Unconventional AI/ML Systems for the Biomedical Sciences

by Thomas Chittenden, PhD, DPhil, PStat, RFSPE

The world has witnessed extraordinary advancements over the past several decades in the collective understanding of biological processes at the molecular, cellular, and organismal levels. However, the current US biomedical research paradigm is no longer sustainable for its current purposes, nor up to the task of leading global innovation and ensuring that future breakthrough science is turned swiftly into benefits for people and patients.

In our 2015 *Ex Laboratorio* piece in *BioEssays*, my fellow researchers and I detailed how the presumption of perpetual growth has led to increasing demand for research monies and thus has placed modern-day bioscientific research at an economic crossroads.[1] While the problems ahead are formidable, we believe biomedical research consortia provide real-world solutions for meeting those challenges. Indeed, consortia are well suited to helping fulfill both halves of the mission of the National Institutes of Health (NIH); that is, "… to seek fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability."[2]

The inherent nature of research consortia makes it possible to overcome common barriers in established biomedical research and drug-development paradigms by promoting global scientific collaboration among academic and industry scientists. These research partnerships allow multi-institutional and independent investigators to collaborate intimately and solve complex problems.[3] To this end, in 2013, I led the founding of the Complex Biological Systems Alliance (CBSA), a non-profit global research consortium dedicated to furthering scientific understanding of biological complexity and the nature and origins of human disease. In 2014, I established the CBSA as a recognized Extreme Science and Engineering Discovery Environment (XSEDE) Campus. Through the XSEDE Campus Champions Program, the Alliance has provided its investigators with high-performance computing resources via direct access to a national consortium of supercomputing facilities, supported by a $100-million National Science Foundation award. This supercomputing platform gives CBSA scientists the means to apply cutting-edge computational discovery tools to their research and thus accelerate scientific publication of their work.

While the success of the CBSA has been significant, the economic lessons learned while directing a non-profit research consortium led me to advocate for the launch of the Genuity AI Research Institute (GAIRI) in the first quarter of 2021. As the $100-million research-and-development (R&D) arm of Genuity Science, GAIRI facilitates uniquely powerful partnerships between global life sciences, technology, healthcare, and academic organizations. Through GAIRI, these organizations can leverage Genuity's pioneering capabilities and work collaboratively to advance novel methods in AI/ML and unconventional computing in the life sciences. At the same time, these organizations can partner with Genuity on revenue-generating projects and spin-offs that efficiently turn scientific discoveries into new products with significant medical and commercial impact. The success of these R&D initiatives has attracted significant partnership interest and funding from the corporate, governmental, and not-for-profit sectors. We have consistently focused on matching this success with a proven track record of advancing diversity, equity, and inclusion within both organizations and the AI capabilities we are developing.

These R&D platforms are providing the means to further our knowledge of the molecular rules governing cellular behavior and driving phenotypes, and they are positioning us to more fully address human disease. For example, we have recently formulated a general strategy for constructing ensemble AI/ML and probabilistic programming methods by integrating *a priori* biological knowledge with multiple high-dimensional phenotype, imaging, and omics data platforms. This approach improves the performance of established classification methods via powerful graph-based feature selection and enhanced semantic interoperability for mapping between multiple biomedical ontologies. On a methodological and theoretical level, this means that we can now reliably and reproducibly identify key genes driving disease etiology. These *in silico* methodologies were experimentally validated via ongoing collaborations with investigators at the Yale University School of Medicine and the Strasbourg University Medical School in France.[4]

Applied to specific diseases, these technologies have recently led us to significant advances in understanding the causal molecular drivers of cardiovascular disease; various cancer states; COVID-19; and several CNS disorders, including Alzheimer's disease. Our novel *in silico* phenotype projection strategies afford robust, reproducible correlations of putative molecular causal dependency structures with digital pathology and other biomedical imaging modalities.[5] In plain terms, we are now mapping the causal biology of disease, not the symptoms. This greatly improves the specificity of patient diagnosis and points to therapeutic interventions rooted in the biology of disease.

While numerous studies highlight the potential of multi-omic analytics to drive discovery based on unique molecular single-cell signatures,[6] several well-documented statistical computing limitations—including large-scale statistical optimization—remain for the analysis of high-dimensional complex biological datasets.

Through a research collaboration with D-Wave Systems and the USC-Lockheed Martin Quantum Computation Center, our AI/ML teams have access to quantum computing hardware. In our 2021 *Patterns* paper, we demonstrate, for the first time, that our novel quantum machine learning (qML) strategies provide competitive classification of actual human cancer types and associated molecular tumor subtypes and "superior" performance with smaller training dataset sizes.[7]

This is, to our knowledge, the first published, replicable evidence for the potential of quantum computing for accelerating our understanding of human biology. Robust classification of small, high-dimensional, genome-wide multi-omics datasets with qML and natural-language processing (NLP) methods provides a potential new avenue to evaluate patient response in early phase clinical drug trials or in other genome-wide datasets with relatively small numbers of patients or animal models. In collaboration with investigators at the University of Strasbourg Medical Center, we have applied qML to uncover putative drivers of SARS-CoV-2-induced acute respiratory distress syndrome (ARDS) in COVID-19 patients. In this study, application of our integrated ensemble AI/ML and structural causal modeling strategy uncovered a complex disease etiology, which led to experimental inhibition of SARS-CoV-2 viral uptake and replication in human lung epithelial cells (manuscript submitted). We are now assessing this disease signature in bacterial sepsis patients to determine whether we have identified a general pathobiological mechanism of ARDS.

Moreover, via collaborations with several large technology companies, we are investigating the utility of novel neuromorphic memristive computing architectures to address current limitations of statistical learning strategies employed on established high-performance computing architectures. This work also involves the development of deep-spiking neural networks, which represent the third generation of

novel deep-learning topologies. Our initial findings indicate even greater potential than qML. Taken together, these unconventional computing architectures and associated AI/ML strategies will inevitably provide real-world solutions for many applications, not just in medicine but across modern society.

The reprinted article that begins on the next page outlines novel, agnostic *in silico*-based strategies for drug discovery and development. I trust this piece will help inform the reader on the utility of these AI/ML and structural causal modeling methodologies, which can be applied to a wide range of research disciplines.

NOTES

1. Curt Balch, et al., "Science and Technology Consortia in U.S. Biomedical Research: A Paradigm Shift in Response to Unsustainable Academic Growth," *Bioessays* 37, no. 2 (November 2014): 119-122, doi:10.1002/bies.201400167; Bruce Alberts, Marc Kirschner, Shirley Tilghman, and Harold Varmus, "Rescuing US Biomedical Research from its Systemic Flaws," *Proceedings of the National Academy of Science* 111, no. 16 (April 2014): 5773-5777, doi:10.1073/pnas.1404402111.

2. National Institutes of Health, "Mission and Goals," https://www.nih.gov/about-nih/what-we-do/mission-goals.

3. Balch, et al.

4. Guangxin Li, et al., "Chronic mTOR Activation Induces a Degradative Smooth Muscle Cell Phenotype," *The Journal of Clinical Investigation* 130, no. 4 (March 2020): 1233-1251, doi:10.1172/JCI131048; Pei-Yu Chen, et al., "Smooth Muscle Cell Reprogramming in Aortic Aneurysms," *Cell Stem Cell* 26, no. 4 (April 2020): 542-557, doi:10.1016/j.stem.2020.02.013; Nicolas Ricard, et al., "Endothelial ERK1/2 Signaling Maintains Integrity of the Quiescent Endothelium," *Journal of Experimental Medicine* 216, no. 8 (Aug 2019): 1874-1890, doi:10.1084/jem.20182151; Pei-Yu Chen, et al., "Endothelial TGF-beta Signalling Drives Vascular Inflammation and Atherosclerosis," *Nature Metabolism* 1, no. 9 (September 2019): 912-926, doi:10.1038/s42255-019-0102-3; Pengchun Yu, et al., "FGF-Dependent Metabolic Control of Vascular Development," *Nature* 545 (2017): 224-228, doi:10.1038/nature22322.

5. Michael Simons, Jeffrey R. Gulcher, and Thomas W. Chittenden, "*In Silico* Phenotype Projection of Endothelial ERK1/2 Signaling," *Aging* 12, no. 11 (June 2020): 10001-10003, doi:10.18632/aging.103529.

6. Balch, et al.; Pei-Yu Chen, et al., "Smooth Muscle Cell Reprogramming"; Pei-Yu Chen, et al., "Endothelial TGF-beta Signalling"; Michael A. Lodato, et al., "Somatic Mutation in Single Human Neurons Tracks Developmental and Transcriptional History," *Science* 350, no. 6256 (October 2015): 94-98, doi:10.1126/science.aab1785; Michael A. Lodato, et al., "Aging and Neurodegeneration Are Associated with Increased Mutations in Single Human Neurons," *Science* 359, no. 6375, (February 2018): 555-559, doi:10.1126/science.aao4426; Michael Simons, Pei-Yu Chen, and Thomas W. Chittenden, "Resilience, Disease and the Age of Single Cell Science," *Aging* 12, no. 3 (February 2020): 2028-2029, doi:10.18632/aging.102850.

7. Richard Y. Li , et al., "Quantum Processor-Inspired Machine Learning in the Biomedical Sciences," *Patterns* 2, no. 100246 (2021), https://www.cell.com/patterns/pdf/S2666-3899(21)00066-0.pdf.

Editor's note: The following article is reprinted in its entirety from the *Journal of Precision Medicine* 7, No. 1 (March 2021), pages 37-42. The original article can be accessed online at https://www.thejournalofprecisionmedicine.com/wp-content/uploads/aortic-chittenden-gulcher.pdf.

Questions on an agnostic AI system: A case study for an aortic aneurysm detection and other applications.

An interview with Tom Chittenden and Jeff Gulcher.

Introduction and Background

Genuity Science is moving away from the old school methods of assessing disease states by individual tests through its artificial intelligence (AI) platform that draws on data from single cells to macroscopic phenotypes. By training AI and machine learning (ML) systems with disease-agnostic methods, Genuity Science can expand the current understanding of biology from the cell to the phenotype level through use of Bayesian networks to perform probabilistic inference in high-dimensional biomedical data and uncover rules for molecular and whole-body governance. Furthermore, these AI/ML systems can generate hypotheses essentially free of confirmation bias, thereby allowing researchers in biology and medical fields to tease out novel mechanisms of action of diseases and therapeutic interventions.

In collaboration with investigators at Yale University Medical School, the Advanced AI Research Laboratory at Genuity Science has extensively studied cells which make up blood vessels and the causal gene drivers that lead to thoracic aortic aneurysms. In an experimental mouse-model system, results from the AI team's time-series analysis of single-cell RNA sequencing (scRNA-seq) data showed that a single cluster of abnormal smooth muscle cells produced thickening of the vessel wall, which led to arteriosclerosis, and subsequently to thoracic aortic aneurysm.

Armed with this information, the team at Yale analyzed expression of specific protein markers by imaging mass cytometry to confirm the finding of a single cluster of abnormal smooth muscle cells in the aorta with consequent remnants of ossification and calcification. The Genuity Science AI team was then able to statistically construct, with the use of conditional probability, a causal gene dependency structure from the mRNA expression signature in that single cluster of cells that bulk RNA-seq methods would have otherwise missed. Their analysis uncovered a novel putative gene driver of disease etiology. Interestingly, a DNA variant within this gene associates with intimal hyperplasia of blood vessels, atherosclerosis, and hypertension in humans.

Along with cardiovascular disease, the Genuity Science AI team has also built integrated multi-omic and single-cell ensemble AI/ML, digital pathology, biomedical imaging, and natural language processing strategies to identify etiological molecular mechanisms of cancer, COVID-19, nonalcoholic steatohepatitis (NASH), and neurodegenerative diseases. Capabilities in oncology applied to 8,200 tumors and 22 cancer types in the Cancer Genome Atlas (TCGA) discriminated any specific tumor from the other 21 types with greater than 99% accuracy. Moreover, the Genuity AI team reports an impressive 76% accuracy rate in predicting pan-cancer patient survival at 60 months. Current state of the art technology is approximately 70% accuracy with patients of a single cancer type.

We asked Tom Chittenden and Jeff Gulcher to address a few questions on the Genuity Science AI platform and its applications in the biomedical sciences.

Q1. Genuity's AI approach claims to be disease agnostic. Can you explain what that means in context of the AI technology, training data sets, and related applications?

A. The Genuity Science AI platform is indeed disease agnostic. Approximately six years of painstaking R&D has produced robust biological domain-specific feature learning for our integrated ensemble AI/ML and probabilistic programming strategies. Our novel applications of these methods afford highly accurate, generalizable pattern recognition within high-dimensional biomedical data. By specifically addressing the significant degrees of correlation bias, feature dependency, and multicollinearity that exist within these data sets, we can now construct highly reproducible, causal gene dependency networks, indicative of the signal transduction cascades that occur within cells and drive phenotypic and pathophenotypic development. Furthermore, our research has shown that we are able to apply these investigative in silico methods to any human disease state.

Along with our research collaborators, we are applying and experimentally validating what we (Genuity) claim to be the most transformative domain-specific, analytical technology currently available. Our technology platform allows users to advance the collective understanding of human biology and to address human disease more fully. We use the tools of mathematics, first, to untangle the molecular and biochemical complexities of cellular behavior that have arisen through natural biological engineering and selection, and second, to build AI/ML strategies capable of uncovering novel mechanisms of human disease etiology. We have recently uncovered putative etiological molecular mechanisms of cardiovascular disease, COVID-19, nonalcoholic steatohepatitis (NASH), Alzheimer's, and cancer, among others.

Q2. On differentiating Genuity's approach: How does Genuity differentiate its AI/ML approach from other algorithms?

A. While we have developed several novel statistical learning approaches for the biomedical sciences, the strategic application of our ensemble AI/ML algorithms is the real differentiator. Statistical optimization is still somewhat problematic in high-dimensional biomedical data, especially for experimental designs with limited sample sizes; for example, datasets with fewer than 100 patients. As such, we have never fully trusted our ability to generate reproducible feature selection and subsequent classification of a single 80/20 partition of data with only one machine or deep learning algorithm. We have built biology domain-specific feature learning and ensemble AI/ML strategies that include multiple partitions of train and test sets with typically six or seven distinct classes of classification algorithms. As an ensemble, these sets provide more generalizable and stable results for downstream analysis and biological interpretation.

Our associated ensemble feature rankings significantly improve our consistent ability to identify gene features that associate with a given disease. This ensemble approach, in turn, provides a strong statistical framework for structural causal modeling within both longitudinal and cross-sectional experimental designs to discover the handful of genes that are drivers of the disease.

In collaboration with Professor Michael Simons and his team at the Yale University School of Medicine, we published experimental evidence1 that validated the robust and consistent efficacy of these supervised AI/ML approaches. The pathobiological insights gained from this work led to the development and application of generative AI models capable of uncovering novel aberrant molecular mechanisms of atherosclerosis2 and thoracic aortic aneurysm3,4 at the single-cell level.

Another major challenge for the AI/ML field is our still relatively limited grasp of human biology. Therefore, knowledge-based AI/ML applications that restrict us only to our current understanding of the human genome and biology do not yield true biological discoveries (see also Q6 for new sources of data through Genuity's collaboration partners). We believe it is also important to point out that while our feature learning approaches are biologically oriented, they do not limit us to our current knowledge base. In other words, our AI/ML strategies can identify putative causal gene signals of unknown function, thus allowing us to advance our collective understanding of human biology.

How would you differentiate Genuity's panomic and phenotypic training sets from other approaches, particularly for its use of molecular data (DNA, RNA, methylation) to imaging and to patient history?

A. We now house some of the largest expertly curated patient cohort data in the world. Our CIO, Hákon Guðbjartsson and his Informatics team in Iceland have developed several advanced analytical software solutions, including a highly scalable database system based on GORdb (genomically ordered relational database) architecture and a novel query syntax. These systems can provide seamless integration of multiple types of clinical, molecular, phenotypic, and biomedical imaging platforms for downstream analyses of patient data with both established statistical methods and advanced AI/ML strategies. Our overall approach provides us with the robust means to identify potential drug targets and disease biomarkers in a much more effective manner than current existing systems.

RNA and DNA methylation (or histone acetylation) states are specific to tissue types. Do you see any tissue-specific signatures identified from AI methods that may provide the basis for classical diagnostics tests? Examples?

A. Our extensive work with TCGA has revealed novel insights into cancer biology and patient survival, especially regarding DNA methylation. We have thus built an ensemble computational intelligence approach based on multi-omics data from approximately 8,200 human tumors that is greater than 99% accurate in predicting one cancer from among the 22 cancer types studied. Somewhat analogous to facial recognition, these supervised AI/ML models confer complex disease recognition, and we have found that DNA methylation is by far the most informative individual omics signal of the five high-dimensional data types assessed. Moreover, we can achieve a 76% accuracy rate in predicting pan-cancer patient survival over 60 months with our associated survival models. **Figure 1** represents an overview of our general ensemble computational intelligence pipeline for multinomial classification and subsequent structural causal modeling and natural language processing of the TCGA annotated cancer types assessed.

Coupled with our findings from the analysis of normal tissue types, we hypothesize that the 22 cancer types evaluated house a strong underlying tissue of origin signal. Thus, we are currently evaluating this information further to determine the efficacy of classifying tumors of unknown origin with similar models.

Are there any other technology differentiators you could cite?

A. Through a research collaboration with Professor Daniel Lidar and his team at the USC-Lockheed Martin Quantum Computing Center in Southern California, we are actively developing statistical optimization strategies for high-dimensional problems with limited data. Based on recent advances in quantum computing, we have evaluated the utility of several unconventional Ising-type ML strategies, including simulated and quantum annealing, for classifying human cancer data. We find Ising-type ML provides competitive classification of human

cancer types and superior performance on smaller training sets of multi-omics patient data. As such, we have extended this multi-institutional collaboration to tackle the SARS-CoV-2 pandemic with Professor Seiamak Bahram and his team at the University of Strasbourg and INSERM in France.

Our preliminary analyses of multi-omics data from a relatively small number of severe COVID-19 ICU patients identified putative causal gene drivers of a complex disease etiology. These empirical findings support future research of unconventional computing in drug target and biomarker discovery, and possibly provide a new avenue to evaluate drug efficacy in early-phase clinical trials with relatively small numbers of patients.

Q3. Can you cite case studies where Genuity identified a condition, signature, or underlying conditions that other analyses missed?

A. In our 2019 JEM[1] paper with Professor Simons' team at Yale University School of Medicine, we gained greater insights into extracellular response kinase (ERK) signaling by applying a novel ensemble computational intelligence strategy (including integrated deep learning and probabilistic programming of bulk RNA-seq data from experimentally perturbed human endothelial cells). We find that this statistical computing framework is highly effective in constructing putative causal gene dependency structures based on conditional probabilities of gene expression states within cross-sectional experimental designs and without the use of prior biological information. The overall goal of the approach is to generate working hypotheses from subsequent hierarchical gene interaction networks for downstream experimental analyses. We are removing the human hand from our analysis– this is completely data driven without using preconceived (or inherently biased) notions about the biology. **Figure 2** illustrates how we implemented our combined deep learning and probabilistic programming pipeline to first assess the molecular consequences of experimentally silencing ERK1/2 in human umbilical vein endothelial cells (HUVECs) in order to subsequently predict and then validate phenotypic cardiovascular abnormality in mice after analogous in-vivo ERK1/2 perturbation.

The structural casual modeling (SCM) of gene expression data in this study revealed that activation of transforming growth factor beta (TGFβ) signaling was most likely driving the state of a gene interaction network, thus putatively linking ERK 1/2 regulation to endothelial-to- mesenchymal transition (EndMT). The model also causally linked other functionally annotated EndMT genes implicated in renal fibrosis and systemic hypertension to TGFβ signaling. As predicted by the AI/ML-derived findings in human endothelial cells, phenotypic and functional assessments of adult mice with endothelial cell specific ablation of ERK 1/2 uncovered severe hypertension, extensive EndMT, multi-organ fibrosis, and progressive renal failure.

The high degree of concordance among our in-silico predictions and these in-vivo biological findings empirically validates that our computational statistics method (we coined as "in silico phenotype projection"[5]) is capable of robustly identifying molecular attributes of signal transduction cascades implicated in cellular behavior. Moreover, as noted in our first 2020 Perspective in Aging,[5] while this AI/ML-based approach is somewhat analogous to the established experimental methods of reverse genetics, it is significantly more effective in predicting complex phenotypes by identifying the causal molecular underpinnings of disease etiology within the context of an entire biological system.

Figure 1, Part a (first half)

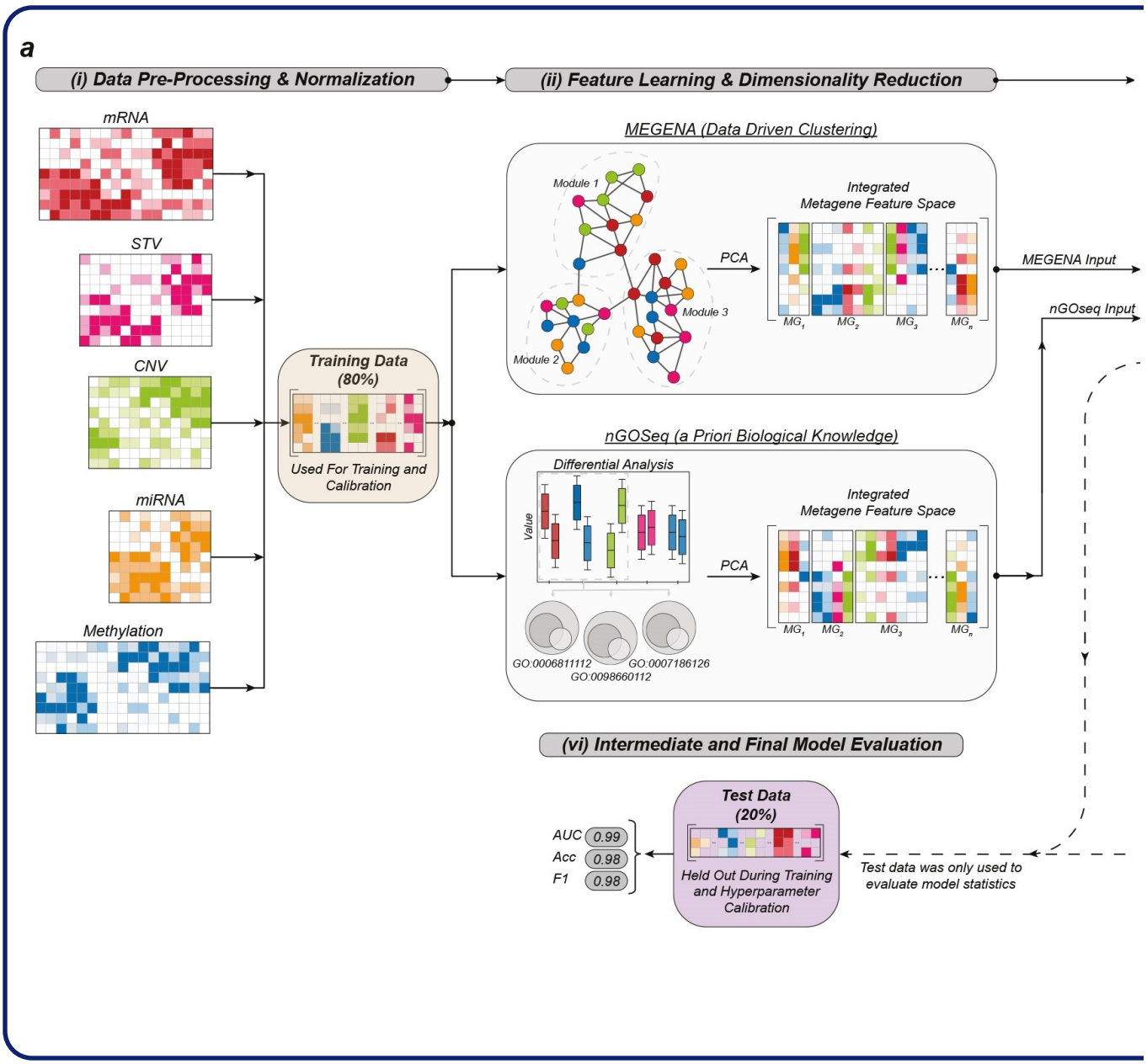


Figure 1, Part a (second half)

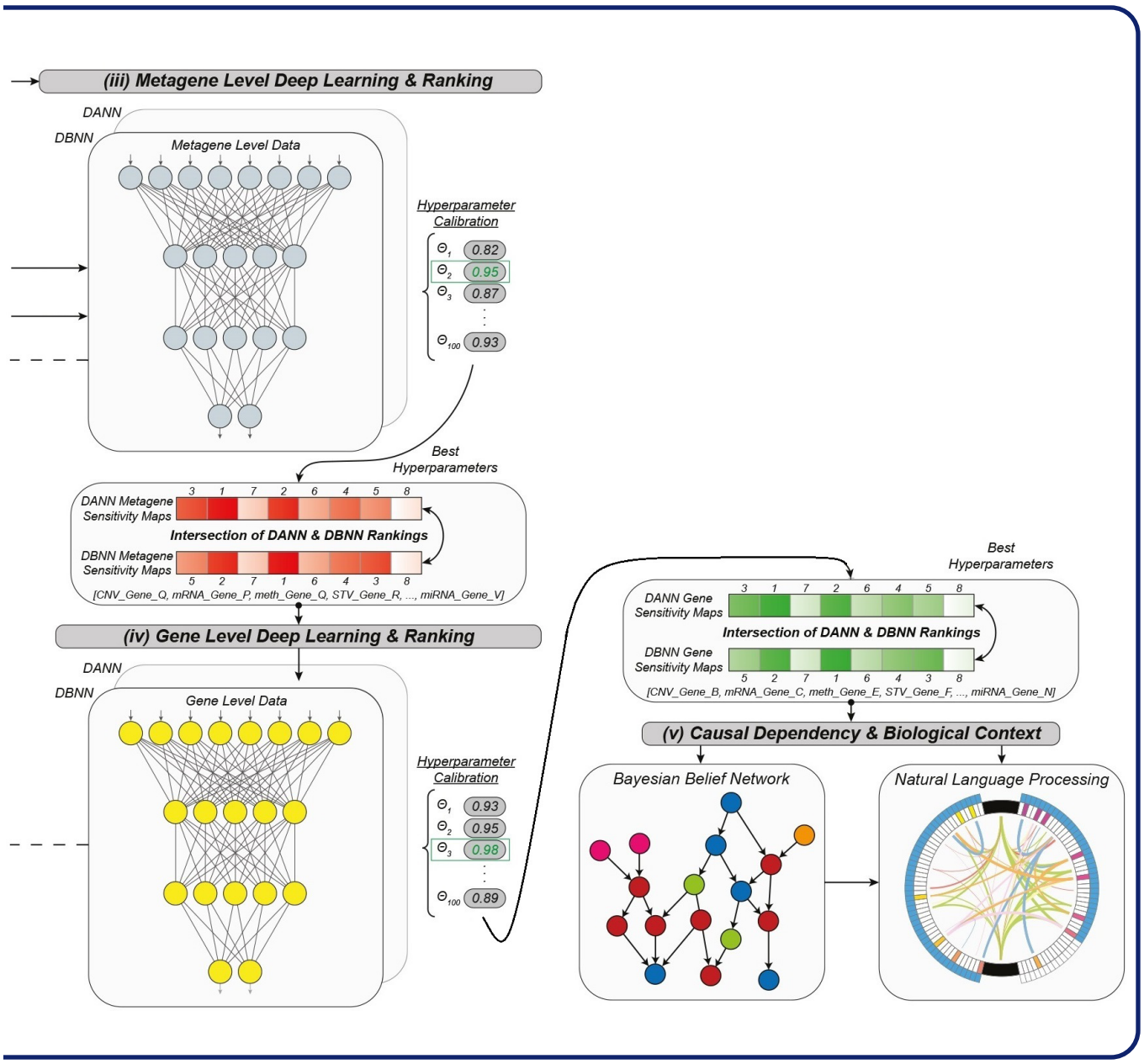


Figure 1, Part b

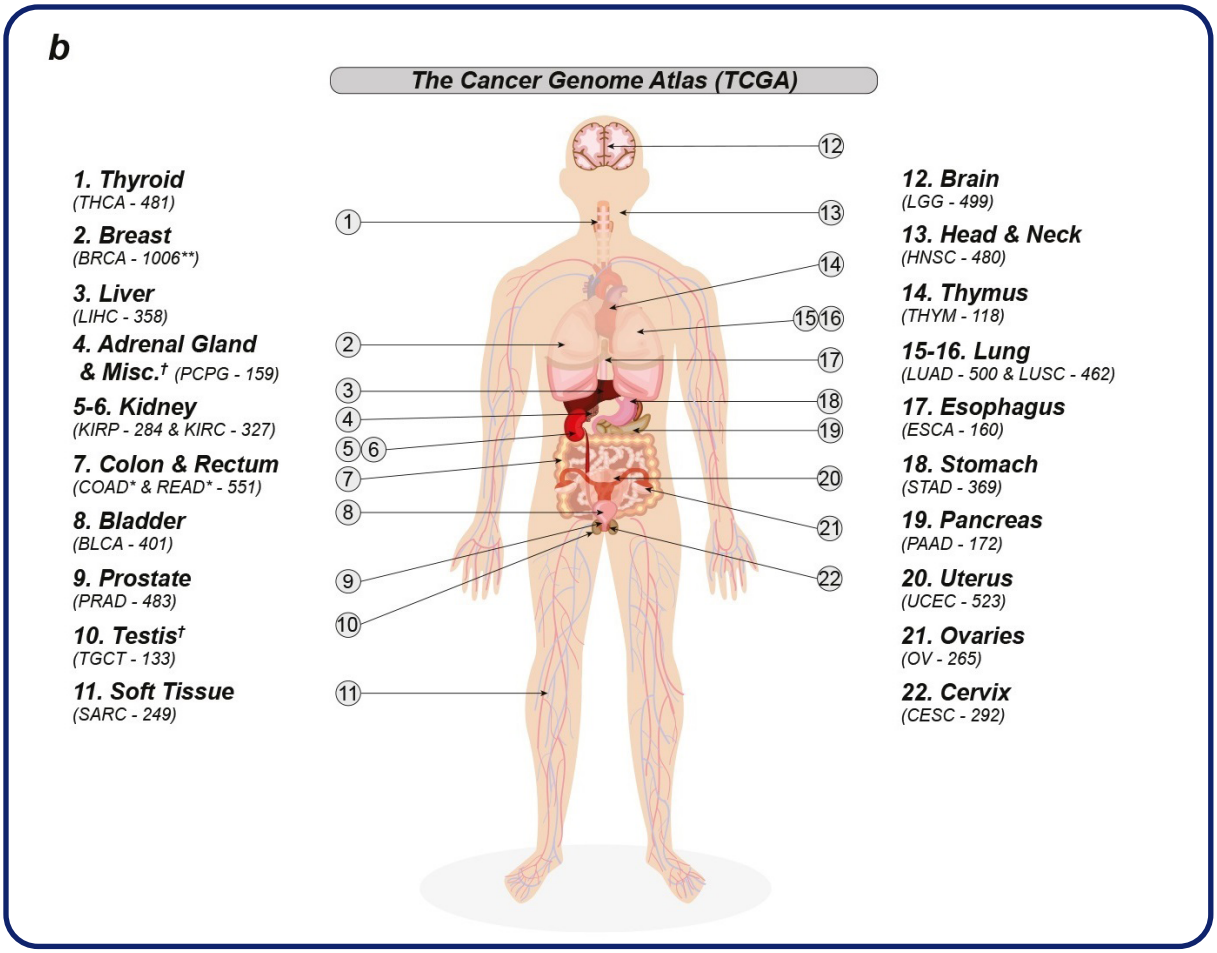


Figure 1: Schematic diagram of our computational intelligence approach for multinomial classification of 22 TCGA cancer types. a) Overall modeling strategy: (i) Five data types, including messenger RNAs (mRNAs), somatic tumor variants (STVs), copy number variations (CNVs), micro-RNAs (miRNAs), and DNA methylation (METH), were downloaded, pre-processed, normalized, and split into training and test data matrices. (ii) Feature learning via a data driven clustering approach (MEGENA) or an a priori biological knowledge based approach (nGOseq) in addition to principal component analysis (PCA) were used to create metagene level data matrices, (iii) Metagene level ensemble approach (with inner cross-validation loop for tuning hyper-parameters) consisting of Deep Artificial Neural Networks (DANNs), Deep Bayesian Neural Networks (DBNNs), and sensitivity maps to determine the most informative MEGENA or nGOseq metagenes, (iv) Genes from the most important metagenes are broken out into gene level data matrices and the ensemble approach (DANNs, DBNNs, and sensitivity maps, with inner cross-validation loop for tuning hyper-parameters) is again used to determine the most informative genes, (v) The most informative genes were used to compute Bayesian Belief Networks (BBN) to predict causal drivers and query biological relevance with natural language processing, (vi) Model evaluation on the held-out test data. b) The 22 cancer types, acronyms, and sample numbers from The Cancer Genome Atlas (TCGA). All available TCGA cancer types were filtered based on total sample number and availability of all five data types. *Colon Adenocarcinoma (COAD) and Rectum Adenocarcinoma (READ) were merged into a single cancer type (CRAD) due to their similarity. **Breast Invasive Carcinoma contains subtypes including ER status (+/-) and Luminal A/B used in subsequent binomial comparisons. † Indicates samples that were excluded from survival analysis. Total sample number was 8,272 for 22 cancers for multinomial classification and 7,822 for 20 cancers for pan-cancer survival analysis (methodology not shown). We thank Nicholas A. Cilfone for initial data analysis and figure development.

Q4. The Genuity website has links to papers with citations about endothelial to mesenchymal transition in atherosclerotic lesions and potential plaque instability.

Could you discuss how Genuity leverages AI to identify lesions and plaques in images, especially as it relates to clonal and subclonal cell populations in lesions and plagues?

How does the AI algorithm identify clonal and subclonal populations in aortic aneurysms?

Can you elaborate on how these results can lead to new pathways to therapeutics and diagnostics?

A. As noted in our second 2020 Aging[6] Perspective, endothelial-to-mesenchymal transition (EndMT) is implicated in several major pathological conditions, including atherosclerosis, pulmonary hypertension, renal dysfunction, and vascular malformations. Therefore, the insights gained from our 2019 JEM[1] paper launched a large single cell RNA sequencing (scRNA-seq) initiative in cardiovascular disease (CVD) in collaboration with Professor Simons' team at Yale University School of Medicine. This initiative is aimed at uncovering novel etiological drivers of atherosclerosis and thoracic aortic aneurysm (TAA), two closely related CVD pathologies.

In our 2019 Nature Metabolism[2] paper, we showed that experimentally induced inhibition of TGFβ signaling in endothelial cells of ApoE knockout mice reversed atherosclerosis. Moreover, nanoparticle-based RNAi targeting of the TGFβ pathway resulted in the same outcome in these animals. To gain a greater understanding of the underlying dysregulated molecular mechanisms of action, we used our own flavor of generative variational autoencoder (VAE) to assess scRNA-seq data from endothelial cells from these same mice.

We have extended the capabilities of conventional VAEs to address the inherent nature of scRNA-seq data as well as other data generated from measurements of processes that follow non-normal distributions. Our generative AI strategy robustly identified a single subpopulation of aberrant cells involved in EndMT and atherosclerotic plaque development. Interestingly, the disruption of TGFβ signaling led to significantly fewer pathogenic endothelial cells. To assess the mechanistic role of EndMT and atherosclerosis in thoracic aortic aneurysm, we then evaluated the contribution of abnormal smooth muscle cell differentiation to TAA disease etiology.

In our 2020 Cell/Stem Cell[4] paper, we applied similar generative AI approaches in the analysis of subpopulation of aberrant cells involved in EndMT and atherosclerotic plaque development. Interestingly, the disruption of TGFβ signaling led to significantly fewer pathogenic endothelial cells. To assess the mechanistic role of EndMT and atherosclerosis in thoracic aortic aneurysm, we then evaluated the contribution of abnormal smooth muscle cell differentiation to TAA disease etiology.

In our 2020 Cell/Stem Cell4 paper, we applied similar generative AI approaches in the analysis of scRNA-seq data over a four-month longitudinal experimental design aimed at identifying the causal molecular drivers of cell fate transition. We were able to quantify the significant loss and concomitant gain of expression of several smooth muscle cell and mesenchymal-like stem cell markers, respectively. Our experimental investigations revealed that the aberrantly reprogrammed subpopulation of smooth muscle cells gave rise to numerous cell types, including adipocytes, chondrocytes, osteoblasts, and macrophage-like cells. Also, based on abrogation of TGFβ signaling and hypercholesterolemia, this pathologic cell population was responsible for the abnormal growth and dilation of the aorta as well the calcification and ossification of the aortic

Figure 2 (first half)

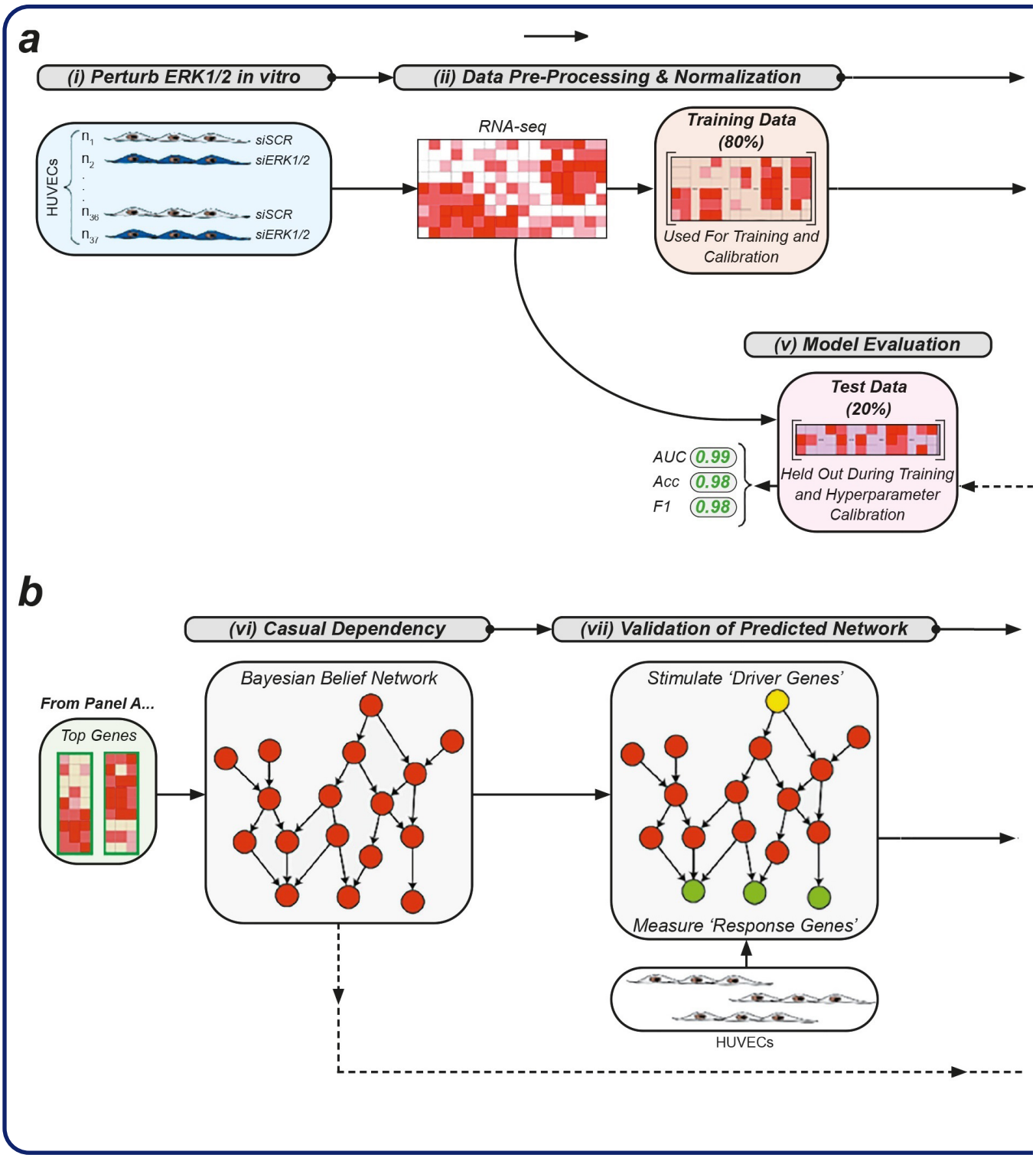
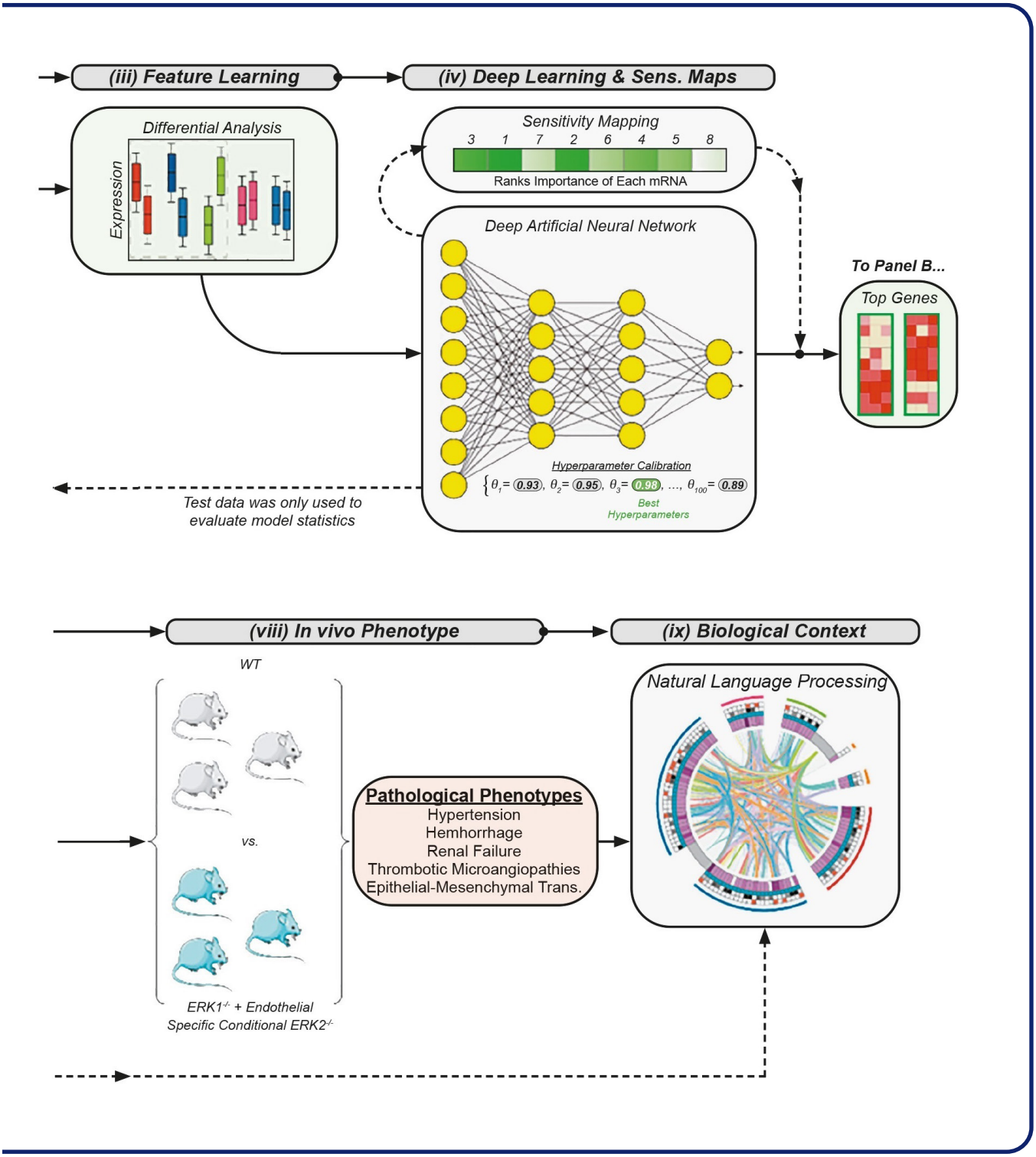


Figure 2 (second half)



wall. This atypical cellular phenomenon ultimately led to TAA in these animals. Imaging mass cytometry then confirmed the scRNA-seq findings by also identifying a single cluster of abnormal smooth muscle cells driving the disruption of the endothelium.

Via structural causal modeling of the scRNA-seq data, we have recently identified several putative drivers of TAA disease pathogenesis. This led us to subsequent identification of a genetic DNA variant within the main causal gene driver that associates with increased intimal thickening of the vascular wall in humans. As indicated by these findings, we believe generative AI modeling of single cell data will lead to robust classification of human disease based on cell differentiation trajectories. Analogous to studies that have used scRNA-seq approaches to classify clonal cell populations in heterogenous tumors, our single cell generative AI results indicate that our robust ability to identify small subpopulations of pathogenic endothelial and smooth muscle cells will lead to a greater collective understanding of the molecular drivers of cardiovascular disease, and thus, in turn, the development of more efficacious CVD therapeutics.

Q5. To what extent can this method identify pre-symptomatic or asymptomatic at-risk populations? Can you cite or speculate on cases that other methods missed (partially or fully) that Genuity would catch?

A. As the pathophysiology of heart failure patients with preserved left ventricular ejection fraction (HFpEF) is currently unresolved and pharmacologic agents for this debilitating condition are also largely ineffective, we are interested in defining the phenotypic heterogeneity that accounts for failed clinical trials. Once clearly delineated, identification of the causal molecular drivers of HFpEF subclasses will afford greater therapeutic efficacy for these patients. Thus, we have extended our single-cell generative AI models to perform deep phenomapping and patient stratification of the 1,400 Irish participants of the Incident Heart Failure Study, STOP-HF (see, e.g., STOP-HF, https://clinicaltrials.gov/ct2/show/NCT00921960, Principal Investigators: Ken McDonald and Mark Ledwidge).

In an initial collaboration with Professors McDonald and Ledwidge and their teams at University College Dublin, we are currently assessing DNA samples and up to 10 years of longitudinal clinical data indicating early and incident heart failure with various cardiac biomarkers and cardiac imaging to uncover unique therapeutically homogeneous HFpEF patient subclasses. One of the major goals of this study to identify genes that impact the early, pre-symptomatic phase of heart failure pathogenesis. In parallel, we are applying our generative AI models to 4,700 HFpEF patients with deep longitudinal clinical data that we have recently whole genome sequenced.
This is by far the largest sequenced cohort of well-characterized HFpEF.

For example, our cohort has an average of 6 serum BNP timepoints per patient that has allowed us to define rates of HF progression by defining the slope of natural log (BNP) vs years. We will use our AI methods to define subsets of HFpEF that progress more quickly than other HFpEF patients and which have poor outcome. While several other studies have attempted to define phenotypic heterogeneity of heart failure patients with older unsupervised ML approaches applied to smaller datasets, we are confident our generative AI models will clearly produce generalizable HFpEF patient stratifications, allowing for subsequent robust identification of causal drug targets and disease biomarkers.

Q6. How does Genuity convey its analysis results to its clients' teams:

Target discovery teams? Drug discovery teams? Clinical trial research teams? To

what extent do these teams have access to the analysis tools for further evaluation?

A. As we are tackling complex problems in the biomedical sciences, it is important to set the stage, as best we can, for success. Thus, we are highly collaborative on all our research and commercial projects. Research project proposals, which constitute highly detailed experimental designs and SOWs are crafted with a great deal of input from our academic and biopharma partners. Data and information along with project updates are shared on a weekly basis in a scientific group meeting forum. Scientists and project managers from both sides of the table share data and project updates as well as discuss findings. Our collaborators are given full access to processed data and analytics over the course of the project. A report, scientifically detailing methods, findings, conclusions, and recommendations is submitted at the close of each project. A decision is usually formalized at that time to collaboratively publish, at least an aspect of our findings, in a peer-reviewed journal.

Genuity Science's biopharma partners may remotely and securely access detailed clinical data, whole genome sequence data, and other omic datasets from patients without personal identifiers. For example, our partners may access our first large batch of 6,000 NASH patients we have recently whole-genome sequenced. Ultimately, we plan to have over 15,000 liver disease patients. All sequence variants, BAM files, QC measures, and variant annotations are stored according to genomic position in the GORdb. Genuity Science's statistical and query tools use that frame of reference, which makes access to data and statistical analysis much faster even with hundreds of thousands of genomes in a single database.

Genuity Science's Sequence Miner interface allows for rapid definition of composite phenotypes and integration with whole genome sequence data and subsequent launch of variant-based and gene-based association algorithms to make disease gene and drug target discoveries. Our cohorts generally have 5 to 10 years of detailed longitudinal data that allows us to define and compare more severe subtypes of disease vs. less severe to find drivers of survival.

About Genuity Science

Genuity Science is a data, analytics and insights organization headquartered in Boston, Massachusetts, with offices in Dublin, Ireland and Reykjavik, Iceland. The company partners with global biopharma and life sciences companies to offer deep end-to-end drug target and biomarker discovery programs aimed at catalyzing precision health and improving the quality of life for patients around the world. Genuity's programs include population-scale, disease-specific data sourcing with detailed longitudinal clinical information, high-quality sequencing, a uniquely scalable genomic and clinical database architecture and tools for analyzing large datasets, and advanced artificial intelligence (AI). For more information, see www.genuitysci.com.

Tom Chittenden, PhD, DPhil, PStat

Dr. Tom Chittenden is Chief Data Science Officer and Founding Director of the Genuity Science Advanced A.I. Research Laboratory. Dr. Chittenden is an Omega Society Fellow and an Accredited Professional Statistician™ with the American Statistical Association. He holds faculty appointments at Boston Children's Hospital and the Harvard Medical School. His work has been published in top-tier scientific journals, including featured articles in Nature and Science. In 2019, Dr. Chittenden was named among the top 100 A.I. Pioneers in Drug Discovery and Advanced Healthcare. He is regarded as one of the world's leading authorities on A.I. and causal statistical machine learning in the biomedical sciences.

Tom holds a PhD in Molecular Cell Biology and Biotechnology from Virginia Tech and a DPhil in Computational Statistics from the University of Oxford. His multidisciplinary postdoctoral training includes experimental investigations in Molecular and Cellular Cardiology from the Dartmouth Medical School; Biostatistics and Computational Biology from the Dana Farber Cancer Institute and the Harvard School of Public Health; and Computational Statistics, Statistical Methodology, and Statistical Machine Learning from the University of Oxford.

Jeff Gulcher, MD, PhD

Dr. Jeff Gulcher is the Chief Scientific Officer for Genuity Science. Before founding Genuity Science, Dr. Gulcher co-founded deCODE genetics. Jeff became Vice President for Research and Development and finally served as the company's Chief Scientific Officer starting in 2003. deCODE was taken public and sold to Amgen for $415 million in 2012. The deCODE launch coincided with a position on staff in the department of neurology at Beth Israel Hospital in Boston and Harvard Medical School, where he served from 1993 to 1998.

Jeff received his PhD and MD from the University of Chicago in 1986 and 1990, respectively, and completed his neurology residency at the Longwood Program of the neurology departments of Brigham and Women's Hospital and Beth Israel Deaconess Medical Center of Harvard Medical School in 1996. He received a bachelor's degree in chemistry/physics from Michigan State University in 1981. He has authored 186 peer-reviewed publications on the genetics of common/complex diseases.

References

1. Endothelial ERK1/2 signaling maintains integrity of the quiescent endothelium, Nicolas Ricard, Rizaldy P Scott, Carmen J Booth, Heino Velazquez, Nicholas A Cilfone, Javier L Baylon, Jeffrey R Gulcher, Susan E Quaggin, Thomas W Chittenden, Michael Simons, J Exp Med 2019 Aug 5; 216(8):1874-1890. doi: 10.1084/ jem.20182151. Epub 2019 Jun 13.

2. Endothelial TGF-β signalling drives vascular inflammation and atherosclerosis, Chen, PY., Qin, L., Li, G. et al. Nat Metab 1, 912–926 (2019). https://doi.org/10.1038/s42255-019-0102-3.

3. Chronic mTOR activation induces a degradative smooth muscle cell phenotype, Guangxin Li et alia, J Clin Invest. 2020; 130(3):1233-1251, https://doi.org/10.1172/JCI131048.

4. Smooth Muscle Cell Reprogramming in Aortic Aneurysms, Pei-Yu Chen, Lingfeng Qin, Guangxin Li, Jose Malagon-Lopez, Zheng Wang, Sonia Bergaya, Sharvari Gujja, Alexander W Caulk, Sae-Il Murtada, Xinbo Zhang, Zhen W Zhuang, Deepak A Rao, Guilin Wang, Zuzana Tobiasova, Bo Jiang, Ruth R Montgomery, Lele Sun, Hongye Sun, Edward A Fisher, Jeffrey R Gulcher, Carlos Fernandez-Hernando, Jay D Humphrey, George Tellides, Thomas W Chittenden, Michael Simons. Cell/Stem Cell, 2020 Apr 2;26(4):542-557.e11. doi: https://doi.org/10.1016/j.stem.2020.02.013.

5. In silico phenotype projection of endothelial ERK1/2 signaling, Michael Simons, Jeffrey R. Gulcher, and Thomas W. Chittenden, Aging (Albany NY). 2020 Jun 15; 12(11): 10001–10003. 2020 Jun 12. doi: 10.18632/aging.103529.

6. Resilience, disease and the age of single cell science, Michael Simons, Pei-Yu Chen, Thomas W Chittenden, Aging (Albany NY), 2020 Feb 10;12(3):2028-2029. https://doi.org/10.18632/aging.102850. Epub 2020 Feb 10. **Ω**

"I work with a lot of scientists, and one of the frustrating things they find is that all this fascinating stuff is being done which doesn't find its way into science fiction. They say, 'Look at the science fact pages - they're so much more imaginative than science fiction.'"

—Geoff Ryman