Computational Approaches to Cancer

A Literature Review for Bioinformaticians, Systems Engineers, and the Impatient

Johan Michalove[*]

February 2026

Abstract

Cancer remains the second leading cause of death globally, killing nearly 10 million people per year. Over the past decade, a convergence of high-throughput sequencing, machine learning, and large-scale data initiatives has transformed oncology from an empirical discipline into a computational one. This review surveys the landscape of computational cancer research across eight domains: genomic variant interpretation, single-cell and spatial transcriptomics, digital pathology, protein structure and drug discovery, liquid biopsy and early detection, large language models in clinical oncology, laboratory automation, and the persistent problem of data equity. For each domain, we describe the state of the art, identify the foundational datasets and models, and highlight open problems where computational expertise is most urgently needed. We argue that the field's most consequential challenges are no longer primarily algorithmic but infrastructural: building systems that are equitable, interoperable, and accessible to the populations that bear the greatest burden of disease.

[*] Department of Information Science, Cornell University. jam844@cornell.edu

Contents

1 Introduction

Cancer is not one disease. It is hundreds of diseases that share a common mechanism: the accumulation of somatic mutations that confer selective advantage to cells, enabling uncontrolled proliferation, immune evasion, and metastatic spread (Hanahan and Weinberg, 2000). The "hallmarks of cancer" framework (Hanahan and Weinberg, 2011; Hanahan, 2022) identifies at least fourteen distinct capabilities that tumor cells acquire through this evolutionary process—from sustaining proliferative signaling to unlocking phenotypic plasticity.

The computational opportunity is enormous. A single tumor genome contains millions of variants. A tumor biopsy processed with single-cell RNA sequencing yields transcriptomic profiles for tens of thousands of individual cells. A whole-slide histopathology image contains approximately 10^{10} pixels. A clinical trial matching problem requires reasoning over thousands of eligibility criteria against a patient's genomic profile, medical history, and demographics.

These are, fundamentally, information processing problems. And they are problems where the gap between data availability and actionable interpretation is growing—not shrinking.

This review is written for computational people who are entering cancer research from adjacent fields: machine learning, systems engineering, information science, robotics, natural language processing. It assumes familiarity with these disciplines and none with biology or medicine beyond a first course. It is organized around the eight domains where computation has the most to offer and the most work to do.

2 Genomic Variant Interpretation

2.1 The Variant Calling Pipeline

The standard cancer genomics pipeline transforms raw sequencing reads into clinically interpretable variants. Tumor and matched normal tissue are sequenced (typically whole-genome or whole-exome), reads are aligned to a reference genome (GRCh38), and somatic variants are called by comparing tumor to normal. The canonical pipeline uses BWA-MEM2 for alignment, GATK Mutect2 for somatic variant calling, and VEP or ANNOVAR for functional annotation (Poplin et al., 2018).

Deep learning has entered variant calling through tools like DeepVariant (Poplin et al., 2018), which frames variant calling as an image classification problem: pileup images of aligned reads are classified by a convolutional neural network. DeepSomatic (Zheng et al., 2025) extends this framework to somatic mutations, supporting whole-genome sequencing, whole-exome sequencing, tumor-normal pairs, tumor-only samples, and formalin-fixed paraffin-embedded (FFPE) tissue across Illumina, PacBio HiFi, and Oxford Nanopore platforms.

2.2 Driver vs. Passenger Discrimination

A tumor may carry 10,000–100,000 somatic mutations. The vast majority are passengers—they occurred during replication but do not contribute to cancer progression. A handful are drivers that confer selective advantage. Distinguishing drivers from passengers remains a core bioinformatics challenge (Chen et al., 2024b).

Current approaches include: frequency-based methods (recurrence across tumors implies driver status), functional impact prediction (SIFT, PolyPhen-2, CADD), network-based approaches (mutations in hub genes of protein interaction networks), and machine learning classifiers trained on known driver/passenger labels. The Cancer Genome Atlas (TCGA) project, spanning 33 cancer types and over 20,000 samples with multi-omic profiling, provides the foundational dataset for driver discovery (The Cancer Genome Atlas Research Network, 2013). The PCAWG consortium extended this to whole-genome analysis of 2,658 tumors across 38 types, identifying 16 structural variant signatures and characterizing chromothripsis as frequently an early event in tumor evolution (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020).

Knowledge bases like COSMIC, ClinVar, and OncoKB annotate known variants with clinical significance and actionability. The AACR Project GENIE consortium provides real-world clinical genomic data from major cancer centers across more than 200,000 samples (AACR Project GENIE Consortium, 2017).

2.3 Open Problems

- **Low variant allele frequency (VAF) detection**: Subclonal driver mutations at $<5\%$ VAF are difficult to distinguish from sequencing artifacts. Better error models and deeper sequencing are needed.

- **Structural variant calling**: Large rearrangements, copy number alterations, and chromothripsis events are poorly captured by short-read sequencing. Long-read technologies (PacBio, Oxford Nanopore) are improving this.
- **Non-coding variant interpretation**: >98% of the genome is non-coding. Regulatory variants that drive cancer through altered gene expression are largely uncharacterized.
- **Variant interpretation across populations**: A variant classified as "of uncertain significance" (VUS) may only be uncertain because the reference population is too narrow—a point we return to in Section 11.

3 Single-Cell and Spatial Transcriptomics

3.1 Single-Cell Revolution

Bulk RNA sequencing measures average gene expression across millions of cells. Single-cell RNA sequencing (scRNA-seq) resolves heterogeneity within a tumor at the individual cell level, revealing subclonal structure, immune microenvironment composition, and cell state transitions that are invisible to bulk methods.

The standard computational pipeline involves: quality control and filtering, normalization, dimensionality reduction (PCA, UMAP), clustering, differential expression between clusters, and cell type annotation. Foundational tools include Seurat (R), Scanpy (Python), and scVI (Lopez et al., 2018), a deep generative model that provides a probabilistic framework for normalization, batch correction, and differential expression simultaneously.

3.2 Foundation Models for Single Cells

The scale of single-cell data (tens of millions of cells across atlases like the Human Cell Atlas) has enabled foundation models:

- **scGPT** (Cui et al., 2024): A generative pre-trained transformer for single-cell multiomics. Pre-trained on >33 million cells. Achieves state-of-the-art performance on cell type annotation, perturbation prediction, and multi-omic integration.
- **Geneformer** (Theodoris et al., 2023): Pre-trained on ∼30 million single-cell transcriptomes from Genecorpus-30M. Uses rank-value encoding of gene expression. Transfer

learning enables predictions of gene dosage sensitivity, chromatin dynamics, and network hierarchy.

- **CellTypist** (Domínguez Conde et al., 2022): A scalable cell type classifier using logistic regression over curated reference atlases. Less architecturally ambitious than the transformer models but pragmatically effective for routine annotation tasks.

3.3 Spatial Transcriptomics

Spatial methods (10x Visium, MERFISH, Slide-seq) add positional information: which cells express which genes *where* within the tissue architecture. This enables analysis of cell-cell communication, spatial organization of the tumor microenvironment, and identification of niches that drive resistance.

Gene regulatory network inference from spatial data is an emerging frontier. SCENIC+ (González-Blas et al., 2023) integrates single-cell transcriptomics with chromatin accessibility to reconstruct enhancer-driven gene regulatory networks, providing mechanistic insight into how cancer cells rewire their regulatory programs.

3.4 Open Problems

- **Scalability**: Atlases are growing to billions of cells. Foundation models must scale accordingly.
- **Perturbation prediction**: Can we predict how a cell will respond to a drug *in silico*? scGPT and Geneformer show early promise, but generalization to unseen perturbations is unresolved.
- **Temporal dynamics**: Current snapshots are static. Inferring trajectories (pseudotime, RNA velocity) from single timepoints introduces substantial uncertainty.
- **Multi-modal integration**: Combining transcriptomic, epigenomic, and proteomic measurements from the same cell is technically possible but computationally challenging.

4 Digital Pathology and Computer Vision

4.1 The Problem

Histopathological examination of tissue biopsies is the gold standard for cancer diagnosis. Pathologists examine hematoxylin and eosin (H&E) stained whole-slide images (WSIs) at 40× magnification, producing gigapixel images ($\sim$100,000 × 100,000 pixels). This process is subjective, time-consuming, and bottlenecked by a global shortage of pathologists.

Computational pathology applies computer vision to WSIs for: tumor detection, grading, subtyping, biomarker prediction, treatment response prediction, and prognosis estimation. The multiple instance learning (MIL) paradigm—treating a slide as a "bag" of thousands of tile-level "instances" with a single slide-level label—has been the dominant framework since CLAM (Lu et al., 2021).

4.2 Foundation Models

2024 marked the "ImageNet moment" for computational pathology with several foundation models trained on millions of WSIs:

- **UNI** (Chen et al., 2024a): A vision transformer (ViT-L) pre-trained on $>$100 million tissue patches from $>$100,000 WSIs across 20 tissue types. Self-supervised learning (DINOv2). General-purpose feature extractor for downstream tasks.
- **CONCH** (Lu et al., 2024): A vision-language foundation model that jointly learns from histopathology images and associated text (pathology reports, captions). Enables zero-shot and few-shot classification, natural language queries over tissue morphology.
- **Virchow** (Vorontsov et al., 2024): Trained on 1.5 million WSIs from Memorial Sloan Kettering. Largest pathology-specific foundation model. Demonstrates clinical-grade performance on rare cancer detection.
- **Prov-GigaPath** (Xu et al., 2024): A whole-slide foundation model from Microsoft Research trained on 170,000 WSIs. Uses a novel slide-level learning objective that operates on the entire slide rather than individual tiles.

4.3 Clinical Validation

Paige AI received the first FDA approval for an AI-based pathology diagnostic in 2021 (prostate cancer detection). PathAI has FDA-cleared tools for PD-L1 scoring (immunotherapy biomarker). These represent the translation frontier: moving from research benchmarks to clinical deployment.

4.4 Open Problems

- **Predicting molecular status from morphology**: Can H&E alone predict ER/PR/HER2 status, MSI status, or specific mutations? Early results suggest yes for some biomarkers, but clinical-grade accuracy is not yet achieved for most.
- **Multi-modal integration**: Combining pathology images with genomic, transcriptomic, and clinical data in a unified model.
- **Generalization**: Models trained on academic medical center data may not generalize to community hospitals, different staining protocols, or scanner types.
- **Explainability**: Pathologists need to understand *why* a model makes a prediction. Attention maps and spatial explanations are active research areas.

5 Protein Structure, Drug Discovery, and Generative Chemistry

5.1 The Structure Revolution

AlphaFold (Jumper et al., 2021) solved the protein structure prediction problem in 2020, achieving experimental accuracy from amino acid sequence alone. AlphaFold 3 (Abramson et al., 2024) extends this to biomolecular complexes: protein-protein, protein-DNA, protein-RNA, and protein-ligand interactions. The AlphaFold Protein Structure Database contains predicted structures for $\sim$200 million proteins.

Protein language models complement structure prediction. ESM-2 (Lin et al., 2023), trained on 65 million protein sequences with 15 billion parameters, learns evolutionary constraints that enable structure prediction, function annotation, and variant effect prediction from sequence alone. Evo (Nguyen et al., 2024) extends language modeling to entire genomes, predicting gene essentiality and regulatory function at the DNA level.

5.2 Drug Discovery

Structure-based drug discovery uses protein 3D structures to identify and optimize small molecules that bind to specific targets. Traditional approaches—molecular docking, molecular dynamics simulation—are computationally expensive and often inaccurate. Deep learning is transforming each stage:

- **Virtual screening**: Given a target protein, score millions of candidate molecules for binding affinity. Neural network scoring functions (e.g., gnina, DeepDTA) outperform traditional force-field-based methods on many benchmarks.
- **Molecular docking**: DiffDock (Corso et al., 2023) uses diffusion models over the space of ligand poses to predict binding conformations. Achieves state-of-the-art on PoseBusters benchmark.
- **De novo molecular generation**: Generative models (VAEs, GANs, diffusion models) propose novel molecules with desired properties. RFdiffusion (Watson et al., 2023) designs novel protein structures by denoising from random coordinates.
- **Drug repurposing**: Network-based and knowledge graph approaches identify existing approved drugs that may be effective against new targets. The discovery of halicin, a novel antibiotic, through deep learning screening of the Drug Repurposing Hub demonstrated the approach's potential (Stokes et al., 2020).
- **End-to-end AI-powered discovery**: Ren et al. (2023) demonstrated the full pipeline: using AlphaFold2-predicted protein structures for hit identification, they discovered a novel CDK20 inhibitor for hepatocellular carcinoma within 30 days of target selection, synthesizing only 7 compounds—the first successful use of AlphaFold-predicted structures for hit identification against a novel target.

5.3 Protein Design

The inverse of structure prediction: given a desired function, design a protein sequence that folds into a structure performing that function. RFdiffusion (Watson et al., 2023) generates protein backbones by diffusion, and inverse folding models (ProteinMPNN) design sequences that fold into those backbones. This enables the design of novel therapeutics including: antibodies, enzymes, receptors, and vaccine antigens.

5.4 Open Problems

- **Binding affinity prediction**: Ranking compounds by binding strength remains unreliable. Current models are better at predicting *whether* a molecule binds than *how strongly*.
- **ADMET prediction**: Absorption, distribution, metabolism, excretion, and toxicity properties determine whether a molecule becomes a drug. Predicting these from structure alone is an unsolved problem.
- **Conformational dynamics**: Proteins are not static. Capturing the ensemble of conformations relevant to drug binding requires either very long molecular dynamics simulations or new ML approaches.
- **Closing the loop**: Computational predictions must be validated experimentally. The design-make-test-analyze cycle remains slow. Self-driving labs (Section 9) aim to accelerate it.

6 Liquid Biopsy and Early Detection

6.1 Circulating Tumor DNA

Tumors shed cell-free DNA (cfDNA) into the bloodstream. Circulating tumor DNA (ctDNA) carries the somatic mutations of the tumor and can be detected through a simple blood draw—a "liquid biopsy." The promise is threefold: early detection (find cancer before symptoms), treatment monitoring (track response without tissue biopsies), and minimal residual disease detection (is any cancer left after surgery?).

The computational challenge is extreme signal-in-noise. In early-stage cancer, ctDNA may constitute <0.01% of total cfDNA. Distinguishing true tumor-derived fragments from sequencing errors, clonal hematopoiesis of indeterminate potential (CHIP), and other biological noise requires sophisticated error suppression and statistical frameworks.

6.2 Multi-Cancer Early Detection

Grail's Galleri test uses targeted methylation sequencing of cfDNA to detect over 50 cancer types from a single blood draw. The CCGA (Circulating Cell-free Genome Atlas) validation study (Klein et al., 2021) demonstrated 51.5% sensitivity across all cancer types at 99.5%

specificity, with cancer signal origin prediction accuracy of 88.7%. The PATHFINDER study (Schrag et al., 2023) prospectively screened 6,621 adults, detecting cancer signals in 1.4% of participants; 48% of confirmed cancers were Stage I or II, and 74% lacked existing recommended screening tests.

Fragmentomics—analyzing cfDNA fragmentation patterns rather than sequences—provides an orthogonal signal. Cristiano et al. (2019) showed that genome-wide cfDNA fragmentation profiles differ between cancer patients and healthy controls, achieving 73–98% sensitivity at 98% specificity across seven cancer types.

6.3 Open Problems

- **Sensitivity for early-stage disease**: Stage I cancers shed the least ctDNA and are the most curable. Improving early-stage sensitivity is the highest-impact problem.
- **Overdiagnosis**: Detecting cancers that would never become clinically significant. Not every detected signal requires treatment. Predicting indolent vs. aggressive disease from cfDNA signatures is unresolved.
- **Multi-modal integration**: Combining methylation, fragmentation, mutation, and protein biomarker signals into a single classifier.
- **Longitudinal modeling**: Tracking ctDNA dynamics over time to predict relapse before clinical recurrence.
- **Cost and access**: Galleri costs $\sim$\$949 per test and is not yet covered by insurance. Equitable access requires dramatic cost reduction.

7 Large Language Models in Clinical Oncology

7.1 Clinical Reasoning

Large language models (LLMs) have demonstrated surprising proficiency on medical knowledge benchmarks. Med-PaLM 2 (Singhal et al., 2023) achieved expert-level performance on USMLE-style questions. GPT-4 passed all three steps of USMLE and performed at or above the level of third-year medical students on clinical reasoning tasks (Nori et al., 2023).

In oncology specifically, LLMs are being applied to:

- **Clinical trial matching**: TrialGPT (Jin et al., 2024) uses LLMs to match patients to clinical trials by parsing complex eligibility criteria in natural language and reasoning about patient records. This is a genuine bottleneck: fewer than 5% of adult cancer patients enroll in clinical trials, and many eligible patients are never identified.
- **Genomic variant interpretation**: LLMs can synthesize evidence from multiple databases (ClinVar, OncoKB, COSMIC, literature) to generate variant interpretation reports, reducing the burden on molecular tumor boards.
- **Radiology and pathology report generation**: Automating the narrative interpretation of imaging and histology findings.
- **Patient-facing communication**: Translating complex genomic reports into language patients can understand.

7.2 Cautions

LLMs hallucinate. In clinical contexts, hallucination can be lethal. The systematic evaluation of LLMs in oncology is in its infancy, and none are FDA-cleared for clinical decision support. The gap between benchmark performance and safe clinical deployment is substantial and requires rigorous prospective validation, integration with existing clinical workflows, and robust uncertainty quantification.

7.3 Open Problems

- **Grounding in evidence**: LLMs must cite specific evidence for clinical claims. Retrieval-augmented generation (RAG) over curated medical knowledge bases is the leading approach.
- **Uncertainty quantification**: A model must know when it does not know. Calibrated confidence estimates for clinical predictions are essential.
- **Regulatory pathway**: How should AI-assisted clinical decisions be regulated? The FDA is developing frameworks, but the landscape is evolving rapidly.
- **Integration with EHR systems**: Practical deployment requires seamless integration with electronic health records, which are notoriously heterogeneous (Epic, Cerner, etc.).

8 Immunotherapy, Neoantigens, and Cancer Vaccines

8.1 The Computational Pipeline

Immunotherapy—particularly immune checkpoint inhibitors targeting PD-1/PD-L1 and CTLA-4—has transformed the treatment of melanoma, non-small cell lung cancer, and a growing number of other malignancies (Ribas and Wolchok, 2018). CAR-T cell therapy has achieved durable remissions in certain blood cancers (June et al., 2018). The computational dimension of immunotherapy centers on three problems: predicting which patients will respond, predicting which neoantigens to target, and designing personalized vaccines.

8.2 Neoantigen Prediction

Somatic mutations produce novel peptides (neoantigens) that can be presented on the cell surface by MHC molecules and recognized by T-cells. The computational pipeline involves: somatic variant calling, HLA typing (each patient expresses a unique set of MHC alleles), peptide-MHC binding prediction, and immunogenicity scoring.

MHCflurry 2.0 (O'Donnell et al., 2020) is an ensemble of neural networks supporting $\sim$15,000 MHC class I alleles, trained on mass spectrometry eluted ligand data. Protein language models fine-tuned for peptide-MHC binding now outperform traditional sequence-based methods, though predicting whether a bound neoantigen will actually trigger a T-cell response remains unreliable.

8.3 Personalized Cancer Vaccines

mRNA neoantigen vaccines represent the most direct application of computational neoantigen prediction. Rojas et al. (2023) reported results from a BioNTech/Memorial Sloan Kettering trial in pancreatic cancer: at 3-year follow-up, 6 of 8 patients with vaccine-induced immune responses remained disease-free, versus 7 of 8 relapse in non-responders. Over 80% of vaccine-induced T-cells were detectable at 3 years. More than 120 personalized cancer vaccine trials are now underway, though manufacturing costs remain above $100,000 per patient.

8.4 Biomarker Prediction

Tumor mutational burden (TMB) and microsatellite instability (MSI-H) are FDA-recognized biomarkers for checkpoint inhibitor eligibility, but single biomarkers are insufficient predictors. Vanguri et al. (2022) demonstrated that multimodal integration of radiology, pathology, and genomics achieves AUC = 0.80 for predicting immunotherapy response in non-small cell lung cancer—significantly outperforming any single modality. Chen et al. (2022) developed PORPOISE, a pan-cancer multimodal deep learning platform that fuses whole-slide images with molecular profiles across 14 cancer types.

8.5 Open Problems

- **Immunogenicity prediction**: Most computationally predicted neoantigens fail to elicit T-cell responses. The gap between MHC binding prediction (largely solved) and immunogenicity prediction (largely unsolved) is the field's central bottleneck.
- **Tumor microenvironment modeling**: Immune response depends on the spatial organization of immune cells, cytokines, and vasculature within the tumor. Spatial transcriptomics is beginning to enable this but computational frameworks are immature.
- **Vaccine manufacturing**: Compressing the timeline from tumor biopsy to personalized vaccine from weeks to days requires computational optimization of the entire pipeline, from neoantigen selection to mRNA sequence design to manufacturing protocols.
- **Solid tumor resistance**: CAR-T therapy works for blood cancers but faces fundamental challenges in solid tumors: immunosuppressive microenvironment, target antigen heterogeneity, and T-cell trafficking. Computational modeling of these barriers is an open frontier.

9 Laboratory Automation and Self-Driving Labs

9.1 The Bottleneck

Computational predictions are only as good as the experimental validation that follows. The design-make-test-analyze (DMTA) cycle in drug discovery typically takes weeks to months per iteration. Automating the “make” and “test” steps—robotic cell culture, automated assay execution, autonomous experimental design—could accelerate this cycle by an order of magnitude.

9.2 Self-Driving Labs

Self-driving laboratories (SDLs) integrate robotic hardware, automated decision-making, and closed-loop experimental optimization (Abolhasani and Kumacheva, 2023). Burger et al. (2020) demonstrated a mobile robot chemist that autonomously performed 688 experiments over 8 days to optimize a photocatalyst, navigating a 10-dimensional search space. In cancer research, analogous systems could automate: drug sensitivity screening, combinatorial therapy optimization, and CRISPR-based functional genomics screens.

9.3 Cloud Labs

Companies like Emerald Cloud Lab and Strateos offer laboratory-as-a-service platforms where experiments are specified programmatically and executed by robotic infrastructure. This decouples computational expertise from physical laboratory access—a researcher at any institution can run experiments through an API. PyLabRobot provides an open-source Python framework for controlling liquid handlers and other lab equipment.

9.4 Open Problems

- **Biological variability**: Chemical synthesis is more reproducible than biological assays. Cell culture, in particular, introduces variability that robotic systems must account for.
- **Cost**: Automated laboratory infrastructure is expensive. Access is limited to well-funded institutions and companies.
- **Active learning**: Designing the next experiment to maximize information gain is an active learning / Bayesian optimization problem that scales poorly with dimensionality.
- **Integration**: Connecting computational models (drug design, resistance prediction) to robotic execution in a closed loop is an engineering challenge as much as a scientific one.

10 Multi-Omics Integration

Cancer is not explained by any single molecular layer. Mutations alter gene expression, which alters protein levels, which alters metabolic flux, which alters cellular behavior. Understanding cancer requires integrating data across the central dogma.

10.1 Methods

Multi-Omics Factor Analysis (MOFA) (Argelaguet et al., 2018) provides an unsupervised framework for identifying shared and data-type-specific axes of variation across multiple omics layers. Cantini et al. (2021) benchmarked 14 joint dimensionality reduction methods on cancer multi-omics data, finding that matrix factorization approaches (MOFA, iCluster+) and deep learning methods (autoencoder-based) each have domain-specific advantages.

Biologically informed neural networks represent a promising direction. Elmarakeby et al. (2021) built P-NET, a neural network whose architecture mirrors the biological hierarchy from genes to pathways to biological processes. Trained on multi-omic prostate cancer data, P-NET predicted treatment resistance and identified interpretable biological mechanisms—a rare instance of a deep learning model that is both performant and biologically interpretable.

10.2 Open Problems

- **Missing data**: Patients rarely have all omics layers measured. Imputation and transfer learning across modalities are active research areas.
- **Temporal integration**: Most multi-omics studies are cross-sectional. Longitudinal multi-omics data (tracking a patient through treatment) would enable causal modeling but is expensive and rare.
- **Clinical utility**: Multi-omics models outperform single-omics models on research benchmarks. Whether this translates to better clinical outcomes in prospective studies is not yet established.

11 Data Equity and Algorithmic Fairness

11.1 The Representation Problem

TCGA—the single most important dataset in cancer genomics—is approximately 77% white, 12% Black, 3% Asian, and 3% Hispanic (Spratt et al., 2016). This demographic composition does not reflect the U.S. cancer burden, let alone the global one. Every model trained on TCGA inherits this bias.

The consequences are concrete. Polygenic risk scores (PRS) developed on European-ancestry populations perform significantly worse—and sometimes in the wrong direction—when applied to African, Asian, and Latin American populations (Martin et al., 2019). A variant classified as "of uncertain significance" may only be uncertain because it has not been observed in a sufficiently diverse reference panel.

Triple-negative breast cancer (TNBC)—the subtype with the worst prognosis, the fewest targeted therapies, and the greatest need for computational research—disproportionately affects Black women. The populations with the greatest burden of disease are the least represented in the data from which precision medicine is derived.

11.2 Algorithmic Bias

Obermeyer et al. (2019) demonstrated that a widely used commercial algorithm for allocating healthcare resources exhibited significant racial bias: at the same risk score, Black patients were considerably sicker than white patients, because the algorithm used healthcare spending as a proxy for health needs, and systemic inequities in healthcare access meant Black patients had lower spending at equivalent levels of illness.

In computational pathology, Vaidya et al. (2024) demonstrated that standard deep learning models exhibited diagnostic biases in 29.3% of tasks across demographic groups, with performance gaps of 3–16% between white and Black patients depending on the task. Applying fairness-aware training frameworks mitigated 88.5% of these disparities—but only when demographic annotations were available, which they often are not. In genomic variant interpretation, the underrepresentation of non-European populations in reference databases directly harms diagnostic accuracy for those populations.

11.3 Federated Learning

Federated learning trains models across distributed datasets without centralizing patient data. Warnat-Herresthal et al. (2021) introduced swarm learning, a decentralized architecture that combines federated learning with blockchain-based peer coordination, demonstrating it on blood cancer transcriptomics. Pati et al. (2022) applied federated learning to brain tumor segmentation across 71 institutions in 6 continents, showing that federated models matched or exceeded those trained on pooled data.

Federated approaches could help address the representation problem: institutions in underrepresented regions can contribute to model training without surrendering patient data. But federation is a technical solution to what is partly a political problem. Data sovereignty frameworks—who controls the data, who benefits from the models, who decides how they are used—require governance, not just cryptography.

11.4 FAIR and CARE Principles

The FAIR principles (Wilkinson et al., 2016)—Findable, Accessible, Interoperable, Reusable—provide a framework for scientific data management. But FAIR was designed for data sharing, not data sovereignty. The CARE principles—Collective Benefit, Authority to Control, Responsibility, Ethics—complement FAIR by centering the rights of data subjects, particularly indigenous and marginalized communities.

In cancer genomics, the tension between FAIR and CARE is acute. Making genomic data maximally accessible accelerates research but may violate the preferences of the communities from whom the data was collected. Resolving this tension requires participatory governance structures that include patient and community voices in data management decisions.

11.5 Open Problems

- **Diverse reference panels**: Building comprehensive variant frequency databases for non-European populations is a prerequisite for equitable precision oncology.
- **Fairness-aware model development**: Techniques for training models that perform equitably across demographic groups, not just on average.
- **Community-governed data commons**: Institutional frameworks that enable data

sharing while respecting community sovereignty.

- **Global infrastructure**: Computational infrastructure for cancer genomics in low- and middle-income countries. The data gap is a death gap.

12 Discussion: From Algorithms to Infrastructure

The eight domains surveyed in this review share a common trajectory. In each, the foundational algorithmic problems—variant calling, image classification, structure prediction, sequence modeling—have seen dramatic progress over the past five years, often driven by the transfer of techniques from the broader machine learning community (transformers, diffusion models, large-scale pre-training, foundation models).

But in each domain, the binding constraint is shifting from *algorithms* to *infrastructure*. The most consequential open problems are not "build a better model" but "build a system that works":

- A liquid biopsy test that costs $949 and is not covered by insurance does not serve the populations most at risk.
- A foundation model for pathology trained on tissue from academic medical centers does not generalize to community hospitals in sub-Saharan Africa.
- A clinical trial matching system that requires structured EHR data does not help patients at institutions with paper records.
- A self-driving lab that costs millions of dollars does not accelerate drug discovery for neglected tropical diseases.

This is not a new observation. Topol (2019) and Rajpurkar et al. (2022) have articulated the gap between AI research and clinical deployment. But the gap is widening because the models are getting more capable while the infrastructure for deploying them equitably is not keeping pace.

For computational people entering cancer research, the implication is clear: the most impactful contributions may not be the most technically glamorous. Building interoperable data pipelines, developing quality control systems for clinical deployment, creating federated learning infrastructure that actually works across institutions with heterogeneous IT systems, and designing governance frameworks for data sovereignty—these are systems engineering problems, and they are where the field needs the most help.

12.1 The Integration Challenge

No single computational modality is sufficient. The patient with breast cancer has: a genome, a transcriptome, a set of histopathology slides, a series of imaging studies, a longitudinal medical record, a demographic profile, a set of social determinants of health, and a personal preference for treatment. Precision oncology aspires to integrate all of these into an individualized treatment recommendation.

We are far from this vision. Current practice involves molecular tumor boards—multidisciplinary teams of oncologists, pathologists, geneticists, and bioinformaticians who manually integrate available data to make treatment decisions. Computational systems that can assist or partially automate this integration, while maintaining interpretability and accountability, represent the next frontier.

12.2 A Note on Reproducibility

Cancer bioinformatics suffers from the same reproducibility challenges as the broader machine learning field, amplified by biological variability. Benchmark datasets may not reflect clinical reality. Pre-processing choices (normalization, filtering, quality control) can dramatically affect results. Model performance on curated benchmarks may not translate to prospective clinical performance.

The field would benefit from: standardized benchmark suites with clinical relevance, mandatory code and data sharing for publications, prospective clinical validation as a requirement for clinical claims, and transparent reporting of failure modes and demographic performance disparities.

13 Conclusion

Cancer bioinformatics is at an inflection point. The convergence of high-throughput sequencing, foundation models, and large-scale data initiatives has created unprecedented opportunities for computational contributions. AlphaFold has solved protein structure prediction. Foundation models for pathology are approaching clinical grade. Liquid biopsy enables cancer detection from a blood draw. Single-cell technologies reveal intratumoral heterogeneity at single-cell resolution. LLMs can parse clinical trial eligibility criteria.

Self-driving labs can automate experimental validation.

The bottleneck has shifted. The most urgent needs are no longer algorithmic breakthroughs but infrastructure: equitable data collection, interoperable systems, federated learning that works across institutional boundaries, cost reduction for clinical deployment, and governance frameworks that respect patient and community sovereignty.

Cancer kills 10 million people per year. The computational tools to reduce that number exist or are within reach. What remains is the engineering, the infrastructure, and the political will to deploy them equitably. For computational people entering this field: the problems are hard, the data is messy, the stakes are absolute, and the work is urgent.

For Ariel, Bea, and those we lost too soon.

References

AACR Project GENIE Consortium (2017). AACR project GENIE: Powering precision medicine through an international consortium. *Cancer Discovery*, 7(8):818–831.

Abolhasani, M. and Kumacheva, E. (2023). The rise of self-driving labs in chemical and materials sciences. *Nature Synthesis*, 2:483–492.

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630:493–500.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marber, J. C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):e8124.

Burger, B., Maffettone, P. M., Gusber, V. V., Ber, C. M., Gryber, G. R., Stabersson, A. J., and Cooper, A. I. (2020). A mobile robotic chemist. *Nature*, 583:237–241.

Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., and Baudot, A. (2021). Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications*, 12:124.

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Song, A. H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al. (2024a). Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30:850–862.

Chen, R. J., Lu, M. Y., Williamson, D. F. K., et al. (2022). Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878.

Chen, S., Francioli, L. C., Goodrich, J. K., Collins, R. L., Kanai, M., Wang, Q., Alföldi, J., Watts, N. A., Vittal, C., Gauthier, L. D., et al. (2024b). A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*, 625:92–100.

Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. (2023). DiffDock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*.

Cristiano, S., Leal, A., Phallen, J., Fiksel, J., Adleff, V., Bruhm, D. C., Jensen, S. Ø., Medina, J. E., Hruban, C., White, J. R., et al. (2019). Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*, 570:385–389.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. (2024). scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21:1470–1480.

Domínguez Conde, C., Xu, C., Jarvis, L. B., Rainbow, D. B., Wells, S. B., Gomes, T., Howlett, S. K., Sherber, O., Gould, J., Kenber, N., et al. (2022). Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594).

Elmarakeby, H. A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., AlDubayan, S. H., Salber, S., Goldfarb, S., Schmit, S. L., et al. (2021). Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598:348–352.

González-Blas, C. B., De Winter, S., Hulselmans, G., Hecker, N., Shalomi, R., and Aerts, S. (2023). SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nature Methods*, 20:1355–1367.

Hanahan, D. (2022). Hallmarks of cancer: New dimensions. *Cancer Discovery*, 12(1):31–46.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100(1):57–70.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. *Nature*, 578:82–93.

Jin, Q., Wang, Z., Floudas, C. S., Chen, F., Gong, C., Bracber, D., Boursi, B., Wang, Y., Wen, H., Morris, J. C., et al. (2024). TrialGPT: Matching patients to clinical trials with large language models. *Nature Communications*, 15:1–12.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589.

June, C. H., O'Connor, R. S., Kawalekar, O. U., Ghassemi, S., and Milone, M. C. (2018). CAR T cell immunotherapy for human cancer. *Science*, 359(6382):1361–1365.

Klein, E. A., Richards, D., Cohn, A., Tummala, M., Lapham, R., Cosgrove, D., Chung, G., Pesber, J., Hubbell, E., Yeatman, T., et al. (2021). Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Annals of Oncology*, 32(9):1167–1177.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smeaton, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15:1053–1058.

Lu, M. Y., Chen, B., Williamson, D. F. K., Chen, R. J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L. P., Gerber, G., et al. (2024). A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874.

Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. In *Nature Biomedical Engineering*, volume 5, pages 555–570.

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51:584–591.

Nguyen, E., Poli, M., Faez, M., Massaroli, S., Briber, D., Chen, B., Hyun, R., Dao, T., and Ré, C. (2024). Sequence modeling and design from molecular to genome scale with Evo. *Science*, 386(6723).

Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

O'Donnell, T. J., Rubinsteyn, A., and Laserson, U. (2020). MHCflurry 2.0: Improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Systems*, 11(1):42–48.

Pati, S., Baid, U., Edwards, B., Sheller, M., Wang, S.-H., Reina, G. A., Foley, P., Grber, A., Myronenko, A., Adewole, K., et al. (2022). Federated learning enables big data for rare cancer boundary detection. *Nature Communications*, 13:7346.

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36:983–987.

Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28:31–38.

Ren, F., Ding, X., Zheng, M., et al. (2023). AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor. *Chemical Science*, 14:1443–1452.

Ribas, A. and Wolchok, J. D. (2018). Cancer immunotherapy using checkpoint blockade. *Science*, 359(6382):1350–1355.

Rojas, L. A., Sethna, Z., Soares, K. C., et al. (2023). Personalized RNA neoantigen vaccines stimulate T cells in pancreatic cancer. *Nature*, 618:144–150.

Schrag, D., Beer, T. M., McDonnell, C. H., Nadauld, L., Dilaveri, C. A., Reid, R., et al. (2023). Blood-based tests for multicancer early detection (PATHFINDER): a prospective cohort study. *The Lancet*, 402:1251–1260.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023). Large language models encode clinical knowledge. *Nature*, 620:172–180.

Spratt, D. E., Chan, T., Waldron, L., Speers, C., Feng, F. Y., Ozel Demirel, D., Jansen, Y., Simko, J., Efstathiou, J. A., Davicioni, E., et al. (2016). Racial/ethnic disparities in genomic sequencing. *JAMA Oncology*, 2(8):1070–1074.

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., et al. (2020). A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702.

The Cancer Genome Atlas Research Network (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45:1113–1120.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Manber, H., Ravber, D., Kamber, R. A., Broadbent, S. E., et al. (2023). Transfer learning enables predictions in network biology. *Nature*, 618:616–624.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25:44–56.

Vaidya, A., Chen, R. J., Williamson, D. F. K., Song, A. H., Jaume, G., Yang, Y., et al. (2024). Demographic bias in misdiagnosis by computational pathology models. *Nature Medicine*, 30:1174–1190.

Vanguri, R. S., Luo, J., Aukerman, A. T., et al. (2022). Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. *Nature Cancer*, 3:1151–1164.

Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., et al. (2024). A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30:2924–2935.

Warnat-Herresthal, S., Schultze, H., Shastry, K. L., Manamohan, S., Muber, S., Garg, V., Srinivasan, R., Bavarva, H., Chillare, C., Ramaswamy, K., et al. (2021). Swarm learning for decentralized and confidential clinical machine learning. *Nature*, 594:265–270.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borber, A. J., Ragotte, R. J., Milles, L. F., et al. (2023). De novo design of protein structure and function with RFdiffusion. *Nature*, 620:1089–1100.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Nature Scientific Data*, 3:160018.

Xu, H., Usuyama, N., Bagber, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al. (2024). A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630:181–188.

Zheng, Z., Li, S., Su, J., et al. (2025). Accurate somatic small variant discovery for multiple sequencing technologies with DeepSomatic. *Nature Biotechnology*.