Contents

1 Confounding in Computational Biology: Examples and Mental Models

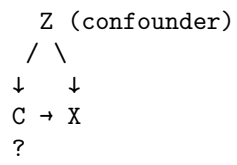
Confounding is the central obstacle to causal inference from observational data. This document provides concrete biological examples to build intuition for recognizing and reasoning about confounders.

> **Key insight:** A confounder is a variable that influences both the treatment and the outcome, creating a spurious association that looks causal but isn't.

1.1 Table of Contents

1.2 The Confounding Pattern

The classic confounding structure looks like this:



Where: - Z influences both C (treatment/exposure) and X (outcome) - We observe a correlation between C and X - But the correlation is (partly or entirely) due to Z, not a direct causal effect

The danger: If we ignore Z, we conclude that C causes X when it may not.

1.3 Example 1: Cell Cycle and Gene Expression

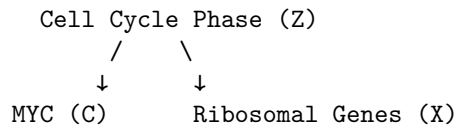
1.3.1 The Setup

- C: MYC expression level
- X: Ribosomal gene expression
- Z: Cell cycle phase (G1, S, G2/M)

1.3.2 What We Observe

Cells with high MYC tend to have high ribosomal gene expression.

1.3.3 The Confounding Story



Cell cycle phase drives both: - **S/G2 phase → high MYC** (MYC is a cell cycle regulator) - **S/G2 phase → high ribosomal genes** (cells preparing to divide need ribosomes)

1.3.4 The Causal Question

> "If I knock down MYC, will ribosomal genes decrease?"

The observational correlation says "yes, strongly." But if cell cycle is the true driver, the intervention effect may be much weaker than expected.

1.3.5 Why This Matters for Drug Discovery

If you're targeting MYC to reduce ribosomal biogenesis (e.g., in cancer), you need to know the *interventional* effect, not the confounded observational correlation.

1.4 Example 2: Batch Effects in scRNA-seq

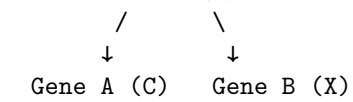
1.4.1 The Setup

- C: Expression of gene A
- X: Expression of gene B
- Z: Sequencing batch (day, lane, reagent lot)

1.4.2 What We Observe

Gene A and gene B are highly correlated across cells.

1.4.3 The Confounding Story



Batch effects create systematic variation: - **Batch 1:** Higher capture efficiency → both genes appear higher - **Batch 2:** Lower capture efficiency → both genes appear lower

1.4.4 The Causal Question

> "Does gene A regulate gene B?"

The batch-driven correlation tells you nothing about regulation. Two completely unrelated genes can appear correlated if batch effects are strong enough.

1.4.5 Why This Matters

This is why Perturb-seq is so valuable—the perturbation is randomized *within* batches, breaking the confounding.

1.5 Example 3: Cell Type Composition

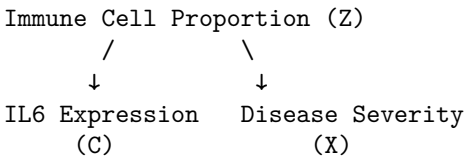
1.5.1 The Setup

- C: Expression of inflammatory gene (e.g., IL6)
- X: Disease severity score
- Z: Proportion of immune cells in sample

1.5.2 What We Observe

Samples with high IL6 expression have worse disease outcomes.

1.5.3 The Confounding Story



More immune cells means: - **Higher IL6** (immune cells express IL6) - **Worse disease** (inflammation drives pathology)

1.5.4 The Causal Question

> "If I block IL6, will disease improve?"

The correlation is real, but the causal effect of IL6 *per se* may be smaller than the correlation suggests. The immune infiltrate is doing many things beyond IL6.

1.5.5 Why This Matters

This is a classic problem in bulk RNA-seq from tissues. Deconvolution methods try to address this, but they don't eliminate the confounding—they just make it visible.

1.6 Example 4: Donor Effects in Human Samples

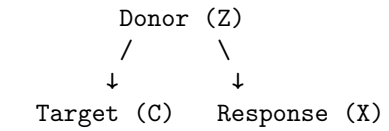
1.6.1 The Setup

- C: Expression of drug target gene
- X: Drug response phenotype
- Z: Donor identity (genetics, age, sex, lifestyle)

1.6.2 What We Observe

Samples with high target expression respond better to the drug.

1.6.3 The Confounding Story



Donor-level factors affect both: - **Genetics → target expression** (eQTLs, regulatory variants) - **Genetics → drug metabolism** (pharmacogenomics)

1.6.4 The Causal Question

> "If I increase target expression, will response improve?"

The donor-driven correlation may reflect shared genetic architecture, not a causal relationship between expression and response.

1.6.5 Why This Matters

This is why patient stratification is hard. Biomarkers that correlate with response may not be causal—they may just be proxies for underlying patient heterogeneity.

1.7 Example 5: Disease Severity

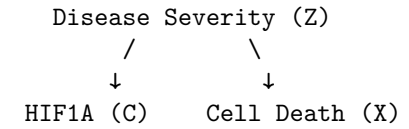
1.7.1 The Setup

- C: Expression of stress response gene (e.g., HIF1A)
- X: Cell death rate
- Z: Overall disease severity / tissue damage

1.7.2 What We Observe

Cells with high HIF1A have higher death rates.

1.7.3 The Confounding Story



Severe disease causes: - **Hypoxia → HIF1A activation** (stress response) - **Tissue damage → cell death** (pathology)

1.7.4 The Causal Question

> "Does HIF1A activation cause cell death?"

HIF1A might actually be *protective*—cells activating HIF1A are trying to survive. The positive correlation with death reflects the severity of the insult, not a causal effect of HIF1A.

1.7.5 Why This Matters

This is a common trap: stress response genes correlate with bad outcomes because they're markers of stress, not causes of harm. Inhibiting them might make things worse.

1.8 Example 6: Tissue Microenvironment

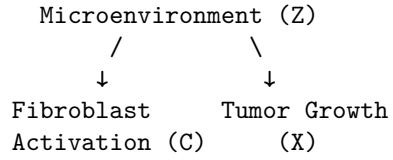
1.8.1 The Setup

- C: Fibroblast activation marker (e.g., SMA)
- X: Tumor growth rate
- Z: Tumor microenvironment state (hypoxia, inflammation, ECM stiffness)

1.8.2 What We Observe

Tumors with activated fibroblasts grow faster.

1.8.3 The Confounding Story



The microenvironment drives both: - **Hypoxia/inflammation → fibroblast activation** - **Hypoxia/inflammation → tumor aggressiveness**

1.8.4 The Causal Question

> "If I inhibit fibroblast activation, will tumors grow slower?"

The fibroblasts might be a bystander effect of the microenvironment, not a driver of growth. Or they might even be tumor-suppressive in some contexts.

1.8.5 Why This Matters

Cancer-associated fibroblasts (CAFs) have been notoriously difficult to target therapeutically, partly because the observational correlations don't reflect simple causal relationships.

1.9 Recognizing Confounders

1.9.1 Questions to Ask

When you see a correlation between C and X, ask:

1. **What could cause both?**
   - Upstream biological processes
   - Technical factors (batch, quality)
   - Patient/donor characteristics
2. **Is the "treatment" actually randomized?**
   - If not, something determined its value
   - That something might also affect the outcome
3. **Would the correlation survive an intervention?**
   - Imagine forcing C to a value
   - Would X still change?

1.9.2 Red Flags

- **Strong correlations without known mechanism**
- **Correlations that vary across datasets**
- **Correlations that disappear after controlling for obvious factors**

- **"Biomarkers" that don't replicate as drug targets**

1.10 What To Do About It

1.10.1 1. Measure and Adjust

If you can measure the confounder, you can (sometimes) adjust for it: - Include Z as a covariate - Stratify by Z - Use propensity score methods

Limitation: Only works for *measured* confounders.

1.10.2 2. Use Perturbation Data

The gold standard: directly intervene on C and measure X. - Perturb-seq - CRISPR screens - Drug treatments

Limitation: Expensive, not always feasible.

1.10.3 3. Find Natural Experiments

Look for situations where C varies for reasons unrelated to X: - Genetic variants (Mendelian randomization) - Batch effects that happen to create useful variation - Time-lagged relationships

Limitation: Requires careful thinking about what's actually randomized.

1.10.4 4. Be Honest About Uncertainty

When you can't eliminate confounding: - Report correlations as correlations - Acknowledge alternative explanations - Design follow-up experiments to test causality

---

1.11 Non-Biological Examples

The confounding pattern appears everywhere—not just in biology. Here are examples from everyday life and astrophysics that illustrate the same fundamental challenge.

1.11.1 Everyday Examples

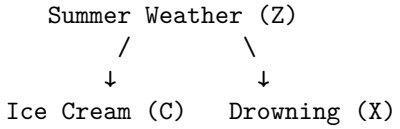
1.11.1.1 Example 7: Ice Cream and Drowning

1.11.1.1.1 The Setup

- C: Ice cream sales
- X: Drowning deaths
- Z: Summer weather (temperature)

1.11.1.1.2 What We Observe Months with high ice cream sales have more drowning deaths.

1.11.1.1.3 The Confounding Story



Hot weather causes: - **More ice cream consumption** (people want cold treats) - **More swimming → more drowning** (people go to pools and beaches)

1.11.1.1.4 The Causal Question

"If we ban ice cream, will drowning deaths decrease?"

Obviously not. Ice cream has no causal effect on drowning—they're both effects of a common cause.

1.11.1.1.5 Why This Matters This is the classic textbook example of confounding. It's absurd when stated explicitly, but similar reasoning errors happen constantly in more complex domains where the relationships aren't as obvious.

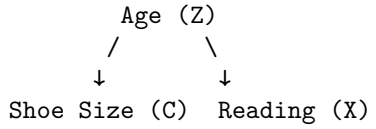
1.11.1.2 Example 8: Shoe Size and Reading Ability

1.11.1.2.1 The Setup

- C: Shoe size
- X: Reading test scores
- Z: Age

1.11.1.2.2 What We Observe Children with larger shoe sizes score higher on reading tests.

1.11.1.2.3 The Confounding Story



Age drives both: - **Older children have larger feet** (growth) - **Older children read better** (education and development)

1.11.1.2.4 The Causal Question

"If we give children bigger shoes, will they read better?"

The correlation is perfect but the causal claim is nonsense. This example shows how strong correlations can be completely non-causal.

1.11.1.2.5 Why This Matters In machine learning, shoe size would be a great *predictor* of reading ability. But for intervention design, it's useless. This distinction between prediction and causation is fundamental.

1.11.2 Astrophysics and Astronomy Examples

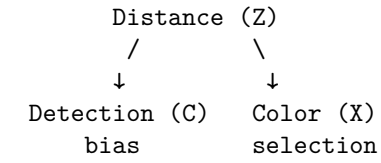
1.11.2.1 Example 9: Galaxy Color and Distance (Malmquist Bias)

1.11.2.1.1 The Setup

- C: Galaxy color (blue vs. red)
- X: Apparent brightness
- Z: Distance from Earth

1.11.2.1.2 What We Observe In flux-limited surveys, distant galaxies appear systematically bluer than nearby galaxies.

1.11.2.1.3 The Confounding Story



Distance creates selection effects: - **Distant galaxies must be intrinsically bright to be detected** (Malmquist bias) - **Intrinsically bright galaxies tend to be blue** (young, star-forming) - **Red galaxies at large distances fall below detection threshold**

1.11.2.1.4 The Causal Question

"Does the universe have more blue galaxies at high redshift?"

The observed color-distance correlation is partly real (cosmic evolution) but partly a selection artifact. Disentangling these requires careful modeling of the selection function.

1.11.2.1.5 Why This Matters This is a classic example of **selection bias**, which is a form of confounding. The "treatment" (being in our sample) is confounded with the outcome (observed properties). Volume-limited samples and careful selection modeling are the astronomical equivalent of randomization.

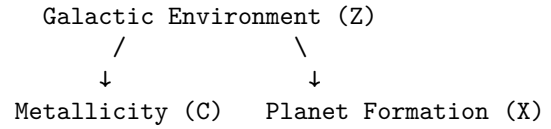
1.11.2.2 Example 10: Stellar Metallicity and Exoplanet Occurrence

1.11.2.2.1 The Setup

- C: Host star metallicity (iron abundance)
- X: Probability of hosting a giant planet
- Z: Galactic environment / stellar age

1.11.2.2.2 What We Observe Metal-rich stars are more likely to host giant planets.

1.11.2.2.3 The Confounding Story



Galactic environment affects both: - **Inner galaxy → higher metallicity** (more enrichment from supernovae) - **Inner galaxy → different planet formation conditions** (disk properties, dynamics)

1.11.2.2.4 The Causal Question

"Does metallicity directly cause giant planet formation?"

The correlation is well-established, but the causal mechanism is debated: - **Core accretion theory:** Higher metallicity → more solid material → easier to form planet cores (direct causation) - **Galactic archaeology:** Metallicity is a proxy for formation environment (confounding)

1.11.2.2.5 Why This Matters This is an active research question in exoplanet science. The answer matters for understanding planet formation physics and for predicting where to search for planets. Different causal models make different predictions for planet occurrence around stars in different galactic environments.

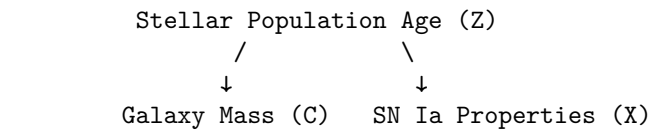
1.11.2.3 Example 11: Supernova Brightness and Host Galaxy Mass

1.11.2.3.1 The Setup

- C: Host galaxy stellar mass
- X: Type Ia supernova brightness (after standardization)
- Z: Progenitor stellar population age

1.11.2.3.2 What We Observe Type Ia supernovae in massive galaxies appear slightly fainter than those in low-mass galaxies, even after standard corrections.

1.11.2.3.3 The Confounding Story



Stellar age confounds the relationship: - **Massive galaxies have older stellar populations** (earlier formation, less ongoing star formation) - **Older progenitors may produce systematically different SNe Ia** (different white dwarf masses, explosion physics)

1.11.2.3.4 The Causal Question

"Does galaxy mass directly affect supernova brightness?"

This matters for cosmology: Type Ia supernovae are "standard candles" used to measure cosmic distances and dark energy. If the mass-brightness correlation is causal, we need to correct for it. If it's confounded by stellar age, we need to correct for age instead.

1.11.2.3.5 Why This Matters Getting this wrong biases measurements of the Hubble constant and dark energy equation of state. The "Hubble tension" (disagreement between different measurements of cosmic expansion) might partly stem from uncontrolled confounding in supernova standardization.

1.12 Summary Table

Example	Confounder Z	Treatment C	Outcome X	Trap
Cell cycle	Cell cycle phase	MYC	Ribosomal genes	Correlation regulation
Batch effects	Sequencing batch	Gene A	Gene B	Technical artifact
Cell type	Immune proportion	IL6	Disease severity	Composition effect
Donor effects	Donor genetics	Target expression	Drug response	Shared genetic basis

Example	Confounder Z	Treatment C	Outcome X	Trap
Disease severity	Tissue damage	HIF1A	Cell death	Marker vs. cause
Microenvironment	TME state	Fibroblast activation	Tumor growth	Bystander effect
Ice cream	Summer weather	Ice cream sales	Drowning deaths	Common cause
Shoe size	Age	Shoe size	Reading ability	Developmental confounder
Galaxy color	Distance	Detection	Color selection	Selection bias
Exoplanets	Galactic environment	Metallicity	Planet occurrence	Environment proxy
Supernovae	Stellar age	Galaxy mass	SN brightness	Cosmological bias

---

1.13 Key Takeaway

Confounding is the rule, not the exception, in biological data.

Every observational correlation you see should be treated as potentially confounded until you have evidence otherwise. The examples above are not edge cases—they are the default situation in computational biology.

This is why perturbation data is so valuable, and why causal inference methods exist: to help us reason carefully about what we can and cannot conclude from the data we have.

1.14 References

- Hernán & Robins, *Causal Inference: What If* — Chapter 7 on confounding
- Pearl, *Causality* — Formal treatment of confounding and adjustment
- Regev et al., "The Human Cell Atlas" — Discusses batch effects and technical confounders
- Rosenbaum, *Observational Studies* — Methods for dealing with confounding