Beyond the Hype: A Simulation Study Evaluating the Predictive Performance of Machine Learning Models in Psychology

Kim-Laura Speck, Kristin Jankowsky, Florian Scharf, and Ulrich Schroeders

University of Kassel

Author Note

Kim-Laura Speck https://orcid.org/0000-0001-9918-4679

Kristin Jankowsky https://orcid.org/0000-0002-4847-0760

Florian Scharf https://orcid.org/0000-0003-1659-4774

Ulrich Schroeders https://orcid.org/0000-0002-5225-1122

We have no conflicts of interest to disclose. We confirm that the work adheres to the American Psychological Association's Ethical Principles of Psychologists and Code of Conduct.

This research was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG) to Kristin Jankowsky (JA 3662/1-1) and Ulrich Schroeders (SCHR 1591/4-1) as part of the Priority Program 2317 "META-REP: A Meta-scientific Programme to Analyse and Optimise Replicability in the Behavioural, Social and Cognitive Sciences".

We embrace the values of openness and transparency in science (http://www.researchtransparency.org/). Therefore, we have published all the scripts necessary to reproduce the simulated data and provide reproducible scripts for all the data analyses reported in this paper at (https://github.com/kimSpeck/MLsim). Correspondence concerning this article should be addressed to Kim-Laura Speck, Holländische Straße 36-38, 34127 Kassel, Germany, Email: kim.speck@uni-kassel.de

Abstract

Although Machine Learning (ML) methods are gaining popularity in psychological research, the debate about their usefulness ranges from hype to disillusionment. The discrepancy between the hopes placed in ML methods and the empirical reality is often attributed to the quality of psychological datasets, which tend to be small and subject to imprecise measurement. In this simulation study, we examined the data requirements necessary for ML methods to perform well. We compared the performance of Elastic Net Regressions with and without prespecified interactions, Random Forests and Gradient Boosting Machines for different data-generating processes (including either interaction, stepwise, or piecewise linear effects) and under various conditions: (a) sample size, (b) number of irrelevant predictors, (c) predictor reliability, (d) effect size, and (e) nature of the data generating model (i.e., linear vs. non-linear effects). We investigated whether the models achieved the maximum predictive performance (i.e., recovery of the systematic variance). There were two main takeaways of our results: First, the maximum possible predictive performance was only achieved under optimal conditions (N = 1,000, perfectly reliable predictors, predominantly linear effects, and an exceptionally large effect size of R^2 = .80), which are arguably rarely met in psychological research. Second, each ML model outperformed the others under certain conditions, but none was consistently superior or entirely robust to suboptimal data characteristics. We stress that data quality fundamentally limits the performance of ML models in a similar way to more traditional regression analyses and discuss the validity of comparing increasingly complex models in psychological research.

Translational Abstract

Machine Learning (ML) methods—including highly flexible models that may detect patterns in data without specifying prior assumptions—have gained considerable attention across scientific fields. They power everyday technologies like face recognition and personalized song or purchase recommendations. Encouraged by these successes, psychologists increasingly apply ML to predict constructs such as personality traits or mental-health symptoms. A common strategy is to apply ML algorithms to the kinds of datasets traditionally used in psychology and benchmark their performance against simpler linear models. Yet such comparisons often disappoint: ML rarely outperforms its conventional counterparts. Using an extensive simulation study, we compared the predictive performance of widely used ML models: regularized regression models with and without interaction terms as well as gradient-boosted trees and random forests. We varied data conditions systematically and found that the more complex, tree-based ML models only outperformed regression models under ideal circumstances: large samples, reliable measurements, very strong effects, and pronounced non-linear relationships. Under more typical psychological research conditions, the simpler regression models performed just as well—or better. Crucially, no model consistently outperformed the others across scenarios. Instead, performance depended on the degree to which the data-generating process aligned with each model's structural assumptions. Our findings suggest that researchers hoping to benefit from ML should prioritize measurement quality, sufficient sample sizes, and thoughtful variable selection. When those foundations are weak, switching to a more complex algorithm unlikely improves predictions. Careful consideration of data quality, quantity, and the underlying psychological structure is essential to choosing an appropriate analytic approach.

Keywords: machine learning, elastic net regression, random forest, gradient boosting machine, simulation study

Beyond the Hype: A Simulation Study Evaluating the Predictive Performance of Machine Learning Models in Psychology

Introduction

During the past decade, machine learning (ML) has gained popularity in many scientific disciplines. In psychology, ML models have been used to predict student retention (e.g., Matz et al., 2023), to predict Big Five personality traits based on behavioral patterns (e.g., Stachl et al., 2020), and to create personalized models of therapy outcomes (e.g., Gomez Penedo et al., 2022; Schwartz et al., 2021). The implicit hope is that ML methods can process large quantities of data without aggregation, pre-selection of predictors, or a priori specification of functional relationships between predictors and outcomes (e.g., Adjerid & Kelley, 2018). By better exploiting the available data (e.g., large scale survey data), it is anticipated that predictive performance in applied settings will improve, and that complex dependencies—hardly accessible to conventional analysis—can be revealed. Initial findings and reports have been promising, suggesting that the use of flexible tree-based methods such as *Random Forests* (RF, Breiman, 2001) or *Gradient Boosting Machines* (GBM, Friedman, 2001) can improve predictive performance.

The potential of ML methods was illustrated by several prominent studies on the prediction of suicidal behavior, which reported substantially higher predictive performance of ML models compared to more traditional approaches such as logistic regression (e.g., Fox et al., 2019; Huang et al., 2020; Walsh et al., 2018). The authors of these studies consistently interpreted their findings as strong evidence for substantial interaction effects between common risk factors for suicidal behavior. However, subsequent analyses revealed that these results were artificially inflated. For example, Jacobucci et al. (2021) demonstrated that the combination of a specific type of validation, namely optimism-corrected bootstrapping (Harrell Jr. et al., 1996), with tree-based methods led to information leakage, thereby artificially inflating performance metrics. Information leakage seems to remain a common issue across various scientific disciplines. For example, Kapoor and Narayanan (2023) identified validation errors leading to information leakage in 294 published studies. This underscores how the seeming superiority of ML methods can arise from methodological flaws, such as information leakage, rather than genuine model performance. These findings emphasize the importance of critically evaluating the assumptions, computational processes, and algorithms underlying ML methods to draw valid conclusions about their capabilities and limitations.

These critical reviews have tempered the initial over-enthusiasm for ML-based prediction, which has even been compared to "magical thinking" by some authors (Iliescu et al., 2022). The current study aims to foster more realistic expectations for ML-based prediction in psychological research by clarifying limits and possibilities through a comprehensive simulation study. Specifically, we evaluate whether psychological data, which are often multidimensional, noisy, and weakly constrained by theory, provide the necessary conditions for ML methods to reach their predictive potential. To this end, we systematically manipulate key data characteristics to mirror psychological research scenarios. In particular, we vary sample size, predictor reliability, the total amount of explained variance and the relative proportion of explained variance attributed to linear or more complex non-linear or interaction effects. A central focus is on the conditions under which ML methods are capable to capture interactions or other non-linear effects, an advantage frequently attributed to ML-based prediction over conventional linear regression models. With these manipulations, we aim to illuminate when ML methods may or may not meet expectations and evaluate the implications for predictive performance that might result from minimizing theoretical assumptions.

Between Theory and Data: The Current State of Modeling in Psychology

Psychology, as a scientific discipline, subsumes a wide range of subfields that differ substantially in the structure, the quantity, and quality of data, as well as in the extent to which theoretical assumptions guide data acquisition and modeling. In some areas, hypotheses derived from formal models of cognitive processes such as learning, decision-making, and memory are tested in a confirmatory manner (e.g., Macmillan & Creelman, 2005; Ratcliff et al., 2016; Viering & Loog, 2023). These formal, often highly parameterized models are typically implemented in within-subject experimental paradigms, where repeated assessments under controlled conditions enable precise modeling of individual cognitive dynamics. In addition, these research areas often collect high-dimensional data sets, for instance, via EEG or (f)MRI recordings in a highly controlled experimental context reducing confounding variance as much as possible. In contrast, other areas of psychology adopt a more exploratory approach, aiming to identify patterns in complex or weakly structured data without strong theoretical assumptions (Borsboom et al., 2021; Eronen & Bringmann, 2021; Wagenmakers et al., 2012). Observational data—such as questionnaire responses, digital trace data, or behavioral logs—are often used for these purposes, particularly in fields like personality psychology or

clinical psychology. In such contexts, flexible modeling techniques, including ML methods, are applied to uncover structure in the data rather than to test predefined hypotheses. The utility of ML methods may vary considerably, depending on factors such as data quantity, data quality, and the reliability of the measurement. In light of this heterogeneity in modeling approaches, the present manuscript focuses on data settings typical of exploratory psychological research and evaluates their suitability for predictive modeling using ML methods.

The emphasis on exploratory research reflects a broader challenge in psychology: many phenomena are investigated in the absence of formal theoretical models. The lack of formalized relationships between variables has led some scholars to frame the reproducibility crisis as, fundamentally, a crisis of theory (Muthukrishna & Henrich, 2019). More generally, our understanding of the underlying relationships and mechanisms within complex causal systems remains limited—for instance, regarding the factors that contribute to multifaceted behavioral outcomes such as psychopathology (Nuijten et al., 2016). Nonetheless, it is often assumed that real-world systems exhibit many small effects which accumulate, cancel out, or interact in non-linear ways (Götz et al., 2022), though this view has been challenged (see Primbs et al., 2023). Thus, in strongly data-driven research environments, which are characterized by large datasets containing heterogeneous variables and a lack of well-specified theoretical assumptions, ML methods are often seen as particularly attractive. Because ML methods do not require researchers to specify the functional form of relationships in advance, they are frequently perceived as an inductive, data-driven counterpart to explanation—or even as a straightforward solution to otherwise intractable modeling problems.[1] Despite their appeal, the application of ML methods in psychology introduces notable challenges. Unlike in computational sciences, where these methods originated, differences in data structure, measurement quality, and model implementation in psychological research can lead to misleading or biased results. It is therefore essential to critically assess the conditions under which ML methods yield valid and informative conclusions (Lavelle-Hill et al., 2025).

In psychology, many studies compare the predictive performance of increasingly complex modeling approaches, such as (regularized) linear regression and more flexible ML methods like RF. This ‘deductive data mining’ approach has been recommended for identifying non-linear or interaction effects through model comparisons (Hong et al., 2020). For instance,

[1] It has been correctly pointed out, however, that ML is rarely a purely inductive process; rather, it often operates abductively, incorporating elements of theory and prior knowledge at various stages including data acquisition, model development, and interpretation (e.g., Jacobucci et al., 2023).

if the RF outperforms (regularized) linear regression, this result is often interpreted as evidence for the presence of non-linear or higher-order interaction effects (e.g., Ali & Ang, 2022). This strategy parallels model fit comparisons common in statistical frameworks like structural equation modeling and linear regression, which have a solid epistemological foundation (Platt, 1964; Rodgers, 2010). In that sense, ML methods may play an important intermediary role between data and theory (Jacobucci et al., 2023; Yarkoni & Westfall, 2017). However, it is important to acknowledge that the potential for improvement from these efforts is ultimately limited by the characteristics of the data. Regardless of how complex an ML method is, there exists a theoretical upper bound to predictive accuracy, determined by the quality and informational content of the data being analyzed. This is a fundamental principle of estimation theory that applies to *any* statistical method (e.g., Bhattacharyya, 1946; Rao, 1992) and equally holds for modern ML methods (Diaz et al., 2022; Seroussi & Zeitouni, 2022; Shalev-Shwartz & Ben-David, 2014; Shalev-Shwartz et al., 2010). That is, even if the true underlying relationships between all variables were known, the performance of any algorithm would still be constrained by the data. In other words, ML methods *cannot* miraculously detect subtle non-linear dynamics in noisy, undersized, or unreliable data, because their performance depends on the same data characteristics that influence traditional statistical models. While hyperparameter tuning and sparsity-inducing regularization techniques (e.g., LASSO or pruning) play a key role in improving generalization and minimizing overfitting, evaluating the suitability of a dataset for non-parametric ML methods (such as GBM or RF) is crucial before comparing different ML models, as they may be outperformed by a well-specified parametric model by leveraging prior knowledge (Wolpert & Macready, 1997). Without such evaluation, poor performance of a flexible method may be misinterpreted as evidence that the underlying data structure is simple or predominantly linear—when in fact it reflects limitations in data quality, sample size, or variable selection.

In this paper, we aim to support researchers in making realistic assessments of whether their datasets may be appropriate for the application of ML methods. We focus on typical use cases in applied psychological research, examining several data conditions by manipulating the following factors: (a) sample size, (b) the number of irrelevant noise variables, (c) predictor reliability, (d) effect size (in terms of variance explained), and (e) effect composition in the data-generating process (i.e., the ratio of linear to non-linear effects). Understanding how these data characteristics influence predictive performance of ML models is critical to ensure

robust, reproducible analyses and generalizable findings. In the following sections, we discuss each of these factors in more detail.

Sample Size

In computational sciences, datasets are typically large, with sample sizes often exceeding 100,000 observations and thousands of variables, primarily collected through automated electronic processes. These vast datasets allow for a data-driven detection of complex patterns and the development of highly accurate predictive models, benefiting from the sheer volume and heterogeneity of data points. Conversely, psychological research often relies on smaller samples: The median sample size in leading personality and social science journals in 2011-2019 ranged between 113 and 330 (Fraley et al., 2022). In clinical psychology, the average sample size is even less favorable. For medical trials, a review of the Cochrane Database of Systematic Reviews (Davey et al., 2011) including nearly 3,000 meta-analyses on mental health found that the median sample size of randomized controlled trials was 63 (with the 25th and 75th percentiles at 36 and 165, respectively). Such small sample sizes pose challenges for uncovering meaningful patterns, enabling satisfying predictive accuracy, and generalizing findings (e.g., Maxwell, 2000, for multiple regression).

Increasing sample size is often recommended to reduce overfitting, where models erroneously treat sample-specific noise as meaningful signal (Anderson & Burnham, 2002). For a given model specification and fixed tuning parameters, larger datasets improve generalization by reducing variance: models become less sensitive to idiosyncrasies in the training data, while bias remains largely unchanged (see Pargent et al., 2023; Yarkoni & Westfall, 2017, for easily accessible illustrations). However, model complexity in ML modeling is not fixed across sample sizes. Rather, many ML methods rely on cross-validation in combination with some form of regularization to adapt model complexity to the dataset at hand. Consequently, smaller training sets should result in more strongly regularized (i.e., simpler and more biased) models, as this configuration better balances the bias-variance trade-off under data constraints. As sample size increases, more complex models can be selected without substantially increasing variance. This reduces bias and improves predictive performance (e.g., Hastie et al., 2009). While larger sample sizes generally contribute to better generalization under fixed model complexity, adaptively allowing for more complex models in larger datasets may drastically increase sample size requirements. Thus, for more flexible approaches such as non-parametric RFs, larger datasets are essential to realize their

full predictive potential – and intuitions for what is a 'large' sample based on experience with conventional methods may not hold for these non-parametric ML methods.

Number of Noise Variables

The impact of noise variables on model performance depends on the relationship between the total number of predictors (P) and the number of available observations (N). A high proportion of noise variables inflates P without contributing meaningful information, increasing the risk of overfitting. When the $N : P$ ratio is low, due to either small sample sizes or an abundance of predictors, the model is more likely to capture noise rather than the true underlying patterns in the data (Guyon & Elisseeff, 2003). This problem is even more pronounced in more complex ML methods, where the large number of predictors increases the risk of overfitting spurious non-linear or higher-order interaction patterns, ultimately impairing predictive performance.

Appropriate regularization via parameter constraints and/or parameter selection can mitigate the impact of noise variables and ensure that the model focuses on the most important effects: For example, Kuhn and Johnson (2013) showed in a small simulation study that the effect of adding irrelevant noise predictors varies between ML methods. Whereas the predictive performance of non-regularized regressions, neural networks, and support vector machines decreased the most when irrelevant noise variables were added, the performances of Elastic Net regressions (ENET) and various tree-based methods were relatively stable irrespective of the number of noise variables. However, for data with much more predictors than observations, that is a highly unfavorable $N : P$ ratio, overfitting may occur despite the use of regularization techniques (e.g Riley et al., 2019a, 2019b).

Reliability of Predictors

Psychological data often rely on self-reports and clinical assessments, which tend to be less reliable than the data typically used in ML models (e.g., song streaming habits or GPS data; Anderson et al., 2021; Wang, 2019). While the median Cronbach's alpha for scales used in articles published in various APA journals is reported to be $\alpha = .85$ in a large meta-science study (Hussey et al., 2023), this figure may not reflect the reliability of data used in ML models for several reasons. First, the reported median reliability usually applies to well-established measures, thus representing an upper bound that is unlikely to be reached by most variables used in predictive modeling. Second, Cronbach's alpha is computed on item sets (scales),

which are more reliable than their indicators. However, often ML models incorporate all available items from psychometric inventories for prediction instead of the more reliable aggregate scores (e.g., McClure et al., 2024; Schroeders et al., 2021; Seeboth & Mõttus, 2018).

The influence of reliability on parameter estimation, particularly in linear regression, is well-established. Specifically, measurement error in predictor variables can negatively impact the predictive accuracy by attenuating the relationship between predictors and outcome. This is commonly referred to as *attenuation bias* and arises from the limitation of the maximal observable correlation between variables, which is constrained by their reliabilities (e.g., Spearman, 1904). It has been demonstrated that measurement error also affects regularized regression models, such as LASSO regression (Sørensen et al., 2015). Moreover, the product of the reliabilities of two predictors determines the lower bound of the reliability of their interaction, thereby exacerbating the issue when dealing with interaction effects (e.g., Jaccard & Wan, 1995; Tosteson et al., 2003). In the context of complex ML algorithms, this challenge becomes even more pronounced when predicting outcomes based on small interaction effects (see section on effect composition), because measurement error and the complexity of modeling interactions can substantially degrade predictive performance.

Fittingly, Jacobucci and Grimm (2020) demonstrated in a simulation study how measurement error affects the ability of GBMs to accurately model a combination of non-linear trigonometric and interaction effects. The authors varied the reliability of the predictor variables (with the levels of .30, .60, and .90) and found that a linear model outperformed the GBM in the condition with the highest measurement error. Even with a reliability index of .60, the GBM did not consistently perform better than the linear model. Given these counter-intuitive findings, we deem it worthwhile to expand these simulation to more realistic data conditions encountered in psychological research to better understand the impact of measurement error on ML predictions.

Overall Effect Size

The overall effect size affects how well the model can capture and leverage the true signal in the data, with stronger effects generally enabling better predictive performance, as they are easier discernible among the noise (Breiman, 2001). A meta-analysis on correlations from the field of individual differences estimated the 25th, 50th, and 75th percentiles to be .11, .19, and .29, respectively (Gignac & Szodorai, 2016). A more recent analysis including

research in social and differential psychology found slightly higher correlations of .12, .24, and .41, respectively (Lovakov & Agadullina, 2021). Overall, effect sizes commonly considered large ($r \geq .5$) are seldom found in observational (non-experimental) studies.

Effect Composition

Theoretically, datasets may contain linear and non-linear (including interaction) effects or a combination of both. However, it is worth noting that interaction effects found in psychological research have typically been small to moderate (e.g., Aguinis et al., 2005; Beck & Jackson, 2022; Sommet et al., 2023; Vize et al., 2023), suggesting that correctly depicting these interactions can be challenging. As a result, there are very few cases of psychological studies where more complex ML models have demonstrated higher predictive accuracy than linear regressions. On a broader level, the ratio of linear to non-linear effects determines the complexity of a model. When non-linear effects such as interactions constitute a significant proportion of the data, more sophisticated models capable of capturing these patterns may play out their strengths (Kuhn & Johnson, 2013).

In sum, these five factors—(a) sample size, (b) number of noise variables, (c) predictor reliability, (d) effect size, and (e) effect composition—are all presumed to contribute, to varying degrees, to the potential of ML models to achieve high predictive performance and to obtain reliable, robust results. While the adverse effects of these factors are well-documented, questions remain regarding the minimal requirements, their interactions, and the extent to which ML methods can handle suboptimal conditions in realistic psychological datasets. For example, a larger sample size may partially offset the negative impact of noise variables or enable researchers to detect smaller non-linear effects. Similarly, the detrimental effect of unreliable predictor variables may depend on their effect sizes. Evaluating these interdependencies can guide researchers in optimizing study designs, particularly when it is not feasible to maximize all factors simultaneously.

About Complex Models and Basic Realities

Various ML methods have been used to predict psychological outcomes, but a common approach is to compare the performance of a baseline linear regression model like Elastic Net Regression (ENET) with a more flexible algorithm like tree-based methods, which can in principle handle more complex relationships. The idea behind the comparison of ML models is to quantify the increment in predictive validity due to modeling more complex relationships

(e.g., Hong et al., 2020; Jankowsky & Schroeders, 2022). Both modeling approaches are popular and offer specific advantages. For example, unlike classical linear regression models, ENET can handle high-dimensional data (i.e., a large number of predictors) and perform feature selection from a large pool of candidate predictors. The strength of tree-based methods is to model complex interactions and non-linear associations without the need for explicitly specifying all effects a priori. In principle, ENET can also include interactions or higher-order effects (e.g., polynomials), but these effects need to be explicitly specified in advance. Essentially, ENET is a parametric method (e.g., Berk et al., 2008), in contrast to tree-based methods, which are non-parametric regression models. In our simulation, we will compare ENET and tree-based methods as representative methods of their respective categories and introduce their foundational concepts in the following sections.

Elastic Net Regression

ENET applies regularization to the estimation of the regression coefficients in a generalized linear model (e.g., linear regression, logistic regression, Poisson regression) to cope with challenges occurring in the context of datasets with a high number of predictors (P) and relatively small sample sizes (N). In case of a continuous outcome, the underlying model is the standard linear regression model denoted as:

$$y = \hat{y} + e = b_0 + Xb + e \tag{1}$$

Here, y is an $N \times 1$ vector of the observed outcome values, $\hat{y}$ is an $N \times 1$ vector of the *predicted* outcome values, and e is an $N \times 1$ vector of residuals computed from the difference between observed and predicted values. In linear regression models, the predicted values are computed from a linear combination of the predictors denoted as Xb, where X is the $N \times P$ *design matrix* and b is a $P \times 1$ vector containing the regression weights $b_1, ..., b_P$. The intercept b_0 is denoted separately because it is typically not regularized. As in classic linear regression models, interaction terms and more complex non-linear terms can be added as additional predictors. For the subsequent presentation, it is important to note that X can contain original variables and their (non-)linear transformations.

Traditionally, linear regression models are fit using the ordinary least squares (OLS) criterion, which can be expressed as:

$$F_{OLS} = \min_{b} \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 = \min_{b} \sum_{n=1}^{N} \left(y_n - \left(b_0 + \sum_{p=1}^{P} x_{np} b_p \right) \right)^2 \tag{2}$$

Regularized estimation of regression models have been proposed to address known limitations of OLS regressions (e.g., overfitting, unstable coefficient estimates, or computational issues with high levels of multicollinearity; Helwig, 2017) by optimizing a modified fitting criterion that adds a penalty term to the OLS criterion:

$$F_{regularized} = F_{OLS} + \lambda \cdot \text{Penalty}(b) \tag{3}$$

Penalty(b) is a function that increases with the magnitude of the regression coefficients (except for the intercept b_0), and λ is a hyperparameter that controls the strength of the penalty term. When optimizing this modified fitting criterion, a balance must be found between approximating the data closely (i.e., minimizing F_{OLS}) and the penalty for large regression weights. The purpose of this approach is to stabilize the estimates of the regression coefficients by exploiting the *bias-variance trade-off*. By introducing a small bias to the parameter estimates, the standard errors can be reduced, leading to estimates that are on average closer to their true values and reducing overfitting (e.g., Helwig, 2017).

Two prominent choices for the penalty function are ridge regression (also known as L_2 regularization, Hoerl and Kennard, 1970) and LASSO (least absolute shrinkage and selection operator) regression (also known as L_1 regularization, Tibshirani, 1996).

$$\text{Penalty}_{\text{Ridge}}(b) = \sum_{p=1}^{P} b_p^2 \tag{4}$$

$$\text{Penalty}_{\text{LASSO}}(b) = \sum_{p=1}^{P} |b_p| \tag{5}$$

Ridge regression tends to shrink all coefficients within a group of correlated predictors toward similar values—a characteristic known as the grouping property (Friedman et al., 2010). Unlike LASSO regression, ridge does not set any coefficients exactly to zero, thus retaining all predictors in the model, even if some are less relevant. In contrast, LASSO regression not only shrinks coefficients but also performs variable selection by setting some coefficients exactly to zero, resulting in a sparser model (Tibshirani, 1996, 2011). This

combination of continuous shrinkage and automatic variable selection makes LASSO particularly useful for identifying the most relevant predictors in high-dimensional data.

LASSO regression may offer inferior predictive performance compared to ridge regression in case of multicollinearity, as LASSO tends to handle highly correlated predictors by selecting one and shrinking the others to zero (Friedman et al., 2010). However, in scenarios with many irrelevant predictors, LASSO can outperform ridge due to its feature selection capability. The *Elastic Net* penalty function combines both the L_1 and L_2 regularization terms (Friedman et al., 2010; Zou & Hastie, 2005) and is defined as:

$$\text{Penalty}_{\text{Elastic Net}}(b) = \alpha \sum_{p=1}^{P} |b_p| + (1 - \alpha) \sum_{p=1}^{P} b_p^2 \tag{6}$$

Here, α is a mixing parameter that controls the balance between the L_1 and L_2 regularization terms. By incorporating both LASSO and ridge penalties, ENET aims to address multicollinearity and perform feature selection, potentially enhancing predictive performance. Notably, two tuning parameters must be determined for ENET regularization: λ, the regularization parameter controlling the overall strength of the penalty, and α, the mixing parameter. Both hyperparameters are typically selected via cross-validation to minimize prediction error.

Tree-based methods

Whereas parametric modeling approaches such as ENET, require specification of the functional form, non-parametric methods, such as the tree-based random forests (RFs) and gradient boosting machines (GBMs), aim to approximate the outcome as an arbitrary function of the predictor values (Berk et al., 2008). This relationship can be expressed as:

$$y = \hat{y} + e = f(X) + e \tag{7}$$

There are various methodologies based on different underlying principles to achieve this goal. RFs and GBMs combine multiple regression trees to produce robust approximations of the outcome. The core idea behind a single regression tree is to recursively partition the data into smaller segments, typically using binary splitting criteria. For instance, the data might first be split into two segments based on whether x_1 falls below a specific threshold; further splits

are then recursively applied within each resulting segment. The final segments, which are not split further, are typically called *end nodes*. Within each end node, a constant predicted outcome value is assigned. Both the splitting variables and the splitting rules are chosen to minimize an overall loss function. In the classification and regression trees (CART) algorithm by Breiman (2017), this is done by minimizing the squared loss function (L_2 loss):

$$L_2 = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 \tag{8}$$

CART optimizes the same fitting criterion as linear regression but generates predictions based on splitting rules rather than a parametric mathematical function. With L_2-loss, commonly applied for continuous outcomes, the predicted value for each observation is the mean of the outcome variable within each end node. This characteristic makes CART both flexible and versatile, as it performs predictor selection and functional approximation in a purely data-driven manner. However, single trees face several significant limitations: (1) they are unstable, meaning small changes in the data (e.g., removing an observation) can lead to completely different tree structures (e.g., Hastie et al., 2009; Strobl et al., 2009), (2) they struggle to approximate smooth functions, (3) they rely on a greedy approach for splitting nodes, considering only the best immediate split based on the splitting criterion, and (4) they are sensitive to irrelevant features, which can lead to splits driven by noise rather than meaningful variables, thereby reducing interpretability and predictive performance. Both RFs and GBMs are ensemble methods specifically designed to aggregate the predictions of multiple individual trees, thereby mitigating these limitations.

Random Forests (RFs)

Random forests address the limitations of individual regression trees by constructing an ensemble of trees. Each tree is trained on a bootstrap sample of the original data, and their predictions are aggregated by averaging (Breiman, 2001), which stabilizes predictions. In addition to training each tree on a different bootstrap sample, RFs select a random subset of candidate predictors at each node when determining the best split. This procedure allows the ensemble to explore more diverse splitting paths and helps to mitigate the greedy nature of the splitting process in individual trees. Averaging the predictions across the ensemble substantially reduces variance and results in a prediction function that can better approximate

smooth relationships compared to single regression trees, which typically struggle with such tasks.

The procedure for fitting a RF based on CART regression trees with M trees can be summarized as follows:

1. Draw M bootstrap samples of size N from the original dataset (sampling with replacement).
2. For each bootstrap sample $m = 1, \dots, M$:
   - Fit a regression tree T_m by recursively partitioning the data to minimize the squared error between observed and predicted values similar to a single CART-Tree.
   - At each split, select the best split from a randomly chosen subset of p predictors (with $p < P$, where P is the total number of predictors).
   - The tree T_m is obtained by solving:

$$T_m = \arg\min_{T} \sum_{n=1}^{N^{(m)}} \left(y_n^{(m)} - f_m(x_n; T) \right)^2$$

     Here, the predicted value $\hat{y}_n^{(m)} = f_m(x_n; T)$ for a given input x_n is the mean outcome of all training observations falling into the same end node as x_n.
   - Store the resulting tree T_m, defined by its sequence of splitting rules and the predicted values assigned to each end node.
3. Compute the final prediction for each observation $n = 1, \dots, N$ by averaging the predictions for the predictor value combination x_n across all M trees:

$$\hat{y}_n = \frac{1}{M} \sum_{m=1}^{M} f_m(x_n; T_m)$$

The performance of a RF depends on the appropriate choice of several hyperparameters, which are typically selected via cross-validation. In particular, two key parameters are (1) the number of candidate predictors randomly considered at each split—commonly set to values such as $\sqrt{P}$ or $\frac{P}{3}$ for classification and regression, respectively (Breiman, 2001)—and (2) the minimum node size, that is, the minimum number of observations required in an end node before tree growth is halted. It is worth noting that the

number of trees M is usually *not* selected via cross-validation, as it does not directly affect the average predictions but rather their variance. A sufficiently large M is essential to ensure stable predictions; however, increasing M beyond this point only raises computational cost without further improving predictive performance (Efron & Hastie, 2021, Chapter 17.1, p. 328).

Gradient Boosting Machines (GBMs)

In contrast to RFs, GBMs do not rely on bootstrapping. Instead, GBMs combine multiple regression trees by successively fitting trees to the residuals of the ensemble of prior trees. From a regression perspective, a single tree typically underfits the true relationship between the predictors and the outcome. By fitting additional trees to the residuals (i.e., the part of the outcome that remains unexplained by the previous trees' predictions), the overall explained variance can be gradually improved. Assuming the dependent variable is continuous and a squared loss is applied, the process of fitting a GBM with $M = 2$ trees can be described as follows:

1. Initially, predict the mean of the outcome for each observation.

$$\hat{y}_n^{(0)} = \bar{y} \tag{9}$$

2. Fit the first tree ($m = 1$). The residuals for the first iteration $e_n^{(1)}$ are calculated as the difference between the true values y_n and the initial predictions $\hat{y}_n^{(0)}$. Next, the first tree, T_1, is fit to minimize the squared error between the residuals, $e_n^{(1)}$, and the predictions of the tree, $f_1(x_n; T)$. The updated predictions, $\hat{y}_n^{(1)}$, are then calculated by adding the new tree's predictions to the initial predictions, $\hat{y}_n^{(0)}$. Importantly, the new predictions are weighted by the shrinkage parameter, ν, which has a value between 0 and 1 (typically closer to 0), effectively shrinking the updated predictions toward the previous predictions.

$$\text{Compute the residuals:} \quad e_n^{(1)} = y_n - \hat{y}_n^{(0)} \tag{10}$$

$$\text{Determine tree parameters:} \quad T_1 = \arg\min_T \sum_{n=1}^{N} \left(e_n^{(1)} - f_1(x_n; T) \right)^2 \tag{11}$$

$$\text{Update the predictions:} \quad \hat{y}_n^{(1)} = \hat{y}_n^{(0)} + \nu \cdot f_1(x_n; T_1) \tag{12}$$

3. The individual steps described in step 2 are then repeated for $m = 2$ and, if $m > 2$, subsequent iterations. In each iteration, the residuals are recalculated, and the tree parameters are updated. The objective of each new boosting iteration is to fit the learner to the residuals, further reducing the remaining error.

$$\text{Compute the residuals:} \quad e_n^{(2)} = y_n - \hat{y}_n^{(1)} \tag{13}$$

$$\text{Determine tree parameters:} \quad T_2 = \arg\min_T \sum_{n=1}^{N} \left(e_n^{(2)} - f_2(x_n; T) \right)^2 \tag{14}$$

$$\text{Compute final predictions:} \quad \hat{y}_n^{(2)} = \hat{y}_n^{(1)} + \nu \cdot f_2(x_n; T_2) \tag{15}$$

After refining the overall predictions successively over the chosen number of iterations, the predictions from the final iteration serve as the model's final predicted outcome values. In each iteration, the described shrinkage serves to avoid overfitting. The parameter ν is usually chosen through cross-validation, analogous to how λ is chosen in regularized regressions. Additional hyperparameters of GBMs that require tuning include the maximum number of trees M, the maximum depth of each individual tree d, and the minimum number of observations within the end nodes. With adequate hyperparameter tuning, GBMs can improve predictions over single trees in terms of both accuracy (i.e., reduced bias) and stability (i.e., reduced variance). In contrast, other tree-based ensemble methods, such as RFs, primarily improve the stability of the predictions (Friedman, 2001).

The Present Study

Insights into how ML methods work, their strengths and limitations, are often scattered across highly technical sources (Efron & Hastie, 2021; Hastie et al., 2009) and not readily accessible to substantive researchers in psychology. One of our central aims, therefore, was to bridge this gap between the methodological literature and applications in psychological research, and to promote a better understanding of the conditions required for robust

prediction—particularly in a field where enthusiasm for ML is growing rapidly. Specifically, we aim to explore the requirements psychological data must meet to enable researchers to achieve a high predictive accuracy. The outcome of interest is the predictive performance of parametric ENET regressions—both with and without interaction effects ($ENET_{int}$ and $ENET_{lin}$, respectively)—and non-parametric tree-based methods (GBMs and RFs) when tested using unseen data randomly sampled from the same population. This performance is evaluated by comparing the achieved R^2 in the testing data with the true simulated R^2. A particular focus is on the interplay of data conditions and their role in achieving high predictive performance. Therefore, we systematically vary sample size, number of irrelevant noise variables, reliability of indicators, the overall effect size and the ratio of linear to interaction or non-linear effects. We also vary the type of "more complex" effects by including three data-generating processes that each combined linear predictors with distinct forms of additional non-linear associations: (1) two-way interaction effects, (2) piecewise linear effects, and (3) dichotomous variables resembling step functions. Doing so allows us to investigate whether and how the predictive performance of different models depends on the functional form of the underlying signal.

Method

Simulation Design

We used a Monte Carlo simulation in which we simulated data based on specific regression models and evaluated the performance of ENET, RF, and GBM across multiple data generating processes (DGPs). We included three DGPs in which the explained variance was split between linear effects and additional non-linear effects. These additional effects were: (1) interactions among linear predictors, (2) piecewise linear effects, and (3) independent dichotomous variables resembling step functions. We also manipulated five key characteristics of the data: First, sample size was varied across three levels, with datasets containing 100, 300, or 1,000 observations, to evaluate how the number of observations influences model performance. These levels reflect typical sample sizes in clinical or personality psychology, with 1,000 observations representing a more favorable scenario, occasionally achieved in panel studies. Second, the number of noise variables was manipulated at two levels (10 and 50) to examine how the presence of irrelevant variables affects the models' ability to differentiate signal from noise and prevent overfitting. Third, reliability was set at three levels (0.6, 0.8, and 1.0) to explore the effects of varying data quality and measurement error on predictive accuracy. Fourth, the proportion of explained variance (R^2) was examined at three

Table 1
Summary of Simulation Parameters

Parameter		Levels (Labels)
Sample size (N)	(3)	100, 300, 1,000
Number of noise variables (# noise)	(2)	10, 50
Reliability (Rel.)	(3)	0.6, 0.8, 1.0
R^2_{sim}	(3)	0.2, 0.5, 0.8
Effect composition (linear vs. other)	(3)	20:80, 50:50, 80:20
Simulated effects		Types of Association
DGPs (linear + ...)	(3)	continuous interactions, piecewise linear, stepwise

levels (0.2, 0.5, and 0.8), reflecting varying strengths of the underlying signal in the data and ranging from realistically small to idealistically large effect sizes. Finally, the effect composition varied in the proportion of total effects between linear and other effects (i.e., continuous interactions, piecewise linear, or stepwise associations, respectively). Specifically, one simulation condition allocated 80% of R^2 to linear effects (and 20% to other effects), another condition split the variance evenly (50:50), and the last condition emphasized interactions, piecewise linear, or stepwise effects (20:80), respectively. Manipulating the effect composition allowed us to compare how ML models perform with different types of association dominating the underlying data. An overview of all the simulated conditions is provided in Table 1.

Data Generation and Simulation Procedure

All predictor and noise variables were randomly sampled from a multivariate normal distribution (e.g., Anderson et al., 1958) with each variable having zero mean. The linear predictor variables were intercorrelated with r = .4, while the noise variables were uncorrelated both with each other and with the predictor variables. The dependent variable (Y) was simulated according to a regression model including linear (main) effects and either interaction effects, piecewise linear, or stepwise associations with the outcome. An overview of all DGPs is provided in Figure 1. For interaction data, we generated the dependent variable according to the following DGP with $\beta_1 = \beta_2 = \beta_3 = \beta_4$ and $\beta_5 = \beta_6 = \beta_7 = \beta_8$.

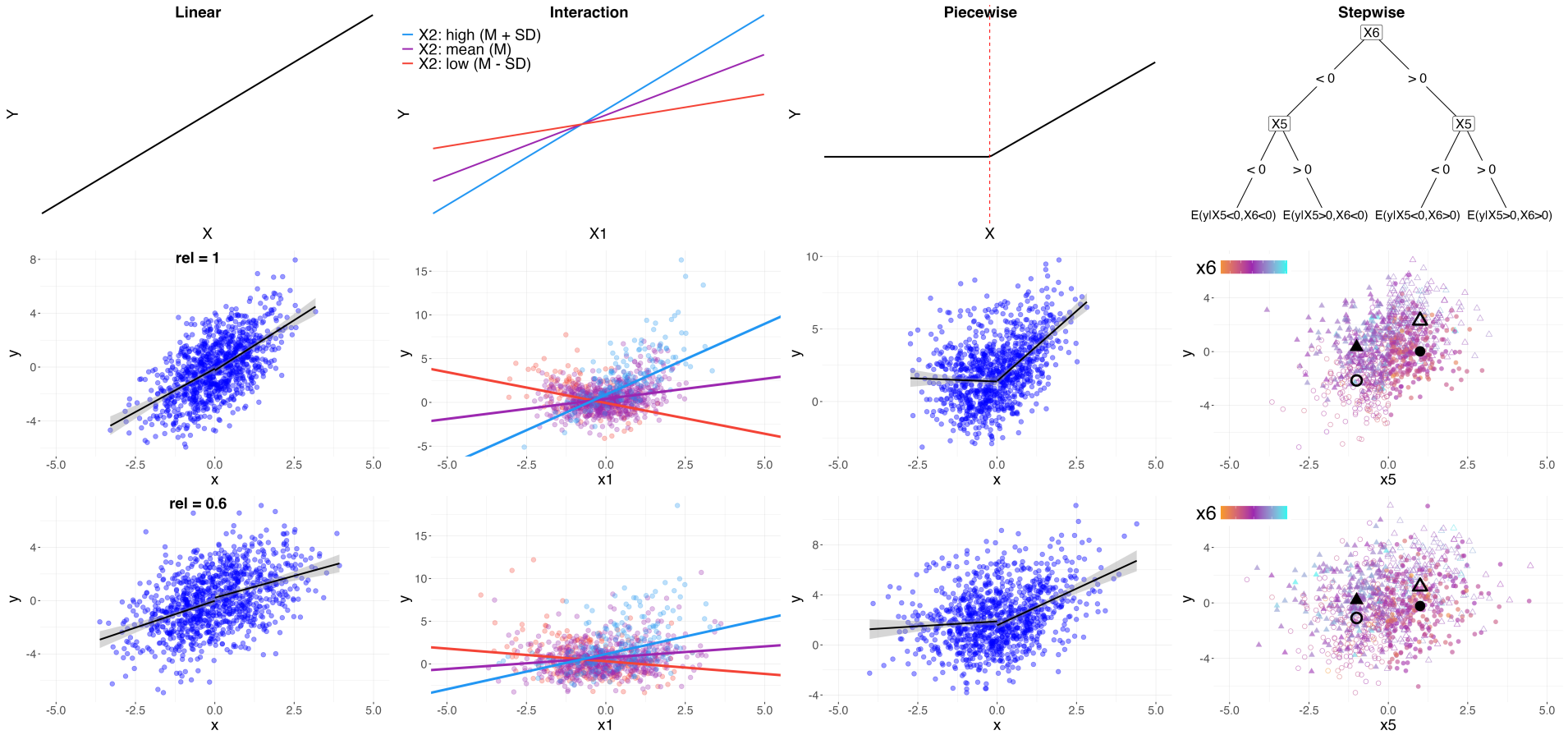
$$
\begin{aligned}
y_i =& \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \\
& \beta_5 x_{1i} x_{2i} + \beta_6 x_{1i} x_{4i} + \beta_7 x_{2i} x_{3i} + \beta_8 x_{3i} x_{4i} + \varepsilon_i \\
& \text{with} \quad \varepsilon_i \sim \mathcal{N}(0, 1)
\end{aligned}
$$

To assess how different ML methods perform under varying data, we added two non-linear DGPs alongside the interactions DGP. While the interactions DGP aligns with the assumptions of ENET, the non-linear DGPs were designed to test the capacity of the tree-based methods (i.e., RF and GBM) to capture more complex associations. In the stepwise DGP, some predictors had discontinuous, stepwise effects that align closely with the recursive partitioning approach used in regression trees. The piecewise linear DGP was inspired by a recent paper on common non-linear effects in psychological data (Simonsohn, 2024), and included predictors with segmented linear effects rather than continuous interactions. For both non-linear DGPs, we sampled variables from the multivariate normal distribution beyond the existing linear predictor variables and the noise variables[2]. Additional variables were uncorrelated with one another and the existing variables.

[2] Because the non-linear DGPs require additional variables for stepwise and piecewise linear effects (whereas the interactions DGP uses existing linear predictors), the total number of predictors is higher in the non-linear DGPs. We kept the number of noise variables constant rather than fixing the total predictor count.

Figure 1

Illustration of Data Generating Processes (DGPs).



Note. All DGPs split the explained variance (R^2_{sim}) between the linear association (first column) and one other non-linear association type (second to fourth columns) depending on the effect composition. The top row illustrates each association schematically; the middle and bottom rows show simulated data ($R^2_{sim} = .8$ with 80% of the variance allocated to the respective association). Data in the middle row illustrates perfect reliability (reliability = 1), whereas data in the bottom row includes measurement error (reliability = .6).

For the stepwise DGP, we first transformed three additional continuous predictor variables x_{pi} into dichotomous variables d_{pi} via median splits ($\psi = \mu = 0$) and subsequently applied effect-coding so that each dichotomous variable had unit variance. Consequently, the model-intercept remained fixed at zero, the Population Grand average of the outcome. For the stepwise DGP, the dependent variable was simulated according to the following equation with $\beta_1 = \beta_2 = \beta_3 = \beta_4$ and $\beta_5 = \beta_6 = \beta_7$.

$$
\begin{aligned}
y_i = &\beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \\
&\beta_5 d_{5i} + \beta_6 d_{6i} + \beta_7 d_{7i} + \varepsilon_i \\
&\text{with} \quad \varepsilon_i \sim \mathcal{N}(0, 1) \\
&\text{with} \quad d_{pi} = \begin{cases} -1, & \text{if } x_{pi} \leq \psi \\ +1, & \text{if } x_{pi} > \psi \end{cases}
\end{aligned}
$$

For the piecewise linear DGP, again, we transformed three additional continuous predictor variables x_{pi} into dichotomous variables d_{pi} via median splits ($\psi = \mu = 0$). We generated the dependent variable using a piecewise linear model with the breakpoint set at $\psi = 0$. Specifically, slopes in the first segment ($x_{pi} \leq \psi$) were fixed to zero (i.e., $\beta_5 = \beta_6 = \beta_7 = 0$), while in the second segment ($x_{pi} > \psi$), the slopes were β_5^*, β_6^* and β_7^* for the respective variables[3]. Therefore, each predictor had a flat (zero) slope below values of ψ, and a nonzero slope above that point. Formally, the DGP was defined as:

[3] A practical advantage of this approach is that the transformation can be conceived as a ReLU transformation applied to the predictor variable, because $((x_{pi} - \psi) \cdot d_{pi} = \max(0, x_{pi} - \psi))$. The resulting transformed variable follows a so-called *Rectified Gaussian* if the original variable was normally distributed, and the reduced variance of the resulting variable can be computed as $var((x_{pi} - \psi) \cdot d_{pi}) = 0.34$ (Beauchamp, 2018).

$$
\begin{aligned}
y_i = &\beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \\
&\beta_5 x_{5i} + \beta_5^* \cdot (x_{5i} - \psi) \cdot d_{5i} + \\
&\beta_6 x_{6i} + \beta_6^* \cdot (x_{6i} - \psi) \cdot d_{6i} + \\
&\beta_7 x_{7i} + \beta_7^* \cdot (x_{7i} - \psi) \cdot d_{7i} + \varepsilon_i \\
&\text{with} \quad \varepsilon_i \sim \mathcal{N}(0, 1) \\
&\text{with} \quad d_{pi} = \begin{cases} 0, & \text{if } x_{pi} \leq \psi \\ 1, & \text{if } x_{pi} > \psi \end{cases}
\end{aligned}
$$

with $\beta_1 = \beta_2 = \beta_3 = \beta_4$, $\beta_5^* = \beta_6^* = \beta_7^*$ and $\beta_5 = \beta_6 = \beta_7 = 0$.

To balance the contribution of linear effects and the other types of association with the outcome according to the specific simulation condition, we derived the regression coefficients through a numerical optimization procedure so that a certain value of R^2_{sim} was achieved for the DGP (see the additional online material at https://github.com/kimSpeck/MLsim/tree/main/onlineMaterial). The residual variance of the outcome was fixed at 1. Note that the outcome variable was simulated based on the set of predictor variables (including interactions and other effects) *without* measurement error. Subsequently, we added measurement error for each predictor or noise variable independently ($x_{\text{with measurement error}} = x_{original} + e_{measurement}$). The measurement error terms were drawn from independent normal distributions with zero mean and variance corresponding to the simulated reliability (i.e., $e_{measurement} \sim \mathcal{N}(0, \frac{1-REL}{REL})$). The (original) predictor's reliability was fixed across all DGPs. However, interaction terms had lower reliability, because combining two correlated variables that each have nonzero measurement error inevitably produces a product term with reduced reliability (Busemeyer & Jones, 1983).

Simulations and analyses were conducted using *R* (Version 4.3.3; R Core Team, 2023). For every simulation condition, 1,000 random training samples were drawn from a multivariate normal distribution (e.g., Anderson et al., 1958) using the *mvtnorm* package (Version 1.2-4; Genz et al., 2021). Additionally, one test sample (N = 1,000) was randomly sampled for every simulated condition.

Analysis Models and Hyperparameter Tuning

For each simulated sample, four models were estimated: (1) an ENET regression using only the original predictor and noise variables to predict Y ($ENET_{lin}$), (2) an ENET regression using the original predictor variables and noise variables as well as all their possible two-way interactions to predict Y ($ENET_{int}$), (3) a GBM using all predictor and noise variables to predict Y and (4) a RF using all predictor and noise variables to predict Y. Hyperparameter tuning for all models was performed using 10-fold cross-validation, applied separately for each training sample. As a result, the selected hyperparameters were not necessarily identical for every sample within a given condition because data-specific hyperparameter tuning is crucial to account for the sampling variance in the training samples. For the cross-validation and estimation of the ENET we used the *glmnet* package (Version 4.1-8; Tay et al., 2023). The tuning grid included a constant vector of α-values between 0 and 1 (20 steps) and a vector of λ-values containing the initial λ values of *glmnet* for $\alpha = 0.5$ (equivalent to the "warm start" implemented by Kuhn and Max (2008) in *caret* as of Version 6.0-94).

For cross-validation and estimation of the RF, we used the *ranger* package (Version 0.17.0; Wright & Ziegler, 2017) as implemented in *caret* (Version 6.0-94; Kuhn & Max, 2008). The tuning grid included the number of random predictors at each node {2, $\sqrt{P}$, $\frac{P}{3}$, $\frac{P}{2}$}, the end node size {5, 10, 20} and the splitting criterion {variance, extratrees}. We used the root mean squared error (*RMSE*) as the optimality criterion to determine the best hyperparameter settings. We used RFs composed of 500 trees in every simulated condition. For cross-validation and estimation of the GBM, we used the *gbm* package (Version 2.1.9; Ridgeway & Developers, 2024) as implemented in *caret* (Version 6.0-94; Kuhn & Max, 2008). The tuning grid included the tree depth {1, 2, 3}, the end node size {5, 10, 20}, the maximal number of trees {5, 10, 20, 30, 40, 50, 80, ..., 500} and the shrinkage rate {0.001, .011, 0.031, 0.051, 0.101, 0.151, 0.201}. Again, we used the *RMSE* as the optimality criterion to determine the best hyperparameter settings.

We applied the "one-standard-error" rule (Breiman et al., 1984; Hastie et al., 2009) to select hyperparameters for all models—the ENET, the RF and the GBM. This approach favors more regularized models (e.g., larger λ for the ENET and smaller tree depth for the GBM) by choosing the simplest model whose performance is within one standard error of the minimum cross-validated error. For the ENET, this meant selecting $\lambda.1se$ and $\alpha.1se$. For the RF and GBM, the rule was applied across the entire hyperparameter grid, including the number of

random predictors at each node, the end node size and the splitting criterion (for RF), and tree depth, the end node size, the maximal number of trees, the shrinkage rate (for GBM). This approach represents a more conservative model selection strategy. By choosing the "one-standard-error" rule, we aimed to avoid overfitting in the model selection (Cawley & Talbot, 2010), thereby balancing between overfitting and capturing as much systematic effect variance as possible.

Analysis Procedure

As a first step, we conducted a mixed ANOVA with the ML model as within-sample factor and all other simulation conditions as between-sample factors to examine the influence of the manipulated variables on the amount of explained variance in the independent test sample. The between-sample factors were the sample size, the number of noise variables, the reliability, the simulated R^2_{sim}, the effect composition (linear vs. other associations), and the DGP. To conduct the ANOVA we used the *afex* package (Version 1.3-1; Singmann et al., 2024). We used Type III sums of squares and sum contrasts for effect coding. We refrained from reporting inferential statistics for the ANOVAs (i.e., F-Statistics and p-values) and instead focused on generalized η^2_G as effect size (Strobl et al., 2024).

Following this initial overview, we illustrate the predictive performance of each ML model for the different DGPs as a function of the most relevant simulation conditions (according to the mixed ANOVA). In applied research, it is common practice to fit and compare several ML models before drawing substantive conclusions. By mirroring this practice—fitting multiple ML models to different simulated DGPs—we aim to identify the conditions under which each model performs best. This approach not only provides insight into the relative strengths of different ML methods but also mimics the practical challenges that researchers face when selecting and interpreting models in empirical studies.

Dependent Measures

Coefficient of Determination

We trained each model on the simulated training dataset and then evaluated the model's performance on an independent test dataset randomly sampled from the same DGP. We evaluated the predictive performance of the models by calculating the coefficient of determination, R^2, on the test data set, because the performance measure is often used in psychology. By comparing the observed R^2 for the test sample to the true simulated R^2, we

evaluated the predictive performance of the model.

Relative R^2 Error

Since the maximal achievable predictive performance of the models is dependent on the different levels of simulated R^2 in the DGPs, we computed the relative deviation of the test R^2 from the simulated R^2 as $\frac{(R^2_{test} - R^2_{sim})}{R^2_{sim}}$. Larger absolute values indicate greater bias (i.e., more discrepancy from the true R^2_{sim}), whereas values closer to zero suggest R^2_{test} closer to R^2_{sim}. We use this measure only in the ANOVA to investigate which simulation parameters affect predictive performance.

Transparency and Openness

To promote transparency and reproducibility, we have made the analysis code and the results openly available at https://github.com/kimSpeck/MLsim. The simulated data and all subsequent analyses are fully reproducible using the same random seed in R. Additional online material can be found in the 'onlineMaterial' folder within the GitHub repository. This study has not been preregistered.

Results

In a first step, we investigated the predictive performance in the independent test sample (i.e., relative R^2 error) across the various simulation conditions to identify manipulations with substantial influence on predictive performance. An ANOVA, using the ML model as within-sample factor and all simulation conditions as between-sample factors, indicated that the predictive performance was affected by all simulation conditions (i.e., reliability: $\eta^2_G = .72$, sample size: $\eta^2_G = .7$, R^2_{sim}: $\eta^2_G = .65$, effect composition: $\eta^2_G = .58$ and DGP: $\eta^2_G = .09$) with the number of noise variables having an almost negligible effect ($\eta^2_G = .05$). In line with expectations, the predictive performance improved with higher reliability, larger sample size, higher R^2_{sim}, and with higher proportions of variance for the linear effects. The ANOVA also revealed interactions between the ML model and nearly all simulated conditions: ML model x DGP: $\eta^2_G = .38$, ML model x sample size: $\eta^2_G = .12$, ML model x effect composition: $\eta^2_G = .0.09$, ML model x R^2_{sim}: $\eta^2_G = .08$, ML model x reliability: $\eta^2_G = .06$)—including some higher order interactions (e.g., ML model x DGP x effect composition: $\eta^2_G = .15$, ML model x DGP x reliability: $\eta^2_G = .06$). These interactions indicated that the effects of the simulation conditions on predictive performance were highly model-dependent. Table 2 shows the results of the between-sample ANOVA for each combination of ML model and DGP to illustrate which

manipulation affects a given ML model's performance under a specific DGP.

Table 2
Between ANOVA Results for the relative R^2 error in every ML model x DGP combination

	Inter				Piecewise				Stepwise			
	trees		ENET		trees		ENET		trees		ENET	
Parameter	RF	GBM	inter	lin	RF	GBM	inter	lin	RF	GBM	inter	lin
Rel.	.80	.69	.72	.44	.82	.75	.61	.69	.85	.80	.58	.67
N	.80	.76	.72	.59	.74	.75	.75	.56	.71	.69	.73	.55
R^2	.66	.59	.54	.26	.76	.78	.68	.58	.76	.76	.68	.60
E.C.	.76	.69	.12	.92	.43	.26	.50	.52	.54	.34	.61	.64
N x R^2	.04	.01	.28	.10	.16	.13	.25	.25	.14	.11	.21	.22
N x #noise	.05	.01	.05	.02	.25	.12	.15	.20	.23	.13	.14	.18
#noise	.35	.12	.20	.05	.07	.01	.02	.00	.04	.00	.03	.00
N x R^2 x E.C.	.16	.11	.06	.12	.04	.02	.04	.06	.04	.02	.05	.05
Rel. x E.C.	.00	.01	.05	.17	.08	.07	.00	.01	.19	.17	.00	.00
N x E.C.	.12	.03	.03	.06	.08	.04	.05	.06	.08	.03	.05	.07
R^2 x Rel.	.03	.04	.02	.00	.06	.13	.05	.01	.10	.15	.04	.01

Note: Parameters are sorted in descending order according to the average η^2_G across all combinations of DGP and ML model. Any parameters for which none of the ANOVAs achieved an $\eta^2_G \geq 0.1$ have been omitted.

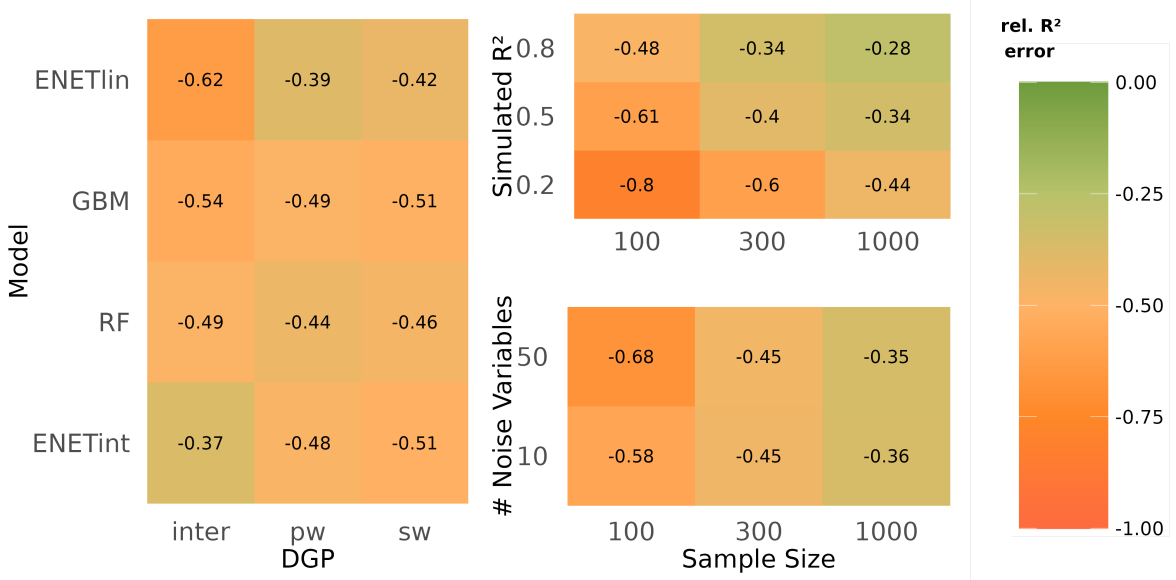
To further examine performance differences, we estimated marginal means of the relative R^2 error for the most prominent interaction effects (see Figure 2). Regarding the largest interaction effect between ML model and the DGP ($\eta^2_G = .38$), the marginal means revealed that no single model outperformed the others across all DGPs (left panel in Figure 2). Similarly, greater proportions of linear effects improved the predictive performance across all ML model x DGP combinations (i.e., ML model x DGP x effect composition: $\eta^2_G = .15$). However, the degree of improvement depended on each model's ability to capture the respective non-linear association (e.g., ENET_{lin} was particularly affected by dominating interaction effects due to its inability to adequately represent such associations). Overall, these findings highlight that the relative advantages of each ML method depend critically on the underlying data structure. This conclusion is consistent with the so-called "No Free Lunch" theorems, which demonstrate that no single model can outperform all others across all possible DGPs (Wolpert & Macready, 1997).

Benefits of increasing sample size depend on characteristics of the data (i.e., R^2_{sim} and the number of noise variables; see Figure 2). Predictive performance (R^2_{test}) improves with sample size at all R^2_{sim} levels, but R^2_{test} profits more from increased sample sizes at lower R^2_{sim} levels (R^2_{sim} x sample size: $\eta^2_G = .13$). Thus, predictive performance is more strongly influenced by sample size when the true variance explained in the data is low. Similarly, having many

noise variables in the data impairs performance more severely at smaller sample sizes (e.g., N = 100), suggesting that larger samples enhance the ability of ML methods to distinguish genuine predictors from irrelevant noise.

Figure 2

Marginal Means of the relative R^2 error for selected simulation conditions



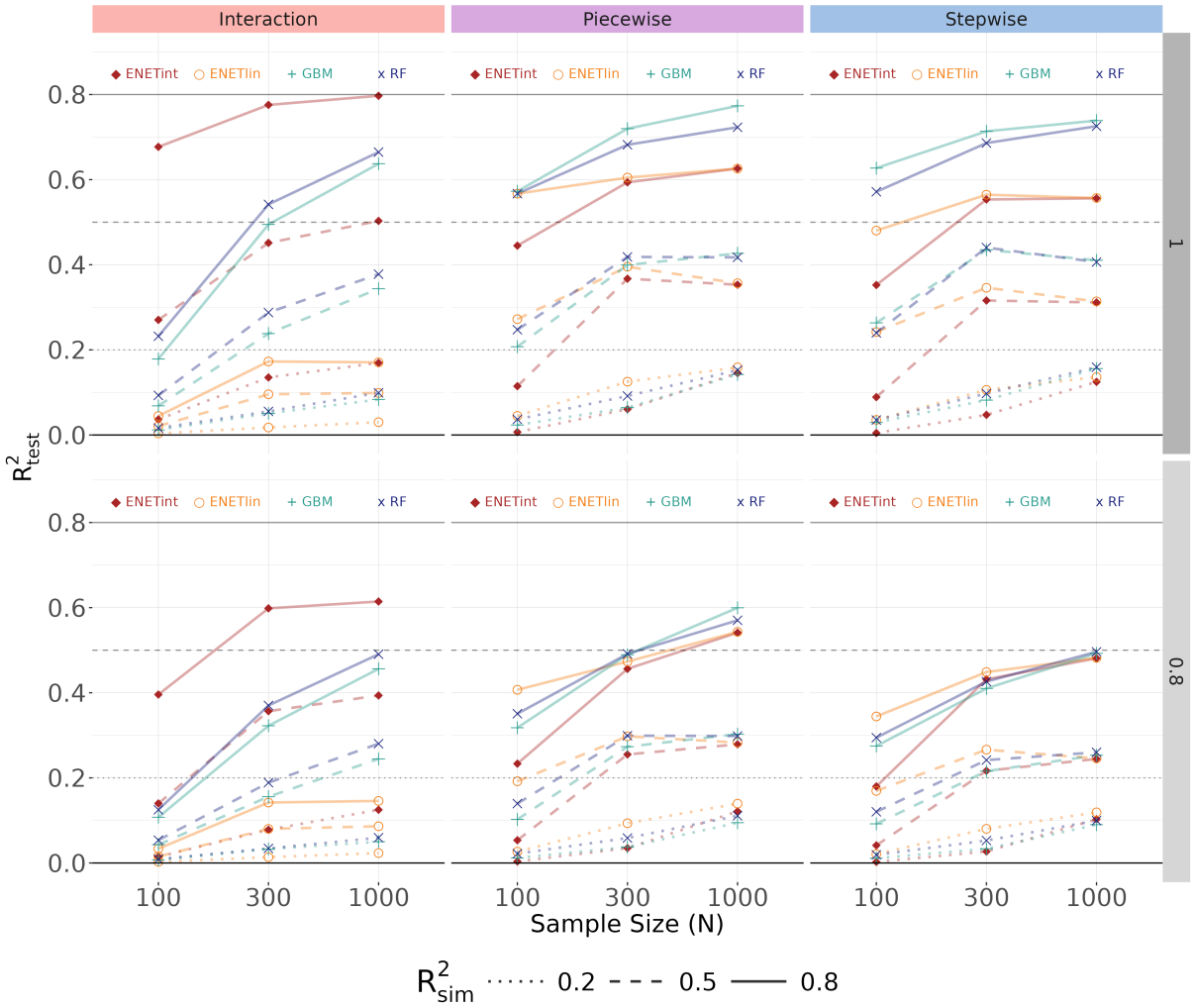
Note. Only the most prominent interaction effects of the mixed ANOVA are displayed. Values represent marginal means in relative R^2 error for the indicated combinations of simulated conditions. Values close to zero (greenish tiles) indicate that R^2_{sim} was nearly recovered; values near one (orange tiles) indicate that almost no variance was explained.

To evaluate the simulation results in more detail, Figure 3 and Figure 4 depict the average achieved R^2_{test} across all samples as a function of the DGP, the ML model, R^2_{sim}, effect composition, reliability and sample size. The influence of the number of noise variables on the R^2_{test} was characterized by a small main effect ($\eta^2_G = .05$) with increasing number of noise variables resulting in slightly lower R^2_{test}. Therefore, we focused on the subset of results with a fixed number of 50 noise variables. We begin by examining simulated conditions that emphasized non-linear effects (i.e., effect composition with 80% of the R^2_{sim} on interactions, stepwise or piecewise associations, respectively; Figure 3). Specifically, we depict the predictive performance of each ML model within each DGP for perfect reliability conditions and for reliability = 0.8. This allows to examine the effect of measurement error on each models performance. We chose to depict conditions with reliability of 0.8 because this is still widely regarded as good or at least acceptable in psychological research. All simulation conditions with reliability of 0.6 yielded consistent conclusions and are detailed in the additional online

material.

Figure 3

Averaged R^2_{test} for DGPs with Dominating Non-Linear Effects



Note. Effect composition is fixed to 80% of R^2_{sim} on non-linear effects. Different DGPs {Interaction, Piecewise, Stepwise} in columns. Reliability of the predictors in rows. Linetypes illustrate different levels of R^2_{sim}. The number of noise variables is fixed to 50.

For the DGP including interactions (left panel, top row of Figure 3), the R^2_{test} of the ENET_{int} approached the theoretical maximum (i.e., R^2_{sim} represented by horizontal lines of different linetypes) only for large sample sizes (N = 1000), in the perfect reliability scenario and for $R^2_{sim} > 0.2$. RF and GBM fell short of the maximal achievable R^2_{sim} although both models are in principle capable of representing continuous interactions. For the interaction DGP, RF showed a slight advantage over the GBM. The ENET_{lin} is limited to capturing linear effects ($R^2 \approx 0.16$ for $R^2_{sim} = 0.8$) and reached this maximum at N = 300. This overall pattern in terms of comparative model performance remained consistent across different levels of R^2_{sim}.

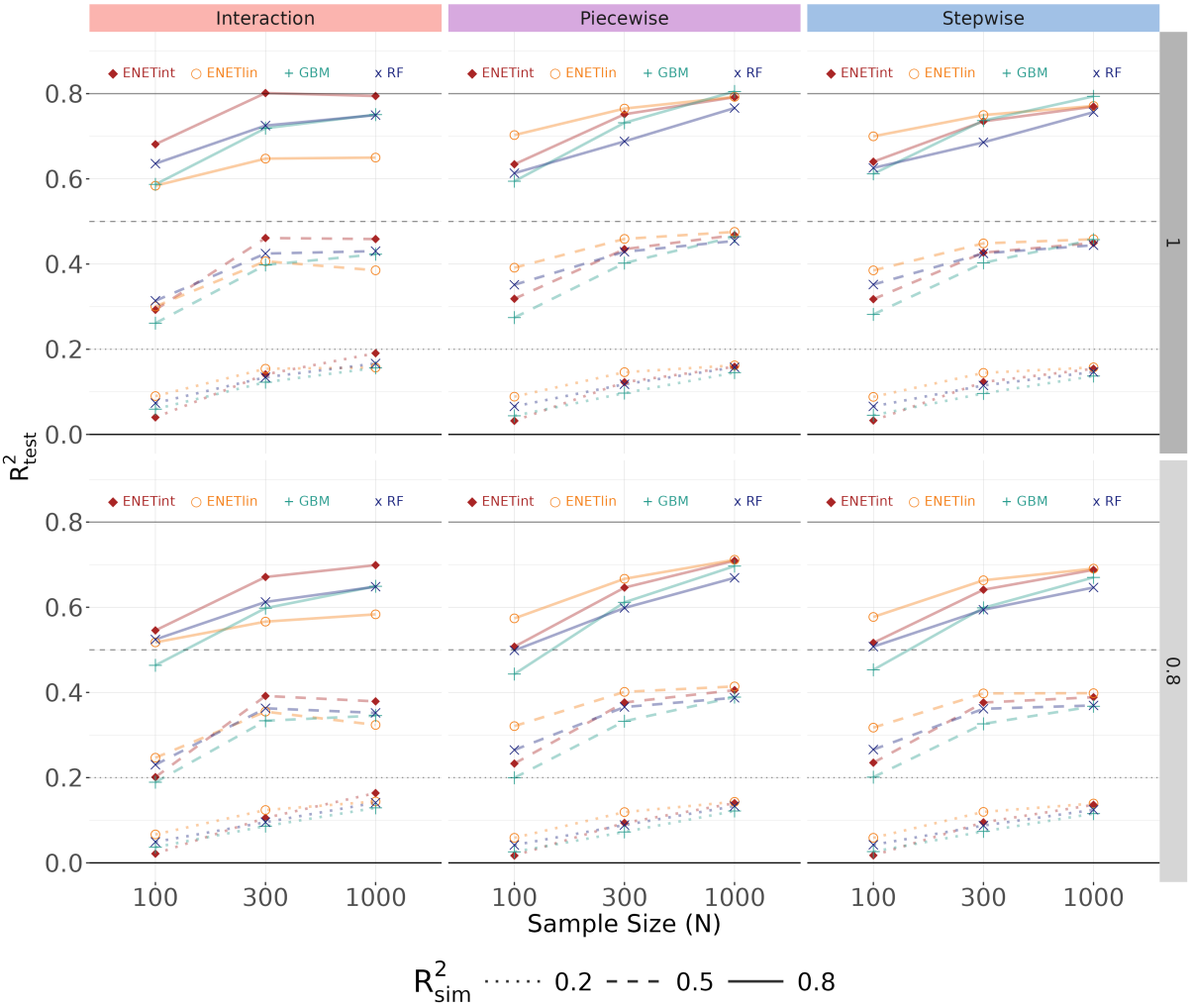
Adding measurement error reduced the predictive performance of all ML models, but did not alter their relative ranking (left panel, bottom row of Figure 3).

For data including piecewise or stepwise effects (top row, middle, and right panel in Figure 3, respectively), both RF and GBM explained more variance than the ENET models under ideal conditions (i.e., $R^2_{sim} \geq 0.5$ with 80% in non-linear effects, reliability = 1 and N = 1000). However, even for N = 1000 and perfect reliability, the GBM and the RF fell short of the ideal R^2_{sim}. As neither the piecewise data nor stepwise data involve interactions, the ENET_{lin} and ENET_{int} performed similarly when the sample size was large. For smaller sample sizes or lower R^2_{sim}, ENET_{lin} explained more variance than ENET_{int}, because the inclusion of irrelevant interaction terms—as was the case for these DGPs and the ENET_{int} model—effectively reduced the signal-to-noise ratio. At lower values of R^2_{sim} and/or for smaller sample sizes, ENET_{lin} matched or exceeded the performance of all other models—including the tree-based methods. When measurement error was added, the predictive performance of the ENET models closely resembled that of the tree-based models for both piecewise and stepwise DGPs, even at large sample sizes (bottom row, middle and right panel in Figure 3, respectively). In the least favorable conditions (i.e., smaller sample sizes, smaller R^2_{sim}), again, the ENET_{lin} outperformed every other model. Piecewise and stepwise data differed in the extent to which the ENET models captured variance, which in turn influenced the advantage gained by tree-based methods. Because stepwise effects aligned more closely with the splitting logic of trees, tree-based models outperformed ENET models by a larger margin in this condition. However, the introduction of measurement error also impaired predictive performance more strongly in the stepwise condition—particularly for the tree-based models—compared to the piecewise condition.

Until now, we reported results from conditions in which non-linear effects dominated, accounting for 80% of the explained variance. We now turn to data dominated by linear effects (i.e., 80% of R^2_{sim} attributed to linear predictors; see Figure 4), a pattern commonly observed in psychological research. When linear effects dominated, differences in predictive performance between models decreased markedly across all DGPs. These reduced differences are consistent with the lower proportion of variance explained by interactions, piecewise linear, or stepwise structures in these scenarios, which limits the potential for predictive improvement through methods that capture such complexities even under ideal conditions. Notably, tree-based methods achieved higher R^2_{test} in conditions with dominating linear effects

compared with dominating non-linear effects. This suggests that tree-based methods could represent linear relationships well, but struggled more with continuous variable interactions. Under ideal conditions (perfect reliability, large sample sizes [N = 1000]), the GBM showed a slight advantage over RF for piecewise and stepwise data (top row, middle and right panel in Figure 4). Nevertheless, predictive performance was quite similar among all methods. However, as sample size, R^2_{sim} or reliability declined, the ENET models outperformed the tree-based approaches - especially ENET_{lin} with its simpler structure. Because the majority of variance was linear, adding measurement error similarly affected all models. Under these more realistic psychological conditions with dominating linear effects and added measurement error, tree-based methods consistently failed to surpass ENET.

Figure 4
Averaged R^2_{test} for DGPs with Dominating Linear Effects.



Note. Effect composition is fixed to 80% of R^2_{sim} on linear effects. Different DGPs {Interaction, Piecewise, Stepwise} in columns. Reliability of the predictors in rows. Linetypes illustrate different levels of R^2_{sim}. The number of noise variables is fixed to 50.

Overall, predictive performance depended strongly on the characteristics of the data, with each modeling approach excelling under specific conditions. Three key patterns emerged: (1) Larger sample sizes, higher reliability, and stronger true effects (R^2_{sim}) consistently improved performance. (2) Under less favorable conditions—such as small sample sizes, weaker true effects, or predictors measured with error—simpler ENET models frequently outperformed tree-based methods. (3) Tree-based approaches outperformed ENET models only when non-linear effects (i.e., piecewise or stepwise effects) were *dominant*, and only in large samples with perfect reliability.

Discussion

Quality, quantity, and structure of datasets in psychology systematically differ from those in computer science, where ML methods were originally developed. Psychological datasets are often smaller and noisier compared to those typically encountered in computational sciences. Consequently, when applying ML methods to psychological data, their performance is not necessarily superior to simpler models—despite the presumed complexity of psychological phenomena. More specifically, several studies comparing different ML models to (penalized) linear regression did not find consistent evidence that more flexible ML methods outperform (penalized) regression on empirical datasets (e.g., Christodoulou et al., 2019; Fokkema et al., 2022; Gravesteijn et al., 2020; Jankowsky et al., 2025; Pargent & Albert-von der Gönna, 2018). Based on these empirical studies, it remained unclear to what extent data characteristics (e.g., sample size, effect size, noise variables, measurement error) contribute to the limited, or even lacking, predictive gains of ML models. To address this gap, we systematically manipulated these characteristics and investigated how they affect ML model performance in a comprehensive simulation study. Specifically, we compared the predictive performance of ENET regressions, RF, and GBM for datasets generated from different DGPs. The following discussion highlights three major implications of our results: (1) the impact of specific data characteristics on prediction, (2) methodological considerations for modeling commonly noisy data in psychology, and (3) the methodological limitations and the extent to which our findings generalize.

Complex Predictive Models Meet Limited Data

Overall, only under the most optimal simulated conditions ($N = 1,000$, perfectly reliable predictors, and large overall effect size of R^2 = .80) did ML models reach the maximum possible predictive performance ($R^2_{sim} \approx R^2_{test}$). As outlined in the introduction, these optimal conditions in terms of sample size and reliability are rarely met in psychological data. In less favorable conditions, all models fell short of the R^2_{sim}, sometimes by a substantial margin. Although we observed that each ML model outperformed the others under certain conditions, none was consistently superior or completely robust to suboptimal data characteristics. Overall, our results illustrate the No Free Lunch theorem stating that no single model performs best across all simulated condition or real-world datasets (Wolpert & Macready, 1997).

While no model is universally optimal, a given model may achieve superior performance in specific contexts when its assumptions align with the actual DGP. In the following, we

discuss how each manipulated data characteristic influenced performance of the investigated models. Our results indicate that all methods benefit greatly from increased sample sizes, particularly when effect sizes are small or the dataset contains many irrelevant (noise) variables (see Figure 2). Due to the models' regularization mechanisms, simply adding more noise variables only had a minor effect (Kuhn & Johnson, 2013). Instead, the predictor-to-sample ratio—which is known to affect the extent of overfitting (Vabalas et al., 2019)—seemed more relevant in terms of predictive performance. This is evident when comparing the performance of $ENET_{lin}$ with $ENET_{int}$ in the datasets with a piecewise or stepwise DGP. In these DGPs, $ENET_{lin}$, which had the better predictor-to-sample ratio, consistently achieved higher prediction and was more robust against overfitting[4]. When applied to small samples with only weak effects, more flexible ML methods seemed to be more prone to overfitting—also due to greater uncertainty and variability in tuning parameter estimation (Riley et al., 2021; Van Calster et al., 2020). Overall, larger samples consistently enhance predictive performance, and thus represent a generally recommended strategy, regardless of the specific ML method used. However, we acknowledge that adding more observations might be challenging and sometimes unfeasible in psychological and clinical research due to funding constraints, genuinely small populations for rare disorders, and labor-intensive procedures (e.g., Button et al., 2013).

Although the influence of measurement error on predictive performance remains underrepresented in the psychological ML literature (see for a similar argument Lavelle-Hill et al., 2025), its importance has already been emphasized in previous studies (e.g., Jacobucci and Grimm, 2020; McNamara et al., 2022). The present simulation underlines the critical role of measurement error because reliability was the most important factor across all combinations of DGP and ML models. Moreover, it should not be assumed that ML models inherently mitigate unreliable measurement. Similar to unregularized multiple regression, these models generally assume that predictor variables are measured without error. In line with Jacobucci and Grimm (2020), we observed performance decreased across all ML models and DGPs as reliability decreased. Although larger samples can improve estimation precision, they do not necessarily offset the effects of low measurement reliability. While regularization can effectively handle irrelevant noise variables (Kuhn & Johnson, 2013), ML models are not

[4] Overfit results are available in the additional online material. Notably, the difference in overfitting between $ENET_{int}$ and $ENET_{lin}$, due to their differences in predictor-to-sample size ratio, was particularly pronounced in the $N = 100$ simulation conditions of the Interactions DGP.

designed to address measurement error in relevant predictor variables. That is why hybrid approaches that combine structural equation modeling (SEM) and ML techniques, such as regularized SEM (Jacobucci et al., 2016) or SEM trees (Brandmaier et al., 2013), have recently been proposed.

Another approach to improve predictor reliability involves data-driven feature engineering techniques, such as using composite indicators or aggregate scores[5] (Kuhn & Johnson, 2013; Stachl et al., 2020). However, this approach remains controversial, since several studies have reported higher predictive performance when individual items are used as predictors (e.g., Achaa-Amankwaa et al., 2021; Fokkema et al., 2022; Mõttus & Rozgonjuk, 2021; Schroeders et al., 2021; Seeboth & Mõttus, 2018). Using items rather than scales as predictors may improve predictive performance, if unique variance that is not shared with other items happens to correlate with the outcome. Importantly, unique item variance is not necessarily measurement error in the sense of being uncorrelated with all other variables, including the outcome. Instead, unique variance may reflect systematic, outcome-relevant information that is not shared with the other items through common variance. In such cases, retaining unique variance may benefit prediction, despite not reflecting the central core of the construct (Lavelle-Hill et al., 2025). Moreover, Rhemtulla et al. (2020) caution that the use of inappropriate factor models can negatively impact predictive performance, further complicating the decision between individual and aggregate predictors. Given the detrimental effect of measurement error in our simulation, future research should focus on addressing unreliability in predictors used within ML methods.

The Shape of Reality: How Effect Composition Impacts Model Performance

In addition to general data characteristics such as sample size, the number of noise variables or reliability, we also examined the effect composition within the DGP. This manipulation touches on the fundamental question about the structure of real-world data: Do linear or non-linear associations explain most of the variance in the outcome variable? Assuming non-linear effects predominate, we investigated how model performance was affected by our manipulations. For interaction data, the $ENET_{int}$ performed well—especially in conditions in which the interactions explained most of the outcome variance. Because the DGP and the ML model specification align perfectly in this case, this result is largely due to the

[5] To avoid data leakage and ensure valid performance, calculating composite scores and aggregating items to scales should be done separately within the training and test sets.

simulation setup. While non-parametric models such as RF and GBM are generally capable of approximating smooth functions and of detecting interaction effects without explicit specification, they struggled to make accurate predictions in data with dominating continuous interactions. This could be an instance of the *XOR problem* which states that an interaction cannot be discovered if the respective marginal ("main") effects are too small (Strobl et al., 2009). Especially for small R^2_{sim} and small sample sizes, the difficulty to detect main effects for the RF and GBM could have limited their ability to capture interactions—despite their large effect sizes—and therefore also limited predictive performance. Supporting this interpretation, we also observed that tree-based methods yielded better predictive performance for dominating linear effects, where marginal ("main") effects are easier to detect due to the simulation setup.

Both RF and GBM are more compatible with data that follow a piecewise or stepwise DGP, and thus showed higher predictive performance compared to ENET. However, both algorithms lose their advantage over the ENET in the presence of measurement error—even when sample sizes are large. This is consistent with McNamara et al. (2022), who also found that GBMs lose their advantage over regularized linear regression when moderate measurement error is present. Thus, to achieve superior predictive performance with more flexible ML methods, the data needs to be measured with perfect reliability and must be paired with either larger sample sizes or pre-trained models. To uncover these complex patterns, non-parametric models like RF and GBM require more data than parametric approaches due to their greater flexibility (e.g., Van Der Ploeg et al., 2014). Because of this data-hungriness, even sample sizes considered large in psychology (e.g., $N = 5{,}000$) may be insufficient for highly flexible non-parametric ML methods. Taken together, the present study emphasizes the importance of study designs that ensure both measurement quality (i.e., high reliability) and data quantity (i.e., adequate sample size) to effectively capture non-linear effects.

Our results highlight that, even when genuinely complex psychological relationships constitute the true DGP, the sample size requirements for detecting such patterns with ML methods are substantial. However, while high complexity may characterize some contexts, it is possible that (almost) linear relationships prevail in others. If data consists mainly of linear effects, a regularized linear regression is well-suited to capture the underlying relationship. Previous empirical research demonstrated that in real-world panel data (N = 2,000), both ENET and RF yielded very similar predictive performance (Pargent & Albert-von der Gönna,

2018). Overall, these findings support the view that, when underlying relationships are largely linear and sample sizes are sufficient, regularized linear models can match the predictive performance of more flexible methods like RF or GBM, especially in the presence of measurement error. In comparing the different DGPs, our simulation results demonstrate that domain knowledge (e.g., explicitly modeling interactions when present in $ENET_{int}$) can yield performance benefits for ML methods that are well aligned with the DGP.

Incorporating accurate knowledge of the DGP can help in optimizing predictive models, especially in the context of limited data or measurement error. Thus, we recommend to more strongly rely on formalization of theories and to explicitly integrate theoretical knowledge into ML models whenever possible, rather than relying solely on non-parametric methods to uncover patterns. Such an understanding of modeling includes theory-driven feature engineering. For instance, presumed continuous effects and interactions can be directly specified in regularized regression models, and tree-based ML methods can also be adapted to leverage prior knowledge about specific interactions. For instance, tree-based methods can be integrated with parametric models to identify subgroups characterized by homogeneous model parameters (Fokkema et al., 2025). Similarly, semi-confirmatory approaches have been proposed in the context of regularized structural equation models (Huang, 2018; Huang, 2020; Huang et al., 2017), and our findings further underscore their potential.

Inferring Complex Relationships From Noisy Psychological Data

Hypotheses about linear and non-linear effects can also be tested in a data-driven fashion using the deductive data mining (DDM) approach suggested by Hong et al. (2020). In DDM, increasingly complex models are compared to uncover the true nature of relationships between predictor variables and outcomes. More specifically, Hong et al. (2020) proposed to compare (regularized) linear regressions with boosted decision trees of increasing tree depths, allowing for non-linear (stump model), first-order or higher-order interaction effects (tree depth > 1). This approach aligns with recommendations from the ML literature, which emphasizes comparing "strong" baseline models to more complex methods (Wolfrath et al., 2024). However, Hong et al. (2020) also mentioned the importance of adequate sample sizes for the DDM approach and note a general lack of comprehensive simulation studies that examine sample size requirements for the DDM approach in the behavioral sciences.

The present simulation study revealed possible boundary conditions of the DDM approach. With sufficiently large samples, a stepwise model comparison of $ENET_{lin}$, $ENET_{int}$,

and GBM might allow correct inferences regarding the presence of first-order interaction effects and the absence of substantial non-linear or higher-order interaction effects in terms of an increased predictive performance. However, when sample sizes are small or when interaction and non-linear effects are weak, the average predictive performance of the models becomes a less reliable proxy of the underlying DGP. For example, in our simulation, GBM did not consistently outperform $ENET_{lin}$ across samples, even when interaction or non-linear effects were present in the DGP. In line with Shmueli (2010), our results show that underspecified models, such as $ENET_{lin}$, can achieve even better predictive performance than correctly specified models in case of weak effects or noisy data. This underscores that the model best suited for prediction may not correspond to the true DGP, and that high predictive performance does not necessarily guarantee the detection of true effects (Shmueli, 2010; Yarkoni & Westfall, 2017). As a result, the validity of the DDM approach depends on samples size and the effect sizes of the interactions or non-linear effects. Assuming that flexible, non-parametric ML methods are inherently robust under more challenging conditions (e.g., $N \leq 500$) can lead to false negatives, erroneously concluding that interaction effects or non-linear effects do not exist in the given data.

Methodological Refinements for Noisy Psychological Data

In this study, we used widely adopted implementations of tree-based methods, which are popular in applied research but are not always ideally suited for noisy psychological data. Below, we outline several methodological refinements that could be explored in future studies to further enhance predictive performance in such contexts. In the implementation of RFs, we followed the original recommendations of Breiman (2001) by growing near-full-depth trees without pruning—a practice that remains the default in many software packages (e.g., Wright & Ziegler, 2017). Instead of tuning tree depth directly, we controlled the individual tree size via cross-validating the end node size. Empirical evidence suggests that treating tree depth as an additional hyperparameter can improve performance (Duroux & Scornet, 2018). In low-reliability simulation conditions with small R^2_{sim}, cross-validating the tree depth or pruning the trees post-growth can provide additional regularization (Zhou & Mentch, 2023), potentially increasing predictive performance[6]. For GBM, too, various regularization techniques can

[6] Tuning tree depth can improve performance but comes at the cost of increased complexity of the tuning grid (Hastie et al., 2009), which may be problematic in small sample size or low signal-to-noise conditions. Zhou and Mentch (2023) showed that tuning tree depth while keeping the number of randomly selected predictors at a default level can benefit prediction for demanding data.

improve model performance beyond tuning parameters like maximum tree depth or shrinkage. Options include regularizing leaf weights (e.g., L1 and L2 regularization in XGBoost; Chen & Guestrin, 2016), dropout techniques (*DART*; Vinayak & Gilad-Bachrach, 2015), and regularizing the base learners in each step of the GBM (e.g., Regularized Gradient Boosting; Cortes et al., 2019). Additionally, using alternative optimization algorithms, such as those leveraging second-order methods (Bottou et al., 2018), can improve training stability and efficiency. Importantly, these refined techniques should not be overhyped in the next wave of "magical thinking", as they will also face the outlined fundamental limitations in predictive performance. Methodological studies consistently demonstrate that both parametric and non-parametric ML methods are fundamentally constrained by data quality and quantity.

Limitations

It is important to acknowledge a few limitations of our simulation study. As with any simulation, our findings are tied to the specific conditions we modeled, and generalization to other data scenarios should be made with caution. First, this applies to our hyperparameter tuning approach, where we used a single strategy (10-fold cross-validation with grid search) across all simulation conditions. Although consistent, this choice may not be optimal for smaller samples, where a more constrained search space or alternative heuristics (e.g., repeated cross-validation or sequential model-based optimization for RF; Probst et al., 2019) might outperform the broad grid search.

Second, we examined a specific subset of conceivable DGPs: We included only linear effects, first-order interactions and non-linear effects following a simple splitting logic (i.e., the piecewise[7] and stepwise DGP), excluding more complex relationships such as higher-order interactions, or other non-linear effects (e.g., polynomial terms). Therefore, the investigated DGPs reflect data structures that are more typical for observational, between-subjects designs, but the results do not necessarily generalize to other data structures. This applies both to experimental, within-subject designs discussed in the introduction, but also to research areas such as neuro-cognitive or psychophysiological research, where more complex, non-linear relationships could be expected. These data types presumably involve non-linear dynamics that may be better suited to the strengths of non-parametric ML methods (e.g., GBM or support vector machines). However, reliably detecting these subtle non-linear dynamics

[7] Note that a piecewise-linear relationship with a single breakpoint can be translated to an interaction between the continuous variable and a dichotomous variable indicating on which side of the breakpoint the observation lies.

would require both high data quantity and data quality, as well. The non-linear effects we selected were intended to illustrate a broader argument: subtle effects, regardless of their specific form, can only be reliably detected when their effect size is sufficiently large, the sample size is adequate, and the measurement is reliable. This principle applies across methods and types of effects. Including additional algorithms, such as neural networks, could provide further insights but would not alter our core conclusion.

Conclusion

The present findings emphasize how data quality and quantity fundamentally constrain the performance of ML methods and stress the importance of considering the data characteristics when choosing ML methods. Three key insights emerge from our simulation study: First, improving the measurement quality of predictor variables is the most effective strategy for enhancing predictive performance. Second, typical sample sizes in psychological research are often insufficient for ML methods. Even sample sizes up to 1,000 often prove insufficient for ML models to approach maximal performance, particularly when data are noisy or effects are small. Third, while comparing methods of varying complexity may help in identifying specific types of effects, the validity of such comparisons critically depends on data quality and quantity. When data quality is too low, simpler linear models can outperform more complex models even when the underlying DGP is dominated by non-linear processes. Substantially better predictions for more complex models only occur under genuinely large non-linear or interaction effects involving highly reliable predictors. Consequently, with relatively simple DGPs and for small sample sizes, complex ML models yield little advantage over simpler (parametric) models. In other words, the combination of self-report measures and small sample sizes commonly found in psychology limits the potential for ML models to outperform simpler parametric approaches. Without substantial improvements in data quality and quantity, non-parametric ML models are unlikely to consistently outperform traditional approaches in psychological research.

Acknowledgements

The authors want to thank two anonymous reviewers and Prof. Dr. Brandmaier as the corresponding editor for their invaluable feedback on early versions of the manuscript.

References

Achaa-Amankwaa, P., Olaru, G., & Schroeders, U. (2021). Coffee or tea? Examining cross-cultural differences in personality nuances across former colonies of the British Empire. *European Journal of Personality*, *35*(3), 383–397. https://doi.org/10.1177/0890207020962327

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*, *73*(7), 899–917. http://dx.doi.org/10.1037/amp0000190

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology*, *90*(1), 94–107. https://doi.org/10.1037/0021-9010.90.1.94

Ali, F., & Ang, R. P. (2022). Predicting how well adolescents get along with peers and teachers: A machine learning approach. *Journal of Youth and Adolescence*, *51*(7), 1241–1256. https://doi.org/10.1007/s10964-022-01605-5

Anderson, D. R., & Burnham, K. P. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd). Springer-Verlag. https://doi.org/10.1007/b97636

Anderson, I., Gil, S., Gibson, C., Wolf, S., Shapiro, W., Semerci, O., & Greenberg, D. M. (2021). "Just the way you are": Linking music listening on spotify and personality. *Social Psychological and Personality Science*, *12*(4), 561–572. https://doi.org/10.1177/1948550620923228

Anderson, T. W., Anderson, T. W., Anderson, T. W., & Anderson, T. W. (1958). *An introduction to multivariate statistical analysis* (Vol. 2). Wiley New York.

Beauchamp, M. (2018). On numerical computation for the distribution of the convolution of N independent rectified gaussian variables. *Journal de la société française de statistique*, *159*(1), 88–111. https://www.numdam.org/item/JSFS_2018__159_1_88_0/

Beck, E. D., & Jackson, J. J. (2022). A mega-analysis of personality prediction: Robustness and boundary conditions. *Journal of Personality and Social Psychology*, *122*(3), 523–553. https://dx.doi.org/10.1037/pspp0000386

Berk, R. A., et al. (2008). *Statistical learning from a regression perspective* (Vol. 14). Springer.

Bhattacharyya, A. (1946). On Some Analogues of the Amount of Information and Their Use in Statistical Estimation. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, *8*(1), 1–14. https://www.jstor.org/stable/25047921

Bohrnstedt, G. W., & Goldberger, A. S. (1969). On the exact covariance of products of random variables. *Journal of the American Statistical Association*, *64*(328), 1439–1442. https://doi.org/10.1080/01621459.1969.10501069

Borsboom, D., Van Der Maas, H. L., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Bottou, L., Curtis, F. E., & Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM review*, *60*(2), 223–311. https://arxiv.org/pdf/1606.04838

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, *18*(1), 71–86. https://doi.org/10.1037/a0030001

Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). *Classification and regression trees*. Taylor & Francis. https://books.google.de/books?id=JwQx-WOmSyQC

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Busemeyer, J. R., & Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin*, *93*(3), 549–562. https://doi.org/10.1037/0033-2909.93.3.549

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, *11*, 2079–2107. http://jmlr.csail.mit.edu/papers/v11/cawley10a.html

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 785–794. https://arxiv.org/abs/1603.02754

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Cortes, C., Mohri, M., & Storcheus, D. (2019). Regularized gradient boosting. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/465636eb4a7ff4b267f3b765d07a02da-Paper.pdf

Davey, J., Turner, R. M., Clarke, M. J., & Higgins, J. P. (2011). Characteristics of meta-analyses and their component studies in the cochrane database of systematic reviews: A cross-sectional, descriptive analysis. *BMC Medical Research Methodology*, *11*, 1–11. https://doi.org/10.1186/1471-2288-11-160

Diaz, M., Kairouz, P., & Sankar, L. (2022). Lower Bounds for the MMSE via Neural Network Estimation and Their Applications to Privacy. https://doi.org/10.48550/arXiv.2108.12851

Duroux, R., & Scornet, E. (2018). Impact of subsampling and tree depth on random forests. *ESAIM: Probability and Statistics*, *22*, 96–128. https://doi.org/10.1051/ps/2018008

Efron, B., & Hastie, T. (2021). *Computer age statistical inference, student edition: Algorithms, evidence, and data science* (Vol. 6). Cambridge University Press.

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Fokkema, M., Henninger, M., & Strobl, C. (2025). One model may not fit all: Subgroup detection using model-based recursive partitioning. *Journal of School Psychology*, *109*, 101394. https://doi.org/https://doi.org/10.1016/j.jsp.2024.101394

Fokkema, M., Iliescu, D., Greiff, S., & Ziegler, M. (2022). Machine learning and prediction in psychological assessment. *European Journal of Psychological Assessment*, *38*(3), 165–175. https://doi.org/10.1027/1015-5759/a000714

Fox, K. R., Huang, X., Linthicum, K. P., Wang, S. B., Franklin, J. C., & Ribeiro, J. D. (2019). Model complexity improves the prediction of nonsuicidal self-injury. *Journal of Consulting and Clinical Psychology*, *87*(8), 684–695. https://doi.org/10.1037/ccp0000421

Fraley, R. C., Chong, J. Y., Baacke, K. A., Greco, A. J., Guan, H., & Vazire, S. (2022). Journal n-pact factors from 2011 to 2019: Evaluating the quality of social/personality journals with respect to sample size and statistical power. *Advances in Methods and Practices in Psychological Science*, *5*(4), Article 25152459221120217. https://doi.org/10.1177/25152459221120217

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2021). *mvtnorm: Multivariate normal and t distributions* [R package version 1.1-3]. https://CRAN.R-project.org/package=mvtnorm

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Gomez Penedo, J. M., Schwartz, B., Giesemann, J., Rubel, J. A., Deisenhofer, A.-K., & Lutz, W. (2022). For whom should psychotherapy focus on problem coping? a machine learning algorithm for treatment personalization. *Psychotherapy Research*, *32*(2), 151–164. https://doi.org/10.1080/10503307.2021.1930242

Götz, F. M., Gosling, S. D., & Rentfrow, P. J. (2022). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives on psychological science*, *17*(1), 205–215. https://doi.org/10.1177/1745691620984483

Gravesteijn, B. Y., Nieboer, D., Ercole, A., Lingsma, H. F., Nelson, D., Van Calster, B., Steyerberg, E. W., Åkerlund, C., Amrein, K., Andelic, N., et al. (2020). Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *Journal of clinical epidemiology*, *122*, 95–107. https://doi.org/10.1016/j.jclinepi.2020.03.005

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182. https://dl.acm.org/doi/10.5555/944919.944968#bibliography

Harrell Jr., F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, *15*(4), 361–387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer. https://doi.org/10.1007/978-0-387-84858-7

Helwig, N. E. (2017). Adding bias to reduce variance in psychological results: A tutorial on penalized regression. *The Quantitative Methods for Psychology*, *13*(1), 1–19. https://doi.org/10.20982/tqmp.13.1.p001

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Hong, M., Jacobucci, R., & Lubke, G. (2020). Deductive data mining. *Psychological Methods*, *25*(6), 691–707. https://doi.org/10.1037/met0000252

Huang, P.-H. (2018). A penalized likelihood method for multi-group structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *71*(3). https://doi.org/10.1111/bmsp.12130

Huang, P.-h. (2020). lslx: Semi-Confirmatory Structural Equation Modeling via Penalized Likelihood. *Journal of Statistical Software*, *93*(7), 1–37. https://doi.org/10.18637/jss.v093.i07

Huang, P.-H., Chen, H., & Weng, L.-J. (2017). A Penalized Likelihood Method for Structural Equation Modeling. *Psychometrika*, *82*(2), 329–354. https://doi.org/10.1007/s11336-017-9566-9

Huang, X., Ribeiro, J. D., & Franklin, J. C. (2020). The differences between suicide ideators and suicide attempters: Simple, complicated, or complex? *Journal of Consulting and Clinical Psychology*, *88*(6), 554–563. https://doi.org/10.1037/ccp0000498

Hussey, I., Alsalti, T., Bosco, F., Elson, M., & Arslan, R. C. (2023). An aberrant abundance of cronbach's alpha values at .70. https://doi.org/10.31234/osf.io/dm8xn

Iliescu, D., Greiff, S., Ziegler, M., & Fokkema, M. (2022). Artificial Intelligence, Machine Learning, and Other Demons. *European Journal of Psychological Assessment*, *38*(3), 163–164. https://doi.org/10.1027/1015-5759/a000713

Jaccard, J., & Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological Bulletin*, *117*(2), 348–357. https://doi.org/10.1037/0033-2909.117.2.348

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, *15*(3), 809–816. https://doi.org/10.1177/1745691620902467

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: a Multidisciplinary Journal*, *23*(4), 555–566. https://doi.org/10.1080/10705511.2016.1154793

Jacobucci, R., Grimm, K. J., & Zhang, Z. (2023). *Machine learning for social and behavioral research* [Paperback published July 11, 2023; Hardcover July 17, 2023]. The Guilford Press.

Jacobucci, R., Littlefield, A. K., Millner, A. J., Kleiman, E. M., & Steinley, D. (2021). Evidence of inflated prediction performance: A commentary on machine learning and suicide research. *Clinical Psychological Science*, *9*(1), 129–134. https://doi.org/10.1177/2167702620954216

Jankowsky, K., Belobrajdic, N., Węziak-Białowolska, D., Bialowolski, P., & McGrath, R. E. (2025). Character strengths as universal predictors of health? using machine learning to examine the predictive validity of character strengths across cultures. *PsyArXiv*. https://doi.org/10.31234/osf.io/fzywe_v2

Jankowsky, K., & Schroeders, U. (2022). Validation and generalizability of machine learning prediction models on attrition in longitudinal studies. *International Journal of Behavioral Development*, *46*(2), 169–176. https://doi.org/10.1177/01650254221075034

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9), Article 100804. https://doi.org/10.1016/j.patter.2023.100804

Kuhn & Max. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Kuhn, M., & Johnson, K. (2013). An introduction to feature selection. In *Applied predictive modeling* (pp. 487–519). Springer New York. https://doi.org/10.1007/978-1-4614-6849-3_19

Lavelle-Hill, R., Smith, G., & Murayama, K. (2025). Bridging traditional statistics and machine-learning approaches in psychology: Navigating small samples, measurement error, non-independent observations and missing data. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.31219/osf.io/6xt82_v3

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *51*(3), 485–504. https://doi.org/10.1002/ejsp.2752

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd). Lawrence Erlbaum Associates Publishers.

Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., Dinu, A., & Stachl, C. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports*, *13*(1), Article 5705. https://doi.org/10.1038/s41598-023-32484-w

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods*, *5*(4), 434–458. https://doi.org/10.1037/1082-989X.5.4.434

McClure, K., Ammerman, B. A., & Jacobucci, R. (2024). On the selection of item scores or composite scores for clinical prediction. *Multivariate Behavioral Research*, *59*(3), 566–583. https://doi.org/10.1080/00273171.2023.2292598

McNamara, M. E., Zisser, M., Beevers, C. G., & Shumake, J. (2022). Not just "big" data: Importance of sample size, measurement error, and uninformative predictors for developing prognostic models for digital interventions. *Behaviour Research and Therapy*, *153*, Article 104086. https://doi.org/10.1016/j.brat.2022.104086

Mõttus, R., & Rozgonjuk, D. (2021). Development is in the details: Age differences in the big five domains, facets, and nuances. *Journal of Personality and Social Psychology*, *120*(4), 1035–1048. https://doi.org/10.1037/pspp0000276

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

Nuijten, M. B., Deserno, M. K., Cramer, A. O., & Borsboom, D. (2016). Mental disorders as complex networks: An introduction and overview of a network approach to psychopathology. *Clinical Neuropsychiatry*, *13*(4/5), 68–76.

Pargent, F., & Albert-von der Gönna, J. (2018). Predictive modeling with psychological panel datapredictive modeling with psychological panel data. *Zeitschrift Für Psychologie*, *226*(4), 246–258. https://doi.org/10.1027/2151-2604/a000343

Pargent, F., Schoedel, R., & Stachl, C. (2023). Best practices in supervised machine learning: A tutorial for psychologists. *Advances in Methods and Practices in Psychological Science*, *6*(3), Article 25152459231162559. https://doi.org/10.1177/25152459231162559

Platt, J. R. (1964). Strong Inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science*, *146*(3642), 347–353. https://doi.org/10.1126/science.146.3642.347

Primbs, M. A., Pennington, C. R., Lakens, D., Silan, M. A. A., Lieck, D. S., Forscher, P. S., Buchanan, E. M., & Westwood, S. J. (2023). Are small effects the indispensable foundation for a cumulative psychological science? a reply to götz et al.(2022). *Perspectives on Psychological Science*, *18*(2), 508–512. https://doi.org/10.1177/17456916221100420

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, *9*(3), Article e1301. https://doi.org/10.1002/widm.1301

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rao, C. R. (1992). Information and the Accuracy Attainable in the Estimation of Statistical Parameters. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics: Foundations and Basic Theory* (pp. 235–247). Springer. https://doi.org/10.1007/978-1-4612-0919-5_16

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in cognitive sciences*, *20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45. https://doi.org/10.1037/met0000220

Ridgeway, G., & Developers, G. (2024). *Gbm: Generalized boosted regression models* [R package version 2.2.2]. https://CRAN.R-project.org/package=gbm

Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell Jr, F. E., Moons, K. G., & Collins, G. S. (2019a). Minimum sample size for developing a multivariable prediction model: Part i–continuous outcomes. *Statistics in medicine*, *38*(7), 1262–1275. https://doi.org/10.1002/sim.7993

Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell Jr, F. E., Moons, K. G., & Collins, G. S. (2019b). Minimum sample size for developing a multivariable prediction model: Part ii-binary and time-to-event outcomes. *Statistics in medicine*, *38*(7), 1276–1296. https://doi.org/10.1002/sim.7992

Riley, R. D., Snell, K. I., Martin, G. P., Whittle, R., Archer, L., Sperrin, M., & Collins, G. S. (2021). Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *Journal of Clinical Epidemiology*, *132*, 88–96. https://doi.org/10.1016/j.jclinepi.2020.12.005

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*(1), 1–12. https://doi.org/10.1037/a0018326

Schroeders, U., Watrin, L., & Wilhelm, O. (2021). Age-related nuances in knowledge assessment. *Intelligence*, *85*, Article 101526. https://doi.org/10.1016/j.intell.2021.101526

Schwartz, B., Cohen, Z. D., Rubel, J. A., Zimmermann, D., Wittmann, W. W., & Lutz, W. (2021). Personalized treatment selection in routine care: Integrating machine learning and statistical algorithms to recommend cognitive behavioral or psychodynamic therapy. *Psychotherapy Research*, *31*(1), 33–51. https://doi.org/10.1080/10503307.2020.1769219

Seeboth, A., & Mõttus, R. (2018). Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate predictions. *European Journal of Personality*, *32*(3), 186–201. https://doi.org/10.1002/per.2147

Seroussi, I., & Zeitouni, O. (2022). Lower Bounds on the Generalization Error of Nonlinear Learning Models. https://doi.org/10.48550/arXiv.2103.14723

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781107298019

Shalev-Shwartz, S., Shamir, O., Srebro, N., & Sridharan, K. (2010). Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, *11*, 2635–2670. https://jmlr.org/papers/v11/shalev-shwartz10a.html

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Simonsohn, U. (2024). Interacting with curves: How to validly test and probe interactions in the real (nonlinear) world. *Advances in Methods and Practices in Psychological Science*, *7*(1), Article 25152459231207787. https://doi.org/10.1177/25152459231207787

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2024). *Afex: Analysis of factorial experiments* [R package version 1.3-1]. https://CRAN.R-project.org/package=afex

Sommet, N., Weissman, D. L., Cheutin, N., & Elliot, A. J. (2023). How many participants do i need to test an interaction? conducting an appropriate power analysis and achieving sufficient power to detect an interaction. *Advances in Methods and Practices in Psychological Science*, *6*(3), Article 25152459231178728. https://doi.org/10.1177/25152459231178728

Sørensen, Ø., Frigessi, A., & Thoresen, M. (2015). Measurement error in lasso: Impact and likelihood bias correction. *Statistica Sinica*, *25*(2), 809–829. https://www.jstor.org/stable/24311046

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101. https://www.jstor.org/stable/1412159

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., Völkel, S. T., Schuwerk, T., Oldemeier, M., Ullmann, T., et al. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, *117*(30), 17680–17687. https://doi.org/10.1073/pnas.1920484117

Strobl, C., Henninger, M., Rothacher, Y., & Debelak, R. (2024). *Simulationsstudien in R. Design und praktische Durchführung*. Springer. https://doi.org/10.1007/978-3-662-70561-2

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Tay, J. K., Narasimhan, B., & Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, *106*(1), 1–31. https://doi.org/10.18637/jss.v106.i01

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(3), 273–282. https://doi.org/https://doi.org/10.1111/j.1467-9868.2011.00771.x

Tosteson, T. D., Buzas, J. S., Demidenko, E., & Karagas, M. (2003). Power and sample size calculations for generalized regression models with covariate measurement error. *Statistics in Medicine*, *22*(7), 1069–1082. https://doi.org/10.1002/sim.1388

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PloS one*, *14*(11), Article e0224365. https://doi.org/10.1371/journal.pone.0224365

Van Calster, B., van Smeden, M., De Cock, B., & Steyerberg, E. W. (2020). Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. *Statistical methods in medical research*, *29*(11), 3166–3178. https://doi.org/10.1177/0962280220921415

Van Der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC medical research methodology*, *14*, 1–13. http://www.biomedcentral.com/1471-2288/14/137

Viering, T., & Loog, M. (2023). The shape of learning curves: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(6), 7799–7819. https://doi.org/10.1109/TPAMI.2022.3220744

Vinayak, R. K., & Gilad-Bachrach, R. (2015). DART: Dropouts meet Multiple Additive Regression Trees. *Artificial Intelligence and Statistics*, *38*, 489–497. https://proceedings.mlr.press/v38/korlakaivinayak15.html

Vize, C. E., Sharpe, B. M., Miller, J. D., Lynam, D. R., & Soto, C. J. (2023). Do the big five personality traits interact to predict life outcomes? systematically testing the prevalence, nature, and effect size of trait-by-trait moderation. *European Journal of Personality*, *37*(5), 605–625. https://doi.org/10.1177/08902070221111857

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on psychological science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2018). Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. *Journal of Child Psychology and Psychiatry*, *59*(12), 1261–1270. https://doi.org/10.1111/jcpp.12916

Wang, G. (2019). Machine learning for inferring animal behavior from location and movement data. *Ecological informatics*, *49*, 69–76. https://doi.org/10.1016/j.ecoinf.2018.12.002

Wolfrath, N., Wolfrath, J., Hu, H., Banerjee, A., & Kothari, A. N. (2024). Stronger baseline models – a key requirement for aligning machine learning research with clinical utility. https://arxiv.org/abs/2409.12116

Wolpert, D., & Macready, W. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. https://doi.org/10.1109/4235.585893

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zhou, S., & Mentch, L. (2023). Trees, forests, chickens, and eggs: When and why to prune trees in a random forest. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *16*(1), 45–64. https://doi.org/10.1002/sam.11594

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Model Setup

Without loss of generality, let the predictor variables $\{x_1, x_2, x_3, x_4\}$ be standardized with zero means, unit variance and correlations ρ. We assume that the predictor variables come from a multivariate normal distribution, that is:

$$(x_{1i}, x_{2i}, x_{3i}, x_{4i}) \sim \mathcal{N}(0, \Sigma) \quad \text{where} \quad \Sigma = \begin{pmatrix} 1 & \rho & \rho & \rho \\ & 1 & \rho & \rho \\ & & 1 & \rho \\ & & & 1 \end{pmatrix} \tag{16}$$

Let y denote the outcome variable related to the predictor variables $\{x_1, x_2, x_3, x_4\}$ in accordance with the following model:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \tag{17}$$

$$\beta_5 x_{1i} x_{2i} + \beta_6 x_{1i} x_{3i} + \beta_7 x_{1i} x_{4i} + \beta_8 x_{2i} x_{3i} + \beta_9 x_{2i} x_{4i} + \beta_{10} x_{3i} x_{4i} + \tag{18}$$

$$\beta_{11} x_{1i}^2 + \beta_{12} x_{2i}^2 + \beta_{13} x_{3i}^2 + \beta_{14} x_{4i}^2 + \varepsilon_i \tag{19}$$

$$\text{with} \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{20}$$

In matrix notation, we can write:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 I) \tag{21}$$

where X is the matrix of predictor values and β is the vector of regression weights.

We model effects for the (linear) predictors and an identical number of interactions between predictors. Thus, the regression weights for two of the possible interactions (i.e., $x_{1i}x_{3i}$ and $x_{2i}x_{4i}$) need to be 0 ($\{\beta_6, \beta_9\} = 0$). At the moment we simulate without polynomials, i.e., $\{\beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}\} = 0$.

Covariance matrix of predictors

Using the results from of Bohrnstedt and Goldberger (1969) and assuming multivariate normality of the predictor variables $\{x_1, x_2, x_3, x_4\}$, it can be shown that the covariance matrix

of the predictors is given by

$$\Sigma_X = \begin{pmatrix} 1 & \rho & \rho & \rho & 0 & 0 & 0 & 0 & 0 & 0 \\ & 1 & \rho & \rho & 0 & 0 & 0 & 0 & 0 & 0 \\ & & 1 & \rho & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & 1+\rho^2 & \rho+\rho^2 & \rho+\rho^2 & \rho+\rho^2 & \rho+\rho^2 & 2\rho^2 \\ & & & & & 1+\rho^2 & \rho+\rho^2 & \rho+\rho^2 & 2\rho^2 & \rho+\rho^2 \\ & & & & & & 1+\rho^2 & 2\rho^2 & \rho+\rho^2 & \rho+\rho^2 \\ & & & & & & & 1+\rho^2 & \rho+\rho^2 & \rho+\rho^2 \\ & & & & & & & & 1+\rho^2 & \rho+\rho^2 \\ & & & & & & & & & 1+\rho^2 \end{pmatrix} \tag{22}$$

It should be noted that this matrix has a block diagonal structure:

$$\Sigma_X = \begin{pmatrix} V_1 & \mathbf{0} \\ \mathbf{0} & V2 \end{pmatrix} \tag{23}$$

Therefore, linear effects and interaction effects as simulated within a linear regression framework are independent from each other.

The correlation matrix of the predictors is given by:

$$R_X = \begin{pmatrix} 1 & \rho & \rho & \rho & 0 & 0 & 0 & 0 & 0 & 0 \\ & 1 & \rho & \rho & 0 & 0 & 0 & 0 & 0 & 0 \\ & & 1 & \rho & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & 1 & \frac{\rho+\rho^2}{1+\rho^2} & \frac{\rho+\rho^2}{1+\rho^2} & \frac{\rho+\rho^2}{1+\rho^2} & \frac{\rho+\rho^2}{1+\rho^2} & \frac{2\rho^2}{1+\rho^2} \\ & & & & & 1 & \frac{\rho+\rho^2}{1+\rho^2} & \frac{\rho+\rho^2}{1+\rho^2} & \frac{2\rho^2}{1+\rho^2} & \frac{\rho+\rho^2}{1+\rho^2} \\ & & & & & & 1 & \frac{2\rho^2}{1+\rho^2} & \frac{\rho+\rho^2}{1+\rho^2} & \frac{\rho+\rho^2}{1+\rho^2} \\ & & & & & & & 1 & \frac{\rho+\rho^2}{1+\rho^2} & \frac{\rho+\rho^2}{1+\rho^2} \\ & & & & & & & & 1 & \frac{\rho+\rho^2}{1+\rho^2} \\ & & & & & & & & & 1 \end{pmatrix} \tag{24}$$

Note: $\sigma^2_{x_1x_2} = 1 + \rho^2$ and $\sigma^2_{x_1x_3} = 1 + \rho^2$, thus $\sigma_{x_1x_2}\sigma_{x_1x_3} = 1 + \rho^2$ instead of

$\sigma_{x_1x_2}\sigma_{x_1x_3} = \sqrt{1+\rho^2}\sqrt{1+\rho^2}$.

Starting from the results of Bohrnstedt and Goldberger (1969) and assuming bivariate normality of x and y, the variance of the product of x and y is.

$$
\begin{aligned}
V(xy) =& E^2(x)V(y) + E^2(y)V(x) + 2E(x)E(y)C(x,y) + V(x)V(y) + C^2(x,y) \\
\sigma^2_{xy} =& (\mu_x\sigma_y)^2 + (\mu_y\sigma_x)^2 + 2\mu_x\mu_y\rho\sigma_x\sigma_y + \sigma^2_x\sigma^2_y + (\rho\sigma_x\sigma_y)^2 \\
=& (\mu_x\sigma_y)^2 + (\mu_y\sigma_x)^2 + 2\mu_x\mu_y\rho\sigma_x\sigma_y + (\sigma_x\sigma_y)^2 + \rho^2(\sigma_x\sigma_y)^2 \\
=& (\mu_x\sigma_y)^2 + (\mu_y\sigma_x)^2 + 2\mu_x\mu_y\rho\sigma_x\sigma_y + (1+\rho^2)(\sigma_x\sigma_y)^2
\end{aligned}
$$

For standardized variables with zero means ($\mu_x = \mu_y = 0$)and unit variance ($\sigma_x = \sigma_y = 1$):

$$
\begin{aligned}
\sigma^2_{xy} =& (\mu_x\sigma_y)^2 + (\mu_y\sigma_x)^2 + 2\mu_x\mu_y\rho\sigma_x\sigma_y + (1+\rho^2)(\sigma_x\sigma_y)^2 \\
=& (0\cdot 1)^2 + (0\cdot 1)^2 + 2\cdot 0\cdot 0\cdot\rho\cdot 1\cdot 1 + (1+\rho^2)(1\cdot 1)^2 \\
=& (1+\rho^2)
\end{aligned}
$$

The covariance of products of random variables from a multivariate normal distribution is denoted by:

$$
\begin{aligned}
C(xy,uv) =& E(x)E(u)C(y,v) + E(x)E(v)C(y,v) + E(y)E(u)C(x,v) + E(y)E(v)C(x,u) + \\
& C(x,u)C(y,v) + C(x,v)C(y,u) \\
=& \mu_x\mu_u\rho_{yv}\sigma_y\sigma_v + \mu_x\mu_v\rho_{yv}\sigma_y\sigma_v + \mu_y\mu_u\rho_{xv}\sigma_x\sigma_v + \mu_y\mu_v\rho_{xu}\sigma_x\sigma_u + \\
& \rho_{xu}\sigma_x\sigma_u\cdot\rho_{yv}\sigma_y\sigma_v + \rho_{xv}\sigma_x\sigma_v\cdot\rho_{yu}\sigma_y\sigma_u
\end{aligned}
$$

For standardized variables with zero means ($\mu_x = \mu_y = \mu_u = \mu_v = 0$) and unit variance

($\sigma_x = \sigma_y = \sigma_u = \sigma_v = 1$):

$$
\begin{aligned}
C(xy, uv) =& \mu_x\mu_u\rho_{yv}\sigma_y\sigma_v + \mu_x\mu_v\rho_{yv}\sigma_y\sigma_v + \mu_y\mu_u\rho_{xv}\sigma_x\sigma_v + \mu_y\mu_v\rho_{xu}\sigma_x\sigma_u + \\
& \rho_{xu}\sigma_x\sigma_u \cdot \rho_{yv}\sigma_y\sigma_v + \rho_{xv}\sigma_x\sigma_v \cdot \rho_{yu}\sigma_y\sigma_u \\
=& 0 \cdot 0 \cdot \rho_{yv} \cdot 1 \cdot 1 + 0 \cdot 0 \cdot \rho_{yv} \cdot 1 \cdot 1 + 0 \cdot 0 \cdot \rho_{xv} \cdot 1 \cdot 1 + 0 \cdot 0 \cdot \rho_{xu} \cdot 1 \cdot 1 + \\
& \rho_{xu} \cdot 1 \cdot 1 \cdot \rho_{yv} \cdot 1 \cdot 1 + \rho_{xv} \cdot 1 \cdot 1 \cdot \rho_{yu} \cdot 1 \cdot 1 \\
=& \rho_{xu} \cdot \rho_{yv} + \rho_{xv} \cdot \rho_{yu}
\end{aligned}
$$

For the present simulation model, we need to differentiate two cases which are represented by the following two examples:

- The products share one of the variables.

$$
\begin{aligned}
C(xy, uv) =& \rho_{xu} \cdot \rho_{yv} + \rho_{xv} \cdot \rho_{yu} \\
C(x_1x_2, x_1x_3) =& \rho_{x_1x_1} \cdot \rho_{x_2x_3} + \rho_{x_1x_3} \cdot \rho_{x_2x_1} \\
C(x_1x_2, x_1x_3) =& 1 \cdot \rho + \rho \cdot \rho \\
C(x_1x_2, x_1x_3) =& \rho + \rho^2
\end{aligned}
$$

- The products do not share any variable.

$$
\begin{aligned}
C(xy, uv) =& \rho_{xu} \cdot \rho_{yv} + \rho_{xv} \cdot \rho_{yu} \\
C(x_1x_2, x_3x_4) =& \rho_{x_1x_3} \cdot \rho_{x_2x_4} + \rho_{x_1x_4} \cdot \rho_{x_2x_3} \\
C(x_1x_2, x_1x_3) =& \rho \cdot \rho + \rho \cdot \rho \\
C(x_1x_2, x_1x_3) =& 2\rho^2
\end{aligned}
$$

Piecewise-linear DGP: second segment variable variance

We sample the original predictor variable from $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = 0$ and $\sigma^2 = 1$. To arrive at the predictor variable for the second segment, we transform X into X^* with

$$x_i^* = (x_i - \psi) \cdot dummy_{x_i} \text{ with } dummy_{x_{pi}} = \begin{cases} 0, & \text{if } x_{pi} \leq \psi \\ 1, & \text{if } x_{pi} > \psi \end{cases}$$

For $\psi = 0$, the given transformation is equivalent to applying the rectified linear unit (ReLU) activation function to the original predictor (or using a rectified Gaussian distribution to sample from).

$$\begin{aligned} x_i^* &= (x_i - 0) \cdot dummy_{x_i} \\ &= max(0, x_i) \end{aligned}$$

Without loss of generality, we can use the probability density function (PDF) of the standard normal distribution $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ and the corresponding cumulative distribution function (CDF) $\Phi(x) = \int_{-\infty}^{x} \phi(t)\, dt$ to calculate the variance of X^*.

The variance of a variable is given by

$$\text{Var}(X^*) \;=\; E\big[(X^*)^2\big] \;-\; \big(E[X^*]\big)^2$$

The expected value of a ReLU transformed variable X^* is given by

$$E[X^*] \;=\; \mu\,\big[1 - \Phi(-\frac{\mu}{\sigma})\big] \;+\; \sigma\,\phi(-\frac{\mu}{\sigma})$$

and its second moment with

$$E\big[(X^*)^2\big] \;=\; (\mu^2 + \sigma^2)\,\big[1 - \Phi(-\frac{\mu}{\sigma})\big] \;+\; \mu\,\sigma\,\phi(-\frac{\mu}{\sigma})$$

with $-\frac{\mu}{\sigma}$ essentially giving the z-score of the breakpoint (Beauchamp, 2018).

By substituting the respective expressions with the simulated true values of the original variable X, we obtain the variance of the variable X^*.