

Genomics and Precision Medicine

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 15 (enrichment reading) of "Healthcare Analytics and AI: Building Systems That Actually Work"
The essentials: what to know, why it matters at the bedside, and how models fail.

One-sentence thesis. The math of precision medicine is sound and the models are dazzling, but the bottleneck is systemic (diverse training data, workflow integration, equitable access), and a genomic model is a hypothesis generator, not an oracle.

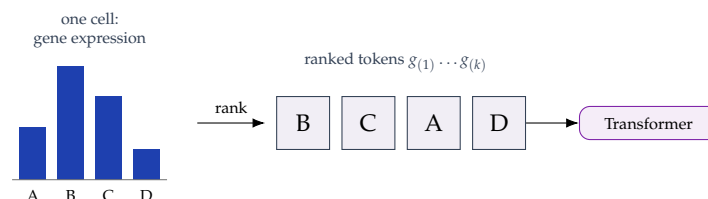
PART 1: Genomic Foundation Models

Cells as Sentences

The trick behind single-cell foundation models is a borrowed one: treat biology like language. A cell is a sentence, a gene is a word, and its expression level is the meaning that word carries in context. Geneformer (Theodoris et al., 2023), pretrained on 30 million single-cell transcriptomes, takes this literally through rank-value encoding: instead of feeding raw expression counts, it ranks the genes within each cell and feeds only the ordering, the tokens $g_{(1)}, g_{(2)}, \dots, g_{(k)}$. The point is that it listens for which voices are loudest in the cellular chorus rather than trusting the exact

decibel reading from every microphone.

That choice buys robustness. Absolute counts drift with sequencing chemistry, but the relative ordering of the most active genes is far more stable, so ranking makes the model resistant to batch effects, the pervasive problem that stops a model trained at one lab from working on another lab's samples. The payoff is real: Geneformer's in silico perturbation predicted gene knockdowns for hypertrophic cardiomyopathy that were then validated by CRISPR deletion in the lab, with one target (mTOR modulation) entering a Phase 2 trial by 2025.



Geneformer's rank-value encoding. Within each cell, genes are ordered by expression and only the ranking is fed to the model. Absolute counts drift with sequencing chemistry; the ordering of the loudest genes is stable, which makes the model robust to batch effects.

Association Is Not Causation

It is tempting to assume a model this fluent in cell biology must understand it. It does not. Sparse-autoencoder analysis of Geneformer and scGPT (Kendiukhov, 2026) found rich biological organization, 29 to 59 percent of features map to known pathways and protein interactions, but minimal causal logic: tested against genome-scale CRISPRi data, only **6.2 percent** of transcription factors showed regulatory-target-specific responses. The model owns a detailed street map of the city but does not understand the traffic lights, road closures, and one-way rules that actually determine how cars move. It knows which

genes co-occur; it does not reliably know which genes control which.

That distinction sets the use cases. For annotation or batch correction, co-expression is enough. For predicting the downstream effect of a gene perturbation, the use case for drug-target discovery, these models capture associations, not mechanisms, and when benchmarked against simple linear baselines they did not win. **Watch out.** Treat every perturbation prediction as a hypothesis that must be validated experimentally; the gap between organized biological knowledge and causal regulatory logic is the gap between annotation and intervention.

The Forward and Inverse Protein Problems

Predicting a protein's 3D shape from its amino-acid sequence, the folding problem, was a fifty-year grand challenge because the search space is astronomical: a single 100-residue protein can fold roughly 10^{47} ways. AlphaFold2 (Jumper et al., 2021) solved this *forward* problem with near-experimental accuracy, and by 2024 its database held over 200 million predicted structures. David Baker's RFDiffusion tackles the *inverse* problem: design a sequence that folds into a shape you specify.

ANALOGY

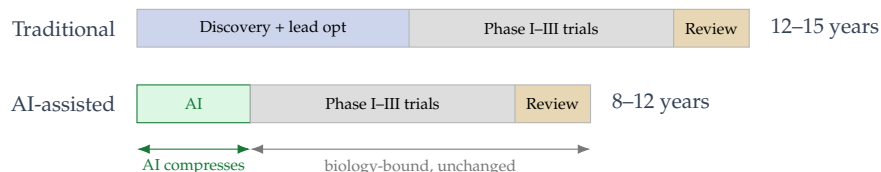
AlphaFold asks, "Given this recipe, what cake will it bake?" RFDiffusion asks, "If I need a cake with this shape and function, what recipe should I write?"

Structure matters clinically because most drugs work by binding a protein, so atomic-resolution shape is the foundation of rational drug design. AlphaFold3 (2024) extended prediction to protein-drug and protein-DNA complexes, and generative pipelines have produced picomolar binders and human-like

antibodies for the 200-billion-dollar antibody market. **Watch out.** AlphaFold predictions are static snapshots while real proteins are dynamic, and RFDiffusion designs are computationally validated only. Model confidence is not experimental truth.

The Timeline-Compression Reality Check

The headlines say AI will collapse drug discovery from decades to months. The reality is more surgical: AI compresses the pipeline *asymmetrically*. It shortens the front end, target identification and lead optimization, which are search-and-prediction problems, from years to months; Insilico Medicine's rentosertib went from AI-identified target to positive Phase 2a results in roughly 30 months. But it is a fast highway bolted onto the front of a road that still ends in a slow mountain pass, and the pass is biology. Phase I, II, and III trials require observing real human bodies over months to years, and regulatory review and manufacturing scale-up cannot be shortcut by any neural network. The honest framing is a reduction from 12 to 15 years down to perhaps 8 to 12, meaningful and worth billions, but not the "decades to months" of the press release.



AI compresses drug discovery asymmetrically. The search-and-predict front end shrinks from years to months (green), but Phase I-III trials and regulatory review are biology-bound and cannot be shortcut. Bar lengths are drawn to one scale and anchored on the end-to-end totals: 12–15 years down to perhaps 8–12, not "decades to months." The split between phases is schematic.

Guide RNA Design and Off-Target Safety

CRISPR-Cas9 cuts DNA wherever a 20-nucleotide guide RNA directs it, which turns guide design into an ML problem: predict which sequences cut efficiently at the intended target and which cut at unintended sites. Think of a guide RNA as a search-and-replace instruction sent to every copy of a very long book; an off-target match is the edit landing on the wrong word in a different chapter. CCLMoff (2026) reached an AUROC of **0.996** for off-target prediction by pretraining an RNA language model before specializing on guide-DNA interactions. **Watch out.** An off-target cut in a tumor-suppressor gene could cause cancer, so this is a domain where model errors have direct physical consequences: explainability (which mismatch positions and chromatin features drive the prediction) is mandatory, and experimental validation of every clinical guide is non-negotiable.

Pharmacogenomics: Right Drug, Right Patient, Right Dose

The same drug at the same dose can heal one patient and endanger another, and the difference is often writ-

ten in a single metabolic gene. The CYP450 enzyme family metabolizes roughly 75 percent of prescription drugs, and genetic variants make a patient a poor metabolizer (drug accumulates to toxic levels), a normal metabolizer, or an ultra-rapid metabolizer (drug cleared before it works). The prescription is a recipe written for an average kitchen; pharmacogenomics reads the actual stove, and some run cold while some run dangerously hot. Codeine is the classic case: it is a prodrug converted to morphine by CYP2D6, and ultra-rapid CYP2D6 metabolizers convert it so fast that standard doses have killed children after tonsillectomy. Clopidogrel in CYP2C19 poor metabolizers carries a **3.6-fold** higher cardiovascular event risk because the drug never activates. **Watch out.** Over 90 percent of people carry at least one actionable variant, but fewer than **5 percent** of prescriptions use genetic testing. The barrier is not the science; it is workflow, reimbursement, and physician education. This is a user-experience problem wearing a data-science costume.

Gene	Drug	Phenotype	Consequence	Frequency
CYP2D6	Codeine	Ultra-rapid	Respiratory depression, death	~7–10% European
CYP2C19	Clopidogrel	Poor metab.	3.6× cardiovascular risk	15–20% East Asian
CYP2C9 / VKORC1	Warfarin	Variable	Bleeding vs clot	Differs by African ancestry
HLA-B*57:01	Abacavir	Carrier	Hypersensitivity	~5–8% European
DPYD	5-FU / capecitabine	Deficient	Severe toxicity	~3–5% European

Tumor Profiling and Liquid Biopsy

Precision oncology matches patients to treatment by what mutated, not by where the tumor sits. Tumor mutational burden (TMB) is the workhorse metric: it counts somatic mutations per megabase, and more mutations mean more neoantigens for checkpoint inhibitors to target. In symbols:

$$\text{TMB} = \frac{\# \text{ somatic mutations}}{\text{Mb sequenced}} \quad (\text{FDA cutoff } 10/\text{Mb})$$

Liquid biopsy reads circulating tumor DNA from a blood draw using three orthogonal signals, mutations, fragment length, and methylation, which is a needle-in-a-haystack task of finding a handful of tumor-derived molecules in a vast background of nor-

The Diversity Gap and Ancestry Portability

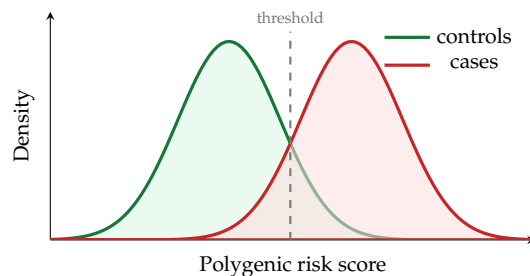
A polygenic risk score is only as representative as the people it was learned from. A PRS aggregates the weighted effects (β_j) of thousands of variants (g_{ij}) into a single risk number:

$$\text{PRS}_i = \sum_{j=1}^M \beta_j g_{ij}$$

The weights come from genome-wide association studies (GWAS), and as of 2024 roughly **79 percent**

mal DNA, exactly what statistical pattern recognition was built for. Methylation even answers not just “is there cancer?” but “where is it?”, because each tissue carries a distinctive chemical fingerprint, and multi-cancer early detection tests (Grail’s Galleri) screen for 50-plus cancers from one draw. But the “deceptively precise” 10/Mb cutoff shifts by a factor of two between labs depending on the panel and variant-calling pipeline, and TMB predicts response well in melanoma and lung cancer but poorly in breast and prostate. **Watch out.** At a 0.5 percent false-positive rate applied to millions of healthy people, the absolute number of false alarms is enormous, each triggering anxiety, imaging, and biopsy; positive predictive value collapses at low prevalence (Bayes’ theorem at population scale), which is why MCED tests target elevated-risk groups rather than universal screening.

of GWAS participants are of European ancestry, 10 percent Asian, 2 percent African, 1 percent Hispanic. A PRS is therefore a tailored suit: fitted on one body type, it hangs badly on every other, and no amount of confidence in the tailoring transfers the fit. Accuracy degrades as ancestry diverges from the European discovery population, dropping by up to **60 percent** for people of African descent, and warfarin dosing algorithms built in European cohorts performed worse than fixed-dose protocols for African American patients because variant frequencies and linkage patterns differ.



A polygenic risk score shifts the case distribution (red) right of controls (green), but the two overlap: a PRS stratifies risk, it does not diagnose. This separation, tuned on European cohorts, degrades by up to 60% when the model is applied across ancestral groups.

Ancestry group	% of GWAS (2024)	PRS accuracy vs European
European	79%	Baseline (1.0)
East Asian	10%	Moderate degradation
African	2%	Up to 60% accuracy loss
Hispanic / Latin American	1%	Substantial degradation
Indigenous, Oceanian, Middle Eastern	~0%	Effectively no signal

Train on European data and evaluate only on European data, and you have built a product for about 16 percent of the global population; deploy it as universal and you cause harm. The technical fix is straightforward, train on diverse data, evaluate across ancestral groups, and report stratified performance, but the structural fix, generating the diverse data in the first place, needs funding no single team can supply.

♥ THE PRECISION-ONCOLOGY REALITY (NCI-MATCH, 6,000+ PATIENTS)

Only 15 to 30 percent of profiled tumors have a matched targeted therapy; resistance emerges within 6 to 12 months; profiling costs 3,000 to 5,000 dollars and access is stratified by race, geography, and insurance; and the genomic, EHR, imaging, and registry data almost never integrate. Building a better matching algorithm is useful. Building the systems around it is essential.

Should You Build This?

Build it when the genomic result changes a decision a clinician can act on today and the workflow to de-

liver it at the moment of prescribing exists. Pharmacogenomics is the clearest case: the CYP450 variants are known, the adverse events they predict are documented, over 90 percent of people carry at least one actionable variant, and fewer than 5 percent of prescribing decisions use them. That gap is workflow, not science, and workflow is buildable. **Think twice when** the model's output is a hypothesis that still requires a wet lab, as perturbation predictions from single-cell foundation models and CRISPR off-target scores both do however high the reported AUROC; when the actionability rate is thin, as in the 15 to 30 percent of profiled tumors with a targetable mutation; and hardest of all before applying a polygenic risk score built on European-ancestry cohorts to a patient outside them. **Who bears the risk:** the patient outside the training population receives a risk estimate that is wrong in a direction no one reports, the ordering clinician owns the interpretation and the liability, and the payer, by deciding whether a several-thousand-dollar test is covered, decides who reaches the technology at all.

Quick Self-Test

1. Why does Geneformer rank genes instead of using raw expression counts, and what problem does that solve?
2. A single-cell model maps 40 percent of its features to known pathways. Why is that not enough to trust its drug-target predictions?
3. What does AI actually compress in the drug-discovery pipeline, and what stays biology-bound?
4. A patient is a CYP2D6 ultra-rapid metabolizer prescribed codeine. What is the danger, and why?
5. A polygenic risk score validated on European cohorts is applied to a patient of African descent. What happens to accuracy, and why?
6. An MCED blood test flags cancer in a healthy 40-year-old with no risk factors. Why should you be skeptical of the flag?

If You Remember One Thing

Precision medicine works best for people whose genomes are already in the training set. Across single-cell models, pharmacogenomics, and polygenic scores, the science is ahead of the infrastructure: the model is the easy part, and diverse data, stratified evaluation, and workflow integration are what decide whether it helps a patient or only a patient who looks like the discovery cohort.

Richard J. Young, Ph.D. · richard@deepneuro.ai · Senior AI Research Scientist, UnitedHealth Group · UNLV Lee Business School · ORCID 0000-0002-1109-7552