Estimation of initial rate from discontinuous progress data

Submitted to Biocatalysis & Biotransformation

Peter J Halling

WestCHEM, Department of Pure & Applied Chemistry, University of Strathclyde, Glasgow G1 1XL, UK. p.j.halling@strath.ac.uk

Supporting Information

1. Approximations to integrated kinetic equations

The Reverse Quadratic equation can actually be obtained as an approximation to some kinetic models, as follows.

The integrated Michaelis-Menten equation (equation 8 in paper) can also be written

$$P - K_M.ln\left(1 - \frac{P}{S_0}\right) = V.t$$

We now take the first two terms in the Taylor series expansion of ln(1-x) to approximate for small conversion (P/S_0)

$$P - K_M.\left(-\frac{P}{S_0} - \frac{P^2}{2S_0^2}\right) = V.t$$

$$P.\left(\frac{S_0 + K_M}{S_0}\right) + K_M.\frac{P^2}{2S_0^2} = V.t$$

$$P + \left(\frac{1}{S_0 + K_M}\right).\frac{K_M.P^2}{2S_0} = \left(\frac{S_0.V}{S_0 + K_M}\right).t = v_0.t$$

which has the Reverse Quadratic form with parameter *a* expressed as a function of K_M and S_0

If we go for the more complicated Michaelis-Menten equation for a reversible reaction:

$$v = \frac{\frac{V_f}{K_S}\left(S - \frac{P}{K_{eq}}\right)}{1 + \frac{S}{K_S} + \frac{P}{K_P}}$$

where K_{eq} is the chemical equilibrium constant, and V_f, K_S and K_P are kinetic parameters of the model. Eliminate S (= S_0 – P)

$$\frac{dP}{dt} = \frac{V_f\left(S_0 - P - \frac{P}{K_{eq}}\right)}{K_S + S_0 - P + \frac{K_S}{K_P}.P}$$

We can re-parameterise by defining

$\propto = 1 + \frac{1}{K_{eq}}$ $\qquad$ $\beta = \frac{K_S}{K_P} - 1$

$$\frac{dP}{dt} = \frac{V_f(S_0 - \propto.P)}{K_S + S_0 + \beta.P}$$

$$\int \frac{K_S + S_0}{S_0 - \propto.P}\, dP + \int \frac{\beta.P}{S_0 - \propto.P} dP = \int V_f.dt$$

Integrate, using the standard result for integral of [x/(a + bx)]:

$$-\frac{K_S + S_0}{\propto}.ln(S_0 - \propto.P) - \frac{\beta.P}{\propto} - \frac{\beta.S_0}{\propto^2}.ln(S_0 - \propto.P) = V_f.t + const$$

By taking P = 0 when t = 0, we find

$$const = -\left(\frac{K_S + S_0}{\propto} + \frac{\beta.S_0}{\propto^2}\right).ln(S_0)$$

Substitute and group terms

$$-\left(\frac{K_S + S_0}{\propto} + \frac{\beta.S_0}{\propto^2}\right).ln\left(\frac{S_0 - \propto.P}{S_0}\right) - \frac{\beta}{\propto}.P = V_f.t$$

Now writing the logarithm term as

$$ln\left(1-\frac{\propto.P}{S_0}\right)$$

and taking the first two terms of the Taylor series expansion of ln(1-x) gives

$$-\left(\frac{K_S+S_0}{\propto}+\frac{\beta.S_0}{\propto^2}\right).\left(-\frac{\propto.P}{S_0}-\frac{\propto^2.P^2}{2S_0^2}\right)-\frac{\beta}{\propto}.P=V_f.t$$

The second and third terms in the brackets multiply to cancel (β/α)*P, leaving

$$\left(\frac{K_S+S_0}{S_0}\right).P+\left(\frac{\propto K_S+\propto S_0+\beta.S_0}{\propto^2}\right).\left(\frac{\propto^2.P^2}{2S_0^2}\right)=V_f.t$$

Multiply both sides by $S_0/(K_S+S_0)$ and make cancellations

$$P+\left(\frac{\propto.K_S+\propto.S_0+\beta.S_0}{2S_0.(K_S+S_0)}\right).P^2=\left(\frac{S_0}{K_S+S_0}\right).V_f.t=v_0.t$$

And again this has the Reverse Quadratic form. The parameter multiplying P^2 is simplified under certain limiting conditions. For an essentially irreversible reaction, $\alpha \rightarrow 1$. If product binding to the enzyme is weaker in comparison to substrate, $K_S/K_P \rightarrow 0$, so $\beta \rightarrow -1$. If both these conditions are satisfied, the numerator becomes simply K_S and we get the same expression derived above for simple Michaelis-Menten kinetics.

Another approximation relates the deactivating enzyme model to the quadratic one. From equation 6 in the paper

$$P=\frac{v_0}{k_d}(1-e^{-k_d.t})$$

using the series expansion of e^x, we get

$$P=\frac{v_0}{k_d}\left[1-\left(1-k_d.t+\frac{k_d^2.t^2}{2}+higher\ terms\right)\right]$$

Neglecting the higher power terms, cancelling 1 and k_d gives

$$P=v_0.t-\frac{k_d.v_0}{2}.t^2$$

Which has exactly the form of the simple quadratic model, with the curvature controlling parameter (*a*) now equal to $-k_d.v_0/2$.

2. Analysis of literature method by Poor (1968)

A spreadsheet was set up to calculate differences and the initial rate according to equation 2 from Poor (1968). This gave the expected rate for perfectly linear and error-free data values. However, addition of even 1% error to product concentration values would often make the calculated rate very different. Often this was because of large higher order terms in the summation of Poor's equation 2 for equally spaced points:

$$v = \frac{\Delta[P]_0}{\Delta t} - \frac{\Delta^2[P]_0}{2!\,(\Delta t)^2}t_1 + \frac{\Delta^3[P]_0}{3!\,(\Delta t)^3}t_1t_2 - \frac{\Delta^4[P]_0}{4!\,(\Delta t)^4}t_1t_2t_3 + \cdots$$

where Δt is the interval between points, at times t_n. Differences $\Delta[P]_0$ etc come from a sequential differencing scheme. If the points start at t = 0, then $t_1 = \Delta t$, $t_2 = 2\Delta t$, etc. Hence $t_1t_2 = 2(\Delta t)^2$, $t_1t_2\ldots t_n = n!(\Delta t)^n$. Substitute in the above equation, and all but a single Δt and the highest integer in the factorial cancel, leaving

$$v = \frac{\Delta[P]_0}{\Delta t} - \frac{\Delta^2[P]_0}{2\Delta t} + \frac{\Delta^3[P]_0}{3\Delta t} - \frac{\Delta^4[P]_0}{4\Delta t} + \cdots$$

So the terms do not decline in magnitude quickly, especially since with some error patterns the differences $\Delta[P]_0$ etc become larger for higher orders. In summary, the Poor method does not appear to be sensible for data with significant errors.

3. Details of Excel macros and files used

All fitting was done by using the built-in Excel Solver to minimise the sum of weighted square deviations of the data points, by varying the initial rate and one or two additional parameters. Fitting was automated using an Excel VBA macro. The standard settings used for the Solver were: GRG Non-linear method; Constraint precision 0.000001; Use automatic scaling; Convergence 0.00001; Forward derivatives; Require bounds on variables.

It was observed that sometimes the Solver would terminate prematurely with a fit that could be easily improved manually by adjustment of the initial rate and curvature control parameter. The problem could not always be resolved by more restrictive convergence criteria or changing the solution method. The problem was usually only found when errors in fitting were assumed to be entirely fractional, and was tracked down to a very strong dependence of the sum of squares on the simple product concentration offset parameter. If this was not correct, the zero time point, assumed to have very small error, would generate a very large weighted square deviation. This high dependence on the offset parameter caused early termination with non-optimal values of the other parameters. So the fitting macro was programmed to find a first solution varying all parameters, then to call the Solver again varying only the initial rate and curvature control (if present) parameters, before a final call with all parameters varied again.

In general the 3-cycle algorithm using the Solver would find the same, visually good, solution whatever the initial estimates provided for parameters. So these were normally left as those fitted to the previous data set. One exception was the curvature control parameter β in the Michaelis-Menten fit (section 3.2). If this started at a very large value, it would sometimes be left unchanged by the Solver, even though a better solution existed with a small value. For any large value, greater than 100 or so, the model defaults to linear progress, so perhaps the Solver algorithm concluded that β had no effect on the fit. To avoid such problems, an initial estimate of 1 was imposed for each new time course. Where the data warranted it, the Solver would then change this to a large value.

The following Excel files used in the study are available for download. They can currently be found at: https://strathcloud.sharefile.eu/d-s2838756205a4c79b

On publication they will be made publicly available in the Strathclyde data repository with a registered DOI.

Simulate initial rate.xlsm – simulates datasets with random errors

Test datasets.xlsx – as used with various fitting algorithms

Linear test datasets.xlsx – used specifically for linear fits

Fit straight.xlsm – fits straight lines

Fit initial rates.xlsm – fits various straight and curved models

Fit quadratic.xlsm – fits quadratic

Fit quad vs linear.xlsm – fits quadratic and linear to different numbers of points

Fit linear with rules.xlsm – uses objective tests to select number of points for linear fit

Linear fit results.xlsx – results from Fit straight.xlsm and analysis of them

Fit results.xlsx– results from Fit initial rates.xlsm and analysis of them

Quad fit results.xlsx– results from Fit quadratic.xlsm and analysis of them

Quad & Lin fit results.xlsx– results from Fit quad vs linear.xlsm and analysis of them

Mod lin fit results.xlsx– results from Fit linear with rules.xlsm and analysis of them

T tests.xlsm – t tests and F tests of results in paper

Poor method.xlsx – test of method of Poor 1968

Fit Initial Rate.xlsm – a rather more user friendly sheet that fits using methods recommended in this paper

4. Effect of random errors on reproducibility of findings

Since the simulation of data involved addition of random error terms, the fitted rates varied between time courses. To see general trends, typically fits were made to 20 time courses with the same true progress (but different errors). To see whether averaging 20 was enough to smooth out major random effects, four different sets were simulated and analysed for one case of relatively high mixed error.

Table 1 shows there are some differences in linear fit results between the replicate sets, as expected. So small differences observed in subsequent analysis may just reflect the random assignment of errors in the simulations. For example, from this table it is clear that the standard deviation is reduced by analysing more points extending to higher conversion. This may also give higher initial rate estimates, but it would be unwise to suggest this from the present study.

Table 1. Reproducibility of normalised rates and standard deviations from simulation approach.

	Set 1	Set 2	Set 3	Set 4
4 points up to 0.025 conversion	0.95 ± 0.15	0.97 ± 0.13	0.95 ± 0.18	0.94 ± 0.16
13 points up to 0.1 conversion	0.981 ± 0.035	0.955 ± 0.038	0.972 ± 0.039	0.980 ± 0.046

Data were simulated for Michaelis-Menten kinetics with K_M equal to initial substrate concentration, using error contributions of 0.04 fractional and 0.0033 dimensionless absolute. Fitting to a straight line used the mixed error assumption. Table shows the mean of fitted rates (with the "true" value set at 1) and their standard deviation, from each set of 20 time courses.

Table 2 shows the extent to which this averaging succeeded when fitting curves, by comparing 4 datasets of 20 each, simulated for the same model (Michaelis-Menten) and high mixed error (type Mh). Since fitted curves commonly give initial rate estimates that are too high, the root mean square difference from the true value is shown as a measure of overall accuracy. As can be seen, there are still noticeable differences in averages between the sets. Hence fairly small differences in further comparisons may not be statistically significant. But fairly small differences are probably also not relevant for advice on methods to be used with real data, in part because they may depend on the underlying kinetic model and error characteristics, which will generally be unknown in real experiments.

Table 2. Reproducibility of average statistics between datasets for identical simulation conditions.

Model	Set 1	Set 2	Set 3	Set 4
Linear to c < 0.1	0.94 ± 0.06 (0.081)	0.97 ± 0.05 (0.058)	0.96 ± 0.06 (0.075)	0.95 ± 0.05 (0.070)
First order	1.03 ± 0.03 (0.037)	1.04 ± 0.02 (0.047)	1.04 ± 0.04 (0.051)	1.04 ± 0.04 (0.045)
Second order	1.13 ± 0.03 (0.13)	1.14 ± 0.03 (0.14)	1.14 ± 0.04 (0.14)	1.14 ± 0.03 (0.14)
Michaelis-Menten	0.98 ± 0.05 (0.055)	1.00 ± 0.06 (0.058)	1.00 ± 0.06 (0.056)	0.99 ± 0.06 (0.056)
Quadratic	0.98 ± 0.05 (0.055)	0.99 ± 0.04 (0.041)	0.99 ±0.05 (0.048)	0.99 ± 0.06 (0.058)
Deactivating	0.98 ± 0.06 (0.058)	0.99 ± 0.04 (0.043)	0.99 ± 0.05 (0.051)	0.99 ± 0.06 (0.061)
Reverse quadratic	0.99 ± 0.07 (0.066)	1.00 ± 0.05 (0.050)	0.99 ± 0.06 (0.057)	1.00 ± 0.07 (0.067)

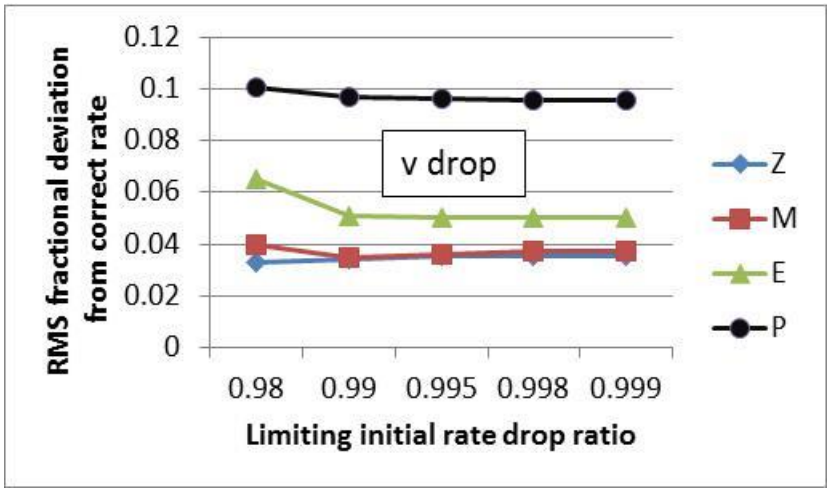
The table shows the mean ± standard deviation for the fitted relative initial rates. In brackets is the RMS deviation of the relative initial rates from the "true" value of 1. Data were simulated for Michaelis-Menten kinetics, initial substrate concentration equal to K_M, with relatively high mixed error, fractional 0.04, absolute 0.0033 dimensionless. Each of 20 time courses in each set had 11 equally spaced data points up to a conversion of about 0.18. Fits used the mixed error assumption. A fifth set showed noticeably higher standard deviations and RMS deviations for the fits to the quadratic, deactivating and reverse quadratic equations. This was due to a single time course in the data set where these equations were fitted with particularly high curvature and initial rates (a known issue with these equations).

5. Methods to select number of points to include in linear fit

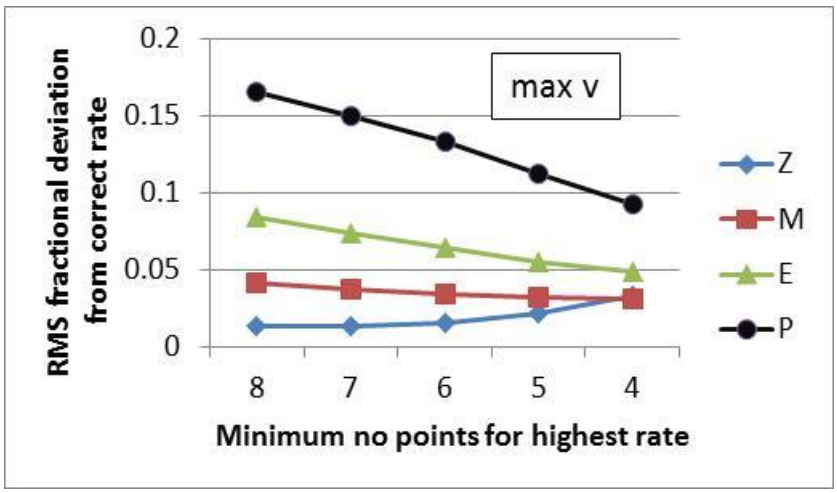
A number of possible methods were tested to select initial data points to use for a linear fit. Each of these methods was tested against the datasets for Zero order, Michaelis-Menten, Equilibrium approach and Product inhibition, covering the range of curvature. All 4 error types were included for each underlying progress type. In each case a minimum number of points was set (usually the first 4), and straight lines were fitted to each possible initial selection of points (first 4, first 5, etc). The algorithm then chose the fit with the largest number of points before a termination condition was reached. In each case the termination condition introduces a parameter that controls the number of data points selected, and hence the initial rate estimated. For each method a range of parameter values for the termination condition were tested, and the results used to select an optimal value.

The graphs below show the RMS deviations of the relative initial rates fitted from the "true" value of 1. Each point shows an overall value calculated from 4 datasets each of 20 time courses, simulated using the model equation shown, but with 4 different error conditions. Simulations used 4 models covering the range of curvature: Z – Zero order, linear; M – Michaelis-Menten kinetics; E – approach to equilibrium; P – product inhibition. Fitting assumed the same error contributions as had been used in simulation of the datasets. In each figure the parameters are ordered so that data points are more likely to be rejected as you go left to right. Reducing the number of data points selected for a linear fit tends to increase the accuracy of rates estimated for strongly curved progress, but to decrease accuracy for truly linear progress. These trends are visible in most of the figures.

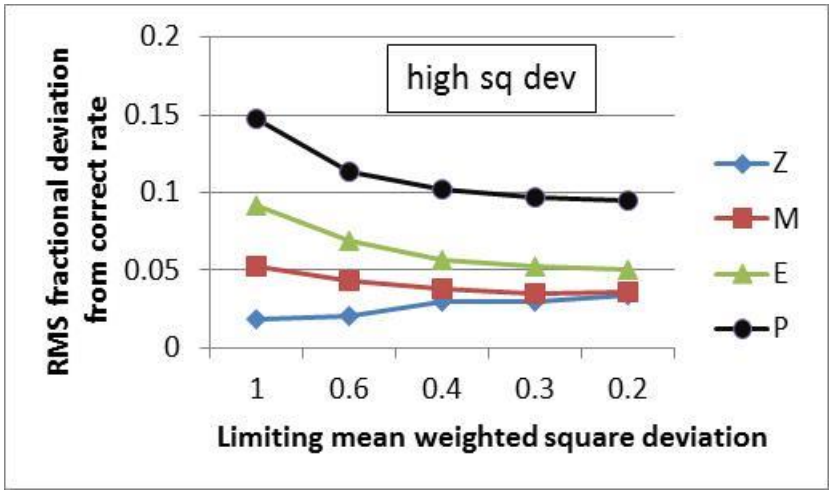
One termination condition was a decrease in fitted initial rate by a factor greater than a chosen limit ("v drop"). Limiting ratios of 0.99 and larger gave similar estimates and deviations, so 0.99 was selected for general use.

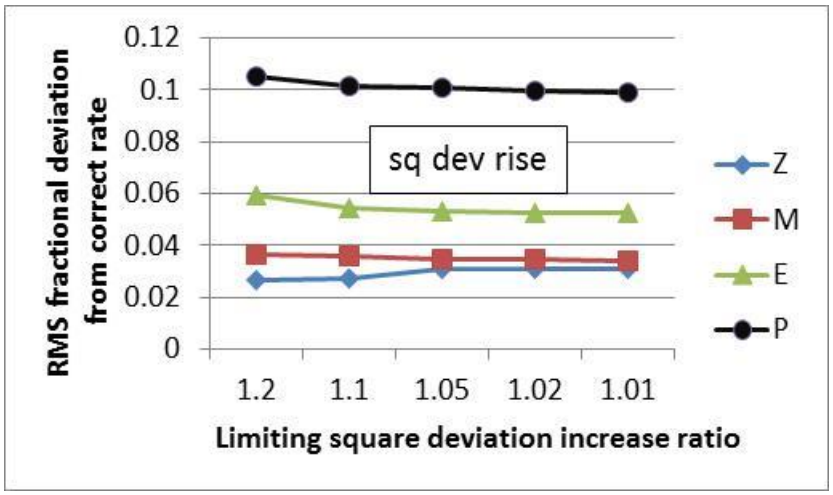


Another method simply selected the highest fitted initial rate, subject to a chosen minimum number of data points ("max v"). Using a minimum of 5 points gave a noticeable improvement in accuracy for truly linear progress, without too much deterioration for curved progress, so was chosen for general use.

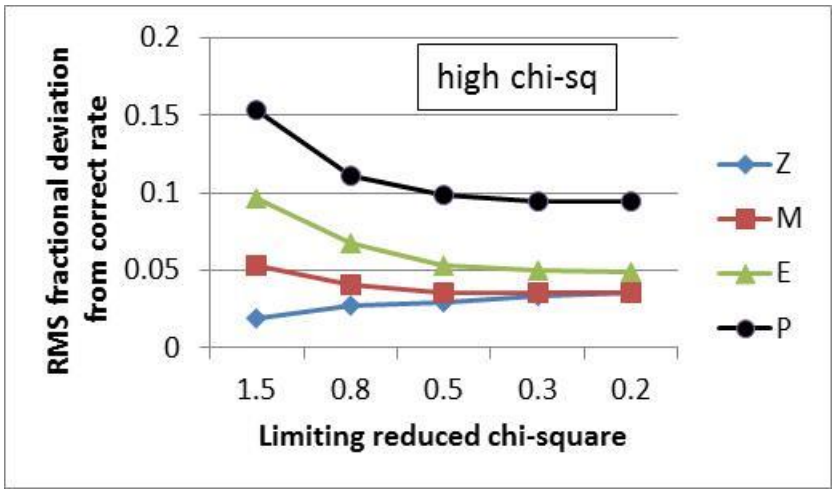


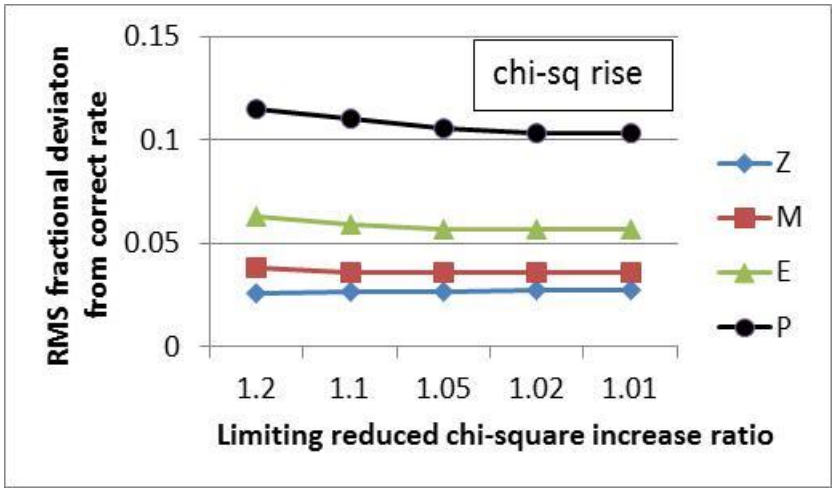
Two conditions looked at the mean weighted square deviation per data point of the fits. Termination could occur when this exceeded a given absolute value ("high sq dev") or increased by more than a chosen factor ("sq dev rise"). In both cases termination also required that there was a drop in fitted rate. For the "high sq dev" method, a limiting mean square deviation of 0.4 looked like a good compromise for general use. In the "sq dev rise" method, the parameter choices had relatively small effects, with essentially no difference between ratios of 1.02 and 1.01, and the former was chosen for general use.



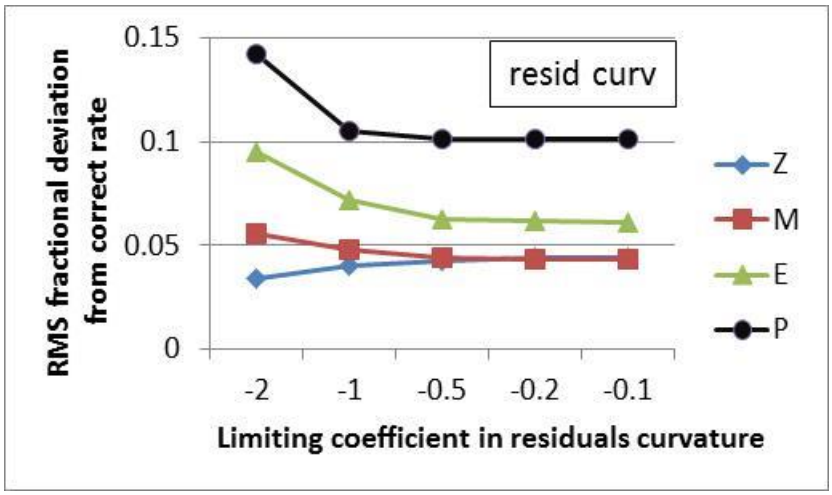


Two methods looked at the reduced chi-square value of the fits. The reduced chi-square is the sum of weighted square deviations divided by the number of degrees of freedom (number of data points minus number of parameters fitted). Hence it is similar too, but larger than, the mean weighted square deviation per data point. Again termination could occur when the fitted rate fell and the reduced chi-square exceeded a given absolute value ("high chisq") or increased by more than a chosen factor ("chisq rise"). For the "high chisq" method, a limiting value of 0.3 looked like a good compromise for general use. In the "chisq rise" method, the parameter choices had relatively small effects, and 1.02 was chosen for general use.





A final condition considered the residuals from the linear fits. For clearly curved progress data the residuals will tend to be negative at the shortest and longest times, and positive in the middle. So residuals were fitted to a quadratic equation, and a large negative value (less than a chosen limit) of the coefficient of the square term indicated the pattern expected for such curvature ("resid curve"). There was little difference between dimensionless parameter values of –0.5 or less negative, and –0.5 was chosen for general use.

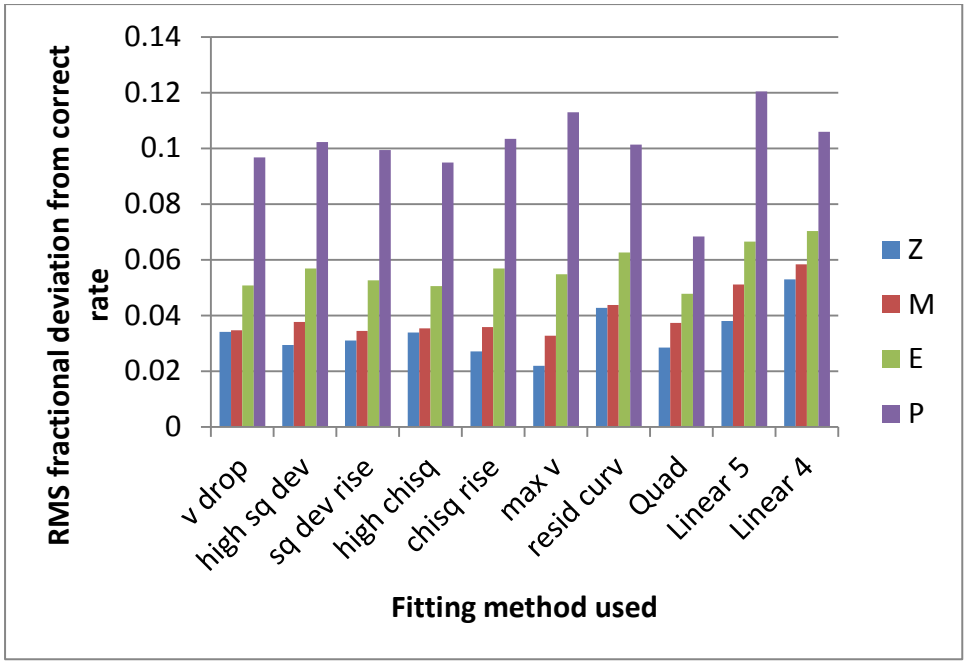


Having chosen the optimal parameter values, the overall performance can be compared for the different approaches to select points for a linear fit. The figure below shows their performance in terms of RMS deviation from the "correct" initial rate (normalised as 1), and also the mean and standard deviation of fitted rates. The overall differences between the various point selection algorithms are not large, with overall RMS deviations being 0.060 (v drop), 0.063 (high sq dev), 0.061 (sq dev rise), 0.059 (high chisq), 0.063 (chisq rise), 0.066 (max v), and 0.067 (resid curv). Any of the methods is arguably superior to the Linear fit to 5 time points in every case (see main paper for fuller details). For the lowest curvature "max v" is probably superior, and "resid curv" is worse than the others. Because the residuals are of course very scattered, the quadratic fit to them would often fluctuate dramatically as extra points were added, leading to some poor point selection choices. This contributes to a generally higher standard deviation in rate estimates.

However, there is a good basis for not recommending some of these approaches. Both the high sq dev and high chisq methods use statistics that depend on the absolute values of the weighted square deviations, which in turn depend on the assumed errors in the data points. In this simulation case these errors were exactly known, but in real experimental practice they may be very uncertain. As a test, the fitting exercise was repeated using the mixed error assumption, where this generated somewhat higher error estimates than had been used in simulation. While methods relying on a fractional rise in square deviations or reduced chi-square were little affected, those using absolute values generally selected several more points to include in the fit. For curved progress this meant a lower and less accurate initial rate estimate – the overall RMS deviation almost doubled.

The resid curve algorithm is also probably not a sensible choice, because of its noted variability and possibly slightly lower accuracy. It also has the problem that the limiting coefficient used has dimensions if plots use product concentrations and times. Hence for a transferable method it would need to be converted to a dimensionless value using the fitted initial rate and starting concentration.

Hence the v drop, max v, sq dev rise and chisq rise methods are recommended for further consideration (see main paper).



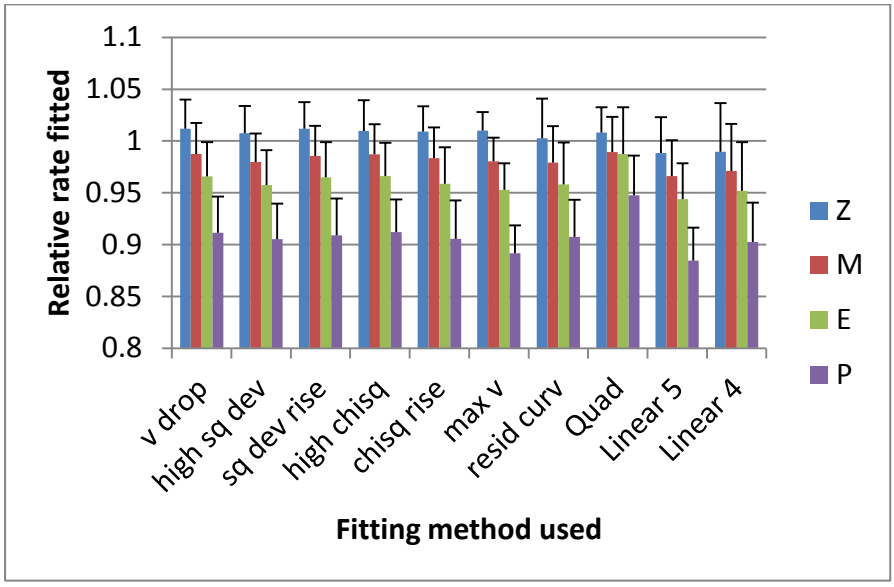


Figure. Performance of various algorithms to select initial points for linear fit. Each bar shows an overall value calculated from 4 datasets each of 20 time courses, simulated using the model equation shown, but with 4 different error conditions. The simulations are listed in order of increasing curvature: Z – Zero order, linear; M – Michaelis-Menten kinetics; E – approach to equilibrium; P – product inhibition. As well as the point selection algorithms, for comparison is shown Linear fits (to the first 4 and first 5 points in each case) and a Quadratic fit to all 11 time points. Fitting of all models used the known error contributions used in simulation. Part A shows the RMS deviations of the relative initial rates fitted from the "true" value of 1. Part B shows the mean values, and the error bars indicate mean standard deviations. A separate standard deviation was calculated for each dataset of 20 time courses which all had the same underlying progress and error – the value plotted is the mean of 4 such standard deviations.