



UNIVERSITY OF
BATH

Department of
Life Sciences

Non-linear regression (least-squares fitting) of assay data with SciDAVis

Dr Charlotte Dodson

Name.....

Reminder of scientific notation

Unit	Exponential form	Typed form of exponential (for Excel and SciDAVis)	Unit relationship to original unit
M	$\times 1$		
mM	$\times 10^{-3}$	1e-03	$1 \text{ mM} = \frac{1}{1,000} \text{ M}$
μM	$\times 10^{-6}$	1e-06	$1 \mu\text{M} = \frac{1}{1,000,000} \text{ M}$
nM	$\times 10^{-9}$	1e-09	$1 \text{ nM} = \frac{1}{1,000,000,000} \text{ M}$
pM	$\times 10^{-12}$	1e-12	$1 \text{ pM} = \frac{1}{1,000,000,000,000} \text{ M}$

SciDAVis workshop

SciDAVis is a data fitting software package used to fit physically meaningful equations to experimental (assay) data. Fitting data in this way enables us to extract parameters such as IC_{50} values and thus to compare the efficacy of different compounds in an experimental assay in a quantitative manner. SciDAVis is free to use and will carry out the basic analyses we need here adequately. Commonly used pay-for software packages which carry out a similar function (and which are more powerful) include GraphPad Prism, SigmaPlot and Origin.

In this workshop we will cover the main principles of non-linear regression (least-squares data fitting). We will apply these principles to some simulated experimental data and fit the data in SciDAVis to an appropriate equation in order to extract IC_{50} values. The aim of the workshop is to enable you to become comfortable using SciDAVis. This will be useful next semester when we come to the virtual drug discovery exercise.

Experimental background

Let's imagine that we are part of a drug discovery team working on new inhibitors for a kinase target. The Medicinal Chemistry team has synthesised some new compounds and we have carried out some activity assays in order to determine the IC_{50} values.

Our assay is designed with a fluorescence read-out where fluorescence \propto [ADP]. ADP is a product of the kinase phosphorylation reaction, so high fluorescence indicates an active (poorly inhibited) kinase.



In our experiment, we have measured the activity of the kinase in the presence of different concentrations of our inhibitor. We have also measured the activity of the uninhibited enzyme (negative control) and the activity of the enzyme in the presence of saturating quantities of a known inhibitor (positive control). We started all our experiments at the same time, let them all proceed for our standard assay time (eg 1 hr), and measured the fluorescence for each condition at the end. We are good experimentalists, and so have carried out our experiments in triplicate.

At the end of all of this we have lots of data and would like to know the IC_{50} values of each of our compounds.

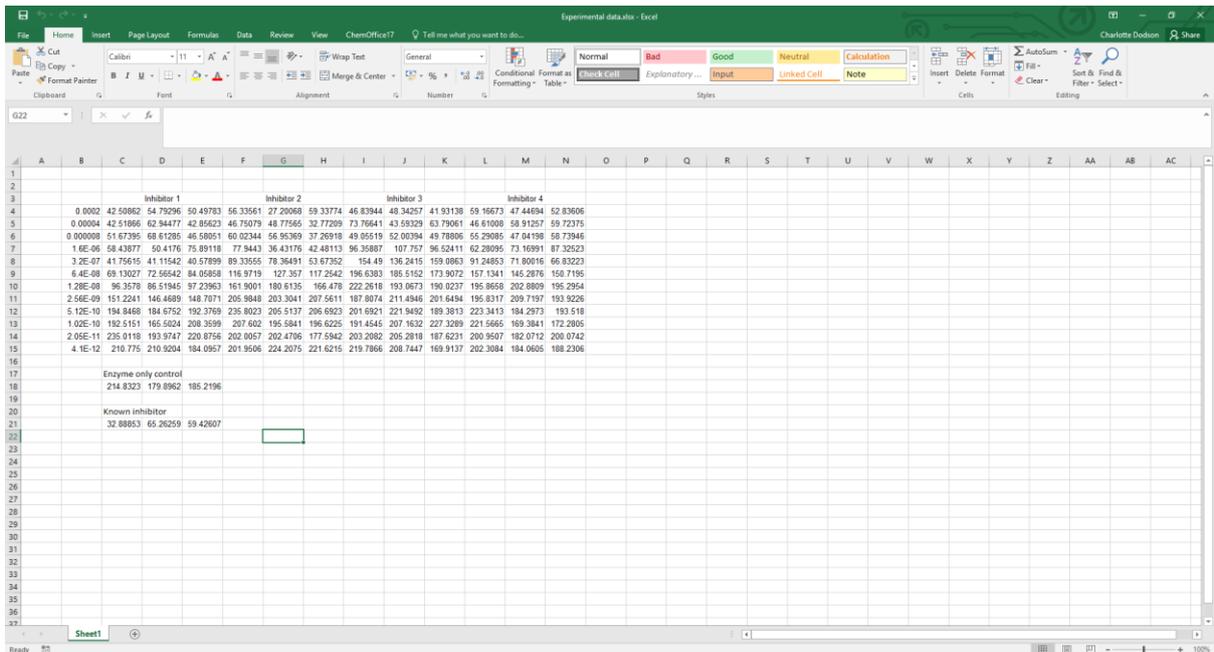
Software requirements

This tutorial assumes that you have SciDAVis installed on your computer (if you need to install SciDAVis, please see p20). It also uses MScExcel to prepare experimental data for fitting in SciDAVis.

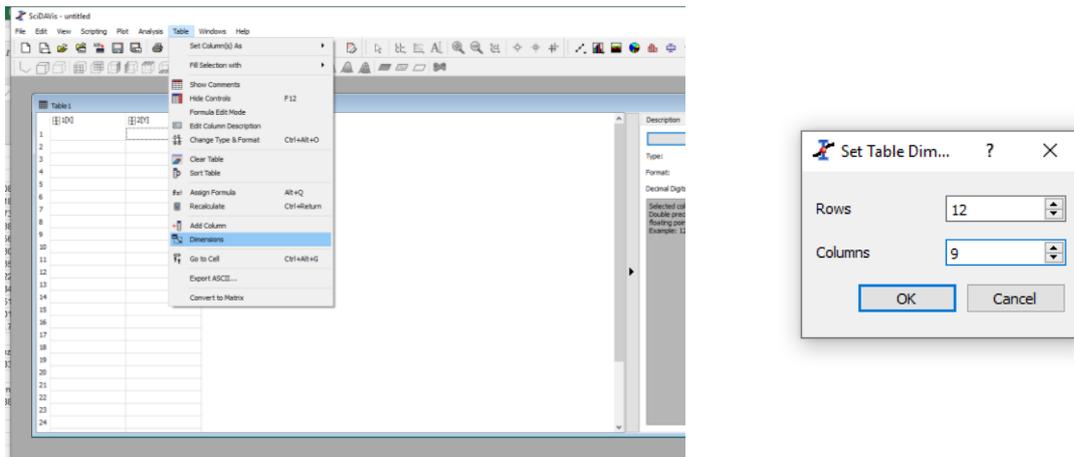
Downloading your data

Download the file *Experimental data.xlsx* or *Experimental data.csv* from the NTF and save the file somewhere where you will be able to find it.

Open *Excel* on your computer, and then open the downloaded data. You will see that the data is in the form of a table. The concentration of each inhibitor is in the left-hand column (column B). The data is then grouped with each set of three columns providing the experimental data for a different inhibitor.



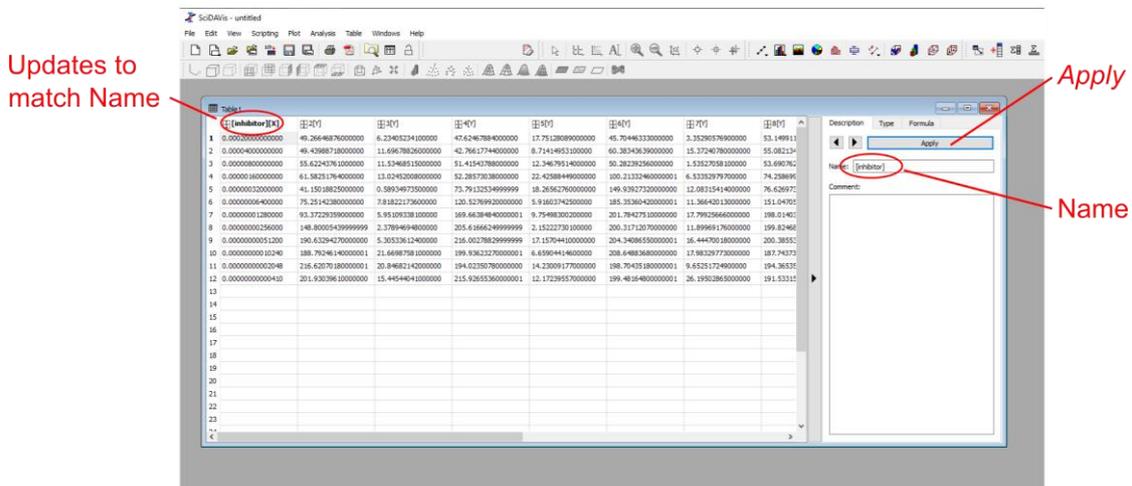
Go to Table -> Dimensions and set the table dimensions to 12 rows and 9 columns.



You are going to need to move your x-values (column B) and your average and standard deviation values from Excel into SciDAVis using Copy and Paste. This is a bit laborious, but I haven't found a better way of doing it. When pasting, make sure that you put the x-values (from column B in your Excel sheet) into the column labelled 1[X] in SciDAVis.

Renaming columns and setting column data type

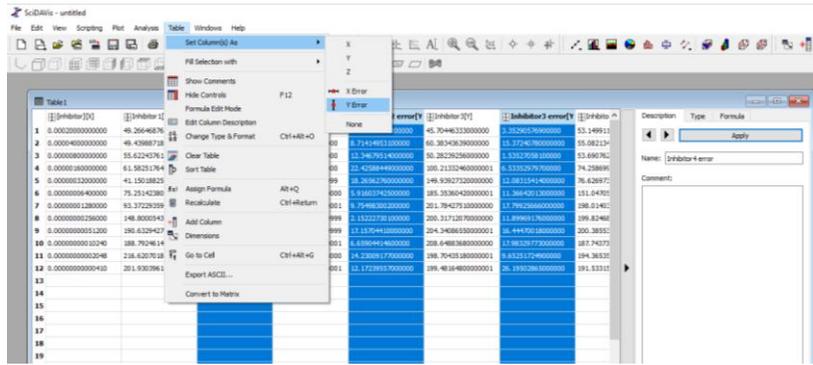
In the right-hand side of your SciDAVis window, click on the *Description* tab. In the *Name* field it will have a number (1 if you are in column 1). Rename column 1 to [inhibitor] and press *Apply*. You will see that the column name has changed in the left hand side of the Table1 window.



Rename your columns so that you know which data is in each column. You need to press *Apply* after each change that you make. You can scroll through all nine columns in your table by pressing the forward and back arrows next to the *Apply* button.

Some of our columns contain data to be plotted on the y-axis of our graph (those with averages from Excel), but around half of them contain error values (those with standard deviation values). We need to specific to SciDAVis which columns are which.

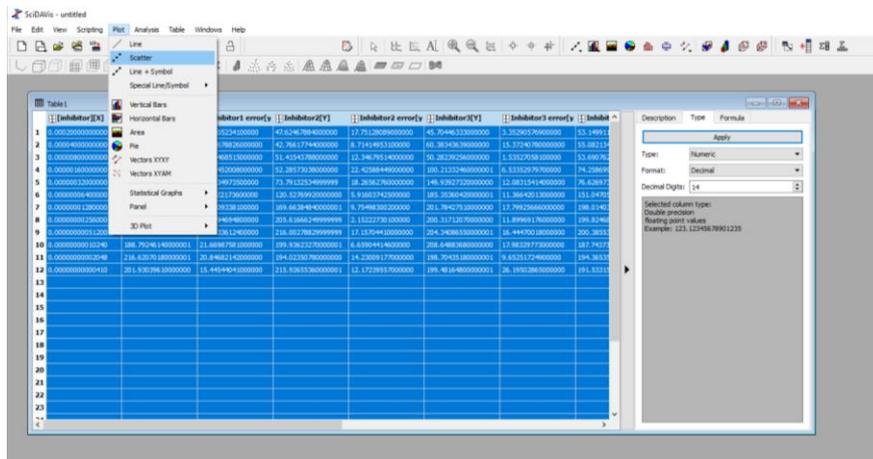
To do this, select your standard deviation columns. You can select multiple columns by holding down the Ctrl key while clicking on the column headers. Then select Table -> Set Column(s) As -> Y Error.



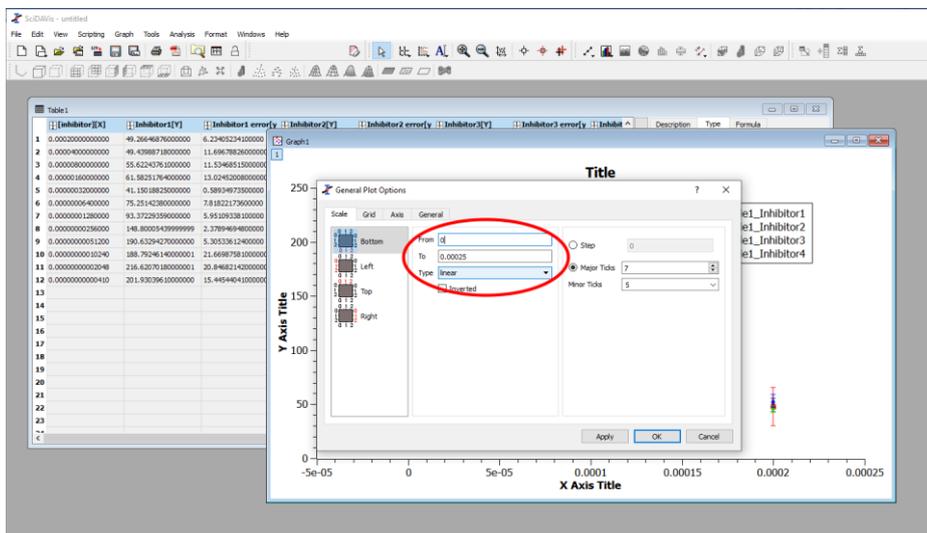
We are now ready to plot the data.

Plotting a scatter graph

Select all the data in your table. Go to Plot -> Scatter. A new window should appear within your SciDAVis window called Graph1. We are going to work in this window, so make it a bit bigger.

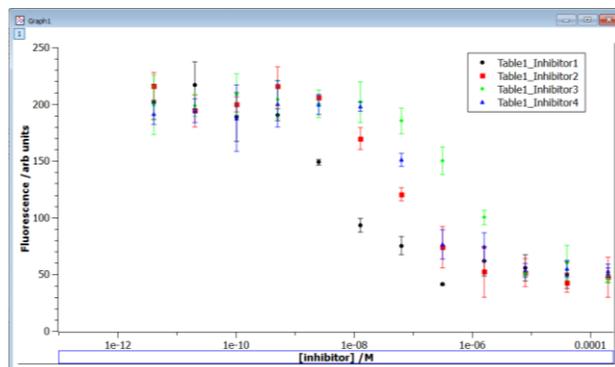


At present, the data in your graph is plotted on a linear x-axis. IC₅₀ graphs are always of activity against $\log([\text{inhibitor}])$, and so we need to change our x-axis to a log scale. Double click on the x-axis to bring up a dialogue box and change the axis type from *linear* to *logarithmic*. Change the limits of the axis (from and to values) to encompass the range of x-data that you have (eg from 1e-13 to 0.00025) and click OK.



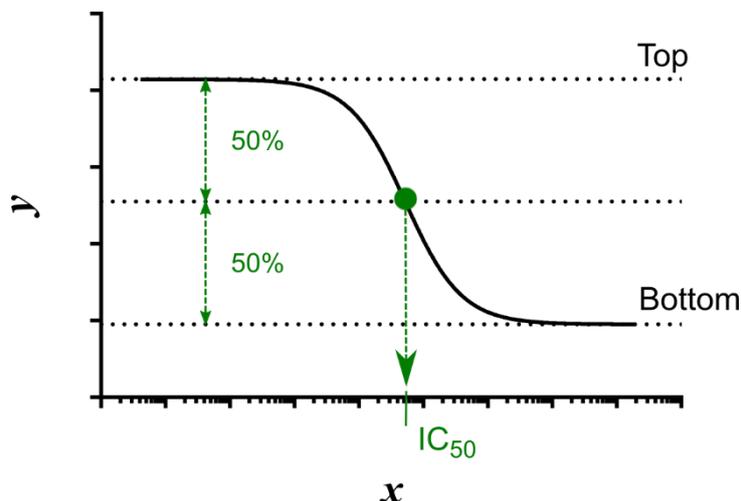
Your graphs should now be roughly S-shaped, decreasing in value as the inhibitor concentration increases. Double click on your axis titles and rename these (NB your inhibitor concentrations are in M, and your fluorescence values are in arbitrary units). If the graph legend is on top of your graph lines, move it.

Your graph might look something like this



What equation are we going to use to fit our data?

IC₅₀ data is a sigmoid curve (the technical term for an S-shaped curve) like that shown below.



The analytical expression (the maths) which describes this is as follows:

$$y = Bottom + \frac{(Top - Bottom)}{1 + \frac{x}{IC_{50}}}$$

In this expression there are three parameters (other than x and y): $Bottom$, Top and IC_{50} . The IC_{50} value is the value of x for which y is exactly half-way up the curve. Our curve has been defined with flat baselines (*ie* there is no slope to the top or bottom of the S). Top is the value of the pre-transition baseline (the y -value of the curve before the curvy part of the S). $Bottom$ is the value of the post-transition baseline (the y -value of the curve after the curvy part of the S).

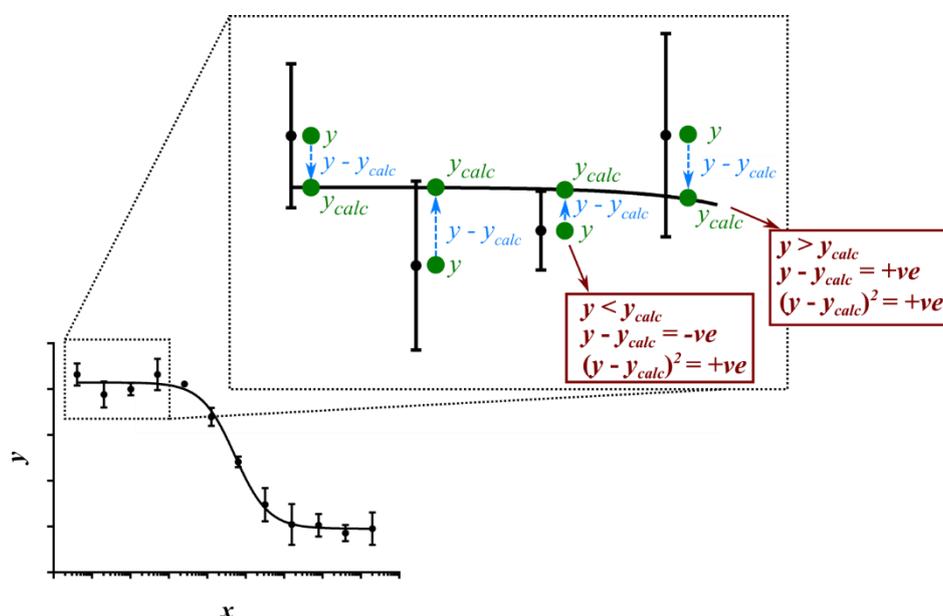
When we are fitting our data, essentially we give the computer values of x and y (our experimental data) and ask it to find the values of Top , $Bottom$ and IC_{50} which best describe our experimental curve. We know that our data has experimental noise (we have error bars and there is scatter), so we are asking the computer to simulate lots and lots of curves (possibly as many as 1000 curves) until it finds the one which has the correct mathematical form and goes as closely as possible to all the data points.

How do we define 'as close as possible to all points'? We're doing this on a computer, so we do it in a quantitative manner with maths.

Least-squares analysis

Whatever line-fitting we are doing, the definition of a good fit (or not) is by least-squares analysis.

We start the process off by giving the computer the equation we want to use and some guesses for the values we don't know (*ie* some guesses for the values of *Top*, *Bottom* and *IC₅₀*). The computer uses the equation to calculate a value of *y* (*y_{calc}*) for each value of *x* in our dataset. It then goes through each data point in turn and subtracts the value of the calculated *y*-value from the experimental one (*ie* it calculates $y - y_{calc}$). In some cases *y_{calc}* will be smaller than *y*, in some cases it will be larger. To stop these two cases cancelling out in the maths later on, the computer then squares the answer (mathematically, what was previously a list of positive and negative numbers is now a list of positive ones).



If the calculated curve is a good fit to the data, the values of *y_{calc}* will be similar to those for *y*. This means that the list of numbers calculated by the computer is a list of small numbers. If the calculated curve is a bad fit to the data, the values of *y_{calc}* will be different to those for *y*. The list of numbers calculated by the computer will therefore be a list of large numbers. In order to make a quantitative assessment of how good the fit is, the computer simply adds up all the numbers in its list.

In maths, the computer computes a parameter (*Total*) and then minimises the value of this parameter.

$$Total = \sum_i (y_i - y_{i_{calc}})^2$$

The computer then makes a change to the value of the parameters *Top*, *Bottom* and *IC₅₀* and repeats the whole procedure. If the total at the end is smaller than the first total (*ie* if the curve is a better fit), it keeps the new set of numbers. If the total at the end is larger, it goes back to the previous set. The computer keeps changing the values and repeating the adding up until either it gets stuck (it returns an error), or it finds a set of parameters which is a very good fit to the data. In the latter

case, it returns the fitted values to the user, together with an estimate of how precisely it can determine each value (*ie* together with an error value).

Least-squares analysis with error bars

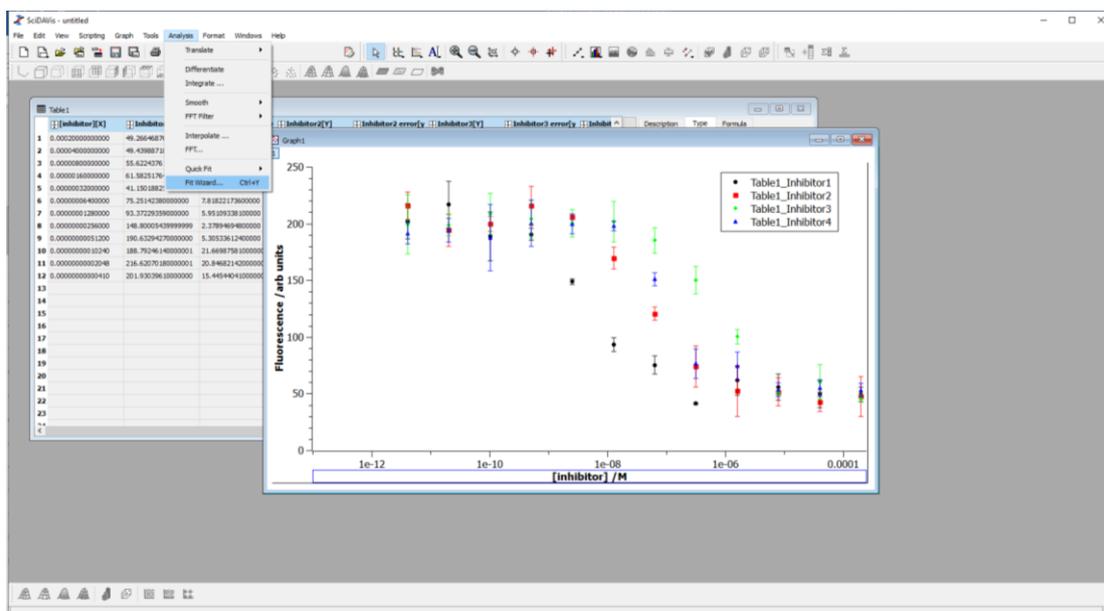
When we are fitting experimental data and we know the error on each measurement, we want to take the error into account when fitting our data. *ie* if we know the value of some of our data points accurately, we want to give these data more weight in our calculation. If we know that the error on our data points is high, we want to give these data less weight in our calculation.

In order to do this mathematically, we include the error value for each point in our calculation of *Total* (we use the error as a weighting factor). There are several ways of doing this. In SciDAVis, we define our error values in one column and tell the program to use the square of this value in our calculation:

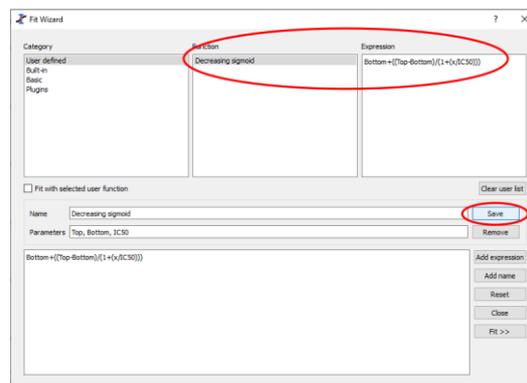
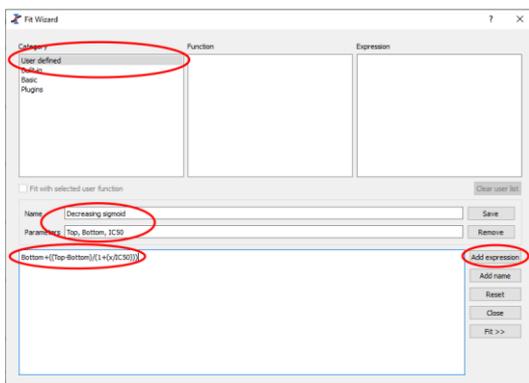
$$Total = \sum_i \frac{(y_i - y_{i_{calc}})^2}{error^2}$$

Entering an equation to fit our data

In order to fit our data, go to Analysis -> Fit Wizard...



This brings up a new dialogue box. We are going to create a new equation and use this to fit our data.



First click on *User defined*. This means that we are going to type our equation into SciDAVis. Next click in the *Name* box and enter a name for your equation. I have called my equation *Decreasing sigmoid*.

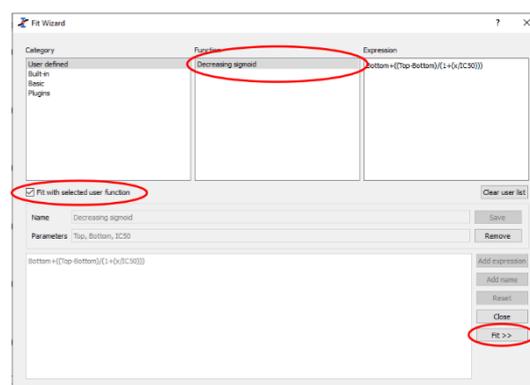
Now click in the *Parameters* box. This is where we need to tell SciDAVis the names of every parameter we are going to use (other than *x* and *y*). Type in all the parameters (in any order) separated by commas. I typed *Top, Bottom, IC50*.

We now need to tell SciDAVis exactly what equation to use to fit the data. To do this, click on *Add expression*, and then type in the large white box. We need to type the right hand side of our equation (everything in the equation on p9 after the = sign). **When typing the equation, it is very important to make sure that the brackets are correct.** I typed $Bottom+(Top-Bottom)/(1+(x/IC50))$.

Finally, we would like SciDAVis to save our equation so that we can use it another time. This isn't so important for this workshop, but if we are going to be fitting lots of IC₅₀ data on different days it saves a lot of typing (and reduces the chances of mis-typing the equation).

To save our equation, click on *Save*. Notice that the equation and its name have now appeared in the two white boxes at the top of the screen.

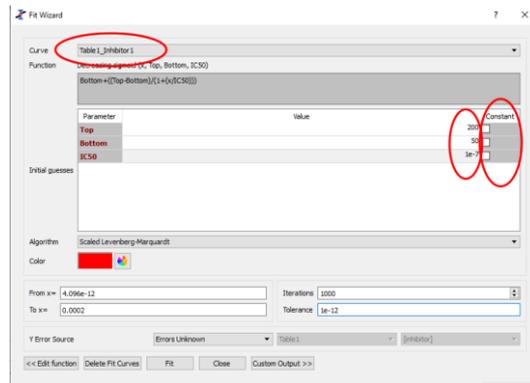
In order to start the fitting procedure, click on your equation in the *Function* box (top middle white box), select the *Fit with selected user function* box and click *Fit >>*.



Fitting our data

If we have saved our equation we only need to enter it once (ever). Every other time we want to fit our data to this equation we go to Analysis -> Fit wizard..., select our equation, click *Fit with selected user function* box and click *Fit >>*.

Once we click *Fit>>* we have a new dialogue box. This is where we tell SciDAVis our initial guess parameters for our data. We need to enter these values in the main box in the centre of the window and we will need to enter new values for each curve we are fitting.



Set the curve to be fit (*Curve* box at the top of the screen) to your first set of data.

Move the dialogue box on your screen until you can see your graph. Have a look at the graph and decide what would be a sensible guess for your values of Top, Bottom and IC50. These values don't need to be very accurate (perhaps ± 50 for Top and Bottom and $\pm 1e01$ for IC50).

Make sure that the check boxes in the *Constant* column are unchecked (we would like SciDAVis to find the best fit value for each of the parameters, not accept the values as a user-set constant which is not changed).

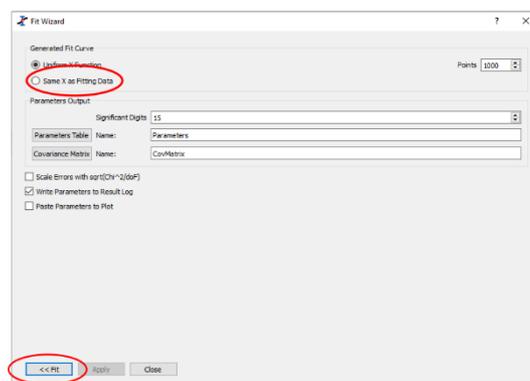
The next few instructions are just to make sure that your graph looks good. They do not change the numerical output of the fitting algorithm at all.

Click on the *Color* box. This is going to be the colour of the line which is plotted on your graph. You can make this any colour you want, but is usually best if it is the same colour as your data points. Select the colour you would like SciDAVis to use for your fit line.

We are going to make sure that the curve plotted on your graph is plotted with a suitable number of points. Click on *Custom Output >>* (at the bottom of the dialogue box).

In the *Generated Fit Curve* box make sure that *Same X as Fitting Data* is selected. What this means is that when your fitted data curve is drawn on the graph, SciDAVis will calculate a point at the same values of x as your input data and then join these points together to give your final fit curve. Note that this does not change the calculated best fit values of the parameters, *only how the line looks on your graph*.

(If *Uniform X Function* is selected, SciDAVis will calculate up to 10^6 values of x at equal distances between the max and min values of your x -axis on a linear scale. Since your data goes from $\sim 10^{-12}$ to $\sim 10^{-2}$ (ie 10 orders of magnitude) this means that the first point will be at $\sim 10^{-12}$ and the next one at $\sim 10^{-8}$. The two points will be joined with a straight line which will not go through your data at all. This will make it almost impossible to do a manual check on whether your data fit is any good.)



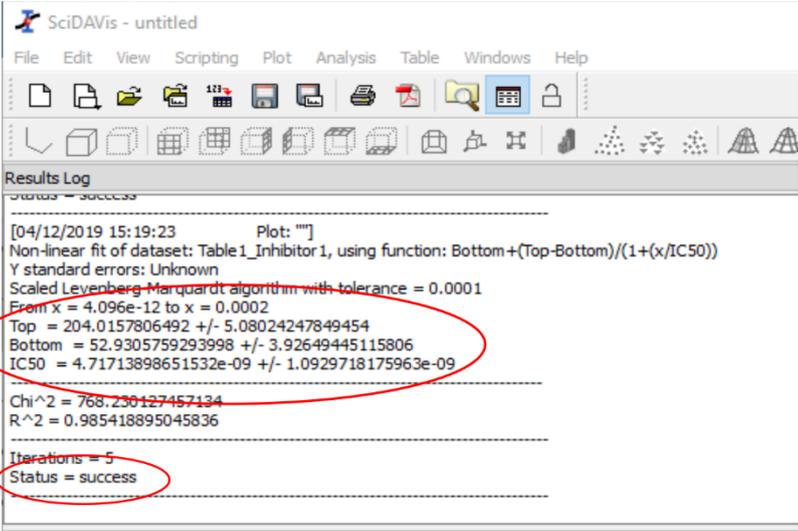
Click *<<Fit* to return to the previous dialogue box.

You are now ready to fit your first data set. Click *Fit* and see what happens.

Looking at the output of a fit

Three things should have changed on your screen.

- 1) In the *Fit Wizard* dialogue box the numbers next to Top, Bottom and IC50 will have changed. The values they have changed to will be the final values that SciDAVis has calculated. If the fit is successful, they will be the best fit parameters. If the fit is unsuccessful, they will be the most recent values that SciDAVis has tried.
- 2) A *Results Log* panel will have appeared towards the top of the SciDAVis window.



```
SciDAVis - untitled
File Edit View Scripting Plot Analysis Table Windows Help
Results Log
-----
[04/12/2019 15:19:23 Plot: ""]
Non-linear fit of dataset: Table1_Inhibitor 1, using function: Bottom+(Top-Bottom)/(1+(x/IC50))
Y standard errors: Unknown
Scaled Levenberg-Marquardt algorithm with tolerance = 0.0001
From x = 4.096e-12 to x = 0.0002
Top = 204.0157806492 +/- 5.08024247849454
Bottom = 52.9305759293998 +/- 3.92649445115806
IC50 = 4.71713898651532e-09 +/- 1.0929718175963e-09
-----
Chi^2 = 768.230127457134
R^2 = 0.985418895045836
-----
Iterations = 5
Status = success
-----
```

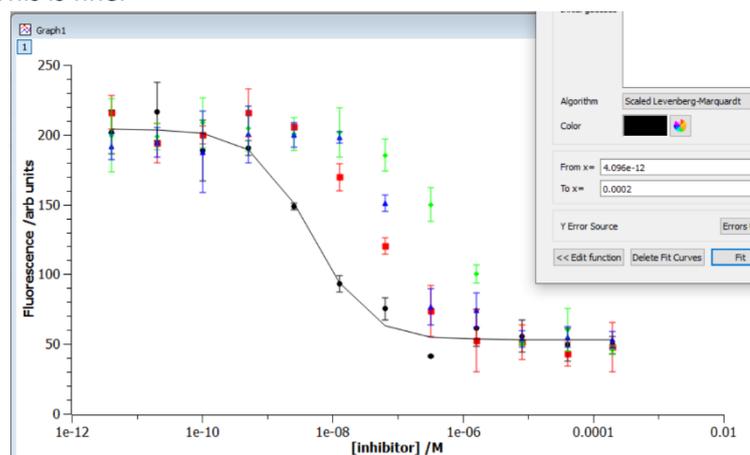
This panel logs everything that SciDAVis does (or tries to do) with your data fitting. The most important piece of information is the one at the bottom. If *Status = success* then all the maths has worked well in fitting your data. **If *Status* is not *success* then you cannot trust any of the numbers you are given by the program. SciDAVis has got stuck in fitting your data** (usually because the guess you gave it wasn't close enough to the true value). In this case, go back to the *Fit Wizard* dialogue box and try some different initial guesses.

If *Status = success*, then the next thing to do is record your best fit parameters. These are given a little above the *Status* information in the form Parameter = value +/- fitting_error. When you record your best fit parameters **always quote the error** associated with them. **Also think about the precision of your values.** There is usually no point in quoting your error to more than one significant figure, and no point in quoting your fitted value to more precision than your error.

In the example above, I would quote my fitted parameters as:

Top = 204 ± 5 arb units
Bottom = 53 ± 4 arb units
IC50 = 5 ± 1 × 10⁻⁹ M = 5 ± 1 nM

- 3) If the fit is successful, a best fit line will have appeared on your graph. It won't be completely smooth because the line shown is joining calculated data points at each value of x in your original data. This is fine.



The first thing you should do is look at your best fit line. Does it go through most of your data points? **If not, stop.** Why not? Has SciDAVis found something that is mathematically right, but physically unrealistic (perhaps because your initial guesses were wrong or the tolerance of the fit is set too high)? **Never accept the results of a data fit where the best fit line does not look as though it is fitting your data. If it looks wrong, something will be wrong.**

If your best fit line goes through your experimental data points then be pleased! You have successfully fit your data and can record your fitted parameters and their errors (from the *Results Log* panel) with confidence.

At this point, you have fit one of your sets of data. Return to p11 and the section *Fitting our data* and repeat the process for the other three curves. Note that you should not need to change the parameters in the *Custom Output >>* box (p13) for the remaining curves (it should default to *Same X as Fitting Data*). When you have finished with the *Fit Wizard* dialogue box, click *Close*.

Question:

What are the IC_{50} values for the four compounds you have been given?

Which inhibitor is most potent?

Saving your data

You have carried out a lot of work on your data and it would be a shame to lose it. Go to File -> Save Project As... and save your Discovery SciDAVis project.

Modifying your graph

The most important thing is to make sure that your data is fit well, but the second most important thing is to present your data well in a graphical form. We have already modified the limits and titles of the x and y axes. Most other aspects of the graph can be modified by double clicking on them.

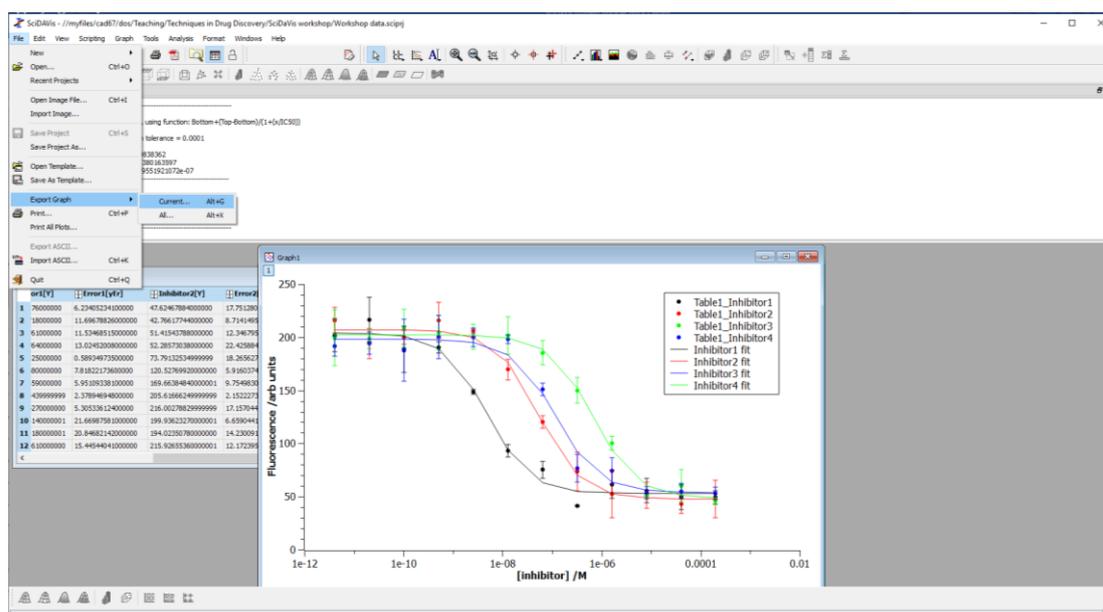
Double clicking on a data set enables you to change (among other things) the symbol used for the data set, the colour of the data set and the colour of the error bars. Slightly annoyingly, you will need to click *Apply* after every change you make.

Have a play with the different options and customise your graph.

Remember that, as a minimum, every scientific graph must have labelled axes (with units) and a key (perhaps in the figure legend) to indicate which data set is which.

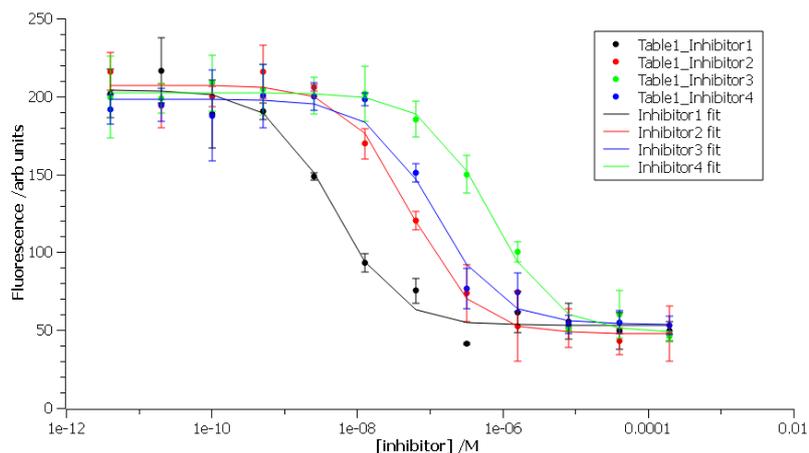
Exporting your graph to use in a report

Now your graph is finalised, we need to export it to use in eg an experimental report. Make sure the graph window is selected and go to File -> Export Graph -> Current...



The default file type is a bitmap image (.bmp). I find bitmap images hard to work with and prefer png files instead. Change *Files of type* to .png and make the image quality as high as possible (under *Advanced*>> if this is not shown). The maximum value of quality is 100. Finally choose a filename and click *Save*.

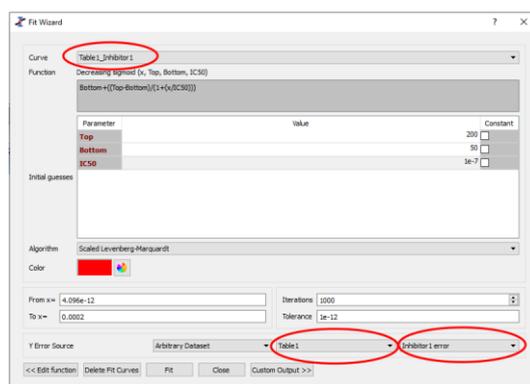
My final graph looks like this:



Including error values

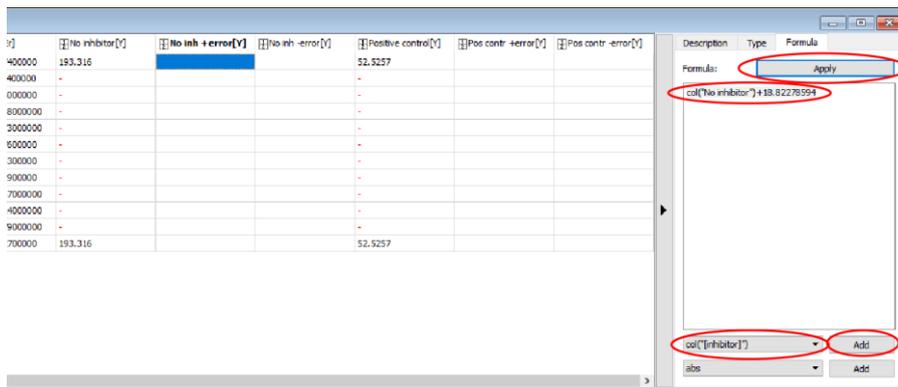
You will notice that we haven't weighted our data points by their error values in fitting our data. Doing a weighted fit is the best practice way of fitting our data, but it turns out that it doesn't always work well in SciDAVis. If you want to explore how to weight your data then return to the Fit Wizard window (Analysis -> Fit Wizard). Select *User defined*, select the equation for your sigmoid curve, check the *Fit with selected user function* box and then click *Fit >>*.

As previously, you will need to enter guess values for Top, Bottom and IC50. In order to weight our data fit by the error on each measurement, click on *Errors unknown* (the value for *Y error source*) and change to *Arbitrary dataset*. We need to tell SciDAVis where to find our error values. Have a look at the data set which is highlighted at the top of the box (in the *Curve* box) and set the Y error source to the matching column. In my case, I am fitting the curve for *Table1_Inhibitor1* so I have set the next two boxes to *Table1* and *Inhibitor1 error* (since this is what I named my error column on p6).



When you have done this, click *Fit* (as you would for a normal fit).

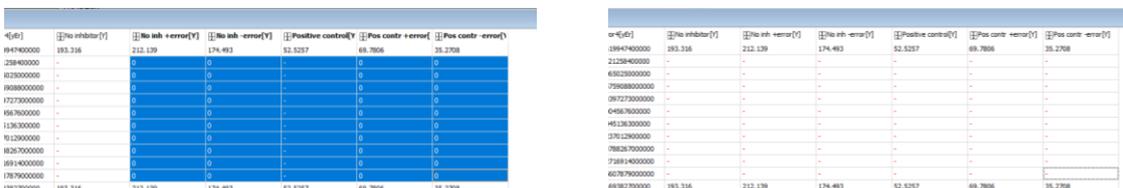
dropdown menu at the bottom of the window and select the column containing your *No inhibitor* values. Click *Add*, and then type + and then the value of the standard deviation you calculated in Excel. My formula ended up looking like: `col("No inhibitor")+18.82278594`. Click *Apply*.



Repeat this for the *No inh -error* column, remembering to subtract the value of the standard deviation.

Now repeat the whole procedure for the *Positive control*.

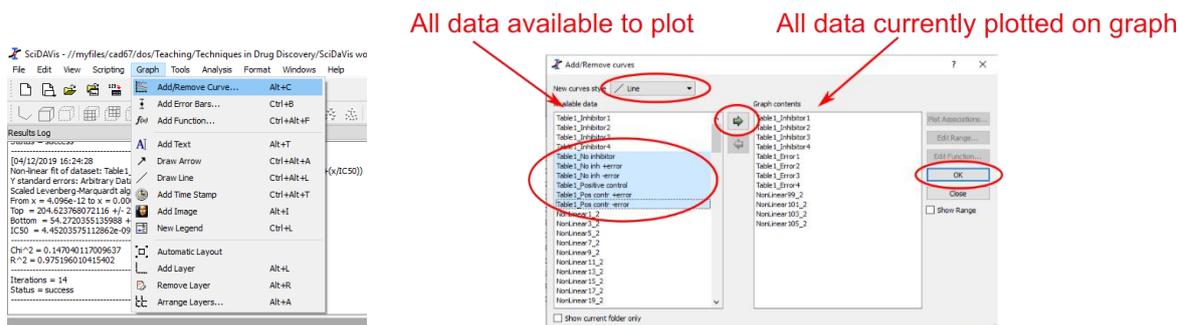
Notice that when calculating formulae SciDAVis has added zeros to the rows where there is no value in the *No inhibitor / Positive control* column. We need to remove these zeros otherwise they will cause problems later on. Select these cells and press the *Delete* key on the keyboard (not backspace). The zeros should all have changed to red dashes.



Plotting positive and negative control values on your graph

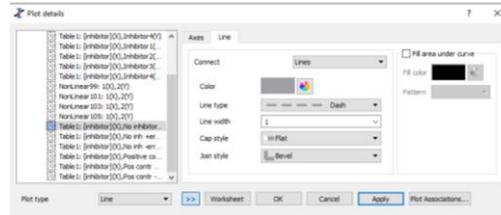
To plot your control values on your graph, click on the graph window. Go to *Graph -> Add/Remove Curve...*. The *Add/Remove curves* dialogue box opens.

The left column of the box lists every set of data that you have (including every curve fit you have carried out). The right column lists all the data that is currently plotted on your graph. The dropdown box at the top enables you to select how new data is plotted on the graph (symbols, joining line etc). Select *Line* in this box, and then select the data columns for your positive and negative controls (and their + / - errors). Click on the arrow pointing right, and then click *OK*.

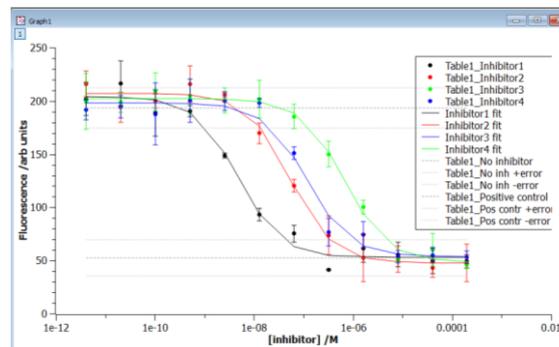


Final formatting of your graph

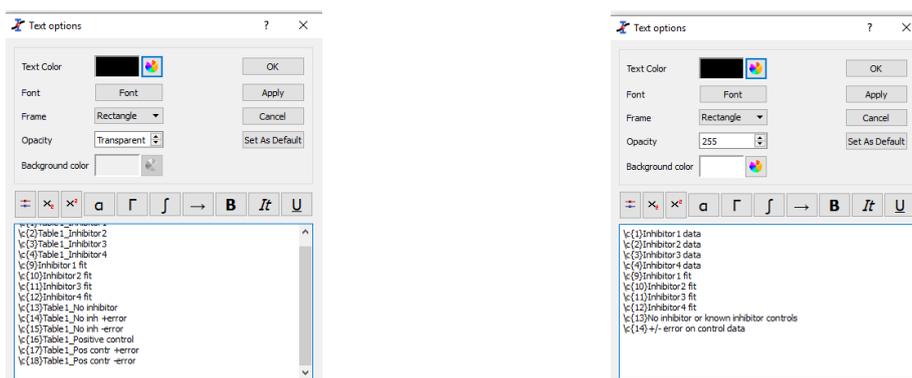
All your data sets have now been added to your graph as straight lines. Your graph probably looks a mess. Double click on one of your straight lines to bring up the *Plot details* dialogue box. The left hand column is a list of all the data sets on your graph. Select one of your straight lines (I have selected *No inhibitor*). You can now adjust the colour and style of line and start to tidy up your graph (I have set the line colour to grey and the line style to a dashed line).



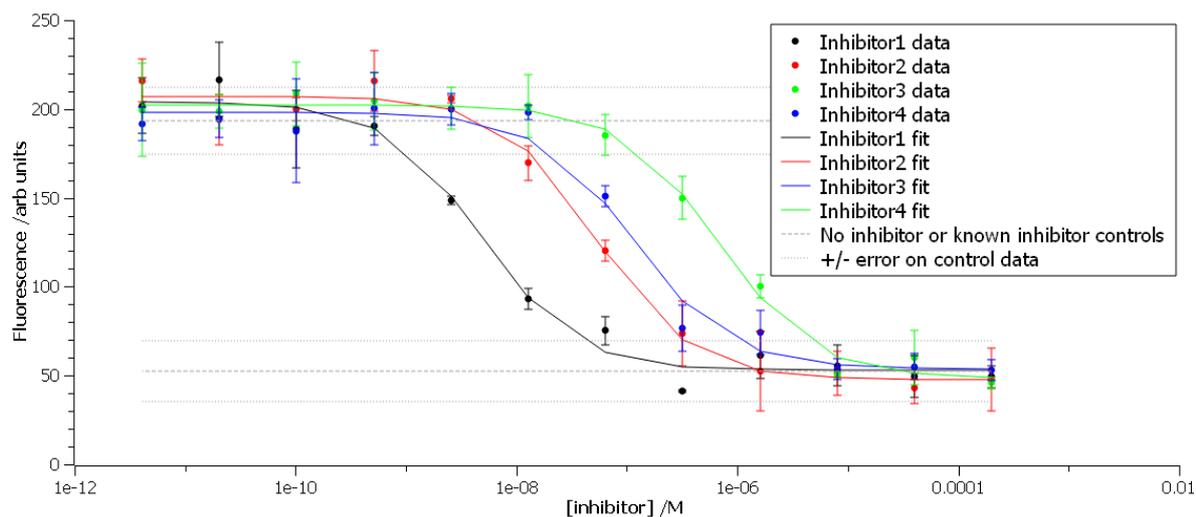
Work your way through your control data. Click *Apply* after each change – SciDAVis will update your graph without closing the dialogue box enabling you to see what your change has done. Finally click *OK*. Your graph might look something like mine below:



Double click on the graph legend to tidy this up. This opens the *Text options* dialogue box. You can type manually in here to change what appears in the legend on your graph. The codes on the very left (eg $\backslash c\{2\}$) tell SciDAVis which line to plot (please don't change these), but everything after the } can be amended as you want. I have amended the text on my figure legend to make it simpler and changed the *Opacity* from *Transparent* to 255 (solid white).



Finally, save your project and export your graph.



Congratulations! You have now reached the end of the workshop. Please keep hold of this booklet – it is likely that you will find it useful in the virtual drug discovery exercise which will run throughout semester 2.

Installing SciDAVis on your own machine

If you want to install SciDAVis on your own machine, it is free to use under version 2 of the GNU General Public Licence. It can be downloaded from <http://scidavis.sourceforge.net> and the licence is available at <http://scidavis.sourceforge.net/about.html>. SciDAVis is available for Windows, Mac and Linux.

The material in this booklet was created using SciDAVis 1.23. It is © Charlotte Dodson and is published under a CC-BY-SA 4.0 licence <https://creativecommons.org/licenses/by-sa/4.0/>.