# Contents

# 1 Identifying and Coping with Confounders: A Practical Guide

Our simulations demonstrate something fundamental: *confounding is not a bug of statistics*—it's a property of the world. Biology, social systems, and ML pipelines all happily generate correlations that look causal until we interrogate the data-generating process.

> **See it in action:** Run `examples/confounding/02_confounding_simulations.py` to explore various confounding scenarios with interactive visualizations.

This tutorial addresses three core questions about confounding, using the same causal language our simulation codebase speaks. We also tie these concepts back to the biological examples covered elsewhere in this project.

---

## 1.1 Table of Contents

---

## 1.2 1. Identifying Potential Confounders

### 1.2.1 State of the Art Thinking

There is no single magic algorithm that "discovers all confounders." Anyone claiming that is selling philosophical snake oil. Instead, SoA methods fall into three complementary mindsets.

**First: causal structure discovery (hypothesis generation, not truth machines).**

These methods try to infer a *graph* from conditional independencies in the data.

Examples:

- PC / FCI algorithms
- NOTEARS and continuous DAG learning
- Additive Noise Models (ANM)

What they do well:

- Suggest *candidate* confounders
- Reveal surprising dependencies
- Generate testable causal hypotheses

What they cannot do:

- Distinguish latent confounders from direct causation without assumptions
- Replace biological or domain knowledge

In our cell-cycle demo, these methods might detect that MYC and ribosomal genes share a parent, but they won't tell us "this parent is the cell cycle" unless that variable is measured.

**Second: representation learning for nuisance variation (very common in genomics).**

Here the idea is: "If I can't name the confounder, maybe I can *absorb* it."

Examples:

- PCA / latent factor models
- scVI, PEER, MOFA
- Autoencoders with nuisance heads

This is exactly how batch effects and cell cycle are often handled in scRNA-seq. You don't always label "batch" explicitly; you infer a latent variable that explains global shifts.

Caution: These methods remove *variation*, not necessarily *confounding*. If the latent factor also carries causal signal, you may overcorrect.

**Third: perturbation-aware designs (the gold standard mindset).**

This isn't an algorithm—it's a philosophy:

- Randomize treatments
- Use CRISPR, drug perturbations, or natural experiments
- Ensure the treatment assignment is independent of confounders

In causal terms: make `T    Z` by design.

Our simulations implicitly show this: once treatment is randomized, many confounders stop being confounders.

**Key takeaway:** Confounder identification is not solved by statistics alone. It's a dialogue between data, assumptions, and scientific context.

---

## 1.3 2. Blocking Confounders to Estimate True Effects

Blocking is a precise causal idea: we want to stop information from flowing along the backdoor path.

In our codebase, the structure is typically:

```
Z → C
Z → X
C → X   (the effect we care about)
```

Blocking means conditioning on something that breaks the `Z → C → X` shortcut.

There are three canonical ways to do this.

**Conditioning directly (regression adjustment).**

You include `Z` as a covariate:

- Linear regression
- Generalized linear models

- Outcome modeling

This works when:

- Z is observed
- The functional form is not wildly misspecified

This is what our job-training example demonstrates: conditioning on age recovers the treatment effect.

**Balancing the treatment assignment (propensity methods).**

Instead of modeling `X | C, Z`, you model `C | Z`.

- Propensity score matching
- Inverse probability weighting
- Stratification

This reframes the question: "Among people who *could plausibly* receive treatment, what happens if they do?"

It's especially useful when:

- Treatment assignment is biased
- Outcome models are complex or noisy

**Doubly robust methods (belt and suspenders).**

These combine both ideas:

- Model the treatment
- Model the outcome

If either model is correct, the estimate is consistent.

This is why doubly robust estimators are beloved in practice—they're forgiving in messy real-world data.

**Biological intuition:**

In our HIF1A example, "blocking disease severity" means conditioning on severity or its proxies (hypoxia markers, inflammation scores), not the stress-response gene itself. Conditioning on the wrong variable opens *collider bias* instead of blocking confounding.

---

## 1.4 3. Estimating Causal Effects Without Full Confounder Control

What if you can't measure or control for all confounders? This is the reality in most biological studies. Fortunately, there are methods that allow us to estimate causal effects even when full confounder control isn't possible—though they require different assumptions. Causality is like conservation of energy: ignorance must be paid for with assumptions.

There are three main approaches:

**Instrumental variables (IVs).**

You find a variable that:

- Affects treatment
- Does not directly affect the outcome
- Is independent of confounders

Classic, powerful, and extremely hard to justify in biology.

**Front-door adjustment.**

You observe a mediator that fully transmits the treatment effect and is not confounded with the outcome.

Rare, but beautiful when it applies.

**Sensitivity analysis and partial identification.**

Instead of pretending you know the truth, you say:

- "If unmeasured confounding is at most this strong…"
- "The true effect lies within this interval…"

This is often the most honest answer in biology and medicine.

**Modern perspective:**

We increasingly aim for *causal robustness*, not point identification:

- Do effects persist across datasets?
- Across perturbations?
- Across environments?

This mindset aligns with cross-context robustness thinking (as explored in our related MetaSpliceAI project): robustness across contexts is itself causal evidence.

---

## 1.5 4. The Unifying Mental Model

Confounding is not an error—it's a shadow cast by missing context.

Our simulations show four archetypes:

- Cell cycle: latent biological state
- Batch effects: technical environment
- Disease severity: upstream pathological driver
- Selection bias: human decision processes

All four share the same DAG skeleton. Once you see that, causal inference stops being a bag of tricks and starts feeling like structural reasoning.

---

## 1.6 5. Methods Roadmap

The following methods and topics are mentioned in this document and represent potential additions to the causal-bio-lab project:

### 1.6.1 Causal Structure Discovery

| Method | Description | Priority |
|---|---|---|
| PC / FCI algorithms | Constraint-based causal discovery from conditional independencies | Medium |
| NOTEARS | Continuous optimization for DAG learning | Medium |
| Additive Noise Models (ANM) | Exploits asymmetry in noise distributions | Low |

### 1.6.2 Representation Learning for Confounders

| Method | Description | Priority |
|---|---|---|
| scVI | Variational inference for scRNA-seq with batch correction | High |
| PEER | Probabilistic estimation of expression residuals | Medium |
| MOFA | Multi-omics factor analysis | Medium |

### 1.6.3 Treatment Effect Estimation

| Method | Description | Priority |
|---|---|---|
| Propensity score matching | Balance treatment groups on observed covariates | High |
| Inverse probability weighting (IPW) | Reweight samples by treatment probability | High |
| Doubly robust estimators | Combine outcome and propensity models | High |

### 1.6.4 Advanced Causal Inference

| Method | Description | Priority |
|---|---|---|
| Instrumental variables | Exploit exogenous variation for identification | Medium |
| Front-door adjustment | Identify effects through mediators | Low |
| Sensitivity analysis | Quantify robustness to unmeasured confounding | High |

### 1.6.5 Future Topics

- When *adjusting* introduces bias (colliders)
- Why representation learning can both help and hurt causality
- How causal thinking dovetails with foundation models and perturbation data in biology

---

## 1.7 Key Takeaway

> **The universe generates correlations for free. Causes are expensive. That's why they're worth the trouble.**