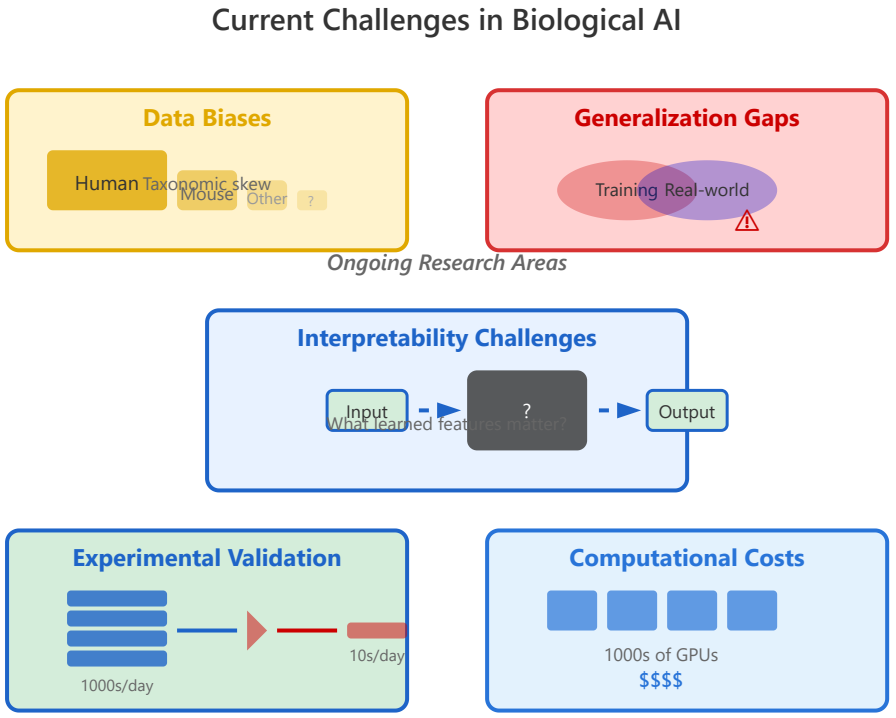
Limitations



Data biases
Taxonomic & functional skew

Generalization gaps
Out-of-distribution failures

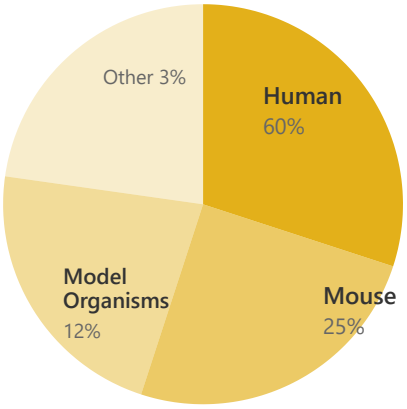
Interpretability challenges
Black box models

Experimental validation
Lab throughput bottleneck

Computational costs
Training & inference expense

1. Data Biases

Taxonomic Distribution in Biological Databases



Underrepresented:

- Non-model organisms
- Rare diseases
- Non-coding regions

The Problem

Biological AI models are trained on highly imbalanced datasets that overrepresent certain species, tissues, and biological processes while underrepresenting others. This creates systematic biases in model predictions.

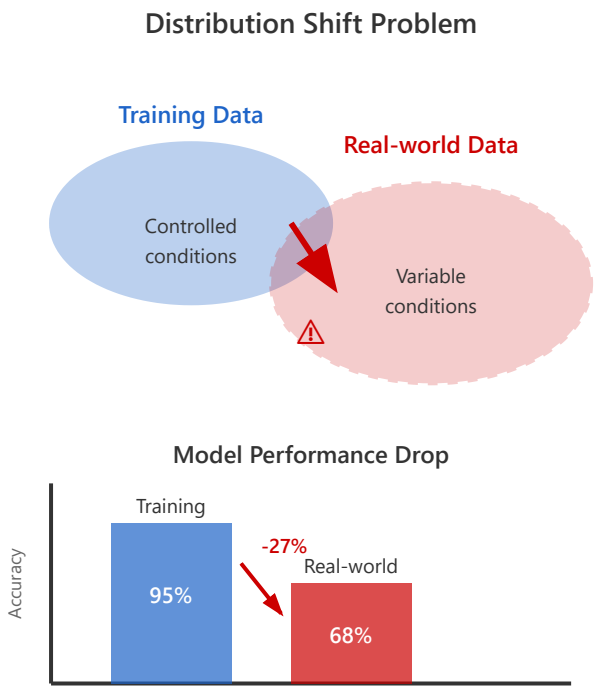
Key Issues

- **Taxonomic skew:** Human and mouse data dominate, while other organisms are severely underrepresented
- **Functional bias:** Well-studied pathways and genes receive more attention than novel or rare functions
- **Tissue bias:** Easily accessible tissues are overrepresented in training data
- **Disease bias:** Common diseases studied more than rare diseases

Example Impact

A protein function prediction model trained primarily on human proteins may fail to accurately predict functions in non-model organisms like extremophiles or plant species, limiting its utility for agricultural or environmental applications.

2. Generalization Gaps



The Problem

Models trained on carefully curated datasets often fail when confronted with real-world data that differs from training conditions. This out-of-distribution (OOD) problem leads to unreliable predictions in practical applications.

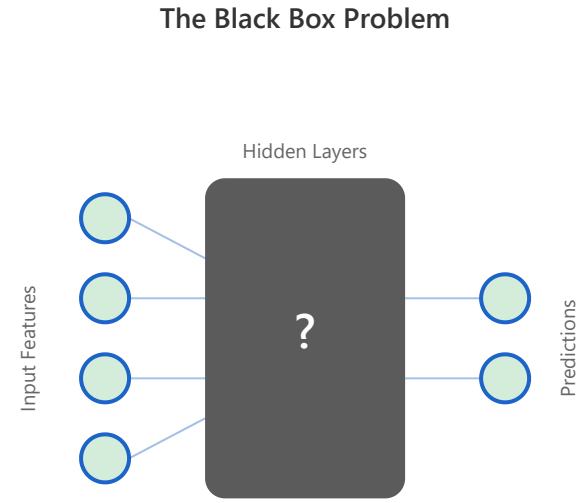
Key Issues

- **Batch effects:** Different experimental protocols and platforms create systematic variations
- **Population diversity:** Models trained on one population may not generalize to others
- **Environmental factors:** Lab conditions differ from natural environments
- **Temporal shifts:** Biological systems evolve and change over time

Example Impact

A drug response prediction model trained on cell lines from European populations may show significantly reduced accuracy when applied to patients of African or Asian ancestry due to genetic and environmental differences.

3. Interpretability Challenges



Key Questions:

- ❓ Which features does the model rely on?
- ❓ Why did it make this specific prediction?
- ❓ Are the learned patterns biologically meaningful?
- ❓ Can we trust it for critical decisions?

The Problem

Deep learning models, while powerful, function as "black boxes" where the relationship between inputs and outputs is opaque. This lack of transparency poses serious challenges for biological applications requiring mechanistic understanding.

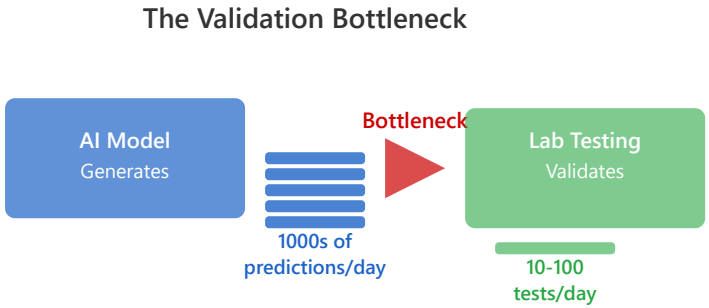
Key Issues

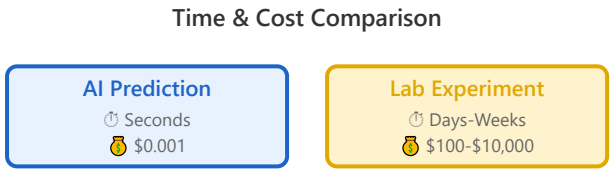
- **Feature attribution:** Difficulty identifying which input features drive predictions
- **Biological plausibility:** Model patterns may not reflect known biological mechanisms
- **Trust and adoption:** Clinicians and researchers hesitant to use unexplainable models
- **Debugging:** Hard to identify and fix systematic errors in model reasoning

Example Impact

A deep learning model predicts a protein will bind to a specific drug target with high confidence, but cannot explain which structural features matter most. Researchers cannot determine if the prediction is based on relevant biochemistry or spurious correlations.

4. Experimental Validation





Consequences:

- Only fraction of predictions can be tested
- Delayed feedback for model improvement
- Risk of deploying unvalidated predictions
- Selection bias in which predictions to test

The Problem

AI models can generate predictions orders of magnitude faster than they can be experimentally validated. This creates a severe bottleneck that limits the practical utility of computational predictions and slows the research cycle.

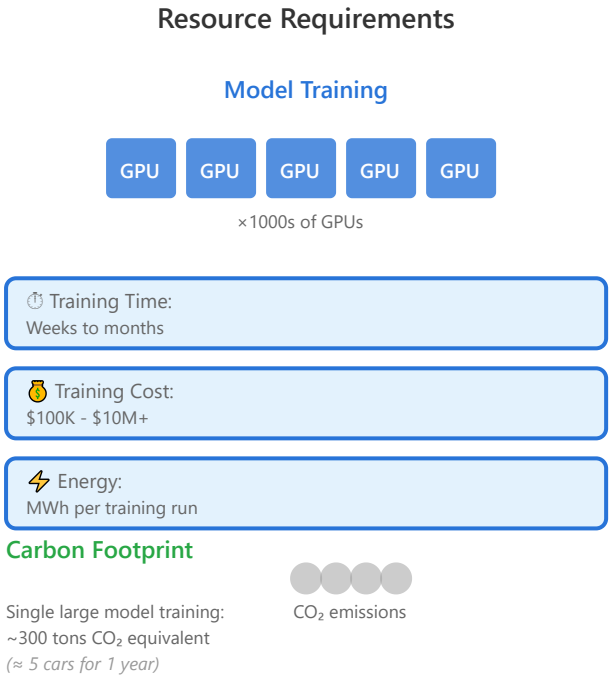
Key Issues

- **Throughput mismatch:** Models predict thousands of hypotheses daily while labs test tens
- **Resource constraints:** Limited lab space, equipment, and trained personnel
- **Time delays:** Experiments take days to months while predictions are instant
- **Cost barriers:** Each validation experiment costs significantly more than predictions

Example Impact

A protein engineering model suggests 10,000 potentially beneficial mutations. The lab can only test 50 mutations per month due to time and cost constraints. It would take over 16 years to validate all predictions, by which time the model may be outdated.

5. Computational Costs



The Problem

Training and deploying state-of-the-art biological AI models requires massive computational resources, creating barriers to entry and raising concerns about sustainability and accessibility of these technologies.

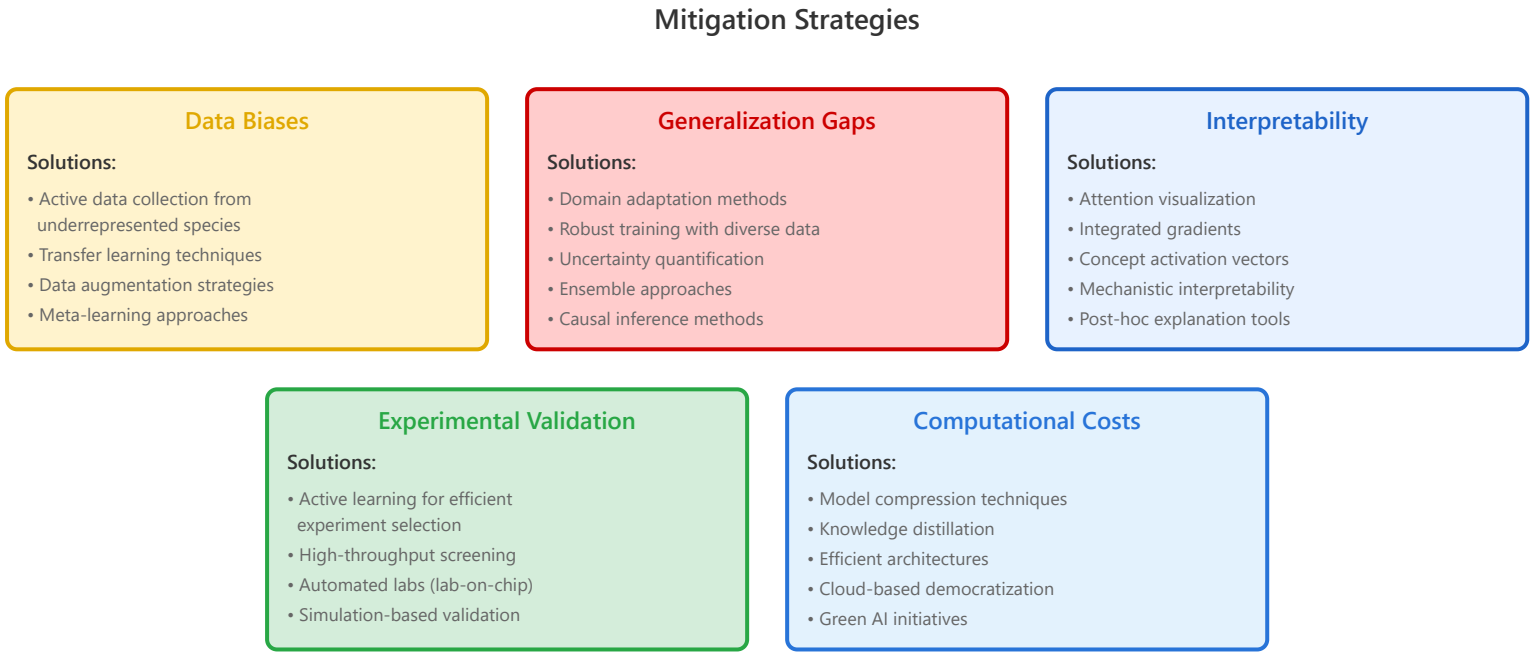
Key Issues

- **Hardware requirements:** Need for thousands of specialized GPUs or TPUs
- **Financial barriers:** Training costs can exceed millions of dollars per model
- **Energy consumption:** Significant electricity use with environmental impact
- **Accessibility:** Only well-funded institutions can develop large models
- **Inference costs:** Running predictions at scale remains expensive

Example Impact

Training a large protein language model like ESM-2 (650M parameters) requires thousands of GPU-hours and costs approximately $200,000. This puts such models out of reach for most academic labs and smaller biotech companies, concentrating power in well-funded institutions.

Addressing the Limitations: Current Approaches



Key Takeaways

While biological AI faces significant limitations, active research is addressing these challenges through methodological innovations, improved experimental design, and technological advances. Success requires interdisciplinary collaboration between computational scientists, experimentalists, and domain experts.

The field is moving toward more robust, interpretable, and accessible AI systems that can reliably contribute to biological discovery and clinical applications.