

# Curve Fitting Best Practice

## Part 1: How to get maximum value from your curve fitting data

Extracting the best results from experimental data during curve fitting and subsequent analysis can seem like a hit and miss process, as factors such as outlying data points, poorly fitting curves and incomplete data sets hamper efforts to achieve accuracy. By understanding the quality of the data and optimizing processes before fitting analysis begins, researchers can improve results accuracy and generate maximum value from their data.

### Preparing your data

Pre-analysis checks help to ensure you have a suitable data set for the curve fitting process.

#### *Generate as much data as possible*

The greater the number of data points that you have in your data set, the greater the confidence in your results data will be. Researchers should ensure that sufficient data points are measured to deliver a good distribution of data, and that if outliers are knocked out, enough data points remain to achieve a meaningful result.

#### *Ensure data is complete*

Before starting the fitting and analysis process, you should ensure that you have a complete data set, which greatly improves the accuracy of your fitting results. A complete data set includes data for each area of the curve to which you will fit your data, ensuring that from a data perspective your curve is well defined. For example, a dose response data set should contain a well defined minimum and maximum, as well as sufficient data points to construct the center of the curve appropriately.

#### *View data in a scatter plot*

Viewing your raw data in a standard X, Y (or X, Y, Z scatter plot) can give you to an instinctive 'feel' for the quality of the data. Clarifying the relationships within the data, an X Y scatter plot immediately indicates whether you have a good range of data. The relationship between the X and Y values also gives insight into the natural 'shape' of the data, for example, whether it's linear, sigmoidal, exponential, concave, convex or bell-shaped.

#### *Enable direct data range comparisons with data normalization*

Data normalization reduces 'noise' in the data and can be used to bring all measurements into a comparable defined range while maintaining its variability. Normalizing data enables direct comparisons between raw data set values that may have been measured across different ranges, for example, a curve that contains CPM values of between 1,000 to 22,000 can be directly compared to another curve with values of 12,000 to 44,000.

In an experimental context, normalizations are used to standardize micro-array data to enable differentiation between real (biological) variations in levels and variations due to the measurement process.

## Why fits fail

Curve fitting, as any mathematical technique, is prone to failure in certain scenarios. IDBS' analysis products, including XLfit®, BioBook® and ActivityBase XE® can be used to both visualize the reasons for failure, described below, and to rectify errors (see part 2 in this best practice series, Resolving Fitting Issues.)

### Not enough data

If you haven't measured enough data points, or have knocked out too many data points, you may have too few remaining data points to get a meaningful fit result.

In all cases for a fitting process, the more data you have, the better the fit quality, and there needs to be a minimum number of data points in the data set to generate a reasonable quality fit.

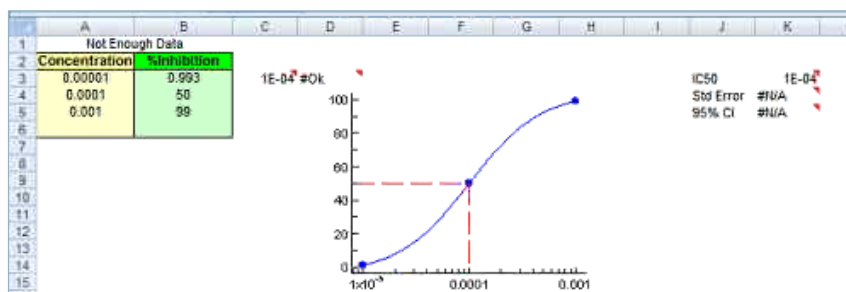


Fig 1: An XLfit graph with insufficient data points to generate error and 95% confidence interval values. Varying different parameters will result in an infinite number of solutions through the three data points which means results are meaningless.

### Poorly defined data

To ensure accurate fitting results, a data set should include data points for all areas of your curve, e.g., for a sigmoidal curve include the lower, mid and upper ranges of the fit.

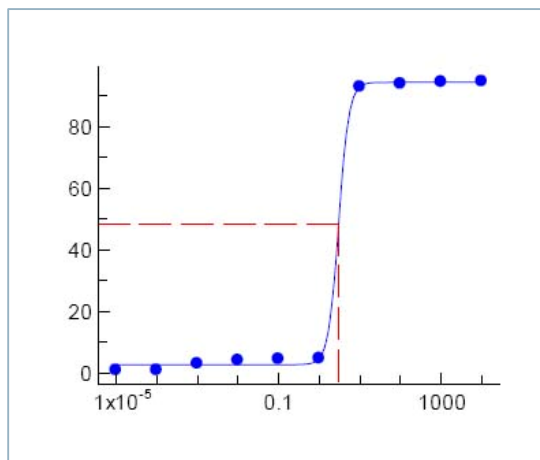


Fig 2: Although the above dose response curve has a well defined lower asymptote (the minimum value), and upper asymptote (the maximum value) with a reasonable number of data points in both, there are no data points in the mid range, resulting in high error values and poor quality IC<sub>50</sub>/EC<sub>50</sub> values.

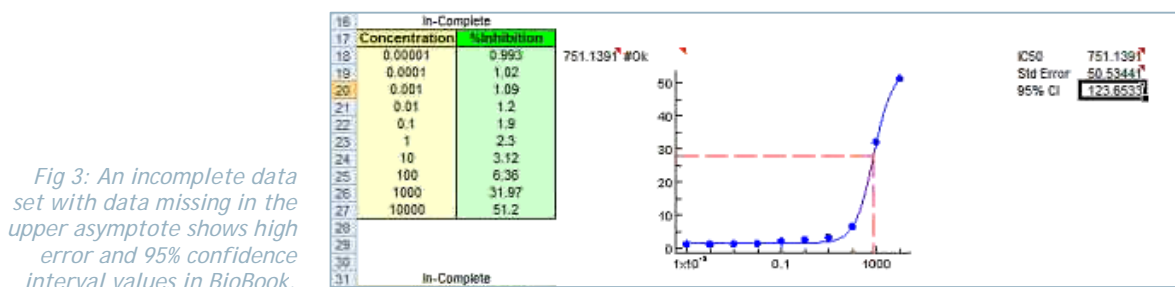


Fig 3: An incomplete data set with data missing in the upper asymptote shows high error and 95% confidence interval values in BioBook.

## Curve Fitting Best Practice Part 1

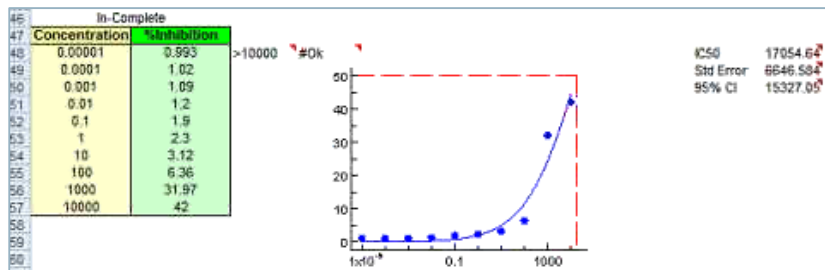


Fig 4: An incomplete data set where the result lies outside the data range shows high error and 95% confidence interval values in BioBook.

### Too many outliers

If your data contains a lot of scatter with a large number of potentially erroneous outlying data points, fitting results will be of a poor quality. These outliers can vary the result significantly. While you can improve the fit quality by knocking out poor values, if you remove a large number of these outliers to the extent that too few data points remain for a reasonable fit, the results will be meaningless.

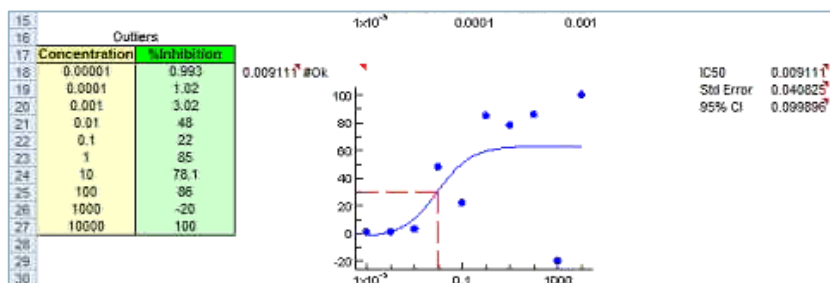
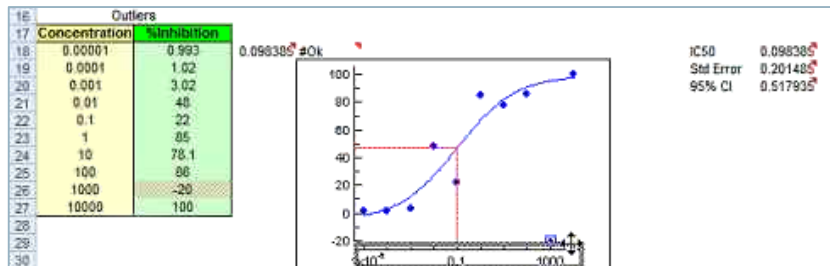


Fig 5: A high level of outliers: knocking out outliers in this case does not decrease the error value or improve fit quality significantly



### Using the wrong model

A model defines the way in which we describe a data set. Before generating experimental data you should know what measurements you want to take and the curve that best represents your data to enable data analysis. However, the model that you are using could be inappropriate, preventing you from achieving the fitting results you require.

### 'Division by zero' errors

Because some model definitions have divisory factors by concentration or X value, if you have a zero X value in your data set, the model will not produce a fitting result.

### The model may have too many parameters

A model defines Y as a function of X along with one or more parameters, which can be varied as part of the fitting process. For example, when fitting your data to a four-parameter model, you can essentially define the complete shape of the curve by altering parameters A, B, C and D and finding the best curve fit for your data.

However, if the model you choose has too many parameters for which you don't have enough data, your final fitting results could be infinite in certain scenarios and as such are meaningless.

### *Poor starting parameter values*

Fitting requires a set of starting parameter values to begin the process. Poor starting parameter values are one of the primary reasons why fits fail.

Because starting parameter values affect the quality of the fit and whether or not it will converge, it is important to assign reasonable start points.

### *Not enough fitting process iterations*

Beginning with a set of starting parameter values for a given model, a typical fitting process alters each parameter value in an iterative process to change the shape of the curve in each successive cycle. When the fitting converges according to convergence criteria, or a set number of iterations are completed, the fitting process should end with what is called a converged fit. You can then extract parameter values from the fit for its final iteration, which provide the required results for the data set. However, in some cases the fit may fail to converge if there are insufficient iterations to reach convergence.

## Summary

Using the best practice guidelines above, you can identify the causes of poor fits and use the information to produce more robust results with greater meaning. By preparing your data properly and analyzing the fit correctly you can have improved confidence that the results produced will be of a significant quality.

In Part 2 we will be discussing the process of QC'ing fits, including an identification process for fixing poor fits. On top of this we will be looking at how these approaches can be strung together to make a robust fitting mechanism capable of high quality consistent results for better decision making.