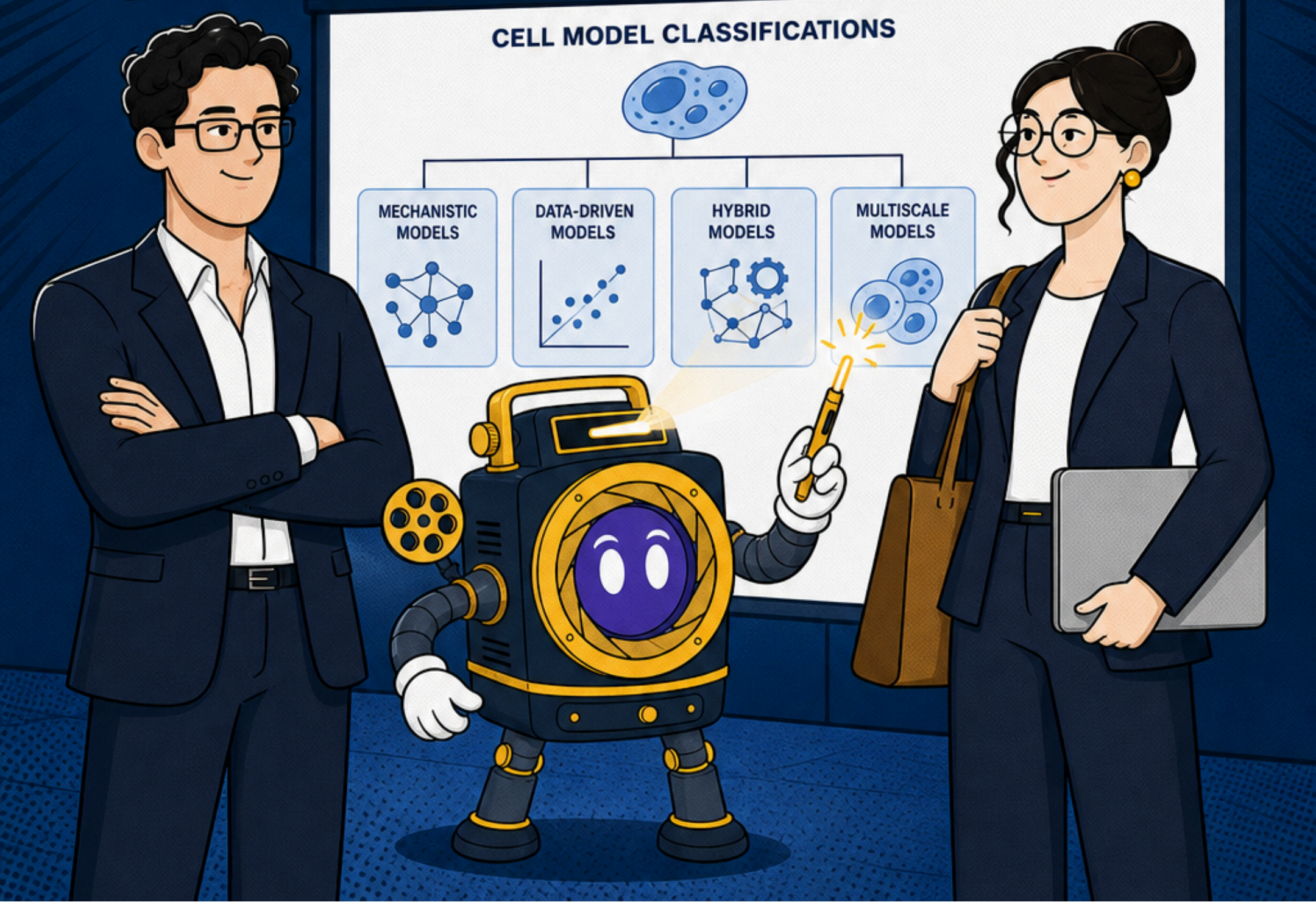
VIRTUAL CELL MODELS EXPLAINED:

WHAT BIOTECH INVESTORS NEED TO KNOW



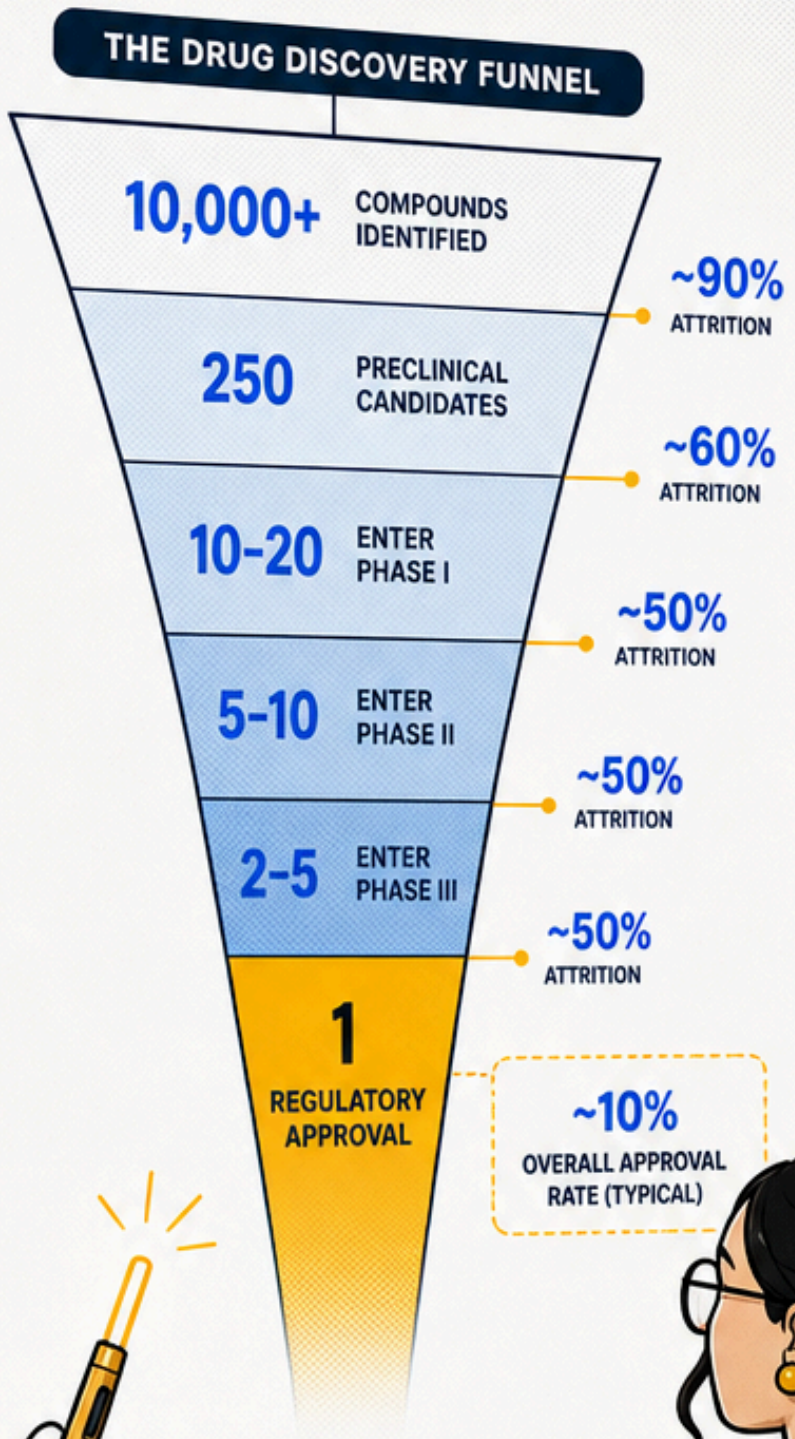
Why virtual cell models exist at all

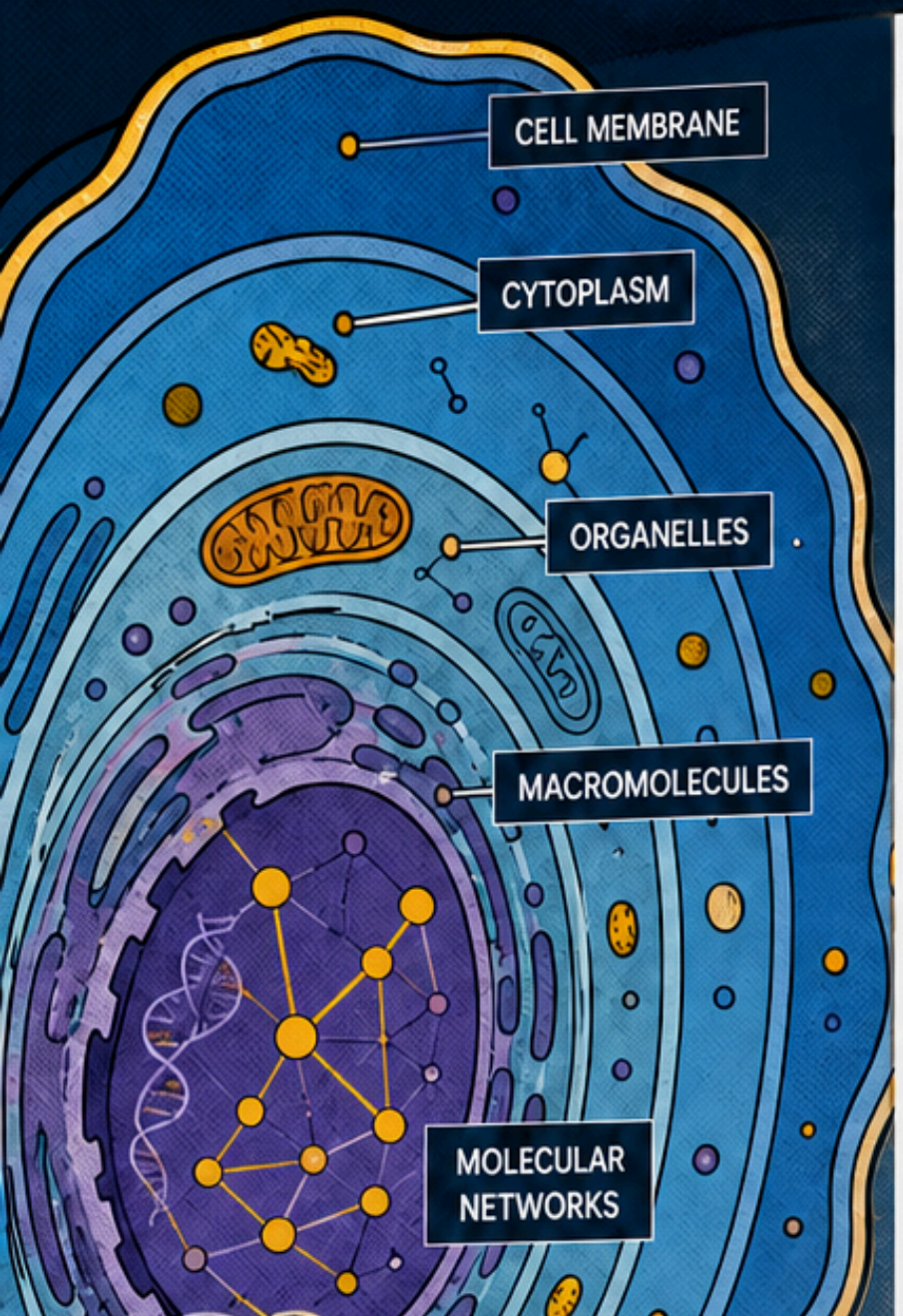
Drug discovery is fundamentally a process of **inferring** the effects of treatments on patients, and would benefit enormously from computational models that can reliably **simulate patient responses**, enabling researchers to generate and test large numbers of therapeutic hypotheses **safely and economically** before initiating costly clinical trials.

The practical problem is **cost and attrition.** Industry analyses consistently show that pharmaceutical R&D spend runs into the **hundreds of billions** of dollars globally each year, and that a large majority of drug candidates fail before reaching patients — with commonly cited estimates suggesting approximately **10%** of candidates entering Phase I trials ultimately gain regulatory approval, though figures vary depending on the starting point of measurement, therapeutic area, and methodology of the analysis. A large portion of that loss occurs before a single human trial begins. Virtual cell models are the attempt to **catch failures earlier, in silico**, before they become expensive wet-lab or clinical failures.

Creating such virtual cells has long been a goal of the computational research community. Recent advances in **AI, computing power, lab automation,** and **high-throughput cellular profiling** are now providing new opportunities for reaching that goal.

The challenge is that **"virtual cell"** covers a wide range of very different approaches, each with different assumptions, different failure modes, and different investment profiles.





The four main model families (part 1 of 4)

1. Whole-cell simulations

The most ambitious category. A whole-cell model attempts to represent every known molecular interaction inside a cell, from gene expression and protein folding through to metabolism and division, as a single integrated computational system.

The goal is to simulate what the cell actually does at a mechanistic level, not just predict one output under one condition. If it works, you can test how a genetic modification ripples across every biological process simultaneously, catching off-target effects that narrower models would miss entirely.

The catch: simulating an entire eukaryotic cell from first principles remains a distant goal. The computational and data requirements are immense. Most published whole-cell models apply to bacteria or minimal synthetic organisms, not human cells. For drug development in oncology, immunology, or rare disease, a true whole-cell model of the relevant human cell type doesn't yet exist at clinical-grade fidelity.

Even partial models (that focus solely on metabolism, while ignoring gene regulation, signalling, and cell-wall biophysics, **have inherent limitations:** they model only the enzymes involved in metabolism, overlooking genes with functions outside metabolic reactions, and fail to capture interactions between different biological processes.

Whole-cell simulations carry the highest theoretical fidelity and the longest, most expensive build cycles. For investors, they represent a bet on whether the team can actually close the gap between the current state of the science and what their model claims to do.

CONSTRAINT-BASED METABOLIC MODELS: THE CORE IDEA

Constraint-based models (CBMs) are among the most extensively published and applied mechanistic model classes in industrial biotechnology. Rather than simulating every cellular process, they focus on metabolic pathways: the chemical reactions that convert inputs into energy, growth, and useful outputs.

Genome-scale metabolic models (GEMs) have become central instruments for mechanistic reasoning in systems biology, enabling *in silico* exploration of cellular phenotypes under genetic and environmental perturbations through steady-state formulations such as flux balance analysis (FBA) and its many derivatives. 22-3

Their appeal lies in a disciplined translation of biochemistry into linear constraints — mass balance, reaction reversibility, and capacity limits — that define a feasible flux space whose optima can be interrogated for growth and other cellular phenotypes. 22-4

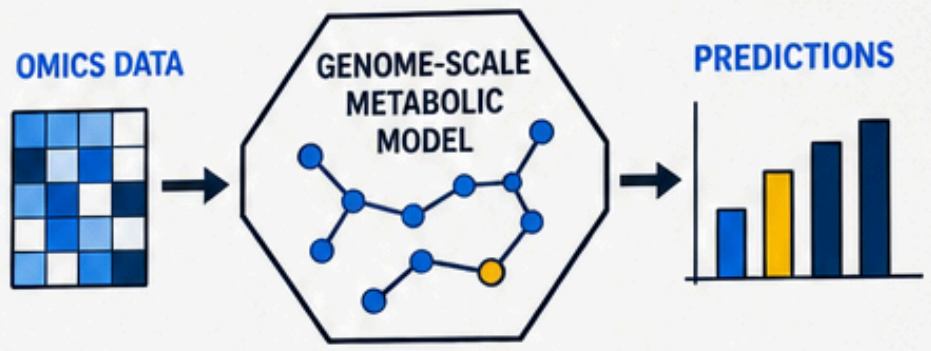
In plain terms: you give the model a starting condition (available nutrients, genetic constraints) and ask what metabolic flows are mathematically possible. It doesn't simulate dynamics in real time. It asks, given these rules, what steady-state is the cell likely to reach?

Genome-scale metabolic modeling is a growing area of computational biology with rich biotechnology applications, including the study of human metabolism for drug development and the design of synthetic microbial communities for health, environmental, and engineering purposes. 17-1

Constraint-based metabolic models:
reliability, limits, and what to ask

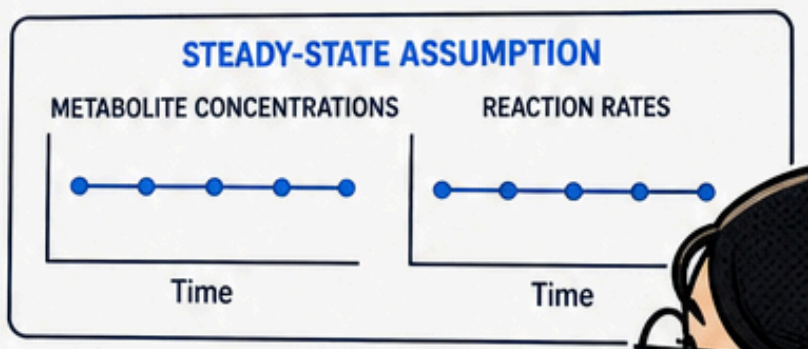
Incorporating omics data into genome-scale metabolic models is a key avenue for improved predictive accuracy.

[17–2]



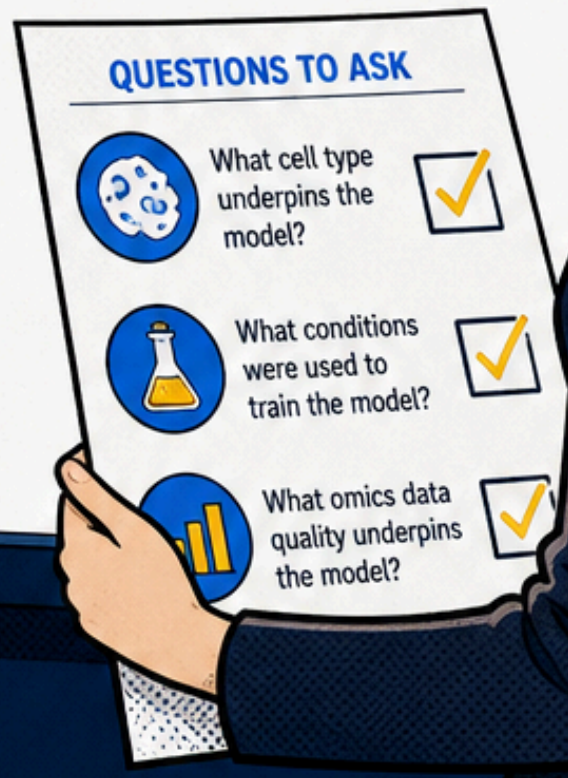
The limitation is meaningful: these methods assume that the system is at **metabolic steady-state**, such that the concentrations of all metabolic intermediates and reaction rates are constant.

[20–10]



Real biology doesn't hold still. Cells respond dynamically to stress, to drugs, to their microenvironment. A steady-state model **won't capture those dynamics reliably.**

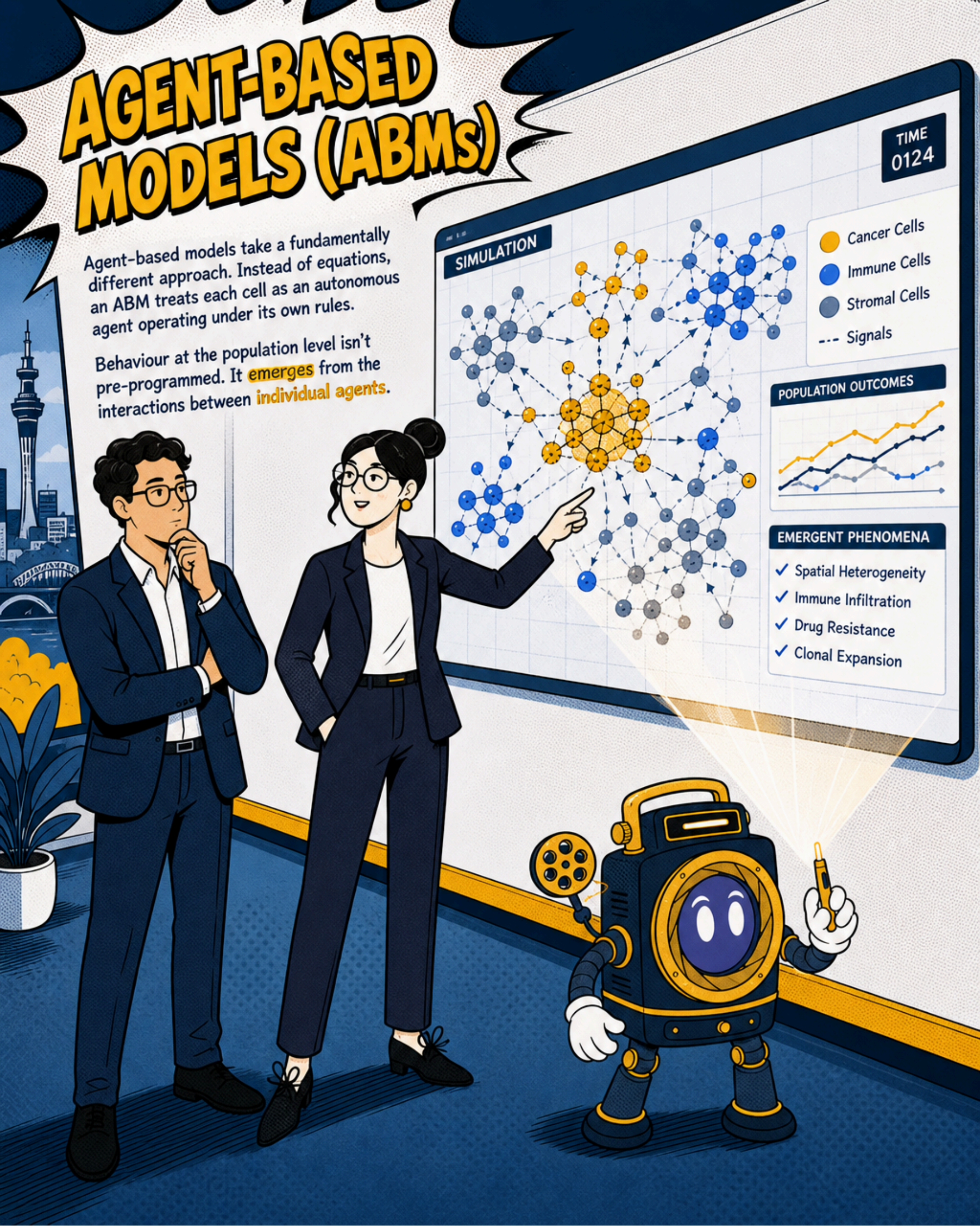
For investors, constraint-based models are a **well-established workhorse** in metabolic engineering. They are faster and cheaper to build than whole-cell simulations, and they have a documented track record in metabolic engineering applications.



AGENT-BASED MODELS (ABMs)

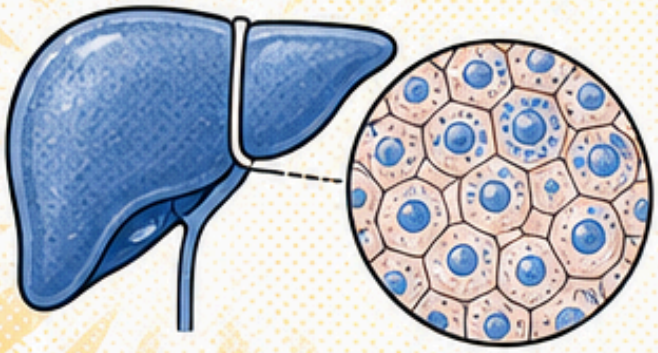
Agent-based models take a fundamentally different approach. Instead of equations, an ABM treats each cell as an autonomous agent operating under its own rules.

Behaviour at the population level isn't pre-programmed. It **emerges** from the interactions between **individual agents**.



MULTISCALE MODELS

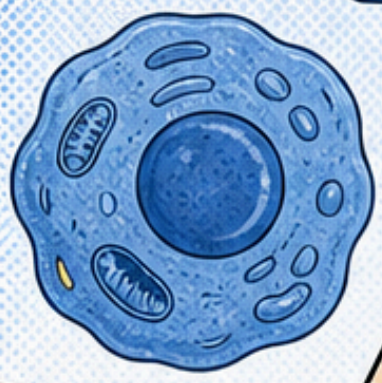
TISSUE / ORGAN LEVEL



OUTCOMES
Function, pathology, treatment response

EMERGENT BEHAVIOUR
Feedback loops, population dynamics

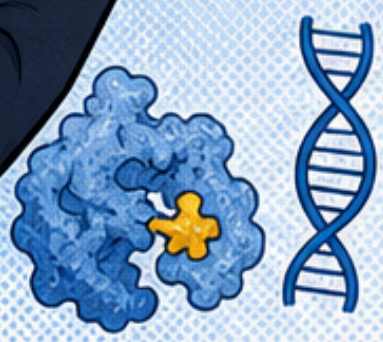
CELLULAR LEVEL



BEHAVIOUR
Metabolism, signalling, division

STATE CHANGES
Pathways, responses, cell fates

MOLECULAR LEVEL



INTERACTIONS
Protein binding, molecular complexes

EXPRESSION
Gene regulation, transcription

THEY CONNECT, NOT REPLACE
A coherent framework across scales

LEVEL MATTERS
Drug targets can behave very differently depending on the biological level

INTEGRATION DRIVES INSIGHT
Multi-scale integration captures emergent events; predictive models anticipate treatment outcomes

ERRORS PROPAGATE
Small errors at the molecular layer can compound into misleading predictions at higher scales

HIGHEST COMPLEXITY, HIGHEST BAR
Credibility comes from validated predictions at each scale before claiming integrated accuracy

THE OMICS DATA PROBLEM

DATA VOLUME ISN'T ENOUGH

More omics data ≠ better models.
Data must be consistent, well-characterised, and contextually appropriate for the cell type and condition.

CONTEXT MATTERS

Mismatched cell types, conditions, or patient populations lead to unreliable outputs—no matter how advanced the model.

BLACK-BOX RISK

Models that fit training data but can't explain predictions mechanistically are harder to validate—and harder to defend.

PHYSICS-INFORMED IS BETTER

Combining data with known biological laws and constraints improves plausibility, interpretability, and trust.

DUE DILIGENCE BOTTOM LINE

Ask about data origin, collection methods, relevance to the disease context, and model performance on held-out data.

MODEL TYPE SHAPES TIMELINE AND VALIDATION COST

DEVELOPMENT TIMELINE

CONSTRAINT-BASED METABOLIC MODELS
Well-characterised organisms

MONTHS

MULTISCALE MODELS
Depends on scales integrated and data availability

MONTHS – YEARS

WHOLE-CELL SIMULATION
Human disease-relevant cell type

YEARS

Investor question:

What does "the model is built" actually mean, and what experimental milestones must follow before it generates commercially relevant predictions?

VALIDATION COSTS

Retrospective and prospective validation are costly, requiring high upfront investments and initial adoption by multiple champions.

Studies are often hard to justify, particularly when they do not replace existing experiments.

This is especially true for ABMs and multiscale models, where emergent predictions are inherently harder to validate against single-experiment benchmarks.

While validation is a significant cost, it is also the primary driver of model credibility, commercial value, and regulatory acceptance — making it a capital allocation decision rather than purely a burden.

Iterative validation, entailing the systematic comparison of computational predictions with experimental data, is essential to enhance model accuracy and biological relevance.

That process takes time, laboratory resource, and money.

REGULATORY COMPLEXITY SHAPES INVESTMENT ASSUMPTIONS

REGULATORY PATHWAY OPTIONS

PRECLINICAL RESEARCH

IND SUBMISSION

REGULATORY REVIEW

WELL-VALIDATED MODEL (AUXILIARY INPUT)	WHOLE-CELL AI SIMULATION (STANDALONE EVIDENCE)
STANDARD PRECLINICAL EVIDENCE PACKAGE	NOVEL EVIDENCE ASSESSMENT
TRADITIONAL REVIEW PATHWAY	ENHANCED REVIEW & SCRUTINY
CLEANER PATHWAY SHORTER TIMELINE LOWER CAPITAL ASSUMPTION	**LONGER PATHWAY** EXTENDED TIMELINE HIGHER CAPITAL ASSUMPTION

12 Cite: 12–6, 12–7, 12–8

A SIMPLE FRAMEWORK FOR READING THE PITCH

1 **What type of model is it, specifically?**
Whole-cell, constraint-based, agent-based, or multiscale? Each carries different maturity and validation requirements.

2 **What omics data is the model trained or constrained on?**
Where was it collected, how was it processed, and how well does it match the target disease and cell type?

3 **What has the model predicted that subsequently matched experimental results?**
Validation against held-out data, not just training data, is the only meaningful signal.

4 **How does the model fit into the regulatory strategy?**
Is it a discovery accelerator, a biomarker qualifier, or is the team claiming it replaces a category of traditional preclinical evidence?

★ **The answers won't always be satisfying. That's fine. The absence of a clear answer to question three is usually the most important thing you learn.**

infrairis

https://startups.infrairis.com