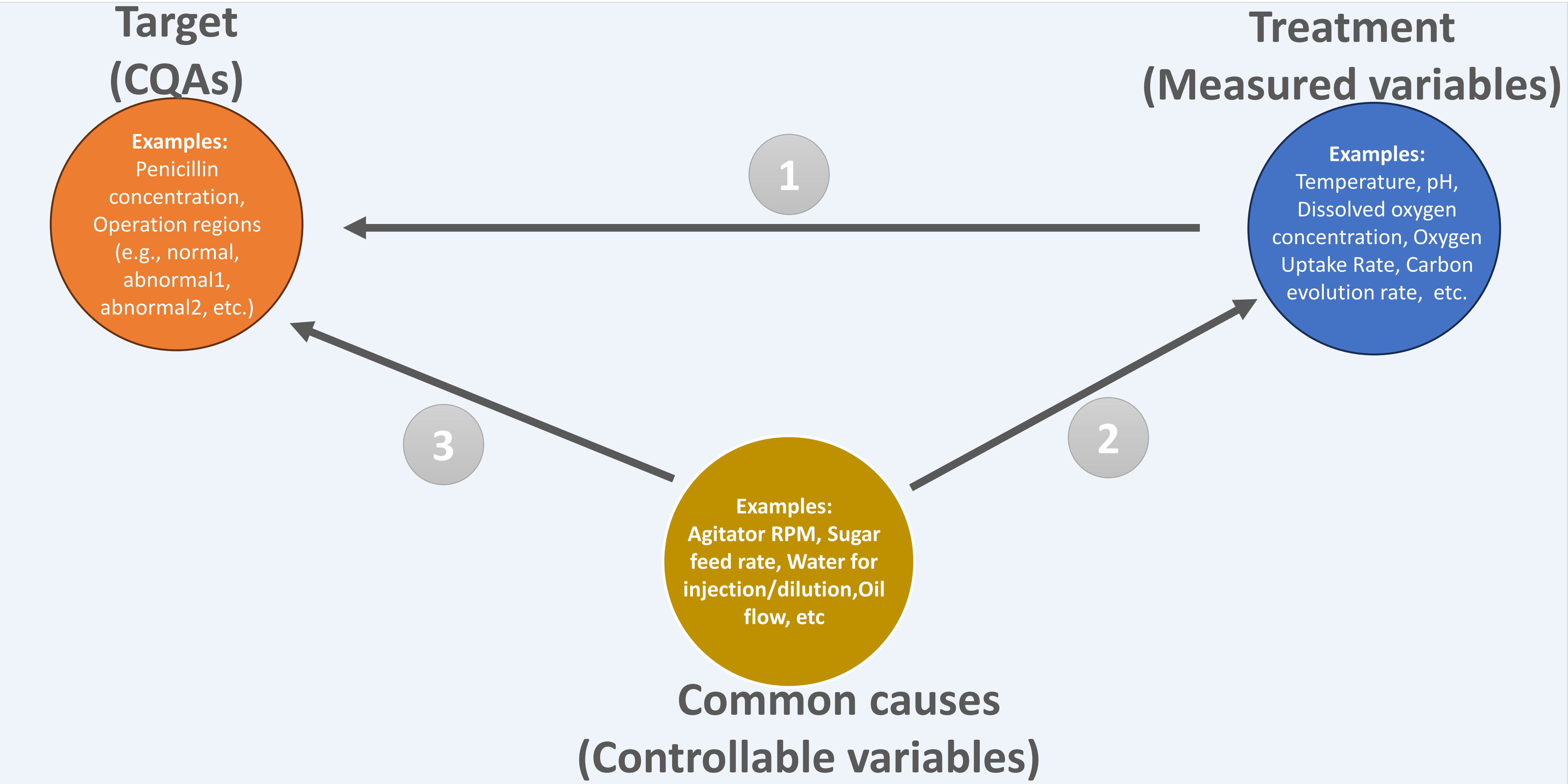
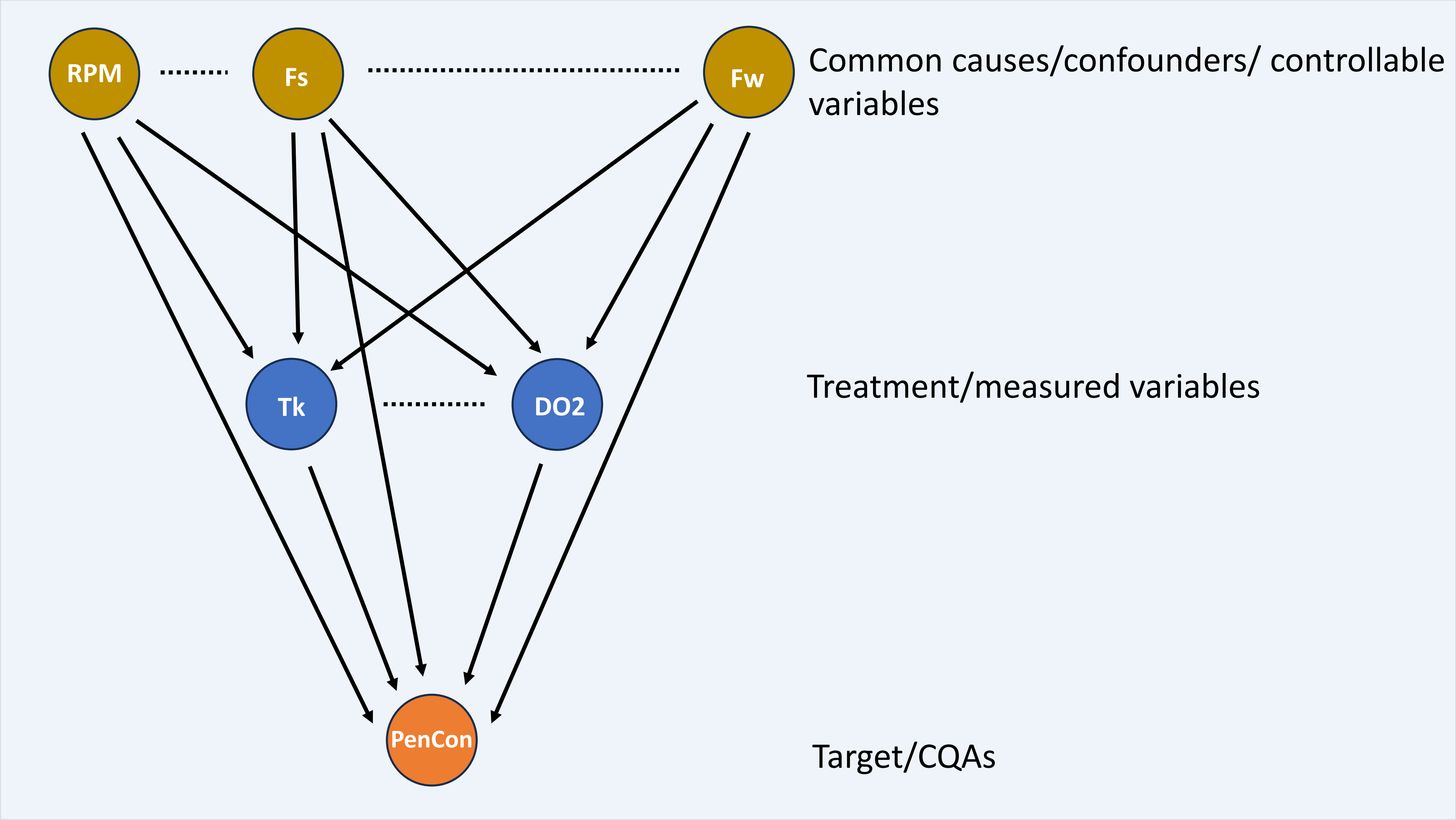
User guide for performing causal analysis

The causal discovery ‘Causal Analysis’ page enables the user to answer the question “How strong is the causal influence from a cause to its direct effect?”.



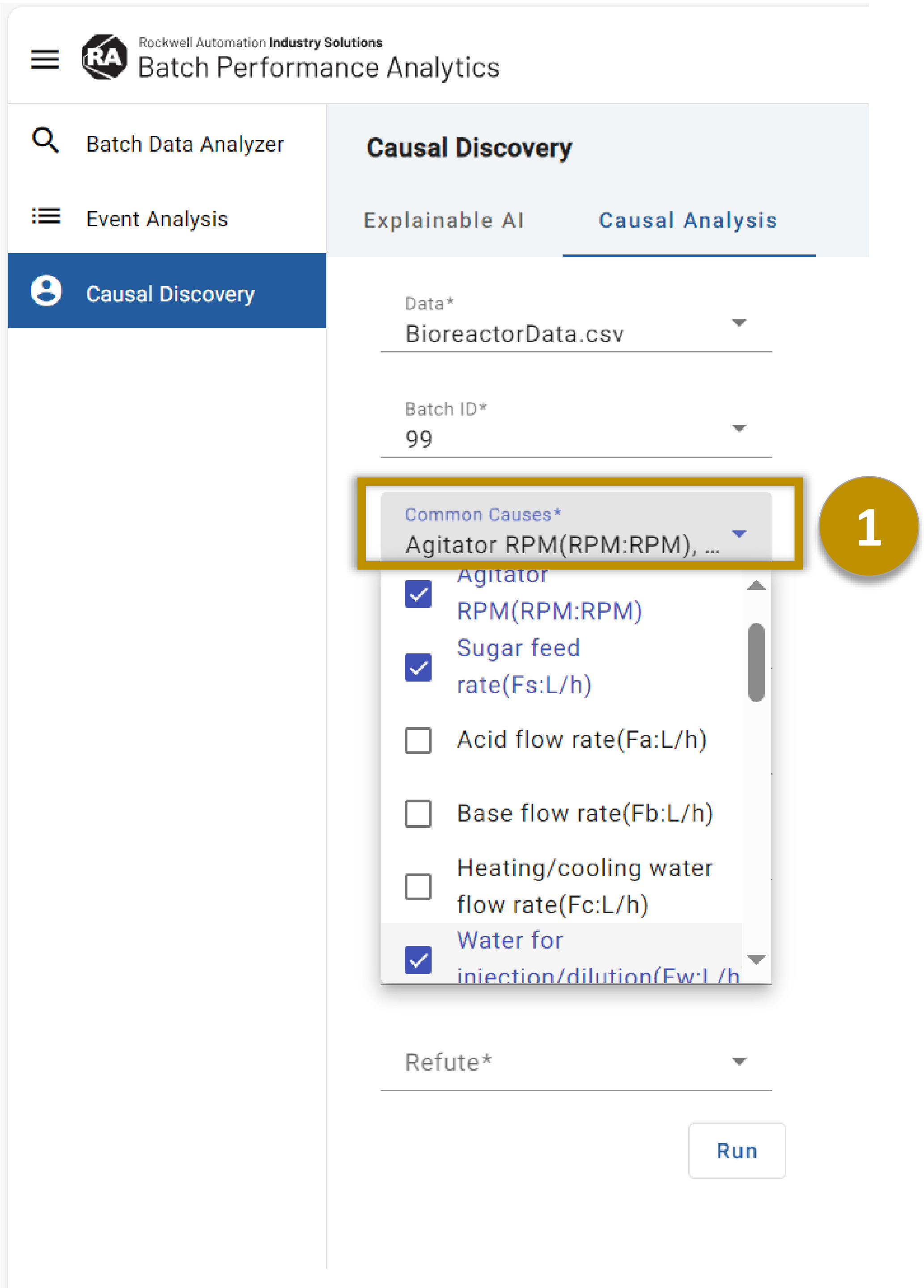
We want to estimate the effect of measured variables on CQAs. Controllable variables is a confounding variable.



Depending on the user's selections at the inputs, the causal graph representing the assumptions encoded in the causal model will follow the template as illustrated in the figure.

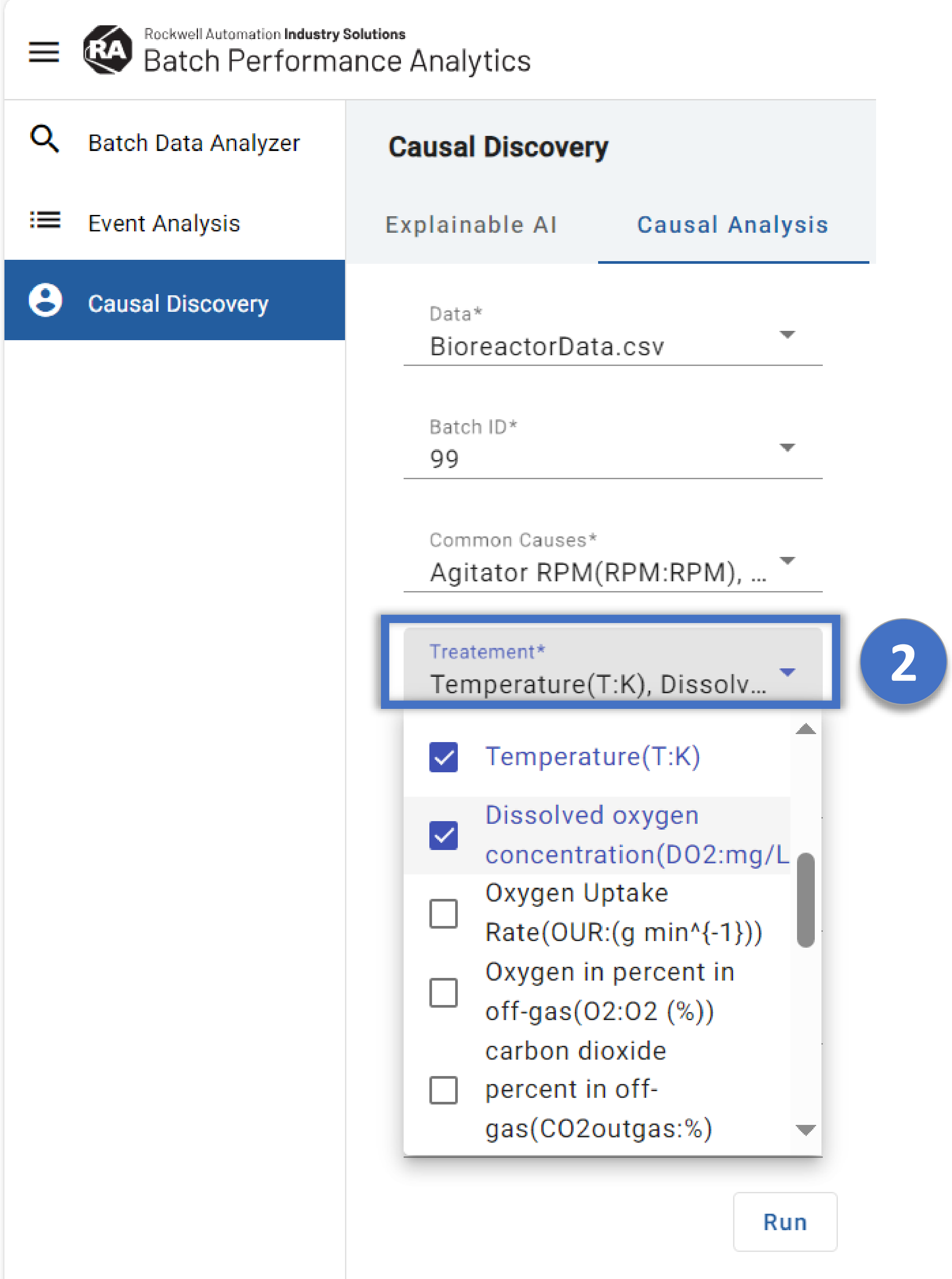
1. **First there is a direct connection between each common cause variable and the treatment variable, as well as a direct connection between each common cause and the target**

2. **There is a direct connection between each treatment variable and the target**



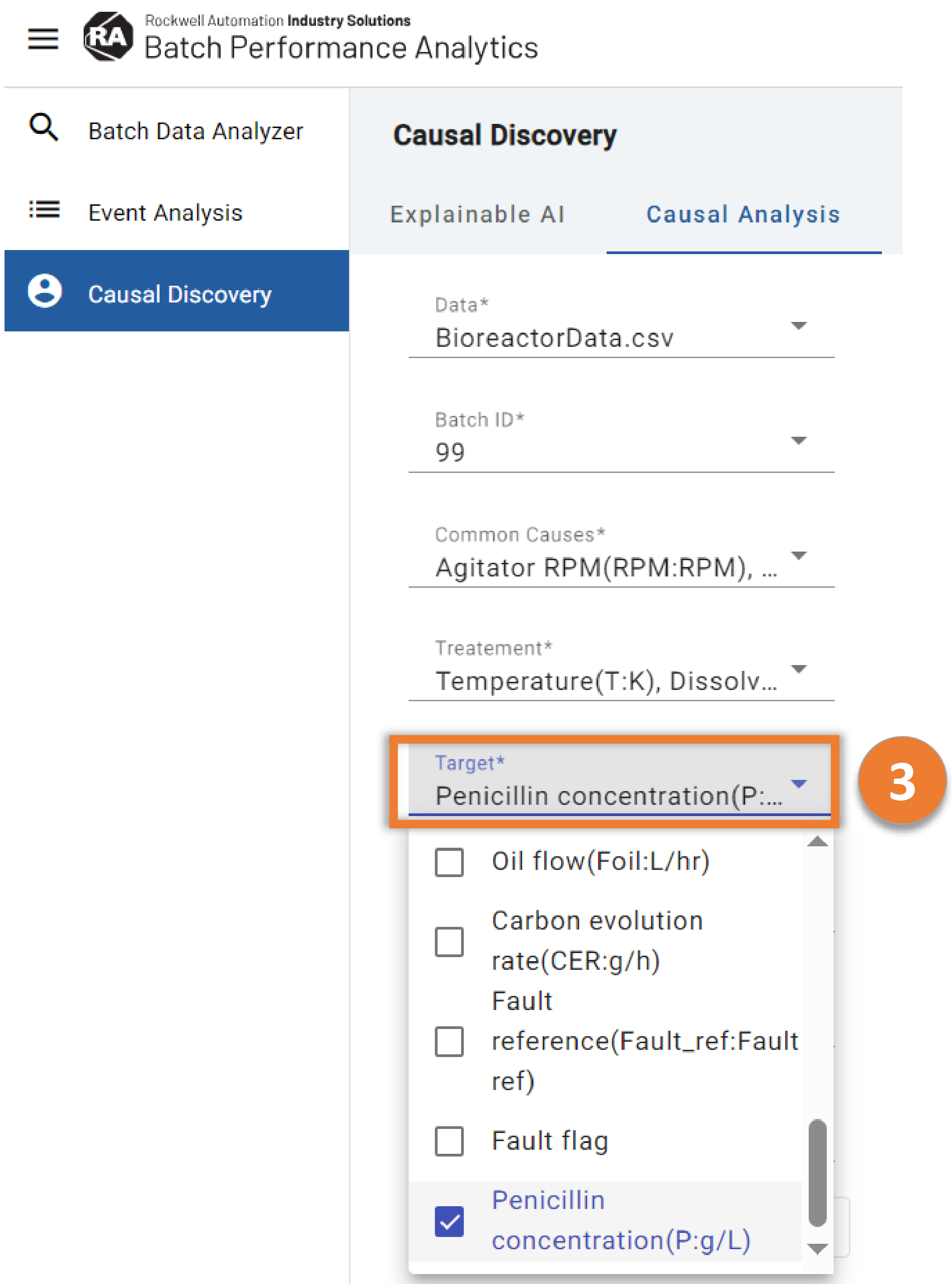
Common Causes

- Those variables that potentially influence both the treatment and target variables.
- Also known as confounding variables
- Single or multi-selection.
- E.g., selection: 'Agitator RPM', 'Sugar feed rate', and 'Water for injection/dilution'



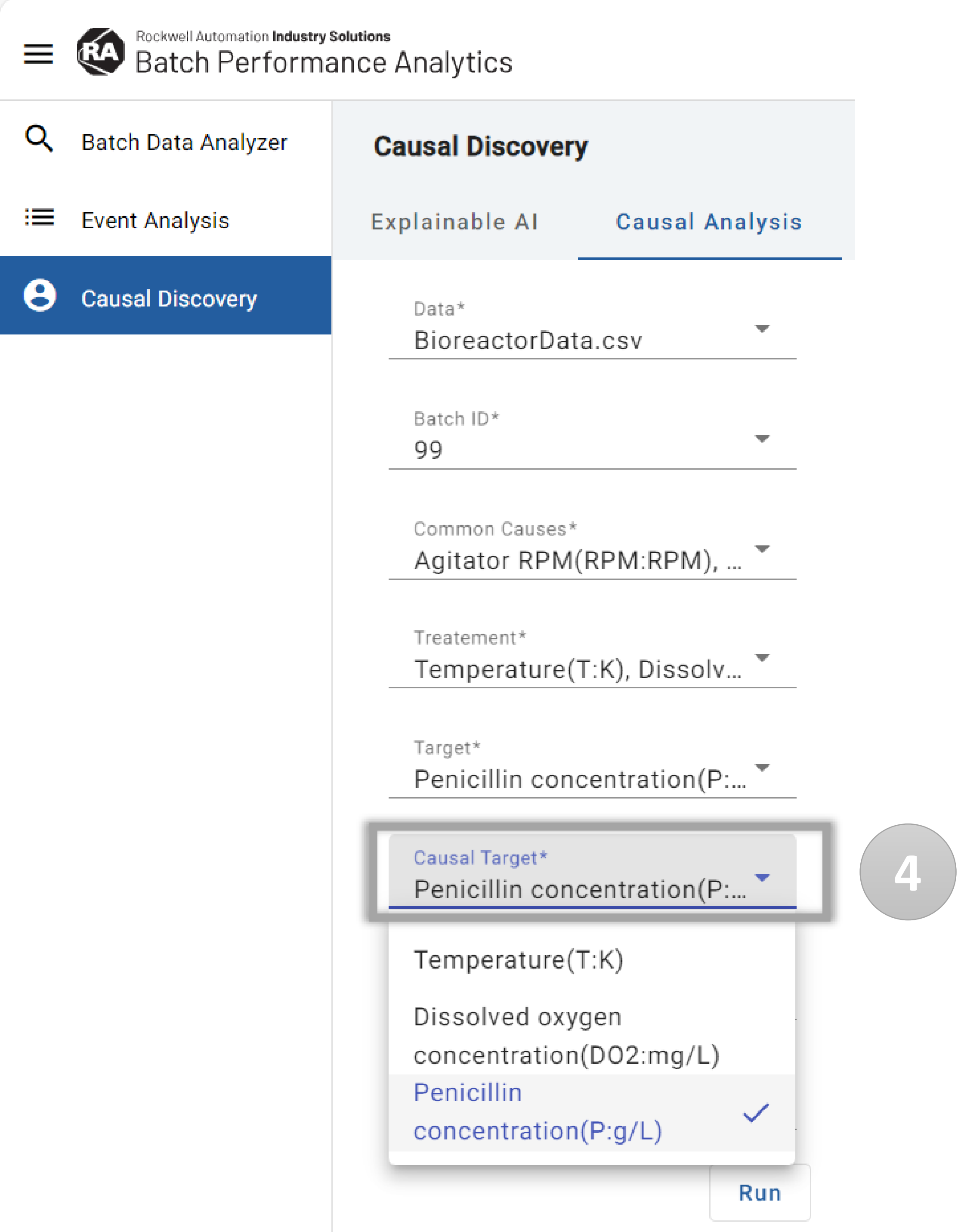
Treatment

- We are usually interested in understanding the effect of the treatment variables on the target
- Single or multi-selection.
- E.g., selection: Temperature, Dissolved oxygen concentration



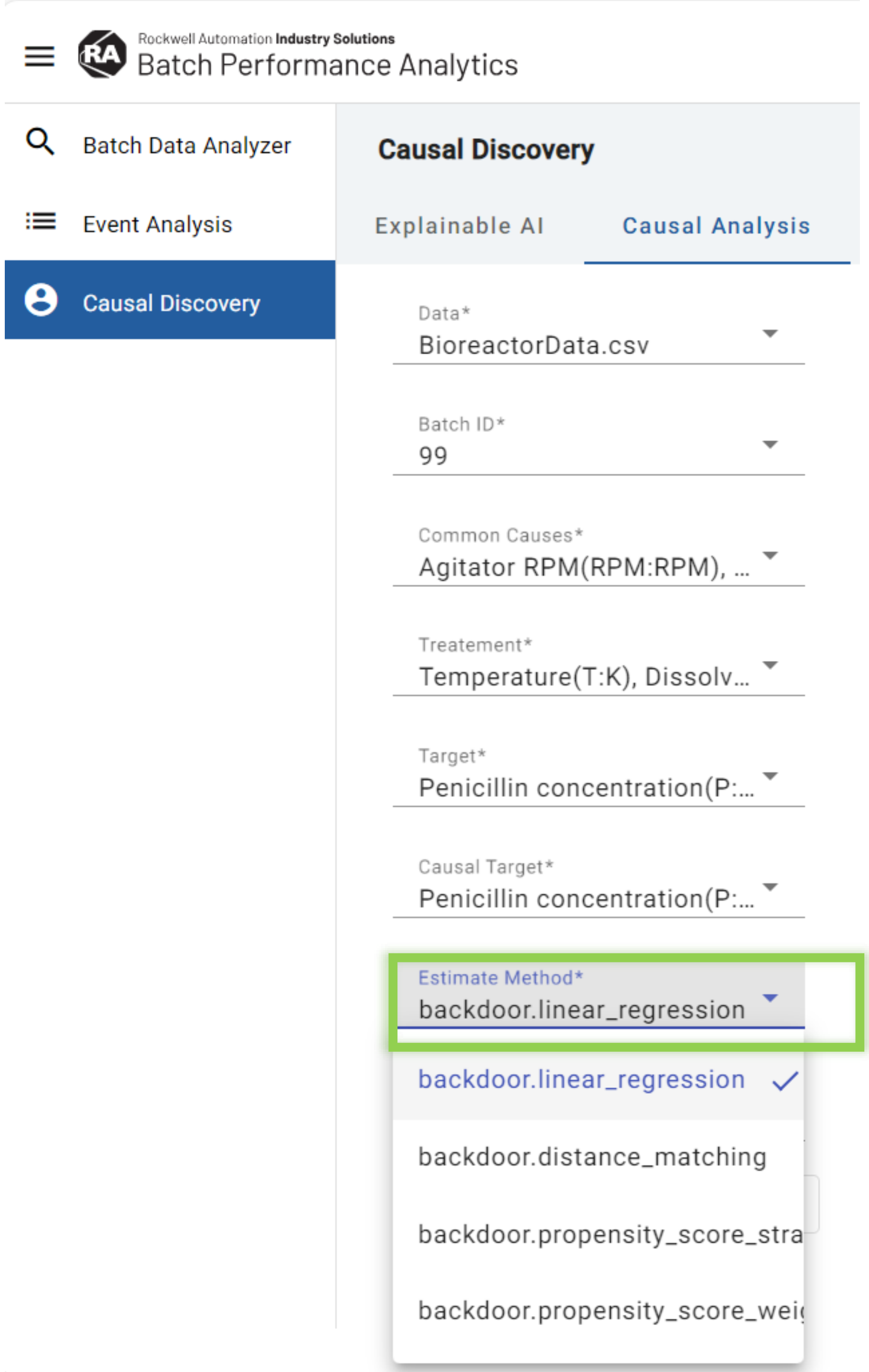
Target

- The variable(s) we want to investigate regarding the influence of treatment variables upon them
- Usually specified as CQAs in BPA
- Single or multi-selection.
- E.g., Penicillin concentration



Causal Target

- This variable (causal target) will be used to generate a graph that quantifies 'how strong is the causal influence from common causes and/or treatment to the causal target'
- Single selection
- Options will comprise users' combined selections from treatment and target.
- If the selected variable is from treatment, the user aims to understand the causal strength between the common causes and the selected treatment. If the selected variable is equal to the target from **3**, the user aims to understand the causal strength between common causes and treatment, on the target.
- E.g., Penicillin concentration



5

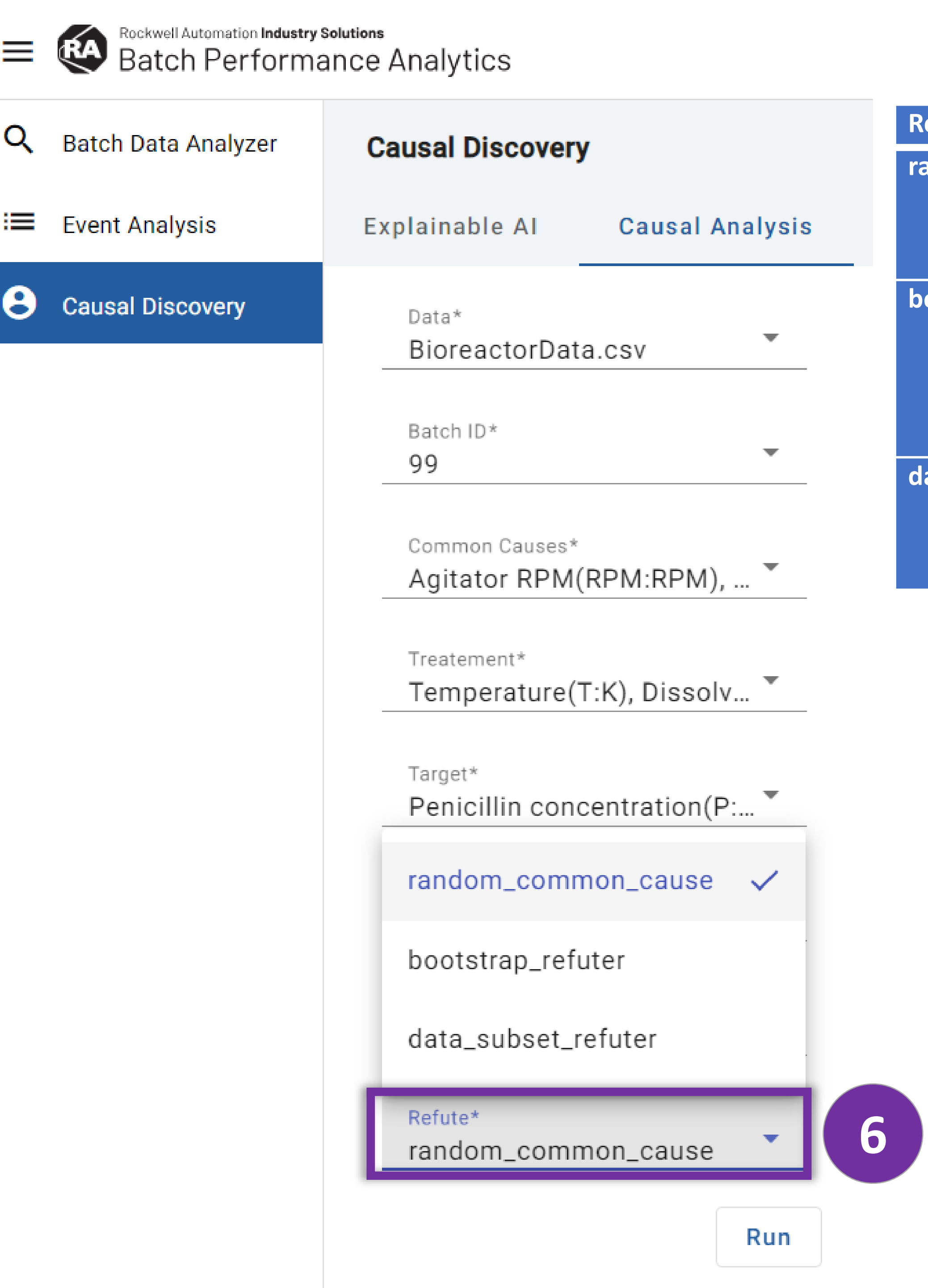
Learning method

- The learning technique used to estimate the relationship among the selected variables in **1, 2,** and **3.**

- Default: backdoor.linear_regression
- Options backdoor.distance_matching, backdoor.propensity_score_stratification, and backdoor.propensity_score_weighting can ONLY handle a single treatment variable and the selected treatment must be a binary variable.

- Single selection

- E.g., Selection: backdoor.linear_regression

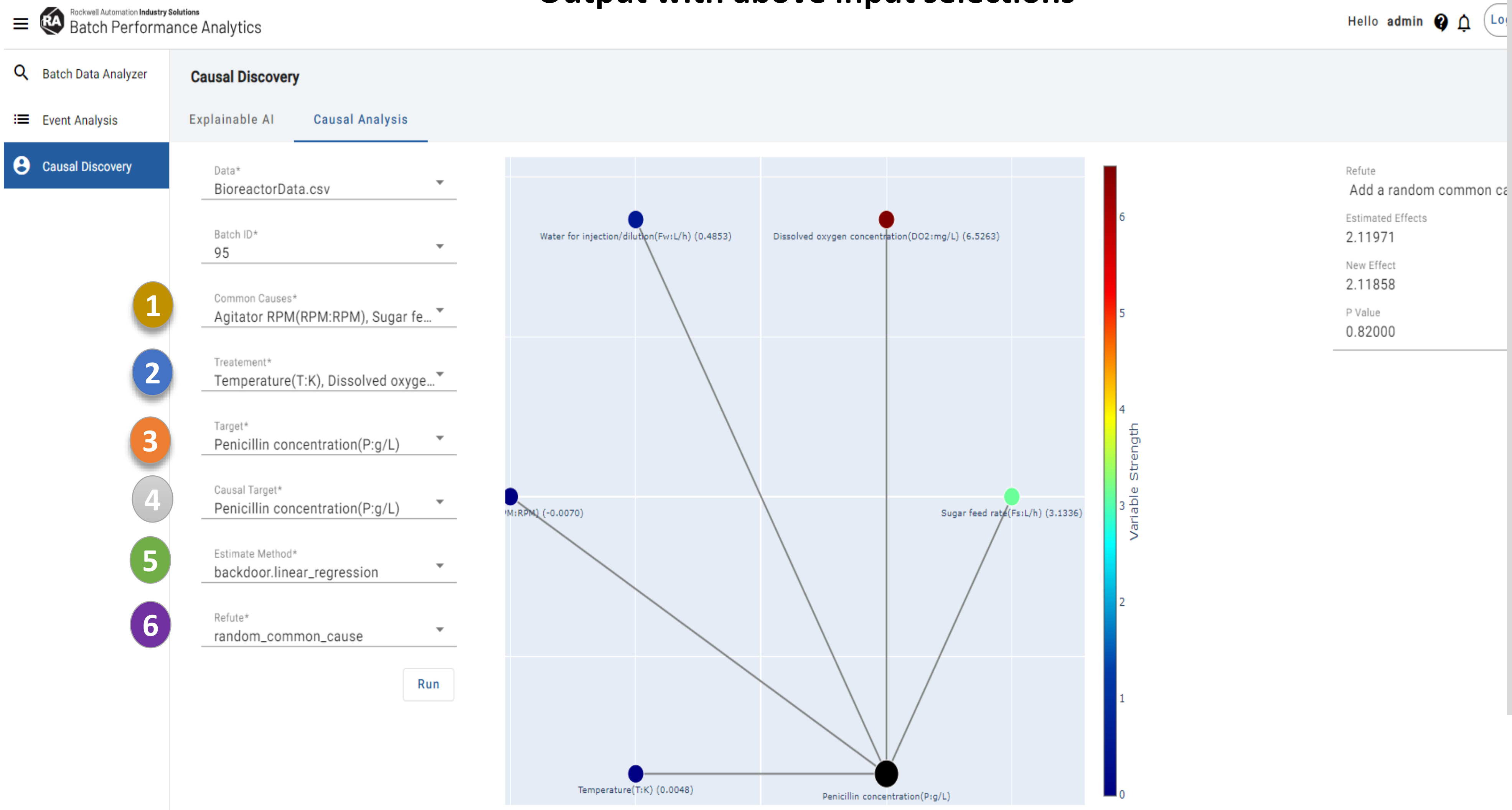


Refute methods	Description	Evaluation (estimated effect vs new effect)
random_common_cause	Adds randomly generated covariates to the data and reruns the analysis to see if the causal estimate changes or not.	The **estimated effect** should not differ significantly from the **new effect** because of a random variable.
bootstrap_refuter	Does the estimated effect change significantly when we replace the given dataset with bootstrapped samples from the same dataset	The **estimated effect** should not differ significantly from the **new effect**
data_subset_refuter	Similar to cross-validation, it creates subsets of data and measures if our causal estimates vary across subsets.	The **estimated effect** should not differ significantly from the **new effect**.

Robustness

- Evaluate or test causal assumptions based on the user input causal graph and assess the validity of the causal relationships inferred from the data
- Single selection

Output with above input selections



- **p-value measures whether the new effect is significantly different from the Estimated effect. In our example, there is no significant difference as denoted by the p-value, which implies that the estimator does not fail this test.**
- **A p-value > 0.05 implies that the estimator is robust.**

For example, with penicillin production data, we observe that the direct influence from Dissolved Oxygen concentration (DO2) to Causal Target (Penicillin Concentration) (~6.3932) is stronger (by ~2 times) than the direct influence from Sugar Feed rate to causal target (Penicillin concentration) (~3.3932.) Roughly speaking, "removing" the arrow from Dissolved Oxygen concentration (DO2) to Penicillin Concentration increases the variance of Penicillin Concentration by ~6.3932 units whereas removing Sugar Feed rate → Penicillin Concentration increases the variance of Penicillin Concentration by ~3.1827units