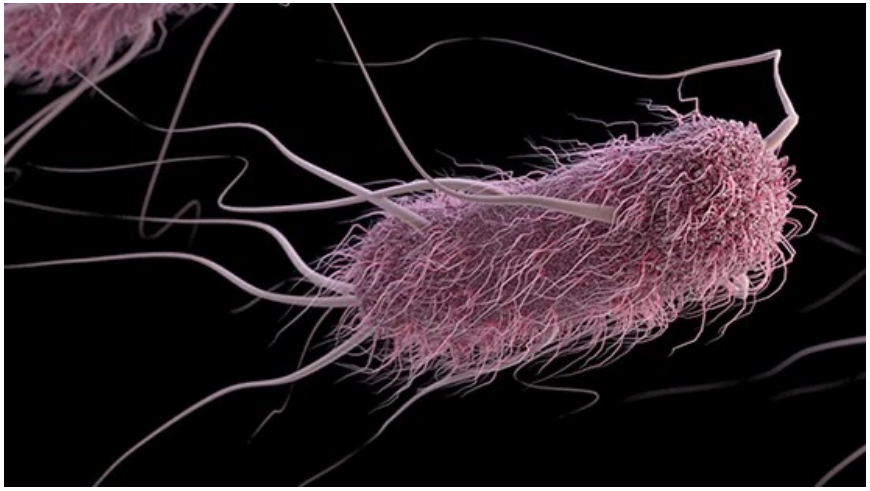
https://systemsbiology.ucsd.edu/imodulons

Machine learning in Medicine beyond AlphaFold and image classifiers

'Omics data' (or high-throughput data)

.... what about simpler organisms?

Gene expression data from a bacterium, *Escherichia coli*



Taken from: www.cdc.gov/ecoli/

- iModulon: gene group identified from patterns in transciptomic datasets
- Regulon: gene group regulated by the same transcriptional regulator (experimentally verified)
- Around 66% of the identified iModulons have significant overlaps with Regulons
- Reasons or biological significance of discrepancies are unclear
- No predictive framework

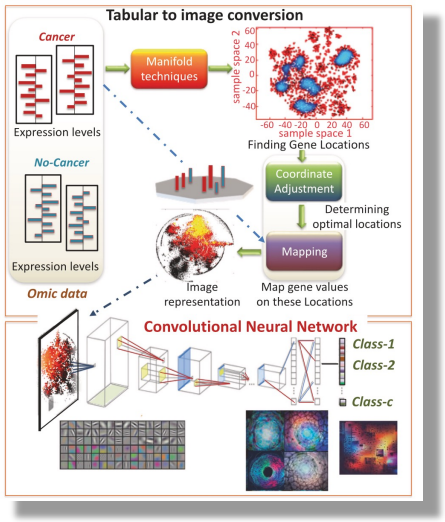
Taken from:

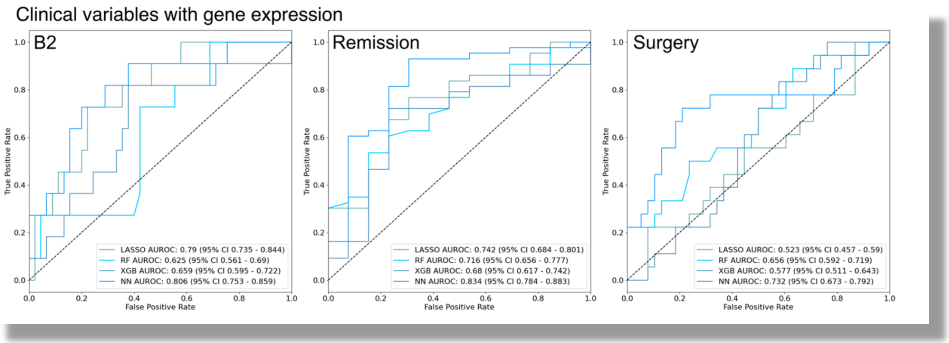
Sastry,... & Palsson (2019). The *Escherichia coli* transcriptome mostly consists of independently regulated modules. *Nat. Commun.* 10, 1–14

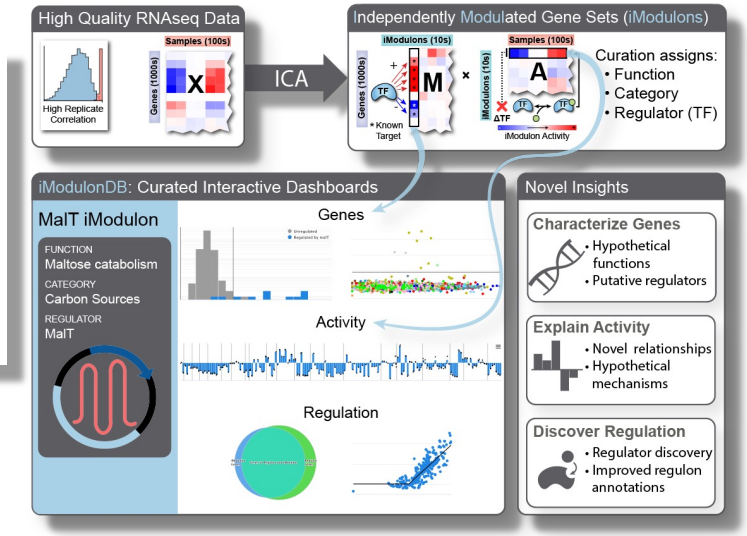
https://systemsbiology.ucsd.edu/imodulons

Machine learning in Medicine beyond AlphaFold and image classifiers

'Omics data' (or high-throughput data)







- **Current status: gene expression patterns (so far) cannot be simulated, explained or predicted**
- **We believe that this is due to the intrinsic complexity of biological systems**

Biological complexity in the context of modeling and algorithms

Biology operates via the interplay of analog (rather gradual) and digital (discrete, symbolic) information.

"The observation I wish to make is this: processes which go through the nervous system may, as I pointed out before, change their character from digital to analog, and back to digital, etc., repeatedly."

von Neumann, J. (1958) The Computer and the Brain. Yale University Press.

"In the future attention undoubtedly will be centered on the genome, and with greater appreciation of its significance as a highly sensitive organ of the cell, monitoring genomic activities and correcting common errors, sensing the unusual and unexpected events, and responding to them, often by restructuring the genome."

McClintock, B. (1983) Nobel prize lecture

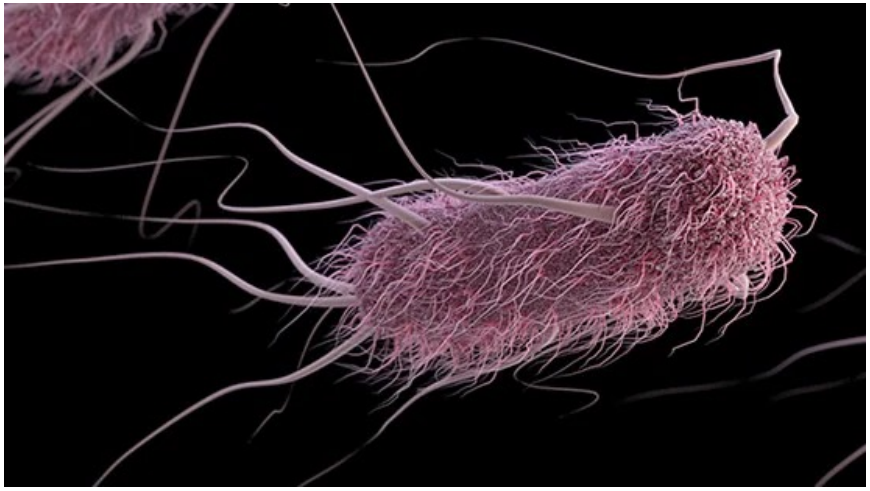
Examples

- Chromatin organization
- Ultrasensitivity in signaling
- Spikes in neurons
- ...

Mathematical and computational approaches (modeling, data analysis, machine learning) are challenged by this interplay of digital and analog information.

Biological complexity in the context of modeling and algorithms

Gene expression data from a bacterium, *Escherichia coli*



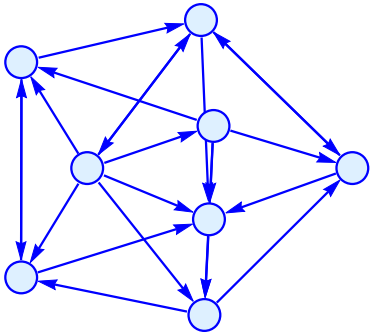
Taken from: www.cdc.gov/ecoli/

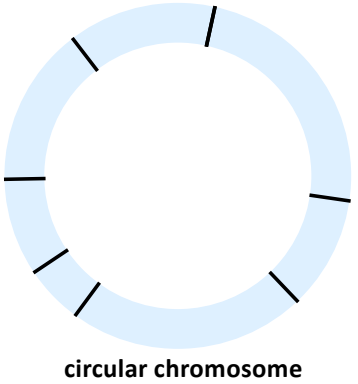
Key question:

Do patterns in the gene expression data reflect the interplay of analog and digital control?

Digital and analog control in bacterial gene regulation

Representations of the biological system



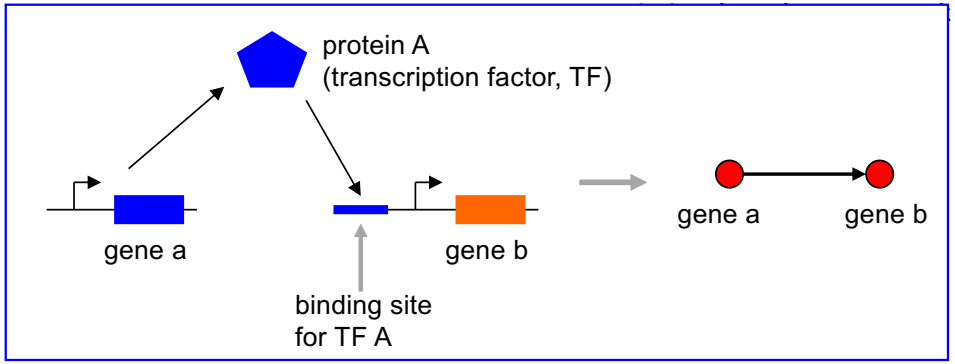


$G = (V, E)$,

$v_i, v_j \in V$ vertices (genes),

$(v_i, v_j) \in E$ directed edges,

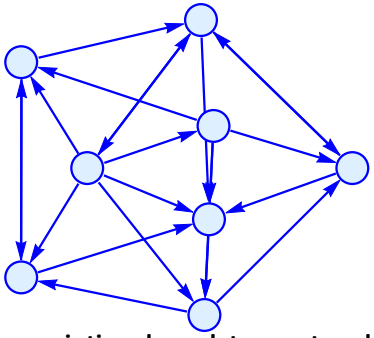
$v_i \rightarrow x_i$ coordinate on a (circular) 1D space



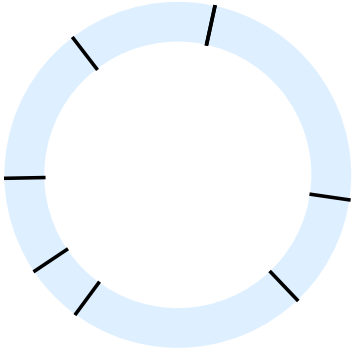
Digital and analog control in bacterial gene regulation

Representations of the biological system

digital



transcriptional regulatory network



circular chromosome

analog

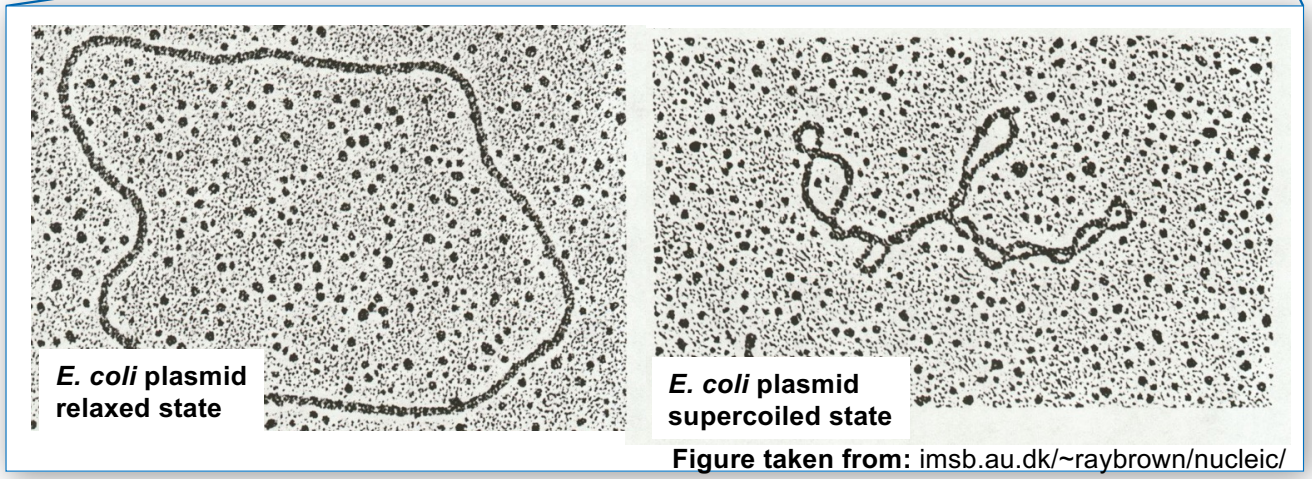
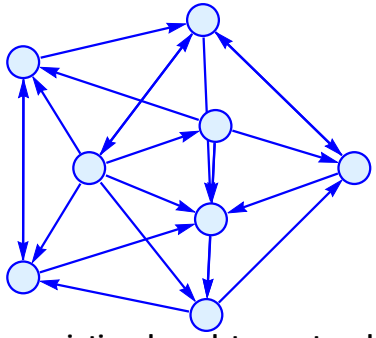


Figure taken from: imsb.au.dk/~raybrown/nucleic/

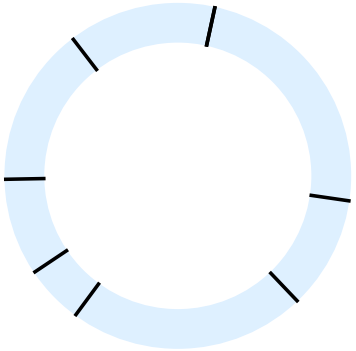
Digital and analog control in bacterial gene regulation

Representations of the biological system

digital



transcriptional regulatory network

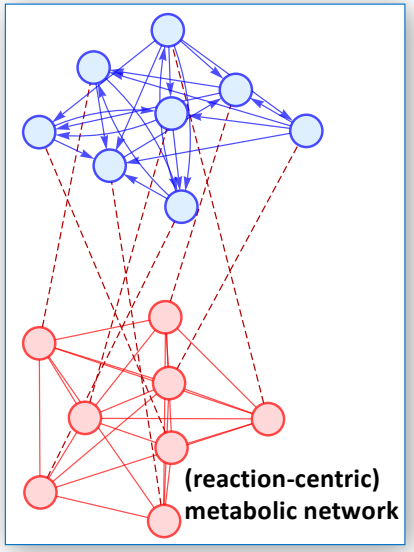


circular chromosome

analog

Side remark: Additional complexity

The gene regulatory network is not acting alone



→ Theory of interdependent networks

Klosik, Grimbs, Bornholdt and Hütt (2017).
Nature Communications 8, 534.

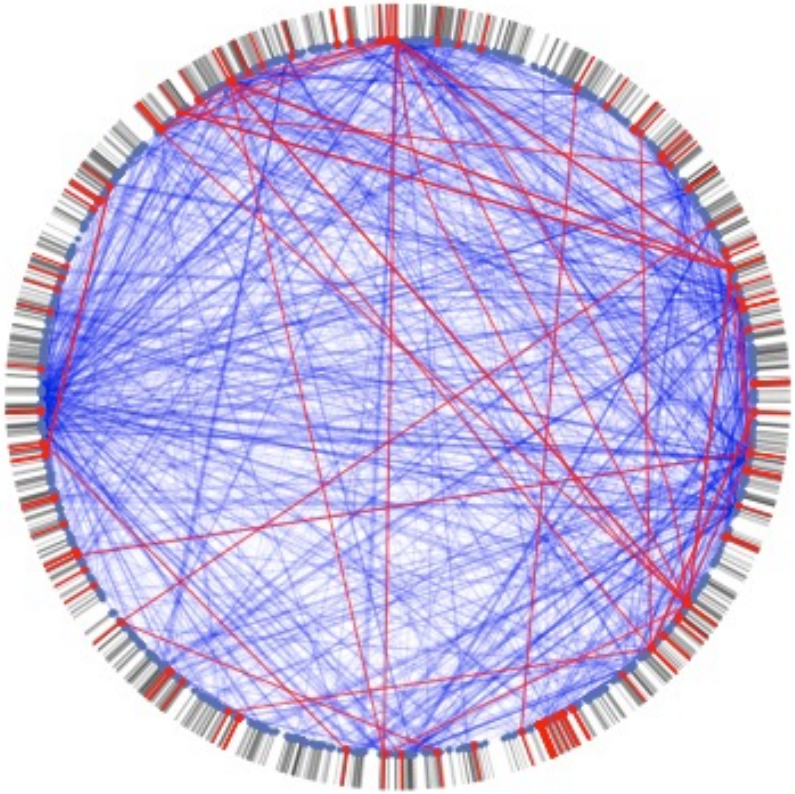
Grimbs, Klosik, Bornholdt, Hütt (2019).
PLoS Computational Biology, 15, e1006962.

Digital and analog control in bacterial gene regulation

Evidence 1: Agreement of gene expression changes with network and chromosome

comparison of expression changes between high and low supercoiling

comparison with TRN



effective TRN

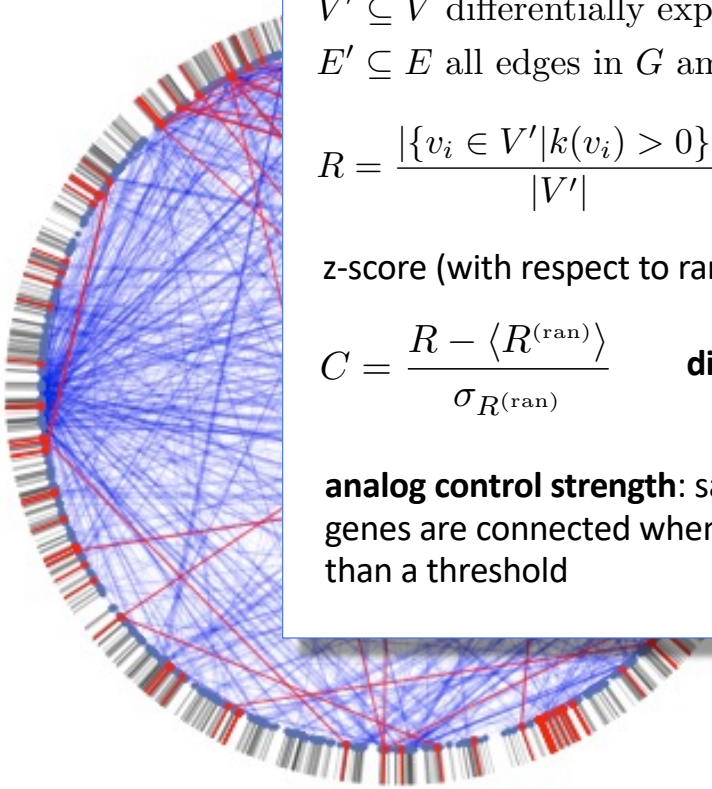
Marr, Geertz, Hütt, Muskhelishvili (2008) BMC Systems Biology 2, 18

Digital and analog control in bacterial gene regulation

Evidence 1: Agreement of gene expression changes with network and chromosome

comparison of expression changes between high and low supercoiling

comparison with TRN



effective TRN

$G' = (V', E')$,

$V' \subseteq V$ differentially expressed genes

$E' \subseteq E$ all edges in G among vertices in V'

$$R = \frac{|\{v_i \in V' | k(v_i) > 0\}|}{|V'|}, \quad k(v_i) \text{ degree of node } v_i$$

z-score (with respect to random vertex sets)

$$C = \frac{R - \langle R^{(\mathrm{ran})} \rangle}{\sigma_{R^{(\mathrm{ran})}}}$$ **digital control strength**

analog control strength: same, but for a network where genes are connected when the (1D) distance is smaller than a threshold

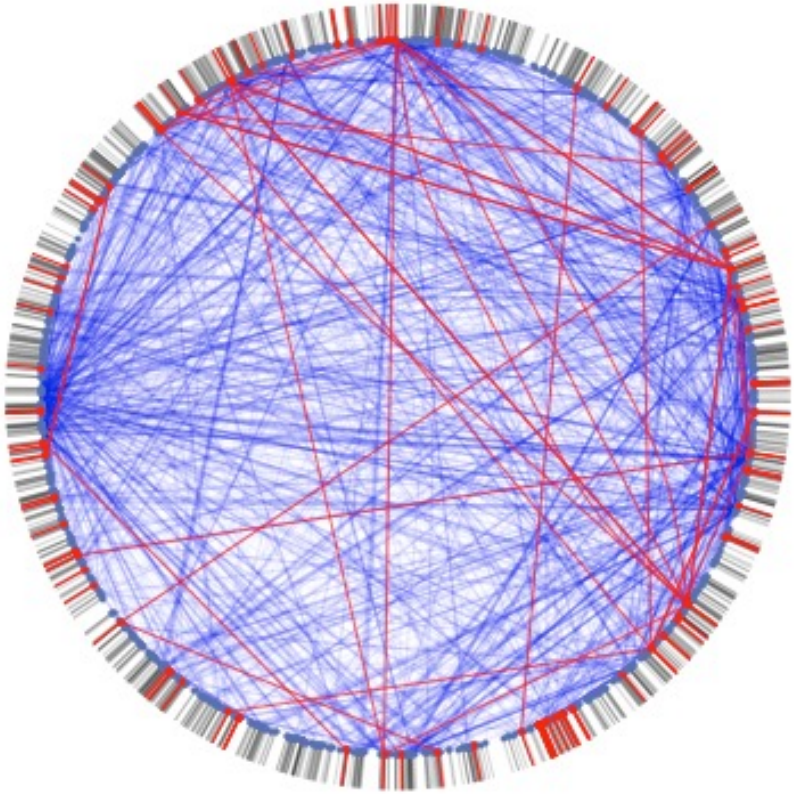
Marr, Geertz, Hütt, Muskhelishvili (2008) BMC Systems Biology 2, 18

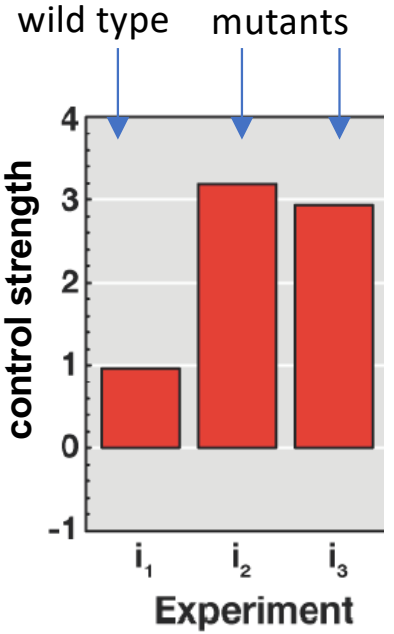
Digital and analog control in bacterial gene regulation

Evidence 1: Agreement of gene expression changes with network and chromosome

comparison of expression changes between high and low supercoiling

comparison with TRN





effective TRN

Marr, Geertz, Hütt, Muskhelishvili (2008) BMC Systems Biology 2, 18

Digital and analog control in bacterial gene regulation

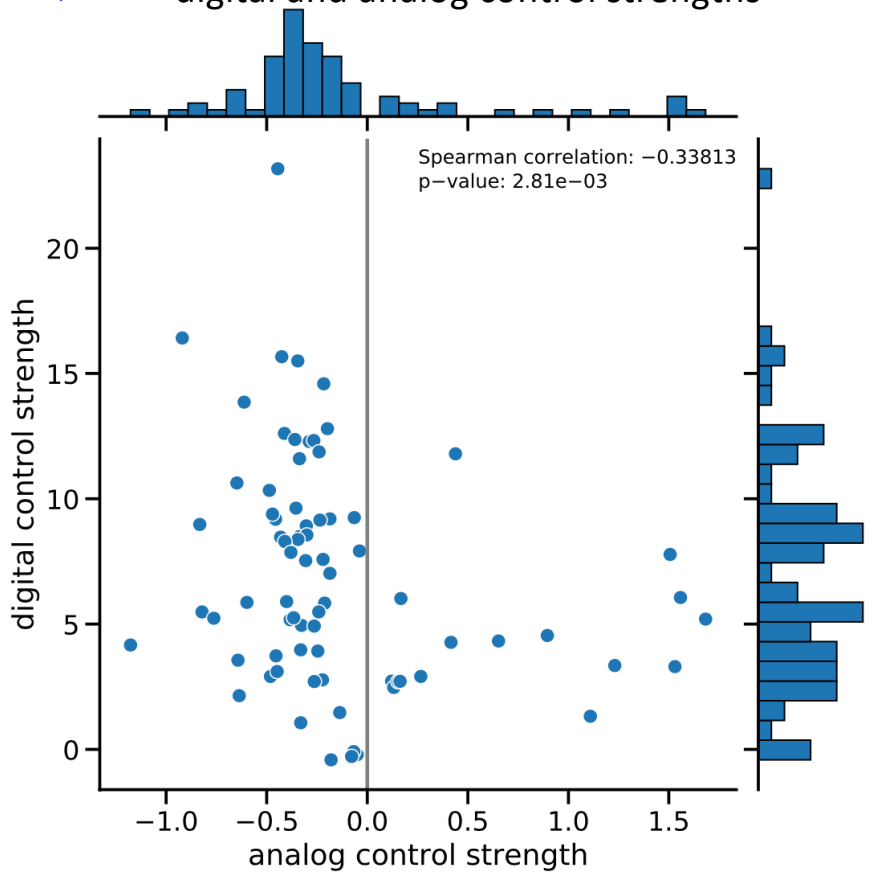
Evidence 1: Agreement of gene expression changes with network and chromosome

104 RNA-Seq data sets for various experimental conditions

→ computation of differentially expressed genes

→ digital and analog control strengths

colombos



- robust against variation of parameters (gene proximity threshold, significance threshold of expression levels)
- valid on the gene and the operon levels
- also found for microarray data

Kosmidis, Jablonski, Muskhelishvili, Hütt (2020) npj Systems Biology and Applications 6:5

Digital and analog control in bacterial gene regulation

Evidence 2: Machine learning classification of gene expression data

Decision Trees

Features	digital vs. analog
Position relative to Ori	unclear
crp binding sites density	analog
hns binding sites density	analog
fis binding sites density	analog
gene proximity network neighbors	analog
TRN regulators	digital
hns is direct regulator	digital
fis is direct regulator	digital
crp is direct regulator	digital

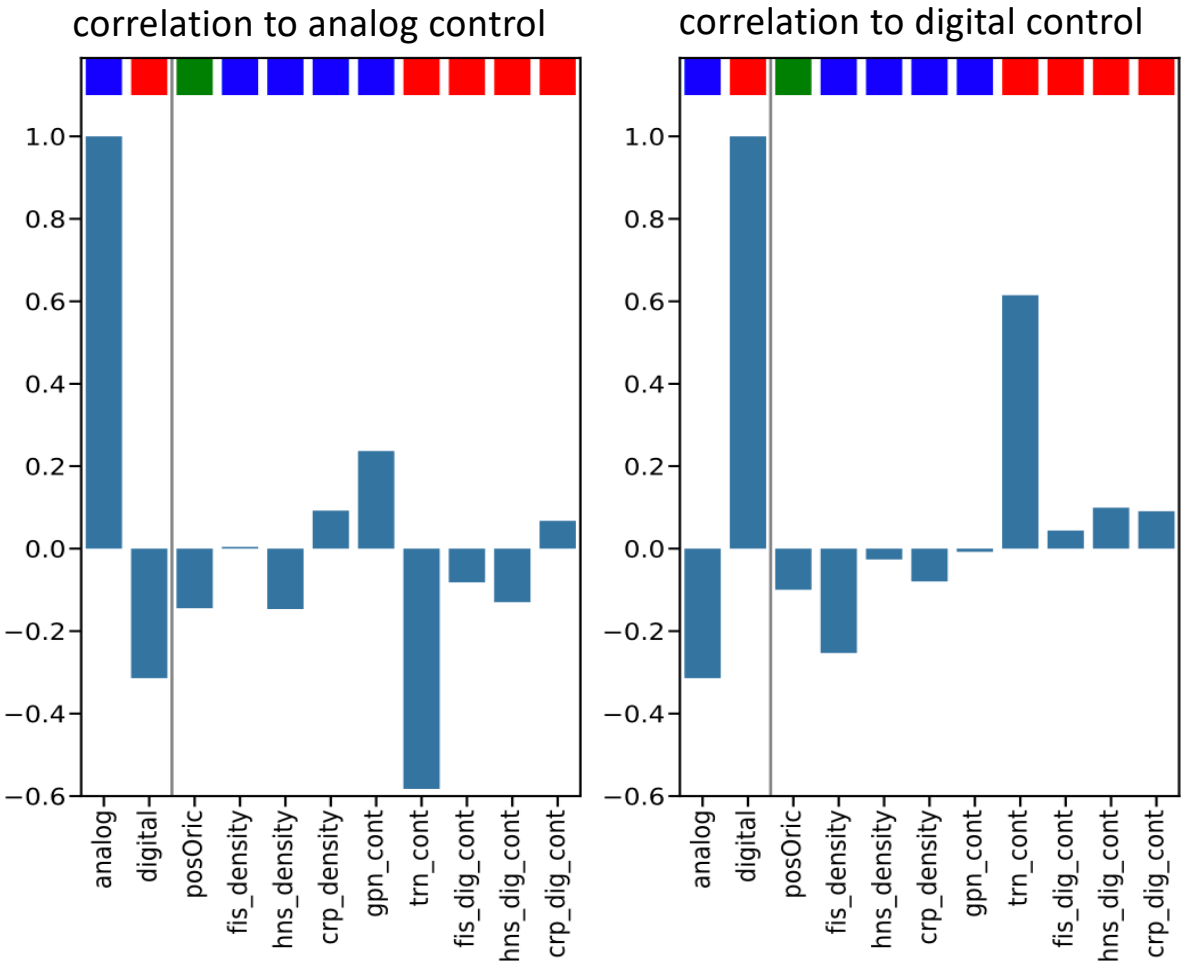
unpacking some 'jargon'

crp, hns, fis are hubs of the transcriptional regulatory network; they also bind at other places of the chromosome and influence 3D organization

Ori is shorthand for 'origin of replication', i.e., the point, the chromosome is duplicated for cell division.

Digital and analog control in bacterial gene regulation

Evidence 2: Machine learning classification of gene expression data

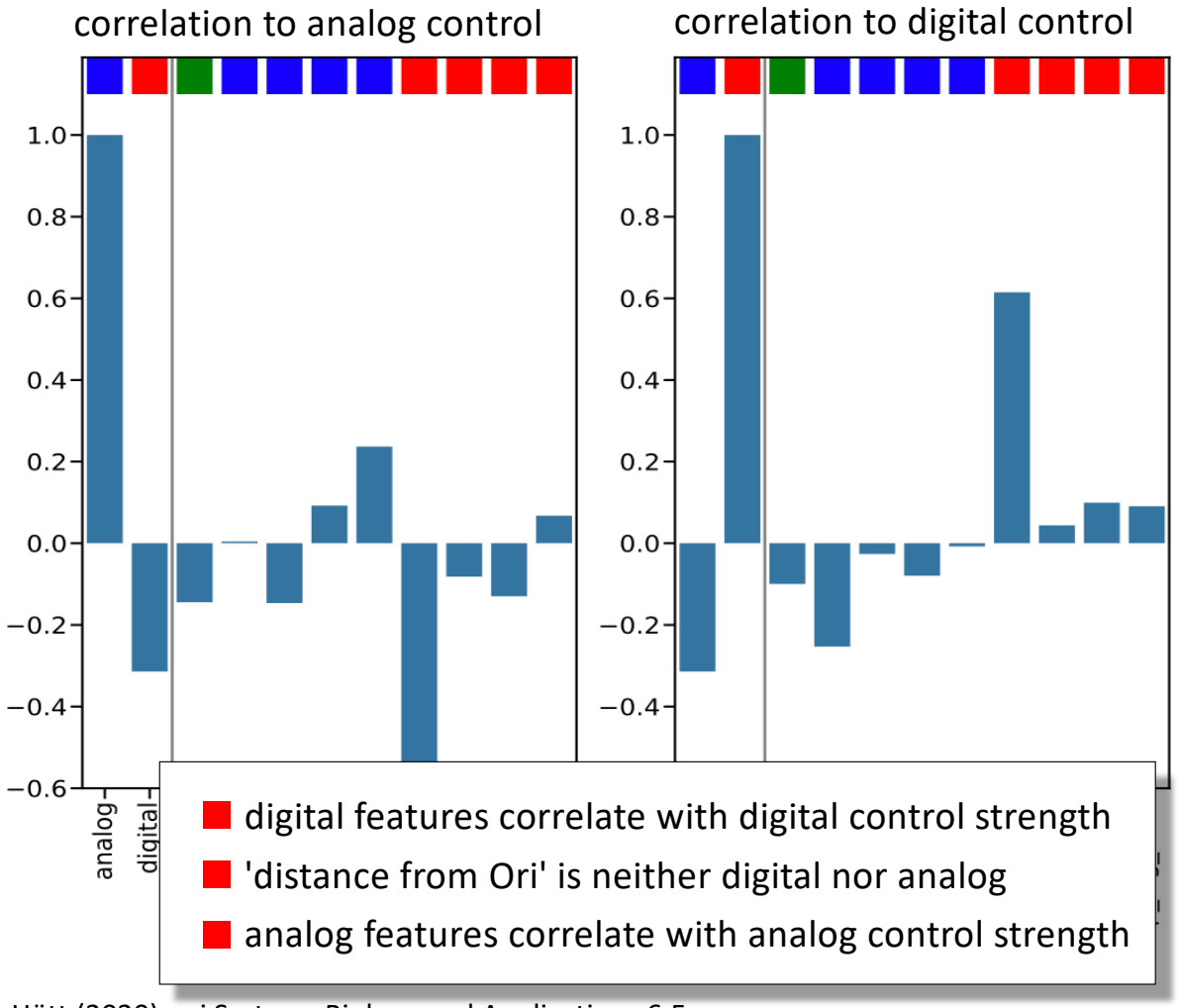


Feature importance for RNA-Seq data

Kosmidis, Jablonski, Muskhelishvili, Hütt (2020) npj Systems Biology and Applications 6:5

Digital and analog control in bacterial gene regulation

Evidence 2: Machine learning classification of gene expression data



Feature importance for RNA-Seq data

Kosmidis, Jablonski, Muskhelishvili, Hütt (2020) npj Systems Biology and Applications 6:5

Digital and analog control in bacterial gene regulation

Evidence 3: 'Wiring economy' in the transcriptional regulatory network

Motivated by

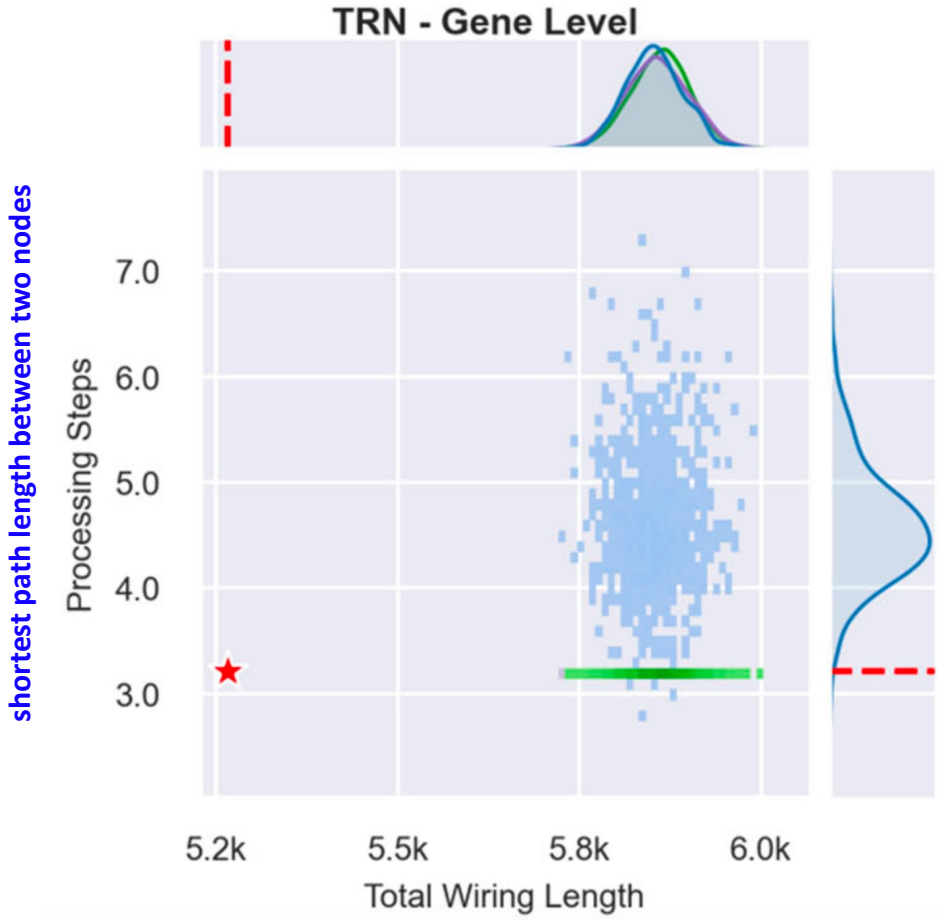
Chen, Y., Wang, S., Hilgetag, C. C. & Zhou, C. Trade-off between multiple constraints enables simultaneous formation of modules and hubs in neural systems. PLoS Comput. Biol. 9, e1002937 (2013).

"Growing evidence shows that neural networks are results from a trade-off between physical cost and functional value of the topology."

"Two obvious but apparently contradictory constraints are **low wiring cost** and **high processing efficiency**, characterized by **short overall wiring length** and a **small average number of processing steps**, respectively."

Digital and analog control in bacterial gene regulation

Evidence 3: 'Wiring economy' in the transcriptional regulatory network



spatial (Euclidean) distance in 2D space between the centers of the nodes (genes)

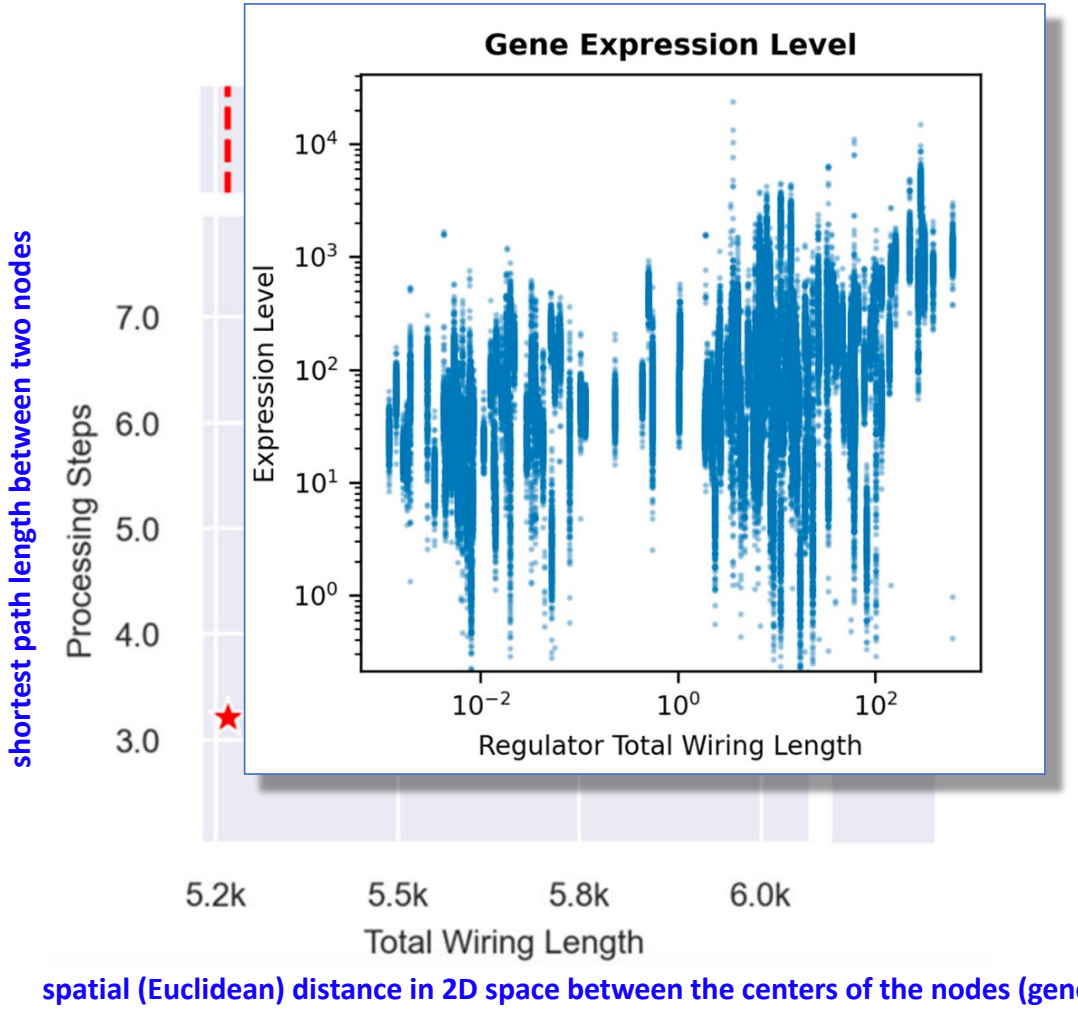
network ('hardware' implementing regulation):

- efficient processing (lower-than-random average number of processing steps)
- very efficient wiring (much lower-than-random total wiring length)

Cakir, Lesne, Hütt (2021). npj Systems Biology and Applications, 7, 49.

Digital and analog control in bacterial gene regulation

Evidence 3: 'Wiring economy' in the transcriptional regulatory network



network ('hardware' implementing regulation):

- efficient processing (lower-than-random average number of processing steps)
- very efficient wiring (much lower-than-random total wiring length)

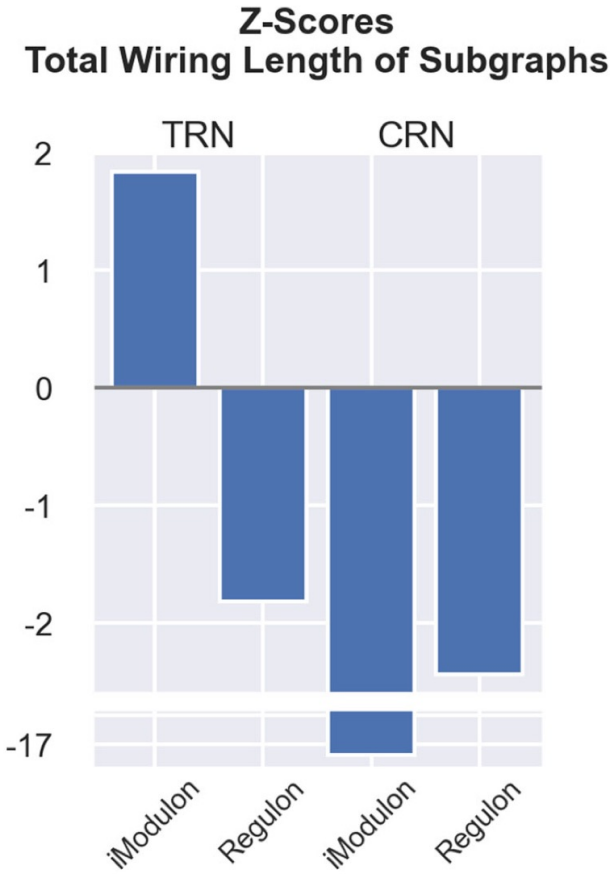
Cakir, Lesne, Hütt (2021). npj Systems Biology and Applications, 7, 49.

Digital and analog control in bacterial gene regulation

Evidence 3: 'Wiring economy' in the transcriptional regulatory network

.... back to the iModulon concept for a moment

- iModulon: gene group identified from patterns in transciptomic datasets
- Regulon: gene group regulated by the same transcriptional regulator (experimentally verified)
- Around 66% of the identified iModulons have significant overlaps with Regulons
- Reasons or biological significance of discrepancies are unclear



- Regulatory perspective (TRN): iModulons: unspecific (low 'wiring economy') in their spatial organization; Regulons: spatially organized.
- Coregulation (CRN; capacity to create coherent activity): both units are spatially tightly clustered

Cakir, Lesne, Hütt (2021). npj Systems Biology and Applications, 7, 49.

Digital and analog control in human diseases

Agreement of transcriptome data with a given biological network

Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature

Haberman et al.

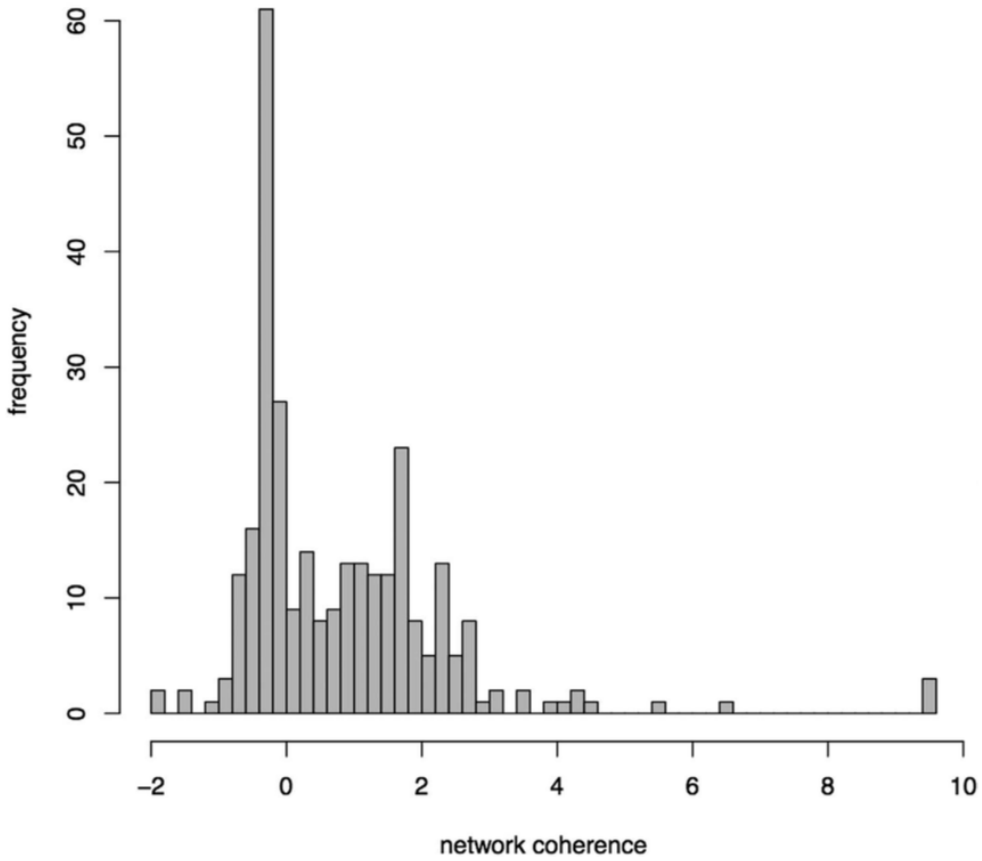
J Clin Invest. 2014;124(8):3617–3633. doi:10.1172/JCI75436.

Some features of the data set

- treatment-naive pediatric patients
  - Crohn's disease (CD)
  - ulcerative colitis (UC)
  - no inflammatory bowel disease (notIBD)
- 321 samples (with an age range from 2 to 17 years)
- gene expression measured via RNA-Seq

Digital and analog control in human diseases

Agreement of transcriptome data with a given biological network



Knecht, Fretter, Rosenstiel, Krawczak and Hütt (2016) Scientific Reports 6, 32584.

$G' = (V', E'),$

$V' \subseteq V$ differentially expressed genes

$E' \subseteq E$ all edges in G among vertices in V'

$$R = \frac{|\{v_i \in V' | k(v_i) > 0\}|}{|V'|}, \; k(v_i) \text{ degree of node } v_i$$

z-score (with respect to random vertex sets)

$$C = \frac{R - \langle R^{(\mathrm{ran})} \rangle}{\sigma_{R^{(\mathrm{ran})}}}$$

digital control strength

control strength

for gene expression *producing* cellular subsystems (gene regulation, chromosomal organization)

network coherence

for gene expression using cellular subsystems (protein interactions, signaling pathways, metabolism)

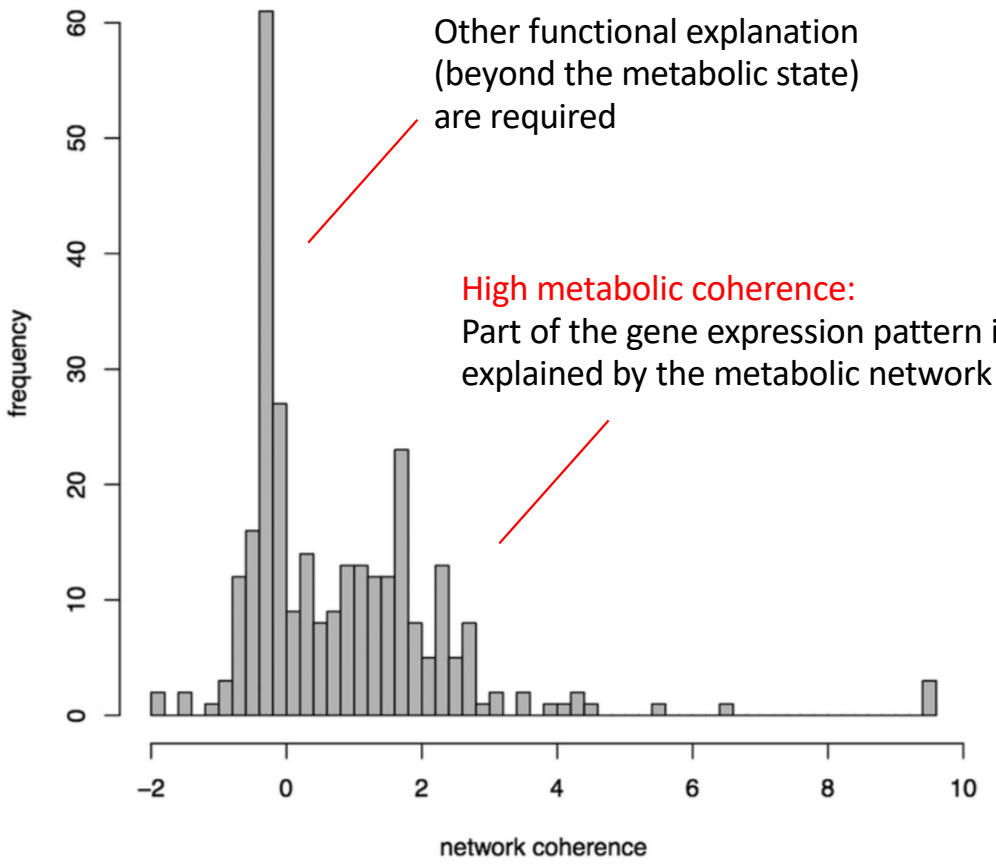
Here: genome-scale metabolic network

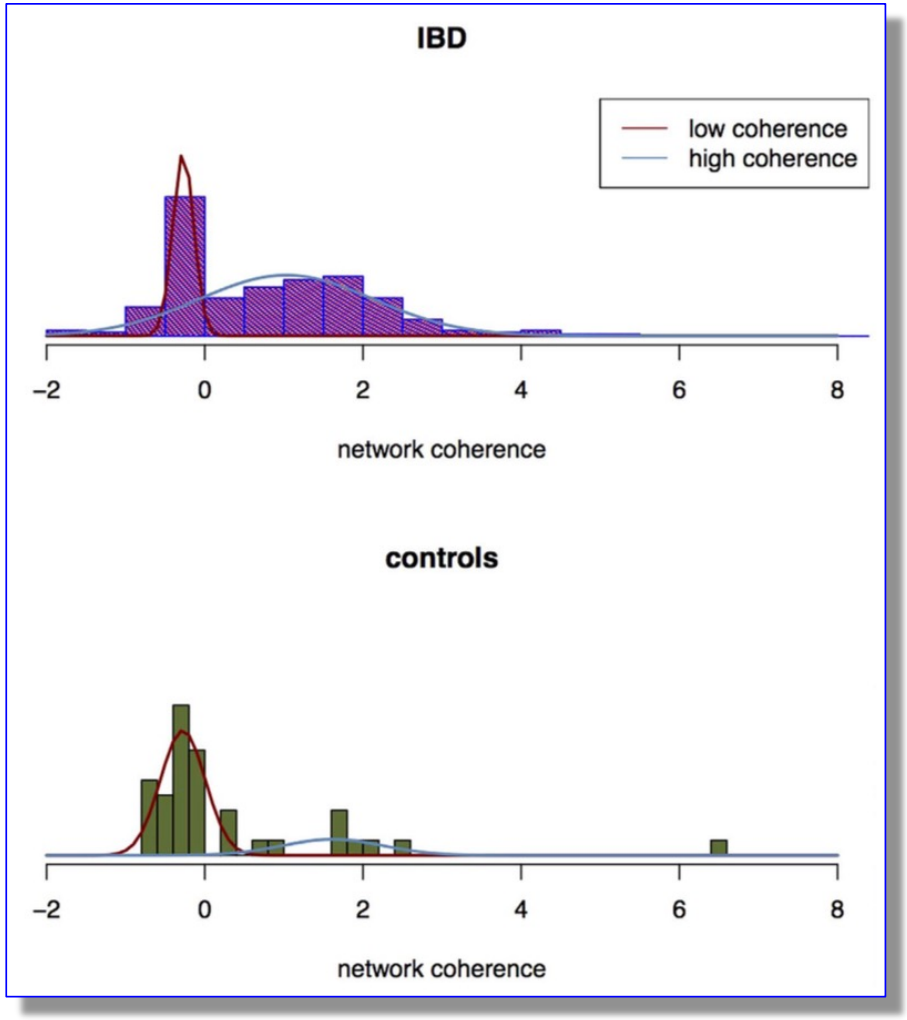
Digital and analog control in human diseases

Agreement of transcriptome data with a given biological network

Low metabolic coherence:
Other functional explanation (beyond the metabolic state) are required

High metabolic coherence:
Part of the gene expression pattern is explained by the metabolic network

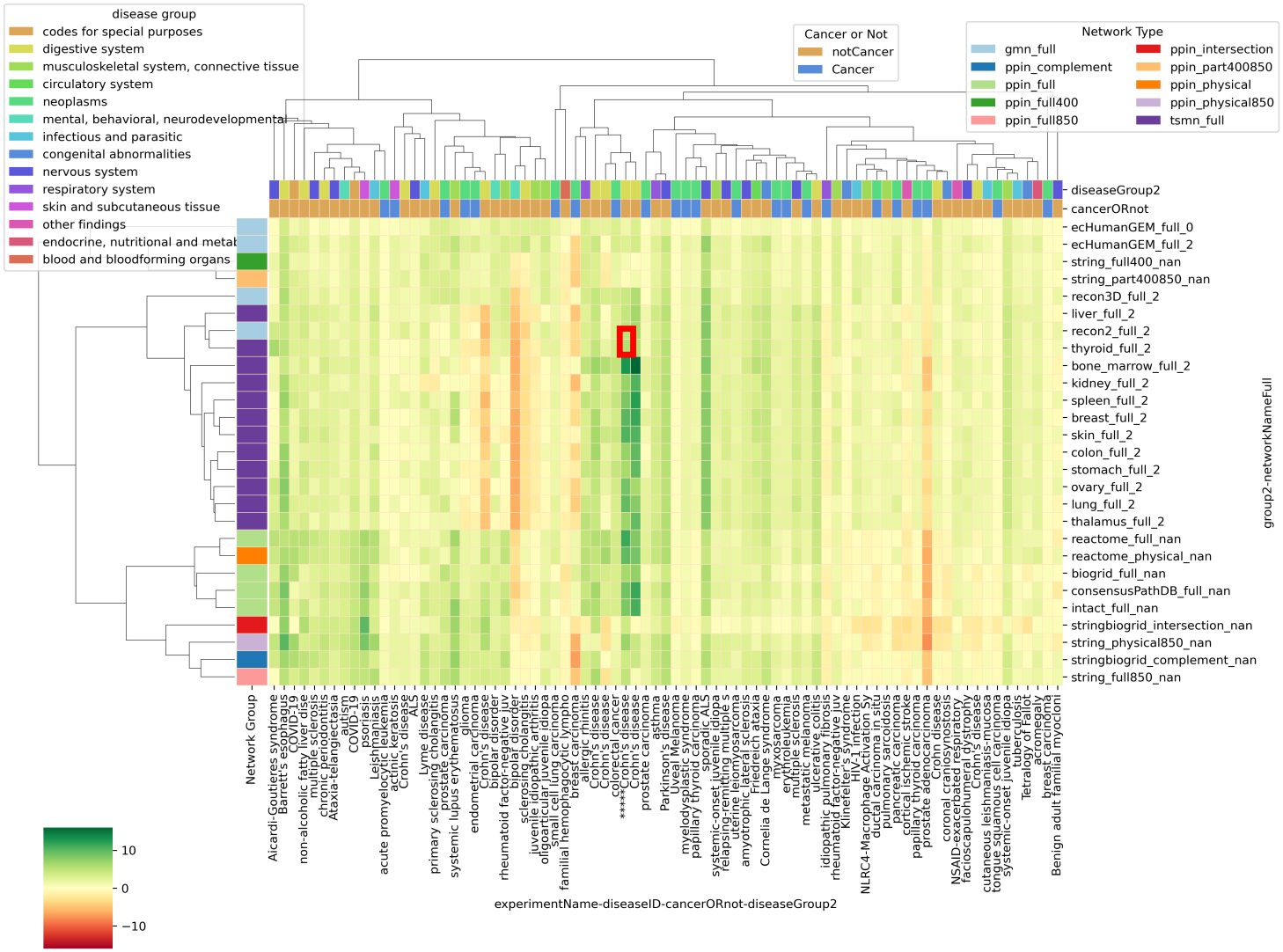




Knecht, Fretter, Rosenstiel, Krawczak and Hütt (2016) Scientific Reports 6, 32584.

Digital and analog control in human diseases

Agreement of transcriptome data with a given biological network



- strong separation of metabolic networks and protein interaction networks
- visible and robust disease clusters
- currently we do not understand these disease clusters

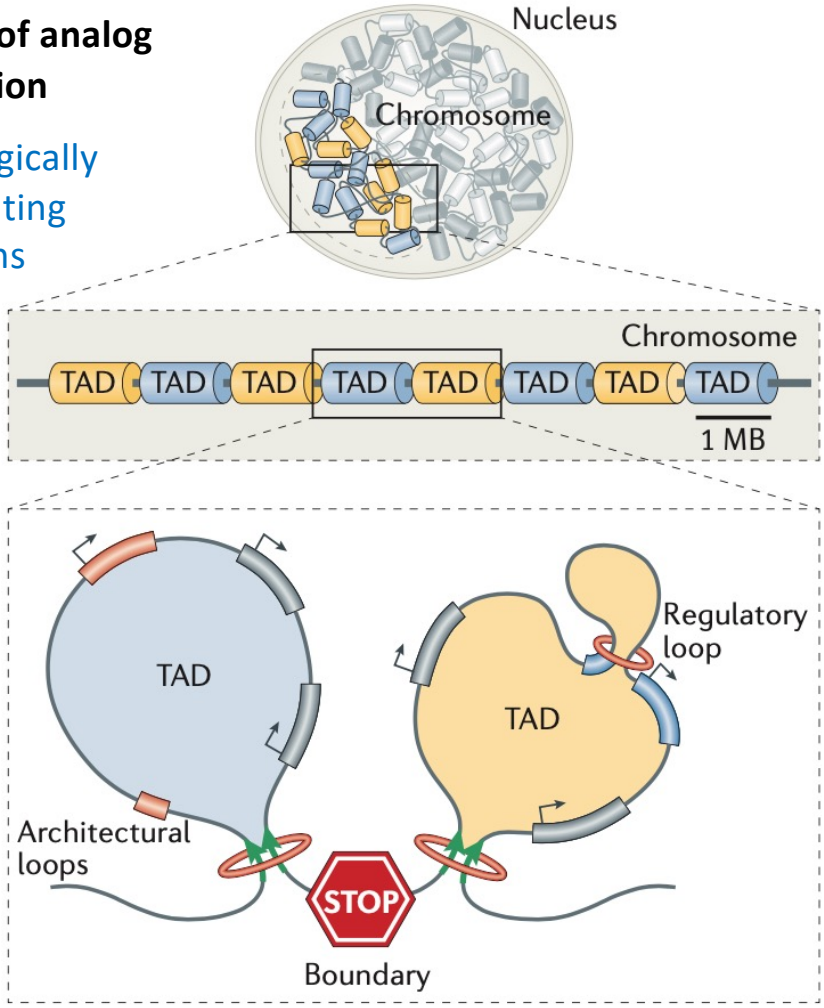
Cakir and Hütt (2024) In preparation.

Digital and analog control in human diseases

Interpretation of disease-associated SNPs via analog information

Concept of analog information

Topologically associating domains (TADs)



Taken from: Krijger and de Laat (2016) Nat. Rev. Mol. Cell. Biol. 17, 771.

Data resources

TAD data

Rao et al. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–1680.

Disease-associated SNPs from GWAS catalog

GWAS Catalog

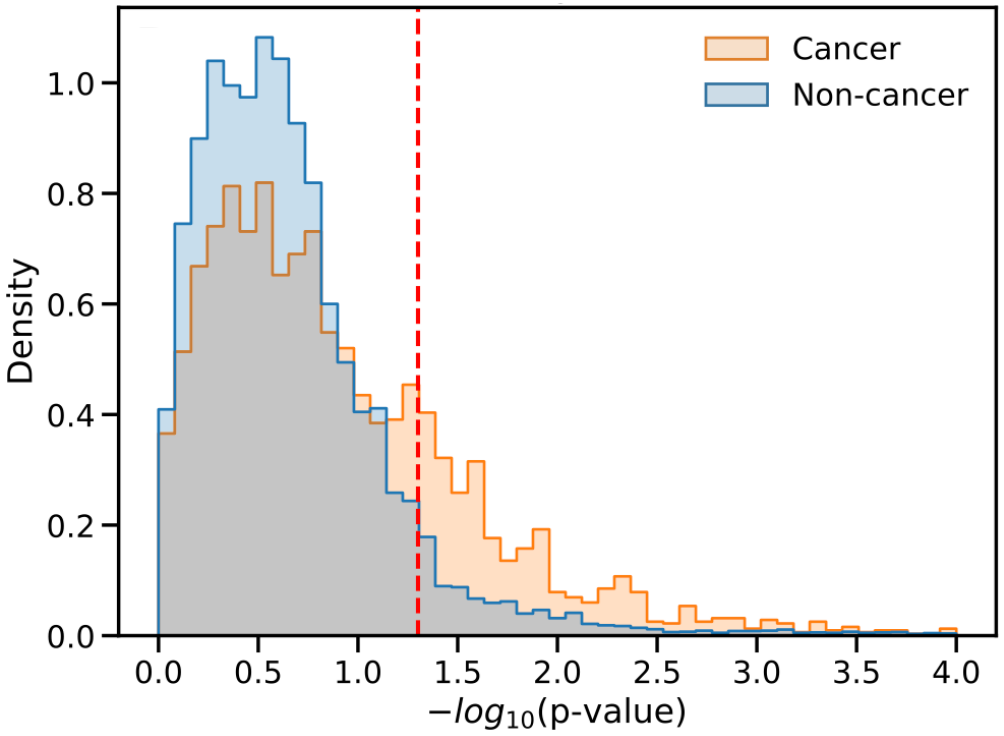
Statistical question

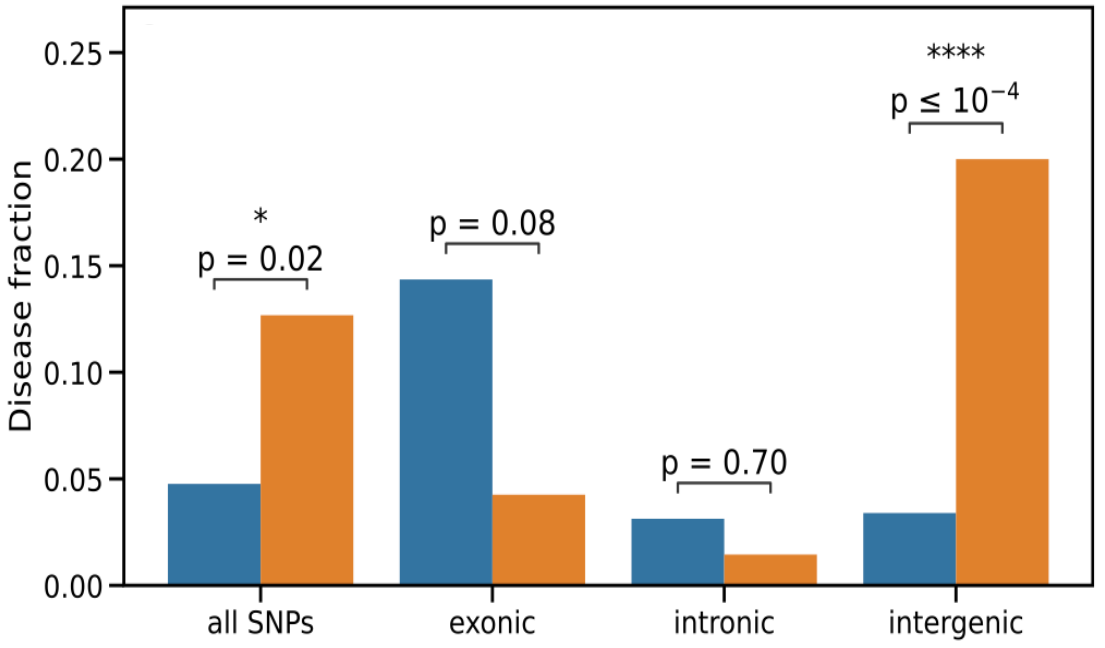
Are there diseases, for which the disease-associated SNPs are significantly often located in TAD boundaries?

Digital and analog control in human diseases

Interpretation of disease-associated SNPs via analog information

enrichment of SNPs in TAD borders





associated with specific genes

SNPs with often unclear gene association

Jablonski, Carron, Mozziconacci, Forné, Hütt, Lesne (2022). Human Genomics, 16(1), 2.

Conclusion

- Omics (high-throughput) data are transformative for biology and medicine
- So far, machine learning has had limited success in interpreting omics data
- The reason might be the interplay of digital and analog information at work in biological systems
- This interplay is relevant across all aspects of biology – from bacteria to human diseases
- Network science is a useful toolbox to make this interplay visible in omics data