Validity of machine learning in biology and medicine increased through collaborations across fields of expertise

Article *in* Nature Machine Intelligence · January 2020

DOI: 10.1038/s42256-019-0139-8

20 authors, including:

Maria Littmann
Technische Universität München

Liel Cohen-Lavi
Ben-Gurion University of the Negev

Evans Kataka
Technische Universität München

Anja Mösch
Technische Universität München

Validity of machine learning in biology and medicine increased through collaborations across fields of expertise

Maria Littmann[1,*,◊], Katharina Selig[2,*,◊], Liel Cohen[3,4], Yotam Frank[5], Peter Hönigschmid[6], Evans Kataka[6], Anja Mösch[6], Kun Qian[7,8], Avihai Ron[9], Sebastian Schmid[10], Adam Sorbie[11], Liran Szlak[12], Ayana Dagan-Wiener[13], Nir Ben-Tal[14], Masha Y. Niv[13,15], Daniel Razansky[9,16,17], Björn W. Schuller[18], Donna Ankerst[2], Tomer Hertz[3,4,19] & Burkhard Rost[1,20]

1 TUM (Technical University of Munich) Department of Informatics, Bioinformatics & Computational Biology - i12, Boltzmannstr. 3, 85748 Garching/Munich, Germany
2 TUM Department of Mathematics, Boltzmannstr. 3, 85748 Garching/Munich, Germany
3 Department of Microbiology, Immunology and Genetics, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel National Institute for Biotechnology in the Negev
4 National Institute of Biotechnology in the Negev, Beer-Sheva, Israel.
5 The Blavatnik School of Computer Science, Tel-Aviv University, Ramat Aviv 69978, Israel
6 TUM Department of Bioinformatics, Wissenschaftszentrum Weihenstephan, Maximus-von-Imhof-Forum 3, 85354 Freising, Germany
7 TUM Chair of Human-Machine Communication, Theresienstraße 90, 80333 Munich, Germany
8 Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
9 Institute for Biological and Medical Imaging, TUM and Helmholtz Center Munich, Germany
10 TUM Chair of Food Chemistry and Molecular Sensory Science, Lise-Meitner Str. 34, 85354 Freising, Germany
11 TUM Chair of Nutrition and Immunology, Gregor-Mendel-Str. 2, 85354 Freising, Germany
12 Weizmann Institute of Science, Rehovot 7610001, Israel
13 The Institute of Biochemistry, Food and Nutrition, The Robert H Smith Faculty of Agriculture, Food and Environment, The Hebrew University, 76100 Rehovot, Israel
14 Department of Biochemistry and Molecular Biology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, 69978 Tel Aviv, Israel
15 The Fritz Haber Center for Molecular Dynamics, The Hebrew University, Jerusalem, Israel
16 Faculty of Medicine and Institute of Pharmacology and Toxicology, University of Zurich, Switzerland
17 Institute for Biomedical Engineering and Department of Information Technology and Electrical Engineering, ETH Zurich, Switzerland
18 Group on Language, Audio & Music (GLAM), Imperial College London, UK
19 Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA.
20 Institute for Advanced Study (TUM-IAS), Lichtenbergstr. 2a, 85748 Garching/Munich, Germany & TUM School of Life Sciences Weihenstephan (WZW), Alte Akademie 8, Freising, Germany & Columbia University, Department of Biochemistry and Molecular Biophysics, 701 West, 168th Street, New York, NY 10032, USA New York Consortium on Membrane Protein Structure (NYCOMPS) & Department of Biochemistry and Molecular Biophysics, Columbia University, 701 West, 168th Street, New York, NY 10032, USA

* Corresponding authors: Maria Littmann: littmann@rostlab.org, http://www.rostlab.org/, Tel: +49-289-17-814 (email rost: assistant@rostlab.org); Katharina Selig: katharina.selig@tum.de

◊ These authors share first authorship

Running Title: ML in Life Sciences
Doc stats 1st: Abstract: 215 words, Text: 2878 words, References: 23, figures: 5; tables: 0; pages: 12
Doc stats R1: Abstract: 193 words, Text: 2893 words, References: 23, figures: 5; tables: 0; pages: 22
Journal: Nature Machine Intelligence

Submitted: 2019/08/09, R1: 2019/11/08

Abstract

Machine Learning (ML) has become an essential asset for the life sciences and medicine. We selected 300 articles describing ML applications from 17 journals sampling 26 different fields between 2011 and 2018. Independent evaluation by two readers highlighted three results. First, only half of the articles shared software, 64% shared data, and 81% applied any kind of evaluation. Although these aspects are crucial to ensure validity and reliability of ML applications, they were met more by publications in lower-ranked journals. Second, the authors' scientific background highly influenced how technical aspects were addressed: reproducibility and computational evaluation methods were more prominent with computational co-authors; experimental proofs more with experimentalists. Third, 73% of the ML applications resulted from interdisciplinary collaborations comprising authors from at least two of the three disciplines: computational sciences, experimental biology, medicine. _deleted_ The data suggested collaborations between computational and experimental scientists to generate more computationally sound and impactful work integrating knowledge. Furthermore, such collaborations provide opportunities to both sides: computational scientists are given access to novel and challenging real-world biological data increasing the scientific impact of their research, and experimentalists benefit from more in-depth computational analyses improving the technical correctness of work.

Key words: machine learning, life sciences, medicine, open access, open data, interdisciplinary research, sustainable research, standardization

Abbreviations used: AI, Artificial Intelligence; **ML**, Machine Learning; **NAR,** Nucleic Acids Research; **NC**, number of citations; **NC/year**: number of citations normalized by number of years since publication; **NEJM,** New England Journal of Medicine; **PNAS**, Proceedings of the National Academy of Sciences.

NOTE editor & reviewers:

Major changes in red font, major deletions marked as _DELETED_

Introduction

Growing importance of Machine Learning (ML). Large amounts of experimental data triggered by technological advances are increasing the interaction between biology, medicine and quantitative sciences[1-3]. For instance, the amount of genome sequencing data is growing exponentially while data storage capacity only grows linearly[4]. Numerous large databases in molecular biology and large clinical datasets increasing through electronic health records call for novel ways to interrogate, analyze, and process biological and biomedical data for gaining biological and medical insights[5].

Machine Learning (ML) automatically identifies patterns and regularities in existing data to accurately predict for unseen data[6]. Despite the complexity of the underlying mathematical concepts, ML has attracted broad attention even outside the research community: querying Google Trends[7] with "machine learning" demonstrated an exponential increase over the last decade (01/2010-02/2019, data not shown). This general rise has been mirrored in many fields of biology and medicine, i.e. the life sciences[8-11] although keeping track with the rapid evolution of artificial intelligence (AI) challenges even those applying ML [12]. Typically, large biological or medical datasets enable the development of ML models that can be used to predict biological or clinical phenotypes through measurements from novel samples.

Quality and validity of ML models hinge upon two primary factors: (1) size, quality and universal validity of data, and (2) the correct development and assessment of the resulting methods[5,13]. Successful ML applications extract generic principles from today's data, allowing the generalization, i.e. accurate prediction, for tomorrow's data. This needs proper extraction and processing of data and features often requiring expert knowledge[14-16]. The development and application of ML models to the life sciences needs expertise from both computational and biological/medical fields. In contrast, ML applications to areas such as object and speech recognition or complex games (including chess and Go) for which task and success are more clearly defined and thus require mainly expertise in ML.

Interdisciplinary research might have more impact. In many fields of science, interdisciplinarity has become crucial to produce groundbreaking results through the integration of approaches from different disciplines[17,18]. Several recent studies [17,19-24] have been investigating the role of interdisciplinarity by automatically extracting tens and hundreds of thousands of publications (e.g. from WoS [25] or PNAS). Toward this end, one definition of *interdisciplinarity* is through the field of the journal in which they are published compared to the journal in which they are cited (the US NSF [26] classifies journals into 14 different disciplines and 143 subdisciplines[17,21]; if published and cited in different fields or subfields, the article is deemed "interdisciplinary" [17,21,24]). Other definitions[19,20] focus on the author's field and define interdisciplinary articles as work published by authors from different fields. So far this definition has been limited to Italian scientists for who there is a public directory mapping researcher and field[19,20].

The scientific impact of an article is usually measured by the number of citations for this article[17,24]. To correct for field- and journal-specific effects that number is often normalized by taking average citation rates and a journal's impact factor into account[23,24]. Since the impact factor is calculated from the number of citations of articles published in this journal[27], articles from higher-ranked journals are expected to have higher citation counts.

All those automated studies allowed the assessment of many articles while being limited to the extraction of only particular type of information. The studies disagree in their findings regarding the importance of interdisciplinary collaborations: one [24] finds no consistent correlation between impact and interdisciplinarity sampling over 750k publications: for some disciplines interdisciplinarity were proportional to citations, for others the relation was reversed xxbr: where is physics in this? Important for next sentence. Another work, focusing on xxbr_number publications from physics[23] found interdisciplinary was proportional to citation rates but only when published in journals with citation rates below average. Yet other studies xxbr_number [17] and xxbr_number publications [20] agreed that interdisciplinary creates higher impact than non-interdisciplinary work. Also, specific collaborations between scientists from related fields leads to higher-impact publications than generic collaborations between scientists from very different fields[20]. Clearly, there is no simple red line leading through all those findings. However, what made us re-open the can and begin our analysis were three other reasons: (1) the focus on the life sciences, not explicitly covered by others, (2) the aim of separating the analysis of scientific quality (soundness) and of impact, and (3) the introduction of a more rigorous definition of interdisciplinarity: instead of proxying by the number of disciplines citing a work, we require experts from different disciplines to co-author a work (incidentally, the same sort of definition was used for the analysis of Italian authors [19,20]).

Focus of this work. Here, we assessed several aspects of ML applications in the life sciences. We started with the selection of 17 journals representing computational/experimental biology and medicine (Materials & Methods: Supporting Online Material, SOM). Amongst all papers published in those 17 journals in the years 2011-2016, keyword searches (Table S1) matched in 4,306 articles, about 2,100 of those were deemed correct hits after a quick expert analysis. From those, initially 250 were randomly selected (*Materials & Methods* SOM; complete list in additional file paper_table.csv, list of identified falsely extracted articles is provided in additional file false_articles.csv). Subsequently, we applied the same selection process and chose another 50 papers from 2018 to verify that the major findings have not changed through the most recent advent of deep learning (xxbr: cite review here). In contrast to previous studies [17,19-24], our assessment focused on ML applications in the life sciences and all information we analyzed was manually extracted from the articles. This allowed, for instance, to correct the 50% false positives from the keyword searches, and also allowed to define interdisciplinarity through the author's background for non-Italians (simply by reading partial CVs for all 1,918 authors of the 250 papers). Each article was classified independently by two of us. These investments limited the number of papers analyzed but allowed a more fine-grained assessment not accessible to automatic extraction. Our focus had several implications, including that all papers reported applications of machine learning to the life sciences, as opposed to more theoretical treatments. In some sense the application of ML (computational sciences) to the life sciences is by definition interdisciplinary. However, we sharpened the perspective by distinguishing expertise from three different disciplines: computational sciences, experimental biology and medicine (expertise of author verified through CV, not through affiliation). Thus, papers could maximally be co-authored by authors from three disciplines, and minimally by one. To simplify, we loosely referred to the case of N=1 as to "non-interdisciplinary" and the case of N>1 as to "interdisciplinary". For some results, we also showed differences between N=2 and N=3.

The correct application of ML requires expertise from those familiar with ML and those familiar with the life sciences, i.e. different disciplines. Thus, we hypothesized articles written by research teams from different disciplines to be more likely to report the necessary evaluation methods ensuring proper implementation of ML methods, to make their data publicly available so others could validate their results, and subsequently, to be accepted in higher-ranked journals and have more citations.

Results and Discussion

Coverage of machine learning varies between journals and fields. 58% of the chosen 250 papers (*Material & Methods* in Supporting Online Material – SOM – for more details on how these articles were selected) appeared in only four of the 17 journals (by occurrence: *Bioinformatics*, *PNAS*, *PLOS Computational Biology* and *BMC Bioinformatics*), i.e. were 2.5–fold over-represented (xxbr: can we compute the surprise, simplest model: 0.58/(4/17) – but I am sure you ladies might come up with something more intelligent;). Most articles were cited fewer than 100 times, and the number of citations was proportional to time passed since publication (Spearman correlation coefficient ρ=-0.22, p-value = 0.03; **Fig. 1**, Fig. S1). The average number of citations for articles from Nature and Science (2011-2016) showed the same trend as that for all 250 articles (Fig. S1). Since the time-dependency obfuscated inter-year comparisons, we normalized by the number of years (*SOM Material & Methods*). As the number of citations correlated with the journal impact factor[28,29] (ρ=0.52, p-value<0.001, **Fig. 1**), all aspects correlating with the impact factor trivially correlated with the number of citations. Normalizing by year and impact factor, removed this correlation. We continued also using the impact factor to assess the visibility of an article as publications in higher-ranked journals tend to be downloaded more often from bioRxiv[30]. Xxbr: do we really need this addition in red? I for one cannot immediately see what exactly you are aiming at, and adding more than one sentence seems a lot. Do you mean that number of citations (btw. should we introduce “names” NC/NCperanno/NCcorrected?) reflect deflection by bioRxiv more than else)?

>>> Fig. 1 <<<

The number of articles differed highly between fields: the top five (*molecular biology* 26%, *genetics* 24%, *medicine* 14%, *oncology* 10% and *neuroscience* 9%) accounted for 76% of the 250 articles (Table S2, Fig. S2). Numbers varied even more by disciplines (author expertise): (xxbr: make sure we are NOT confusing discipline=1,2,3 interdisciplinary and field=previous sentence; whatever we settle on: we HAVE to stick to it to avoid further complications): Computational scientists co-authored 88% of all articles, and 95% of those from *genetics* (Fig. S3). Experimental biologists co-authored 70% of all and 59% in *medicine*. Physicians were primarily involved in articles from *medicine* and *oncology* (Fig. S3). Numbers of citations were largely similar for all fields (Fig. S4) but articles focusing on *medicine*, *neuroscience,* and *oncology* tended to be published in higher impact journals (Fig. S4). While the disciplines *experimental biologist* and *physician* correlated positively with impact factor (ρ=0.30/p-value<0.001, ρ=0.26/p-value<0.001, respectively), *computational science* correlated negatively (ρ=-0.30/p-value<0.001; Fig. 1). Computational scientists might focus more on methods, experimental biologists and physicians more on new data that tend to be highly cited in the life sciences.

Three levels of interdisciplinarity. By definition, all the papers analyzed applied methods from computational fields to the life sciences, i.e. were intrinsically interdisciplinary. Most likely all 300 papers analyzed would have been considered “interdisciplinary” by automated analyses checking from which field/discipline the article was quoted. To generate a more detailed lens, we distinguished three disciplines (computational scientists, experimental biologists, and physicians) and introduced interdisciplinarity as a number ranging from 1-3 depending on how many disciplines were represented by the authors of the work. Most of the 250 papers were co-authored by two disciplines (one: 27%, two: 53%, three: 20%). Given these levels, we could classify all papers according to their level of interdisciplinarity and differentially analyze the key indicators: validity (evaluation and sharing) and impact (NC: number of citations, NC/year, NC/year*journal impact factor).

Scientific validity higher with experts participating in collaboration. We proxied the **validity** of papers describing the application of machine learning (ML) methods to biology & medicine (the life sciences) through six different indicators. The first four relate to whether or not the method was assessed in ways needed to ascertain that it works as promised (or at all). We asked: did the authors use cross-validation (V1: binary value), more than one single measure for performance (V2: integer), additional test sets (V3: binary value), and additional experimental verification (V4: binary value). While method evaluation might correctly estimate performance for unseen data without V4, it appears impossible to accomplish this simple objective without V1-V3. The second two indicators related to sharing methods and results. These were sharing data (V5), programs and codes (V6) through publicly available sites. Typically, reviewing ML applications by journal reviewers and the public at large requires availability of data and programs in a form beyond what is available through what can be squeezed into writing. In ML it is almost impossible to imagine the development of the best possible method without any assessment (V1-V3). What if someone might have decided not to publish that assessment? On top, should the aspects of sharing (V5, V6) better be termed "reproducibility" than "validity". Given the rules of proper scientific conduct, we answer both questions in the negative arguing that without making the evaluation available or making the content of a publication reproducible, the work should either not be published as an application of ML or should be considered as invalid.

Evaluation methods (e.g. cross-validation), usage of independent test sets, and/or independent experimental proofs reduce the chance of overfitting and enhance the applicability of the model to future data. Indeed, 80% of the articles with only computational authors, applied some evaluation methods or independent tests; compared to 41% of those written by "experimentalists" (experimental biologists & physicians; Fig. 4). However, most articles written solely by experimentalists provided independent experimental proof (55%), so did 16% of those from only computational co-authors (Fig. 4). The corresponding numbers for interdisciplinary collaborations between computational and experimental scientists (level of interdisciplinarity≥2) were between these two extremes: 67% evaluated their methods, 43% provided independent experimental proof, suggesting that such collaborations facilitate experimental and computational validation. On the flip side: 19% of all articles did not provide any evaluation; this number rose as high as 34% without computational co-authors (Fig. 4). To put this most clearly: 19-34% of the papers should have never been accepted, because applications of ML without evaluation resemble "experiments" with no output and/or no measurement.

>>> Fig. 4 <<<

Several evaluation metrics are required to assess the performance of ML applications (e.g. precision, recall, accuracy or confusion matrices). 6% of all articles used no evaluation metric, 53% used one or two, and 6% used over five (Fig. S7). Although, more metrics do not necessarily imply better assessment, even for binary predictions (separation of two classes/classifications), we have to consider the predictive power of the model for both classes separately, i.e. minimally need two evaluation metrics. More complex problems require more evaluation metrics. Typically, clearly more than two metrics are needed to show different strengths and weaknesses of a prediction method. To put the number five metrics (6% used ≥5), none of the ML applications with >400 citations published by the most senior co-author of this manuscript used fewer than eight different metrics.

Slightly more than half (52%) of the articles compared their method to others; this again dropped to 21% without computational co-authors (p-value = 0.001; Fig. 2). Although method comparisons are crucial for validation, they might add complications leading to acceptance in lower-ranked journals (Fig. 3C) and possibly to lower impact (Fig. 3A; although adjusting also by impact factor suggested a slight pay-off from method comparisons: Fig. 3B).

>>> Fig. 3 <<<

Reproducibility is a major pillar of science[31-33] partially relying on making data and methods publicly available. It is particularly critical for ML applications because many minor technical details may invalidate results[27]. Overall, 64% of the articles shared their data (V5; with large variation between journals: from NAR=89% to NEJM=8%, Fig. S5) reflecting the general trend that articles from *medicine* shared data the least (Fig. S6). We could not establish whether or not this related to sensitive patient data. While all journals encourage data sharing, many do not enforce it.

Overall, 68 % of the articles with computational scientists shared data, opposed to 31% without (p-value < 0.001; Fig. 2 _deleted_). 57% of the articles relied on data extracted from public resources or previous articles. However, 22% of those who did, did not publish their data. Data sharing was highest for collaborations with computer scientists (xxbr: as discussed: have not quite seen this plot, we should look separately at data sharing -ds- for computer scientists CS, exp biol EB, clinicians/physicians CP, and interdisciplinary=2 (EB+CP vs EB+CS + CP+CS), and interdisciplinary=3 (all), here I argue that we see interdisciplinary+CS always best, if not: we'll have to refine the statement).

>>> Fig. 2 <<<

Collaborations of scientists with different expertise somehow cited more often. Interdisciplinary collaborations of researchers from different fields seem increasingly important to generate new ideas and results[34,35]. The higher the level of interdisciplinarity, the higher the NC/year (number of citations divided by number of years since publication; ρ=0.22, p-value=0.02; Fig. 1, Fig. S8A) and the higher the impact factor (ρ=0.24, p-value=0.002; Fig. 1, Fig. S8C). When adjusting the number of citations also by impact factor, the correlation was no longer significant (Fig. 1, Fig. S8B) suggesting that interdisciplinary articles were cited more mainly because they were published in higher-ranked journals (Fig. S8C). The correlation between impact factor and level of interdisciplinarity (Fig. S8C) suggested that authors profit from collaborations.

>>> Fig. 5 <<<

Closer analysis of the correlation between interdisciplinarity and impact refined the message: distinguishing just two groups (*computational* and *experimental*), revealed NC to be higher for research teams of only experimental scientists (Fig. 5A), and this outcome was largely caused by physicians (xxbr: point to a figure that makes that point or keep it out?). The results for impact factor and NC adjusted by impact factor suggested that the higher NC originated essentially from physicians publishing in higher-ranked journals (Fig. 5B and C).

Did scientific validity (evaluation and sharing) correlate with impact? Computational evaluations correlated negatively with the impact factor (ρ=-0.31, p-value<0.001); using no evaluation method correlated positively with the impact factor (ρ=0.23, p-value=0.004), but we could not detect a significant relationship between impact factor and experimental proof (Fig. 1). Since all articles analyzed here focus on applications, the absence of proper evaluation - independent of the focus of a paper - clearly contradicts good scientific conduct.

Data sharing was not rewarded by increases in NC or NC/year (Fig. 3A), although adjusting also by impact factor hinted at a tendency that sharing leads to more citations (Fig. 3B). Thus, although data sharing is crucial to ascertain validity and reproducibility, it is not incentivized by increased visibility. In fact, there was no significant difference in the impact factor (Fig. 3C).

Software sharing also did not correlate with NC/year (Fig. 3A; the trend changed toward more cited when adjusting NC by impact factor: Fig. 3B). On the contrary, **not** sharing software seemed to lead to acceptance of articles in higher-ranked journals, but again the difference was not significant (Fig. 3C). Certainly, method sharing is crucial for reproducibility and for the impact of a method on science. Therefore, we were surprised that program sharing appeared neither crucial for visibility nor acceptance in the research community as proxied by citations and journal rank. Ultimately, this might shed light on the limitations of such measures to evaluate scientific impact.

More computational scientists involved in 2018. Artificial intelligence (AI) and ML are so rapidly evolving that papers published from 2011-2016 might simply not be up-to-date enough to capture the newest trends. We attempted to address this worry by analyzing another 50 articles describing ML applications to the life sciences published in 2018 (selected and analyzed largely by the same criteria as the other 250, Material & Methods in SOM for details). The major differences were: fewer publications without computational scientists (6% 2018 vs. 12% 2011-2016), and program sharing rose (70% vs. 50%). Although data sharing did not change significantly (68% vs. 64%), those papers that shared data were cited more often and accepted to higher-ranked journals, we could not detect a significant difference (Fig. S9). T
Other aspects did not change significantly, neither program sharing, nor the fact that papers sharing programs tended to be published in lower-ranked journals (Fig. S9), nor the proxies for impact (e.g. NC, NC/year, impact factor, NC normalized by impact factor). Overall, the most substantial change was that computational scientists contributed more often in 2018. This might reflect the increasing complexity of realizing ever more popular deep learning-type solutions of ML.

Limitations. Although our analysis revealed interesting insights, some issues remain to be addressed in the future. First of all, thoroughly analyzing more than 300 articles will render the conclusions more valid. The problem might be one between the Scylla of two few papers and Charybdis of too unreliable analysis. Our solution fell victim of the first, while other solutions xxbr quote-may be NOT incl. the Italians here fell victim of the later. Secondly, we proxied impact and visibility through number of citations and the impact factor. However, other factors are also influencing the number of citations that can seem superficial and can be controlled by the authors[36] and it is hard to compensate for these factors. Using the impact factor for measuring scientific impact has been criticized in the literature and the increasing use of social media might increase the visibility of research independent of the journals impact factor[37,38]. Thirdly, the scope of a journal might influence the description of ML applications. Journals focusing on methodologies are more likely to require certain standards in ML, those focusing on biological and medical relevant novelties are less likely to specifically ask for methodological details. Xxbr: remove (argument: this is peer-reviewed; put up or shut up!) ~~Also, the assumption that articles not reporting evaluations did not evaluate is over-simplified.~~ _DELETED_ Fourthly, we considered any publicly available information to assign author fields, but could not account for paid statisticians not listed as authors. A variety of medical scientists from pathologists to clinicians were all simplified as *physician* ignoring large differences in scientific training. These simplifications might lead to under-estimate computational expertise in publications. Furthermore, we considered data and program availability as stated in the articles, but did not attempt to contact authors to obtain those if not available. Finally, since several aspects in our analysis correlated with the impact factor and they also correlated with each other, confounding factors might influence the results and these interrelationships are difficult to separate.

Xxbr: the following I suggest to remove, NOT because I don't like it, but because (1) it doesn't really fit here and I do not know how to replace it, and (2) too long anyway. Reason to keep: is simple and some of it is bla & others like more bla than I! ~~For research teams with only computational expertise, contributions from physicians or colleagues with expertise in wet lab experiments can help to add new data, find biologically relevant applications and interpretations of the results, and increase the relevance of ML applications leading to more visibility of conducted research because it might be accepted in higher-ranked journals. Involving computational scientists in their work does not increase the visibility of research for physicians or experimental biologists because this work is rather accepted in lower-ranked journals. However, they might benefit from colleagues with knowledge in computer science to add evaluation methods, bring a greater variety of tools, and help with the interpretation of the scientific and statistical significance of results. Therefore, the results focus more on technical aspects making it possibly less intuitive for a broader research community but increasing its scientific value by achieving more technical correctness.~~

~~Most likely, with the introduction of new high-dimensional datasets and high-throughput technologies, the need for collaborations will increasingly grow. As the enforcement of data and program transparency will increase, ML methods in biology and medicine will have to be implemented more carefully. While using the impact factor to measure the success of a scientific article currently does not show an advantage of collaborations for experimental scientists (Fig. 5C), we suggest that these collaborations will become more frequent and impactful in the near future.~~

Conclusions

We analyzed 250 articles describing ML applications to the life sciences published 2011-2016 and another 50 published in 2018 in 17 journals from 26 different biological/medical fields (SOM). This diversity of fields was mirrored by the diversity of how machine learning was applied. Reproducibility and correct evaluation of results are crucial to ascertain validity and reliability of ML applications. Surprisingly, many articles did not focus on these aspects: 50% shared no software, 36% shared no data, and 19% applied no evaluation. In fact, an entire third (34%) of the articles only written by experimentalists described no evaluation. While we hypothesized that ensuring validity of ML applications would be necessary to achieve high visibility of the research, we found the opposite: more valid work was often published in lower-ranked journals attracting fewer citations (Fig. 1, Fig. 3).

In general, how these technical aspects were addressed was highly influenced by the authors' scientific background: Reproducibility and evaluation were more prominent with computational scientists as co-authors (Fig. 2, Fig. 4 _deleted_), while articles co-authored by experimentalists more frequently provided independent experimental proof (Fig. 4). Thus, collaborations of authors from different disciplines provided more opportunity for higher quality results integrating knowledge from various fields of expertise.

We hypothesized that collaborative research should also be cited more often and be accepted in higher-ranked journals. However, this was only true for computational scientists who profited from collaborating with experimentalists, in particular physicians, by getting accepted in higher impact factor journals (Fig. 5C).

One of the most substantial challenges for AI and ML is a comprehensive, adequate evaluation; incorrect application of such tools can lead to drawing false conclusions or to overestimating the predictive power of a method. Collaborations between computational and experimental scientists substantially increased the correctness of evaluations and the likelihood of reproducibility. Thus, increased the scientific validity of published research, a good incentive to focus on such collaborations to improve ML applications that will advance the life sciences in the future.

Acknowledgments

Thanks to T. Karl and I. Weise (both TUM) for invaluable help with technical and administrative aspects of this work. Thanks to the TUM Graduate School (in particular Z. Zhang) for organizing the Summer School, to the TUM (in particular Dr. H. Keidel and President W. Herrmann) for substantial support on several levels, to the Weizmann Institute, Tel Aviv University, Technion, and Hebrew University for financial and general support; thanks also to the enlightening talks by D. Cremers (TUM), M. Linial (IAS Israel, Hebrew University), Y. Ofran (Bar-Ilan University); thanks to PubMed for providing easy access to published articles and supporting automatic access; thanks to the maintainers of Biopython for providing excellent code to access various databases and process biological data. Particular thanks also to both reviewers for their extremely helpful and plentiful criticisms substantially improving this work. Last, not least, thanks to all maintainers of public databases and to all experimentalists who enabled this analysis by making their data publicly available. This work was supported by grant 2003403.

Author's contribution

M.L. and K.S. performed the major part of data analysis and of writing the manuscript. M.L. created and adapted the pre-defined list of articles. K.S. generated figures and performed statistical tests. L.C. assisted in finding interesting correlations in the data by performing complex analyses and statistical test and in generating figures. M.L, K.S., L.C., Y.F., P.H, E.K., A.M., K.Q., A.R., S.S., A.S., L.S., and A. D.-W. participated in the summer school where the idea for this work was developed, were involved in agreeing on the goals and analysis methods of this work, were involved in data analysis by collecting data from the pre-defined list of articles, and assisted in writing the manuscript. M.L., K.S., and A.M. collected the data for 2018. N.B.-T., M.Y.N, D.R., and B.W.S. supervised the work over the entire time and proofread the manuscript. D.A. provided valuable comments especially regarding statistical analysis and was involved in manuscript writing. T.H. and B.R. initiated and supervised the summer school where the idea for this project was developed. T.H. provided important comments to refine the analysis and contributed to manuscript writing. B.R. supervised and guided the work over the entire time and proofread the manuscript. All authors read and approved the final manuscript.

References

1 Bleicher, K. H., Bohm, H. J., Muller, K. & Alanine, A. I. Hit and lead generation: beyond high-throughput screening. *Nat Rev Drug Discov* **2**, 369-378, doi:10.1038/nrd1086 (2003).

2 Sulakhe, D. *et al.* High-throughput translational medicine: challenges and solutions. *Adv Exp Med Biol* **799**, 39-67, doi:10.1007/978-1-4614-8778-4_3 (2014).

3 Howard, J. Quantitative cell biology: the essential role of theory. *Mol Biol Cell* **25**, 3438-3440, doi:10.1091/mbc.E14-02-0715 (2014).

4 Cook, C. E. *et al.* The European Bioinformatics Institute in 2016: Data growth and integration. *Nucleic Acids Res* **44**, D20-26, doi:10.1093/nar/gkv1352 (2016).

5 Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Min* **10**, 35, doi:10.1186/s13040-017-0155-3 (2017).

6 Cios, K. J., Kurgan, L. A. & Reformat, M. Machine learning in the life sciences. *IEEE Eng Med Biol Mag* **26**, 14-16 (2007).

7 *Google Trends*, <https://trends.google.de/trends> (

8 Rost, B., Radivojac, P. & Bromberg, Y. Protein function in precision medicine: deep understanding with machine learning. *FEBS Lett* **590**, 2327-2341, doi:10.1002/1873-3468.12307 (2016).

9 Webb, S. Deep learning for biology. *Nature* **554**, 555-557, doi:10.1038/d41586-018-02174-z (2018).

10 Min, S., Lee, B. & Yoon, S. Deep learning in bioinformatics. *Brief Bioinform* **18**, 851-869, doi:10.1093/bib/bbw068 (2017).

11 Larranaga, P. *et al.* Machine learning in bioinformatics. *Brief Bioinform* **7**, 86-112 (2006).

12 Frank, M. R., Wang, D., Cebrian, M. & Rahwan, I. The evolution of citation graphs in artificial intelligence research. *Nature Machine Intelligence* **1**, 79-85 (2019).

13 Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **55**, 78-87, doi:10.1145/2347736.2347755 (2012).

14 Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of Theoretical Biology* **273**, 236-247 (2011).

15 Ioannidis, J. P. *et al.* Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* **383**, 166-175, doi:10.1016/S0140-6736(13)62227-8 (2014).

16 Gron, A. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. (O'Reilly Media, Inc., 2017).

17 Chen, S., Arsenault, C. & Larivière, V. Are top-cited papers more interdisciplinary? *Journal of Informetrics* **9**, 1034-1046 (2015).

18 Cummings, J. & Kiesler, S. Organization Theory and the Changing Nature of Science. *Journal of Organization Design* **3** (2014).

19 Abramo, G., D'Angelo, C. A. & Di Costa, F. Authorship analysis of specialized vs diversified research output. *Journal of Informetrics* **13**, 564-573 (2019).

20 Abramo, G., D'Angelo, C. A. & Di Costa, F. Do interdisciplinary research teams deliver higher gains to science? *Scientometrics* **111**, 317-336, doi:10.1007/s11192-017-2253-x (2017).

21 Chen, S., Arsenault, C., Gingras, Y. & Larivière, V. Exploring the interdisciplinary evolution of a discipline: the case of Biochemistry and Molecular Biology. *Scientometrics* **102**, 1307-1323, doi:10.1007/s11192-014-1457-6 (2015).

22 Xie, Z., Li, M., Li, J., Duan, X. & Ouyang, Z. Feature analysis of multidisciplinary scientific collaboration patterns based on PNAS. *EPJ Data Science* **7**, 5, doi:10.1140/epjds/s13688-018-0134-z (2018).

23 Rinia, E. J., van Leeuwen, T. N. & van Raan, A. F. J. Impact measures of interdisciplinary research in physics. *Scientometrics* **53**, 241-248, doi:10.1023/a:1014856625623 (2002).

24 Larivière, V. & Gingras, Y. On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology* **61**, 126-131, doi:10.1002/asi.21226 (2010).

25 *Thomson Reuter's Web of Science*, <http://wokinfo.com> (

26 *U. S. National Science Foundation*, <https://www.nsf.gov/> (

27 Wallach, J. D., Boyack, K. W. & Ioannidis, J. P. A. Reproducible research practices, transparency, and open access data in the biomedical literature, 2015-2017. *PLoS Biol* **16**, e2006930, doi:10.1371/journal.pbio.2006930 (2018).

28 *Impact Factors*, <https://www.scijournal.org/> (

29 *Journal Citation Reports*, <https://clarivate.com/products/journal-citation-reports/> (

30 Abdill, R. J. & Blekhman, R. Tracking the popularity and outcomes of all bioRxiv preprints. *Elife* **8**, doi:10.7554/eLife.45133 (2019).

31 Berger, B. *et al.* ISCB's Initial Reaction to The New England Journal of Medicine Editorial on Data Sharing. *PLoS Comput Biol* **12**, e1004816, doi:10.1371/journal.pcbi.1004816 (2016).

32 Drazen, J. M. Data Sharing and the Journal. *N Engl J Med* **374**, e24, doi:10.1056/NEJMe1601087 (2016).

33 Longo, D. L. & Drazen, J. M. Data Sharing. *N Engl J Med* **374**, 276-277, doi:10.1056/NEJMe1516564 (2016).

34 Mind meld. *Nature* **525**, 289-290, doi:10.1038/525289b (2015).

35 Nissani, M. Ten cheers for interdisciplinarity: The case for interdisciplinary knowledge and research. *The Social Science Journal* **34**, 201-216 (1997).

36 van Wesel, M., Wyatt, S. & ten Haaf, J. What a difference a colon makes: how superficial factors. *Scientometrics* **98**, 1601-1615 (2014).

37 Fitzgerald, R. T. & Radmanesh, A. Social media and research visibility. *AJNR Am J Neuroradiol* **36**, 637, doi:10.3174/ajnr.A4054 (2015).

38 Patton, R. M., Stahl, C. G. & Wells, J. C. Measuring Scientific Impact Beyond Citation Counts. *D-Lib Magazine* **22**, 5-5 (2016).

Figure legends

Xxbr: I might want to go over the following captions in the next round….

Fig. 1: Spearman correlation coefficients for numeric and binary variables. We assessed the correlation between the different criteria using the Spearman correlation and tested the significance at a level of 0.05. Significant p-values are displayed using * for p-value < 0.05, ** for p-value < 0.01 and *** for p-value <0.001 after adjusting for multiple testing using the Benjamin-Hochberg procedure. Blank squares denote that the correlation is non-significant. Citations adj. (Year + Imp. Fct.) denote the citations adjusted by year and impact factor.

Fig. 2: More sharing and method comparison with computational scientists. The involvement of a computational scientist was highly correlated with _DELETED_ sharing the data, _DELETED_ making the program available, or _DELETED_ performing a comparison with other methods. Percentage of articles with data or program available or performance of a comparison with other methods with 95% percentile bootstrap confidence intervals split by whether a computational scientist was involved.

Fig. 3: Sharing and method comparison hardly impact citations. A. Number of citations adjusted by year were not influenced by data or program availability. Comparing the developed method to others led to a small, decrease in the number of citations. B. Adjusting also by impact factor showed a small trend towards higher citations when data or program were available, or a comparison to other methods was performed. C. The impact factor was higher for articles that did not make data or program available, or compared their method to others.

Fig. 4: Method testing depends on author expertise. _DELETED_ Articles involving a computational scientist applied a computational evaluation method more often than articles with only an experimentalist (physician or biologist). _DELETED_ Articles co-authored by an experimentalist provided experimental proof more often than without such a co-author. _DELETED_ Providing no evaluation method was more common among articles written solely by experimentalists. Percentage of articles with computational evaluation methods, experimental proof or no evaluation methods are shown with 95% percentile bootstrap confidence intervals split by author background.

Fig. 5: Adjusted number of citations and impact factor for different collaborations. A. The number of citations adjusted by year is slightly higher for articles solely written by experimentalists compared to articles involving computational scientists. B. Adjusting also by impact factor removes this difference. This suggests that the higher number of citations for experimentalists was mainly caused by the fact that their work got accepted in higher-ranked journals. C. _DELETED_ Impact factor was higher for articles only published by experimentalists (biologists and/or physicians) than for articles involving also computational scientists.

Fig. 1: Spearman correlation coefficients for numeric and binary variables

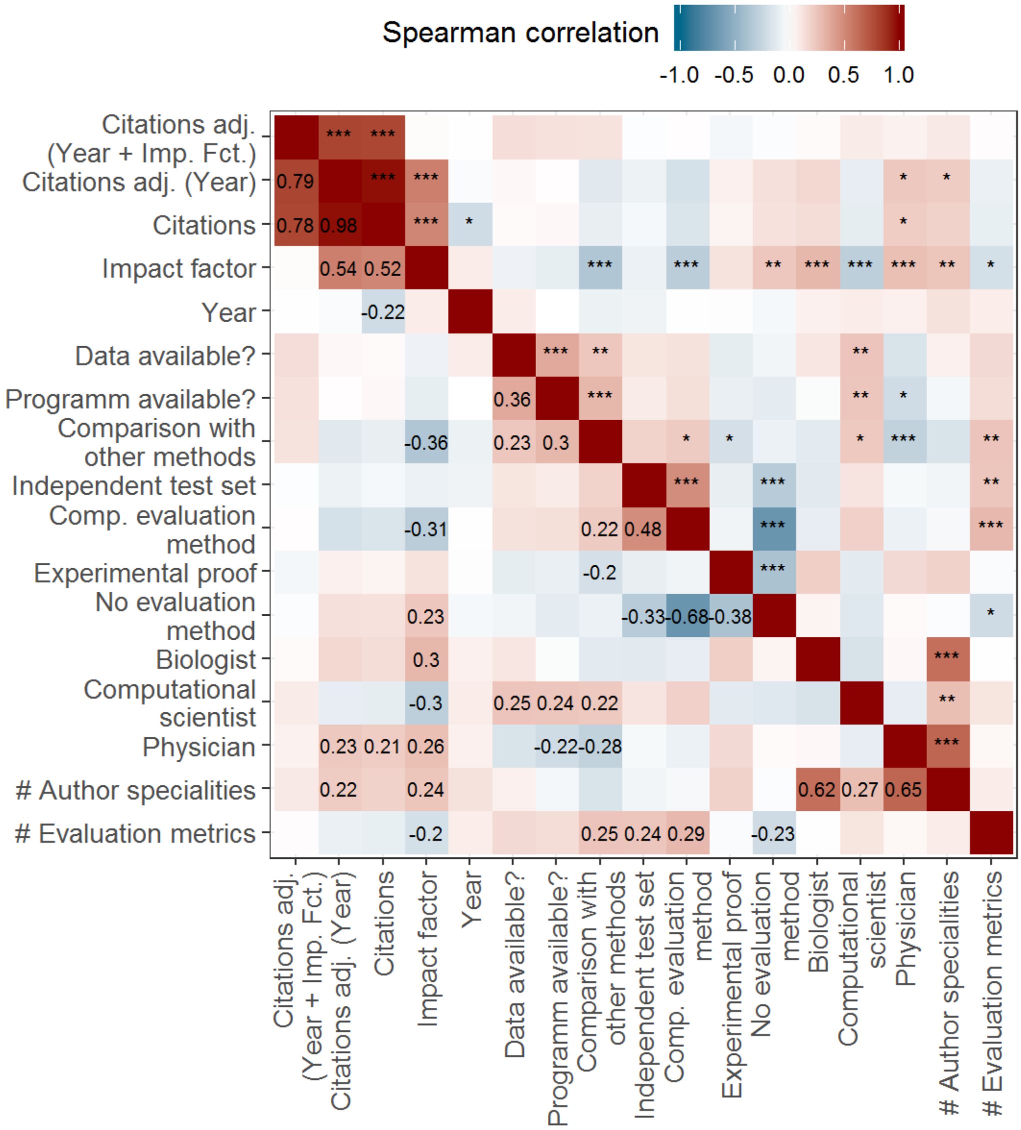


Fig. 1: Spearman correlation coefficients for numeric and binary variables. We assessed the correlation between the different criteria using the Spearman correlation and tested the significance at a level of 0.05. Significant p-values are displayed using * for p-value < 0.05, ** for p-value < 0.01 and *** for p-value <0.001 after adjusting for multiple testing using the Benjamin-Hochberg procedure. Blank squares denote that the correlation is non-significant. Citations adj. (Year + Imp. Fct.) denote the citations adjusted by year and impact factor.

Fig. 2: Method validation depends on author expertise

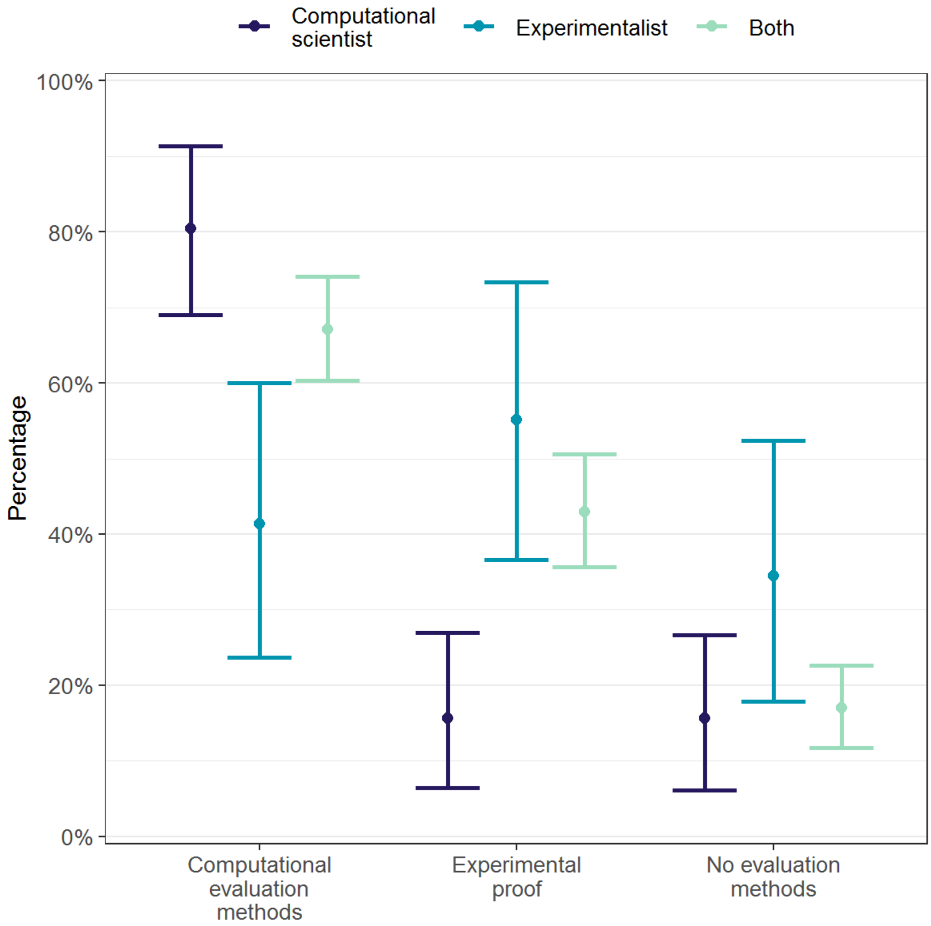
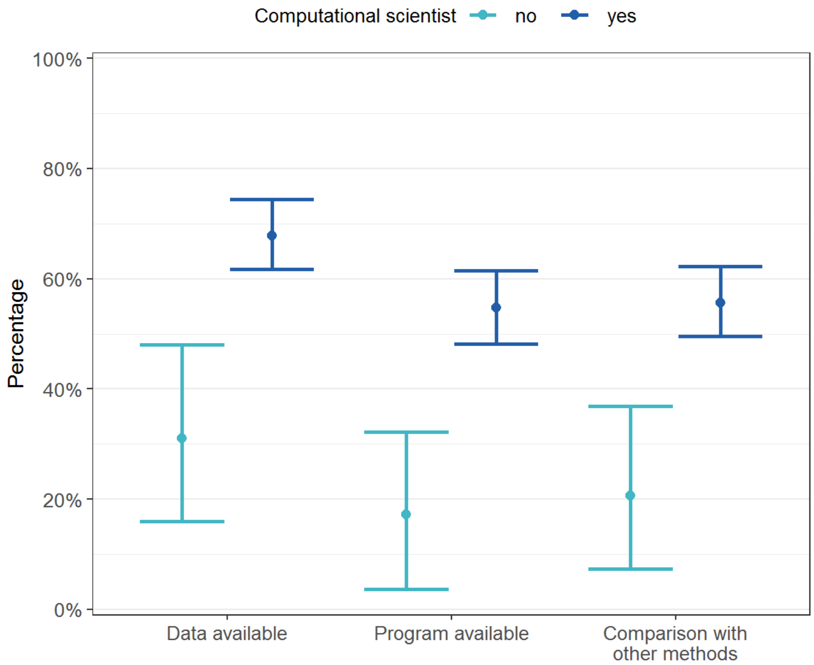


Fig. 2: Method validation depends on author expertise. _DELETED_ Articles involving a computational scientist applied a computational evaluation method more often than articles with only an experimentalist (physician or experimental biologist). _DELETED_ Articles co-authored by experimentalists provided experimental proof more often than those without. _DELETED_ Providing no evaluation method was more common among articles written solely by experimentalists. Percentage of articles with computational evaluation methods, experimental proof or no evaluation methods are shown with 95% percentile bootstrap confidence intervals split by author background.

Xxbr: may be join Fig. 2+3 into one with two panels (assuming you can compress the two to fit together onto a page, i.e. each being a column wide (7.5 cm)

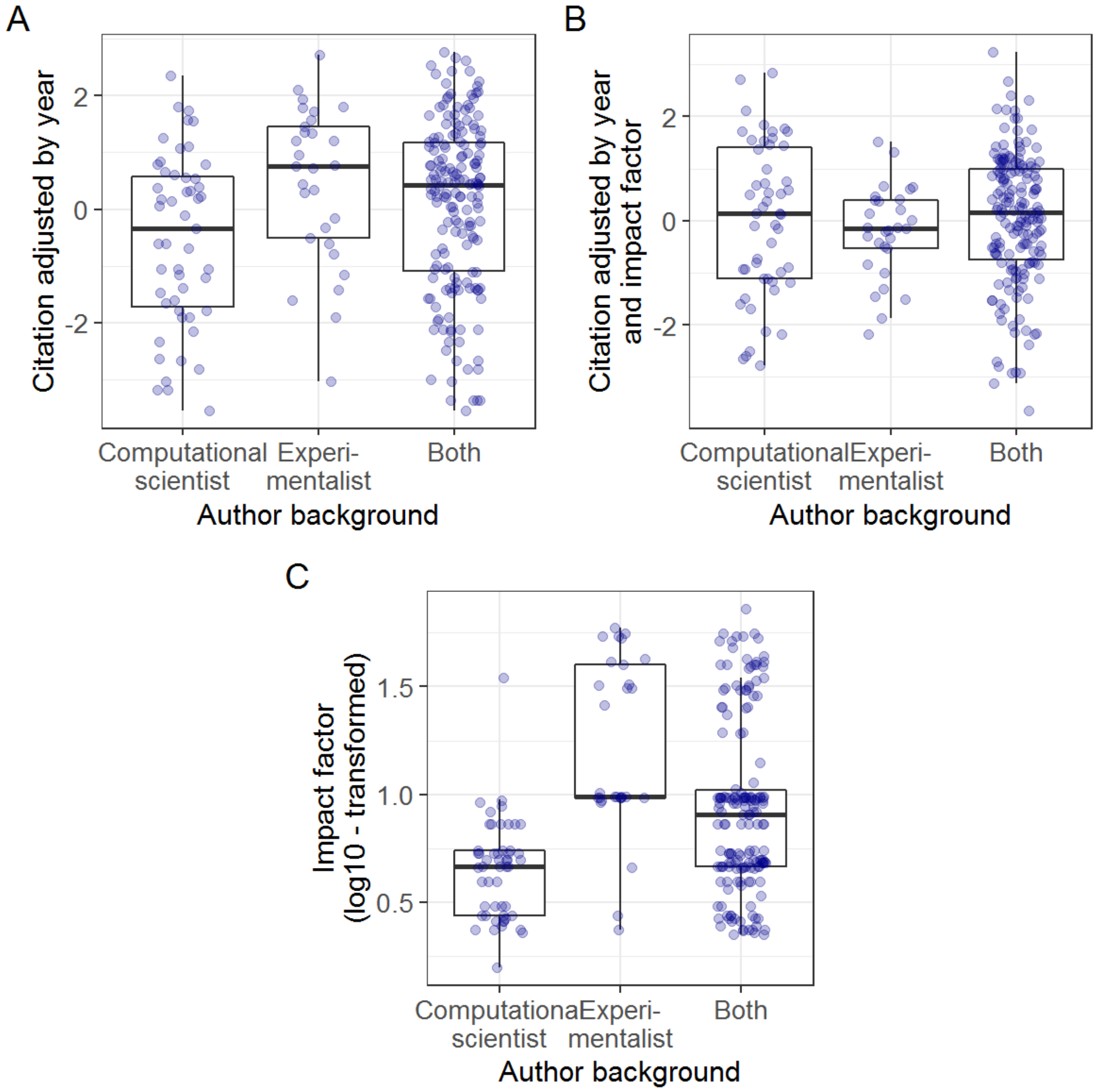
Fig. 3: More sharing and method comparison with computational scientists



Xxbr: we really ought to optimize this figure: it will be totally fine as a one-column (column=7.5 cm including legends and all); if I shrank it down to that size, we'd loose the font: too small, and there is way too much white space: make thicker lines, bigger dots, less wide bars for errors, points closer to each other in x-axis. Coloring: 1 be careful to use color consistently between figures, already inconsistent between 2 and 3; may be full black line (yes) and dashed gray: no; you could use the same scheme in Fig. 2 but would have to come up with a new one (blue?) for "both". Btw. Fig. 5 ought to follow that concept and Fig. 4 should NOT interfere – which might speak against blue for both in Fig. 2…

Fig. 3: More sharing and method comparison with computational scientists. The involvement of a computational scientist was highly correlated with _DELETED_ sharing the data, _DELETED_ making the program available, or _DELETED_ performing a comparison with other methods. Percentage of articles with data or program available or performance of a comparison with other methods with 95% percentile bootstrap confidence intervals split by whether a computational scientist was involved.

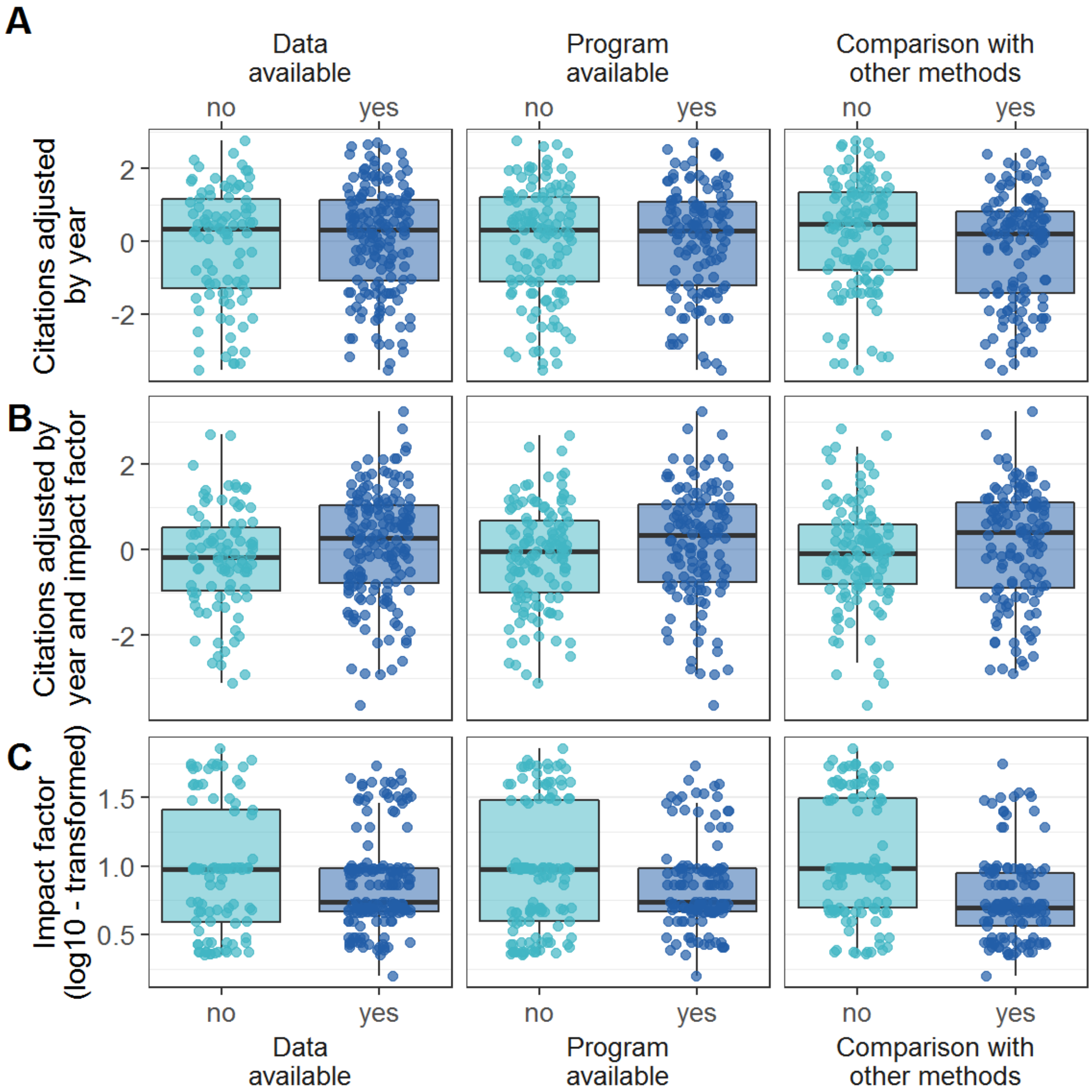
Fig. 4: NC and impact factor not consistently higher for collaborations



Xxbr: I assume that this will be either one figure spanning two columns (then have them all 3 on one, guess that might work), or 3 under each other as one column.

Fig. 4: NC and impact factor not consistently higher for collaborations. A. The number of citations adjusted by year was slightly higher for articles solely written by experimentalists compared to articles involving computational scientists. **B.** Adjusting also by impact factor removed this difference. This suggests that the higher number of citations for experimentalists was mainly caused by the fact that their work got accepted in higher-ranked journals. **C.** _DELETED_ Impact factor was higher for articles only published by experimentalists (biologists and/or physicians) than for articles with computational scientists.

Fig. 5: Sharing and method comparison hardly impact citations



A. Number of citations adjusted by year were not influenced by data or program availability. Comparing the developed method to others led to a small, decrease in the number of citations. B. Adjusting also by impact factor showed a small trend towards higher citations when data or program were available, or a comparison to other methods was performed. C. The impact factor was higher for articles that did not make data or program available, or compared their method to others.