DOME: recommendations for supervised machine learning validation in biology

DOME is a set of community-wide recommendations for reporting supervised machine learning–based analyses applied to biological studies. Broad adoption of these recommendations will help improve machine learning assessment and reproducibility.

Ian Walsh, Dmytro Fishman, Dario Garcia-Gasulla, Tiina Titma, Gianluca Pollastri, ELIXIR Machine Learning Focus Group, Jennifer Harrow, Fotis E. Psomopoulos and Silvio C. E. Tosatto

With the steep decline in the cost of many high-throughput technologies, large amounts of biological data are being generated and made accessible to researchers. Machine learning (ML) has come into the spotlight as a very useful approach for understanding cellular[1], genomic[2], proteomic[3], post-translational[4], metabolic[5] and drug discovery data[6], with the potential to result in ground-breaking medical applications[7,8]. This is clearly reflected in the corresponding growth of ML publications (Fig. 1), reporting a wide range of modeling techniques in biology. While ideally ML methods should be validated experimentally, this happens only in a fraction of the publications[9]. We believe that the time is right for the ML community to develop standards for reporting ML-based analyses to enable critical assessment[10] and improve reproducibility[11,12].

Guidelines or recommendations on how to appropriately construct ML algorithms can help to ensure correct results and predictions[13,14]. In biomedical research, communities have defined standard guidelines and best practices for scientific data management[15] and reproducibility of computational tools[16,17]. On the ML community side, there is demand for a cohesive and combined set of recommendations with respect to data, the optimization techniques, the final model, and evaluation protocols as a whole.

A recent comment highlighted the need for standards in ML[18], arguing for the adoption of on-submission checklists[10] as a first step toward improving publication standards. Through a community-driven consensus, we propose a list of minimal requirements asked as questions to ML implementers (Box 1) that, if followed, will help to assess the quality and reliability of reported methods more faithfully. We have focused on data, optimization, model and evaluation (DOME) as each component of an ML implementation usually falls within one of these four topics. We do not propose new specific solutions, only recommendations (Table 1). A reporting checklist is also provided (Box 1). Our recommendations are made primarily for the case of supervised learning in biological applications in the absence of direct experimental validation, as this is the most common type of ML approach used. We do not discuss how ML can be used in clinical applications[19,20]. It also remains to be determined whether the DOME recommendations can be extended to other fields of ML, like unsupervised, semisupervised and reinforcement learning.

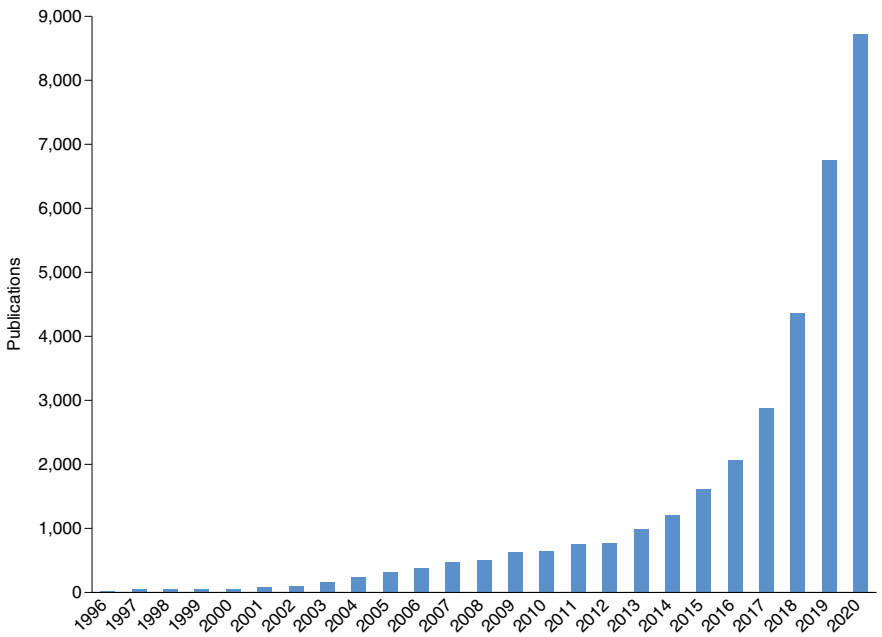

Fig. 1 | Exponential increase of ML publications in biology. The number of ML publications per year is based on Web of Science from 1996 onwards using the topic category for "machine learning" in combination with each of the following terms: "biolog*", "medicine", "genom*", "prote*", "cell*", "post translational", "metabolic" and "clinical".

Development of the recommendations

The recommendations outlined below were initially formulated through the ELIXIR Machine Learning Focus Group after the publication of a Comment calling for the establishment of standards for ML in biology[18]. ELIXIR, initially established in 2014, is now a mature intergovernmental European infrastructure for biological data and represents over 220 research organizations in 22 countries across many aspects of bioinformatics[21]. Over 700 national experts participate in the development and operation of national services that contribute to data access, integration, training and analysis for the research community. Over 50 of these

Box 1 | Structuring a Methods section for supervised machine learning approaches

Here we suggest a list of questions that authors should address in the Methods sections of manuscripts describing supervised ML approaches, in order to conform to the DOME recommendations and ensure a high quality of ML analysis.

Data (this section should be repeated separately for each dataset)

- *Provenance*: What is the source of the data (database, publication, direct experiment)? If data are in classes, how many data points are available in each class—for example, total for the positive (N_{pos}) and negative (N_{neg}) cases? If regression, how many real value points are there? Has the dataset been previously used by other papers and/or is it recognized by the community?
- *Data splits*: How many data points are in the training and test sets? Was a separate validation set used, and if yes, how large was it? Are the distributions of data types in the training and test sets different? Are the distributions of data types in both training and test sets plotted?
- *Redundancy between data splits*: How were the sets split? Are the training and test sets independent? How was this enforced (for example, redundancy reduction to less than *X*% pairwise identity)? How does the distribution compare to previously published ML datasets?
- *Availability of data*: Are the data, including the data splits used, released in a public forum? If yes, where (for example, supporting material, URL) and how (license)?

Optimization *(this section should be repeated separately for each trained model)*

- *Algorithm*: What is the ML algorithm class used? Is the ML algorithm new? If yes, why was it chosen over better known alternatives?
- *Meta-predictions*: Does the model use data from other ML algorithms as input? If yes, which ones? Is it clear that training data of initial predictors and meta-predictor are independent of test data for the meta-predictor?
- *Data encoding*: How were the data encoded and preprocessed for the ML algorithm?
- *Parameters*: How many parameters (p) are used in the model? How was p selected?
- *Features*: How many features (f) are used as input? Was feature selection performed? If yes, was it performed using the training set only?
- *Fitting*: Is p much larger than the number of training points and/or is f large (for example, in classification is $p \gg (N_{pos} + N_{neg})$ and/or $f > 100$)? If yes, how was overfitting ruled out? Conversely, if the number of training points is much larger than p and/or f is small (for example, $(N_{pos} + N_{neg}) \gg p$ and/or $f < 5$), how was underfitting ruled out?
- *Regularization*: were any overfitting prevention techniques used (for example, early stopping using a validation set)? If yes, which ones?
- *Availability of configuration*: Are the hyperparameter configurations, optimization schedule, model files and optimization parameters reported? If yes, where (for example, URL) and how (license)?

Model *(this section should be repeated separately for each trained model)*

- *Interpretability*: Is the model black box or interpretable? If the model is interpretable, can you give clear examples of this?
- *Output*: Is the model classification or regression?
- *Execution time*: How much time does a single representative prediction require on a standard machine (for example, seconds on a desktop PC or high-performance computing cluster)?
- *Availability of software*: Is the source code released? Is a method to run the algorithm—such as executable, web server, virtual machine or container instance—released? If yes, where (for example, URL) and how (license)?

Evaluation

- *Evaluation method*: How was the method evaluated (for example cross-validation, independent dataset, novel experiments)?
- *Performance measures*: Which performance metrics are reported? Is this set representative (for example, compared to the literature)?
- *Comparison*: Was a comparison to publicly available methods performed on benchmark datasets? Was a comparison to simpler baselines performed?
- *Confidence*: Do the performance metrics have confidence intervals? Are the results statistically significant to claim that the method is superior to others and baselines?
- *Availability of evaluation*: Are the raw evaluation files (for example, assignments for comparison and baselines, statistical code, confusion matrices) available? If yes, where (for example, URL) and how (license)?

The above description is provided in table format in Supplementary Table 1, together with two fully worked out examples (Supplementary Tables 2 and 3).

experts involved in the field of ML have established the ELIXIR Machine Learning Focus Group (https://elixir-europe.org/focus-groups/machine-learning), which held meetings to develop and refine recommendations based on a broad consensus.

Scope of the recommendations

The recommendations cover four major aspects of supervised ML according to the DOME acronym. The key points and rationale for each aspect of DOME are described below and summarized in Table 1. Box 1 provides an actionable checklist (with the recommendations codified as questions), which we suggest authors use as a guide when reporting ML-based methods in manuscripts.

Data. State-of-the-art ML models are often capable of memorizing all the variation in training data. Such models when evaluated on data that they were exposed to during training would create the illusion of mastering the task at hand. However, when tested on an independent set of data (termed a test or validation set), the performance would seem less impressive, suggesting low generalization power of the model. To tackle this problem, initial data should be divided randomly into non-overlapping parts. The simplest approach is to have independent training and testing sets (and possibly a third validation set). Alternatively, the cross-validation or bootstrapping techniques

Table 1 | Supervised ML in biology: concerns, the consequences they impart and recommendations

Broad topic	Be on the lookout for	Consequences	Recommendation(s)
Data	• Inadequate data size & quality • Inappropriate partitioning, dependence between train and test data • Class imbalance • No access to data	• Data not representative of domain application • Unreliable or biased performance evaluation • Cannot check data credibility	• **Use independent optimization (training) and evaluation (testing) sets**. This is especially important for meta algorithms, where independence of multiple training sets must be shown to be independent of the evaluation (testing) sets. • **Release data, preferably using appropriate long-term repositories, and include exact splits**. • Offer sufficient evidence of data size & distribution being representative of the domain.
Optimization	• Overfitting, underfitting and illegal parameter tuning • Imprecise parameters and protocols given	• Reported performance is too optimistic or too pessimistic • The model models noise or misses relevant relationships • Results are not reproducible	• **Clarify that evaluation sets were not used for feature selection, preprocessing steps or parameter tuning**. • **Report indicators on training and testing data that can aid in assessing the possibility of under- or overfitting; for example, train vs. test error**. • **Release definitions of all algorithmic hyperparameters, regularization protocols, parameters and optimization protocol**. • For neural networks, release definitions of training and learning curves. • Include explicit model validation techniques, such as *N*-fold cross-validation.
Model	• Unclear if black box or interpretable model • No access to resulting source code, trained models & data • Execution time impractical	• An interpretable model shows no explainable behavior • Cannot cross compare methods & reproducibility, or check data credibility • Model takes too much time to produce results	• **Describe the choice of black box or interpretable model. If interpretable, show examples of interpretable output**. • Release documented source code + models + executable + user interface/webserver + software containers. • Report execution time averaged across many repeats. If computationally tough, compare to similar methods.
Evaluation	• Performance measures inadequate • No comparisons to baselines or other methods • Highly variable performance	• Biased performance measures reported • The method is falsely claimed as state-of-the-art • Unpredictable performance in production	• **Compare with public methods & simple models (baselines)**. • **Adopt community-validated measures and benchmark datasets for evaluation**. • Compare related methods and alternatives on the same dataset. • Evaluate performance on a final independent held-out set. • **Use confidence intervals/error intervals and statistical tests to gauge prediction robustness**.

Key recommendations are bolded.

that choose a new training/testing split multiple times from the available data are often considered a preferred solution[22].

Overlap of training/testing data splits is particularly troublesome to overcome in biology. For example, in predictions on entire gene and protein sequences, independence of training and testing could be achieved by reducing the number of homologs in the data[10,23]. Modeling enhancer–promoter contacts requires a different criterion, for example, not sharing one endpoint[24]. Modeling protein domains might require the multidomain sequence to be split into its constituent domains before homology reduction[25]. In short, each area of biology has its own recommendations for handling overlapping data issues, and previous literature is vital to putting forward a strategy. In Box 1, we propose a set of questions under the category 'data splits' that should help to evaluate potential overlap between training and testing data.

Reporting statistics on the dataset size and distribution of data types can help show whether there is a good domain representation in all sets. Simple plots and/or tables showing the number of classes (classification), a histogram of real values binned (regression) and the different types of biological molecules in the data are vital pieces of information for each set. Further, in classification, inclusion of methods that address imbalanced classes[26,27] is also needed if the class frequencies show as much. Models trained on one dataset may not be successful in dealing with data coming from adjacent but not identical datasets, a phenomenon known as covariance shift. The scale of this effect has been demonstrated in several recent publications—for example, for prediction of disease risk from exome sequencing[28]. Although covariance shift remains an open problem, several potential solutions have been proposed in the area of transfer learning[29]. Moreover, the problem of training ML models that can generalize

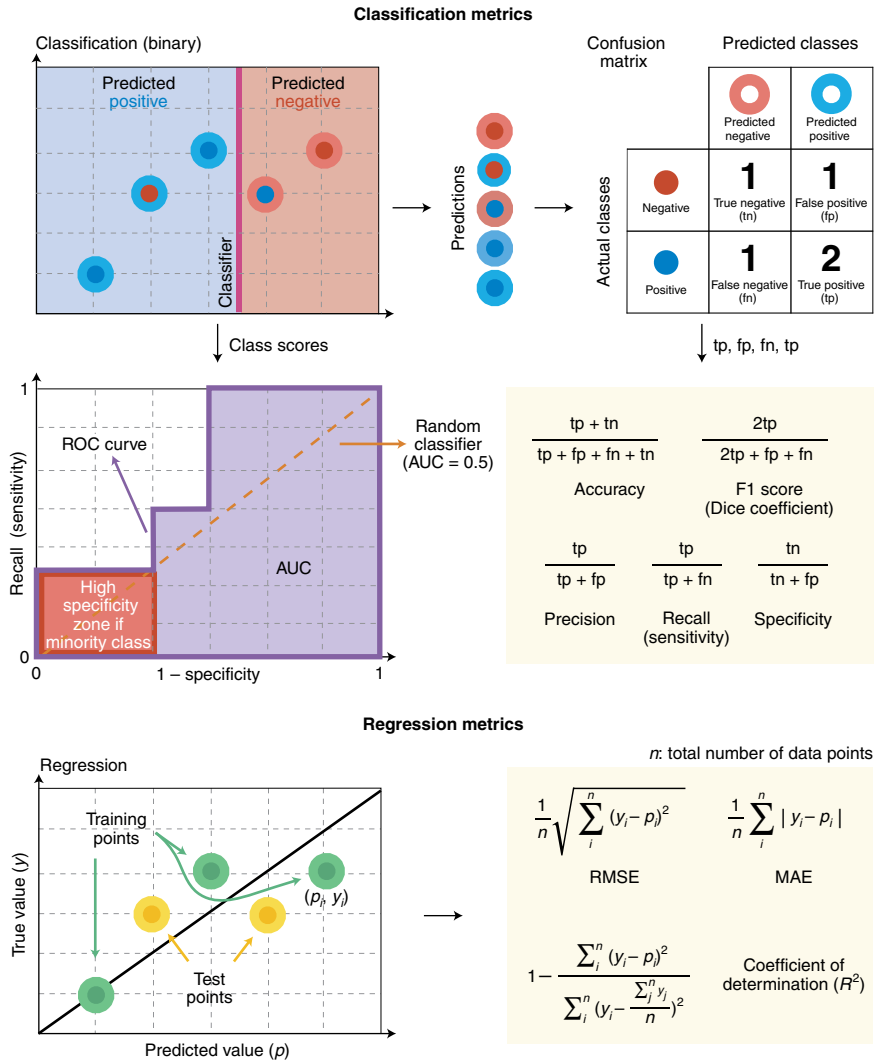

Fig. 2 | Metrics for ML. Top and middle: classification metrics. For binary classification, true positives (tp), false positives (fp), false negatives (fn) and true negatives (tn) together form the confusion matrix. As all classification measures can be calculated from combinations of these four basic values, the confusion matrix should be provided as a core metric. Several measures (shown as equations) and plots should be used to evaluate the ML methods. For descriptions of how to adapt these metrics to multi-class problems, see ref. [35]. Bottom: regression metrics. ML regression attempts to produce predicted values (p) matching experimental values (y). Metrics (shown as equations) attempt to capture the difference in various ways. Alternatively, a plot can provide a visual way to represent the differences. It is advisable to report all these measures in any ML work. ROC, receiver operating characteristic; AUC, area under the ROC curve; RMSE, root mean squared error; MAE, mean absolute error.

well on small training data usually requires special models and algorithms[30].

Lastly, it is important to make as much data available to the public as possible[12]. Having open access to the data used for experiments, including precise data splits, would ensure better reproducibility of published research and as a result will improve the overall quality of published ML papers. If datasets are not readily available in public repositories, authors should be encouraged to find the most appropriate vehicle—for example, ELIXIR deposition databases or Zenodo—to guarantee the long-term availability of such data.

Optimization. Optimization, also known as training, refers to the process of changing values that constitute the model (parameters and hyperparameters), including preprocessing steps, in a way that maximizes the model's ability to solve a given problem. A poor choice of optimization strategy may lead to issues such as over- or underfitting[31]. A model that has suffered severe overfitting will show an excellent performance on training data while performing poorly on unseen data, rendering it useless for real-life applications. On the other side of the spectrum, underfitting occurs when very simple models capable of capturing only straightforward dependencies between features are applied to data of a more complex nature. Algorithms for feature selection[32] can be employed to reduce the chances of overfitting. However, feature selection and other preprocessing actions come with their own recommendations. The main one is to abstain from using non-training data for feature selection and preprocessing—a particularly hard issue to spot for meta-predictors, which may lead to an overestimation of performance.

Finally, the release of files showing the exact specification of the optimization protocol and the type of parameters or hyperparameters are a vital characteristic of the final algorithm. Lack of documentation, including limited accessibility to relevant records for the parameters, hyperparameters and optimization protocol, may further compound the understanding of the overall model performance.

Model. Equally important aspects related to ML models are their interpretability and reproducibility. Interpretable models can infer causal relationships from the data and can output logical reasoning for each of their predictions. They are especially relevant in areas of discovery such as drug design[6] and diagnostics[33]. Conversely, black box models often give accurate predictions but may not provide insight in a way humans can understand into why they made the predictions. Both interpretable and black box models are discussed in more detail elsewhere[34]. However, developing recommendations on the choice of black box or interpretability is not straightforward as both have their merits. The main recommendation would be that there is a statement as to whether the model type is black box or interpretable (Box 1), and if it is interpretable, clear examples of interpretable output should be given.

Reproducibility is a key component for ensuring research outcomes can be further used and validated by the wider community. Poor model reproducibility extends beyond the documentation and reporting of the involved parameters, hyperparameters and optimization protocol. Lacking access to the various components of a model (source code, model files, parameter configurations

and executables), as well as having steep computational requirements for executing the trained models to generate predictions based on new data, can make reproducibility of the model either limited or practically impossible.

Evaluation. There are two types of evaluation scenarios in biological research. The first is the experimental validation of the predictions made by the ML model in the laboratory. This is highly desirable but beyond the scope of many ML studies. The second is a computational assessment of the model performance using established metrics. The following deals with the latter. There are a few possible risks in computational evaluation.

To start with performance metrics—that is, the quantifiable indicators of a model's ability to solve the given task—there are dozens of metrics available[35] for assessing different ML classification and regression problems. The plethora of options available, combined with the domain-specific expertise that might be required to select the appropriate metrics, can lead to the selection of inadequate performance measures. Often, there are critical assessment communities advocating certain performance metrics for biological ML models—for example, Critical Assessment of Protein Function Annotation (CAFA)[3] and Critical Assessment of Genome Interpretation (CAGI)[28]—and we recommend that a new algorithm should use metrics from the literature and community-promulgated critical assessments. In the absence of literature, the ones shown in Fig. 2 could be a starting point.

Once performance metrics are decided, methods published in the same biological domain must be cross-compared using appropriate statistical tests (for example, Student's *t*-test) and confidence intervals. Then, to prevent the release of ML methods that appear sophisticated but perform no better than simpler algorithms, baselines should be compared to the 'sophisticated' method and proven to be statistically inferior (for example, as in comparison of shallow vs. deep neural networks).

Open areas and limitations of the proposed recommendations

The primary goal of this work is to define best practices that can be of use in writing of ML-related papers while remaining agnostic as to the actual underlying solutions. We also expect that our proposed recommendations will be useful for peer reviewers of biological studies that use ML. Our intent is to trigger a discussion in the wider ML community leading to future work addressing possible solutions.

Several key issues related to reproducibility (for example, data are not published, data splits are not reported and model source code with its final parameters and hyperparameters are not released) can be aided by workflow systems that automate multistep processes to help to ensure that they are completely reproducible by tracking model parameters and exact versions of the source code and libraries. Examples of commonly used workflows include Galaxy[36] and Nextflow[37]. Another de facto standard practice in software engineering is using version control systems such as Github to create an online copy of the source code, which can also include parameters and documentation. Similar version control systems exist for datasets. Public repositories can store experimental data on demand on a long-term basis, enabling long-term reproducibility of the experiment. Existing software engineering tools can be used to address many of the DOME recommendations.

Although having further, more topic-specific recommendations in the future will undoubtedly be useful, in this work we aim to provide a first version that should be of general interest. Adapting the DOME recommendations to address the unique aspects of specific topics and domains would be a task of those particular communities. For example, having guidelines for data independence is tricky because each biological domain has its own set of guidelines for this. Nonetheless, we believe it is relevant to at least have a recommendation that authors describe how they achieved data split independence. Discussions on the correct independence strategies are needed for all of biology. Given constructive consultation processes with ML communities, relying on our own experience, it is our belief that this Comment can be useful as a first iteration of the recommendations for supervised ML in biology. This will have the added benefit of kickstarting community discussion with a coherent but rough set of goals, thus facilitating the overall engagement and involvement of key stakeholders. Topics to be addressed by communities include how to adapt DOME to entire pipelines and to unsupervised, semisupervised, reinforcement and other types of ML. For instance, in unsupervised learning, the evaluation metrics shown in Fig. 2 would not apply and a completely new set of definitions would be needed. Another debate, as AI becomes more commonplace in society, is that ML algorithms differ in their ability to explain learned patterns back to humans. Humans naturally prefer actions or predictions to be made with reasons given. This is the black box vs. interpretability debate, and we point those interested to excellent reviews in refs. [38–41] as a starting point for thoughtful discussions.

Finally, we address the governance structure by suggesting a community-managed governance model similar to that of the open-source initiatives[42]. Community-managed governance has been used in initiatives such as Minimum Information About a Microarray Experiment (MIAME)[43] or the Proteomics Standards Initiative (PSI) Molecular Interaction (MI) format[44]. This sort of structure ensures continuous community consultation and improvement of the recommendations in collaboration with academic (CLAIRE; see https://claire-ai.org/) and industrial (Pistoia Alliance; see https://www.pistoiaalliance.org/) networks. More importantly, this can be applied in particular to ML communities working with specific problems requiring more detailed guidelines—for example, imaging or clinical applications. We have set up a website (https://www.dome-ml.org/) where news and upcoming events will be posted to provide a platform for governance and community involvement around the DOME recommendations. As the recommendations and minimal requirements evolve over time, a version history will be available on the website. A template supplementary checklist in human-readable (spreadsheet) and machine-readable (YAML) format, as well as software for the automatic conversion of a YAML file into a human-readable one, are available from a dedicated GitHub repository (https://github.com/MachineLearning-ELIXIR/dome-ml).

Conclusion

The objective of our recommendations is to increase the reproducibility and clarity of ML methods for the reader, the experimentalist, the reviewer and the wider community. We accept that these recommendations are not complete and should be viewed as a first iteration of a consensus-based community discussion. One of the most pressing issues is to agree to a standardized data structure to describe the most relevant features of the ML methods being presented. As a first step in addressing this issue, we recommend including an ML summary table, derived from Box 1, in manuscripts describing ML-based studies (Supplementary Table 1). We recommend including the following sentence in the Methods section of a manuscript: "To support the reproducibility of the machine learning method of this study, the machine learning summary table (Table *N*) is

included in the supporting information as per DOME recommendations (https://doi.org/10.1038/s41592-021-01205-4)."

We believe that the development of standardized reporting guidelines has the potential to make a major impact in increasing the quality of publishing ML methods. First, the current disparity among manuscripts in reporting key elements of the ML method can make reviewing and assessing the ML method challenging. Second, certain performance measures and essential statistics that may affect the validity of the publication's conclusions are sometimes not mentioned at all. Third, there are unexplored opportunities associated with meta-analysis of ML datasets. Access to large sets of data can both enhance the comparison between methods and facilitate the development of better-performing methods while reducing unnecessary repetition of data generation. We believe that our recommendations to include a "machine learning summary table" and to make datasets available will greatly benefit the ML community and improve its standing with the intended users of these methods. ❐

Ian Walsh[1,27], Dmytro Fishman[2,27], Dario Garcia-Gasulla[3], Tiina Titma[4], Gianluca Pollastri[5], ELIXIR Machine Learning Focus Group*, Jennifer Harrow[6]✉, Fotis E. Psomopoulos[7]✉ and Silvio C. E. Tosatto[8]✉

[1]*Bioprocessing Technology Institute, Agency for Science, Technology and Research, Singapore, Singapore.* [2]*Institute of Computer Science, University of Tartu, Tartu, Estonia.* [3]*Barcelona Supercomputing Center (BSC), Barcelona, Spain.* [4]*School of Information Technologies, Tallinn University of Technology, Tallinn, Estonia.* [5]*School of Computer Science, University College Dublin, Dublin, Ireland.* [6]*ELIXIR Hub, Wellcome Genome Campus, Hinxton, UK.* [7]*Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece.* [8]*Department of Biomedical Sciences, University of Padua, Padua, Italy.* [27]*These authors contributed equally: Ian Walsh, Dmytro Fishman.*

**A list of authors and their affiliations appears at the end of the paper.*

✉*e-mail: jen.harrow@elixir-europe.org; fpsom@certh.gr; silvio.tosatto@unipd.it*

Published online: 27 July 2021
https://doi.org/10.1038/s41592-021-01205-4

References

1. Baron, C. S. et al. *Cell* **179**, 527–542.e19 (2019).
2. Libbrecht, M. W. & Noble, W. S. *Nat. Rev. Genet.* **16**, 321–332 (2015).
3. Radivojac, P. et al. *Nat. Methods* **10**, 221–227 (2013).
4. Franciosa, G., Martinez-Val, A. & Olsen, J. V. *Nat. Biotechnol.* **38**, 285–286 (2020).
5. Yang, J. H. et al. *Cell* **177**, 1649–1661.e9 (2019).
6. Vamathevan, J. et al. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
7. Rajkomar, A., Dean, J. & Kohane, I. *N. Engl. J. Med.* **380**, 1347–1358 (2019).
8. Anonymous. *Nat. Mater.* **18**, 407 (2019).
9. Littmann, M. et al. *Nat. Mach. Intell.* **2**, 18–24 (2020).
10. Walsh, I., Pollastri, G. & Tosatto, S. C. E. *Brief. Bioinform.* **17**, 831–840 (2016).
11. Bishop, D. *Nature* **568**, 435 (2019).
12. Hutson, M. *Science* **359**, 725–726 (2018).
13. Schwartz, D. *Essays Biochem.* **52**, 165–177 (2012).
14. Piovesan, D. et al. *PLOS Comput. Biol.* **16**, e1007967 (2020).
15. Wilkinson, M. D. et al. *Sci. Data* **3**, 160018 (2016).
16. Sandve, G. K., Nekrutenko, A., Taylor, J. & Hovig, E. *PLOS Comput. Biol.* **9**, e1003285 (2013).
17. Grüning, B. et al. *Cell Syst.* **6**, 631–635 (2018).
18. Jones, D. T. *Nat. Rev. Mol. Cell Biol.* **20**, 659–660 (2019).
19. Norgeot, B. et al. *Nat. Med.* **26**, 1320–1324 (2020).
20. Luo, W. et al. *J. Med. Internet Res.* **18**, e323 (2016).
21. Harrow, J. et al. *EMBO J.* **40**, e107409 (2021).
22. Kohavi, R. *Artif. Intell.* **14**, 1137–1145 (1995).
23. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. *Protein Sci.* **1**, 409–417 (1992).
24. Xi, W. & Beer, M. A. *PLOS Comput. Biol.* **14**, e1006625 (2018).
25. Zhou, X., Hu, J., Zhang, C., Zhang, G. & Zhang, Y. *Proc. Natl Acad. Sci. USA* **116**, 15930–15938 (2019).
26. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
27. He, H., Bai, Y., Garcia, E. A. & Li, S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. *IEEE Int. Joint Conf. Neural Networks* 1322–1328 (IEEE, 2008).
28. Daneshjou, R. et al. *Hum. Mutat.* **38**, 1182–1192 (2017).
29. Pan, S. J. & Yang, Q. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
30. Vinyals, O., Blundell, C., Lillicrap, T. & Wierstra, D. *Adv. Neural Inf. Process. Syst.* **29**, 3630–3638 (2016).
31. Mehta, P. et al. *Phys. Rep.* **810**, 1–124 (2019).
32. Guyon, I. & Elisseeff, A. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003).
33. He, J. et al. *Nat. Med.* **25**, 30–36 (2019).
34. Rudin, C. *Nat. Mach. Intell.* **1**, 206–215 (2019).
35. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. & Nielsen, H. *Bioinformatics* **16**, 412–424 (2000).
36. Goecks, J., Nekrutenko, A. & Taylor, J. *Genome Biol.* **11**, R86 (2010).
37. Di Tommaso, P. et al. *Nat. Biotechnol.* **35**, 316–319 (2017).
38. Arrieta, A. B. et al. *Inf. Fusion* **58**, 82–115 (2020).
39. Guidotti, R. et al. *ACM Comput. Surv.* **51**, 1–42 (2018).
40. Adadi, A. & Berrada, M. *IEEE Access* **6**, 52138–52160 (2018).
41. Holm, E. A. *Science* **364**, 26–27 (2019).
42. O'Mahony, S. J. *Manag. Gov.* **11**, 139–150 (2007).
43. Brazma, A. et al. *Nat. Genet.* **29**, 365–371 (2001).
44. Hermjakob, H. et al. *Nat. Biotechnol.* **22**, 177–183 (2004).

Acknowledgements

The work of the Machine Learning Focus Group was funded by ELIXIR, the research infrastructure for life-science data. IW was funded by the A*STAR Career Development Award (project no. C210112057) from the Agency for Science, Technology and Research (A*STAR), Singapore. D.F. was supported by Estonian Research Council grants (PRG1095, PSG59 and ERA-NET TRANSCAN-2 (BioEndoCar)); Project No 2014-2020.4.01.16-0271, ELIXIR and the European Regional Development Fund through EXCITE Center of Excellence. S.C.E.T. has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie Grant agreements No. 778247 and No. 823886, and Italian Ministry of University and Research PRIN 2017 grant 2017483NH8.

Author contributions

I.W., D.F., J.H., F.E.P. and S.C.E.T. guided the development, writing and final edits. All members of the ELIXIR Machine Learning Focus Group contributed to the discussions leading to the recommendations and writing of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-021-01205-4.

Peer review information *Nature Methods* thanks Jeremy Goecks, Amalio Telenti and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

ELIXIR Machine Learning Focus Group

Emidio Capriotti[9], Rita Casadio[9,10], Salvador Capella-Gutierrez[3], Davide Cirillo[3], Alessio Del Conte[8], Alexandros C. Dimopoulos[11], VictoriaDominguezDelAngel[12], JoaquinDopazo[13], Piero Fariselli[14], José Maria Fernández[3], Florian Huber[15], Anna Kreshuk[16], Tom Lenaerts[17], Pier Luigi Martelli[9], Arcadi Navarro[18,19,20], Pilib Ó Broin[21], Janet Piñero[18,22], Damiano Piovesan[8], Martin Reczko[11], Francesco Ronzano[18,21], Venkata Satagopam[23], Castrense Savojardo[9], Vojtech Spiwok[24], Marco Antonio Tangaro[10], Giacomo Tartari[10], David Salgado[25], Alfonso Valencia[3,19] and Federico Zambelli[26]

[9]*Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy.* [10]*Istituto di Biomembrane, Bioenergetica e Biotecnologie Molecolari (IBIOM), National Research Council (CNR), Bari, Italy.* [11]*Institute for Fundamental Biomedical Science, Biomedical Sciences Research Center "Alexander Fleming", Athens, Greece.* [12]*Centre National de Recherche Scientifique, University Paris-Saclay, IFB, Gif-sur-Yvette, France.* [13]*Clinical Bioinformatics Area, Fundación Progreso y Salud, Sevilla, Spain.* [14]*Department of Medical Sciences, University of Turin, Turin, Italy.* [15]*Netherlands eScience Center, Amsterdam, the Netherlands.* [16]*European Molecular Biology Laboratory (EMBL), Heidelberg, Germany.* [17]*Université Libre de Bruxelles, Vrije Universiteit Brussel and Interuniversity Institute of Bioinformatics in Brussels, Brussels, Belgium.* [18]*Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain.* [19]*Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain.* [20]*Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain.* [21]*School of Mathematics, Statistics & Applied Mathematics, National University of Ireland, Galway, Ireland.* [22]*Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain.* [23]*Luxembourg Centre for Systems Biomedicine, University of Luxembourg and ELIXIR-Luxembourg, Luxembourg, Luxembourg.* [24]*Department of Biochemistry and Microbiology, University of Chemistry and Technology, Prague and ELIXIR-Czech Republic, Prague, Czech Republic.* [25]*Aix Marseille University, INSERM, MMG UMR1251, Marseille, France.* [26]*Department of Biosciences, University of Milan, Milan, Italy.*