

The City College of New York

CSC 59867-E: Senior Project II

AI Agents for Decision Making in the Real World

Biomaterials, Cellular Biology Simulations, and Disease Modeling

By Saptarashmi Bandyopadhyay
Email: sbandyopadhyay@ccny.cuny.edu
Assistant Professor of Computer Science
City College of New York and Graduate Center at the City University of New York

Saptarashmi Bandyopadhyay

September 30, 2026 CSC 59867 - E

Today's Agenda

Part 01	Macromolecular Design & Automated Binders (BindCraft)
Part 02	Cellular Perturbations & Gene Interaction Dynamics (GEARS)
Part 03	Multi-Scale World Models (AI-Designed Digital Organisms)
Part 04	Closed-Loop Disease Modeling

opadhyay

BindCraft

Saptarashmi Bandyopadhyay

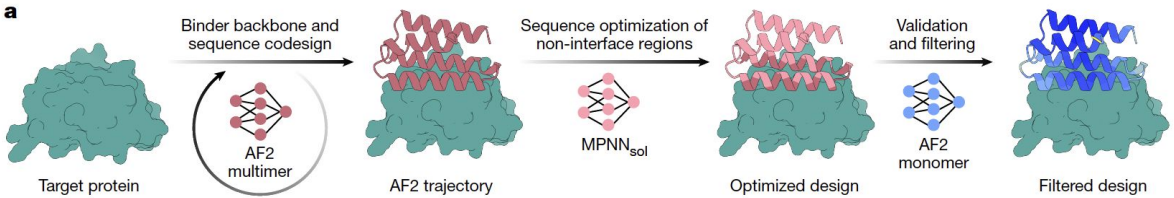
Beyond Smallecules (Small Molecules)

Combinatorial state explosion: Moving from 2D graphs to continuous 3D coord. space $SE(3)^N$ ($> 10^4$ Da).

Rigid-scaffold failure: Traditional physics-based docking assumes rigid targets, yielding < 0.1% experimental hit rates.

The inverse problem: Forward models predict structure from sequence ($s \rightarrow x$); binder design requires generating sequences that fold and bind a designated target pocket.

Induced-fit necessity: High-affinity binding requires modeling backbone and side-chain flexibility rather than static crystal coordinates.



Pacesa et al., 2025 (BindCraft), Fig. 1a.

Lopamudra Bandyopadhyay

Loss-Guided Backbone Hallucination w/ BindCraft

- **Backpropagating Through Structure Predictors:** BindCraft uses ColabDesign to backpropagate error gradients through AlphaFold2 weights from a random sequence, iteratively shaping the binder backbone.
- **Composite Structural Loss Landscape:** A multi-objective loss balances confidence, internal compaction, and interface contact density through a linear combination:

$$\mathcal{L}_{\text{design}} = w_1\mathcal{L}_{\text{pLDDT}} + w_2\mathcal{L}_{\text{i_pTM}} + w_3\mathcal{L}_{\text{PAE}}^{\text{binder}} + w_4\mathcal{L}_{\text{PAE}}^{\text{interface}} + w_5\mathcal{L}_{\text{contact}} + w_6\mathcal{L}_{\text{gyration}} + w_7\mathcal{L}_{\text{helicity}}$$

- **Preventing Structural Pathologies:** A "negative helicity loss" penalizes periodic i → i+3 backbone hydrogen bonds to avoid trivial all-α-helical bundles.
- **Staged Annealing Protocol:** Optimization moves through four stages: continuous logit relaxation, temperature-decayed softmax probabilities, straight-through estimation, and discrete greedy single-site mutation.

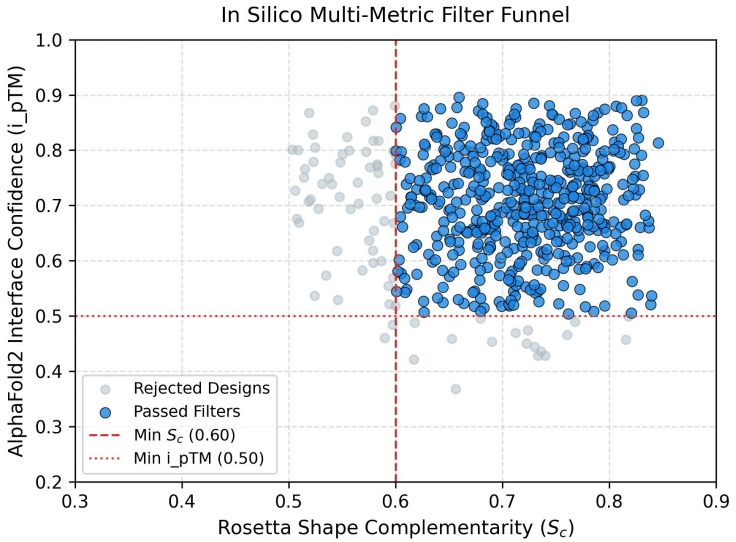
Sootanoshini Banerjopadhyay

Sequence Co-Design and *In Silico* Triage

Expression and solubility gating: Redesigns non-interface surface and core residues using soluble ProteinMPNN (MPNN_sol) while freezing the binding interface within 4 Å.

Decoupled monomer validation: Re-predicts complexes using AlphaFold2 monomer weights to remove multimer prediction bias.

Multi-metric filter funnel: Enforces simultaneous physical cutoffs: complex pLDDT > 0.8, i_PAE < 0.35, shape complementarity S_c > 0.60, and buried unsatisfied hydrogen bonds < 4.

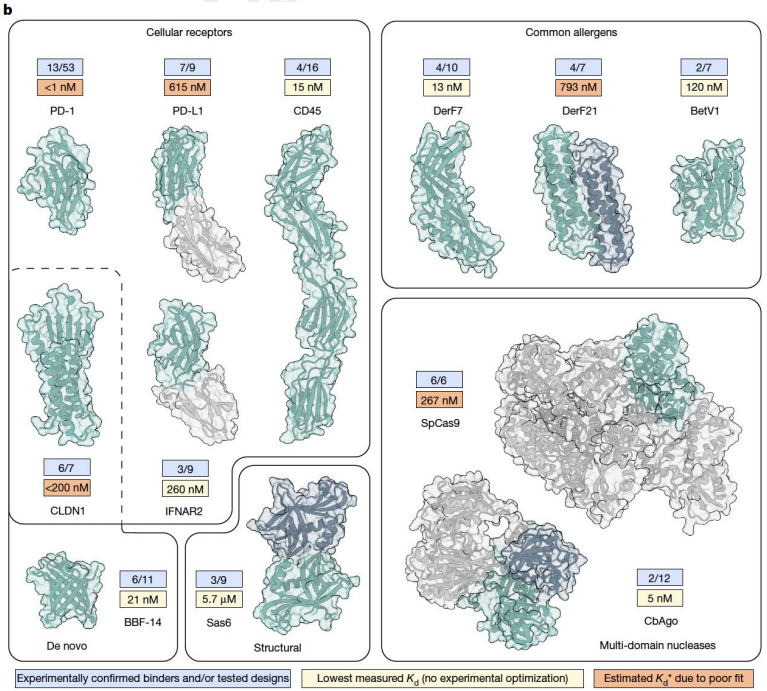


Plotting synthetic (simulated) distributions of i_pTM vs. Shape Complementarity (S_c), illustrating the Pareto filtration boundary separating viable binders from non-functional artifacts.

Santarashmi Bandyopadhyay

BindCraft Experimental Validation

- **High one-shot hit rates**: Achieved 10%–100% experimental success across 12 diverse targets without wet-lab optimization rounds.
- **Undruggable interface inhibition**: Engineered nanomolar blockers against large nucleic acid channels in SpCas9 (REC1 domain) and CbAgo nucleases.
- **De novo target generalizability**: Generated sub-nanomolar binders against the synthetic β-barrel BBF-14 (K_d = 20.9 nM), demonstrating zero reliance on natural co-evolutionary templates.
- **Crystallographic validation**: Solved X-ray structures (e.g., Der f7-binder2 at 2.2 Å resolution) verified atomic backbone alignment within RMSD_Cα ≤ 1.7 Å.

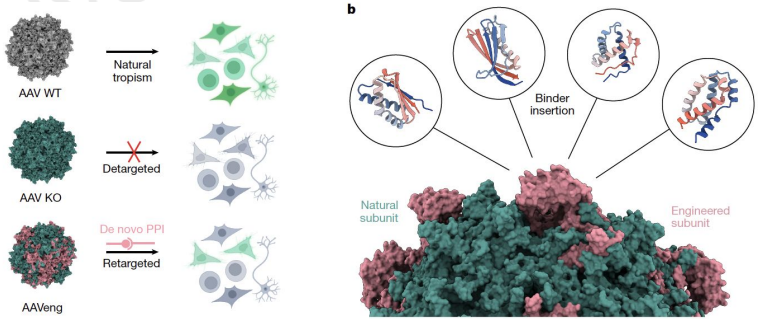


Precision Targeting of Binding Sites

Eliminating viral broad tropism: Reprograms Adeno-Associated Viruses (AAVs) to prevent off-target delivery and systemic liver toxicity.

Terminal distance constraint: Penalizes N-to-C terminal distance (< 5 Å) to ensure seamless miniprotein insertion into the capsid VR-V loop.

High-affinity cell-specific homing: Engineered capsids bearing HER2 and PD-L1 binders achieved up to 100-fold selective transduction into receptor-overexpressing target cells.



Pacesa et al., 2025 (*BindCraft*), Fig. 5a, b.

Schematic of AAV capsid retargeting via binder insertion into the VR-V loop near the threefold symmetry axis, showing wild-type detargeting and selective on-target transduction.

Saptarashmi Bandyopadhyay

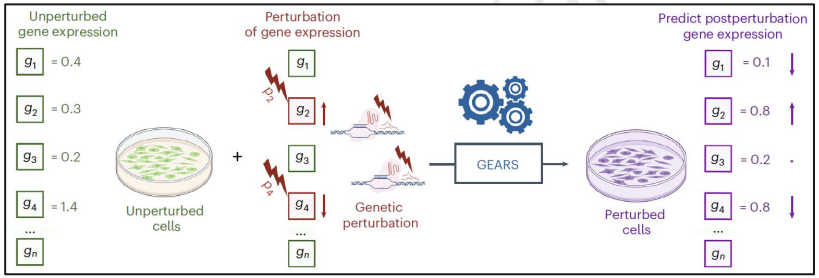
Gene activation and repression simulator (GEARS)

Saptarashmi Bandyopadhyay

High-Dimensional Cell State Transitions

High-dimensional cell state: Formulates cellular phenotype as an expression vector $x \in \mathbb{R}^K$ across $K \approx 20{,}000$ human genes.

Extreme measurement sparsity: Overcomes > 90% zero-count technical dropouts inherent to single-cell RNA-seq (scRNA-seq).



Perturbations as state transitions: Genetic interventions act as operators mapping baseline controls to perturbed profiles ($x_{ctrl} \rightarrow x_{pert}$).

Combinatorial assay bottleneck: Exhaustive wet-lab testing of all pairwise combinations across 20,000 genes requires $\approx 2 \times 10^8$ assays, necessitating in silico simulators.

Roohani et al., 2024 (*GEARS*), Fig. 1a.

Conceptual diagram illustrating unperturbed cells with baseline expression vectors undergoing genetic perturbation and transitioning into perturbed cell states.

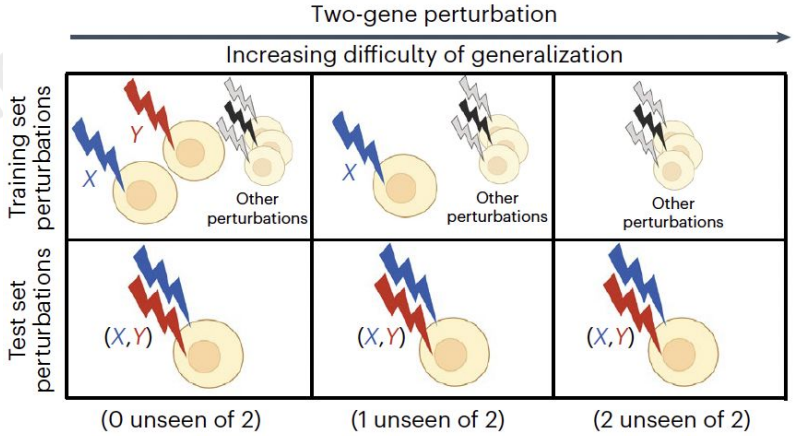
Aparasmni Bandyopadhyay

Multi-Gene Perturbations & Non-Additivity

Generalization tiers: Categorizes test difficulty into 0/2 unseen (both perturbed individually in training), 1/2 unseen, and 2/2 unseen (neither perturbed in training).

Failure of linear superposition: Naive additive models ($\Delta x_AB = \Delta x_A + \Delta x_B$) fail to capture non-linear genetic interactions like epistasis or synthetic lethality.

Knowledge graph grounding: Leverages Gene Ontology (GO) pathway networks to infer responses for previously unperturbed genes.



Roohani et al., 2024 (*GEARS*), Fig. 2e.

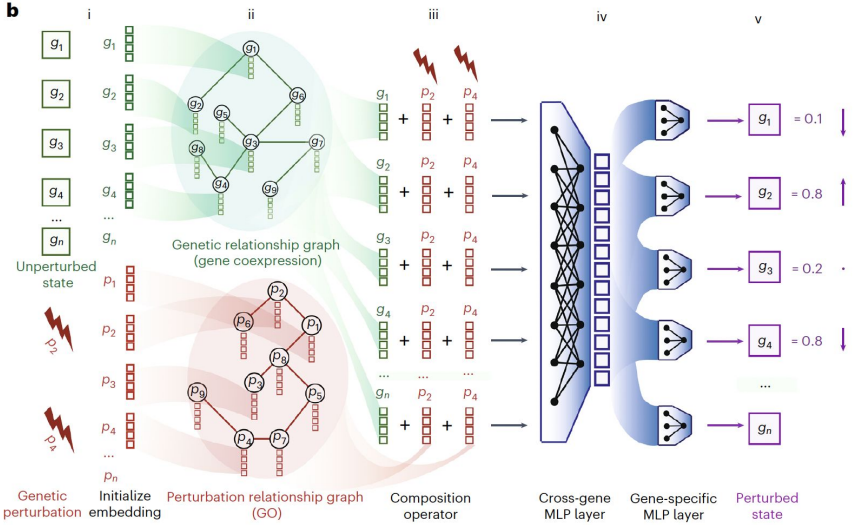
Diagram of the three generalization classes for two-gene combinatorial perturbations (0/2, 1/2, and 2/2 unseen) showing increasing degrees of out-of-distribution difficulty.

Graph-Enhanced Perturbation Learning

Dual biological GNNs: Integrates basal coexpression priors (G_gene) and functional pathway similarity graphs (G_pert) via graph message passing.

Compositional sum operator: Combines multi-gene perturbation vectors into a shared cross-gene latent state via a non-linear MLP.

Direction-aware autofocus loss: Penalizes differentially expressed gene errors exponentially while enforcing correct positive or negative regulatory direction.



Roohani et al., 2024 (GEARS), Fig. 1b.

GEARS network wiring diagram: from initial embedding vectors through GNN message passing on coexpression and GO graphs, compositional summation, cross-gene layer, to gene-specific output heads.

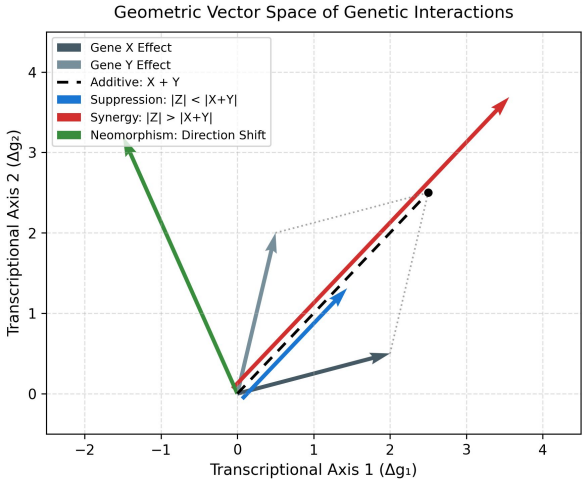
Saptarshmi Bandhyay

The Geometric Landscape of Genetic Interactions

Five non-additive classes: Quantifies synergy (|z| > |x+y|), suppression (|z| < |x+y|), neomorphism, redundancy, and epistasis.

High-precision candidate retrieval: Improves top-10 interaction detection precision by > 40% over existing latent-space models (CPA).

Empirical biological concordance: Accurately predicts non-additive expression shifts in PTPN12 + ZBTB25 knockouts, matching wet-lab Perturb-seq profiles.



Illustrative example of Additive, Synergistic, Suppressive, and Neomorphic genetic interactions in expression space (not from real data).

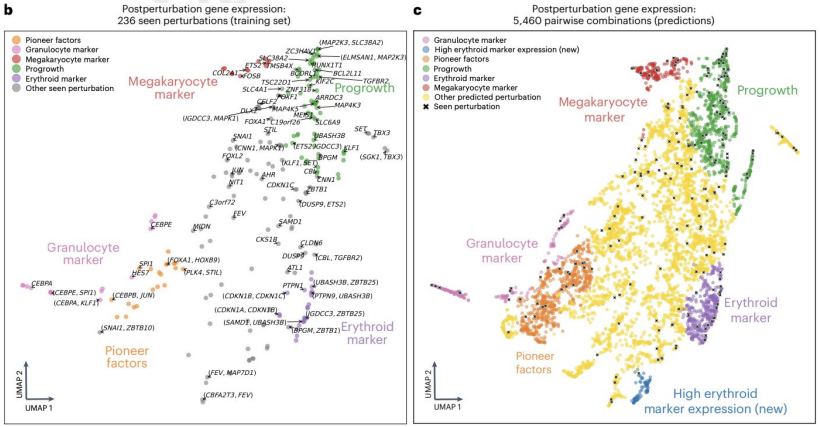
Sahtarashmi Band Madhyay

In Silico Phenotype Discovery

Combinatorial manifold mapping: Screened all 5,151 pairwise combinations of 102 single genes in human cells entirely in silico.

Emergent phenotype identification: Predicted unobserved cellular clusters (e.g., distinct high-erythroid marker expression) confirmed by fitness screens.

Uncertainty-guided acquisition: Uses Bayesian log-variance heads (σ_u^2) to flag unconfident predictions where gene nodes lack graph connectivity.



Roohani et al., 2024 (GEARS), Fig. 4b, c.

UMAP embeddings comparing the 236 experimentally seen perturbations in training against the 5,460 GEARS-predicted pairwise combinations, highlighting the discovery of the new "High erythroid marker" cluster.

Soptarshi Bandyopadhyay

AI-Driven Digital Organism (AIDO)

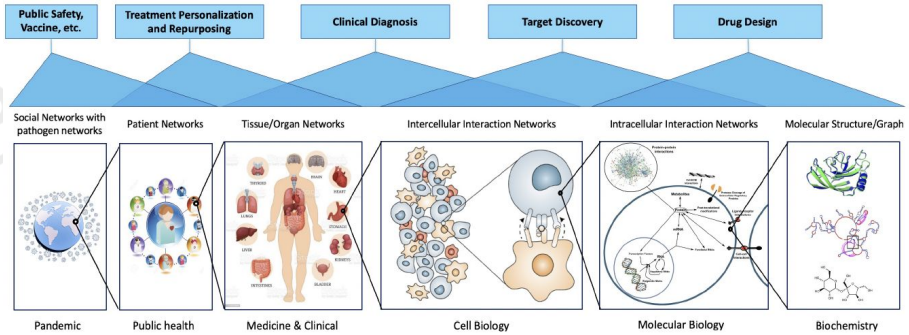
Saptarashmi Bandyopadhyay

Moving Beyond Single-Purpose Bio-Simulation

Beyond single-modality models: Unifies isolated point models (DNA, protein language models, single-cell transformers) into an interconnected system.

Multiscale biological hierarchy: Links sequence and 3D molecular structure to cellular networks, tissue microenvironments, and clinical phenotypes.

Three-stage engineering roadmap: Executes Divide-and-Conquer (modality FMs) → Connect the Dots (graphs and shared tokens) → Piece It All Together (cross-scale feedback).



Song et al., 2024 (AIDO), Fig. 1.

Multi-scale biological hierarchy diagram spanning from Biochemistry (molecular structure) → Molecular Biology → Cell Biology → Medicine & Clinical → Public Health / Populations.

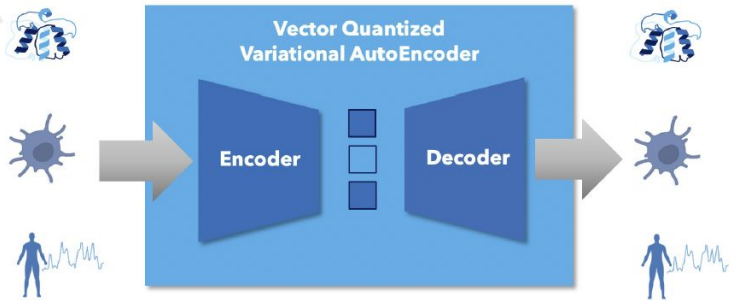
Soptarshi Bhattacharyopadhyay

Biological Tokenization & VQ-VAEs

Structure-aware sequence tokenization: Uses codon-aligned triplets (3 nucleotides → 1 amino acid) for coding regions and BPE for non-coding sequences.

3D geometric quantization: Encodes continuous protein coordinates into discrete structural tokens using Vector Quantized Variational Autoencoders (VQ-VAEs).

High-dimensional phenotypic codebooks: Maps sparse 20,000-D transcriptomes and longitudinal continuous glucose monitoring (CGM) time-series into shared operational latent vectors.



Song et al., 2024 (*AIDO*), Fig. 4.

Q-VAE architectural flow: continuous input (protein backbone coordinates, single-cell profiles, time-series phenotypes) passing into an Encoder → Vector Quantized Codebook → Decoder.

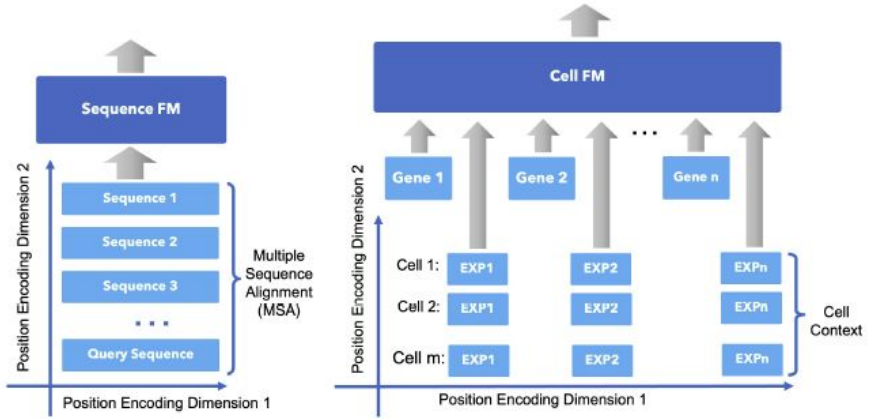
Soutarashmi Bandyopadhyay

Connecting 2D Positional Encodings & Graphs

2D Rotary Position Encodings (RoPE): Simultaneously encodes residue sequence index and evolutionary depth across Multiple Sequence Alignments (MSAs).

Spatial multicellular modeling: Adapts 2D RoPE to encode intrinsic gene expression alongside physical 2D/3D neighbor cell coordinates.

Differentiable Computation Graphs (DCG): Connects foundation models across scales via Message Passing Neural Networks that preserve signed regulatory logic (+/-).



Song et al., 2024 (*AIDO*), Fig. 7.

Description: Two-dimensional positional encoding architecture showing 2D RoPE applied across Multiple Sequence Alignments (left) and multi-cell spatial microenvironment transcriptomics (right).

Holistic System Alignment & Cross-Scale Feedback

Hierarchical embedding propagation: Latent representations flow upward from sequence models to cell-state transformers and clinical phenotype heads.

Top-down gradient backpropagation: Clinical supervisory signals propagate gradients backward through cellular and molecular layers.

Multi-task joint loss balancing: Integrates weighted task losses across molecular, cellular, and clinical tiers to prevent catastrophic forgetting.

$$\mathcal{L}_{\text{AIDO}} = \sum_{m \in \text{Molecules}} \lambda_m \mathcal{L}^{(m)}_{\text{struct}} + \sum_{c \in \text{Cells}} \lambda_c \mathcal{L}^{(c)}_{\text{pert}} + \sum_{p \in \text{Phenotypes}} \lambda_p \mathcal{L}^{(p)}_{\text{clinical}}$$

Sapteshmi Bandyopadhyay

Closed-Loop Disease Modeling

Saptarashmi Bandyopadhyay

Disease Modeling and Adaptive Control

The failure of maximum tolerated dose: Continuous high-dose chemotherapy eradicates sensitive populations, enabling rapid competitive release of drug-resistant clones.

Evolutionary game dynamics: Formulates cancer progression as Lotka-Volterra competition between treatment-sensitive and resistant sub-clones.

Closed-loop adaptive therapy: Reinforcement learning policies titrate drug timing to maintain sensitive populations that competitively suppress resistant cell growth.

Saptarshi Bandyopadhyay

Bottlenecks and Guardrails for Sim2Real Biology

In vitro assays fail to replicate in vivo tissue perfusion, immune microenvironments, and metabolic clearance:

- **Knowledge graph disconnection**: GEARS predictive accuracy drops when target genes lack dense connectivity in Gene Ontology networks.
- **Dual-use biosecurity firewalls**: Automated binder and sequence generators must be gated by strict synthesis screening firewalls to prevent toxin misuse.

Saptarashmi Bandyopadhyay

Lecture Summary & Next Lecture Topic

- **Macromolecular agency (BindCraft)**: Direct AlphaFold2 backpropagation enables one-shot protein binder design without high-throughput screening.
- **Cellular simulation (GEARS)**: GNNs over biological knowledge graphs predict non-additive genetic interactions and emergent phenotypes.
- **Unified multiscale modeling (AIDO)**: VQ-VAEs and differentiable computation graphs connect sequence, structure, and cellular dynamics into a digital organism.

Next lecture preview: Transitioning to multi-agent economic systems: **Financial Trading Agents, Stock Portfolio Orchestration, and High-Frequency Market Dynamics**.

Saptarashmi Bandyopadhyay

Questions?

Saptarashmi Bandyopadhyay