




mlr3extralearners: Expanding the mlr3 Ecosystem with Community-Driven Learner Integration

Sebastian Fischer^{1,2,3}, John Zobolas⁴, Raphael Sonabend¹⁵, Marc Becker^{2,3}, Michel Lang¹, Martin Binder^{2,3}, Lennart Schneider^{2,3}, Lukas Burk^{2,3,5,6}, Patrick Schratz¹⁸, Byron C. Jaeger¹³, Stephen A Lauer⁷, Lorenz A. Kapsner⁸, Maximilian Mücke², Zezhi Wang⁹, Damir Pulatov¹⁴, Keenan Ganz¹⁰, Henri Funk^{3,11,12}, Liana Harutyunyan¹⁷, Pierre Camilleri¹⁶, Philipp Kopper³, Andreas Bender^{2,3}, and Bernd Bischl^{2,3}

¹ TU Dortmund University, Germany ² Department of Statistics, LMU Munich, Germany ³ Munich Center for Machine Learning (MCML), Germany ⁴ Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, Norway ⁵ Leibniz Institute for Prevention Research and Epidemiology (BIPS), Bremen, Germany ⁶ Faculty of Mathematics and Computer Science, University of Bremen, Germany ⁷ Certilytics, Inc., 9200 Shelbyville Rd, Louisville, KY, 40222, USA ⁸ Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany ⁹ Department of Statistics and Finance/International Institute of Finance, School of Management, University of Science and Technology of China, Hefei, Anhui, China ¹⁰ School of Environmental and Forest Sciences, University of Washington, Seattle ¹¹ Department of Geography, LMU Munich, Germany ¹² Statistical Consulting Unit StaBLab, LMU Munich, Germany ¹³ Wake Forest University School of Medicine, Department of Biostatistics and Data Science, Division of Public Health Sciences Winston-Salem, North Carolina ¹⁴ University of North Carolina Wilmington ¹⁵ OSPO Now ¹⁶ multi, 8 passage Brûlon, 75012 PARIS, France ¹⁷ ServiceTitan, Inc., Glendale, California ¹⁸ devXY GmbH

DOI: 10.xxxxxx/draft

Software

- Review 
- Repository 
- Archive 

Editor: Richard Littauer 

Reviewers:

- @ritika-giri
- @kelly-sovacool

Submitted: 24 March 2025

Published: unpublished

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Summary

The `mlr3` ecosystem is a versatile toolbox for machine learning (ML) in R (R Core Team, 2019) that is targeted towards both practitioners and researchers (Bischl et al., 2024). The core `mlr3` package (Lang et al., 2019) defines the standardized interface for ML, but its goal is not to implement algorithms. This is, e.g., done by the `mlr3learners` extension (Lang, Au, et al., 2024) that connects 21 stable learning algorithms from various R packages to the `mlr3` ecosystem that serve as a good starting point for many ML tasks. In addition, `mlr3extralearners` is a *community-driven* package that integrates many more methods. The package currently wraps **85 different ML algorithms** from many different R packages, for tasks such as classification, regression, and survival analysis. This enables users to seamlessly access and utilize these learners directly within their workflows. One of the strengths of `mlr3` is the design and implementation of large-scale benchmark experiments. For example, datasets for such experiments can be easily obtained from the OpenML¹ repository (Vanschoren et al., 2014) via the `mlr3oml` package (Lang & Fischer, 2024). Furthermore, strong support for parallelization, including execution on high-performance computing clusters via `batchtools` (Lang et al., 2017) and its `mlr3` integration `mlr3batchmark` (Becker & Lang, 2024), is available and well documented (Fischer et al., 2024). In combination, these tools allow for large-scale empirical investigations, which has, for example, been used to collect and analyze data about hyperparameter landscapes of ML algorithms (Binder et al., 2020). An overview of all `mlr3` learners, including those introduced through `mlr3extralearners`, is available on the `mlr3`

¹<https://openml.org>

43 website².

44 Beyond accessibility, `mlr3extralearners` also allows `mlr3` users to easily connect their own
45 algorithms to the interface. This **enriches each learner with extensive metadata** about its hyper-
46 parameter space, prediction types, and other key attributes. Furthermore, `mlr3extralearners`
47 includes robust mechanisms for **quality assurance**, such as regular automated sanity checks and
48 verification tests that ensure learner parameters are consistent and up-to-date with the latest
49 versions of their underlying R packages. In order to allow the integration of learners that are not
50 available on CRAN, the package is hosted on the `mlr` R-universe³. By providing a standardized
51 interface and comprehensive metadata for each learner, `mlr3extralearners` enhances the
52 FAIRness (findability, accessibility, interoperability, and reusability) of ML algorithms within
53 the R ecosystem (Wilkinson et al., 2016).

54 Statement of Need

55 ML often requires practitioners to navigate a diverse array of modeling problems, each
56 with unique demands such as predictive performance, prediction speed, interpretability, or
57 compatibility with specific data types and tasks. To address this challenge, packages like
58 `mlr3`'s predecessor `mlr` (Bischl et al., 2016), `caret` (Kuhn, 2008), and more recently `parsnip`
59 (Kuhn & Vaughan, 2024) from the `tidymodels` ecosystem (Kuhn & Wickham, 2020) were
60 designed to provide unified interfaces for simplifying model experimentation. For instance,
61 `parsnip` provides a clean and consistent way to define models, enabling users to experiment
62 with different algorithms without dealing with the nuances of underlying package syntax.
63 Similarly, the `mlr3` ecosystem aims to streamline model selection and experimentation, making
64 it a versatile toolbox for ML in R.

65 Within this ecosystem, `mlr3extralearners` plays a crucial role by providing a comprehensive
66 collection of external ML algorithms integrated into the `mlr3` framework. This ensures that
67 users can access a wide variety of learners to meet their needs and choose the most appropriate
68 algorithm for their particular problem. While connecting new learners to `mlr3` is straightforward
69 and can be done on a per-need basis, integrating them into `mlr3extralearners` benefits
70 the broader community by avoiding redundant effort and ensuring accessibility for all users.
71 Additionally, contributions to `mlr3extralearners` are reviewed by the package maintainers,
72 providing a layer of quality assurance. This review process ensures that integrated learners
73 work as expected and adhere to the high standards of the `mlr3` ecosystem.

74 Beyond its utility for users, `mlr3extralearners` also offers significant advantages for developers
75 of ML packages. By integrating a new algorithm into the `mlr3` ecosystem, developers can
76 immediately make their methods accessible to a wider audience. This integration facilitates
77 seamless tuning (Becker et al., 2024) and preprocessing (Binder et al., 2021) through the
78 broader `mlr3` framework, enhancing the usability and impact of their work.

79 Features

80 The core functionality of `mlr3extralearners` is to integrate new learners into the `mlr3`
81 ecosystem, allowing users to access a wide array of learning algorithms through a unified syntax
82 and standardized interface. However, the advantages of `mlr3extralearners` go well beyond
83 simple integration.

84 Metadata

85 One core feature of the `mlr3` ecosystem is that it annotates learners with extensive metadata.

²<https://mlr-org.com/learners.html>

³<https://mlr-org.r-universe.dev>

- 86 ■ **Hyperparameter management:** The hyperparameter spaces of learners are defined
87 using ParamSet objects from the paradox package (Lang, Bischl, et al., 2024). Each
88 hyperparameter is explicitly typed, with annotations for feasible values. This ensures
89 valid configurations and simplifies tasks like hyperparameter tuning.
- 90 ■ **Task and prediction types:** Learners are categorized with respect to their task type (e.g. as
91 classification, regression or survival analysis (Sonabend et al., 2021)) and prediction
92 types (e.g. probabilities or response predictions). This allows users to easily identify
93 suitable learners for their specific modeling tasks.
- 94 ■ **Standardized properties:** Learners are annotated with detailed attributes, including the
95 types of features they can process and their support for functionalities such as feature
96 selection, importance scoring, handling missing values, or monitoring performance on a
97 separate validation set during training among others. This allows users to have a clear
98 understanding of a learner's capabilities and limitations and assess if it aligns with the
99 specific requirements of their workflows, reducing trial-and-error and streamlining the
100 modeling process.

101 Functional Correctness

102 Integrating learners from diverse R packages poses challenges, on the one hand because changes
103 in upstream APIs need to be reflected in mlr3extralearners and on the other hand because
104 we want to ensure a high level of quality of algorithms connected to mlr3. mlr3extralearners
105 addresses both points through automated checks:

- 106 ■ **Interface consistency:** The package regularly verifies that each learner adheres to the
107 expected interface of the latest released version of its upstream function. When new
108 parameters are introduced or existing ones changed or removed, the tests fail until the
109 parameter sets are updated accordingly.
- 110 ■ **Automated testing:** In general, writing unit tests for ML algorithms is challenging,
111 because of edge-cases, numeric errors, and the fact that the input to these algorithms
112 can be arbitrary datasets. Aimed at addressing these challenges, mlr3extralearners
113 performs regular automated tests on all learners. These tests include sanity checks that,
114 e.g., verify that the learners produce sensible predictions for simple randomly generated
115 datasets. Furthermore, the tests also validate the learners' metadata annotations, such
116 as whether a learner can actually handle missing values or is able to produce importance
117 scores. In the past, these tests have detected bugs in some upstream packages and we
118 have subsequently notified their authors.

119 Simplified Integration of New Learners

120 To streamline the addition of new learners, mlr3extralearners provides robust support tools:

- 121 ■ **Code templates:** Predefined templates are available for both the learner implementation
122 and associated test files. Contributors can utilize these templates through an R function
123 that accepts learner metadata and generates new R code files based on the templates.
124 This approach pre-fills as much information as possible, minimizing the input required
125 from the contributor. Note that these templates can also be used when learners are only
126 used locally for specific projects and not contributed to mlr3extralearners.
- 127 ■ **Guides and resources:** The package website⁴ contains an extensive tutorial, as well as a
128 curated list of common issues encountered during learner integration, making the process
129 accessible for contributors of all experience levels. Additionally, every integrated learner
130 includes a simple example of usage in the documentation, ensuring that users can quickly
131 understand how to utilize the learner effectively within the mlr3 ecosystem.

⁴<https://mlr3extralearners.mlr-org.com>

Community Impact and Future Directions

`mlr3extralearners` is a direct result of the contributions from a diverse community of authors and developers. The authors of this paper themselves have been actively involved in integrating learners, providing quality assurance, and maintaining the package's infrastructure. Their contributions, such as the addition of learners for specialized tasks like survival analysis and high-dimensional data, highlight the impact that thoughtful integration has on the `mlr3` ecosystem. This ongoing effort illustrates the transformative potential of **community-driven development**, ensuring that `mlr3extralearners` continues to grow as a dynamic and inclusive repository for ML algorithms.

Future work will also focus on expanding the ecosystem with more deep learning methods through `mlr3torch` (Fischer & Binder, 2025), which aims to seamlessly integrate deep learning models and neural network architectures within the `mlr3` framework.

Acknowledgements

Sebastian Fischer is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 460135501 (NFDI project MaRDI). John Zobolas received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016851, project PANCAIM.

References

- Becker, M., & Lang, M. (2024). *mlr3batchmark: Batch experiments for 'mlr3'*. <https://CRAN.R-project.org/package=mlr3batchmark>
- Becker, M., Lang, M., Richter, J., Bischl, B., & Schalk, D. (2024). *mlr3tuning: Hyperparameter optimization for 'mlr3'*. <https://github.com/mlr-org/mlr3tuning>
- Binder, M., Pfisterer, F., & Bischl, B. (2020). Collecting empirical data about hyperparameters for data driven AutoML. *Democratizing Machine Learning Contributions in AutoML and Fairness*, 93.
- Binder, M., Pfisterer, F., Lang, M., Schneider, L., Kotthoff, L., & Bischl, B. (2021). *mlr3pipelines - Flexible Machine Learning Pipelines in R*. *Journal of Machine Learning Research*, 22(184), 1–7. <http://jmlr.org/papers/v22/21-0281.html>
- Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., & Jones, Z. M. (2016). *MLr: Machine learning in r*. *Journal of Machine Learning Research*, 17(170), 1–5.
- Bischl, B., Sonabend, R., Kotthoff, L., & Lang, M. (Eds.). (2024). *Applied machine learning using mlr3 in R*. CRC Press. ISBN: 9781032507545
- Fischer, S., & Binder, M. (2025). *mlr3torch: Deep learning with 'mlr3'*. <https://CRAN.R-project.org/package=mlr3torch>
- Fischer, S., Lang, M., & Becker, M. (2024). Large-scale benchmarking. In B. Bischl, R. Sonabend, L. Kotthoff, & M. Lang (Eds.), *Applied machine learning using mlr3 in R*. CRC Press. https://mlr3book.ml-org.com/large-scale_benchmarking.html
- Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, 28, 1–26.
- Kuhn, M., & Vaughan, D. (2024). *Parsnip: A common API to modeling and analysis functions*. <https://github.com/tidymodels/parsnip>

- 174 Kuhn, M., & Wickham, H. (2020). *Tidymodels: A collection of packages for modeling and*
175 *machine learning using tidyverse principles*. <https://www.tidymodels.org>
- 176 Lang, M., Au, Q., Coors, S., Schratz, P., & Becker, M. (2024). *mlr3learners: Recommended*
177 *learners for 'mlr3'*. <https://CRAN.R-project.org/package=mlr3learners>
- 178 Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio,
179 G., Kotthoff, L., & Bischl, B. (2019). mlr3: A modern object-oriented machine learning
180 framework in R. *Journal of Open Source Software*, 4(44), 1903. [https://doi.org/10.21105/](https://doi.org/10.21105/JOSS.01903)
181 [JOSS.01903](https://doi.org/10.21105/JOSS.01903)
- 182 Lang, M., Bischl, B., Richter, J., Sun, X., & Binder, M. (2024). *Paradox: Define and work with*
183 *parameter spaces for complex algorithms*. <https://CRAN.R-project.org/package=paradox>
- 184 Lang, M., Bischl, B., & Surmann, D. (2017). Batchtools: Tools for r to work on batch systems.
185 *Journal of Open Source Software*, 2(10), 135.
- 186 Lang, M., & Fischer, S. (2024). *mlr3oml: Connector between 'mlr3' and 'OpenML'*. <https://CRAN.R-project.org/package=mlr3oml>
- 187
- 188 R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation
189 for Statistical Computing. <https://www.R-project.org/>
- 190 Sonabend, R., Király, F. J., Bender, A., Bischl, B., & Lang, M. (2021). mlr3proba: an R
191 package for machine learning in survival analysis. *Bioinformatics*, 37(17), 2789–2791.
192 <https://doi.org/10.1093/BIOINFORMATICS/BTAB039>
- 193 Vanschoren, J., Van Rijn, J. N., Bischl, B., & Torgo, L. (2014). OpenML: Networked science
194 in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2), 49–60.
- 195 Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A.,
196 Blomberg, N., Boiten, J.-W., Silva Santos, L. B. da, Bourne, P. E., & others. (2016). The
197 FAIR guiding principles for scientific data management and stewardship. *Scientific Data*,
198 3(1), 1–9.