Clean-up and analysis of small datasets can distort conclusions

Royal Truman

Many important scientific conclusions are based on small datasets with considerable measurement error, especially in areas relevant to the origin of life debate. Including a few very inaccurate or miscategorized data values could produce seriously flawed reconstructions, chronologies or relationships. But removal of data ('cleanup') can also eliminate putative outliers which could invalidate a flawed model. Expertise is often subjective, with warring views producing seemingly compelling proof for their view. Choice of experiments to perform, prior beliefs and reasons for selecting specific mathematical treatments to apply are rarely communicated to the readers of professional papers. Especially problematic is when convictions are use to 'massage' the data, leading to results then argued to demonstrate the validity of the prior belief.

Quantitative scientific data needs to be scrutinized by a subject matter expert (SME) for plausibility before and after mathematical tools are applied and conclusions are published. Erroneous data can lead to flawed mathematical equations of which decision-makers may not be aware. Correcting data requires prior knowledge, and cleaning up datasets can be quite subjective despite the best intentions.

In the past, when I worked in data science, I often encountered outliers that were clearly inconsistent with my empirical mathematical models. In many cases it was possible to trace the error back to the data sources, where explanations included, for example, a decimal point that had been accidentally shifted. But what about the cases where we doubt the validity of some data but have no means to decide if it is wrong?

In the literature, we often encounter examples of dates being recalibrated because the researcher believed more strongly in his or her presuppositions than in the data available. In bioinformatics, gene or protein sequences that disagree with phylogenetic relationships can be ignored or removed from the dataset. Is this dishonest, or simply a routine matter of data cleanup? This is an important question especially in those cases where the amount of data available is very limited.

There are many cases of important decisions relying on a small dataset. Examples include hominid fossils, amino-acid-containing meteorites, putative pseudo-genes, tree-ring series to calibrate ^{13}C ages, and so on.

Case study: glycine condensation

I recently examined some data from a paper published by Cronin *et al.*,[1] and, like all data scientists, I have a compulsive need to 'play with' quantitative data. In this *Nature Communications* paper, the team determined the concentration and size of poly-glycine using dehydration–hydration cycles. Parameters tested included initial concentration of glycine (Gly, 10^{-4}–10^{-1} M), dehydration times (1–96 h), number of dehydration cycles (1–4), temperature (90–130°C), pH (2.15–10), and concentration of NaCl (0–1 M).

This is useful data because it can help predict the largest Gly_n oligomer formed, using optimal settings. This is relevant for origin of life research, which hopes to account for a natural origin of large peptides. After calculating the theoretically largest Gly_n that could form, one could then extrapolate to more plausible abiotic conditions. Indeed, I concluded that the largest Gly_n would have been much smaller than formed under optimized laboratory conditions.[2]

While analyzing the data from the Cronin *et al.* paper, I noticed that I was instinctively applying judgment when evaluating the reported data and my mathematical fits. So, is bias always wrong? I decided to share some simple examples from my own *modus operandi* to illustrate several points that are relevant to the origins debate. This led to the *observations* shown in table 3.

I. Using data as is, without data transformation or cleanup

The maximum concentration of Gly_n having $n > 13$ could be estimated by extrapolation using the data in table 1.[1] Higher concentrations resulted after two cycles, but additional cycles led to chemical decomposition, so I decided to examine the data for cycle 2, figure 1.

Plotting the data shows this will be easy to model, figure 1A. In figure 1B I used a logarithmic function. Suppose one believes there are compelling reasons why a logarithmic relationship must be correct, and suspect the reported concentrations for Gly_{12} and Gly_{13} that do not follow this relationship very well might be flawed, figure 1B. Visual inspection shows $Glys_{12}$ and $Glys_{13}$ fall above

the empirical curve, whereas Gly_5–Gly_{11} all fall below the fitted curve. I re-examined the IP-HPLC traces and concluded that considerable doubt could be raised about the accuracy of the concentrations reported for Gly_{12} and Gly_{13}.

I regenerated the logarithmic curve with the cleaned-up dataset (i.e. lacking these two allegedly wrong data points). The correlation coefficient R^2 jumped from 0.979 (figure 1B) to an impressive $R^2 = 0.991$, based on a new regression equation $14.028–5.814\ln(n)$. This could be correctly or mistakenly provided as evidence that glycine oligomers larger than $n = 12$ will not be produced under these conditions, since replacing n with a value ≥ 12 leads to a negative % (*observation 1*).

Suppose instead that we trust all the data and now generate a third order polynomial function, Figure 1C, with $R^2 = 0.995$, which seems compelling and permits extrapolation to higher values of n. However, a negative % yield results for $n \geq 15$, which is physically absurd and indicates the fitted equation should not be extrapolated to high Gly_n values. But, in other studies, it is possible an analyst would have no reason to suspect he or she had overfitted the data set (*observation 2*). As an alternative example, suppose a fourth order polynomial would be offered, also having $R^2 = 0.995$. Now predicted values for $n \geq 14$ are no longer negative, but begin to increase steadily. Since the reported % yield of Gly_{13} was greater than of Gly_{11}, this might seem mathematically plausible. But a chemist would know this is not reasonable. In cases where the analyst lacks a deep understanding of the underlying physical reality, seemingly excellent equations offered could make nonsense predictions.

Table 1. Oligomer concentrations after number of hydration–dehydration cycles at 130°C after 24 h. Yields calculated as a percentage of the glycine (Gly) starting material.[1]

Oligomer	Cycle 1	Cycle 2	Cycle 3	Cycle 4
Gly_2	13.96	10.26	9.42	8.36
Gly_3	10.4	7.7	6.46	5.41
Gly_4	7.61	5.95	5.11	4.41
Gly_5	5.11	4.23	3.53	3.07
Gly_6	3.64	3.71	3.37	3.03
Gly_7	1.91	2.07	1.94	1.67
Gly_8	1.91	2.05	1.07	0.64
Gly_9	1.09	1.3	0.77	0.66
Gly_{10}	0.81	0.93	0.81	0.74
Gly_{11}	0.2	0.32	0.28	0.26
Gly_{12}	0.4	0.56	0.4	0.38
Gly_{13}	0.11	0.34	0.3	0.25

Performing data transformations

A quarter of the y values of the dataset are less than a tenth the size of the largest value, figure 1A. The regression algorithm minimizes the square of the difference between predicted and reported data, so the largest concentrations will dominate the resulting empirical equation. This is fine if the goal is to predict yields of Gly_n for small values of n, but here the opposite is true; we would like to extrapolate to $n \geq 14$. Therefore, I took the natural log of the y values and plotted them against n, figure 2A. The new linear regression equation has an $R^2 = 0.961$, which is not as high as obtained before (figures 1B and 1C). But now there will be better agreement between measured and predicted values in the larger n region. Importantly, this confirmed that, when extrapolating to $n \geq 14$, one no longer obtains negative yields nor increasing yields at high values of n (*observation 3*).

The plot in figure 2A makes clear that data point Gly_{11} is suspect. It makes no sense for the yield of Gly_{11} to be less than for Gly_{12} and about the same as Gly_{13}; see table 1. Ideally, additional laboratory measurements could be performed to resolve contradictions, but often this now instead requires behind-the-scenes decision making. Data delivered for analysis is often final.

Perhaps the plot in figure 2A should not be perfectly linear but display a slight downward trend. And since we wish to extrapolate to larger values of n, we are reluctant to forfeit our end point at Gly_{13}. If we exclude the Gly_{12} data point, we obtain a miniscule increase in R^2, figure 2B. Suppose we retain Gly_{12} and exclude Gly_{11} instead, since figure 2A reveals this to be the obvious outlier. This time the improvement in R^2 is rather dramatic (figure 2C), and using the equation leads to very reasonable-looking predictions, figure 2D. (There are statistical principles that can be used to decide which outliers are statistically significant based on assumptions

of the probability distribution of the errors, but this is not our topic here.) The message is simply that researchers routinely exclude data that they believe are flawed, and this is rarely apparent to those reading the reports. Sometimes data exclusions should have been done, but other times not.

The equations in figures 2A, B, and C were used to predict values for Gly_{14} to Gly_{20}, table 2.

Selecting one data point or the other as being an outlier produced a significant relative difference in predicted yield of Gly_{20} (table 2). This effect can be especially significant when large outliers are involved. Major conclusions could be communicated that are flawed because of incorrectly excluded data (*observation 4*). This becomes problematic when the outliers are chosen for removal in a way that strengthens what a researcher believes or wishes to be true. If one wishes to emphasize that large Gly_n won't form, then removing Gly_{12} is a temptation, whereas if one wishes to claim large oligomers are not so difficult to produce naturalistically, then removing Gly_{11} would be the option of choice.

Selective choice of experiments to perform

Another example of data bias arises in the selection of experiments to be conducted. In the rich amount of data available in ref. (1), we find the data shown in figure 3. Once the cycle time (i.e. duration of the dehydration phase) increases to longer times, chemical decomposition occurs that decreases the % yield of oligomer. This is shown for 110°C and 130°C.

Suppose only experiments at 90°C or less were chosen for analysis, knowing that chemical degradation would be a problem. No malice need be imputed. The researchers could simply be exercising good judgment to use their time and funding wisely. However, based on the now incomplete picture, their readers or sponsors might surmise that continually increasing the dehydration time would steadily increase the yield of oligomers. Based on the green line in figure 3, there would be no reason to suspect otherwise (*observation 5*).[3]

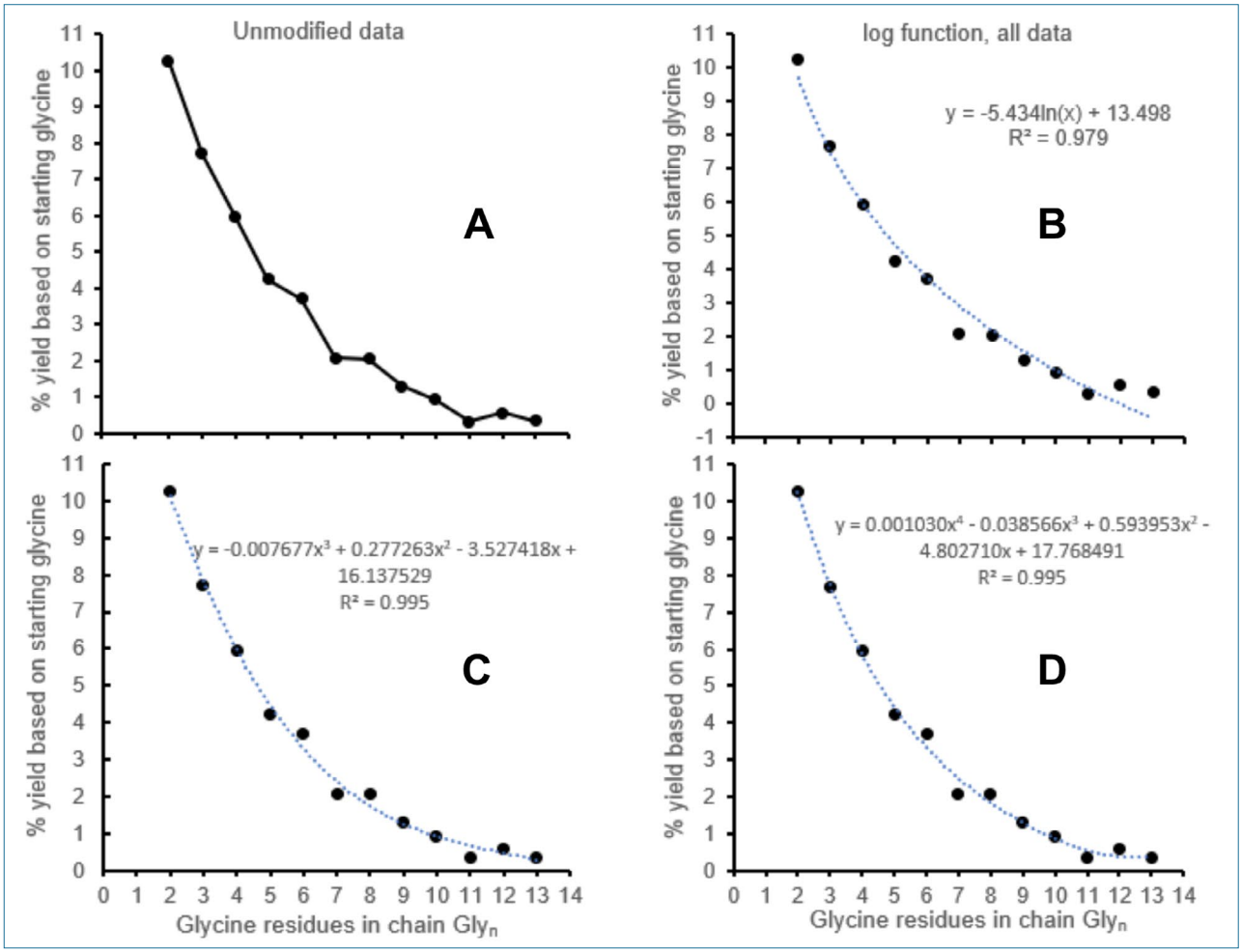

Figure 1. % yield glycine oligomers of size $n = 2–13$. **A**. Plot using all the data. **B**. Fitting of ln (% yield) *vs* Gly_n. **C**. Modelling % yield using a third order polynomial equation. **D**. Modelling % yield using a fourth order polynomial equation.

The intention of this research was to demonstrate that larger peptides could form in water than believed so far, lending more credibility to a natural origin of proteins. But why would the large peptides formed no longer remain exposed to degrading heat for much longer than 100 h? The cycle durations shown in figure 3 could have been extended to ensure the reader does not overlook what would have occurred naturally. For example, experiments with cycle times out to 500 h at 130°C would demonstrate that larger peptides would be almost entirely degraded.

Homochirality

There are many variants of this kind of bias. Which topic to research is an example. There is a plethora of papers addressing how the origin of biochemical homochirality might be solved naturalistically. Wildly overstated abstracts and summary statements, coupled with irresponsible journalistic sensationalism produces a general feeling that "someone has found a plausible solution, or with all the promising ideas one will be found" (*observation 6*). They hope that readers will forget that they said the same thing before; now they tacitly admit that the previous claim is no longer believed.

Suppose a comparable amount of effort was being devoted to finding all the experimental and theoretical ways amino acids in free or bound form could racemize. The flood of papers would cement the consensus that amino acid racemization is how nature works. In fact, that is why amino-acid racemization is widely used as a dating method.

Ancient biomaterials

Few researchers deliberately search for biological remains in fossils allegedly millions of years old or measure ^{14}C in diamonds which also allegedly formed millions of years ago. Neither are they focusing resources to examine alternative dating methods which could indicate the earth or life on

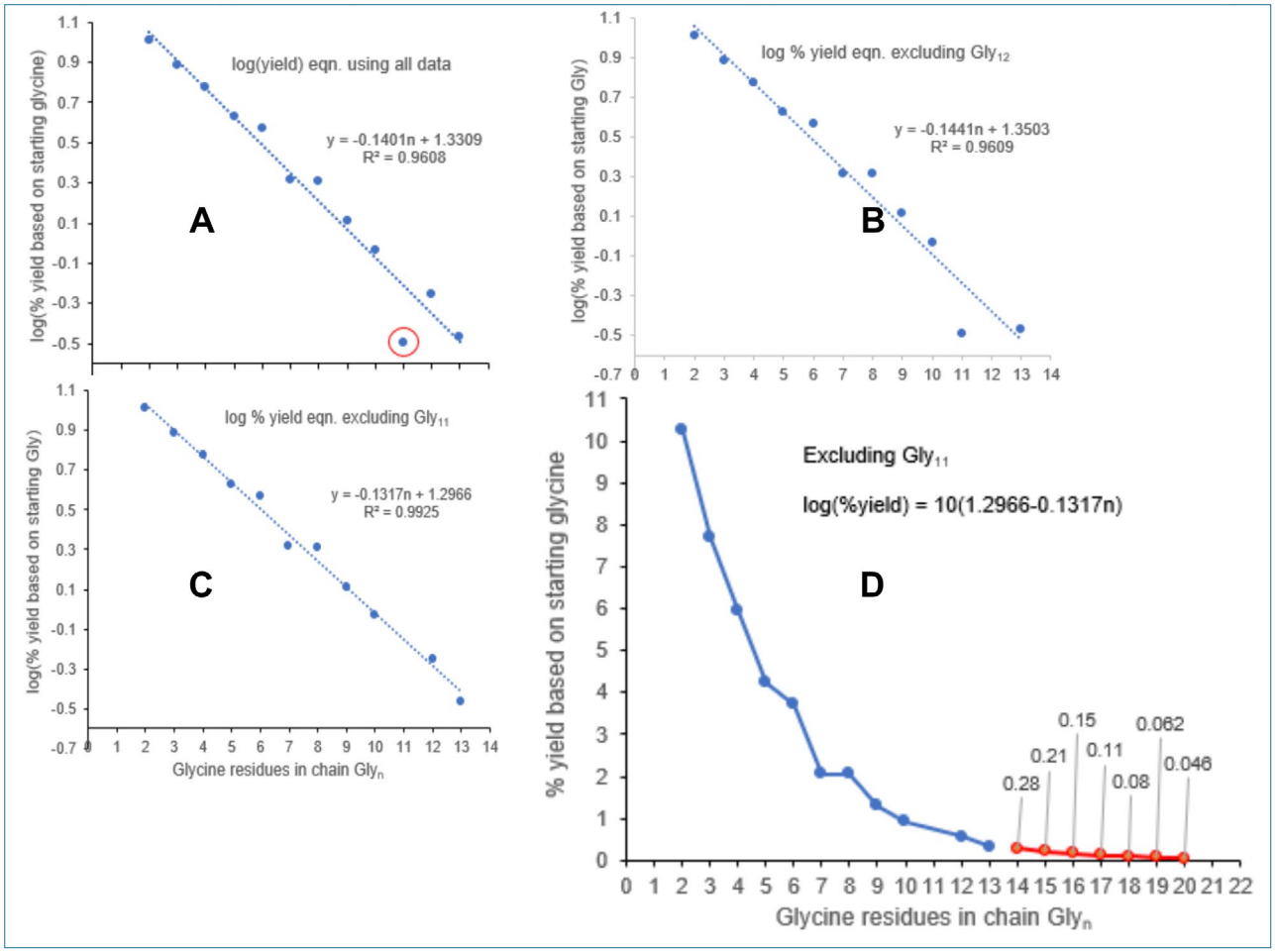

Figure 2. Analysis of the same data used in figure 1 after transforming to the ln of the % yield glycine oligomers. **A**. New regression equation using all the data. **B**. New regression equation after excluding data point Gly_{12}. **C**. New regression equation after excluding data point Gly_{11} **D**. Predicted values for *n* 14–20 using the regression equation from 2C.

Table 2. Predicted values for Gly_{14} to Gly_{20} after excluding one or no data points from the dataset. Yields calculated as a percentage of the Glycine (Gly) starting material.

No. of residues, n	Using all the data	Without the Gly_{12} data point	Without the Gly_{11} data point
14	0.23	0.22	0.28
15	0.17	0.15	0.21
16	0.12	0.11	0.15
17	0.089	0.080	0.11
18	0.064	0.057	0.084
19	0.047	0.041	0.062
20	0.034	0.029	0.046

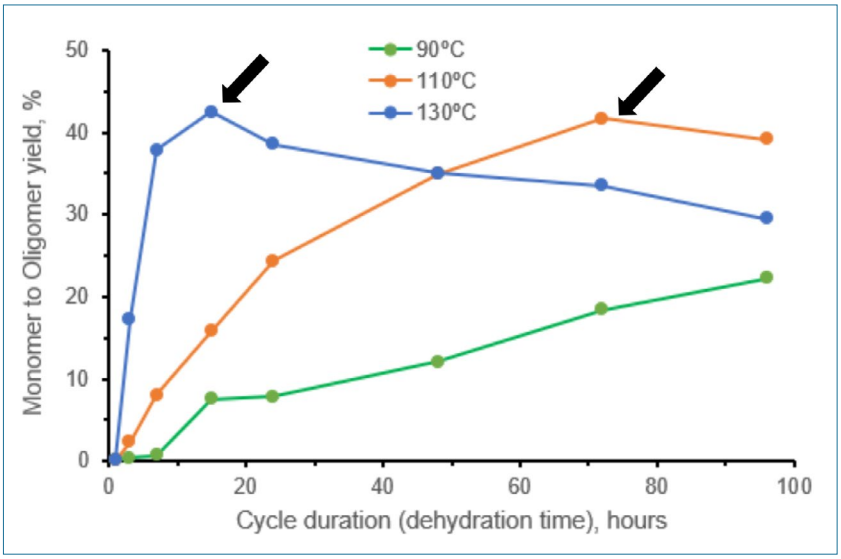

Figure 3. Three temperatures were chosen to experiment with. By avoiding other conditions to test and report, attention is not drawn to results inimical to the researcher's goals.

Earth might be recent. These are obvious research projects young-earth creationists would think of and wish to carry out. Clearly, more Bible-believing students need to become scientists.

Discussion

I cannot think of any paper dealing with origin of life topics having measured laboratory data where I did not wish that specific other tests would also be conducted. Given the dismayingly growing number of cases where important scientific and medical experiments cannot be validated, it is becoming ever more important to critically question how conclusions are being reached and communicated.[4,5] Those working with other premises will often think of alternative ways of interpreting data, or experiments leading to entirely different insights.

There are some subtleties to *observation 5*. In the example in figure 3, the sum of oligomers was reported but the goal of the project was to find the best parameter settings to produce the largest Gly_n. Larger oligomers were produced in mere fractions of a percent, so a sum of all oligomers is not truly addressing the question of interest. It is absurd to imply that pure amino acids would be present at 130°C somewhere for just a few hours before fleeing to the safety of much colder water to avoid degradation. Environments of much lower temperatures where decomposition would be minimized over time are more realistic, so the obvious experiments would be to determine Gly_n distribution at much lower temperatures. This would ensure that the correct facts are available to reach well-reasoned conclusions.

Observation 7 addresses the use of experimental details which camouflage facts that should be more honestly emphasized. An example is the use of glycine for condensation studies to show how large peptides might be formed naturalistically.[6] Glycine is the only proteinogenic amino acid that is not chiral, so is unable to form D- and L-enantiomers, and thus racemize. Poly-glycines cannot produce folded proteins. Another example involves the use of average dates obtained from different dating methods on the same sample to claim good agreement, whereas the range of values obtained for the same sample could be so great that serious doubt about the alleged agreement should exist. For example, one could exclude a particular measurement or two and the average values suddenly no longer agree at all. I often encounter published tables of data where, for each row, one reads 'average of *n* measurements', where *n* is variable. Which measurements were excluded and why is rarely explained.

Unfortunately, there are also examples of unethical bias. Data obtained that contradicts the researcher's thesis might not be reported, distorted, or downplayed (*observation 8*). It might be excluded entirely from the final report or could

Table 3. Insights on how bias affects analysis and reporting of data

No.	Observations
1	Presuppositions guide model building and which data to retain. Contradictory data might be wrong but should be reported.
2	Excellent mathematical fits to a dataset are no guarantee that extrapolations will be accurate. Variables not integrated in the model, or inadequately so, can lead to bad predictions.
3	Analysts often transform the original data for mathematical reasons. The consequences of the pre-processing are rarely communicated to others.
4	Excluding an unexpected result from consideration could prevent a new discovery. Sometimes a strong mathematical relationship results because key points are discarded. Identifying which data is erroneous has a big impact when the dataset is small. It is usually easy to find a reason for excluding a data point. However, none of the data which conform to expectations are challenged.
5	Experiments to perform are based on presuppositions and a desired outcome. For the non-specialist in the subject area this creates the impression that only those outcomes will occur. Parameter values are often selected near the best-case scenario. These optimized experiments establish a pattern in the mind of the reader as to what is expected to occur.[3]
6	Researchers select topics to explore. For example, who would write a research proposal for funding by the US National Science Foundation to find ways radioactive dating could lead to a false illusion of deep time?
7	Results can be reported in aggregated manners which ignore the uncertainty in measurements. For example, older values might reflect better what a researcher believes is true than newer measurements. Justifications are easy to find ("it has been contaminated since the former measurements"), so average values could simply be reported.
8	Biases can arise when the research project was funded by an entity with a strong agenda. Reporting 'bad' data along with the 'good' results could make the research team look inept and jeopardize further funding and publication.

be presented as a rare curiosity of allegedly no significance. Being inimical to the researcher's goals, little effort is invested to determine if the outlier is reproducible and, if so, what causes it.

Conclusions

Data clean-up and questioning which data could be flawed are a necessary part of research. Two kinds of errors could arise: data gets included in the models that should not have been or data gets excluded that should not have been. Presuppositions can be so strong that contradictory data is simply dismissed. An example is the view that only naturalist explanations are real, and these can explain all aspects of life. This is an assumption Dr Sivanesan has criticized in depth in his recent book.[7]

Using biased data clean-up to 'prove' a cherished belief can lead to circular reasoning. For example, if our presupposition is that a logarithmic function reflects the true underlying physics (figure 1B), and we remove the two rightmost values (or correct them in some *post-facto* manner), it would be incorrect to then use this new dataset and a new logarithmic fit to 'prove' Gly_n cannot be produced above a certain size. The origin of life literature is replete with this kind of error. Data are recalibrated or dismissed according to deep-time assumptions and this new 'data' is then used to claim that the facts speak for an ancient earth.

References

1. Rodriguez-Garcia, M., Surman, A.J., Cooper, G.J.T., Suárez-Marina, I., Hosni, Z., Lee, M.P., and Cronin, L., Formation of oligopeptides in high yield under simple programmable conditions, *Nature Communications* **6**(8385):1–6, 2015.
2. Truman, R., Racemization of amino acids: part 3—Condensation to form oligopeptides, *J. Creation* **36**(2):81–89, 2022.
3. Lenski, R.E., Ofria, C., Pennock, R.T., and Adami, C., The evolutionary origin of complex features, *Nature* **423**:139–144, 2003.
4. Ioannidis, J.P.A., Why most published research findings are false, *PLOS Medicine* **2**(8):e124, 2005.
5. Begley, C.G. and Ioannidis, J.P.A., Reproducibility in science: improving the standard for basic and preclinical research, *Circulation Research* **116**(1):116–126, 2015.
6. Ogata, Y., Imai, E.-I., Honda, H., Hatori, H.K., and Matsuno, K., Hydrothermal circulation of seawater through hot vents and contribution of interface chemistry to prebiotic synthesis, *Orig. Life Evol. Biosphere* **30**:527–537, 2000.
7. Sivanesan, N., *Objections to Evolution*, Poland Sp. z.o.o., Wroclaw, 2020.

Royal Truman *has bachelor's degrees in chemistry and in computer science from State University of New York; an M.B.A. from the University of Michigan (Ann Arbor); a Ph.D. in organic chemistry from Michigan State University; and a two-year post-graduate 'Fortbildung' in bioinformatics from the Universities of Mannheim and Heidelberg. He works in Germany for a European- based multinational.*