Before a golem is born…

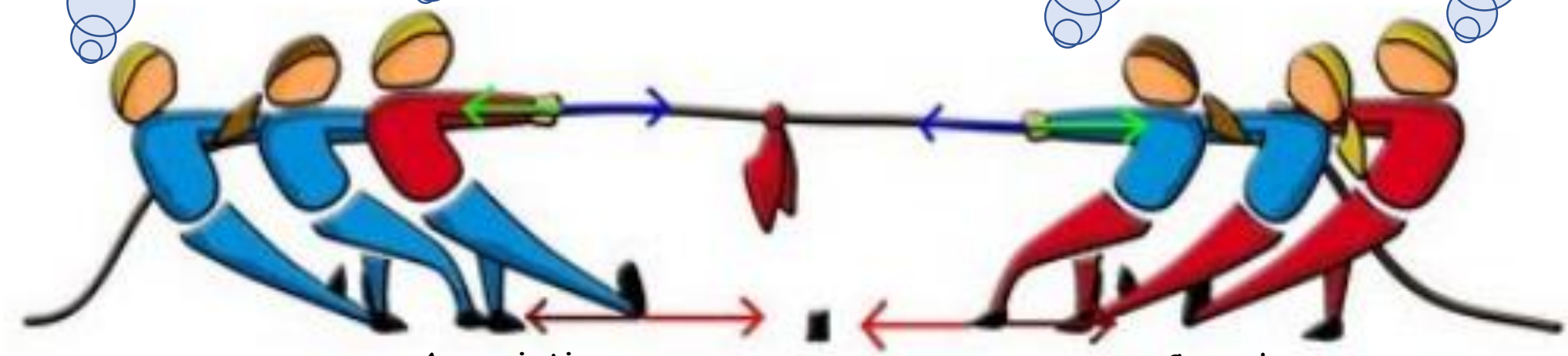
Methods to distinguish between causal and association = Directed Acyclic Graph

Its validity relies on information outside the data

Successful prediction does not require correct causal identification

The point is to produce understanding. We need more than just prediction……we need causal understanding

It is needed to design the collection of data and the structure of our statistical model



Association

Causal

Can we use LAI for as a selection trait for soybean seed protein concentrations?

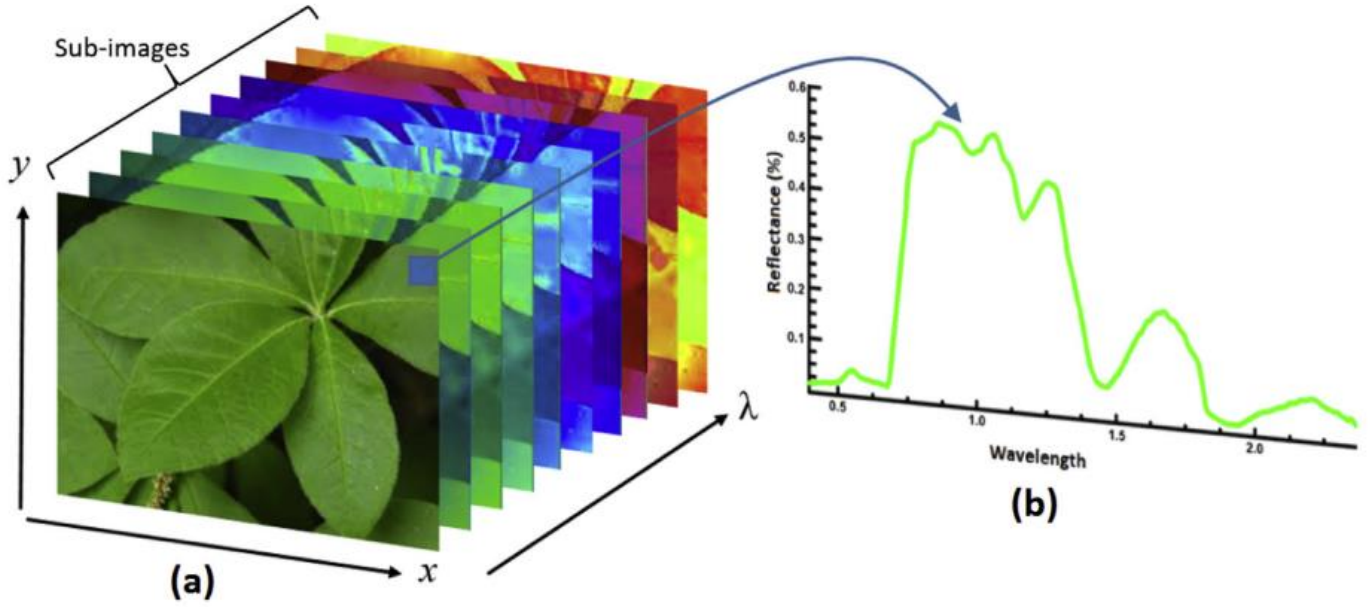
Is there a relationship between LAI and Protein Concentration?

Is that relationship causal?

Why I am interested in the LAI- seed protein association?

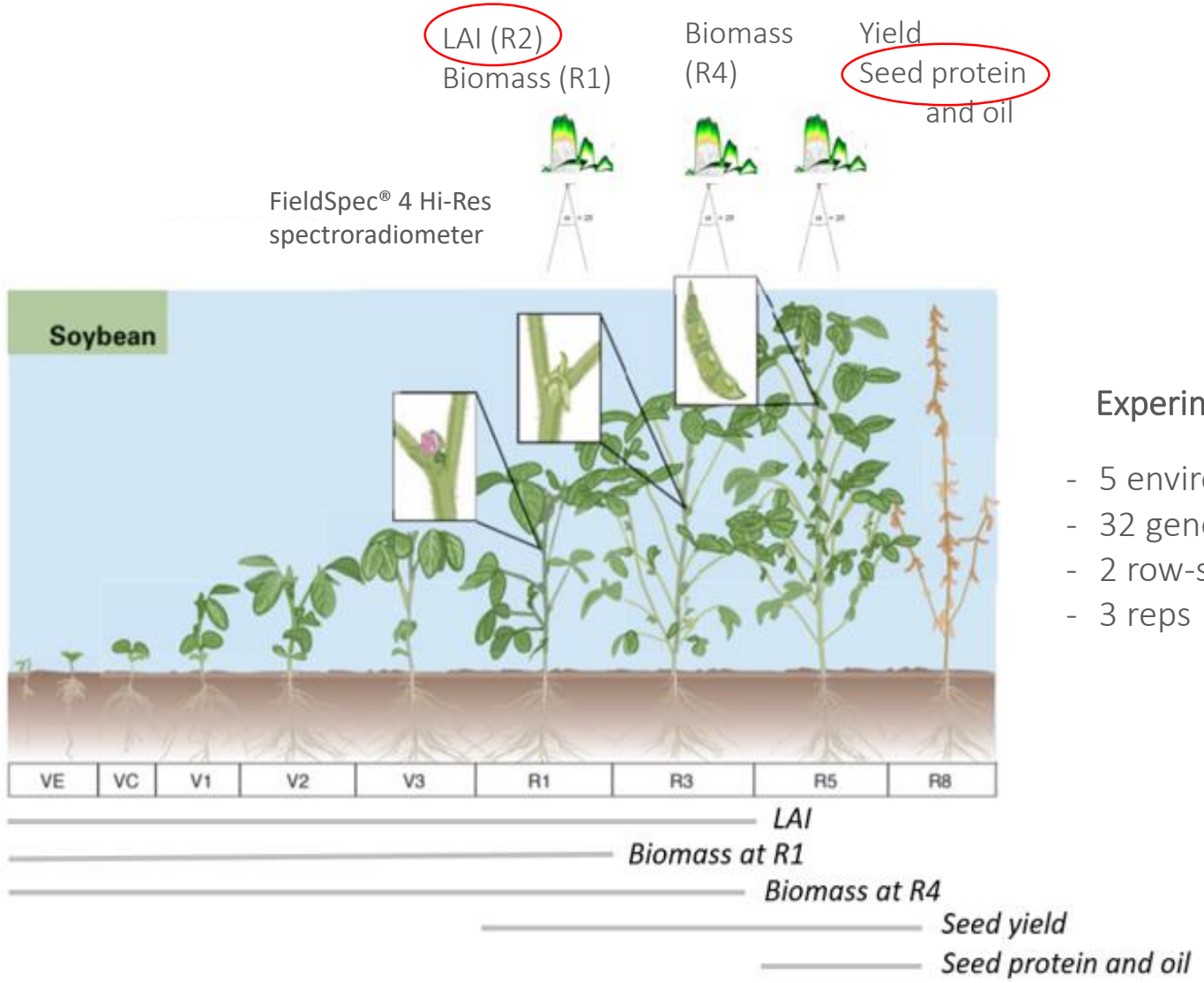
Predicting soybean crop variables using high-throughput hyperspectral imaging

From a breeding perspective……HI allows phenotyping in extensive areas in a faster, non-destructive, and more cost-effective manner compared to direct measurements.



What crop variable we can best predict using the hyperspectral range (350-2500 nm)?

Predicting soybean crop variables using high-throughput hyperspectral imaging



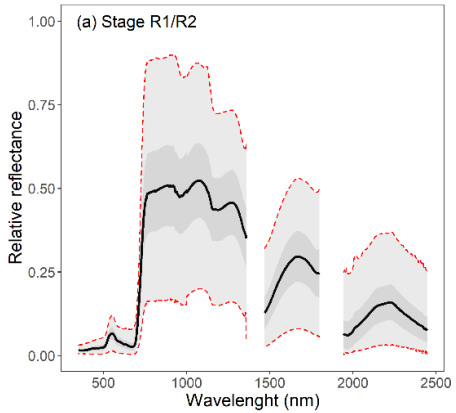
Experimental design

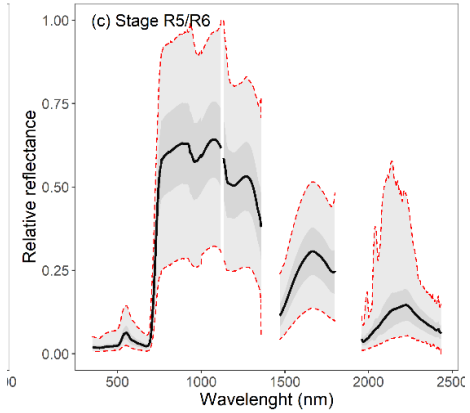
- 5 environments (yr./site)
- 32 genotypes
- 2 row-spacing (15 and 30 in)
- 3 reps

Data from Dr. Singh's experiments

Predicting soybean crop variables using high-throughput hyperspectral imaging

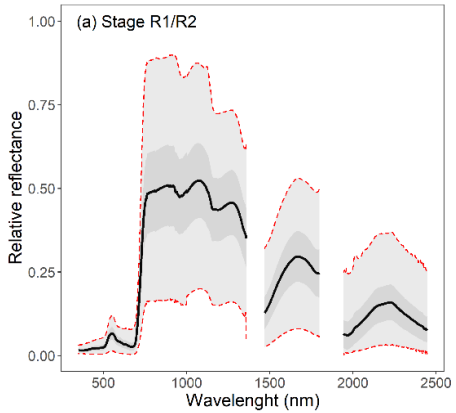
Predictors

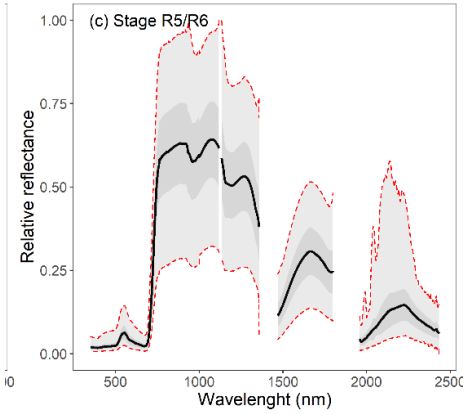




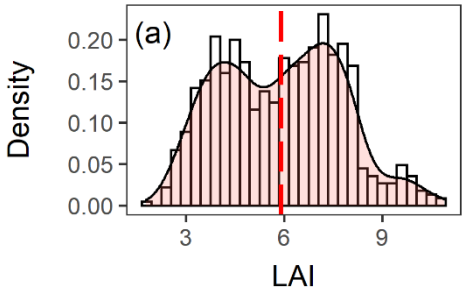
Predicting soybean crop variables using high-throughput hyperspectral imaging

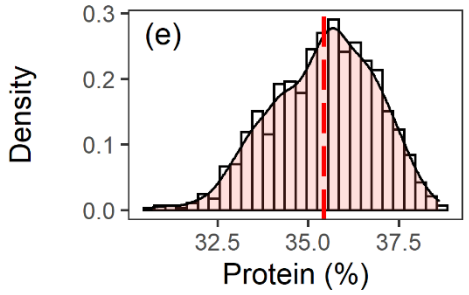
Predictors





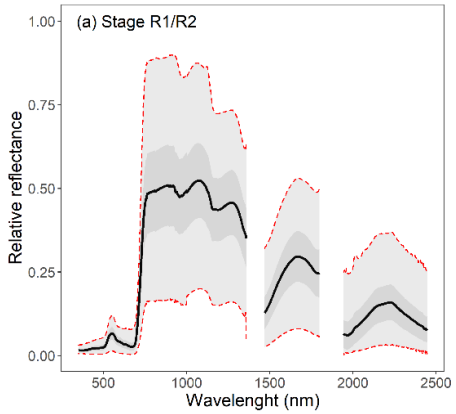
Observed variables

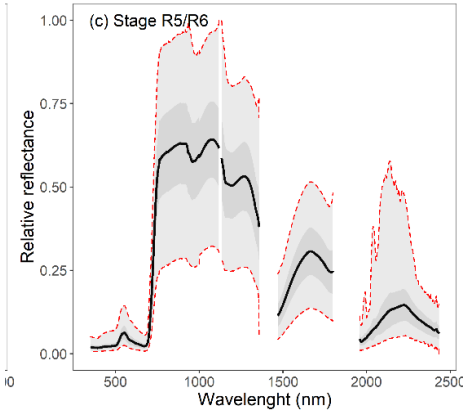




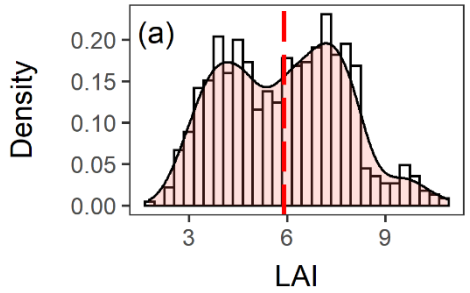
Predicting soybean crop variables using high-throughput hyperspectral imaging

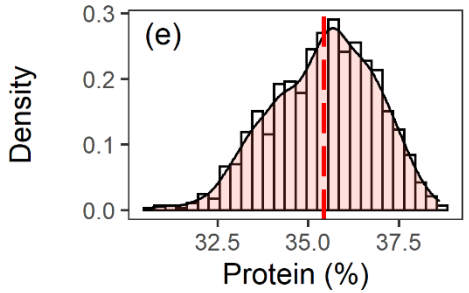
Predictors



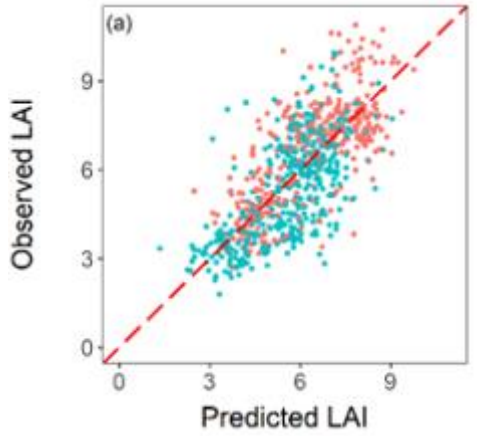


Observed variables

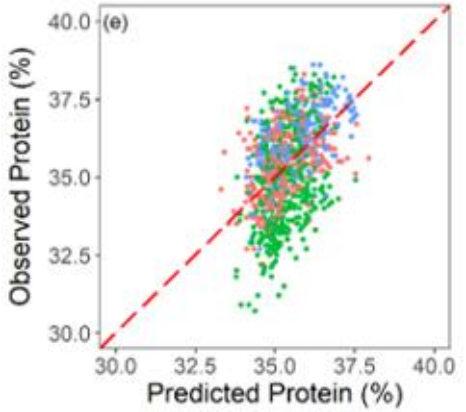




Predicted vs Observed values



ME= 0.5



ME= 0.2

Why I am interested in the LAI- seed protein association?

Because…..LAI is the best variable we can extract from hyperspectral images and there is evidence of LAI and Protein % association

Studying the total causal effect of the LAI on Protein concentration

Total causal effect of the LAI on Protein concentration

A way of describing qualitative causal relations among variables:

DAGs = DIRECTED ACYCLIC GRAPH

DAGs in R…….

- **dagitty: Graphical Analysis of Structural Causal Models**

- **ggdag: Analyze and Create Elegant Directed Acyclic Graphs**
is built on top of 'dagitty', makes it easy to tidy and plot 'dagitty' objects using 'ggplot2' and 'ggraph'

Total causal effect of the LAI on Protein concentration

What variables to include or not include in the DAG…….Here is the recipe:

Total causal effect of the LAI on Protein concentration

What variables to include or not include in the DAG…….Here is the recipe:

1. List all the paths connecting LAI (the potential cause of interest) and seed protein concentration (the outcome).

Total causal effect of the LAI on Protein concentration

What variables to include or not include in the DAG…….Here is the recipe:

1. List all the paths connecting LAI (the potential cause of interest) and seed protein concentration (the outcome).

2. Classify each path by whether is open or close. A path is open unless it contains a collider.

Total causal effect of the LAI on Protein concentration

What variables to include or not include in the DAG…….Here is the recipe:

1. List all the paths connecting LAI (the potential cause of interest) and seed protein concentration (the outcome).

2. Classify each path by whether is open or close. A path is open unless it contains a collider.

3. Classify each path by whether it is a backdoor path. A backdoor path has an arrow entering LAI

Total causal effect of the LAI on Protein concentration

What variables to include or not include in the DAG…….Here is the recipe:

1. List all the paths connecting LAI (the potential cause of interest) and seed protein concentration (the outcome).

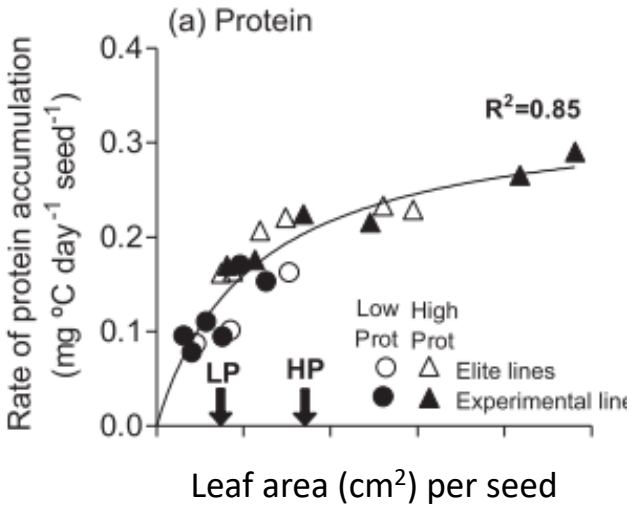
2. Classify each path by whether is open or close. A path is open unless it contains a collider.

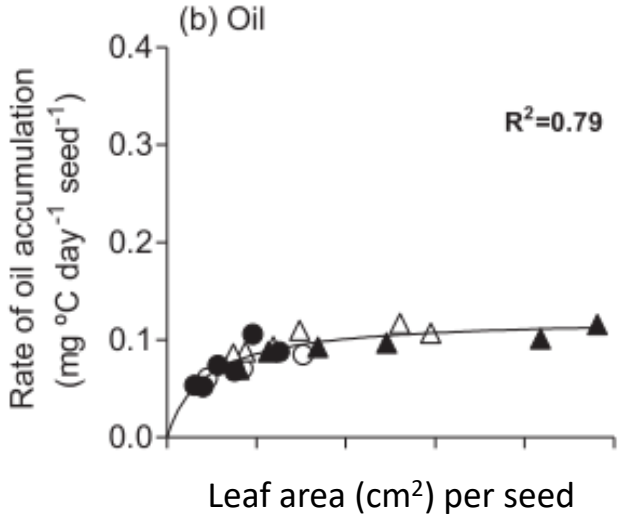
3. Classify each path by whether it is a backdoor path. A backdoor path has an arrow entering LAI

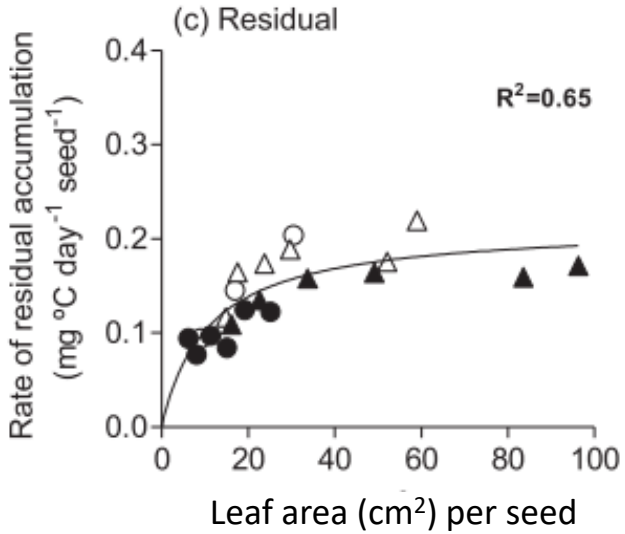
4. If there are any open backdoor paths, decide which variable(s) to condition on to close it (if possible)

Total causal effect of the LAI on Protein concentration

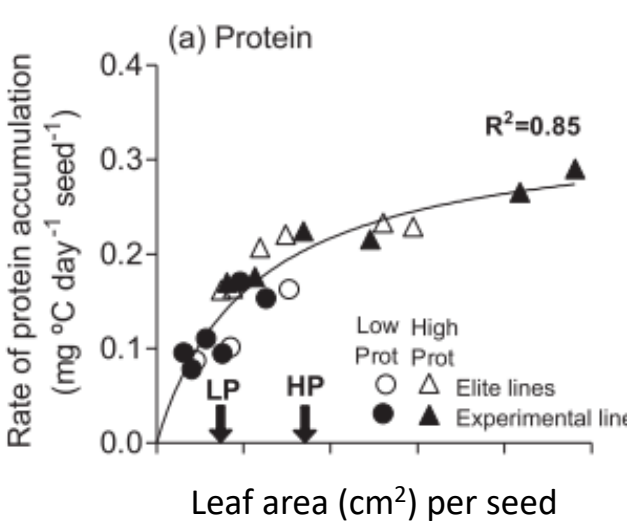
LAI per seed as a proxy for assimilates supply per seed.

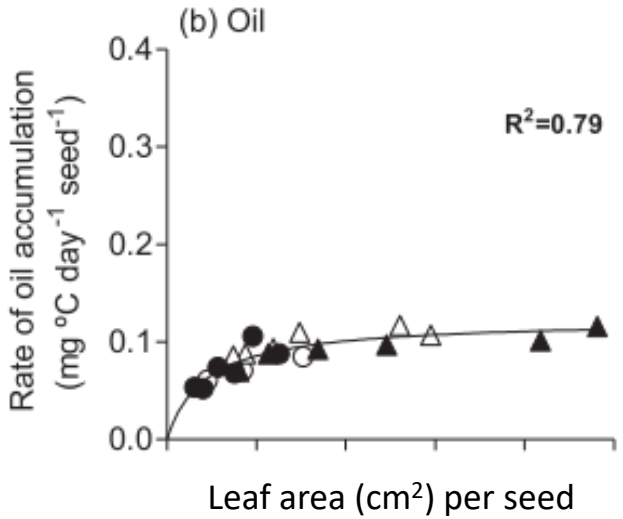


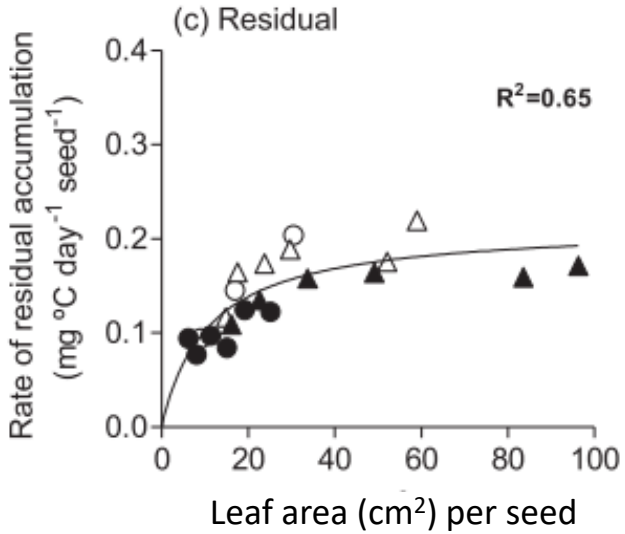




Total causal effect of the LAI on Protein concentration





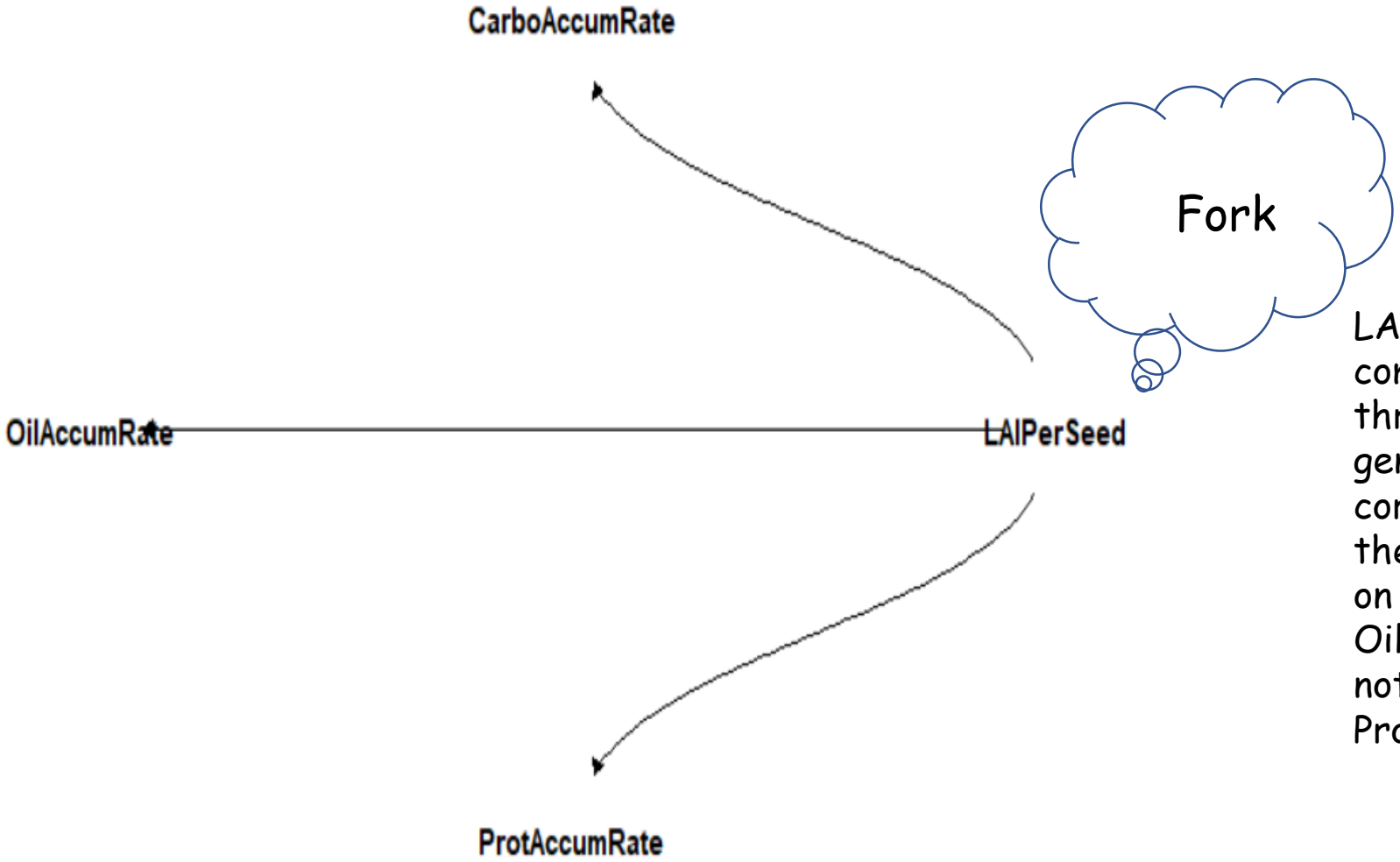


ggdag package

```
small_dag1 <- dagify(CarboAccumRate ~ LAIPerSeed,
                     ProtAccumRate ~ LAIPerSeed,
                     OilAccumRate ~ LAIPerSeed,
                     exposure = "LAIPerSeed",
                     outcome = "ProtAccumRate")

ggdag(small_dag1, layout = "circle", text_col = "black", node_size = 16, edge_type = "diagonal", node = FALSE, text_size
= 4)+
  theme_dag_blank()
```

Total causal effect of the LAI on Protein concentration

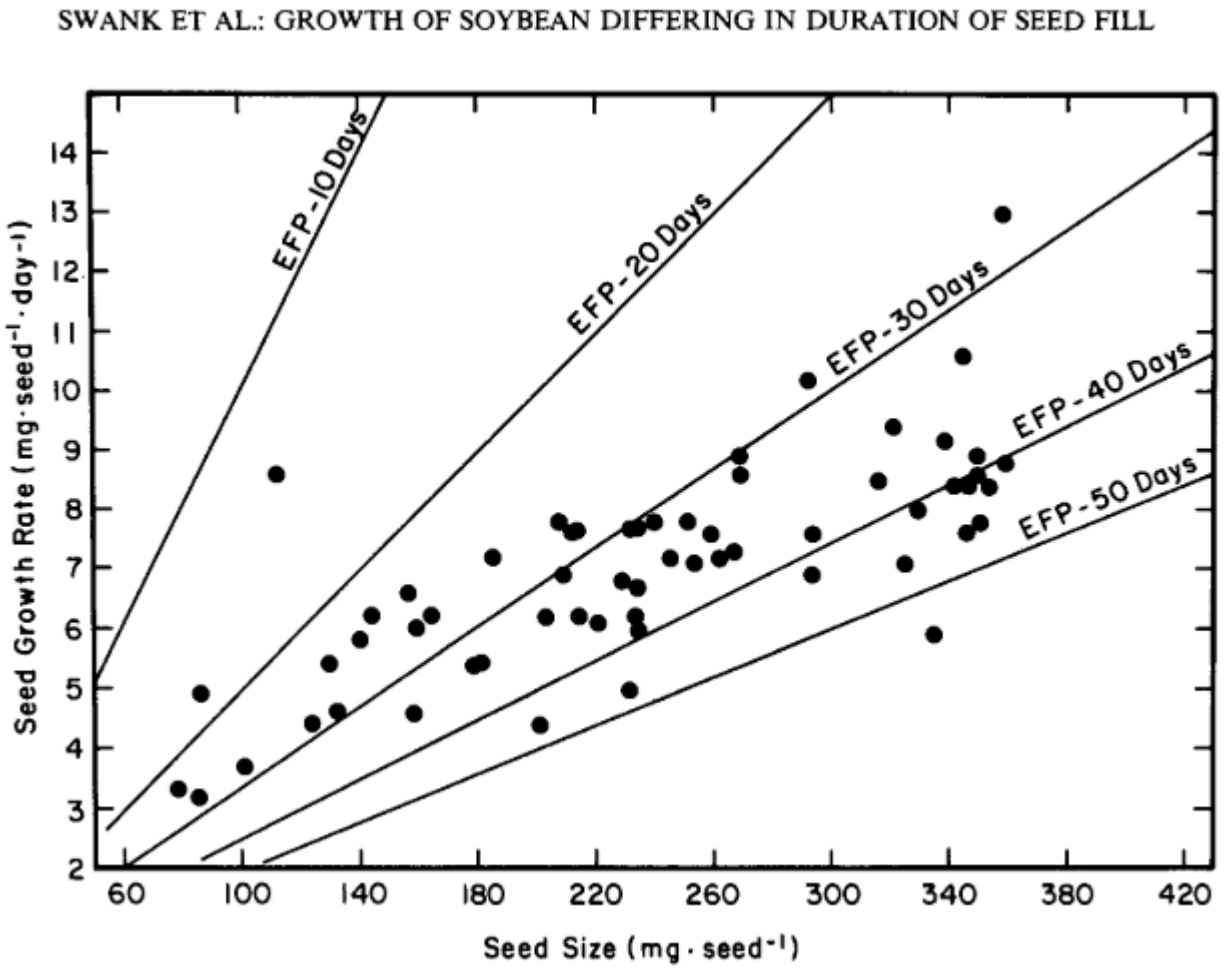
A fork is a classic confounder



LAIPerSeed is a common cause of three variables, generating a correlation between them. If we condition on LAIPerSeed, OilAccumRate tell us nothing more about ProtAccumRate.

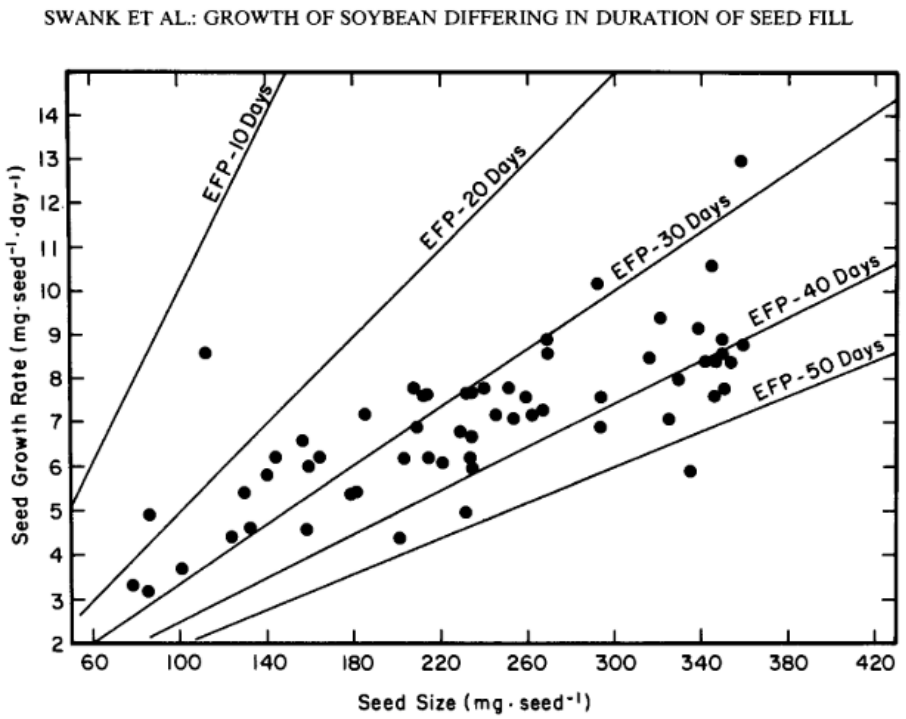
Total causal effect of the LAI on Protein concentration

Content = rate x duration



Total causal effect of the LAI on Protein concentration

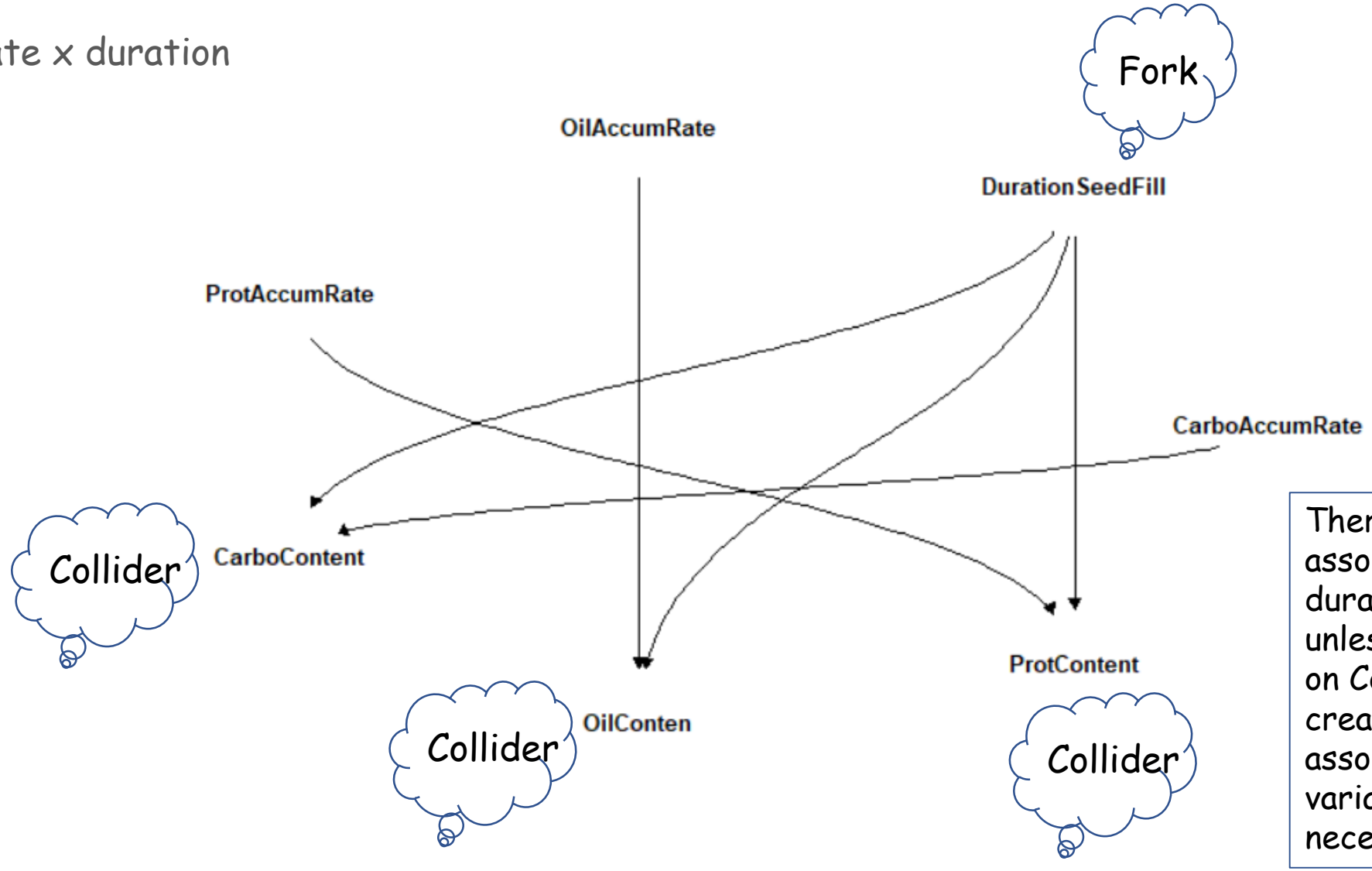
Content = rate x duration




```
small_dag2 <- dagify(CarboContent ~ CarboAccumRate + DurationSeedFill,
                     ProtContent ~ ProtAccumRate + DurationSeedFill,
                     OilConten ~ OilAccumRate + DurationSeedFill,
exposure = "ProtAccumRate",
outcome = "ProtContent")

ggdag(small_dag2, layout = "circle", text_col = "black", node_size = 16, edge_type = "diagonal", node = FALSE, text_size
= 3.5)+
  theme_dag_blank()
```

Total causal effect of the LAI on Protein concentration

Content = rate x duration



There is no association between duration and Rate unless we condition on Content. A collider creates statistical associations between variables, not necessarily causal.

Total causal effect of the LAI on Protein concentration

Concentration is a mathematical construct that relates the content of a particular component to the total weight of the seed (i.e. the sum of all components).

Protein Concentration (%) =

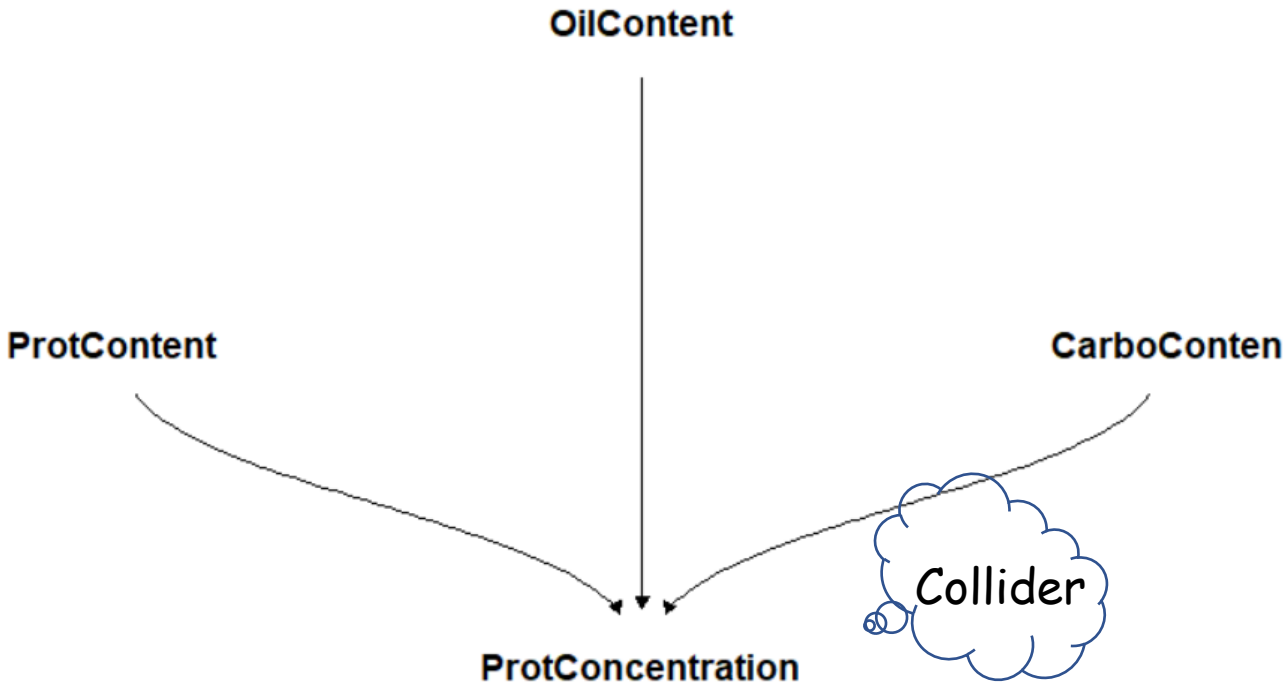
(mg protein/mg dry weight) * 100

mg of Protein + Oil + Carbohydrates

```
small_dag3 <- dagify(ProtConcentration ~ CarboContent + ProtContent + OilContent,
exposure = "ProtContent",
outcome = "ProtConcentration")

ggdag(small_dag3, layout = "circle", text_col = "black", node_size = 16, edge_type = "diagonal", node = FALSE, text_size
= 3.5)+
  theme_dag_blank()
```

Total causal effect of the LAI on Protein concentration

Knowing protein concentration, then learning about Protein Content also give you information about Oil and Carbo Content. The same works in reverse



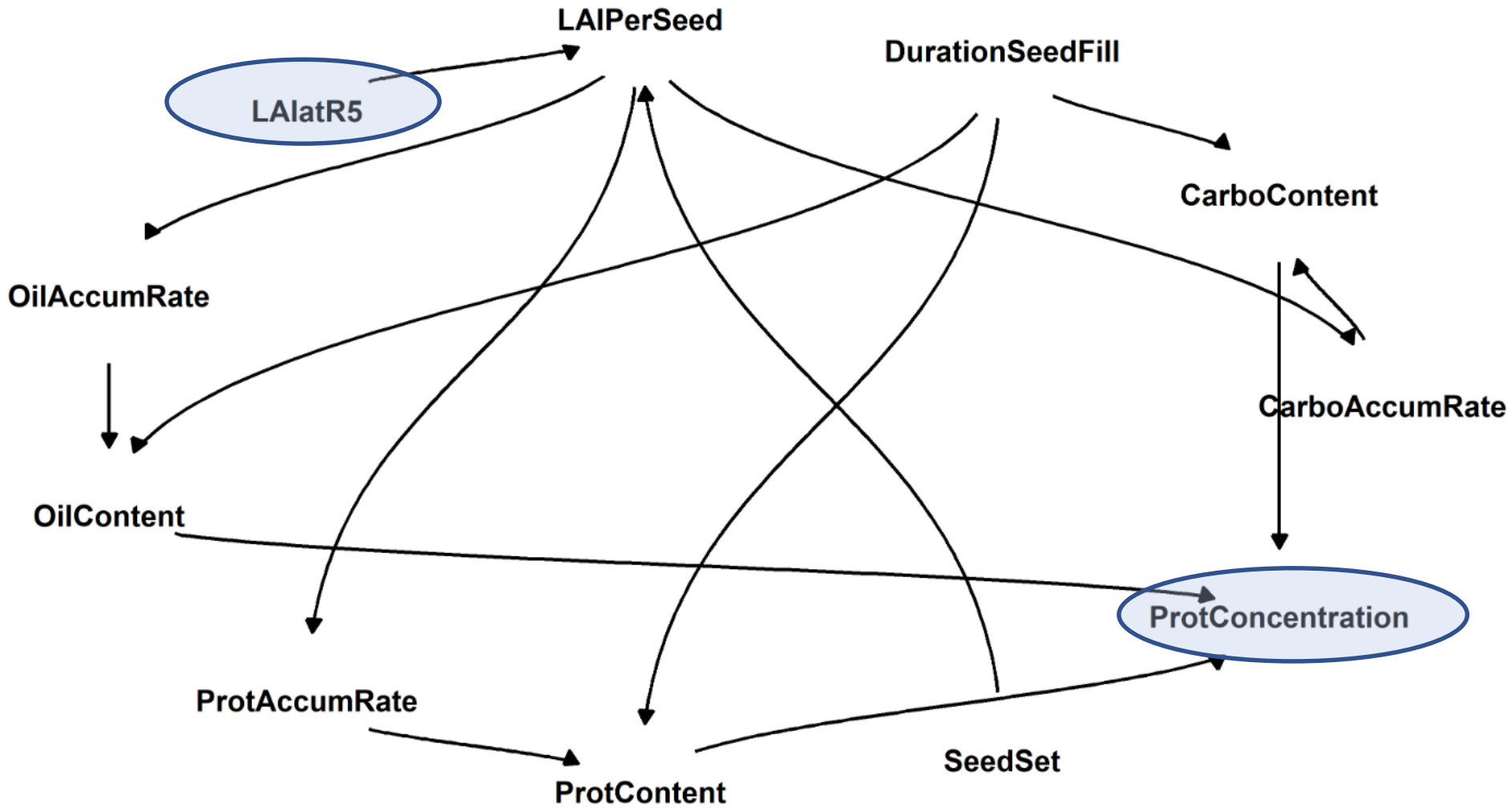
This is the reason for a negative association between ProtContent and OilContent

Total causal effect of the LAI on Protein concentration

All paths together.....

```
bigger_dag2 <- dagify(ProtConcentration ~ CarboContent + ProtContent + OilContent,
                      CarboContent ~ CarboAccumRate + DurationSeedFill,
                      ProtContent ~ ProtAccumRate + DurationSeedFill,
                      OilContent ~ OilAccumRate + DurationSeedFill,
                      CarboAccumRate ~ LAIPerSeed,
                      ProtAccumRate ~ LAIPerSeed,
                      OilAccumRate ~ LAIPerSeed,
                      LAIPerSeed ~ LAIatR5 + SeedSet,
                      exposure = "LAIatR5",
                      outcome = "ProtConcentration")

ggdag(bigger_dag2, layout = "circle", text_col = "black", node_size = 16, edge_type = "diagonal", node = FALSE,
text_size = 3.5)+
  theme_dag_blank()
```

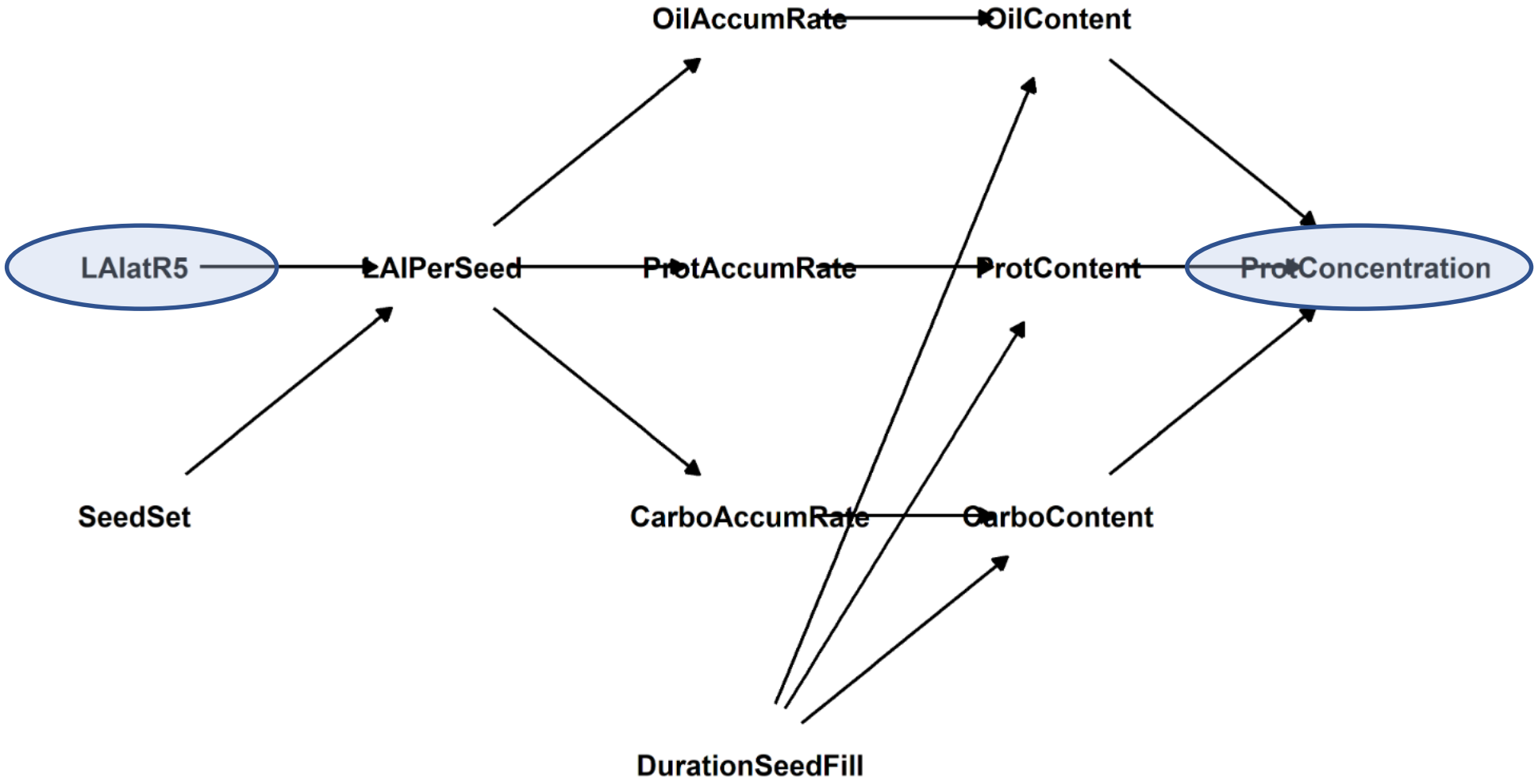
Total causal effect of the LAI on Protein concentration



Total causal effect of the LAI on Protein concentration

```r
# Modify the DAG format using coordinates

coords <- list(
  x = c(LAIatR5 = 1, SeedSet = 1, LAIPerSeed = 2, CarboAccumRate = 3, ProtAccumRate = 3, OilAccumRate = 3,
DurationSeedFill = 3, ProtContent = 4, OilContent = 4, CarboContent = 4, ProtConcentration = 5),


  y = c(SeedSet = -1, LAIatR5 = 0, LAIPerSeed = 0, ProtAccumRate = 0, OilAccumRate = 1, CarboAccumRate = -1, ProtContent
= 0, OilContent = 1, CarboContent = -1, DurationSeedFill = -2, ProtConcentration = 0)
)

coord_df <- coords2df(coords)
coords2list(coord_df)

bigger_dag2.2 <- dagify(ProtConcentration ~ CarboContent + ProtContent + OilContent,
                        CarboContent ~ CarboAccumRate + DurationSeedFill,
                        ProtContent ~ ProtAccumRate + DurationSeedFill,
                        OilContent ~ OilAccumRate + DurationSeedFill,
                        CarboAccumRate ~ LAIPerSeed,
                        ProtAccumRate ~ LAIPerSeed,
                        OilAccumRate ~ LAIPerSeed,
                        LAIPerSeed ~ LAIatR5 + SeedSet,
                        exposure = "LAIatR5",
                        outcome = "ProtConcentration")

coordinates(bigger_dag2.2) <- coords2list(coord_df)
ggdag(bigger_dag2.2, text_col = "black", text_size = 3.5, node = FALSE)+
  theme_dag_blank()+expand_plot()
ggsave("bigger_dag2.2.png", dpi = 500)
```

Total causal effect of the LAI on Protein concentration



Total causal effect of the LAI on Protein concentration

- Many potential predictor variables, why don't we add all of them?

Total causal effect of the LAI on Protein concentration

- Many potential predictor variables, why don't we add all of them?
    - Multicollinearity = Strong association among predictor variables. None of the variables will be associated with the outcome. Ex = including both legs to predict height

Total causal effect of the LAI on Protein concentration

- Many potential predictor variables, why don't we add all of them?
  - Multicollinearity = Strong association among predictor variables. None of the variables will be associated with the outcome. Ex = including both legs to predict height
  - Post-treatment bias = Included variables bias. Mistaken inference for including variables. Ex = including fungus and fungicide treatment to predict growth

Total causal effect of the LAI on Protein concentration

- Many potential predictor variables, why don't we add all of them?
  - Multicollinearity = Strong association among predictor variables. None of the variables will be associated with the outcome. Ex = including both legs to predict height
  - Post-treatment bias = Included variables bias. Mistaken inference for including variables. Ex = including fungus and fungicide treatment to predict growth
  - Collider bias = When a variable is a common consequence of other variables. It generates statistical associations between the predictors. Ex = trustworthiness and newsworthiness influencing selection of proposals for funding.

Total causal effect of the LAI on Protein concentration

Some variables are independent from others under certain conditions= Conditional Independencies

- Which variables should be associated (or not) with one another?
- Which variables become dis-associated when we condition on some other set of variables?

Total causal effect of the LAI on Protein concentration

Some variables are independent from others under certain conditions= Conditional Independencies

- Which variables should be associated (or not) with one another.
- Which variables become dis-associated when we condition on some other set of variables.

Conditioning on a variable (Z) means learning its value and then asking if a variable X adds any additional information about Y.

If I know the value of Z, should I include X in the model to know Y??

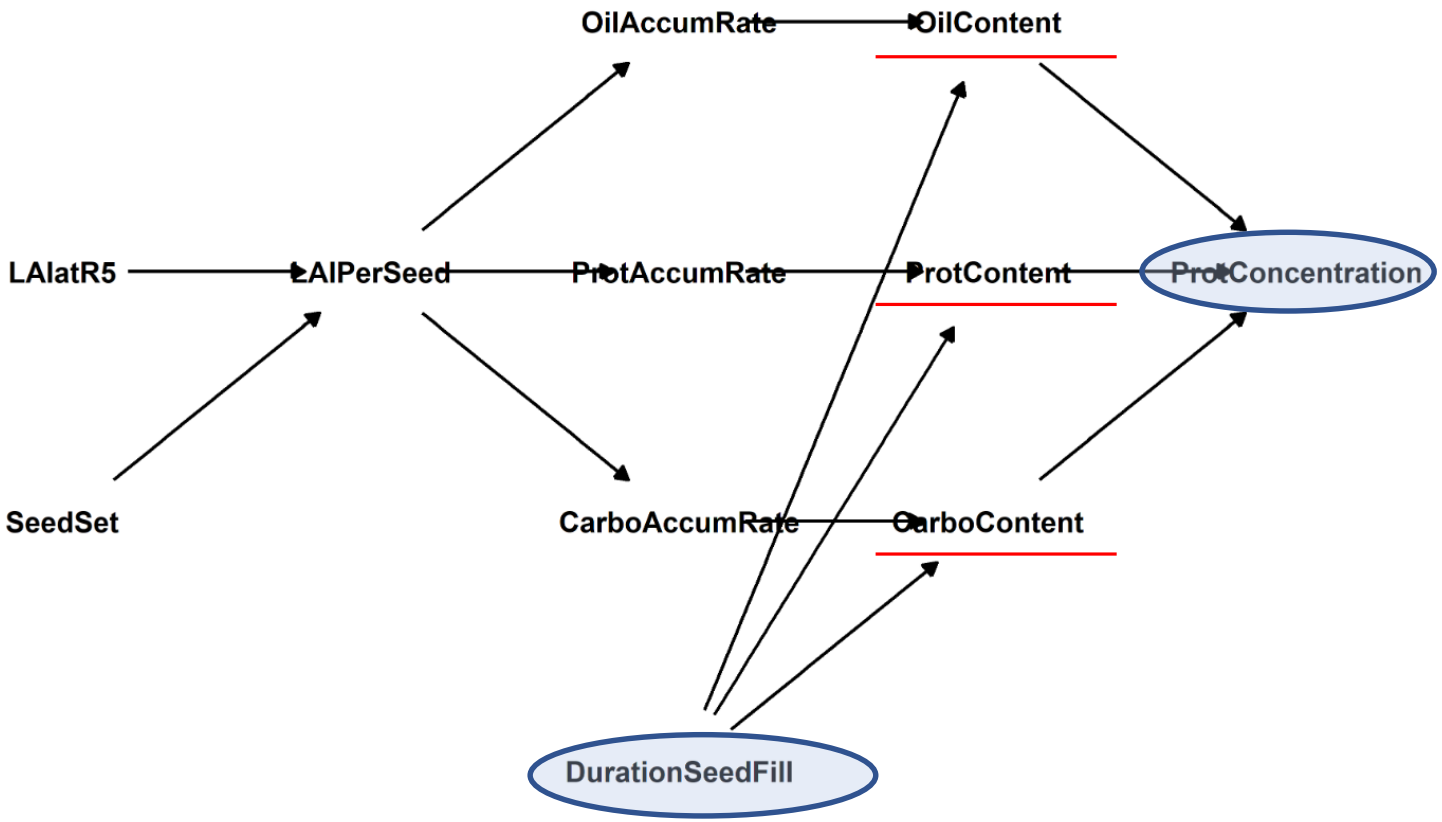
$$Y \perp\!\!\!\perp X|Z$$

Total causal effect of the LAI on Protein concentration

Examples……

```
impliedConditionalIndependencies(bigger_dag2)
```

```
DrSF _||_ PrtCnc | CrbC, OlCn, PrtCnt
```

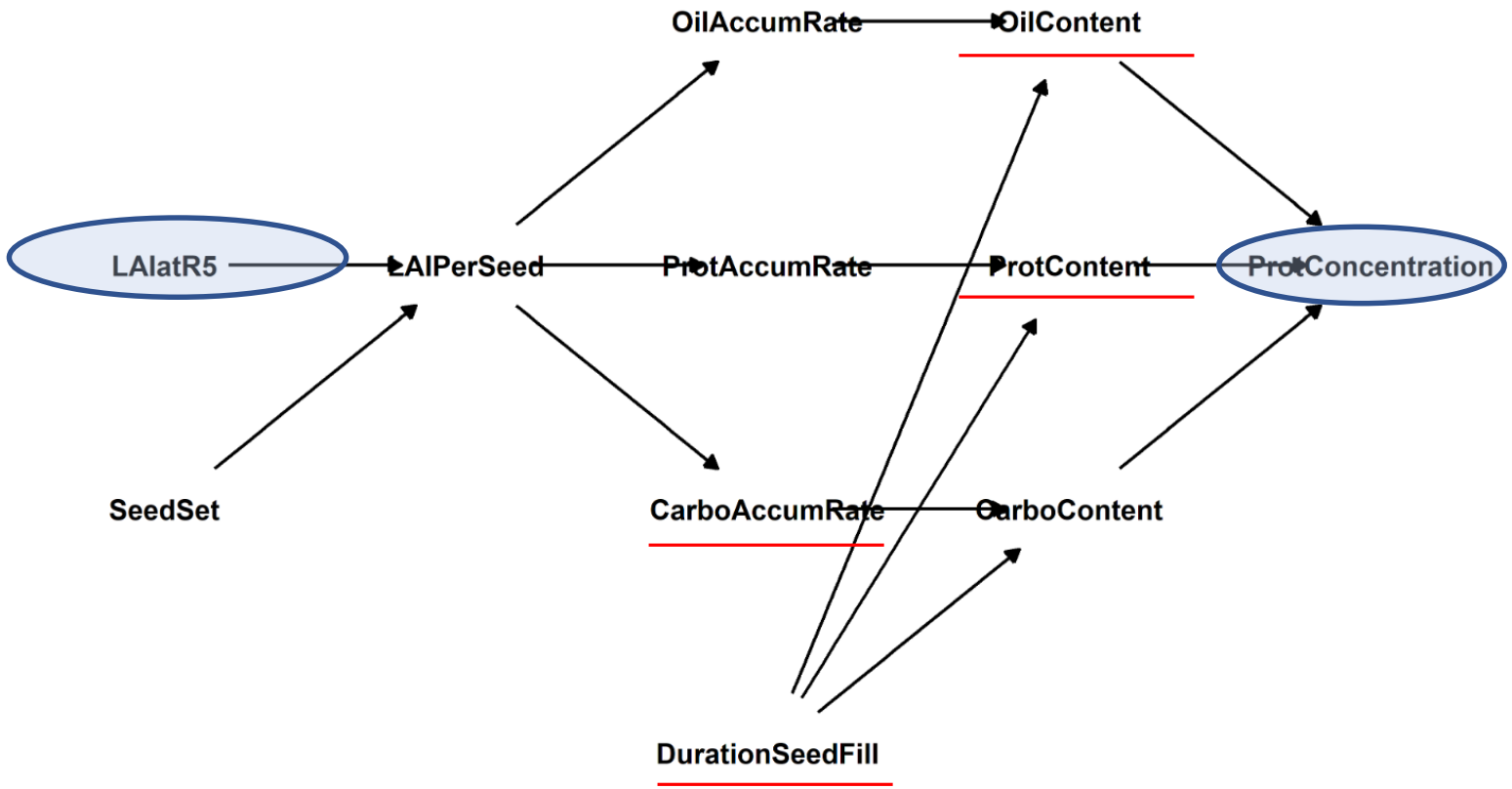


Total causal effect of the LAI on Protein concentration

Examples……

```
impliedConditionalIndependencies(bigger_dag2)
```

```
LAIR _||_ PrtCnc | CrAR, DrSF, OlCn, PrtCnt
```



Total causal effect of the LAI on Protein concentration

Shutting the backdoor: Blocking confounding paths between a predictor X and an outcome Y. We do not want spurious associations

Which path is open? All paths are open, unless they have a collider.

Analyze the graph and find the variables to control for in order to block the backdoor paths

```
adjustmentSets(bigger_dag2, exposure = "LAIatR5", outcome = "ProtConcentration")
```

```
{}
```

No backdoor paths

Total causal effect of the LAI on Protein concentration

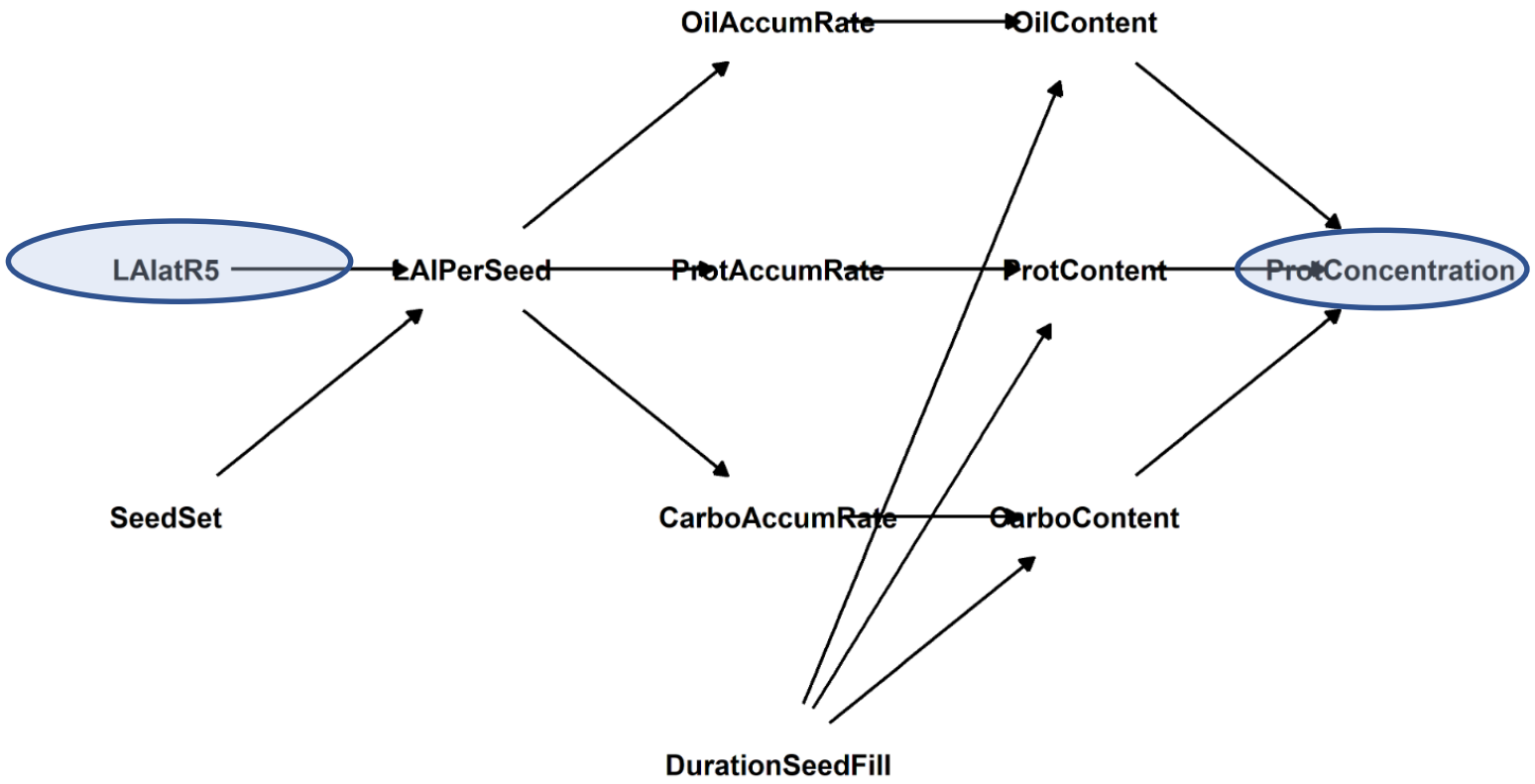
What variables to include or not include in the DAG…….Here is the recipe:

1. List all the paths connecting LAI (the potential cause of interest) and seed protein concentration (the outcome).

2. Classify each path by whether is open or close. A path is open unless it contains a collider.

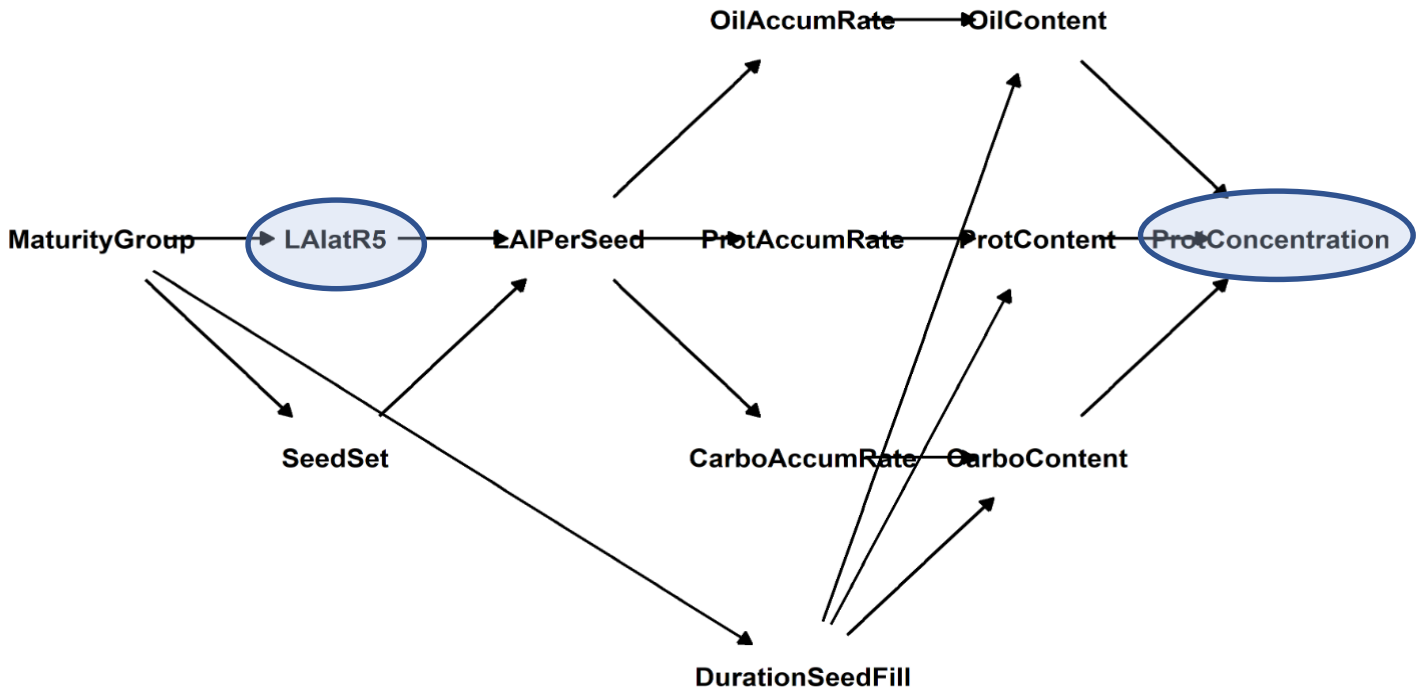
3. Classify each path by whether it is a backdoor path. A backdoor path has an arrow entering LAI

Total causal effect of the LAI on Protein concentration



Total causal effect of the LAI on Protein concentration

But, what if I include Maturity Group into the model…….




```
adjustmentSets(bigger_dag3)
```

```
{ DurationSeedFill, SeedSet }
{ MaturityGroup }
```