

Manual: 7.1.2. Process of Modeling

Going from a dataset to a finished model is a process involving several steps. Some of these steps are automated in the software tools and thus invisible to the user but we describe them here for understanding.

First, we must choose the formula that will most likely be capable of representing the data we have. If we are dealing with an ideal gas, for example, then the ideal gas law will be a good choice. If the gas is not ideal, then this law will only work in approximation. We must be aware of the inherent limitations in the formula we initially choose. Here it is important to note that neural networks can be proven to be capable of representing any dataset with great accuracy. This is why we choose neural networks as the formula of choice.

Second, we choose the data that the model will be trained on. For this to make sense, we must understand that the parameters in a mathematical model are not known at this point. Going back to the ideal gas law, we would not know the value of the ideal gas constant R . To determine the value of the parameter, we need historical data for all variables in the formula. Computing the best value for the parameter(s) of the formula given the historical data for the variables is called training the model.

Third, we check our work. It is common to divide a dataset into two components. One is used to train the model. The other is used to verify that the model gives sensible results even on data that has not been used to train it. This is called the validation data. If the parameter values determined previously give accurate results for the validation data, we may consider the model good and the modeling process finished.