PERSPECTIVE
https://doi.org/10.1038/s42256-019-0139-8
nature machine intelligence

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

Maria Littmann[1,27*], Katharina Selig[2,27*], Liel Cohen-Lavi[3,4], Yotam Frank[5], Peter Hönigschmid[6], Evans Kataka[6], Anja Mösch[6], Kun Qian[7,8], Avihai Ron[9,10], Sebastian Schmid[11], Adam Sorbie[12], Liran Szlak[13], Ayana Dagan-Wiener[14], Nir Ben-Tal[15], Masha Y. Niv[14,16], Daniel Razansky[9,10,17,18,19,20], Björn W. Schuller[21], Donna Ankerst[2], Tomer Hertz[3,22,23] and Burkhard Rost[1,24,25,26]

Machine learning (ML) has become an essential asset for the life sciences and medicine. We selected 250 articles describing ML applications from 17 journals sampling 26 different fields between 2011 and 2016. 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 crucial for ensuring the validity of ML applications, these aspects were met more by publications in lower-ranked journals. Second, the authors' scientific backgrounds 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, biology, and medicine. The results suggested collaborations between computational and experimental scientists to generate more scientifically sound and impactful work integrating knowledge from both domains. Although scientifically more valid solutions and collaborations involving diverse expertise did not correlate with impact factors, 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.

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, analyse 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 of the research community: querying Google Trends[7] with "machine learning" demonstrated an exponential increase over the past decade (January 2010–February 2019, data not shown). This general rise has been mirrored in many fields of biology and medicine—that is, 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.

[1]Department of Informatics, Bioinformatics and Computational Biology, Technical University of Munich, Garching/Munich, Germany. [2]Department of Mathematics, Technical University of Munich, Garching/Munich, Germany. [3]National Institute for Biotechnology in the Negev, Ben-Gurion University of the Negev, Be'er-Sheva, Israel. [4]Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Be'er-Sheva, Israel. [5]The Blavatnik School of Computer Science, Tel-Aviv University, Ramat Aviv, Israel. [6]Department of Bioinformatics, Wissenschaftszentrum Weihenstephan, Technical University of Munich, Freising, Germany. [7]Chair of Human-Machine Communication, Technical University of Munich, Munich, Germany. [8]Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, Tokyo, Japan. [9]Institute for Biological and Medical Imaging, Helmholtz Center Munich, Neuherberg, Germany. [10]Faculty of Medicine, Technical University of Munich, Munich, Germany. [11]Chair of Food Chemistry and Molecular Sensory Science, Technical University of Munich, Freising, Germany. [12]Chair of Nutrition and Immunology, Technical University of Munich, Freising, Germany. [13]Weizmann Institute of Science, Rehovot, Israel. [14]The Institute of Biochemistry, Food and Nutrition, The Robert H. Smith Faculty of Agriculture, Food and Environment, The Hebrew University, Rehovot, Israel. [15]Department of Biochemistry and Molecular Biology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel. [16]The Fritz Haber Center for Molecular Dynamics, The Hebrew University, Jerusalem, Israel. [17]Faculty of Medicine, University of Zurich, Zurich, Switzerland. [18]Institute of Pharmacology and Toxicology, University of Zurich, Zurich, Switzerland. [19]Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Zurich, Switzerland. [20]Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich, Switzerland. [21]Group on Language, Audio and Music, Imperial College London, London, UK. [22]The Shraga Segal Department of Microbiology and Immunology, Ben-Gurion University of the Negev, Be'er-Sheva, Israel. [23]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [24]Institute for Advanced Study, Garching/Munich, Germany. [25]School of Life Sciences, Technical University of Munich, Freising, Germany. [26]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA. [27]These authors contributed equally: Maria Littmann, Katharina Selig. *e-mail: littmann@rostlab.org; katharina.selig@tum.de

Quality and validity of ML models hinge on two primary factors: (1) size, quality and universal validity of data; and (2) the correct development and assessment of the resulting models[5,13]. Successful ML applications extract generic principles from today's data, allowing the generalization—that is, 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/Weiqi) for which task and success are more clearly defined require mainly expertise in ML.

Collaborations across fields of expertise

Throughout science, interdisciplinarity has become important to break new grounds[17,18]. Several recent studies[17,19–24] investigated the role of interdisciplinarity by automatically extracting tens and hundreds of thousands of publications (for example, from Web of Science or the *Proceedings of the National Academy of Sciences*). Toward this end, one definition of interdisciplinarity is as follows: if an article is published and cited in different fields or subfields (for example, the US National Science Foundation classifies journals into 14 different disciplines and 143 subdisciplines[17,21]), the article is deemed 'interdisciplinary'[17,21,24]. Others define interdisciplinary as articles published by authors from different disciplines, an approach so far limited to Italian scientists due to a public directory mapping Italian researchers to disciplines[19,20].

The scientific impact of an article is usually measured by its number of citations[17,24]. To correct for field- and journal-specific effects, that number is normalized by time (years since publication) and by the journal's impact factor[23,24]. Since the impact factor is calculated from the number of citations of articles published in this journal[25], 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 a particular type of information. The studies disagree in their findings regarding the importance of interdisciplinary collaborations: one finds no consistent correlation between impact and interdisciplinarity from sampling over 750,000 publications: for some disciplines, interdisciplinarity was proportional to citations; for others (including physics) the relation was reversed[24]. Another work, focusing on more than 15,000 publications from physics, found interdisciplinarity was proportional to citation rates but only when published in journals with citation rates below average[23]. Yet other studies, based on 751,766[17] and 71,633 publications[20], agreed that interdisciplinary work creates higher impact than non-interdisciplinary work. Also, specific collaborations between scientists from related fields lead to higher-impact publications than generic collaborations between scientists from very different fields[20]. Clearly, there is no simple common thread running through all of those findings. However, what made us revisit this question and begin our analysis were three other reasons: (1) the focus on ML and the life sciences, not explicitly covered by others; (2) the aim of separating the analysis of scientific quality (soundness) from 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, a definition similar to the one 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 (see Supplementary Information). Among all papers published in those 17 journals in the years 2011–2016, keyword searches (Supplementary Table 1) matched 4,306 articles, where about 2,100 of those were deemed correct hits based on the observed false positive rate for a subset of articles. From those, initially 250 were randomly selected (see Supplementary Information; complete list in Supplementary Dataset 1, list of identified falsely extracted articles is provided in Supplementary Dataset 2). 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[9,10]. In contrast to previous studies[17,19–24], our assessment focused on ML applications in the life sciences and all information was manually extracted from the articles. This allowed, for instance, to correct the 50% false positives from the keyword searches, and also to define interdisciplinarity through the authors' scientific backgrounds by reading partial CVs for 1,918 authors of the 250 papers. Each article was classified independently by two of us. These investments limited the number of papers analysed but allowed a more fine-grained assessment not accessible to automatic extraction.

Our focus had several implications, including that all papers reported applications of ML 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. Thus, we could sharpen the perspective by distinguishing the expertise contributing to the application of ML to the life sciences with authors from potentially three disciplines: computational sciences, biology and medicine (expertise of author verified through CV, not through affiliation). The number of different disciplines presented in the author list proxied the level of interdisciplinarity with values from 1 to 3.

We proxied the validity of papers describing the application of ML methods to biology and medicine through six different indicators. The first four relate to whether 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 or other evaluation methods (V1: binary value), more than one single measure for performance (V2: integer), additional test sets (V3: binary value) or 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, let alone to develop the best possible method. The last two indicators related to sharing methods and results. These were sharing data (V5), programmes and codes (V6) through publicly available sites. Typically, reviewing ML applications by journal reviewers and the public at large requires availability of data and programmes in a form beyond what is available through description of methods.

The correct application of ML requires expertise from those familiar with ML and those familiar with the life sciences, that is, 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

Three levels of interdisciplinarity. By definition, all the papers analysed applied methods from computational fields to the life sciences—that is, were intrinsically interdisciplinary. All 250 papers analysed might 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, biologists and physicians) and introduced interdisciplinarity as a number ranging from one to three depending on how many disciplines were represented by the authors of the work. Most of the 250 papers were co-authored

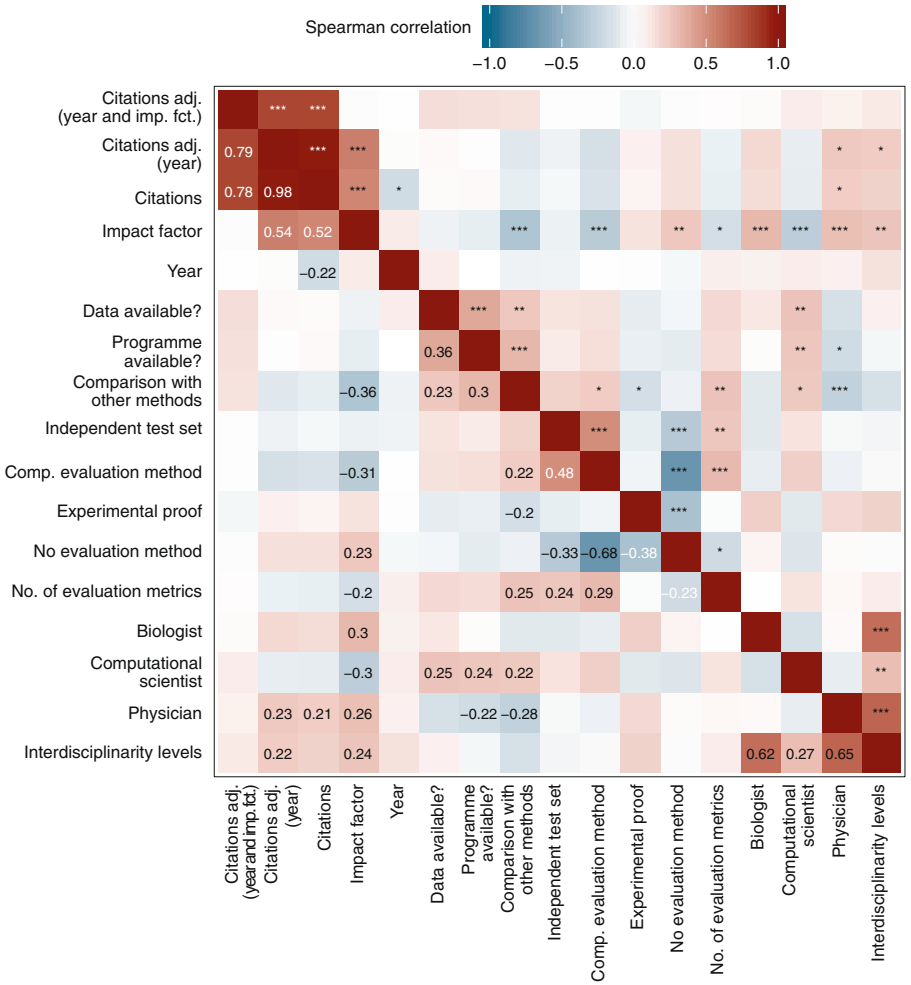

Fig. 1 | Spearman correlation coefficients for numeric and binary variables. Correlation between the different criteria of 250 articles using the Spearman correlation tested at a significance level of 0.05. Significant *p*-values are displayed using * for *p*-value < 0.05, ** for *p*-value < 0.01 and *** 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) and citations adj. (year and imp. fct.) denote the citations adjusted by year and by year and impact factor, respectively.

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 analyse the key indicators: validity (evaluation and sharing) and impact (number of citations (NC); NC adjusted by year, equation (1) in Supplementary Information; impact factor, and NC adjusted by year and impact factor, equation (2) in Supplementary Information).

58% of the chosen 250 papers (see Supplementary Information for more details on how these articles were selected) appeared in only four of the 17 journals (by occurrence: *Bioinformatics*, *Proceedings of the National Academy of Sciences*, *PLOS Computational Biology* and *BMC Bioinformatics*; see additional results in Supplementary Information, including Supplementary Figs. 1, 2, 3 and 4, for more details)—that is, were 2.5-fold over-represented. While the disciplines of biologist and physician correlated positively with impact factor ($\rho = 0.30$/p-value < 0.001, $\rho = 0.26$/p-value < 0.001, respectively), computational science correlated negatively ($\rho = -0.30$/p-value < 0.001; Fig. 1). Computational scientists might focus more on methods, while biologists and physicians focus more on new data that tend to be highly cited in the life sciences.

Scientific validity higher with experts participating in collaboration. Evaluation methods (for example, cross-validation), usage of independent test sets, and/or 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' (biologists and physicians; Fig. 2a). However, most articles written solely by experimentalists provided experimental proof (55%), so did 16% of those from only computational co-authors (Fig. 2a). The corresponding numbers for interdisciplinary collaborations between computational and experimental scientists (level of interdisciplinarity ≥ 2) were between these two extremes: 67% evaluated their methods and 43% provided 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. 2a).

Several evaluation metrics are required to assess the performance of ML applications (for example, precision, recall, accuracy or confusion matrices). 6% of all articles used no evaluation metric, 53% used one or two, and 6% used over five (Supplementary Fig. 5). 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—that is, minimally we need two evaluation metrics. More complex problems require more evaluation metrics.

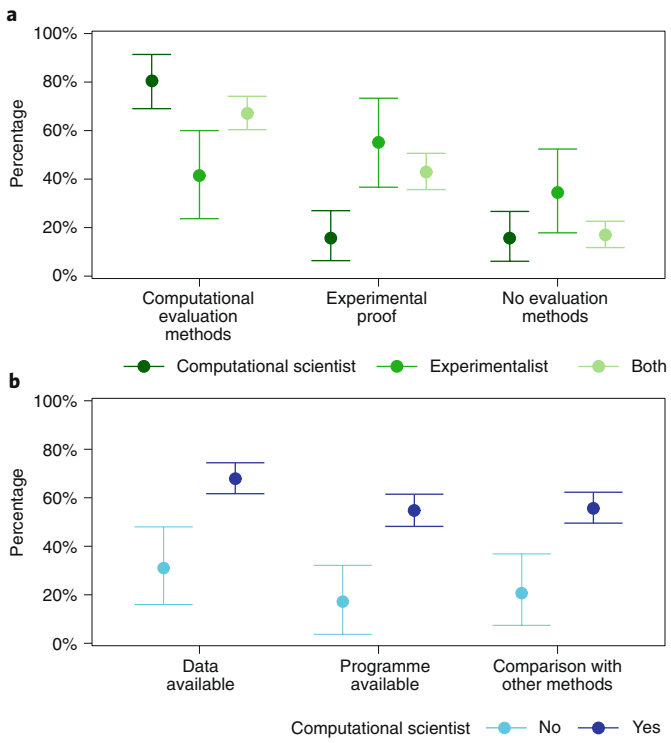

Fig. 2 | Method validation, comparison and data and programme sharing depends on author expertise. Percentages of 250 articles by evaluation methods, data or programme sharing and comparison with other methods split by authors' backgrounds are shown with 95% percentile bootstrap confidence intervals based on 1,000 bootstraps. **a**, Articles involving a computational scientist applied a computational evaluation method more often than articles with only an experimentalist (physician or biologist). Articles co-authored by experimentalists provided experimental proof more often than those without. Providing no evaluation method was more common among articles written solely by experimentalists. **b**, The involvement of a computational scientist was highly correlated with sharing the data, making the programme available, or performing a comparison with other methods.

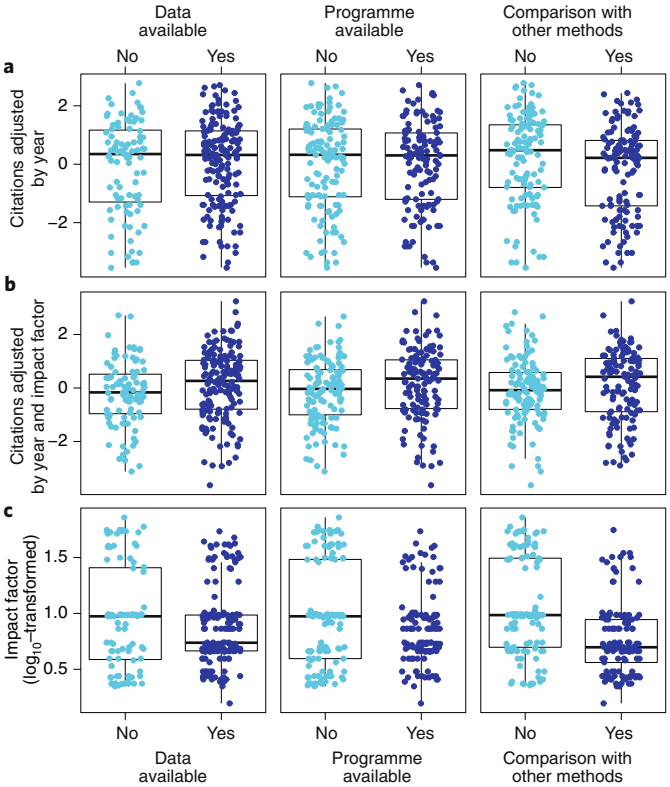

Fig. 3 | Sharing and method comparison hardly impact citations. Boxplots of adjusted citations and $\log_{10}$-transformed impact factor of 250 articles split by data or programme sharing and comparison with other methods. Vertical bars indicate largest (smallest) value within 1.5 times the interquartile range above (below) the third (first) quartile. **a**, Number of citations adjusted by year were not influenced by data or programme availability. Comparing the developed method to others led to a small decrease in the number of citations. **b**, Adjusting by impact factor as well showed a small, but non-significant, trend towards higher citations when data or programme were available, or a comparison to other methods was performed. **c**, The impact factor was slightly higher for articles that did not make data or programme available, or compared their method to others, where only the last difference was statistically significant.

Typically, clearly more than two metrics are needed to show different strengths and weaknesses of a prediction method.

About half (52%) of the methods were compared to others; this again dropped to 21% without computational co-authors (p-value = 0.001; Fig. 2b). Although crucial for validation, method comparisons might make descriptions more complex, leading to rejection from higher-ranked journals (Fig. 3c) and possibly to lower impact (Fig. 3c), although adjusting by impact factor as well suggested a slight pay-off from method comparisons in terms of citations (Fig. 3b).

Reproducibility is a pillar of science[26–28], partially relying on making data and methods publicly available. It is particularly critical for ML applications because many minor technical details may invalidate results[25]. Overall, 64% of the articles shared their data (with large variation between journals: from *Nucleic Acids Research* = 89% to *New England Journal of Medicine* = 8%; Supplementary Fig. 6), reflecting the general trend that articles from medicine shared data the least (Supplementary Fig. 7). We could not establish whether this is 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. 2b). 57% of the articles relied on data extracted from public resources or previous articles; however, 22% of those that did, did not publish their data. Data sharing was highest for collaborations with computer scientists (Fig. 2b).

Experimentalists 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 focusing more on technical aspects; while computational scientists benefit from the access to new data, domain knowledge and experimental verification of the results. Therefore, collaborative work will generate more scientifically sound and impactful work.

Collaborations of scientists with different expertise were somehow cited more often. Interdisciplinary collaborations of researchers from different fields seem increasingly important to generate new ideas and results[29,30]. The higher the level of interdisciplinarity, the higher the NC adjusted by year ($\rho = 0.22$, p-value = 0.02; Fig. 1, Supplementary Fig. 8) and the higher the impact factor ($\rho = 0.24$, p-value = 0.002; Fig. 1, Supplementary Fig. 8). When adjusting NC by impact factor as well, the correlation was no longer significant (Fig. 1, Supplementary Fig. 8), suggesting that interdisciplinary articles were cited more mainly because they were published in higher-ranked journals (Supplementary Fig. 8). The correlation between

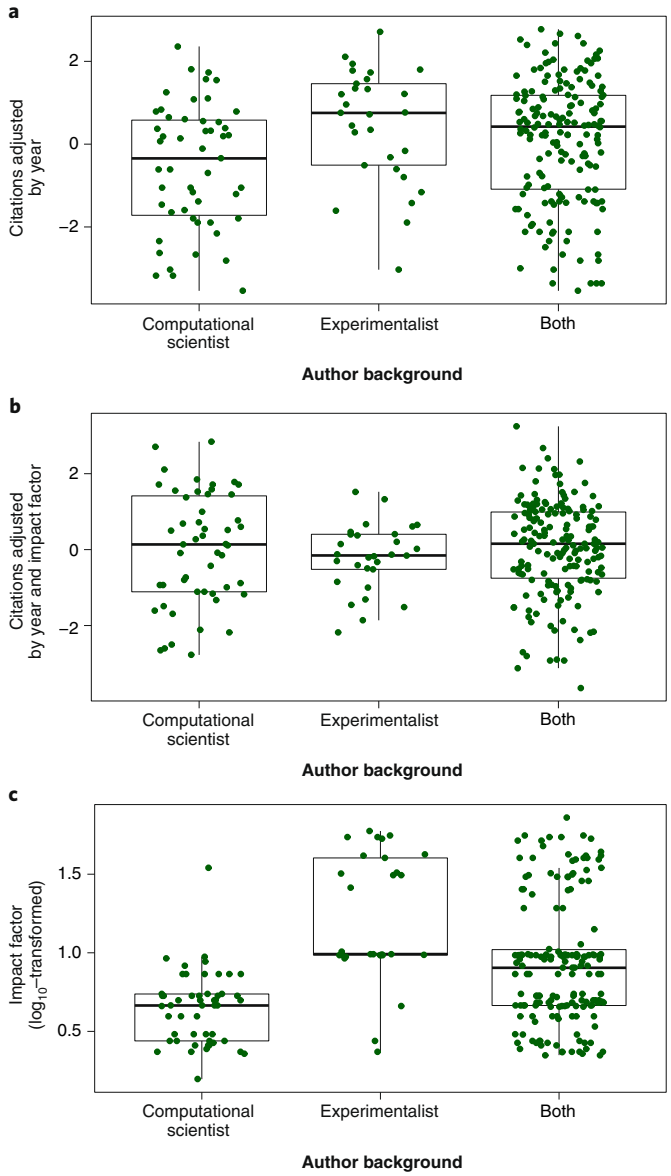

Fig. 4 | Number of citations and impact factor not consistently higher for collaborations. Boxplots of adjusted citations and $\log_{10}$-transformed impact factor of 250 articles split by authors' backgrounds. Vertical bars indicate largest (smallest) value within 1.5 times the interquartile range above (below) the third (first) quartile. **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 by impact factor as well 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**, Impact factor was higher for articles only published by experimentalists (biologists and/or physicians) than for articles with computational scientists.

impact factor and level of interdisciplinarity (Supplementary Fig. 8) suggested that authors profit from collaborations.

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. 4a). The results for impact factor and NC adjusted by impact factor and year suggested that the higher NC originated essentially from experimentalists publishing in higher-ranked journals (Fig. 4b,c). For research teams with only computational expertise, contributions from experimentalists 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.

Did scientific validity (evaluation and sharing) correlate with impact? Computational evaluations correlated negatively with the impact factor ($\rho = -0.31$, p-value < 0.001); using no evaluation method correlated positively with the impact factor ($\rho = 0.23$, p-value $= 0.004$), but we could not detect a significant relationship between impact factor and experimental proof (Fig. 1). Since all articles analysed 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 adjusted by year (Fig. 3a), although adjusting by impact factor as well hinted at a tendency for sharing to lead 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 adjusted by year (Fig. 3a); the trend changed toward more cited when adjusting NC by impact factor as well (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 programme 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. 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 issue by analysing another 50 articles describing ML applications to the life sciences published in 2018 (selected and analysed largely by the same criteria as the other 250; see Supplementary Information for details; complete list in Supplementary Dataset 3). The major differences were: fewer publications without computational scientists (6% 2018 versus 12% 2011–2016), and programme sharing rose (70% versus 50%). Although data sharing did not change significantly (68% versus 64%), those papers that shared data were cited more often and accepted to higher-ranked journals, but we could not detect a significant difference (Supplementary Fig. 9). Other aspects also did not change, neither the fact that papers sharing programmes tended to be published in lower-ranked journals (Supplementary Fig. 6) nor the correlation between number of involved disciplines and the proxies for impact (for example, NC adjusted by year, impact factor, and NC adjusted by year and 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, thoroughly analysing more than 300 articles will render the conclusions more valid. Second, we proxied impact and visibility through number of citations and the impact factor. However, the number of citations can be influenced by other factors that can seem superficial and can be controlled by the authors[31], and it is hard to compensate for these ones. 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 journal's impact factor[32,33]. Third, 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 biologically and medically relevant novelties are less likely to specifically ask for methodological details. Fourth, we considered any publicly available information to assign author disciplines 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 underestimating computational expertise in publications. Furthermore, we considered data and programme 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 that correlated with the impact factor also correlated with each other, confounding factors might influence the results and these interrelationships are difficult to separate.

Conclusions

We analysed 250 articles describing ML applications to the life sciences published 2011–2016 and another 50 articles published in 2018 in 17 journals from 24 different biological/medical fields (see Supplementary Information for more information). This diversity of fields was mirrored by the diversity of how ML 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 backgrounds: reproducibility and evaluation were more prominent with computational scientists as co-authors (Fig. 2), while articles co-authored by experimentalists more frequently provided experimental proof (Fig. 2). 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 by getting accepted in higher impact factor journals (Fig. 4c).

One of the most substantial challenges for 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, interdisciplinary collaborations increased the scientific validity of published research. As the enforcement of data and programme 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. 4c), we suggest that these collaborations will become more frequent and impactful in the near future.

Received: 9 August 2019; Accepted: 6 December 2019;
Published online: 13 January 2020

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 (2003).
2. Sulakhe, D. et al. High-throughput translational medicine: challenges and solutions. *Adv. Exp. Med. Biol.* **799**, 39–67 (2014).
3. Howard, J. Quantitative cell biology: the essential role of theory. *Mol. Biol. Cell.* **25**, 3438–3440 (2014).
4. Cook, C. E. et al. The European Bioinformatics Institute in 2016: data growth and integration. *Nucl. Acids Res.* **44**, D20–26 (2016).
5. Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Mining* **10**, 35 (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. *Google* https://trends.google.de/trends (2019).
8. Rost, B., Radivojac, P. & Bromberg, Y. Protein function in precision medicine: deep understanding with machine learning. *FEBS Lett.* **590**, 2327–2341 (2016).
9. Webb, S. Deep learning for biology. *Nature* **554**, 555–557 (2018).
10. Min, S., Lee, B. & Yoon, S. Deep learning in bioinformatics. *Brief. Bioinform.* **18**, 851–869 (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. *Nat. Mach. Intell.* **1**, 79–85 (2019).
13. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **55**, 78–87 (2012).
14. Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **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 (2014).
16. Gron, A. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (O'Reilly Media, 2017).
17. Chen, S., Arsenault, C. & Larivière, V. Are top-cited papers more interdisciplinary? *J. Informetr.* **9**, 1034–1046 (2015).
18. Cummings, J. & Kiesler, S. Organization theory and the changing nature of science. *J. Org. Des.* **3**, 1–16 (2014).
19. Abramo, G., D'Angelo, C. A. & Di Costa, F. Authorship analysis of specialized vs diversified research output. *J. Informetr.* **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 (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 (2015).
22. Xie, Z., Li, M., Li, J., Duan, X. & Ouyang, Z. Feature analysis of multidisciplinary scientific collaboration patterns based on PNAS. *EPJ Data Sci.* **7**, 5 (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 (2002).
24. Larivière, V. & Gingras, Y. On the relationship between interdisciplinarity and scientific impact. *J. Am. Soc. Inform. Sci. Technol.* **61**, 126–131 (2010).
25. 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 (2018).
26. Berger, B. et al. ISCB's initial reaction to the New England Journal of Medicine editorial on data sharing. *PLoS Comput. Biol.* **12**, e1004816 (2016).
27. Drazen, J. M. Data sharing and the journal. *N. Engl. J. Med.* **374**, e24 (2016).
28. Longo, D. L. & Drazen, J. M. Data sharing. *N. Engl. J. Med.* **374**, 276–277 (2016).
29. Mind meld. *Nature* **525**, 289–290 (2015).
30. Nissani, M. Ten cheers for interdisciplinarity: the case for interdisciplinary knowledge and research. *Soc. Sci. J.* **34**, 201–216 (1997).
31. van Wesel, M., Wyatt, S. & ten Haaf, J. What a difference a colon makes: how superficial factors. *Scientometrics* **98**, 1601–1615 (2014).
32. Fitzgerald, R. T. & Radmanesh, A. Social media and research visibility. *Am. J. Neuroradiol.* **36**, 637 (2015).
33. Patton, R. M., Stahl, C. G. & Wells, J. C. Measuring scientific impact beyond citation counts. *D-Lib Magazine* **22**, 5 (2016).

Acknowledgements

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 H. Keidel and W. Herrmann) for substantial support on several levels including financing the summer school, 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. Last, but 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 no. 640508 from the Deutsche Forschungsgemeinschaft (DFG).

Author contributions

M.L. and K.S. performed the major part of data analysis and of writing the manuscript. M.L. created and adapted the predefined 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 predefined 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.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-019-0139-8.

Correspondence should be addressed to M.L. or K.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.