OPEN ACCESS Freely available online

Fungal Genomics & Biology

ISSN: 2165-8056

Commentary

Synthetic Mycological Data Sets for AI-Driven Disease Monitoring

Eriko Bender*

Department of Life Sciences, University of Cartagena, Cartagena, Colombia

DESCRIPTION

Artificially generated mycology data-commonly referred to as synthetic mycological datasets-offer a transformative new avenue for training AI systems to detect, classify, and forecast fungal diseases. These datasets replicate real-world fungal morphology, growth stages, and environmental dynamics in a controlled, scalable, and privacy-safe manner. They underpin diagnostic tools, environmental surveillance systems, and outbreak prediction platforms-especially for rare pathogens, early growth phases, or biosafety-level organisms. Examples such as time-aligned growth simulations and vision-language paired corpora illustrate how synthetic datasets support deep learning, zero-shot inference, and morphological analysis without requiring patient-derived or hazardous materials.

Synthetic collections simulate the progression of fungal growth in spatially and temporally coherent image frames. Beginning with dormant spores, transitioning through germ tubes, branching hyphae, and culminating in mature mycelial networks, these sequences maintain structural alignment across frames. This enables deep learning models-such as convolutional neural networks or vision transformers-to effectively learn temporal morphology and dynamic patterns. To facilitate semantic reasoning, synthetic datasets sometimes pair generated images with textual descriptions like "branched septate hyphae at 48h". This alignment supports vision-language models to interpret morphology from text prompts-a powerful tool for zero-shot inference on novel species or growth conditions without real image examples. While less established in fungal research, synthetic simulation of tabular features (e.g., spore diameter, radial growth rate, gene expression under stress) enables structured model training for outbreak risk, resistance prediction, or phenotype classification. Generative models-such as GAN-inspired or transformer-based simulators-can produce these datasets while preserving statistical realism.

Using time-aligned synthetic sequences, AI systems can recognize key fungal developmental milestones-such as early germination or branching architecture-important for diagnosing active infection or environmental colonization in microscopy contexts. By training with descriptive prompts and synthetic images, AI can interpret unseen fungal morphologies described textually, increasing robustness and interpretability-especially valuable when encountering unusual growth forms or previously unseen species. Synthetic datasets can emulate rare pathogens and drug-resistant strains, such as *Candida auris* or *Cryptococcus* variants, offering models exposure to morphologies that seldom appear in real datasets but are crucial for early outbreak detection. AI models can be stress-tested using synthetic outlier scenarios-for example, accelerated sporulation under heat stress or high-spore environmental surges-supporting proactive detection tool development and system resilience planning.

Synthetic imagery may omit real-world micro-artifacts like staining variability or imaging noise. Strategies such as domain adaptation, mixed real-synthetic fine-tuning, and adversarial training help close this gap. Including structured metadata-for example, incubation time, substrate type, temperature, fungal species-is crucial to support robust model generalization across sampling conditions. Synthetic image-text resources are maturing, but synthetic generation of transcriptomic, proteomic, or resistance phenotype data remains a frontier requiring further development. Excessive reliance on synthetic features can reduce real-world validity. Ensuring model validation on held-out real datasets and mixing in real samples is essential for generalizability.

One effective strategy is pretraining AI models on large, synthetic datasets, followed by fine-tuning on a small set of annotated real-world data. This approach accelerates learning, improves generalization, and reduces dependence on costly real data. Environmental surveillance tools-such as air or soil microscopy systems-can then leverage pretrained models directly, augmented by occasional real-world calibration.

Simulating gene expression under temperature stress, antifungal exposure, or host-pathogen interactions to support phenotype prediction and virulence modeling. Pair fungal image simulations with synthetic environmental metadata-like humidity or particulate levels-to build predictive ecological risk models. Enhancing realism by emulating microscopic stain patterns, illumination variability, or sensor noise found in real clinical labs. Establish community platforms where models trained on synthetic data are evaluated on standardized real test

Correspondence to: Eriko Bender, Department of Life Sciences, University of Cartagena, Cartagena, Colombia, E-mail: erikobender123@gmail.com

Received: 03-Mar-2025, Manuscript No. FGB-25-38376; **Editor assigned:** 05-Mar-2025, PreQC No. FGB-25-38376 (PQ); **Reviewed:** 19-Mar-2025, QC No. FGB-25-38376; **Revised:** 26-Mar-2025, Manuscript No. FGB-25-38376 (R); **Published:** 02-Apr-2025, DOI: 10.35248/2165-8056.25.15.274

Citation: Bender E (2025). Synthetic Mycological Data Sets for AI-Driven Disease Monitoring. Fung Genom Biol. 15:274.

sets, promoting best practices and setting performance expectations.

CONCLUSION

Synthetic mycological datasets-ranging from time-aligned fungal growth sequences to vision-language paired corpora-offer a vital new paradigm in AI-based fungal disease monitoring. They provide scalable, privacy-safe, and richly annotated alternatives to scarce clinical or environmental data. By supporting morphology-based classification, zero-shot inference, rare pathogen simulation, and outbreak-aware modeling, these datasets lay a strong foundation for future diagnostic and surveillance systems. With improvements in metadata richness, integration of omics-level data, and hybrid training pipelines, synthetic data approaches promise to boost early detection, broaden pathogen coverage, and enhance AI interpretability in medical, ecological, and One Health domains. In an age of emerging fungal threats and global health vulnerabilities, synthetic data innovation is both timely and transformative.