

1 **An empirical analysis of enzyme function reporting for experimental reproducibility:**
2 **missing/incomplete information in published papers**

3
4 Peter Halling^a, Paul F Fitzpatrick^b, Frank M Raushel^c, Johann Rohwer^d, Santiago Schnell^e, Ulrike
5 Wittig^f, Roland Wohlgemuth^g, Carsten Kettner^h

6 ^aWestCHEM, Dept P & A Chemistry, University of Strathclyde, Glasgow, G1 1XL, United Kingdom

7 ^bThe University of Texas Health Science Center, San Antonio, TX, USA

8 ^cDepartment of Chemistry, Texas A&M University, College Station, TX, USA

9 ^dDepartment of Biochemistry, University of Stellenbosch, Stellenbosch, South Africa

10 ^eDepartment of Molecular & Integrative Physiology, and Department of Computational Medicine &
11 Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA

12 ^fHeidelberg Institute for Theoretical Studies (HITS gGmbH), Heidelberg, Germany

13 ^gSigma-Aldrich, Member of Merck Group, Industriestrasse 25, CH-9470 Buchs, Switzerland

14 ^hBeilstein-Institut, Frankfurt am Main, Germany

15
16 Email addresses: p.j.halling@strath.ac.uk; fitzpatrickp@uthscsa.edu; raushel@chem.tamu.edu;
17 jr@sun.ac.za; schnells@umich.edu; ulrike.wittig@h-its.org; roland.wohlgemuth@sial.com;
18 ckettner@beilstein-institut.de;

19
20 Short title: Omissions in published enzyme function reports

21
22
23 Corresponding author: Peter Halling, Dept P & A Chemistry, University of Strathclyde, Glasgow, G1
24 1XL, United Kingdom; p.j.halling@strath.ac.uk; +44 141 548 2683; ORCID: 0000-0001-5077-4088

25
26 Declarations of interest: none.
27

28 **Abstract**

29 A key component of enzyme function experiments is reporting of considerable metadata, to allow
30 other researchers to replicate, interpret properly or use fully the results. This paper evaluates the
31 completeness of enzyme function data reporting for reproducibility. We present a detailed
32 examination of 11 recent papers (and their supplementary material) from two leading journals. We
33 found that in every paper we were not able to collect some critical information necessary to

34 reproduce the enzyme function findings. Study of 100 papers used by the SABIO-RK database
35 confirmed some of the most common omissions: concentration of enzyme or its substrates, identity
36 of counter-ions in buffers. A computer system should be better at preventing such omissions,
37 helping secure the scientific record. Many of the omissions found would be trapped by the currently
38 available version of STRENDA DB.

39

40 **Keywords:** enzyme function; metadata; reproducibility; transparency; data quality; standardization

41

42

Introduction

43 Studies of enzyme function are important in many areas, including biochemistry, cell biology,
44 chemical biology and applied biocatalysis. The comprehensive description of experiments on enzyme
45 function is necessary but not trivial, as many factors can affect the results obtained. Unfortunately
46 full description cannot be guaranteed using a simple checklist, as the factors that need to be
47 reported depend on the precise experiment carried out and the type of measurements performed.
48 Most scientists working in the field have experience of finding papers with missing essential
49 information about experimental details, but this often only comes to light when attempting to
50 repeat the experiments described. Some previous studies have shown that key information is rather
51 commonly missing (Wittig et al, 2014ab). Through extensive interactions with the biochemistry
52 community, the Standards for Reporting Enzymology Data (STRENDA) Commission, supported by the
53 Beilstein Institut, has established reporting guidelines with the aim to support authors to
54 comprehensively report kinetic and equilibrium data from their investigations of enzyme activities
55 (<http://www.beilstein-strenda.org>). The goal is providing a seamless mechanism through which
56 scientists and journals can jointly ensure reliable reporting of data (Apweiler et al., 2005; Apweiler et
57 al., 2010; Gardossi et al., 2010; Tipton et al. 2014).

58 These guidelines are now recommended by more than 50 international biochemistry journals which
59 publish work on enzyme function. In addition, the guidelines are registered with FAIRsharing.org (a
60 resource on data and metadata standards, inter-related to databases and data policies) as part of
61 the FAIRDOM Guidelines for Systems Biology. However, none of these journals enforce data
62 reporting in compliance with the guidelines. Full reporting is clearly important if others wish to
63 repeat the experiments. But it can be equally if not more important to those who simply wish to
64 compare the reported results or reuse them for purposes like modelling metabolic pathways. The
65 values of kinetic parameters can be very dependent on the precise conditions under which they are

66 measured. Hence accurate and comprehensive reporting of conditions is essential to make results
67 useful to users in fields such as systems biology, synthetic biology and applied biocatalysis.

68 In this work, we seek to test the efficacy of current reporting in the recent literature by
69 systematically sampling papers investigating enzyme function in two high profile journals. Over the
70 last few years, a shift is occurring in the research community towards sharing to any reader all data
71 necessary to understand, assess, and extend the conclusions of the manuscript. For example, *Science*
72 suggests using established community repositories and databases to host data (American
73 Association for the Advancement of Science. Science journals: Editorial Policies. Available at
74 www.sciencemag.org/authors/science-journals-editorial-policies). We test the availability of critical
75 enzyme assay data to allow replication of experiments, using the STRENDA Guidelines. We found
76 that in all of the papers sampled, we were unable to find at least one piece of critical data.

77 **Approach adopted**

78 Our empirical sample comprised of 11 papers that report studies on enzyme function in two selected
79 leading journals, which we studied very carefully. Having arbitrarily chosen a recent date, we worked
80 backwards to find all papers including significant enzyme function studies, getting 6 and 5 from the
81 two journals. No further selection was made, so these papers should be a representative sample.
82 These 11 papers and their Supplementary Materials were then studied in great detail, and assessed
83 using the STRENDA Reporting Guidelines. All 11 papers were found to have some issues, of which
84 further details are given below. Clearly some are more serious than others, but all should be avoided
85 when the requirement of a comprehensive description of an enzymology experiment should be met.
86 The object of this study is not to show deficiencies in the papers of a particular journal or individual
87 authors. Undoubtedly similar problems could be found by detailed study of papers from any journal
88 and most authors. Hence references are not given below, and the quoted text has irrelevant sections
89 removed to prevent easy identification by searches. Full references and quotes have been made
90 available to editors and referees on a confidential basis. We do also give full details of some previous
91 publications by co-authors of the present paper, who have identified similar issues with their own
92 work! In addition, to get more meaningful statistics, we have looked for selected omissions in
93 papers used as sources for the SABIO-RK database on enzyme function (<http://sabiork.h-its.org/>)
94 (Wittig et al, 2018), extending the previous study (Wittig et al, 2014ab). For this purpose we have
95 analysed the 100 most recent papers taken from the 10 journals that are the most common sources
96 for this database.

97 The objective of our approach is to show how easy it is for authors, referees, editors and others to
98 fail to spot missing information. Avoiding omissions in reporting data is a job that electronic systems
99 do much better than the human brain. It is for this reason that the Beilstein Institut, with guidance
100 from the STRENDA Commission, has created a web-based system (STRENDA DB) that incorporates
101 the STRENDA guidelines. The STENDA-DB data submission form automatically checks the manuscript
102 data prior or during the publication process in compliance with the guidelines. The data entered
103 normally become publicly available in a database only after the corresponding article has been peer-
104 reviewed and published in a journal. The current released version is freely available for use at
105 <https://www.beilstein-strenda-db.org/strenda/>, and has been described in more detail (Swainston et
106 al, 2018). As this tool develops it should gain ever more sophisticated algorithms to handle the
107 admittedly quite complicated rules for what details are needed for each type of experiment. But
108 many of the issues noted below would already be trapped by the current version of STRENDA DB, as
109 indicated in the table showing details. This is done to illustrate better the omissions in the papers
110 analysed, rather than to show comprehensively what the current version of STRENDA DB can and
111 cannot do.

112

113 **Details of issues found in recent published papers**

114 We evaluated by careful inspection whether each paper provided information to carry out
115 replication and judged that all papers missed some details critical to make the results potentially
116 reproducible. The missing data issues are given in the following Table 1. In every case we looked
117 carefully to see if apparently missing information could be found elsewhere in the paper or the
118 Supplementary Materials, but did not find it. Extracts shown may be from the paper itself or its
119 Supplementary Materials. In each case we have removed some unnecessary detail from the
120 quotations to make the sources hard to identify. Deletions are indicated by, while substitutions
121 have been made when ..[enzyme].. or ..[substrate].. replaces the full name.

122 As referees of manuscripts, we will quite often spot issues of this type and require their correction.
123 But as a referee, the focus is naturally on the scientific interest and quality of the paper, and the
124 extent to which conclusions are supported by the results presented. In the present cases, where we
125 did not need to be concerned with such topics, we could focus on these issues of completeness in
126 the description of the experiments. As noted, this is ultimately a job better suited to a computational
127 system, leaving referees to deal just with the scientific quality, for which human judgment is
128 essential. It is of course important to assess whether the experimental design was sensible and the

129 analysis of the results appropriate. But this is beyond the scope of our current paper. We simply
 130 focus on a more basic issue: does the paper fully report what was done? If essential details are
 131 omitted, then complete further expert evaluation is impossible.

132

133

134 **Table 1. Missing information found in 11 sample papers.** The table also illustrates where the use of
 135 STRENDA DB would ensure proper reporting.

Issues found in papers	Where could STRENDA DB help?
<p>Paper 1</p> <p>In the description of a single turnover experiment, we find “After the pre-incubation periods, the samples were heated to 37°C and reaction was initiated by mixing 1:1 with 25 mM HEPES, pH 7.5, 50 mM KCl, 16 mM MgCl₂, 1 mM DTT, 0.1 mg/mL BSA and 5% glycerol, to a final concentration of 2 μM enzyme and 5 nM substrate.” HEPES buffers require a positive counter-ion, usually Na⁺ or K⁺, but the choice can affect enzyme function. The inclusion of KCl in the buffer here suggests there may be monovalent cation effects on this enzyme.</p> <p>A reviewer has pointed out that the high temperature coefficient of HEPES buffers will also lead to a significant pH change in experiments with temperature varied between 20 °C and 37 °C. (See further details for paper 5).</p>	<p>STRENDA DB requires naming of a specified compound for all buffer components. If the user follows the instructions given when the compound is assigned as a buffer, full counter-ion details will be given.</p>
<p>Paper 2</p> <p>An assay is described “to determine the initial rate of O₂ consumption in solutions employing 0.2 μM ..[enzyme].., dissolved atmospheric oxygen (i.e., 0.250 mM), and sodium formate (100 mM) in 50 mM acetate (pH 3.6).” The wording suggests that sodium formate might have been added to an acetate buffer previously adjusted to pH 3.6. Acetate at pH 3.6 has rather weak buffering capacity, whereas formate is close to its pKa and so will dominate buffering – hence the</p>	<p>STRENDA DB requires the final assay pH, which should be the value after all components are added.</p>

<p>final pH would end up quite far above 3.6. Of course the formate might have been adjusted to pH 3.6, but this is not clear. Furthermore, the counter-ion to the acetate is not specified – probably Na⁺, but can't be sure.</p> <p>Results from assays are reported as apparent k_{cat} and k_{cat}/K_M values. But there is no information on the range of substrate concentrations studied to estimate these values. From the values presented we can see that K_M must be much lower than the fixed value of substrate concentration quoted elsewhere.</p>	<p>STRENDA DB requires entry of buffer counter-ions if help recommendations are followed.</p> <p>STRENDA DB requires entry of the range of concentrations studied for any substrate that is varied.</p>
<p>Paper 3</p> <p>A part of the study is reported as “Test reactions using ..[enzyme].. (1 mg/mL), 2 mM ..[substrate], 50 mM HEPPS (pH 8.0), and 1 mM of various nucleotide-linked sugars were allowed to incubate for 12 h at 37 °C.” Again the counter-ion in the HEPPS buffer is not stated.</p> <p>For determination of kinetic parameters we have “The 1.5 mL reaction mixtures contained 5 mM ..[substrate].., 50 mM HEPPS (pH 8.0), 0.28 μM ..[enzyme].., and ..[substrate 2].. concentrations ranging from 0.01–2.0 mM.” Again the buffer counter-ion is not stated. (This is a single time point assay, and the reaction time is not stated here, but a comment in the Results section confirms that this was within the linear range.)</p>	<p>STRENDA DB requires entry of buffer counter-ions if help recommendations are followed.</p> <p>STRENDA DB requires entry of the buffer counter-ions if help recommendations are followed.</p>
<p>Paper 4</p> <p>A plot of the Michaelis-Menten equation for the effect of substrate concentration has its y axis labelled just “[hour⁻¹]”.</p> <p>This may be an initial linear reaction rate presented as mol product (mol enzyme)⁻¹ hour⁻¹ but this is not made clear.</p>	<p>STRENDA DB does not currently allow entry of initial rate values, although this is a planned enhancement. Units for all results entered must be selected from a pre-set list which should have</p>

	unambiguous meanings.
<p>Paper 5</p> <p>The legend for a figure showing effects on k_{cat} reads “The concentration of NADH was 80 μM and concentration of ..[substrate].. was increased as needed with increasing temperature to maintain substrate saturation. The data was normalized relatively to that in the buffer without TMAO. All measurements were performed in triplicate with average values reported. Buffer: 100mM Na-phosphate, pH 7.2 with or without TMAO.” No information is given about the actual substrate concentrations studied, and the evidence that enzyme saturation was ensured so that k_{cat} values were obtained. Furthermore, the value of k_{cat} used for normalisation is not stated, so no absolute values can be deduced.</p> <p>Studies of the effect of temperature over a significant range are reported, but without necessary detail on the treatment of pH. Was buffer pH set at some reference temperature, then heated or cooled as required? This can cause significant changes in actual pH. Or were pH values actually set/measured at each assay temperature?</p>	<p>STRENDA DB requires entry of the range of concentrations studied for any substrate that is varied.</p> <p>To make an independent confirmation of saturation would require initial rate data for various concentrations of both substrates, which cannot currently be entered in STRENDA DB. Another planned enhancement is specification of the equation which has been fitted to the data, which would clarify the meaning of k_{cat} values.</p> <p>STRENDA DB requires that k_{cat} values are reported in absolute units, selected from a preset list.</p> <p>Fuller details of the treatment of pH are a planned enhancement.</p>
<p>Paper 6</p> <p>Assays use a “standard buffer, consisting of 50 mM HEPES (pH 7.5), 150 mM KCl, 10 mM $MgCl_2$, and 5 mM β-mercaptoethanol (β-ME).” The description misses the counter-ion in the HEPES buffer, and the presence of KCl again makes one suspect monovalent cation effects.</p> <p>For a pre-steady state study “In a typical experiment, one syringe contained enzyme at a final concentration of 5 μM ... monomer, 8 units/mL PPIase, a saturating concentration of</p>	<p>STRENDA DB requires entry of buffer counter-ions if help recommendations are followed.</p> <p>STRENDA DB requires entry of the buffer type, concentration and pH for every assay. If these are the same as the previous assay, default</p>

<p>..[substrate], and 25 μM (final) ..[substrate 2] (if present) in reaction buffer, while the second syringe contained 100 μM [$\alpha\text{-}^{32}\text{P}$]ATP in reaction buffer.” Unlike “standard buffer”, “reaction buffer” is not defined elsewhere in the paper or Supplementary Information.</p> <p>The last extract, and also a figure legend, refer to “a saturating concentration of ..[substrate].”, but no information is given as to what this was, and how the authors know it to be saturating.</p> <p>A table reports kinetic parameters for both substrates, without indicating whether these are true or apparent values. For a two substrate enzyme there will be one true k_{cat} value and a true K_{M} value for each substrate. It is possible to obtain apparent K_{M} and k_{cat} values when one substrate concentration is held constant and the other is varied. However, these values depend on the arbitrary choice of fixed concentration for the other substrate, and so are only meaningful if this is stated. Because two k_{cat} values are stated, it is likely that the values are apparent ones, but the paper gives no information on the chosen fixed substrate concentrations.</p> <p>In a table and the text there are reported rates of a pre-steady state process stated to show exponential and linear phases. Rates of both these phases are given in units of s^{-1}. For an exponential process the meaning is clear as an apparent first order rate constant. But for the linear phase the meaning needs to be clarified – probably it means mol reaction per mol enzyme per s.</p>	<p>values can be left in place – but these can be edited where necessary.</p> <p>STRENDA DB requires entry of substrate concentrations, either a fixed value or a range. To make an independent confirmation of saturation would require initial rate data for various substrate concentrations, which cannot currently be entered in STRENDA DB.</p> <p>A planned enhancement to STRENDA DB is the entry of the equation fitted to data in order to estimate kinetic parameters. STRENDA DB currently requires entry of substrate concentrations used, either as a fixed value or a range.</p> <p>STRENDA DB does not currently allow entry of pre-steady state kinetic parameters, but this is a possible extension.</p>
<p>Paper 7.</p> <p>Describing kinetic experiments on the enzyme, the authors state “After suitable period of time (2-80 min, depending on</p>	<p>The current STRENDA DB version</p>

<p>the activity), the reactions were simultaneously quenched". This implies that time courses were studied, which would allow estimation of initial rates and from them, kinetic parameters. But inspection of the results indicates that most k_{cat} and K_M values quoted are actually based on a single time point. However, it seems that progress was not actually fully linear over the time range used, as evidenced by a plot showing a rather poor fit to the Michaelis-Menten equation when rates assumed constant over different time periods are combined. A relatively subtle problem, but we really need raw time and conversion data to allow better analysis of kinetics.</p> <p>A figure gives "rate" in units of min^{-1}. This may be a turnover frequency (mol product formed per mol catalyst per min), but this is not clear. Rates are normally reported as concentration or molar amount per unit time.</p>	<p>would not check this, indeed it does not currently allow entry of single time point data. Future enhancements do include more details of data from which k_{cat} and K_M values are estimated. (Relevant details might be captured by the existing free text "Methodology" field.)</p> <p>STRENDA DB does not currently accept rate data, but this is a planned enhancement. Clear units are enforced for all data entered.</p>
<p>Paper 8</p> <p>In describing an assay, the authors write "30 pmol of 5' fluorescein labeled primer (.....) were annealed with 30 pmol of template (.....) and 0.4 μg of polymerase by heat denaturation at 90°C for 1 minute and allowing to cool to room temperature. Reactions were initiated by the addition of "start" mix which contained 50mM Tris-HCl (pH8.4), 10 mM $(\text{NH}_4)_2\text{SO}_4$, 10 mM KCl, 2 mM MgSO_4 and 200 μM dNTPs." Here amounts of nucleic acid and template are presented, but with no indication of assay volume. Hence the reaction cannot be repeated with the correct concentrations. The same problem is found in two further assay descriptions in this paper.</p> <p>In another assay there is no mention of the amount or concentration of enzyme used, but the results are given in</p>	<p>STRENDA DB requires concentrations for all assay components as well as a defined assay temperature</p> <p>STRENDA DB requires enzyme concentration (and also a statement</p>

<p>units of nmol/s - this is meaningless without knowledge of the amount of enzyme present.</p> <p>In a radiochemical assay the authors report “Reactions (100 μL) consisting of assay buffer, 1 mM MgSO₄, and 500 nM duplex were initiated by variable amounts of α-P32-dCTP (0.003-400μM), which was diluted 1:400 in unlabeled dCTP.” This is not completely clear as to the final dCTP concentration. Is this 0.003-400 μM, or is it actually 400 times higher? (The latter is unlikely but possible).</p> <p>Two figures show values of k_{cat} and K_M without units.</p> <p>No error estimates are given for reported k_{cat} and K_M values, and the number of replicates is not stated.</p>	<p>on how this was measured).</p> <p>STRENDA DB requires concentrations of all substrates.</p> <p>STRENDA DB requires units for all kinetic parameters.</p> <p>STRENDA DB requires error estimates. The possibility to report more detail on the basis for these error estimates is a possible enhancement.</p>
<p>Paper 9</p> <p>An assay buffer is described as “50 mM GPT buffer (50 mM glycine, 50 mM phosphate, and 50 mM Tris), pH 7.6”. This is not clear. It could be a mixture of glycine and Tris base adjusted to pH with phosphoric acid, but the total phosphate concentration in this case would be less than 50 mM. Hence there are probably also other counter-ions present, such as Na⁺ and/or Cl⁻.</p> <p>In a later assay description, no buffer is mentioned. Since at least two different buffers are reported for other enzyme assays, we cannot guess which is used (or whether it’s actually a third).</p> <p>Studies of the effect of temperature over a significant range</p>	<p>STRENDA DB requires entry of buffer counter-ions if help recommendations are followed.</p> <p>STRENDA DB requires entry of the buffer type, concentration and pH for every assay. If these are the same as the previous assay, default values can be left in place – but these can be edited where necessary.</p> <p>Details of the treatment of pH are</p>

<p>are reported, but without necessary detail on the treatment of pH (see paper 5).</p>	<p>planned as enhancements for future releases of STRENDA DB.</p>
<p>Paper 10</p> <p>In an assay description we find “The mixtures were diluted into 50 mM Tris-HCl, 5 mM NaCl, pH 7 and mM ascorbic acid” A value is missing before “mM ascorbic”.</p> <p>Later, the authors report “The mixtures were diluted into 50 mM Tris-HCl, 5 mM NaCl, pH 7 and [the compound] under investigation (HEPES, MES and MOPS @ 500 mM, ascorbic acid @ 100 mM, or Tris alone @ 50 mM) to a final concentration of 16.7 nM protein and a final volume of 300 μL.” The wording suggests that test compounds were added to a buffer previously adjusted to pH 7. Adding these as a single compound could substantially alter pH. But perhaps the additives were pH adjusted, in which case details of the counter-ions present are missing.</p> <p>In describing an assay there is “The mixtures were diluted into 500 mM HEPES, pH 7”, but no mention of the cation in the HEPES buffer.</p> <p>Figures in the paper and in Supplementary Materials show that the progress of the reaction studied is clearly not linear. However, data is presented as turnover frequency (min^{-1}) averaged over the full experimental period. Really what should be reported is observed number of turnovers after a given reaction time.</p>	<p>STRENDA DB requires concentrations for all assay components.</p> <p>STRENDA DB requires entry of full details of buffer species and counter-ions, provided help recommendations are followed. If these species were assigned as “Other components”, user would still have to select a defined chemical compound to enter.</p> <p>STRENDA DB requires entry of assay pH, which should be the final value in the assay.</p> <p>STRENDA DB requires entry of buffer counter-ions if help recommendations are followed.</p> <p>STRENDA DB does not currently accept input of single time point data, but this is a planned enhancement.</p>
<p>Paper 11</p> <p>An assay mixture is described as comprising “20 mM MOPS-KOH, pH 7, 8 mM MgSO_4, 3 mM ATP”. In this case where the ATP concentration is within an order of magnitude of the</p>	<p>STRENDA DB requires entry of a defined chemical compound for all assay components. Hence the</p>

buffer concentration, the salt form or counter-ions of the added ATP should be given. The counter-ions may have a significant effect, and a significant pH effect is also possible.	appropriate salt of ATP would have been selected. The final assay pH is also required, which should be the value after all components are added.
---	--

136 Footnote to table: A reviewer has pointed out that the IUBMB Recommendations for ‘Symbolism
137 and Terminology in Enzyme Kinetics (Recommendation 1981)’
138 <http://www.sbcs.qmul.ac.uk/iubmb/kinetics/index.html> imply that k_{cat} should only be used where it
139 can be calculated using the measured active site concentration in the enzyme preparation used. This
140 would imply that the presentation of results as k_{cat} in papers 2, 3, 4 and 8 is wrong. Since papers 2-4
141 do make clear the way protein concentration has been measured, their experiments are
142 reproducible in this respect. However, their reported k_{cat} values may only set a minimum for the true
143 microscopic rate constant.

144

145

146 **Data omissions in published papers by authors of this article**

147 The papers analysed in Table1 are presented anonymously, because we do not suggest they are in
148 any way unusual in having these deficiencies. Our next step is to show similar omissions in some
149 published papers by co-authors of this article to illustrate the complexity of reporting enzyme
150 function data.

151

152 **Table 2. Omissions in published papers by authors of this article.** We also show where STRENDA DB
153 would have avoided those omissions.

Issues found in papers	Where could STRENDA DB help?
Michele Cianci, Bartlomiej Tomaszewski, John R. Helliwell, and Peter J. Halling (2010) “Crystallographic Analysis of Counterion Effects on Subtilisin Enzymatic Action in Acetonitrile” J. Am. Chem. Soc. <i>132</i> , 2293–2300. In Materials & Methods, under “Cs Soaking, Cross-Linking, and Acetonitrile	

<p>Wash” there is “So the soak solution was prepared freshly by mixing 2 volumes of 25% aqueous glutaraldehyde with 3 volumes of 250 mM Na-cacodylate buffer pH 6.5 containing 25% v/v glycerol and, where required, 1.5 M salt. Crystals were soaked in this for 15 min, with vortexing in the case of suspensions of multiple crystals for activity assay. Washing in acetonitrile followed immediately.” The presentation of the buffer and pH does not make it clear whether:</p> <p>a) 250 mM Na-cacodylate pH 6.5 was first prepared, then 25% glycerol was added, or</p> <p>b) pH 6.5 is an apparent pH reading in the aqueous-glycerol mixture from an electrode calibrated in dilute aqueous buffer, or</p> <p>c) pH 6.5 refers to an absolute pH scale established for 25% aqueous glycerol, with a protocol using appropriate standards in this mixture.</p>	<p>STRENDA DB requires entry of the pH, which should be the final value in the assay mixture. Further details of the treatment of pH are a planned enhancement.</p>
<p>Richard S. Hall, Rakhi Agarwal, Daniel Hitchcock, J. Michael Sauder, Stephen K. Burley, Subramanyam Swaminathan and Frank M. Raushel (2010) “Discovery and Structure Determination of the Orphan Enzyme Isoxanthopterin Deaminase” <i>Biochemistry</i> 49, 4374-4382. Under “Measurement of Enzymatic Activity”, there is “The coupling system contained 7.4mM alpha-ketoglutarate, 0.4mM NADH, 6 units of glutamate dehydrogenase, and 100 mM HEPES (pH8.5) in a final volume of 0.25mL.” The cation present in the HEPES buffer is not stated. Also, the range of substrate concentrations used to measure the kinetic constants was not clearly reported.</p>	<p>STRENDA DB requires entry of buffer counter-ions if help recommendations are followed. In addition, STRENDA DB requires entry of substrate concentrations, either a fixed value or a range.</p>
<p>Wolfgang E. Schäfer, Johann M. Rohwer and Frederik C. Botha (2004) “A kinetic study of sugarcane sucrose synthase”. <i>Eur. J. Biochem.</i> 271, 3971–3977.</p> <p>In the description under “SuSy assays”, the assay temperature is not specified.</p>	<p>STRENDA DB requires the assay temperature.</p>

Table 1 does not provide error estimates for K_i values.

STRENDA DB requires error estimates for all kinetic constants entered.

154

155

Missing data statistics of papers used as sources for SABIO-RK

156

The SABIO-RK database is an expert-curated collection of enzyme kinetic values, of particular value

157

in modelling multi-enzyme networks and metabolism (Wittig et al, 2018). Wittig et al. (2014ab)

158

recently investigated publications on enzyme function with the aim to detect omissions that

159

complicate the efficient data extraction for SABIO-RK. This analysis showed a number of issues that

160

prevent readers from repetition of the experiments such as missing temperature or pH values at

161

which the assays have been carried out. The major focus however was on problems that

162

complicated the abstraction of data into the database, such as the absence of unique identifiers for

163

enzymes (e.g. UniProtKB codes). Such issues would often not prevent a dedicated expert from

164

reproducing the experiments, but are a serious impediment for database curators, and researchers

165

who wish to use data from many publications, for example in systems biology models. Other

166

problems of this type identified were:

167

i. required information is scattered in different parts of the paper, and between text, tables

168

and figures

169

ii. no use of recognised unique Identifiers or controlled vocabularies for source organism,

170

compounds used in assay, etc

171

iii. a reference (which often leads to a sequence of further references) must be consulted to

172

find details of methods

173

iv. values are given in non-standard units, e.g. mg/mL rather than M

174

v. assay concentrations are not stated, but must be calculated from separate values of amount

175

and volume

176

vi. full reaction catalysed is not stated, e.g. products not named

177

It should be noted that the consistent use of STRENDA DB will also deal with issues of this type.

178

In this paper, however, we focus on issues that would definitely prevent repetition of the work

179

(Table 3). These will clearly also affect the meaning and transferability of the results reported. It is

180

clear that omissions of this type are common. Previously analysis by Wittig et al (2014ab) indicated

181

that such omissions are not becoming less frequent over time. The larger test set of papers analysed

182 here allows us to make more confident comments about the most common types of omission, which
183 hence should be priority targets for checking.

184

185 Table 3: Statistics for omissions based on recent papers used in SABIO-RK database

Missing information	Found in papers (%)
No reaction pH	4
No temperature	11
States "room temperature"	5
No information on buffer	3
No enzyme concentration	45
No substrate concentrations	11
"Tris buffer" with no counter-ion stated	22

186

187 The 100 papers analysed for this table were all from one of the 10 journals most represented in the
188 SABIO-RK database (J Biol Chem, Biochemistry, Biochem J, Eur J Biochem, Biochim Biophys Acta,
189 Arch Biochem Biophys, J Bacteriol, FEBS Lett, Plant Physiol, Proc Natl Acad Sci U S A). They are the
190 100 most recent (as of September 2017), with the earliest dating from 2008.

191

192 The most common omission is clearly the failure to report the concentration of the enzyme used.
193 Authors may have allowed for this in reporting a V or k_{cat} value, but the actual value used is
194 necessary, for example to be sure that studies were in the range where the reaction rate is
195 proportional to enzyme concentration.

196 The next most common omission is failing to report the counter-ion used in a buffer, which was also
197 identified in 5 out of the 11 papers examined more closely in Table 1. Any ions present in the
198 reaction medium can affect the enzyme behaviour, so these should be identified. Common lab
199 parlance will refer to "phosphate buffer" or "Tris buffer", but this should not allow counter-ions to
200 be omitted in formal reporting.

201 Failure to report substrate concentrations used (usually a range) was also found in 4 out of the 11
202 papers. When estimating K_M or to "ensure saturation", these are needed to test the validity of the
203 conclusions. In the case of two-substrate enzymes it is not trivial to ensure saturation with both
204 substrates, and kinetic constants obtained varying only one substrate are typically apparent values
205 dependent on the fixed concentration of the other substrate.

206 Some other issues identified in several of the 11 papers would not have been spotted in our analysis
207 of the SABIO-RK papers.

208 A value of pH may be given, but its meaning is not clear. In two papers it is not clear whether this is
209 the final reaction pH or that of a buffer to which significant acidic or basic compounds are then
210 added. Ideally the final reaction pH should be reported. But at least it should be completely obvious
211 at what stage the quoted pH value was measured or set. In another two papers there are issues
212 about the treatment of pH for experiments conducted at a significant range of temperatures. The
213 most common protocol is to measure or set the pH of a buffer at a reference temperature (typically
214 about 20 °C), then heat or cool to the chosen assay temperature. This can be reproduced, but is
215 theoretically not ideal, because the pH will change significantly with temperature, particularly for
216 some buffer species. A better option, but less used, is to calibrate the pH electrode with an
217 appropriate standard at each temperature, then measure or adjust buffer pH at the same
218 temperature. When it is not clear which approach is adopted, the experiments cannot be
219 reproduced precisely.

220 In two papers kinetic parameters are given without units or with only relative values. Using values
221 relative to some reference can often be a useful way to present results, but the absolute reference
222 value should always be stated, so that absolute values under any conditions can be calculated if
223 wanted. Papers like these would not be incorporated in SABIO-RK, as the parameters are not usable.

224

225

Conclusions

226 We find that 11 recent papers reporting research on enzyme function, chosen non-selectively, all do
227 not provide enough information to reproduce their findings. Our detailed analysis is backed up by
228 statistics from 100 papers used as sources for the SABIO-RK database. Therefore, we can be
229 confident that a large fraction of published papers have similar critical omissions important for
230 replication purposes. It is clear that the field would benefit from systems ensuring that necessary
231 data and meta-data are properly recorded, so that others can use results or repeat experiments in
232 confidence. It is obvious that paper-based reporting guidelines will not be widely applied, neither by
233 the authors nor by the journals, as discussed in the introduction. However, the use of an electronic
234 tool such as STRENDA DB should increase the chances of the adoption of and reporting in
235 compliance with guidelines. Moving towards deposit of data, at the time of publication, in open and
236 trusted repositories appears to be a natural and obvious step in research and publication policy. This
237 should improve the quality, openness and transparency of experimental enzyme function data in the

238 literature, especially if planned enhancements are implemented, including algorithms that validate
239 datasets as complete.

240

241

References

242 Apweiler R, Cornish-Bowden, A, Hofmeyr J-HS, Kettner C, Leyh TS, Schomburg D and Tipton KT
243 (2005) The importance of uniformity in reporting protein-function data. *Trends Biochem. Sci.* **30**:11-
244 12.

245 Apweiler R, Armstrong R, Bairoch A, Cornish-Bowden A, Halling PJ, Hofmeyr J-HS, Kettner C, Leyh TS,
246 Rohwer J, Schomburg D, Steinbeck C and Tipton KT (2010) A large-scale protein-function database.
247 *Nature Chemical Biology* **6**: 785. DOI:10.1038/nchembio.460.

248 Gardossi L, Poulsen PB, Ballesteros A, Hult K, Švedas VK, Vasić-Rački Đ, Carrea G, Magnusson A,
249 Schmid A, Wohlgemuth R, Halling PJ (2010) Guidelines for reporting of biocatalytic reactions. *Trends*
250 *Biotechnol.* **28**, 171-180.

251 Neil Swainston, Antonio Baici, Barbara M. Bakker, Athel Cornish-Bowden, Paul F. Fitzpatrick, Peter
252 Halling, Thomas S. Leyh, Claire O'Donovan, Frank M. Raushel, Udo Reschel, Johann M. Rohwer,
253 Santiago Schnell, Dietmar Schomburg, Ming-Daw Tsai, Hans V. Westerhoff, Ulrike Wittig, Roland
254 Wohlgemuth and Carsten Kettner (2018) "STREND A DB: enabling the validation and sharing of
255 enzyme kinetics data" *FEBS Journal*, **285**, 2193-2204. DOI: 10.1111/febs.14427

256 Keith F. Tipton, Richard N. Armstrong, Barbara M. Bakker, Amos Bairoch, Athel Cornish-Bowden,
257 Peter J. Halling, Jan-Hendrik Hofmeyr, Thomas S. Leyh, Carsten Kettner, Frank M. Raushel, Johann
258 Rohwer, Dietmar Schomburg, Christoph Steinbeck (2014) Standards for Reporting Enzyme Data: The
259 STREND A Consortium: What it aims to do and why it should be helpful. *Perspectives in Science* **1**(1-
260 6):131.-137. DOI:10.1016/j.pisc.2014.02.012

261 Ulrike Wittig, Renate Kania, Meik Bittkowski, Elina Wetsch, Lei Shi, Lenneke Jong, Martin
262 Golebiewski, Maja Rey, Andreas Weidemann, Isabel Rojas, Wolfgang Müller (2014a) "Data extraction
263 for the reaction kinetics database SABIO-RK" *Perspectives in Science* **1**, 33–40

264 Ulrike Wittig, Maja Rey, Renate Kania, Meik Bittkowski, Lei Shi, Martin Golebiewski, Andreas
265 Weidemann, Wolfgang Müller and Isabel Rojas (2014b) "Challenges for an enzymatic reaction
266 kinetics database" *FEBS Journal* **281**, 572–582

267 Ulrike Wittig, Maja Rey, Andreas Weidemann, Renate Kania, Wolfgang Müller (2018) SABIO-RK: an
268 updated resource for manually curated biochemical reaction kinetics. *Nucleic Acids Res*
269 **46**(D1):D656-D660

270

271 This research did not receive any specific grant from funding agencies in the public, commercial, or
272 not-for-profit sectors.