Perspective

Towards the construction of a virtual yeast

https://doi.org/10.1038/s41586-026-10574-9

Received: 17 July 2025

Accepted: 21 April 2026

Published online: 1 July 2026

Check for updates

Liujia Qian[1,2,3], Zizhuo Zhou[1,2,3], Peijie Zhou[4,5,6,7], Zhen Dong[1,2,3], Xudong Zhang[4,7], Zhenwu Dai[1,2,3], Zhangyang Gao[8], Siqi Sun[9], Kevin R. Roy[10,11], Shuaiyao Wang[1,2,3], Nicola Zamboni[12], Charles Boone[13,14], Michael Costanzo[13,14], Jianzhe Li[4,7], Gianni Liti[15], Jia-Xing Yue[16], Markus Ralser[17,18,19,20], Evan Williams[21], Mattia Zampieri[22], Heng Jiang[1,2,3], Tailin Wu[23], Yalin Wang[24], Feiran Li[25], Joseph Schacherer[26], Rui Sun[1,2,3], Zhaoxing Li[1,2,3], Yaming Deng[1,2,3], Yi Chen[1,2,3], Zhiping Xie[27], Huiqiang Lou[28], Xiaowen Wang[29], Linhai Xie[29,30], Han Wen[5,31,32,33], Liangyi Chen[34], Kai Lei[35,36,37], George Rosenberger[38], Xue Cai[1,2,3], Yingrui Wang[1,2,3], Qi Xiao[1,2,3], Huaizong Shen[35,37,39], Gaowen Liu[40], Lei Ma[41], Brenda Andrews[13,14], Hui Lu[42,43,44], Kiryl Piatkevich[35,45,46], Yi Zhu[1,2,3], Lei Bai[8], Yizhi Cai[47,48], Yuping Chen[49,50], Weinan E[4,5,7,51], Ge Gao[52], Fuchu He[29,30], Luonan Chen[53], Stan Z. Li[54], Hongwu Ma[55,56], Liang Qiao[57], Lars M. Steinmetz[11,58,59,60], Leihan Tang[61], Tang Tang[62], Xiaofan Zhang[1,2,3], Jing Yang[30,63,64,65], Yifan Yang[61], Kaicheng Yu[66], Jianyang Zeng[67], Yefeng Zheng[68], Bowen Zhou[8,69] & Tiannan Guo[1,2,3] ✉

To advance the computational simulation of cellular life, we propose a virtual yeast, an artificial intelligence (AI)-driven agent that models eukaryotic cellular behaviours by integrating multimodal biological data, mechanistic reasoning and active experimentation using *Saccharomyces cerevisiae* as a genetically tractable and data-rich model system. Cellular complexity is decomposed into eight function-centred modules, spanning genetic, metabolic and structural systems, each realized as a domain-specific AI tool coordinated through a large language model-based orchestration layer. Built on three data pillars, namely, mechanistic knowledge, subcellular architecture and dynamic states, the system integrates representation learning and generative modelling within a closed-loop learning pipeline that autonomously designs and executes experiments. The virtual yeast serves as both a conceptual and an operational platform to optimize biosynthetic pathways, support the generation and prioritization of hypotheses across diverse cellular processes, and accelerate target discovery. By coupling biological realism with autonomous AI reasoning, the virtual yeast establishes a generalizable blueprint for constructing virtual eukaryotic cells and advancing synthetic biology.

Recent advances in AI are reshaping the life sciences. Landmark achievements such as AlphaFold, now extended to biomolecular interaction mapping[1], together with models such as Geneformer[2], Evo[3] and AlphaGenome[4], illustrate how learning from large-scale biological data enables extrapolation into uncharted spaces. Yet, individual molecules studied in isolation cannot capture the complexity of life, because the cell is its fundamental unit. Constructing predictive virtual cells is therefore a critical frontier for simulating disease, revealing mechanisms and accelerating therapeutic discovery and bioengineering[5]. Whole-cell modelling began with minimal organisms such as *Mycoplasma genitalium*[6] and later extended to *Escherichia coli*[7], using hybrid frameworks that combine constraint-based metabolism with mechanistic, stochastic and rule-based description of cellular processes. In yeast, genome-scale metabolic models (GEMs) provided stoichiometric reconstructions of biochemical networks[8], initially enabling predictions of gene essentiality, nutrient utilization and growth phenotypes. Contemporary GEMs integrate multi-omics data with enzymatic and thermodynamic constraints, to predict metabolic fluxes and growth phenotypes[9–11]. However, classical rule-based approaches struggle to integrate fragmented datasets and often generalize poorly across diverse cellular states.

Predictive AI frameworks and machine learning offer a transformative path forward and already constitute early forms of AI-driven virtual cells by learning cellular state transitions directly from large-scale data and forecasting responses to perturbations. Emerging AI systems, including large-scale transformers trained on single-cell atlases[12], neural ordinary differential equation models of perturbation proteomics[13], genome-to-proteome predictive maps[14] and language-model approaches for transcriptomic interpretation[15], demonstrate that data-driven models can infer cellular dynamics and forecast responses to genetic and chemical perturbations across contexts. Despite these advances, current models remain constrained by reliance on a few environmental conditions, single modalities, limited spatial resolution and insufficient incorporation of perturbational time-series data to capture transient dynamics. Addressing these limitations requires three data pillars: a priori biochemical and genetic knowledge, spatially resolved cell architecture and dynamic multi-omics perturbation datasets, resources that remain comparatively scarce[16].

Mass spectrometry-based proteomics and metabolomics now enable high-sensitivity, high-throughput measurements of spatiotemporal

A list of affiliations appears at the end of the paper.

Box 1

Why *S. cerevisiae* provides a strategic foundation for virtual cell development

S. cerevisiae offers a uniquely tractable eukaryotic platform for developing and validating AIVC architectures. This choice reflects a staged strategy.

(1) Compact yet fully eukaryotic architecture
- Encapsulates core eukaryotic compartmentalization (the nucleus, mitochondria, endoplasmic reticulum, Golgi, vacuole and stress-responsive organelles) in a 3–10-μm cell[87] (Fig. 1).
- Exhibits pronounced subcellular spatial heterogeneity, including polarized growth, asymmetric budding[106,107], organelle inheritance and dynamic membrane remodelling (Fig. 1).
- Provides accessible modelling of cross-organelle coordination and spatiotemporal cellular functions, which are foundational elements of any virtual cell.

Although lacking tissue organization, yeast forms structured colonies, flocculent aggregates and biofilms under defined conditions, allowing controlled analysis of cell–cell coordination.

(2) Systematic genetic toolkit
- Possesses extensive, community-developed resources including gene deletion libraries (YKO) covering essentially all non-essential genes[68,108] and their prototrophic strains[109], overexpression libraries (GAL-GST)[110,111], proteome-scale GFP-tagged localization collections[73,112,113], multiplexed precision genome-editing libraries[114,] and extensive genetic interaction maps generated by SGA and related platforms[115–117], phenomics (more than 14,500 genome-wide phenotypic screens)[118] and large-scale genotype–proteome-fitness mapping in controlled meiotic populations[14], revealing genetic interactions, networks and buffering pathways.
- Benefits from comprehensive genomic resources including the SGD[30,119,120], genome evolution across 1,011 *S. cerevisiae* isolates[39], well-characterized reference strains (for example, S288C/BY series)[121] and near-complete telomere-to-telomere assemblies[42,122,123].

Although mammalian systems now support powerful CRISPR-based perturbation platforms, few provide comparably dense genetic interaction maps or systematically standardized perturbation-to-phenotype landscapes within a single species. Yeast uniquely integrates scalable perturbation tools, from gene deletions to multiplexed precision editing and high-throughput variant effect assays, with comprehensive interaction networks, offering a coherent foundation for mechanistic modelling.

(3) Conserved eukaryotic cellular logic
- Overall sequence conservation between yeast and humans is limited[124–126], and yeast does not model multicellular organismal complexity. However, core cellular processes, including cell cycle regulation, DNA repair, vesicle trafficking and metabolic control, are largely conserved at pathway and functional module levels[127]. Yeast thus serves as a controlled eukaryotic baseline for validating architectural and perturbation-driven modelling principles.

(4) Versatile synthetic biology chassis
- Proven platform for engineering biofuels, pharmaceuticals and metabolites due to genetic tractability, rapid growth and well-characterized metabolism[128].
- Drives synthetic genomics (Sc2.0/SCRaMbLE, Sc3.0)[129,130] and pathway reconstruction[131,132].
- Enables synergistic integration of computational modelling (GEMs) with synthetic biology for predictive design[133].

A virtual yeast therefore has immediate value as a closed-loop optimization system.

(5) Extensive multi-omics integration
- Supported by curated multiscale databases such as genomics (SGD and ScRAPdb)[30,134], metabolomics (YMDB)[31], proteomics and interactomics (Yeast interactome[38], 5k Yeast Proteome[44] and Yeast PeptideAtlas[135]), knowledge-based functional network (YEASTNET)[32] and imaging (LoQAtE and CYCLoPs)[33,34].
- Subject of landmark integrative studies including near-saturation genetic and protein interaction maps[35,37,38]; genome, proteome or metabolome profiling across genetic perturbations[43–47]; pangenome analyses[39,41]; and metal ion homeostasis studies[49].

Among eukaryotes, yeast remains one of the most systematically perturbed and quantitatively profiled organisms across molecular scales. Establishing architectural principles within this controlled single-species framework enables clear attribution of predictive advances and provides a transferable blueprint for progressively more complex systems.

molecular dynamics[17,18], while AI is advancing towards multimodal foundation models pretrained on integrated omics datasets to represent cellular states[5,19]. Together, these developments enable experimentally grounded virtual cells capable of simulating molecular and cellular behaviours across physiological contexts. A pivotal question is which organism should serve as the inaugural platform. The budding yeast *S. cerevisiae*, with its rapid life cycle, compact yet fully eukaryotic cellular architecture, genetic tractability and extensive multi-omics resources, remains a premier system for eukaryotic biology (Box 1). Building on existing resources while systematically generating new spatiotemporal datasets, a yeast AI-driven virtual cell (AIVC) could serve as a mechanistically grounded and experimentally verifiable prototype for a family of transferable virtual cell systems. In this Perspective, we outline the data, modelling and experimental foundations required to construct a virtual yeast, and describe an integrated framework that unifies spatial architecture, dynamic perturbation responses and a priori biological knowledge. Together, these position virtual yeast as a blueprint for predictive, experimentally testable virtual cells across eukaryotes.

Virtual yeast as an AI agent

Our virtual yeast project aims to establish a predictive, multiscale model of *S. cerevisiae* with unprecedented spatial and temporal resolution. To render cellular complexity computationally tractable, we decompose the cell into eight interconnected functional modules, implemented as specialized AI tools or functional skills coordinated within a unified agent framework (Fig. 1). These comprise: (1) membrane systems, coordinating endomembrane architecture, trafficking and lipid synthesis across organelles; (2) genetic hubs, integrating nuclear organization, genome stability, transcriptional control and cell cycle progression

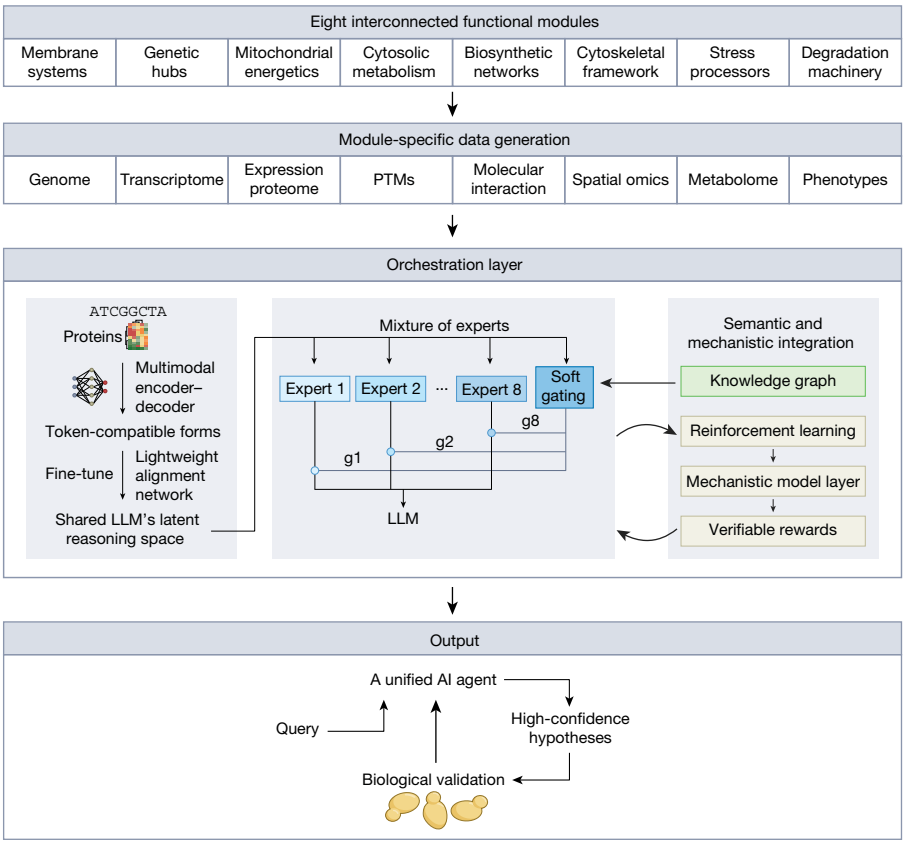

Fig. 1 | Conceptual roadmap of the virtual yeast AI agent. The virtual yeast integrates biological modularization, multimodal data representation and mechanistic reasoning into a unified AI agent framework that emulates cellular regulatory and information-processing dynamics. Eight functional modules, including membrane systems, genetic hubs, mitochondrial energetics, cytosolic metabolism, biosynthetic networks, cytoskeletal framework, stress processors and degradation machinery, are implemented as specialized AI tools trained on module-specific datasets. These tools are coordinated by an LLM-based planning and reasoning layer that aligns heterogeneous inputs through multimodal encoder–decoder bridges and lightweight alignment networks, embedding them into a unified multimodal context adapted to yeast biology. The agent dynamically decomposes complex biological queries into subtasks routed to relevant tools, enabling cooperative reasoning across biological functions and data modalities. In addition to architectural strategies such as sparse expert activation (for example, MoE) used within the underlying language model to support scalable computation, the system's routing primarily reflects task-dependent tool selection based on the biological query. Mechanistic grounding is achieved through knowledge graph constraints and external mechanistic verifiers that provide verifiable rewards during reinforcement learning, forming a feedback loop that refines predictions and preserves biological consistency. The resulting AI agent supports predictive and generative tasks, producing experimentally testable hypotheses that are validated and iteratively incorporated, establishing a closed loop between modelling and biological experimentation. PTM, post-translational modification.

within a regulatory context; (3) mitochondrial energetics, governing oxidative phosphorylation, redox balance and ageing-associated metabolic dynamics; (4) cytosolic metabolism, encompassing central carbon metabolism, amino acid and nucleotide biosynthesis, nutrient sensing and metabolic storage; (5) biosynthetic networks, managing protein synthesis, folding, post-translational modification and turnover through endoplasmic reticulum–Golgi and cytosolic quality-control pathways; (6) cytoskeletal framework, integrating actin, microtubules and cell wall remodelling to support morphogenesis, intracellular transport and spatial organization; (7) stress processors, including stress granules, processing bodies and oxidative stress detoxification systems that mediate adaptive reprogramming and RNA quality control; and (8) degradation machinery, comprising proteasomal and vacuolar pathways executing autophagy, proteolysis and organelle recycling.

These modules operate as execution units within an integrated virtual cell that mirrors eukaryotic organization: specialized compartments perform distinct processes while exchanging state information through constrained interfaces. This function-centred modularization complements recent virtual-cell frameworks that organize models across biological scales (molecular–cellular–multicellular) and integrate interconnected foundation models spanning molecular networks to higher-level cellular and tissue states[5,20]. Beyond representation learning, this architecture supports perception, reasoning and adaptive intervention simulation through iterative refinement. Aligning model boundaries with functional execution units enhances biological interpretability and computational tractability: many genetic, environmental and pharmacological perturbations originate within specific functions before propagating across scales. This architecture therefore enables attribution of perturbation effects, temporal ordering of responses and identification of compensatory cross-organelle coupling, which are challenges for monolithic models.

Each module is trained primarily on measurements most relevant to its biological role, while leveraging shared datasets when appropriate. This ensures that representations are grounded in functionally informative evidence rather than imposed architectural assumptions. Accordingly, each module operates as a data-driven, domain-specific AI tool built using representation learning and generative algorithms tailored to its context (Fig. 1). For example, membrane trafficking models learn from dynamic imaging and vesicle-tracking data, whereas nuclear function models rely on sequence-based and transcriptomic representations. Experimental design is modularized to minimize redundancy while preserving cross-module dependencies for integrative reasoning.

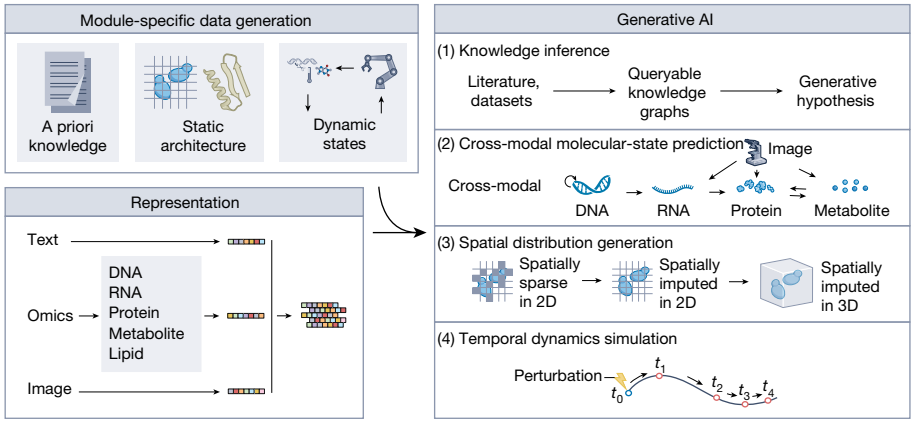

Fig. 2 | Architectural framework for each functional module of the virtual yeast agent. Functional module-specific data, including a priori knowledge, static architectural information and dynamic molecular states, are integrated into a unified multimodal embedding space through representation learning, which supports downstream generative modelling. Fragmented literature and curated databases are structured into queryable knowledge graphs that support mechanistic inference. Multimodal inputs (text, omics and image) can be used to support cross-omics prediction and molecular-state translation across scales. Sparse spatial omics and images are reconstructed into 3D molecular and organelle-scale architectures via geometry-aware modelling. Unpaired endpoint measurements inform constrained dynamical simulations, whereas conditional generative models extrapolate responses to unseen perturbations. Although the framework is comprehensive, practical virtual yeast models need not include all components. Instead, selected modules can be developed for specific applications, enabling experimentally tractable, function-oriented virtual cells rather than full digital twins. The framework therefore provides a flexible basis for building predictive and controllable modules.

To maintain physical plausibility, each module is further regularized by biophysical constraints, including stoichiometric balance, kinetic laws and topological connectivity, derived from mechanistic frameworks such as GEMs and force-balance models.

All modules integrate three information sources into a coordinated modelling framework: a priori biological knowledge, spatial architecture and temporal evolution[16] (Fig. 2). To operationalize this integration, a multi-layered alignment strategy is required, in which specialized encoders process diverse data types and align representations as structured context for the orchestration layer[21]. Sequence foundation models such as Evo[3] or ESM[22], fine-tuned on yeast-specific corpora (Box 2), embed DNA and protein sequences, whereas omics data matrices are encoded using variational autoencoders or self-supervised models that capture statistical dependencies and gene or pathway-level relationships. Relational priors and interaction networks are incorporated as relation-aware conditioning or graph-structured priors to preserve biological context. Generative models then reconstruct cellular organization and dynamics, infer state transitions, and predict unobserved configurations. Diffusion models[23] and flow-matching approaches[24] infer state transitions from discrete experimental snapshots. Together, these representation learning and generative layers enable modular intelligence that learns both cellular state and its evolution. Module-specific models (typically approximately 10^7–10^8 parameters each) are coordinated by a higher-level orchestration layer based on large language models (LLMs), forming an integrated AI agent that orchestrates tasks across specialized tools via tool calling (Fig. 1). Rather than operating as a monolithic model, the system dynamically invokes module-specific capabilities depending on the biological query, enabling adaptive reasoning across cellular functions and data modalities. In addition, architectural strategies such as mixture of experts (MoE)[25,26] or related sparse-computation schemes may be used within the underlying language model to improve scalability (Fig. 1). This design parallels multimodal scientific foundation models such as Intern-S1 (ref. 27) and may be fine-tuned from pretrained systems or trained de novo using adapter-based alignment.

Task routing is performed through an adaptive planning layer that selects relevant modules based on modality, task structure and intermediate representations[28] (Fig. 1). In practice, this may be implemented through explicit planning, semantic routing or tool-selection policies. Selective module invocation activates only a subset of capabilities and computational paths per query, enabling efficient specialization across omics, imaging, structural data and functional modules. Routing decisions are logged during inference, providing transparency regarding which tools and biological modalities were engaged and enabling empirical validation through activation statistics and modality-tool associations.

To ensure biological fidelity, planning and routing are further constrained by ontologies and knowledge graphs encoding causal relationships among genes, pathways and organelles (Fig. 1). The system is optimized through reinforcement learning from verifiable rewards, in which mechanistic models serve as external validators providing objective feedback[29]. This feedback allows the system to iteratively refine predictions against experimental constraints, transforming pretrained representations into an adaptive, hypothesis-generating agent. To further disentangle causal mechanisms from statistical correlations, we plan to incorporate a causality-oriented framework that formulates regulatory structure discovery as a continuous optimization problem. This approach may improve causal inference beyond directly perturbed genes and enable counterfactual prediction across molecular layers and functional modules.

Building functional intelligence

Bridging modular design and implementation, the computational infrastructure of the virtual yeast defines how each functional module is instantiated through specific data sources and algorithmic backbones. Each module comprises two tiers: non-generative embeddings encoding biological knowledge, and generative models that impute missing information, extrapolate beyond training distributions, infer transitions and synthesize unseen cellular states. Together, these layers translate heterogeneous biological data into a coherent computational organism capable of predictive reasoning across scales.

Non-generative representation backbone

The development of a virtual yeast requires systematic integration of decades of biological knowledge within modern AI frameworks[16]. Cellular information processing spans both the vertical logic of

central dogma flow and horizontal networks of molecular interactions. Although existing models capture many horizontal relationships, they struggle with vertical integration across scales, for example, linking transcriptional regulation to metabolic flux or metabolite binding to protein conformational change, limitations not resolved by single-scale tools such as AlphaFold[1].

This phased strategy anchors the AIVC architecture in pretrained models selected for their capacity to address distinct hierarchical gaps, from genomic-level to systems-level scales. Rather than merely cataloguing existing architectures, the framework adapts and fine-tunes general-purpose foundation models for *S. cerevisiae* (Box 2), creating an interconnected knowledge scaffold that supports multidimensional inference.

To bridge biological scales, hybrid neural networks incorporate dual coordination hubs: one capturing within-type relationships (for example, gene–gene interactions) and another linking across types (for example, gene–protein regulatory cascades). These hierarchies integrate nucleic acids, proteins and metabolic pathways within unified training frameworks that use instruction tags to encode modality, condition and task, enabling cross-scale prediction and perturbation modelling within a shared backbone[19]. This design enhances biological fidelity while balancing complexity and scalability in whole-cell modelling.

Harnessing biological knowledge

Yeast benefits from deeply curated multi-omics resources, including the *Saccharomyces* Genome Database (SGD) for genomic annotation[30], YMDB for metabolism[31], YeastNET for protein interaction networks[32] and imaging repositories such as LoQAtE and CYCLoPs[33,34]. Large-scale efforts have produced a global genetic interaction map comprising approximately 1 million interactions across 23 million double mutants[35,36], integrative structural insights from genetic interaction profiles[37] and comprehensive interactome maps[38]. Multi-omics studies span pangenome diversity, growth phenotypes and molecular traits across 1,011 isolates[39–42], knockout adaptation signatures[43] to proteomic response atlases[44], mitochondrial assemblies with approximately 90% coverage[45] and evolutionary metabolomic landscapes[46,47].

Multiple datasets reveal pervasive context dependence: aneuploidy tolerance varies across strains[41], 18.5% of knockout phenotypes differ among genetic backgrounds[48], and the role of metal ion biology remains a major gap in genotype–phenotype prediction[49]. Such pleiotropy and environmental modulation necessitate genotype-aware integration frameworks that account for background effects on variant penetrance and expressivity. The breadth of available knowledge, from subcellular architecture to population variation, demands generative approaches that are capable of synthesizing fragmented, context-contingent evidence into predictive simulation of cellular behaviours and phenotypes.

Knowledge-guided generative modelling

Knowledge resources and generative models jointly support two complementary functions: structured knowledge assembly and cross-modal prediction (Fig. 2). Knowledge assembly systems such as BioGPT[50] integrate dispersed literature into structured knowledge graphs, whereas INDRA combines machine-read evidence with curated databases to generate mechanistic pathways[51]. Although not directly embedded within the yeast AIVC, these systems could provide structured biological priors that enhance interpretability and support hypothesis generation[52]. When coupled with experimental validation (for example, targeted assays to test model-predicted kinase–substrate interactions), this closes the iterative discovery loop.

Cross-modal predictors, including SPIDER[53] and scLinear[54], enable transcriptome-to-proteome or modality-to-modality inference, extending constraint-aware metabolome–proteome translation frameworks[55]. True generative architectures instead learn joint distributions, enabling the synthesis of biologically plausible cellular states. For example, conditional diffusion models such as Stem generate full gene expression profiles from histology images while preserving stochastic variability[56]. This represents a paradigm shift: regression models interpolate within observed data, whereas generative systems such as Stem synthesize novel, biologically consistent configurations[56], preserving variation that matches empirical data.

Static architecture of virtual yeast

The static architecture pillar establishes a spatial blueprint by integrating structural knowledge across scales (micron to Ångstrom) into a unified computational atlas. Leveraging the multimodal resources of *S. cerevisiae*, including morphological atlases, spatial omics and near-atomic cryo-electron tomography (cryo-ET), this framework preserves spatial hierarchies essential for modelling cellular logic (Fig. 2).

Sequencing and mass spectrometry-based spatial omics

Sequencing-based and mass spectrometry-based spatial omics extend this hierarchy by enabling in situ mapping of dozens to thousands of biomolecules across scales[57]. Given the small size of yeast cells (3–10 µm in diameter), high spatial resolution is essential. At the RNA level, Stereo-seq combines DNA nanoball-patterned arrays with in situ capture to achieve large-field imaging at cellular resolution[58], potentially resolving organelle-specific mRNA localization during stress. For proteins and lipids, proximity labelling (BioID and APEX) and organelle purification have resolved compartmentalized networks[59,60]. Photo-biotinylation enables spatially resolved proteomics without genetic engineering[61]. Related photocatalytic and light-activatable labelling strategies extend spatial proteomics to primary samples and controlled manipulation of organelle-localized interactions[62,63]. These imaging-integrated strategies enable interrogation of organelle-specific or stress-responsive proteomes without labour-intensive construction of endogenously tagged libraries.

Expansion microscopy further bridges resolution gaps. Methods such as ExPRESSO and TEMI enable multiplexed imaging of proteins, lipids and metabolites at subcellular and single-cell resolution[64,65]. GAMSI further improves MALDI resolution for nanoscale lipid mapping[66]. For discovery proteomics, FAXP combined with laser capture microdissection allows quantitative profiling of approximately 2,368 proteins from single nuclei[67]. Together, these approaches enable high-resolution quantification of organelle-specific proteoforms and lipid distributions.

Morphology and fluorescence microscopy

Multiscale imaging defines the structural constraints on molecular interactions. Differential interference contrast microscopy has documented phenotypes across more than 5,000 deletion mutants[68] and natural isolates[69]. Fluorescence microscopy of over 4,000 haploid deletion strains has mapped subcellular organization[70]. Time-resolved high-content imaging combined with deep learning profiled autophagy dynamics genome wide across 5,919 mutants[71].

For molecular mapping, spatial transcriptomics approaches such as SeqFISH+ enable simultaneous imaging of thousands of chromosomal loci, chromatin marks and RNA transcripts within individual nuclei[72]. GFP-tagged strain libraries have resolved localization of 4,156 proteins[73] and 400 lipid regulators[74], whereas dynamic analyses using CycleNET and DeepLoc link proteome reorganization to cell cycle stages[75]. Super-resolution methods including STED and PALM achieve nanoscale visualization of spindle pole body duplication and contractile ring formation[76]. DNA-PAINT approaches such as SUM-PAINT extend multiplexing to less than 15-nm resolution, enabling single-molecule spatial proteomics and discovery of new structures[77]. Collectively, these techniques enable dissection of spatial regulation of cellular processes, and adapting them to the smaller dimensions of yeast will

Box 2

Representative foundation models and their adaptation for yeast AIVC

A diverse set of AI models and methods provides the computational backbone for constructing AIVCs, spanning genomic, transcriptomic, proteomic, metabolic, chemical and imaging modalities. Their application to yeast requires varying degrees of domain adaptation and integration with organism-specific datasets.

At the genomic level, models such as Nucleotide Transformer[136], Evo[3], DeepSEA[137], AlphaGenome[4] and Enformer[138] capture regulatory grammars and long-range chromatin interactions. As these models are predominantly trained on non-yeast genomes, adaptation to yeast requires fine-tuning using yeast-specific resources, including genome sequences, chromatin immunoprecipitation followed by sequencing, assay for transposase-accessible chromatin using sequencing, and transcription factor-binding profiles from databases such as the SGD, YEASTRACT+ and YeasTSS, as well as comprehensive interaction maps in *Schizosaccharomyces pombe*[139].

Transcriptome-focused models, including Geneformer[2], scFoundation[140], scGPT[141] and GET[142], encode gene expression patterns, gene–gene interactions and transcriptional dynamics. Although trained mainly on human data, these models benefit from domain adaptation using yeast transcriptomic datasets, such as those from the 1002 Genomes Project, ScRAPdb, YeastNet and stress-response studies.

Protein-level modelling is enabled by frameworks such as AlphaFold[1], ESM[22], DiffDock[143], EquiBind[144] and PIPR[145], which predict protein structures, complexes and molecular interactions. These models are largely transferable across species but require validation and contextualization using yeast-specific data, including protein–protein interactions from BioGRID and yeast interactome studies, structural and imaging evidence from EMPIAR, proteomics evidence from PaxDb and Yeast PeptideAtlas, and kinase annotation from YeastKinome.

For small-molecule representation, models such as MolCLR[146] and Uni-Mol[105] encode chemical properties and conformations. These approaches are generally transferable and typically require minimal fine-tuning, but their application benefits from yeast-relevant compound datasets, including metabolites, stressors and inhibitors curated in YMDB, MOSAIC, the HIP HOP database and the Metabolic Atlas.

GEMs[104] provide a complementary mechanistic framework, linking genes to metabolic reactions through gene–protein-reaction associations, while constraining metabolic fluxes based on reaction stoichiometry. This framework involves light parameter refinement to account for strain-specific flux distributions and the incorporation of newly characterized reactions. This refinement leverages strain-resolved genomic, proteomic, metabolomic and fluxomic datasets from resources such as the Metabolic Atlas, PaxDb, the 5k Yeast Proteome project, YMDB, the 1002 Genomes Project and the SGD.

Finally, imaging-based architectures and models, including U-Net[147], Cellpose[148], CytoSelf[149] and IMPA[150], can segment cells and encode cellular morphology and protein localization patterns from microscopy data. These models require adaptation to yeast-specific imaging features, particularly those associated with cell wall structure and morphology, using datasets from TheCellVision, YeastGFP, SCMD2, LoQAtE, PARPAL and EMPIAR.

Together, these complementary model classes enable multiscale representation and integration, forming the foundation for yeast AIVC construction.

require continued methodological innovation to achieve comparable spatial resolution and throughput.

Cryo-ET and volume electron microscopy

Complementing optical imaging, cryo-ET visualizes native ultrastructure at near-atomic resolution, providing ground-truth data on molecular organization within crowded cellular environments[78]. Advances in focused ion beam milling and subtomogram averaging enable sub-nanometre reconstructions of chromatin organization, autophagy and ribosome assembly[79]. However, cryo-ET remains technically demanding and low throughput, and is therefore predominantly applied to targeted validation or high-impact structural discoveries.

Volume electron microscopy reconstructs mesoscale architectures across larger volumes at nanometre resolution[80], contextualizing cryo-ET insights within whole-cell organization[81]. Gigavoxel-scale segmentation quantifies organelle relationships across hundreds of yeast cells[82].

These datasets exhibit inverse scaling: higher spatial resolution often limits throughput but yields more precise structural information. Such precision anchors quantitative models, whereas structural priors derived from high-resolution data help to deconvolute lower-resolution datasets. Sub-organelle maps therefore refine single-cell profiles, establishing a closed knowledge loop across spatial scales.

Compartmentalization and cell states

Cellular organization is constrained by finite physical volumes that dynamically synchronize with cell cycle progression and stress responses. Examples include endoplasmic reticulum expansion during the unfolded protein response and peroxisome proliferation during fatty acid metabolism. Compartmentalization enables functional specialization via vesicular trafficking and membrane contact sites, where localization dictates activity. Rapid structural remodelling, such as minute-scale Golgi turnover, requires that virtual yeast modelling integrates static architecture with trafficking networks and compartmental states. Emerging modalities such as expansion microscopy and AI-driven multimodal analysis further enhance real-time, label-free structural mapping.

Generative AI for spatial modelling

Although non-generative models encode spatial priors, generative approaches reconstruct holistic architectures from fragmented inputs (Fig. 2). Spatially aware transformers, inspired by masked autoencoders[83], may predict molecular distributions from sparse spatial omics or partial cryo-ET lamellae. Diffusion models may probabilistically reconstruct protein localization from sparse expansion proteomics data, generating more complete cellular spatial maps. PLATO similarly integrates microfluidic strip sampling with transfer learning to reconstruct approximately 25-µm-resolved proteomic maps across tissue sections[84]. More recently, AI-empowered expansion proteomics has further advanced this concept by combining expansion-based strip-resolved proteome measurements with convolutional neural network-based spatial inference to enable whole-slide, single-cell-resolved spatial proteomics[85]. Although specific to mammalian tissues, these examples illustrate how AI can bridge sparse measurements with high-resolution

atlases and suggest that analogous strategies could be developed for yeast at finer scales. Together, these advances help to address a critical constraint in yeast AIVC development: the relatively low throughput of mass spectrometry-based spatial proteomics.

Dimensional lifting approaches, such as depth-aware neural radiance fields, convert two-dimensional (2D) fluorescence images into volumetric reconstructions[86]. Beyond visualization, three-dimensional (3D) synthesis enables functional prediction, as protein complexes operate within defined geometric constraints[87]. Reconstructing their 3D organization supports mechanistic inference of biosynthetic efficiency and metabolic state. Together, these advances anchor the virtual yeast spatial blueprint as a quantitatively resolved scaffold on which subsequent dynamic modules can operate with structural fidelity.

Dynamic states of cellular life

If static architecture defines the spatial grammar, dynamic states encode the temporal logic through which cells interpret internal and external signals. This pillar links structural blueprints to functional plasticity by mapping how perturbations rewire molecular networks across timescales. Central to this effort is an active-learning strategy in which the AIVC predicts cellular responses and prioritizes high-impact perturbations, forming a self-reinforcing experiment cycle. Coupled with automation, this framework enables efficient exploration of perturbational landscapes.

Genetic perturbation dynamics

S. cerevisiae remains uniquely suited for systematic perturbation biology, offering unmatched density of genotype–phenotype measurements, which have revealed deep cellular interdependencies across scales. High-resolution imaging of deletion mutants has demonstrated how gene loss reshapes morphology: 673 homozygous deletions were classified into 7 morphological categories using differential interference contrast microscopy, linking shape abnormalities to disrupted processes[68]. Automated fluorescence imaging of cellular structures across 4,718 haploid mutants quantified 254 morphological parameters, enabling functional gene annotation by phenotypic similarity[70]. Perturbation of 5,919 mutants under nitrogen shifts further uncovered a hierarchical autophagy network, with deep learning distinguishing ultrasensitive, hyposensitive and hyperactive dynamic responses[71].

These phenotypic signatures are complemented by molecular dissection. Transcriptomic analyses of 1,484 regulator deletions[88], 3,500 knockouts under osmotic stress[89] and natural isolates[41] revealed coordinated gene expression programs shaping cell states. Genome-wide deletion proteomics has shown how gene loss reconfigures protein abundance and interaction networks, exposing pathway interdependence and compensatory mechanisms[44,90]. Metabolomic profiling of approximately 5,000 mutants has demonstrated that one-third of yeast genes reshape the amino acid metabolome, particularly those governing chromatin, translation and transport[47]. Integrative genetic interaction mapping of 700,000 double mutants delineated functional hierarchies among 1,471 genes and highlighted RNA-processing pathways in translational control[91]. Together, these multi-omic perturbation datasets chart the dynamic cellular wiring that links genotype to phenotype.

Chemical perturbation dynamics

Chemical perturbations add complementary dimensionality, probing small-molecule–gene–environment interactions through bioactive compounds, metabolites and nutrient shifts. Screens of mutant libraries have revealed therapeutic targets and metabolic adaptations[92]. For example, targets of over 80 bioactive compounds were identified through chemogenomic profiling[93]. Phosphoproteomic analyses across 101 stress conditions mapped stress-responsive signalling circuits[94], whereas high-flow scanning SWATH profiling of 16 drugs has distinguished mechanistic signatures for azoles, statins and antifolates[95]. Expansion to 3,250 compounds resolved 45 major response patterns linking genes, chemicals and biological processes[96]. Such studies define a multidimensional chemical response atlas for the virtual yeast.

Environmental perturbation dynamics

Physical perturbations further illuminate adaptive logic. A conserved environmental stress response underscores network interconnectedness[97]. Genome-wide screens under ionizing radiation have uncovered over 100 loci influencing tolerance, many involved in conserved DNA repair, chromatin remodelling and checkpoint functions, with parallels to human cancer genes[98]. These findings situate yeast within a conserved framework of stress adaptation while reinforcing its utility as a systems-level testbed.

Despite this progress, challenges remain. Current perturbation maps are constructed solely under standard laboratory conditions, lack full combinatorial coverage and sufficient temporal resolution to resolve intermediate states. Multi-omics integration across perturbations remains computationally demanding, and most datasets average population responses, obscuring single-cell heterogeneity. Future advances will require higher-throughput temporal sampling, single-cell resolution and machine learning models that are capable of predicting cellular plasticity under unseen perturbations[99]. Ultimately, incorporating experimental evolution data may allow prediction of long-term adaptive trajectories, with implications for biomedicine and bioengineering.

Active learning experimentation

To address combinatorial complexity, an autonomous closed-loop framework coupling large-scale yeast perturbation profiling with active learning is required (Fig. 3a). This system grounds the 'AI flywheel' in an experimental pipeline that iteratively improves virtual modelling through data-guided experimental design (Fig. 3a).

In Data v0.1, perturbations were selected to maximize genotype–environment diversity (Fig. 3b). From 969 natural strains with paired genomic and transcriptomic profiles, a diverse core set was chosen using a core-set algorithm in a fused latent space integrating evolutionary (Evo 2 embeddings[3]) and transcriptional features. Each strain was profiled under varied carbon and nitrogen sources, temperatures and chemical stressors. The dataset comprises over 15,000 time-resolved proteomic profiles across 6 time points per perturbation, paired with early and late metabolomic measurements and growth curves, establishing a multimodal foundation for AIVC training.

Data acquisition integrates automated perturbation handling, growth monitoring and standardized multi-omics processing (Fig. 3b), including robotics-assisted proteomics and metabolomics analysis. Future iterations will incorporate sensor-guided and vision-guided feedback, with LLM-driven AI agents coordinating real-time monitoring, troubleshooting and experimental iteration, moving towards a fully autonomous pipeline.

On this foundation, the WAY Active Learning framework (WAY-AL v0.1) integrates probabilistic modelling, generative modelling and adaptive experiment selection (Fig. 3b). Generative models, including variational autoencoders, diffusion models, flow-based models and generative transformers, act as virtual experimenters to simulate probabilistic cellular trajectories. To identify gaps in current knowledge of the model, these predictions are coupled with ensemble-based or dropout-based uncertainty quantification methods that evaluate predictive reliability across conditions. Advanced acquisition strategies, driven by reinforcement learning[100] or Bayesian optimization[101], then select perturbations maximizing expected information gain for subsequent rounds (Data v0.2). By prioritizing high-uncertainty or information-rich conditions, the system maximizes model improvement under limited experimental budgets, paralleling autonomous

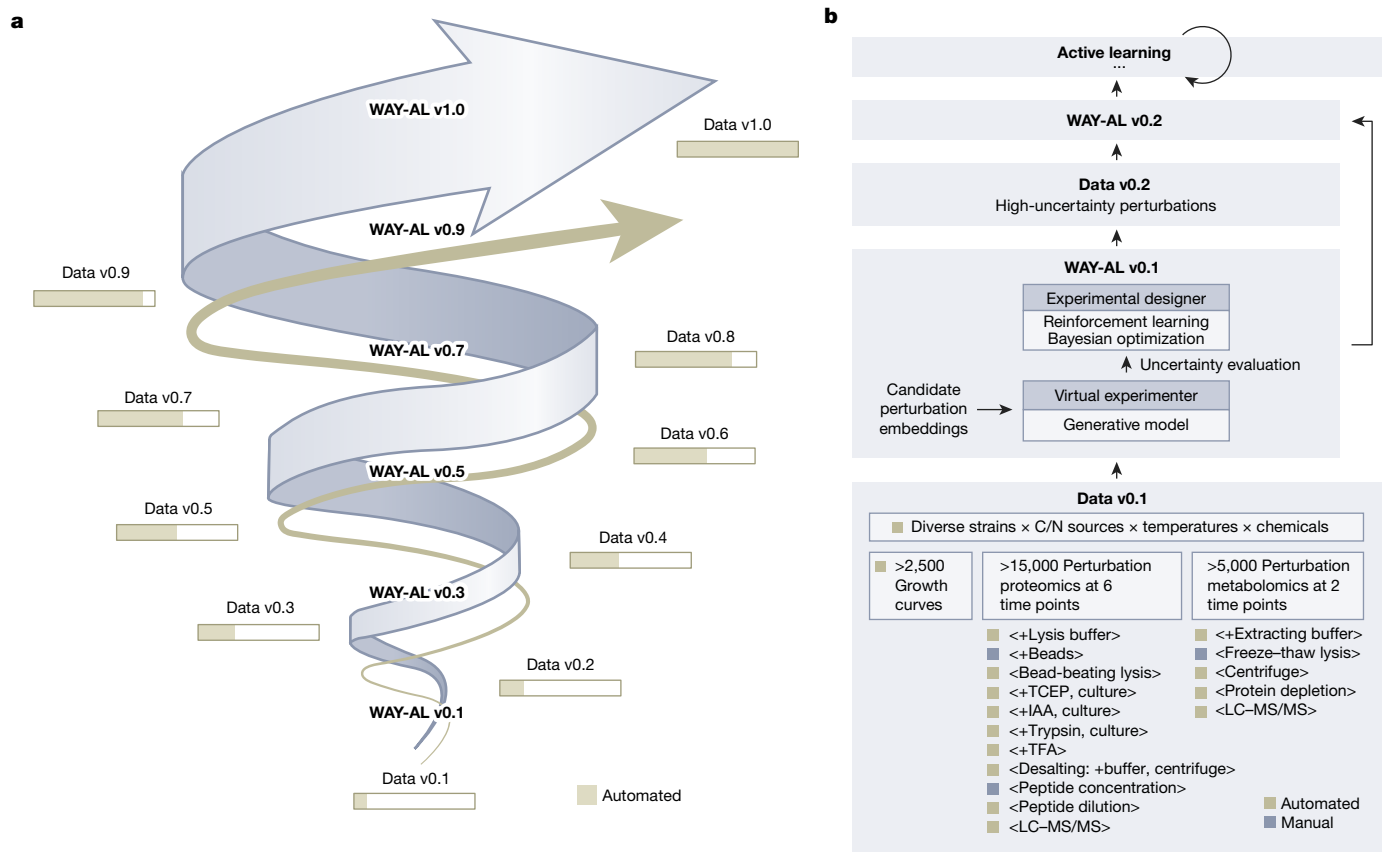

Fig. 3 | Closed-loop active learning drives iterative refinement of virtual yeast models. a, Schematic of the AI flywheel that enables autonomous model refinement. Initial multi-omics datasets (Data v0.1) establish a baseline virtual yeast model (WAY-AL v0.1), which identifies high-impact perturbations predicted to maximize information gain about poorly characterized biological processes. Automated platforms execute these targeted perturbations through robotic multi-omics profiling, generating updated datasets that refine the model. Iterative cycles progressively improve both the virtual model and the experimental design, forming a self-reinforcing framework that accelerates mapping of cellular plasticity while compressing discovery timelines. **b**, Implementation of the metabolic and synthetic biology module in *S. cerevisiae* (Data v0.1). Twelve genetically and functionally diverse strains are selected from a panel of 969 based on integrated evolutionary and transcriptional embeddings, and subjected to over 200 environmental and chemical perturbations. Across these conditions, over 15,000 time-resolved proteomic profiles, over 5,000 metabolomic measurements and paired growth curves have been generated. The schematic illustrates the level of automation across experimental steps: the light brown boxes indicate fully or semi-automated procedures, whereas the dark grey boxes denote processes pending automation. These datasets form the training foundation for WAY-AL v0.1, which integrates probabilistic modelling, generative learning and adaptive experiment selection to guide subsequent perturbations (Data v0.2) through a closed model–experiment feedback loop. C/N, carbon/nitrogen; IAA, iodoacetamide; LC–MS/MS, liquid chromatography–tandem mass spectrometry; TCEP, tris(2-carboxyethyl)phosphine; TFA, trifluoroacetic acid.

chemical platforms such as ChemOS[102,103], but extending them to cellular biology. This closed-loop structure progressively enhances model accuracy and experimental informativeness, creating a self-reinforcing discovery cycle (Fig. 3a).

Metabolic and synthetic biology tool

To illustrate how the eight functional modules of the virtual yeast operate, we highlight the metabolic and synthetic biology tool as a prototype implementation. This module integrates targeted perturbation experiments with representation learning and generative AI models to capture and predict dynamic metabolic regulation. Its goal is to explain how genetic and environmental variables jointly reshape metabolic fluxes and compound synthesis, thereby enabling rational strain design for optimized metabolite production.

The metabolic core provides the energetic and biochemical foundation of the virtual yeast, integrating cytosolic, mitochondrial and peroxisomal pathways governing energy conversion, redox balance and small-molecule metabolism. Its construction rests on three pillars: a priori knowledge, spatial architecture and dynamic state[16]. Mechanistic priors derive from yeast-GEM[104], YMDB 2.0 (ref. 31) and SGD[30], forming a multimodal scaffold of reaction stoichiometry, pathway topology and regulatory logic. Spatial information from cryo-ET, expansion-based spatial proteomics and fluorescence imaging anchors reactions within organelles, capturing compartment-specific fluxes and contact-dependent regulation. Dynamic states are iteratively refined through the AI flywheel (Data v0.1–v1.0), which uses active learning to design perturbations under nutrient and stress conditions (Fig. 3).

Computationally, the module adopts a hybrid architecture (Fig. 2). Genomic embeddings derived from Evo 2 (ref. 3) and fine-tuned on *S. cerevisiae* population data (Box 2) encode genetic capacity and regulatory structure. Environmental contexts are represented through nutrient and compound embeddings analogous to Uni-Mol[105], whereas proteomic and metabolomic embeddings summarize molecular states. Conditional generative models, including variational autoencoders, diffusion models and transformer-based architectures, map genotype and perturbation inputs to proteomic and metabolomic outputs. GEMs constrain flux distributions within this learned latent space, ensuring biochemical plausibility.

Performance can be benchmarked against GEMs[104], which accurately recapitulate growth phenotypes across genetic and environmental conditions (mean $R^2 \approx 0.84$), yet primarily operate at the level of flux

optimization and do not quantitatively resolve intracellular metabolite concentrations or dynamic biomass composition. Data-driven models based on enzyme abundance provide partial resolution at the metabolite level. For instance, a machine-learning model trained on proteomics data from 97 kinase knockout strains achieved cross-validated $R^2 \approx 0.55$ in predicting metabolite concentrations[55]. By integrating environmental embeddings and dynamic modelling with paired proteomic–metabolomic perturbation datasets, the AIVC framework aims to extend these approaches towards more generalizable and quantitative prediction of metabolite responses, with particular relevance to high-value and industrially relevant metabolites. This module thus exemplifies how domain-specific AI components can translate structured perturbation data into actionable synthetic biology strategies.

Predictive deliverables of virtual yeast

The virtual yeast is both a conceptual framework and an operational prediction platform. It generates quantitative, testable outputs across synthetic and fundamental biology, linking in silico design and automated validation. In synthetic biology, metabolic and biosynthetic tools simulate flux redistribution under genetic and environmental perturbations to identify yield bottlenecks and optimal design strategies. Benchmarked against GEMs[104] and enzyme-expression-based predictors[55], the AIVC framework is expected to enhance metabolite-level inference and support inverse strain design. In fundamental biology, the coordinated agent models spatiotemporal dynamics during cell cycle progression, meiosis and ageing, generating testable hypotheses on lifespan trade-offs and conserved regulatory nodes. In compound screening and target validation, the platform treats yeast as an experimentally tractable eukaryotic system for systematic evaluation of chemical libraries, identifying synthetic-lethal interactions, pathway vulnerabilities and resistance circuits.

The metabolic and synthetic biology tool represents the first operational milestone. Large-scale perturbation datasets integrating proteomes, metabolomes and growth curves under systematically varied conditions (Fig. 3) are being generated to model growth and metabolite yields. This will establish a benchmarked predictive layer in the phase I study.

Subsequent milestones include horizontal integration of modules (for example, lipid metabolism and trafficking to capture cross-compartment coupling) and vertical expansion towards ageing and compound screening. These developments will advance iteratively, with each module inheriting shared embeddings and evaluation frameworks established in the metabolic prototype. In the phase II study, validated agents for 2–3 organelle systems are envisioned. The phase III study will be focused on development and integration of all functional modules into a unified whole-cell AI agent. With sustained support, development of virtual yeast may span 5–10 years.

Consortium architecture

The virtual yeast initiative is organized around four interdependent cores enabling end-to-end development of predictive AI agents. The biology core establishes mechanistic ground truth and validates predictions. The data and technology core coordinates multimodal data generation, including but not limited to omics and imaging upon perturbations. The AI core develops hybrid computational frameworks integrating embeddings and generative models. The applications core translates predictions into synthetic biology (for example, metabolic flux optimization), fundamental biology (such as meiotic dynamics) and compound discovery.

Module-specific working groups integrate biological vignettes into model design. For example, the stress processors group constructs nutrient-specific, temperature-specific and genotype-specific intervention matrices, generates time-resolved multi-omics data, and trains the agent to simulate stress responses across organelles, including mitochondrial–nuclear–vacuolar coupling. Comparable workflows support metabolic flux rerouting under nutrient limitation and prediction of age-associated trafficking defects arising from organelle dysfunction. These scenarios serve as executable testbeds for evaluating perception, routing and causal simulation beyond static phenotype prediction, and enable comparison with existing models[5,20]. Evaluation relies on expert-defined benchmarks using experimentally tractable phenotypes, assessing generalization, recovery of regulatory relationships and generation of testable hypotheses.

Conclusion and outlook

S. cerevisiae provides a practical and useful entry point for developing predictive virtual cells with relevance to biomedicine and biotechnology. As a eukaryote, it shares many conserved cellular components and processes with human systems, yet remains a tractable single-cell organism supported by extensive datasets and experimental resources. Here we outlined a biological framework and technical path towards a virtual yeast that integrates a priori knowledge, spatial organization and perturbation-driven dynamics. Although many emerging AIVC efforts are led primarily by advances in AI and large-scale modelling, our perspective emphasizes starting from cellular function, experimental data and testable hypotheses, with AI serving as an integrating engine. We emphasize not only a biology-first perspective but also biology-driven active learning that continuously refines data generation and computational reasoning strategies.

Several principles shape this effort. First, a virtual yeast is not equivalent to an ideal digital twin of the cell. Rather than attempting exhaustive representation of all molecules and modalities, practical virtual yeast models can focus on defined goals, such as predicting metabolite production or understanding processes such as stress responses or membrane trafficking, while remaining experimentally testable. Second, we propose organizing virtual-cell intelligence around function-centred AI modules that together form an integrated agent. Compared with models structured mainly by molecular type or scale, functional modules align more naturally with cellular operations and perturbations, although integrating multimodal datasets across modules remains challenging. Third, biological complexity must be acknowledged: interactions among modules, links across data modalities, genetic background effects among yeast strains, incomplete variant-level phenotypic coverage, and the integration of experimental measurements with knowledge bases all introduce uncertainty and demand interpretable modelling.

Looking ahead, initial generations of virtual yeast models may emerge within the next 5–10 years. These systems are unlikely to be digital twins of yeast, and the gap between models and living cells will remain substantial. During the project, rapid advances in AI and lab-in-a-loop architectures are likely to reshape both methods and implementation strategies, therefore the approaches discussed in this roadmap may evolve. Even so, pursuing virtual yeast is expected to yield experimentally useful AI tools, for example, in synthetic biology and mechanistic discovery, and provide a transferable blueprint for more complex eukaryotic systems.

1. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
2. Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
3. Brixi, G. et al. Genome modelling and design across all domains of life with Evo 2. *Nature* **652**, 1349–1361 (2026).
4. Avsec, Z. et al. Advancing regulatory variant effect prediction with AlphaGenome. *Nature* **649**, 1206–1218 (2026).
5. Bunne, C. et al. How to build the virtual cell with artificial intelligence: priorities and opportunities. *Cell* **187**, 7045–7063 (2024).
This perspective article defines the concept of the AIVC and outlines the design principles and collaborative strategies needed to realize AI-driven, multiscale simulations of a living system.

6. Karr, J. R. et al. A whole-cell computational model predicts phenotype from genotype. *Cell* **150**, 389–401 (2012).
This pioneering study presents the first mechanistic whole-cell model that integrates all molecular processes of *M. genitalium* using diverse mathematical formalisms to unify fundamentally different cellular processes and experimental measurements, enabling genotype-to-phenotype prediction.
7. Macklin, D. N. et al. Simultaneous cross-evaluation of heterogeneous *E. coli* datasets via mechanistic simulation. *Science* **369**, eaav3751 (2020).
8. Lu, H., Kerkhoven, E. J. & Nielsen, J. Multiscale models quantifying yeast physiology: towards a whole-cell model. *Trends Biotechnol.* **40**, 291–305 (2022).
9. Sanchez, B. J. et al. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol. Syst. Biol.* **13**, 935 (2017).
10. Elsemman, I. E. et al. Whole-cell modeling in yeast predicts compartment-specific proteome constraints that drive metabolic strategies. *Nat. Commun.* **13**, 801 (2022).
11. Oftadeh, O. et al. A genome-scale metabolic model of Saccharomyces cerevisiae that integrates expression constraints and reaction thermodynamics. *Nat. Commun.* **12**, 4790 (2021).
12. Adduri, A. K. et al. Predicting cellular responses to perturbation across diverse contexts with State. Preprint at *bioRxiv* https://doi.org/10.1101/2025.06.26.661135 (2025).
This study introduces State, a scalable AI framework trained on perturbation transcriptomic data, from over 100 million cells, that predicts cellular responses across unseen contexts, advancing the dynamic modelling of perturbation effects central to AIVC development.
13. Sun, R. et al. A perturbation proteomics-based foundation model for virtual cell construction. Preprint at *bioRxiv* https://doi.org/10.1101/2025.02.07.637070 (2025).
This study introduces ProteinTalks, a neural ordinary differential equation-based foundation model trained on 38 million perturbed protein measurements, which learns cellular protein network dynamics to predict drug efficacy, synergy and resistance, laying a proteome-centric foundation for virtual cell development.
14. Jakobson, C. M. et al. A genome-to-proteome map reveals how natural variants drive proteome diversity and shape fitness. *Science* **390**, eadu3198 (2025).
15. Rizvi, S. A. et al. Scaling large language models for next-generation single-cell analysis. Preprint at *bioRxiv* https://doi.org/10.1101/2025.04.14.648850 (2025).
16. Qian, L., Dong, Z. & Guo, T. Grow AI virtual cells: three data pillars and closed-loop learning. *Cell Res.* **35**, 319–321 (2025).
This perspective article proposes the 'three data pillars'—a priori knowledge, static architecture and dynamic states—as the foundation for building AIVCs, and introduces closed-loop active learning systems to autonomously refine virtual cell models.
17. Guo, T., Steen, J. A. & Mann, M. Mass-spectrometry-based proteomics: from single cells to clinical applications. *Nature* **638**, 901–911 (2025).
18. Bauermeister, A., Mannochio-Russo, H., Costa-Lotufo, L. V., Jarmusch, A. K. & Dorrestein, P. C. Mass spectrometry-based metabolomics in microbiome investigations. *Nat. Rev. Microbiol.* **20**, 143–160 (2022).
19. Cui, H. et al. Towards multimodal foundation models in molecular cell biology. *Nature* **640**, 623–633 (2025).
20. Song, L., Segal, E. & Xing, E. Toward AI-driven digital organism: multiscale foundation models for predicting, simulating and programming biology at all levels. Preprint at https://arxiv.org/abs/2412.06993 (2024).
21. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nat. Med.* **28**, 1773–1784 (2022).
22. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
23. Yang, L. et al. Diffusion models: a comprehensive survey of methods and applications. *ACM Comput. Surv.* **56**, 1–39 (2023).
24. Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M. & Le, M. Flow matching for generative modeling. Preprint at https://arxiv.org/abs/2210.02747 (2022).
25. Shazeer, N. et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. Preprint at https://arxiv.org/abs/1701.06538 (2017).
26. Fedus, W., Zoph, B. & Shazeer, N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.* **23**, 5232–5270 (2022).
27. Bai, L. et al. Intern-s1: a scientific multimodal foundation model. Preprint at https://arxiv.org/abs/2508.15763 (2025).
28. Li, Y. et al. Uni-moe: scaling unified multimodal llms with mixture of experts. *IEEE Trans. Patt. Anal. Mach. Intell.* **47**, 3424–3439 (2025).
29. Mroueh, Y. Reinforcement learning with verifiable rewards: GRPO's effective loss, dynamics, and success amplification. Preprint at https://arxiv.org/abs/2503.06639 (2025).
30. Cherry, J. M. et al. SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.* **26**, 73–79 (1998).
31. Ramirez-Gaona, M. et al. YMDB 2.0: a significantly expanded version of the yeast metabolome database. *Nucleic Acids Res.* **45**, D440–D445 (2017).
32. Kim, H. et al. YeastNet v3: a public database of data-specific and integrated functional gene networks for *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **42**, D731–D736 (2014).
33. Breker, M., Gymrek, M., Moldavski, O. & Schuldiner, M. LoQAtE—localization and quantitation atlas of the yeast proteome. A new tool for multiparametric dissection of single-protein behavior in response to biological perturbations in yeast. *Nucleic Acids Res.* **42**, D726–D730 (2014).
34. Koh, J. L. et al. CYCLoPs: a comprehensive database constructed from automated analysis of protein abundance and subcellular localization patterns in *Saccharomyces cerevisiae*. *G3* **5**, 1223–1232 (2015).
35. Costanzo, M. et al. A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, aaf1420 (2016).
This landmark study maps nearly 1 million genetic interactions in yeast, revealing the global wiring diagram of cellular function and establishing a quantitative framework for decoding genotype-to-phenotype relationships.
36. Usaj, M. et al. TheCellMap.org: a web-accessible database for visualizing and mining the Global Yeast Genetic Interaction Network. *G3* **7**, 1539–1549 (2017).
37. Braberg, H. et al. Genetic interaction mapping informs integrative structure determination of protein complexes. *Science* **370**, eaaz4910 (2020).
38. Michaelis, A. C. et al. The social and structural architecture of the yeast protein interactome. *Nature* **624**, 192–200 (2023).
Using high-throughput affinity purification–mass spectrometry, this work generates a nearly complete yeast protein–protein interaction map, uncovering the dense and modular architecture of the cellular interactome.
39. Peter, J. et al. Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **556**, 339–344 (2018).
This population-scale genomic analysis of 1,011 yeast isolates reveals the evolutionary trajectories and domestication history of *S. cerevisiae*, providing a comprehensive resource for genotype–phenotype studies.
40. Caudal, E. et al. Pan-transcriptome reveals a large accessory genome contribution to gene expression variation in yeast. *Nat. Genet.* **56**, 1278–1287 (2024).
41. Muenzner, J. et al. Natural proteome diversity links aneuploidy tolerance to protein turnover. *Nature* **630**, 149–157 (2024).
42. Loegler, V. et al. From genotype to phenotype with 1,086 near telomere-to-telomere yeast genomes. *Nature* **648**, 649–658 (2025).
43. Puddu, F. et al. Genome architecture and stability in the *Saccharomyces cerevisiae* knockout collection. *Nature* **573**, 416–420 (2019).
44. Messner, C. B. et al. The proteomic landscape of genome-wide genetic perturbations. *Cell* **186**, 2018–2034.e21 (2023).
45. Schulte, U. et al. Mitochondrial complexome reveals quality-control pathways of protein import. *Nature* **614**, 153–159 (2023).
46. Tengolics, R. et al. The metabolic domestication syndrome of budding yeast. *Proc. Natl Acad. Sci. USA* **121**, e2313354121 (2024).
47. Mulleder, M. et al. Functional metabolomics describes the yeast biosynthetic regulome. *Cell* **167**, 553–565.e12 (2016).
48. Galardini, M. et al. The impact of the genetic background on gene deletion phenotypes in *Saccharomyces cerevisiae*. *Mol. Syst. Biol.* **15**, e8831 (2019).
49. Aulakh, S. K. et al. The molecular landscape of cellular metal ion biology. *Cell Syst.* **16**, 101319 (2025).
50. Luo, R. et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.* **23**, bbac409 (2022).
51. Bachman, J. A., Gyori, B. M. & Sorger, P. K. Automated assembly of molecular mechanisms at scale from text mining and curated databases. *Mol. Syst. Biol.* **19**, e11325 (2023).
52. Qu, S. et al. Automating exploratory multiomics research via language models. Preprint at https://arxiv.org/abs/2506.07591 (2025).
53. Chen, R., Zhou, J. & Chen, B. Imputing abundance of over 2,500 surface proteins from single-cell transcriptomes with context-agnostic zero-shot deep ensembles. *Cell Syst.* **15**, 869–884.e6 (2024).
54. Hanhart, D., Gossi, F., Rapsomaniki, M. A., Kruithof-de Julio, M. & Chouvardas, P. ScLinear predicts protein abundance at single-cell resolution. *Commun. Biol.* **7**, 267 (2024).
55. Zelezniak, A. et al. Machine learning predicts the yeast metabolome from the quantitative proteome of kinase knockouts. *Cell Syst.* **7**, 269–283.e6 (2018).
56. Zhu, S., Zhu, Y., Tao, M. & Qiu, P. Diffusion generative modeling for spatially resolved gene expression inference from histology images. Preprint at https://arxiv.org/abs/2501.15598 (2025).
57. Bressan, D., Battistoni, G. & Hannon, G. J. The dawn of spatial omics. *Science* **381**, eabq4964 (2023).
58. Chen, A. et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777–1792.e21 (2022).
59. Christopher, J. A., Geladaki, A., Dawson, C. S., Vennard, O. L. & Lilley, K. S. Subcellular transcriptomics and proteomics: a comparative methods review. *Mol. Cell. Proteomics* **21**, 100186 (2022).
60. Hein, M. Y. et al. Global organelle profiling reveals subcellular localization and remodeling at proteome scale. *Cell* **188**, 1137–1155.e20 (2025).
61. Chen, Y.-D. et al. Microscopy-guided subcellular proteomic discovery by high-speed ultra-content photo-biotinylation. Preprint at *bioRxiv* https://doi.org/10.1101/2023.12.27.573388 (2023).
62. Liu, Z. et al. Bioorthogonal photocatalytic proximity labeling in primary living samples. *Nat. Commun.* **15**, 2712 (2024).
63. Zhu, Y. et al. Genetically encoded bioorthogonal tryptophan decaging in living cells. *Nat. Chem.* **16**, 533–542 (2024).
64. Bai, Y. et al. Expanded vacuum-stable gels for multiplexed high-resolution spatial histopathology. *Nat. Commun.* **14**, 4013 (2023).
65. Zhang, H. et al. TEMI: tissue-expansion mass-spectrometry imaging. *Nat. Methods* **22**, 1051–1058 (2025).
66. Chan, Y. H. et al. Gel-assisted mass spectrometry imaging enables sub-micrometer spatial lipidomics. *Nat. Commun.* **15**, 5036 (2024).
67. Dong, Z. et al. Spatial proteomics of single cells and organelles on tissue slides using filter-aided expansion proteomics. *Nat. Commun.* **15**, 9378 (2024).
68. Giaever, G. et al. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
69. Ohnuki, S. et al. Phenotypic diagnosis of lineage and differentiation during sake yeast breeding. *G3* **7**, 2807–2820 (2017).
70. Ohya, Y. et al. High-dimensional and large-scale phenotyping of yeast mutants. *Proc. Natl Acad. Sci. USA* **102**, 19015–19020 (2005).
71. Chica, N. et al. Time-resolved functional genomics using deep learning reveals a global hierarchical control of autophagy. *Nat. Cell Biol.* **28**, 465–479 (2026).
72. Takei, Y. et al. Integrated spatial genomics reveals global architecture of single nuclei. *Nature* **590**, 344–350 (2021).
73. Huh, W. K. et al. Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
74. Natter, K. et al. The spatial organization of lipid synthesis in the yeast *Saccharomyces cerevisiae* derived from large scale green fluorescent protein tagging and high resolution microscopy. *Mol. Cell. Proteomics* **4**, 662–672 (2005).

75. Litsios, A. et al. Proteome-scale movements and compartment connectivity during the eukaryotic cell cycle. *Cell* **187**, 1490–1507.e21 (2024).
76. Bond, C., Santiago-Ruiz, A. N., Tang, Q. & Lakadamyali, M. Technological advances in super-resolution microscopy to study cellular processes. *Mol. Cell* **82**, 315–332 (2022).
77. Unterauer, E. M. et al. Spatial proteomics in neurons at single-protein resolution. *Cell* **187**, 1785–1800.e16 (2024).
This study develops SUM-PAINT, a high-throughput super-resolution method achieving virtually unlimited multiplexing at sub-15-nm resolution and revealing a new VGLUT1$^+$ Gephyrin$^+$ synapse subtype.
78. Nogales, E. & Mahamid, J. Bridging structural and cell biology with cryo-electron microscopy. *Nature* **628**, 47–56 (2024).
79. Young, L. N. & Villa, E. Bringing structure to cell biology with cryo-electron tomography. *Annu. Rev. Biophys.* **52**, 573–595 (2023).
80. Peddie, C. J. et al. Volume electron microscopy. *Nat. Rev. Methods Primers* **2**, 51 (2022).
81. McCafferty, C. L. et al. Integrating cellular electron microscopy with multimodal data to explore biology across space and time. *Cell* **187**, 563–584 (2024).
82. Nesic, N. et al. Automated segmentation of cell organelles in volume electron microscopy using deep learning. *Microsc. Res. Tech.* **87**, 1718–1732 (2024).
83. He, K. et al. Masked autoencoders are scalable vision learners. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (eds Dana, K. et al.) 16000–16009 (IEEE, 2022).
84. Hu, B. et al. High-resolution spatially resolved proteomics of complex tissues based on microfluidics and transfer learning. *Cell* **188**, 734–748.e22 (2025).
This study presents PLATO, a microfluidics-based and AI-based framework enabling high-resolution spatial proteomics across whole tissues and revealing distinct tumour subtypes in human breast cancer.
85. Wang, S. et al. Multimodal AI-enabled mass spectrometry-based expansion proteomics for whole-slide at single-cell resolution. Preprint at *LangTaoSha* https://doi.org/10.65215/LTSpreprints.2026.02.20.000134 (2026).
86. Shih, M.-L., Su, S.-Y., Kopf, J. & Huang, J.-B. 3D photography using context-aware layered depth inpainting. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (eds Liu, C. et al.) 8028–8038 (IEEE, 2020).
87. Hammer, S. K. & Avalos, J. L. Harnessing yeast organelles for metabolic engineering. *Nat. Chem. Biol.* **13**, 823–832 (2017).
88. Kemmeren, P. et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* **157**, 740–752 (2014).
89. Nadal-Ribelles, M. et al. A single-cell resolved genotype-phenotype map using genome-wide genetic and environmental perturbations. *Nat. Commun.* **16**, 2645 (2025).
90. Ozturk, M. et al. Proteome effects of genome-wide single gene perturbations. *Nat. Commun.* **13**, 6153 (2022).
91. Decourty, L., Malabat, C., Frachon, E., Jacquier, A. & Saveanu, C. Investigation of RNA metabolism through large-scale genetic interaction profiling in yeast. *Nucleic Acids Res.* **49**, 8535–8555 (2021).
92. di Bernardo, D. et al. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.* **23**, 377–383 (2005).
93. Parsons, A. B. et al. Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast. *Cell* **126**, 611–625 (2006).
94. Leutert, M., Barente, A. S., Fukuda, N. K., Rodriguez-Mias, R. A. & Villen, J. The regulatory landscape of the yeast phosphoproteome. *Nat. Struct. Mol. Biol.* **30**, 1761–1773 (2023).
95. Messner, C. B. et al. Ultra-fast proteomics with Scanning SWATH. *Nat. Biotechnol.* **39**, 846–854 (2021).
96. Lee, A. Y. et al. Mapping the cellular response to small molecules using chemogenomic fitness signatures. *Science* **344**, 208–211 (2014).
97. Gasch, A. P. et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11**, 4241–4257 (2000).
98. Bennett, C. B. et al. Genes required for ionizing radiation resistance in yeast. *Nat. Genet.* **29**, 426–434 (2001).
99. Klein, D. et al. CellFlow enables generative single-cell phenotype modeling with flow matching. Preprint at *bioRxiv* https://doi.org/10.1101/2025.04.11.648220 (2025).
100. DeMeo, B. et al. Active learning framework leveraging transcriptomics identifies modulators of disease phenotypes. *Science* **390**, eadi8577 (2025).
101. Xu, Y. et al. LUMI-lab: a foundation model-driven autonomous platform enabling discovery of ionizable lipid designs for mRNA delivery. *Cell* **189**, 1620–1635.e25 (2026).
102. Burger, B. et al. A mobile robotic chemist. *Nature* **583**, 237–241 (2020).
This pioneering demonstration of an autonomous mobile robotic chemist establishes a closed-loop system for self-driven experimentation, foreshadowing active-learning frameworks for AIVC evolution.
103. Coley, C. W. et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566 (2019).
104. Zhang, C. et al. Yeast9: a consensus genome-scale metabolic model for *S. cerevisiae* curated by the community. *Mol. Syst. Biol.* **20**, 1134–1150 (2024).
105. Ji, X. et al. Uni-mol2: exploring molecular pretraining model at scale. Preprint at https://arxiv.org/abs/2406.14969 (2024).
106. Hartwell, L. H. Yeast and cancer. *Biosci. Rep.* **24**, 523–544 (2004).
107. Neiman, A. M. Sporulation in the budding yeast *Saccharomyces cerevisiae*. *Genetics* **189**, 737–765 (2011).
108. Winzeler, E. A. et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
109. Mulleder, M. et al. A prototrophic deletion mutant collection for yeast metabolomics and systems biology. *Nat. Biotechnol.* **30**, 1176–1178 (2012).
110. Zhu, H. et al. Global analysis of protein activities using proteome chips. *Science* **293**, 2101–2105 (2001).
111. Sopko, R. et al. Mapping pathways and phenotypes by systematic gene overexpression. *Mol. Cell* **21**, 319–330 (2006).
112. Ross-Macdonald, P. et al. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413–418 (1999).
113. Weill, U. et al. Genome-wide SWAp-Tag yeast libraries for proteome exploration. *Nat. Methods* **15**, 617–622 (2018).
114. Roy, K. R. et al. Multiplexed precision genome editing with trackable genomic barcodes in yeast. *Nat. Biotechnol.* **36**, 512–520 (2018).
115. Tong, A. H. et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–2368 (2001).
116. Pan, X. et al. dSLAM analysis of genome-wide genetic interactions in *Saccharomyces cerevisiae*. *Methods* **41**, 206–221 (2007).
117. Boone, C., Bussey, H. & Andrews, B. J. Exploring genetic interactions and networks with yeast. *Nat. Rev. Genet.* **8**, 437–449 (2007).
118. Turco, G. et al. Global analysis of the yeast knockout phenome. *Sci. Adv.* **9**, eadg5702 (2023).
119. Engel, S. R. et al. *Saccharomyces* Genome Database provides mutant phenotype data. *Nucleic Acids Res.* **38**, D433–D436 (2010).
120. Engel, S. R. et al. *Saccharomyces* Genome Database: advances in genome annotation, expanded biochemical pathways, and other key enhancements. *Genetics* **229**, iyae185 (2025).
121. Song, G. et al. Integration of new alternative reference strain genome sequences into the *Saccharomyces* Genome Database. *Database* **2016**, baw074 (2016).
122. Yue, J. X. et al. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat. Genet.* **49**, 913–924 (2017).
123. O'Donnell, S. et al. Telomere-to-telomere assemblies of 142 strains characterize the genome structural landscape in *Saccharomyces cerevisiae*. *Nat. Genet.* **55**, 1390–1399 (2023).
124. Kachroo, A. H. et al. Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science* **348**, 921–925 (2015).
125. Laurent, J. M. et al. Humanization of yeast genes with multiple human orthologs reveals functional divergence between paralogs. *PLoS Biol.* **18**, e3000627 (2020).
126. Persson, E. & Sonnhammer, E. L. L. InParanoiDB 9: ortholog groups for protein domains and full-length proteins. *J. Mol. Biol.* **435**, 168001 (2023).
127. Kachroo, A. H., Vandeloo, M., Greco, B. M. & Abdullah, M. Humanized yeast to model human biology, disease and evolution. *Dis. Model. Mech.* **15**, dmm049309 (2022).
128. Paddon, C. J. & Keasling, J. D. Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. *Nat. Rev. Microbiol.* **12**, 355–367 (2014).
129. Richardson, S. M. et al. Design of a synthetic yeast genome. *Science* **355**, 1040–1044 (2017).
130. Dai, J., Boeke, J. D., Luo, Z., Jiang, S. & Cai, Y. Sc3.0: revamping and minimizing the yeast genome. *Genome Biol.* **21**, 205 (2020).
131. Luo, X. et al. Complete biosynthesis of cannabinoids and their unnatural analogues in yeast. *Nature* **567**, 123–126 (2019).
132. Nielsen, J. & Keasling, J. D. Engineering cellular metabolism. *Cell* **164**, 1185–1197 (2016).
133. Li, G. et al. Yeast metabolism adaptation for efficient terpenoids synthesis via isopentenol utilization. *Nat. Commun.* **15**, 9844 (2024).
134. Miao, Z. et al. ScRAPdb: an integrated pan-omics database for the *Saccharomyces cerevisiae* reference assembly panel. *Nucleic Acids Res.* **53**, D852–D863 (2025).
135. King, N. L. et al. Analysis of the *Saccharomyces cerevisiae* proteome with PeptideAtlas. *Genome Biol.* **7**, R106 (2006).
136. Dalla-Torre, H. et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* **22**, 287–297 (2025).
137. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
138. Avsec, Z. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
139. Skribbe, M. et al. A comprehensive *Schizosaccharomyces pombe* atlas of physical transcription factor interactions with proteins and chromatin. *Mol. Cell* **85**, 1426–1444.e8 (2025).
140. Hao, M. et al. Large-scale foundation model on single-cell transcriptomics. *Nat. Methods* **21**, 1481–1491 (2024).
141. Cui, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).
142. Fu, X. et al. A foundation model of transcription across human cell types. *Nature* **637**, 965–973 (2025).
143. Corso, G., Stärk, H., Jing, B., Barzilay, R. & Jaakkola, T. DiffDock: diffusion steps, twists, and turns for molecular docking. In *The Eleventh International Conference on Learning Representations* (eds Nickel, M. et al.) (ICLR, 2023).
144. Stärk, H., Ganea, O., Pattanaik, L., Barzilay, R. & Jaakkola, T. EquiBind: geometric deep learning for drug binding structure prediction. In *Proc. 39th International Conference on Machine Learning* (eds Chaudhuri, K. et al.) 20503–20521 (PMLR, 2022).
145. Chen, M. et al. Multifaceted protein-protein interaction prediction based on Siamese residual RCNN. *Bioinformatics* **35**, i305–i314 (2019).
146. Wang, Y., Wang, J., Cao, Z. & Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.* **4**, 279–287 (2022).
147. Falk, T. et al. U-Net: deep learning for cell counting, detection, and morphometry. *Nat. Methods* **16**, 67–70 (2019).
148. Pachitariu, M. & Stringer, C. Cellpose 2.0: how to train your own model. *Nat. Methods* **19**, 1634–1641 (2022).
149. Kobayashi, H., Cheveralls, K. C., Leonetti, M. D. & Royer, L. A. Self-supervised deep learning encodes high-resolution features of protein subcellular localization. *Nat. Methods* **19**, 995–1003 (2022).
150. Palma, A., Theis, F. J. & Lotfollahi, M. Predicting cell morphological responses to perturbations using generative modeling. *Nat. Commun.* **16**, 505 (2025).

Acknowledgements We acknowledge the National Natural Science Foundation of China (grant nos. U24A20476 and 92259201); the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0533300); the National Key R&D Program of China (grant nos. 2022YFF0608403 and 2020YFE0202200); the National Natural Science

Foundation of China (32088101); the 'Pioneer' and 'Leading Goose' R&D Program of Zhejiang (grant no. 2023C03056); the State Key Laboratory of Medical Proteomics (SKLP-K202406); the Shanghai Municipal Science and Technology Major Project (2025SHZDZX026D07); the Zhejiang Provincial Natural Science Foundation of China (LMS26C050002 and LQ24C050002); and the National Natural Science Foundation of China (grant no. 32401239). Funding was provided by the State Key Laboratory of Medical Proteomics (SKLP-Y202403); the National Natural Science Foundation of China (12288101, 8206100646 and T2321001); the Clinical Medicine Plus X-Young Scholars Project, Peking University; the Fundamental Research Funds for the Central Universities (PKU2025PKULCXQ031); SNF Sinergia (CRSII5_189952); Novartis Forschungsstiftung (FN24-0000000612); the Desirée and Niels Yde Foundation (543-23); the National Natural Science Foundation of China (32470878, 32122032 and 31970750); the Zhejiang Provincial Natural Science Foundation (QKWL25H0901); the National Natural Science Foundation of China (grant no. 32470663); the Guangdong Pearl River Talents Program (grant no. 2019QN01Y183); the National Institutes of Health grant (R01HG012446); the Young Talents Program of Sun Yat-sen University Cancer Center (grant no. YTP-SYSUCC-0042); the Ministry of Science and Technology of the People's Republic of China (2024YFA0916903); the National Natural Science Foundation of China (32122042 and 32071208); the Zhejiang Provincial Natural Science Foundation (DQ24C050001); the National Natural Science Foundation of China (nos. 12371485, T2341007, T2350003 and 12131020); the Science and Technology Commission of Shanghai Municipality (no. 23JS1401300); the Zhejiang Province Vanguard Goose-Leading Initiative (no. 2025C01114); the National Science and Technology Major Project (grant no. 2022ZD0115004); The Luxembourg National Research Fund with the German Research Foundation (INTER/DFG/23/18289476/TRIUMPH); and the National Natural Science Foundation of China (grant no. 32270796). We thank the Proteomic Navigator of the Human Body (π-HuB) Project for support, and R. Aebersold and M. Mann for helpful discussions.

Author contributions T.G. conceived and supervised the study, provided overall conceptual guidance and critically revised the manuscript. L. Qian drafted the manuscript and performed comprehensive revisions of the text and figures. Z. Z., P.Z. and Z.G. contributed to the AI sections, including content development and revision. Z. Dong contributed to the spatial omics-related content. Xudong Zhang contributed to the active learning-related content. Z. Dai curated and organized the datasets. K.R.R., C.B., M.C., G. Liti, J.-X.Y., M.R., E.W., F.L., J.S., R.S., Z.X., H. Lou, K.L., G. Liu, B.A., Y. Cai, Yuping Chen, H.M., L.M.S., L.T. and Y.Y. contributed to the yeast biology-related content, including domain expertise, discussion and revision. S.S., J.L., H.J., T.W., Z.L., Y.D., Yi Chen, H.W., L.M., H. Lu, L.B., W.E., G.G., Luonan Chen, S.Z.L., Xiaofan Zhang, K.Y., J.Z., Y. Zheng and B.Z. contributed to AI algorithms and computational methodology, including technical input and revision. S.W., N.Z., M.Z., Yingrui Wang, X.W., L.X., Liangyi Chen, G.R., X.C., Yalin Wang, Q.X., H.S., K.P., Y. Zhu, F.H., L. Qiao, T.T. and J.Y. contributed to omics and imaging technologies, including data interpretation, technical development and manuscript revision. All authors reviewed and approved the final manuscript.

Competing interests T.G. and Y. Zhu are shareholders of Westlake Omics. T.T. is a shareholder of Wuhan Metware Biotechnology. The remaining authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Tiannan Guo.
Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Affiliated Hangzhou First People's Hospital, State Key Laboratory of Medical Proteomics, School of Medicine, School of Future Biomedicine, Westlake University, Hangzhou, China. [2]Westlake Center for Intelligent Proteomics, Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, China. [3]Research Center for Industries of the Future, School of Life Sciences, Westlake University, Hangzhou, China. [4]Center for Machine Learning Research, Peking University, Beijing, China. [5]AI for Science Institute, Beijing, China. [6]National Engineering Laboratory for Big Data Analysis and Applications, Beijing, China. [7]Center for Data Science, Peking University, Beijing, China. [8]Shanghai Artificial Intelligence Laboratory, Shanghai, China. [9]Research Institute of Intelligent Complex Systems, Fudan University, Shanghai, China. [10]Stanford Genome Technology Center, Stanford University, Palo Alto, CA, USA. [11]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [12]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland. [13]The Donnelly Centre, University of Toronto, Toronto, Ontario, Canada. [14]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [15]Institute of Research on Cancer and Ageing of Nice (IRCAN), Faculté de Médecine, Nice, France. [16]State Key Laboratory of Oncology in South China, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou, China. [17]Department of Biochemistry, Charité-Universitätsmedizin Berlin, Berlin, Germany. [18]Exploratory Diagnostic Sciences, Berlin Institute of Health at Charité, Berlin, Germany. [19]Center for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [20]Max Planck Institute for Molecular Genetics, Berlin, Germany. [21]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg. [22]Department of Biomedicine, University of Basel, Basel, Switzerland. [23]Department of Artificial Intelligence, School of Engineering, Westlake University, Hangzhou, China. [24]Biomedical Research Core Facilities, Westlake University, Hangzhou, China. [25]Institute of Biopharmaceutical and Health Engineering, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China. [26]Université de Strasbourg, CNRS, Inserm, IGBMC UMR 7104-UMR-S 1258, Illkirch, France. [27]State Key Laboratory of Microbial Metabolism and Joint International Research Laboratory of Metabolic and Developmental Sciences, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China. [28]South China Hospital, Health Science Center, Guangdong Key Laboratory of Genome Instability and Disease Prevention, Shenzhen University School of Medicine, Shenzhen, China. [29]State Key Laboratory of Medical Proteomics, National Center for Protein Sciences (Beijing), Research Unit of Proteomics Driven Cancer Precision Medicine (Chinese Academy of Medical Sciences), Beijing, China. [30]International Academy of Phronesis Medicine (Guangdong), Guangzhou, China. [31]DP Technology Co. Ltd., Beijing, China. [32]Beijing Advanced Center of RNA Biology (BEACON), Peking University, Beijing, China. [33]State Key Laboratory of Medical Proteomics, Beijing, China. [34]State Key Laboratory of Membrane Biology, Beijing Key Laboratory of Cardiometabolic Molecular Medicine, Institute of Molecular Medicine, Center for Life Sciences, College of Future Technology, Peking University, Beijing, China. [35]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, China. [36]Key Laboratory of Growth Regulation and Translational Research of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou, China. [37]Institute of Biology, Westlake Institute for Advanced Study, Hangzhou, China. [38]Bruker Switzerland AG, Faellanden, Switzerland. [39]Zhejiang Key Laboratory of Structural Biology, School of Life Sciences, Westlake University, Hangzhou, China. [40]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. [41]National Biomedical Imaging Center, College of Future Technology, Peking University, Beijing, China. [42]State Key Laboratory of Microbial Metabolism, Joint International Research Laboratory of Metabolic and Developmental Sciences, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China. [43]SJTU-Yale Joint Center of Biostatistics and Data Science, National Center for Translational Medicine, MoE Key Lab of Artificial Intelligence, AI Institute Shanghai Jiao Tong University, Shanghai, China. [44]Shanghai Engineering Research Center for Big Data in Pediatric Precision Medicine, NHC Key Laboratory of Medical Embryogenesis and Developmental Molecular Biology & Shanghai Key Laboratory of Embryo and Reproduction Engineering, Shanghai, China. [45]School of Life Sciences, Westlake University, Hangzhou, China. [46]Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Hangzhou, China. [47]Manchester Institute of Biotechnology, University of Manchester, Manchester, UK. [48]Generative and Synthetic Genomics, Wellcome Sanger Institute, Cambridge, UK. [49]Shenzhen Institute of Synthetic Biology, Chinese Academy of Sciences, Shenzhen, China. [50]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. [51]School of Mathematical Sciences, Peking University, Beijing, China. [52]State Key Laboratory of Gene Function and Modulation Research, Biomedical Pioneering Innovative Center (BIOPIC) and Beijing Advanced Innovation Center for Genomics (ICG), School of Life Sciences, Center for Bioinformatics (CBI), Peking University, Beijing, China. [53]School of Mathematical Sciences and School of AI, Shanghai Jiao Tong University, Shanghai, China. [54]AI Laboratory, Research Center for Industries of the Future, Westlake University, Hangzhou, China. [55]Biodesign Center, State Key Laboratory of Engineering Biology for Low-Carbon Manufacturing, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, China. [56]National Center of Technology Innovation for Synthetic Biology, Tianjin, China. [57]Department of Chemistry, Fudan University, Shanghai, China. [58]Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. [59]DZHK (German Centre for Cardiovascular Research), Heidelberg, Germany. [60]Stanford Genome Technology Center, Palo Alto, CA, USA. [61]Center for Interdisciplinary Studies, Westlake University, Hangzhou, China. [62]Wuhan Metware Biotechnology Co. Ltd., Wuhan, China. [63]Guangzhou National Laboratory, Guangzhou International Bio Island, Guangzhou, China. [64]State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences Beijing, Beijing Institute of Lifeomics, Beijing, China. [65]School of Pharmaceutical Sciences, Guangzhou Medical University, Guangzhou, China. [66]Autolab, Westlake University, Hangzhou, China. [67]School of Engineering, Research Center for Industries of the Future, Westlake University, Hangzhou, China. [68]Medical Artificial Intelligence Laboratory, Westlake University, Hangzhou, China. [69]Department of Electronic Engineering, Tsinghua University, Beijing, China. ✉e-mail: guotiannan@westlake.edu.cn