Supplementary Data to the Paper:

Improved Outcome Prediction Across Data Sources Through Robust Parameter Tuning

Nicole Ellenbach[*,1,2], Anne-Laure Boulesteix [1], Bernd Bischl [3], Kristian Unger [2,4], Roman Hornung [1]

[1] Institute for Medical Information Processing, Biometry and Epidemiology, University of Munich, Munich, Germany

[2] Research Unit Radiation Cytogenetics, Helmholtz Zentrum Munich, German Research Center for Environmental Health GmbH, Neuherberg, Germany

[3] Department of Statistics, University of Munich, Munich, Germany

[4] Department of Radiation Oncology, University Hospital, University of Munich, Munich, Germany

[*]To whom correspondence should be addressed: nellenbach@ibe.med.uni-muenchen.de

A Overview shown in the main paper extended by the procedures that include the external data set for training

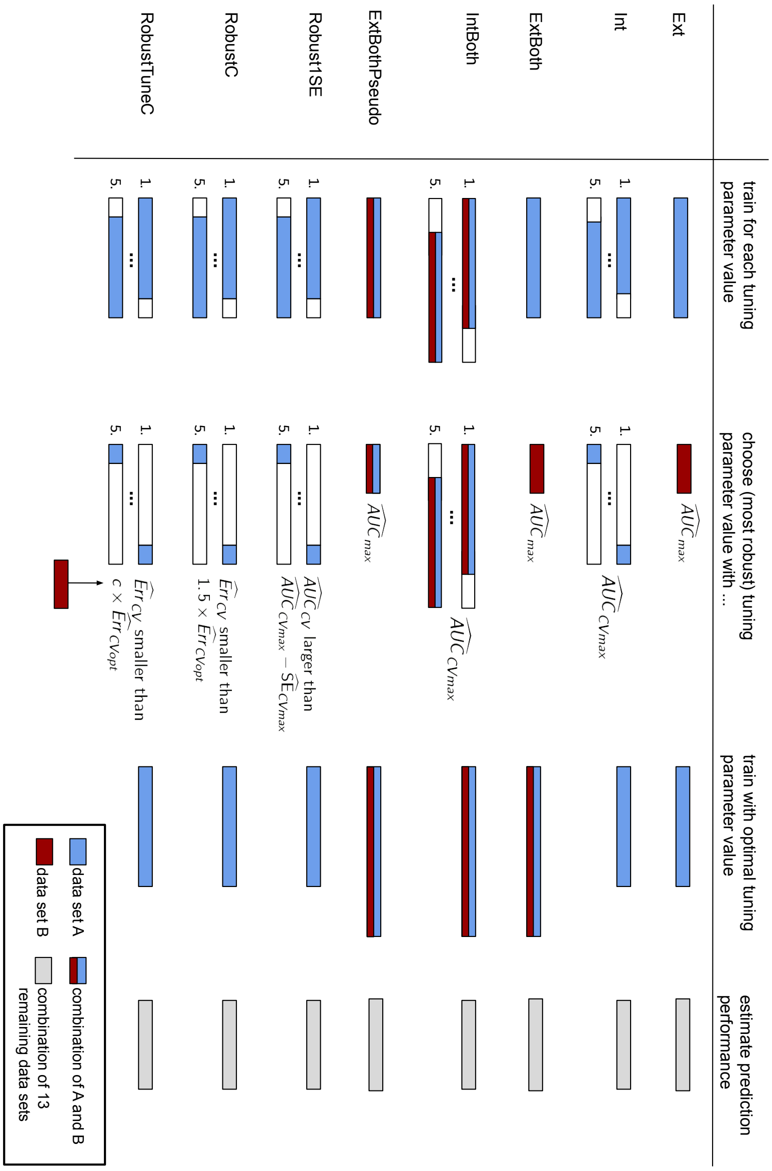

Fig. S1: Overview of the practically motivated approaches for external/ internal tuning and the procedures for robust tuning including the procedures that include the external data set for training. See Section 2.5.3 and 2.5.4 of the main paper for more details on `RobustC` and `RobustTuneC`

B Procedures using both A and B for training

The procedures `Ext` and `Int` described in Section 2.5.1 of the main paper use the external data set only for choosing the optimal tuning parameter value (`Ext`) or not at all (`Int`). It might, however, be worthwhile to use data set B also for training. We included two such procedures in our comparison study.

In the first one, `ExtBoth`, external tuning is performed (as in `Ext`). Then A and B are combined with batch effects adjustment using ComBat (Johnson et al., 2007) to finally train the prediction rule with the optimal tuning parameter value.

In the second variant, `IntBoth`, A and B are first combined (again performing ComBat to adjust for batch effects). Then internal tuning is applied to the combined data (i.e., `Int` is applied to the combination of A and B). This method is followed by researchers who are only aware of internal tuning and want to use all observations (from A and B) for training the prediction rule.

Finally, we also include an additional variant, `ExtBothPseudo`, of the first approach `ExtBoth`. This procedure would not be used in practice. Its aim is to assess whether it is important that, in `ExtBoth`, the external data set B comes from a different distribution than A. To assess this, we proceed as follows. As with `IntBoth` we first combine A and B adjusting for batch effects using ComBat. Subsequently, we randomly split the combined data into two parts. The first part has the same size n_A as data set A and the second part has the same size n_B as data set B. Subsequently, the `ExtBoth` procedure is applied to these two parts, which play the role of training and (pseudo) external tuning data set, respectively. In order to reduce the dependency of the results of `ExtBothPseudo` on the specific random splitting, we again repeat the procedure for 10 random splits. The 10 optimized tuning parameter values and the 10 obtained AUC values are subsequently averaged. This approach is essentially the same as `ExtBoth`, except that the training and (pseudo) external tuning data set follow the same (mixture) distribution.

C Extended results of the preliminary study of the conceptual comparison of external and internal tuning

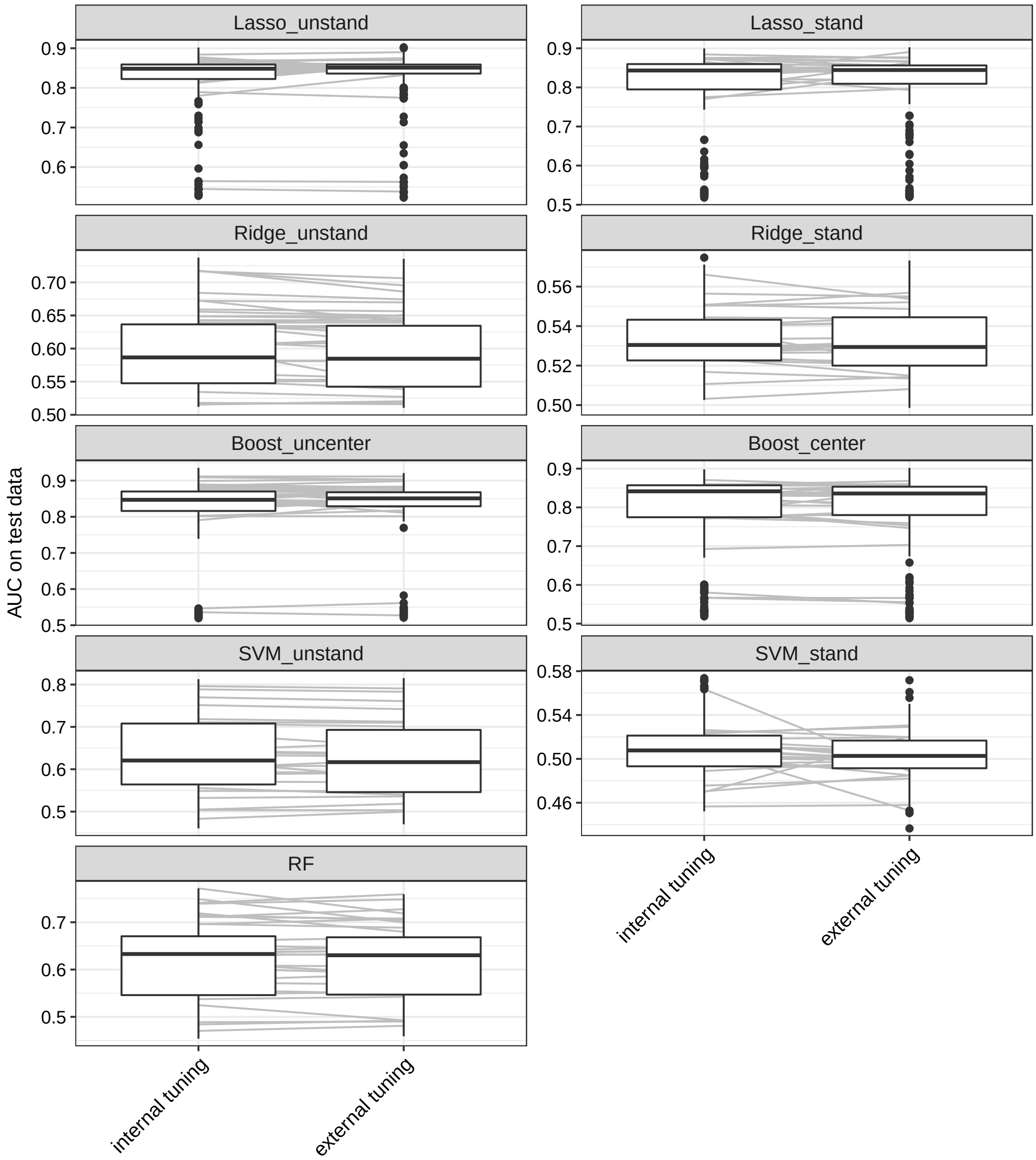

Fig. S2: Extended results of the preliminary study. Prediction performance estimates based on independent test data in the conceptual comparison of external and internal tuning. The gray lines connect the values of pairs that share the same training data sets, where in each case, for the sake of clarity, we do not show a line for each of the pairs, but merely for a random subset of 30 pairs

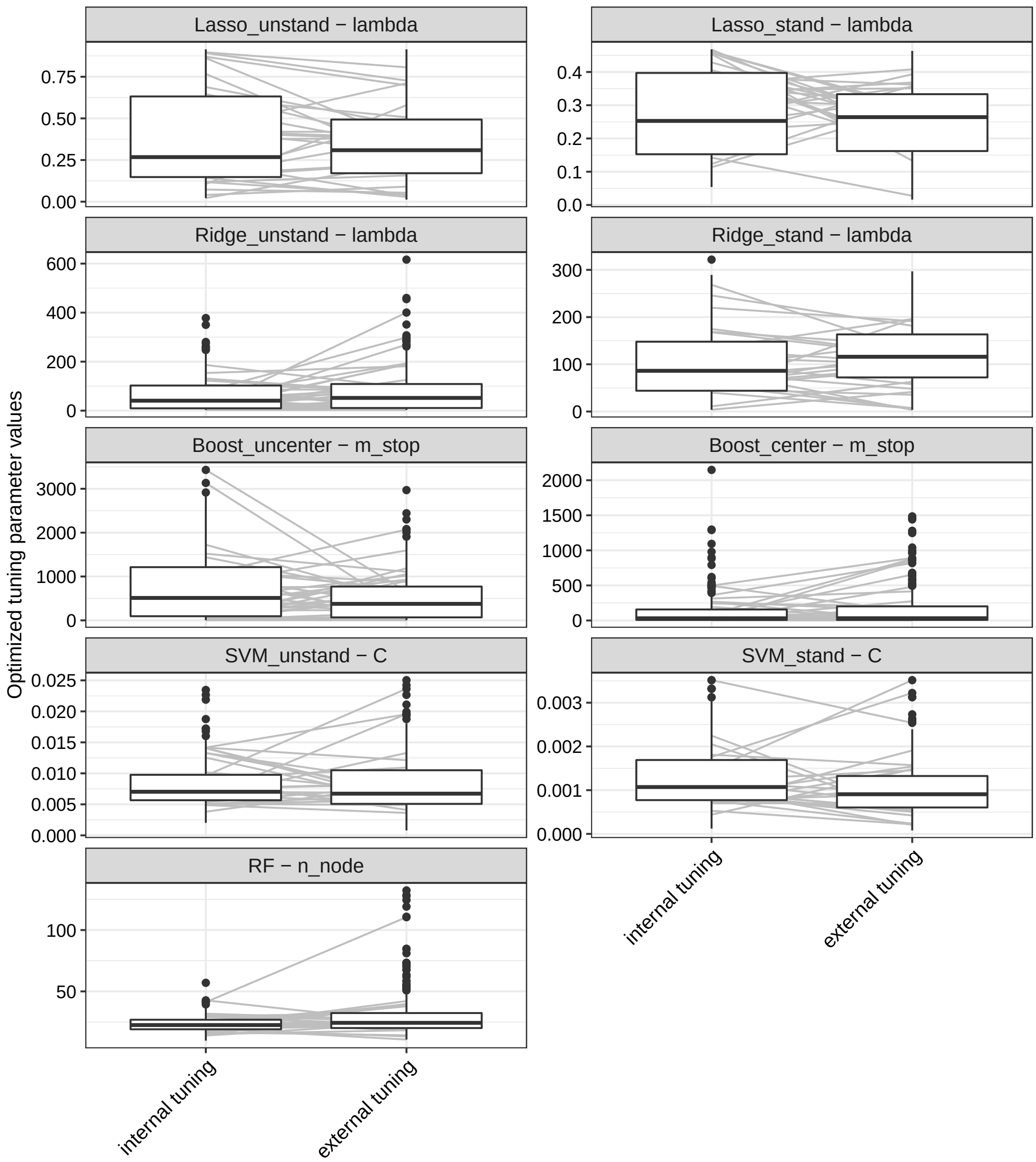

Fig. S3: Extended results of the preliminary study. Chosen tuning parameter values in the conceptual comparison of external and internal tuning. The gray lines connect the values of pairs that share the same training data sets, where in each case, for the sake of clarity, we do not show a line for each of the pairs, but merely for a random subset of 30 pairs

D Extended results of the main study. Prediction performance estimates for various practically motivated approaches to external, internal, and robust tuning

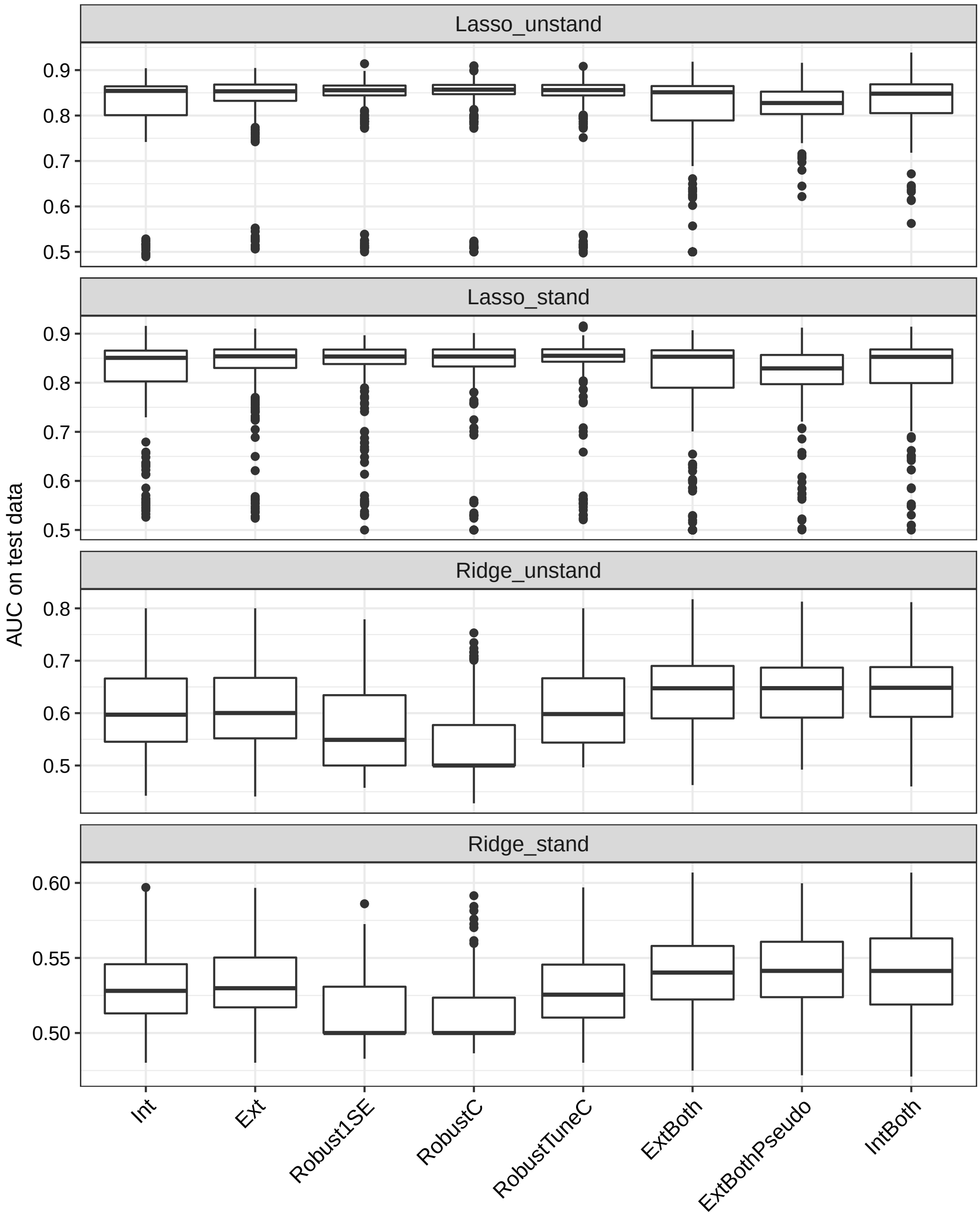

Fig. S4: Extended results of the main study. Prediction performance estimates for various practically motivated approaches to external, internal, and robust tuning - I

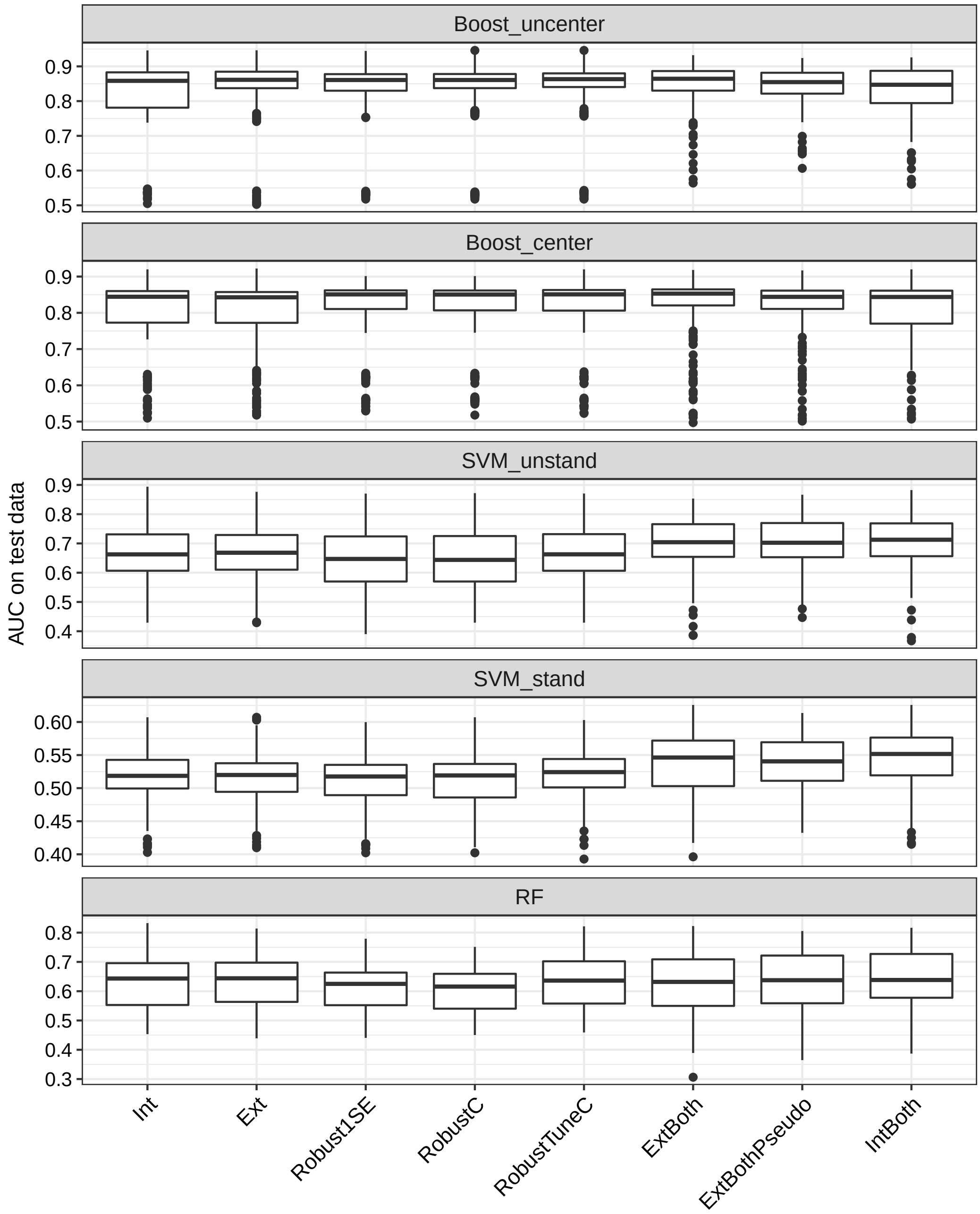

Fig. S5: Extended results of the main study. Prediction performance estimates for various practically motivated approaches to external, internal, and robust tuning - II

E Extended results of the main study. Chosen tuning parameter values for various practically motivated approaches to external, internal, and robust tuning

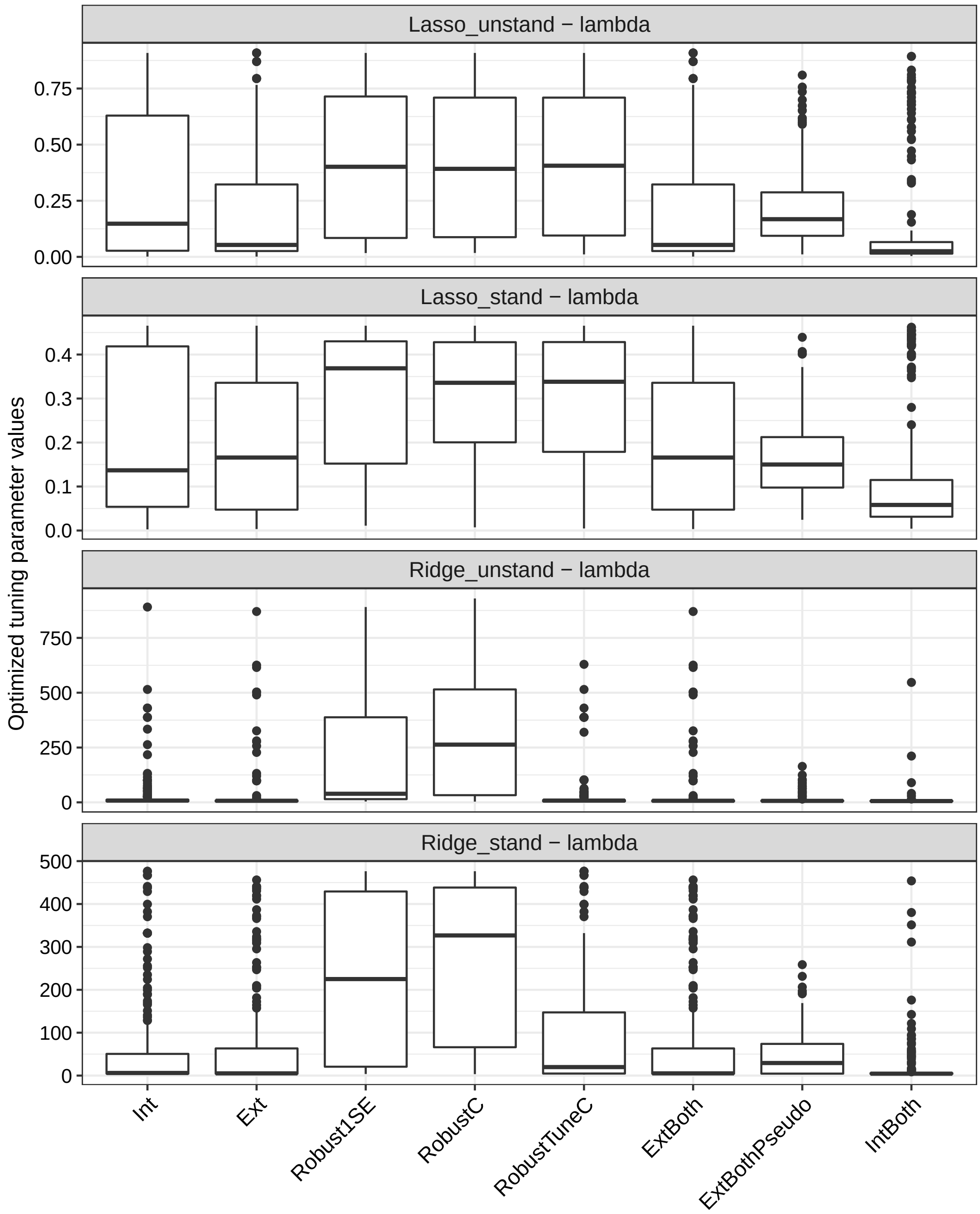

Fig. S6: Extended results of the main study. Chosen tuning parameter values for various practically motivated approaches to external, internal, and robust tuning - I

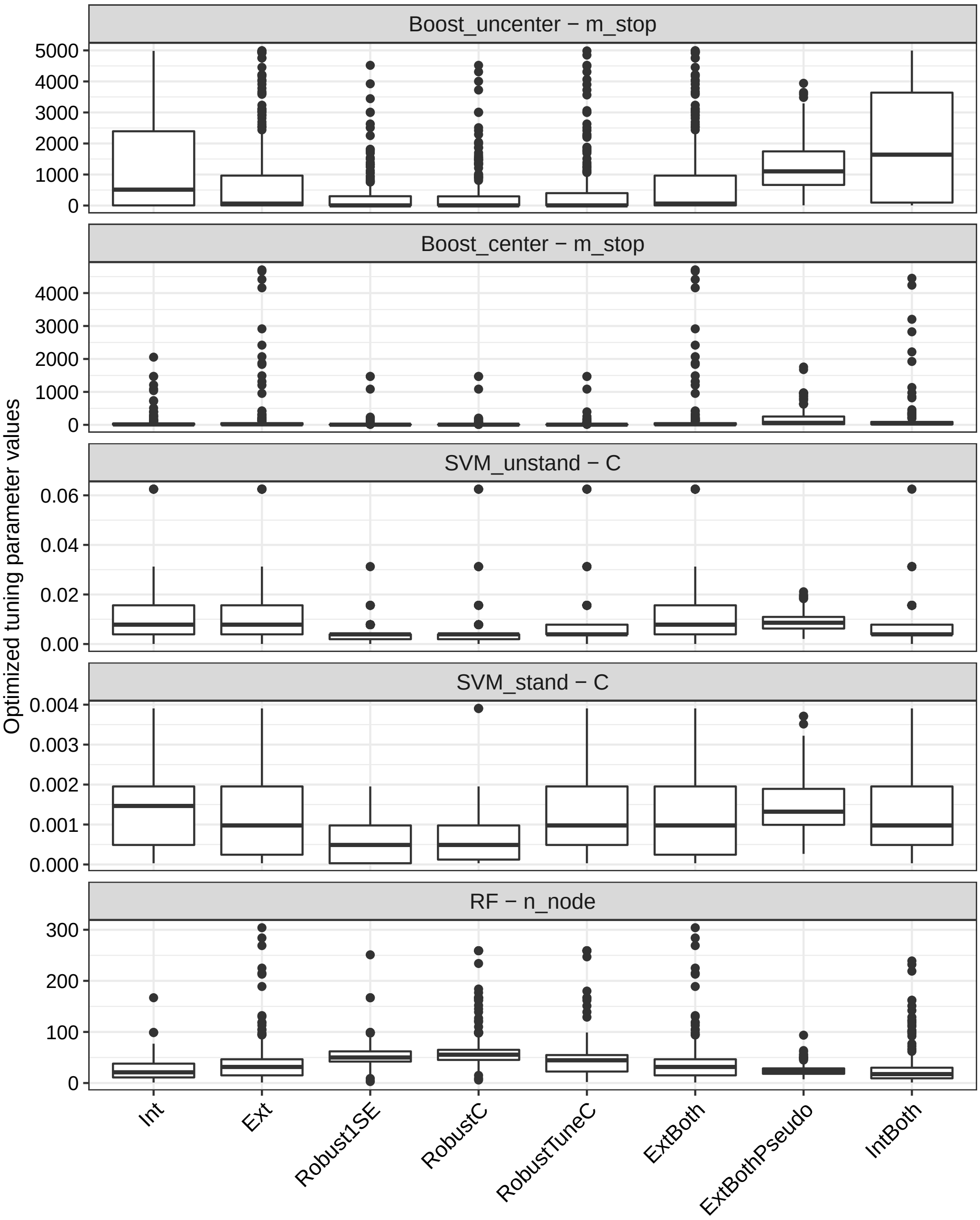

Fig. S7: Extended results of the main study. Chosen tuning parameter values for various practically motivated approaches to external, internal, and robust tuning - II

F RobustTuneC: Chosen c values

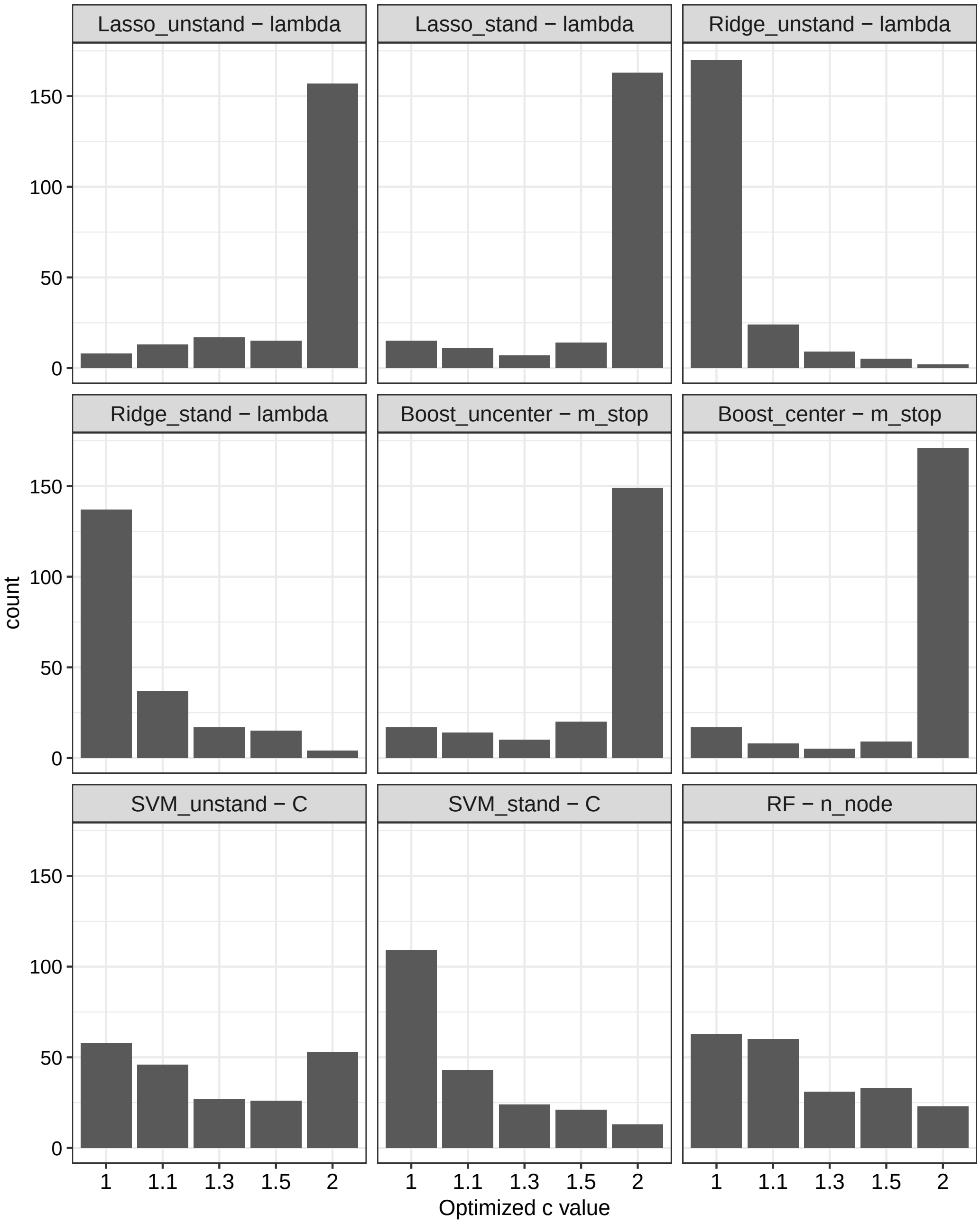

Fig. S8: Frequencies of c values chosen from the grid used for RobustTuneC

G Discussion of the results obtained with the procedures that include the external data set for training

The three approaches that include the external data set for training, `ExtBoth`, `ExtBothPseudo`, and `IntBoth` (see Section B), perform slightly better than the other approaches for some of the prediction methods (see Section D). More precisely, in cases in which the cross-study prediction performance is strong overall, these approaches do not perform better than competitors. `ExtBoth` performs slightly better than `ExtBothPseudo` for some prediction methods, while for others there are no notable differences. The latter suggests that it is not very beneficial to perform external tuning to choose the tuning parameter value and subsequently combine the training data set with the external data set for training the prediction rule. For some prediction methods `ExtBoth` outperforms `IntBoth`, while there are no systematic differences for the remaining prediction methods. `ExtBothPseudo` seems to perform better than `IntBoth` for `Boost_uncenter` and `Boost_center`. This is, however, likely due to the fact that the variance of the AUC estimates is reduced for `ExtBothPseudo`, as in the case of this approach the AUC estimates do not represent AUC values obtained from single evaluations of the test data. Instead, they are obtained by averaging the AUC values obtained from 10 iterations, where each of these corresponds to a different split into "pseudo training data set" and "pseudo external data set", see again Section B.

H Extended results of the additional study: optimistic bias by using the external data set for both tuning and prediction performance estimation

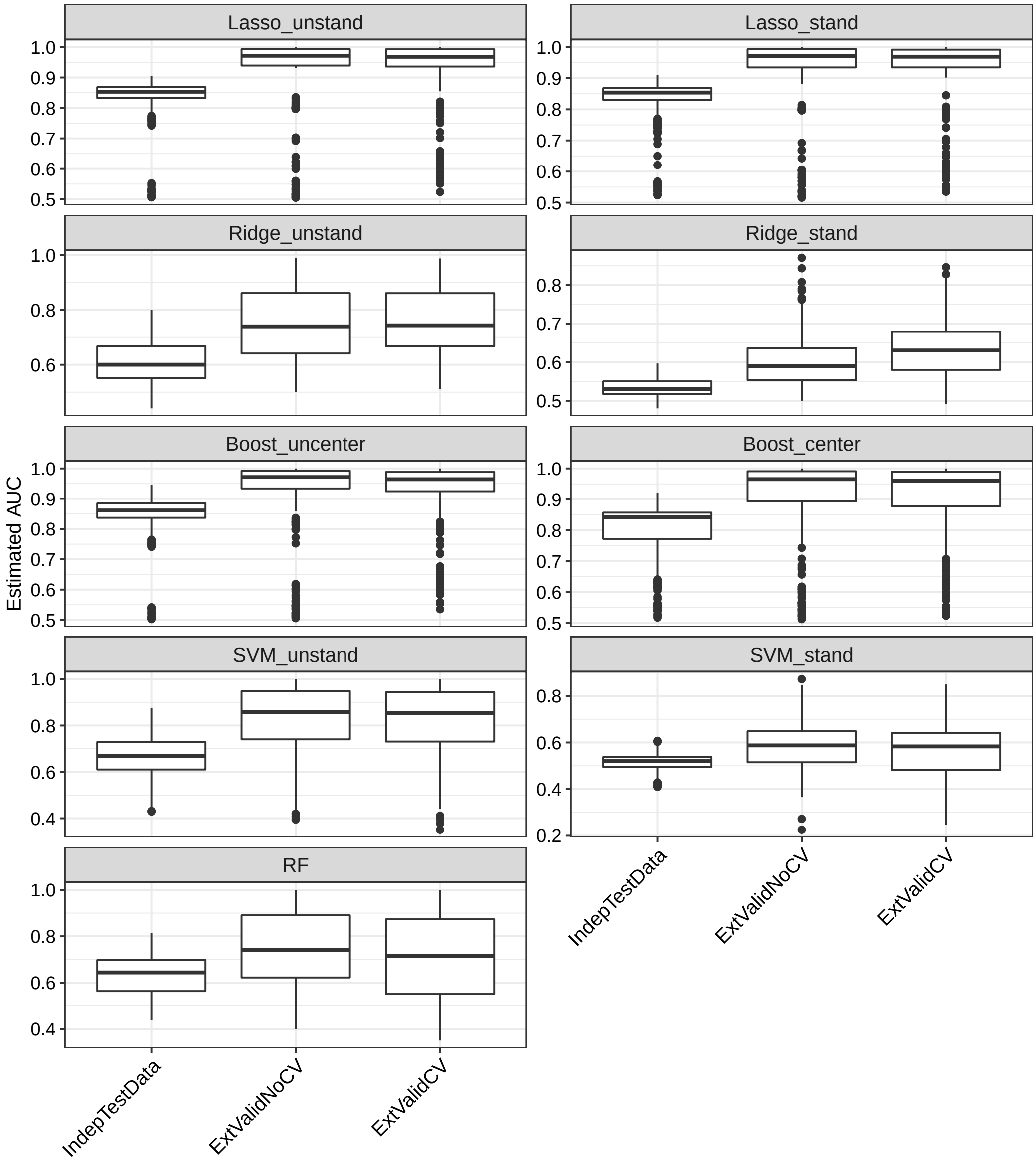

Fig. S9: Extended results of the additional study. Prediction performance estimates based on independent test data `IndepTestData` (left) compared with estimates `ExtValidNoCV` (middle) and `ExtValidCV` (right)

References

Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8:118–127, DOI 10.1093/biostatistics/kxj037, URL https://doi.org/10.1093/biostatistics/kxj037, https://academic.oup.com/biostatistics/article-pdf/8/1/118/25435561/kxj037.pdf