Prior Knowledge on Features: Should You Listen to Your Experts?

Bjarne C. Hiller Martin Becker

2026

Biomedical prediction tasks often have far more features than samples, which leaves standard regression models underdetermined, so that they fit the training data but fail on new data. Prior knowledge about which features matter, for example importance scores provided by domain experts, promises to restrict the hypothesis space through *adaptive shrinkage*: features believed to be relevant are penalized less. Many methods implement this idea, but each was evaluated by its authors on their own data and protocol. We review these methods and compare 15 of them, using their reference implementations, in an open and reproducible benchmark on six real datasets with prior knowledge and on synthetic data with a prior of controlled quality. On the real datasets, no method gains significantly from the prior compared with a permuted prior, so whether the prior knowledge helps cannot be verified with 28 to 133 samples. On synthetic data, where the prior is known to be informative, methods that use the scores directly profit strongly, most of all in high dimensions, whereas methods that bin the scores into groups barely use them. Experts who scored the features of two clinical datasets agree on the ranking of the features but show distinct response styles, and some response styles, such as acquiescent or random ratings, remove most of the benefit of an informative prior. Neither single experts nor their disagreement as a measure of uncertainty improve on the experts' median. Integrating the reference implementations required a number of workarounds, which we document to support their reproducible use. Taken together, our results cast doubt on whether adaptive shrinkage guided by prior knowledge on features can be recommended in practice, unless its benefit is verified against a permuted prior.

1. Introduction

Standard linear models fail when there are more features than samples. Such models are underdetermined: infinitely many hypotheses fit the training data equally well, but

most of them predict new data poorly. This setting is the rule rather than the exception in biomedicine, where the data stem from small patient cohorts but describe a high-dimensional biological state, such as thousands of single-cell signalling responses [3], transcripts [52] or methylation sites [27].

Prior knowledge can restrict the hypothesis space. Informed Machine Learning (iML) [50] integrates knowledge from an independent source into the learning pipeline, so that hypotheses that agree with the knowledge are preferred. For linear and logistic regression, the most common entry point is the regularization of the coefficients. Penalizing large coefficients during training shrinks them towards zero, hence the name *shrinkage* [46, 65]. If prior knowledge on the importance of each feature is available, the strength of the penalty can be set per feature (*adaptive shrinkage*): the coefficients of presumably irrelevant features are kept small, while presumably relevant features may take large coefficients. A prominent example is the immunological Elastic Net (iEN) [16], in which five immunologists scored the expected relevance of each of the 792 to 960 mass-cytometry features of two clinical datasets, and the scores guide an elastic net.

The methods have not been compared on common ground. Since the adaptive lasso [64], numerous methods have been proposed that differ in the kind of knowledge they accept (feature scores, groups, several sources), in how the knowledge enters the penalty, and in how its influence is estimated (cross-validation, empirical Bayes or full Bayes). Each was evaluated by its authors, on their own datasets and with their own protocol, typically against an uninformed baseline only, a setting in which new methods tend to look favourable [9, 54]. Whether the reported gains stem from the knowledge, and how robust they are to imperfect knowledge, such as the subjective ratings of human experts, remains unclear.

Contributions.

- We derive adaptive shrinkage from regularized regression, distinguish the types of prior knowledge on features and the ways in which the methods estimate the influence of the prior, and give an overview of 31 methods (section 2).
- We present `pkbench`, an open R benchmark that wraps 15 such methods, using their reference implementations, as `parsnip` models and compares them with nested cross-validation, shared folds, permuted-prior controls and corrected significance tests on six real and two synthetic tasks (section 3). We document the workarounds that the integration of the reference implementations required.
- In four experiments, we study when the prior helps: depending on its informativeness, on the number of samples and features, on the response styles of the experts who rate the features, and on the use of multiple experts as a measure of uncertainty (section 4).
- From the results, we derive recommendations for practitioners (section 5).

2. Prior Knowledge Guided Shrinkage

2.1. From Regression to Adaptive Shrinkage

Regression is underdetermined if there are more features than samples. We consider n samples with features $\mathbf{x}_i \in \mathbb{R}^p$ and outcomes y_i, collected in $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{y}$. A generalized linear model, such as linear or logistic regression, relates the outcome to the linear predictor $\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}$, and maximum likelihood estimates the coefficients as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta_0, \boldsymbol{\beta}} \; -\ell(\beta_0, \boldsymbol{\beta}; \mathbf{X}, \mathbf{y}), \tag{1}$$

where ℓ is the log-likelihood. For the linear model, this is least squares with the solution $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ for centred data. If $p > n$, however, $\mathbf{X}^\top \mathbf{X}$ is singular: infinitely many coefficient vectors fit the training data perfectly, and most of them predict new data poorly.

Shrinkage penalizes large coefficients. Regularization adds a penalty on the size of the coefficients to (1),

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta_0, \boldsymbol{\beta}} \; -\ell(\beta_0, \boldsymbol{\beta}; \mathbf{X}, \mathbf{y}) + \lambda \sum_{j=1}^{p} \left(\alpha |\beta_j| + \tfrac{1-\alpha}{2} \beta_j^2 \right), \tag{2}$$

with the ridge penalty for $\alpha = 0$ [21], the lasso for $\alpha = 1$, which also selects features by setting coefficients exactly to zero [46], and the elastic net (EN) in between [65]. The penalty shrinks the coefficients towards zero and trades bias for a lower variance; with a quadratic part ($\alpha < 1$), the solution is unique even if $p > n$. A single penalty λ, usually chosen by cross-validation [18], shrinks all coefficients alike.

In the Bayesian view, the penalty is a prior. The penalty in (2) is the negative logarithm of a prior distribution of the coefficients, and the penalized estimate is the mode of their posterior distribution. The ridge penalty corresponds to independent normal priors $\beta_j \sim \mathcal{N}(0, \tau^2)$ with a prior variance τ^2 inversely proportional to λ, and the lasso penalty to Laplace priors [38]. Other priors express sparsity more directly, such as the spike-and-slab prior, a mixture of a point mass at zero and a wide distribution [33, 19], or the horseshoe prior [13]. In each case, the strength of the shrinkage is a hyperparameter of the prior: the narrower the prior, the more strongly a coefficient is pulled towards zero.

Adaptive shrinkage uses feature-specific penalties. Adaptive shrinkage replaces the single penalty by feature-specific penalties λw_j,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta_0, \boldsymbol{\beta}} \; -\ell(\beta_0, \boldsymbol{\beta}; \mathbf{X}, \mathbf{y}) + \lambda \sum_{j=1}^{p} w_j \left(\alpha |\beta_j| + \tfrac{1-\alpha}{2} \beta_j^2 \right), \qquad w_j = f(s_j; \boldsymbol{\theta}), \tag{3}$$

which reduces to the EN for $w_j \equiv 1$. In the Bayesian view, each coefficient receives its own prior; for the ridge penalty, its variance is proportional to $1/w_j$. The weights can be derived from the data: the adaptive lasso and EN set $w_j = |\hat{\beta}_j^{(0)}|^{-\gamma}$ from a pilot estimate [64, 66]. We are interested in weights derived from prior knowledge, i.e. knowledge from a source independent of the training data [50]: a score $s_j \in [0, 1]$ of the expected importance of feature j, where larger means more important, and a mapping f, decreasing in s_j, whose parameters $\boldsymbol{\theta}$ determine the influence of the prior. Prior knowledge on features is thus prior knowledge in the Bayesian sense: it shapes the prior of each coefficient. Wiel and Wieringen [56] recently reviewed such methods as *guided adaptive shrinkage*; what we call prior knowledge on features, e.g. the relevance of each feature, is called *co-data* there, for complementary data on the features, a term used in this setting since Wiel et al. [57]. Many of the methods discussed below were developed by van de Wiel and colleagues [57, 34, 37, 36, 55, 11, 40].

2.2. Types of Prior Knowledge on Features

Prior knowledge on features comes from different sources. Domain experts can rate the expected relevance of each feature [16]; external studies provide p-values or effect estimates of the features for a related outcome [57, 2]; databases provide annotations such as pathways or gene sets; multi-omics data come in blocks of data types [10, 25]; and, recently, large language models have been used to score features [63]. Depending on its source, the knowledge may be wrong, imprecise or uncertain, e.g. when several experts disagree.

The knowledge differs in its granularity. We distinguish five types, which structure our literature overview (Figure 1 and Table 1): (a) *Structure only*: groups or graphs of features without a statement about their importance, used by the group lasso [61], its sparse and latent variants [42, 22] and network-constrained penalties [29]. Such knowledge determines which coefficients are shrunk together, but not how strongly. (b) *Data-driven weights* use no external knowledge but the same mechanism as the methods below, as in the adaptive lasso and EN [64, 66] and the adaptive group lasso [51]. (c) *Importance of individual features*: either one score per feature, mapped to a penalty by a fixed function with a tuned strength [5, 16, 63], or a matrix $\mathbf{Z} \in \mathbb{R}^{p \times q}$ of meta-features with a learned mapping, e.g. $w_j = \exp(\mathbf{z}_j^\top \boldsymbol{\theta})$ [62] or a learned feature weighting [45]. (d) *Importance of groups*: a partition of the features into groups or blocks with group-specific penalties or priors [57, 34, 36, 49, 10, 25]. (e) *Multiple sources* of type (c) or (d) at once, each with a learned weight [37, 11, 40].

The knowledge can set the scale or the mean of the prior. Most methods use the knowledge to set the scale of the prior, i.e. how strongly a coefficient is shrunk, which only requires a statement about its importance. iEN, for example, rescales each feature by $\exp(-(1-s_j)\,\varphi)$, which is equivalent to penalties that grow exponentially as the score decreases [16]. Other methods use the direction of the effects: hierarchical regression

[59] and xrnet [53] shift the mean of the prior towards a linear function of the meta-features, and transreg [40] calibrates and combines signed prior effect estimates. Such knowledge must state the direction of an effect, which importance scores do not provide. Finally, the methods differ in whether the influence of the knowledge is fixed up to a single strength parameter or learned in detail from the data, which allows a method to ignore an uninformative source, but requires more hyperparameters.

2.3. Estimating the Hyperparameters

The influence of the prior, $\boldsymbol{\theta}$, is estimated together with the global penalty λ and the mixing α; below, we write $\boldsymbol{\theta}$ for all hyperparameters. The methods use three approaches [56].

Cross-validation (CV) splits the training data into K folds, fits the model for each candidate value of the hyperparameters on $K-1$ folds, and selects the value with the lowest loss L on the held-out folds [43],

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{k=1}^{K} L\big(\mathbf{y}_{(k)}, \hat{\mathbf{y}}_{(k)}(\boldsymbol{\theta})\big), \tag{4}$$

where $\hat{\mathbf{y}}_{(k)}(\boldsymbol{\theta})$ are the predictions for fold k of the model fitted without it. CV makes no assumptions about the model and targets predictive performance directly. However, it searches a grid whose size grows exponentially with the number of hyperparameters, so that in practice only a few are tuned, such as λ, α and one prior strength (iEN) or the penalty ratios of a few blocks (IPF-lasso). With few samples, the CV estimate is noisy, and the performance of the selected model must be estimated in a further, outer loop to avoid an optimistic bias [48]. fwelnet is a hybrid: it tunes λ by CV and learns $\boldsymbol{\theta}$ jointly with the coefficients [45].

Empirical Bayes (EB) treats the penalties as parameters of the prior and estimates them from the data by maximizing the marginal likelihood, in which the coefficients are integrated out [14, 31],

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \int p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta})\, \pi(\boldsymbol{\beta} \mid \boldsymbol{\theta})\, d\boldsymbol{\beta}. \tag{5}$$

For the linear model with normal priors $\beta_j \sim \mathcal{N}(0, \tau_j^2(\boldsymbol{\theta}))$, the integral has a closed form, $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{X}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\mathbf{X}^\top + \sigma^2\mathbf{I})$ with $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \operatorname{diag}(\tau_1^2, \ldots, \tau_p^2)$; otherwise, it is approximated, e.g. by a variational approximation [34], or replaced by moment estimators [57, 37]. The model is then fitted with $\hat{\boldsymbol{\theta}}$ plugged in. Because a numerical optimization replaces the grid search, EB can estimate one hyperparameter per group or per source of prior knowledge at low cost, as in GRridge, gren, squeezy, ecpc and xtune. It ignores the uncertainty in $\hat{\boldsymbol{\theta}}$, however, and can overfit if there are many hyperparameters relative to the number of samples; ecpc therefore regularizes their estimation [37].

Full Bayes also places a prior on the hyperparameters and infers their joint posterior with the coefficients,

$$p(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}) \propto p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta})\, \pi(\boldsymbol{\beta} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}), \tag{6}$$

so that predictions average over the uncertainty in $\boldsymbol{\theta}$ instead of conditioning on a single estimate. The posterior is sampled by Markov chain Monte Carlo (MCMC), e.g. Gibbs sampling [19], which is exact in the limit but slow in high dimensions, or approximated by variational Bayes (VB), which fits a factorized distribution by optimization [7]; VB scales to thousands of features but tends to underestimate the posterior uncertainty. Fully Bayesian methods for prior knowledge on features include graper and varbvs (VB) [49, 12], iBMU (MCMC) [39] and infHS (both) [11]. In short, CV is the most direct but least scalable approach, EB scales to many hyperparameters at low cost, and full Bayes accounts for their uncertainty at the highest cost.

2.4. Overview of Methods

In a literature review, we identified 27 methods for adaptive shrinkage guided by prior knowledge, or for its building blocks (Figure 1; details in Figure 6). The field has moved from fixed penalty factors tuned by CV towards learning the influence of the knowledge from the data, often by EB, and towards multiple sources of prior knowledge. Table 1 lists these methods together with the EN and four further methods of our benchmark, and whether and why each is included in the benchmark (section 3). Most methods are distributed as R packages, several only on GitHub or in the CRAN archive.

3. Benchmark

Methods. `pkbench` includes 15 methods that use prior knowledge on features, and the EN and the adaptive EN as baselines without it (Table 1). We selected methods with a working R implementation that accept feature scores, either directly, as a matrix of meta-features, or as groups derived from the scores, and included two non-linear methods based on random forests. The prior-weighted EN and RF are simple reference points of our own that pass transformed scores to the penalty factors of `glmnet` [18] and to the split-variable weights of `ranger` [60]. Each method is implemented as a `parsnip` model [26] whose engine is the package with the reference implementation, so that all methods share one interface and data format. For iEN and Know-GRRF, we additionally re-implemented the method on `glmnet` and `RRF`, respectively, to tune it on the same inner folds as the other methods; the benchmark runs both engines. The prior is translated for each method at fit time: scores are rescaled to $[0, 1]$, methods that learn the mapping f receive the score as a column of $\mathbf{Z}$, and group methods receive five quantile groups of the scores (ties are never split). We checked in the code of every method, including the reference packages, that a larger score never leads to a stronger penalty.

Table 1: Methods for adaptive shrinkage guided by prior knowledge (Figure 1) and further methods of our benchmark, by kind of knowledge (subsection 2.2). Penalty: grp-L2 group norm, S&S spike-and-slab, HS horseshoe, RF random forest; tuning: FB full Bayes, IC information criterion; s_j: score of feature j; $\mathbf{Z}$: matrix of prior knowledge. Benchmark: ✓ included (package), otherwise the reason for exclusion.

Knowledge	Method	Penalty	Tuning	How the prior enters	Benchmark
(a) structure only	Group lasso [61]	grp-L2	CV	groups shrunk jointly	no importance
	Latent group lasso [22]	grp-L2	CV	overlapping groups shrunk jointly	no importance
	Network-constrained [29]	L1+graph	CV	coefficients smoothed over a graph	no importance
	Sparse group lasso [42]	L1+grp-L2	CV	group and feature penalty factors from s_j	✓ sparsegl [30]
(b) data-driven	Elastic net [65]	EN	CV	–	✓ glmnet [18] (baseline)
	Adaptive lasso [64]	L1	CV	w_j from a pilot estimate	covered by adaptive EN
	Adaptive EN [66]	EN	CV	w_j from a ridge pilot estimate	✓ glmnet (baseline)
	Adaptive group lasso [51]	grp-L2	IC	group weights from a pilot estimate	no prior knowledge
(c) feature-level	Hierarchical regression [59]	L2	EB	prior mean $\mathbf{Z}\boldsymbol{\theta}$	predecessor of xrnet
	Weighted lasso [5]	L1	CV	penalty factors from external data	covered by prior-weighted EN
	Prior-weighted EN (ours)	EN	CV	$w_j = (s_j + \varepsilon)^{-\gamma}$	✓ glmnet
	varbvs [12]	S&S	VB	prior inclusion log-odds linear in s_j	✓ varbvs
	iBMU [39]	S&S	MCMC	inclusion probabilities from $\mathbf{Z}$	model search; package archived
	prior LASSO [23]	L1	CV	fit to predictions of a prior model	needs prior effects or a variable set
	xrnet [53]	L1/L2	CV	coefficients shrunk towards $\mathbf{Z}\boldsymbol{\theta}$	✓ xrnet
	iEN [16]	EN	CV	features rescaled by $\exp(-(1 - s_j)\varphi)$	✓ iEN; glmnet
	xtune [62]	L1/EN	EB	$w_j = \exp(\mathbf{z}_j^\top \boldsymbol{\theta})$	✓ xtune
	fwelnet [45]	EN	CV	w_j a learned function of $\mathbf{Z}$	✓ fwelnet
	LLM-Lasso [63]	L1	CV	penalty factors v_j^η from LLM scores	same model as prior-weighted EN
(d) group-level	PAM & PPLS [44]	grp-wise	CV	group-specific penalties	no software found
	GRridge [57]	L2	EB	group-specific ridge penalties	✓ GRridge
	IPF-lasso [10]	L1	CV	block-specific penalty factors	✓ ipflasso
	priority-lasso [25]	L1	CV	blocks fitted in order of priority	uses only an order of blocks
	graper [49]	S&S/L2	VB	group-specific priors	✓ graper
	gren [34]	EN	EB	group-specific EN penalties	✓ gren (classification)
	multiridge [55]	L2	CV/EB	one ridge penalty per data type	group ridge, as GRridge
	squeezy [36]	L2→EN	EB	group-specific EN penalties	✓ squeezy
(e) multiple sources	ecpc [37]	L2	EB	prior variances from several sources	✓ ecpc
	transreg [40]	L1/L2	CV	calibrated prior effects, stacked	needs signed prior effects
	infHS [11]	HS	FB	horseshoe scales regressed on $\mathbf{Z}$	could not be run
non-linear	Prior-weighted RF (ours)	RF	CV	split variables sampled $\propto (s_j + \varepsilon)^\gamma$	✓ ranger [60]
	Know-GRRF [20, 17]	RF	AIC	regularization coefficients $\propto s_j^\gamma$	✓ KnowGRRF; RRF

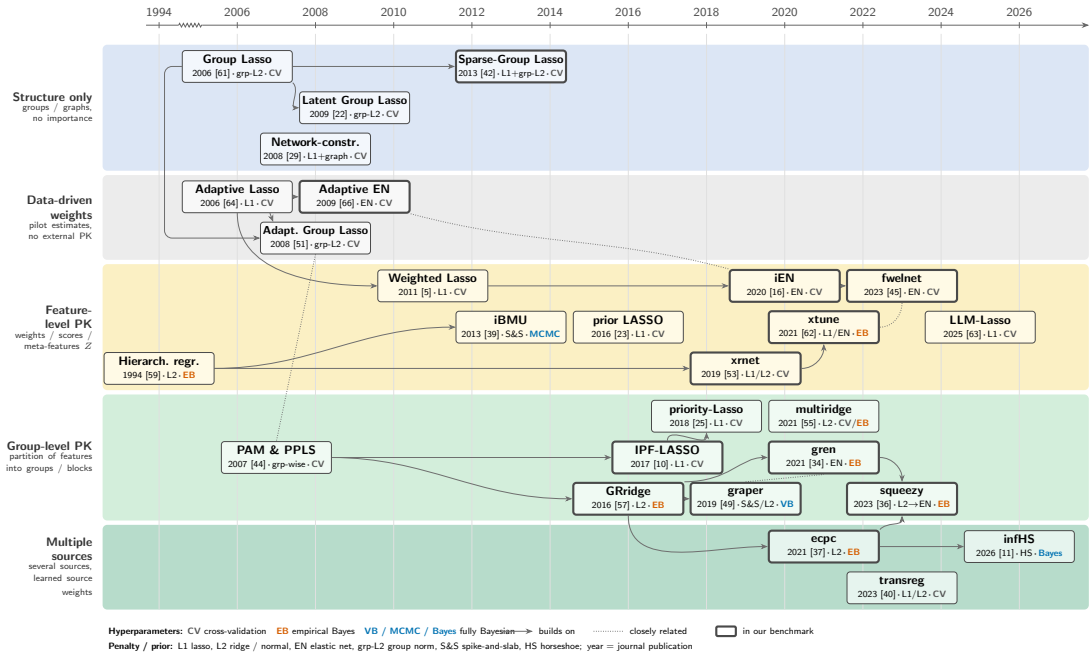


Figure 1: Methods for adaptive shrinkage guided by prior knowledge, by the kind of knowledge (lanes) and year of publication. Each box lists the penalty or prior and how the hyperparameters are estimated. Arrows: a method builds on another. Bold frame: included in our benchmark. A detailed version with short descriptions is given in Figure 6.

Tasks. The benchmark has six real tasks with prior knowledge and two synthetic ones (Table 2). Two clinical mass-cytometry datasets [16] come with scores of five immunologists for every feature; the prior is the median of their scores. The four other real tasks are the example datasets of the GRridge, CoRF and gren packages, whose prior comes from external studies. The synthetic tasks have Gaussian features in correlated blocks of 10 ($\rho = 0.5$) and 20 features with an effect of random sign; the outcome is linear or logistic. Five simulated experts score every feature as $q \cdot o_j + (1 - q) \cdot u_j$, a mix of the oracle score o_j (0.5 to 1 by effect size for the 20 features with an effect, 0 otherwise) and uniform noise u_j, with qualities $q \in \{0.9, 0.8, 0.7, 0.5, 0.7\}$; the fifth expert reports cubed scores.

Evaluation protocol. We use 10 repeats of 5-fold cross-validation (CV); the folds are stratified by the outcome, and grouped by patient for LTP. Each learner is tuned on the four training folds only, by inner 5-fold CV on folds shared by all learners or by its own EB or Bayesian procedure, and refit on all training folds. Within a repeat, the out-of-fold predictions are pooled into one metric, Spearman's ρ for regression and the ROC AUC for classification, because test folds of five to six samples give unstable per-fold metrics. As a control, every learner is also fitted with a *permuted* prior, which

Table 2: Tasks of the benchmark. The prior score is larger for features expected to be more important.

Task	n	p	Outcome	Prior knowledge
LTP [16]	54	960	gestational age	median of 5 immunologists (18 women $\times$ 3 time points)
CHP [16]	28	792	periodontitis	median of 5 immunologists
verlaat [57]	44	9 491	cervical lesions	$-\log_{10} p$ of an external methylation study
LNM [2]	133	12 838	lymph node metastasis	$-\log_{10} p$ of an external expression study
colon [34]	88	2 114	therapy benefit	expression level in an external study (3 levels)
wurdinger [6, 57]	81	18 410	breast vs. colorectal	platelet-expressed gene list (0/1)
synthetic	100	1 000	linear; binary	median of 5 simulated experts of known quality

keeps the distribution of the scores but destroys their link to the features; a learner uses the prior meaningfully only if the true prior beats the permuted one. Differences are tested one-sided with the corrected repeated k-fold CV t-test [35, 8] on the 50 paired fold differences, and p-values are adjusted over the learners of a task [4]. We call a difference significant if the adjusted $p < 0.05$; because the corrected test is conservative for small n, we report differences that are significant only by a Wilcoxon signed-rank test [58] on the 10 repeats as *trends*.

Implementation. The benchmark is an R package with a `targets` pipeline [28] and package versions pinned by `renv`. All learners and prior conditions share the outer splits, inner folds and permutations, and all randomness is seeded per task and split. The main benchmark and the four experiments comprise about 73 000 fits and 1 500 CPU hours. The code will be made available upon publication.

Integrating research software. Bringing 15 research packages into one framework required a number of workarounds, which we list with their effect in Table 6. They fall into a few groups. *Availability*: five packages are distributed only on GitHub and three only in the CRAN archive, and some could not be installed in the standard way. *Interfaces*: three packages (ecpc, squeezy, IPF-lasso) do not accept predefined folds and draw their own inner folds, and some prediction functions failed in our setting, so we predict from the fitted coefficients. *Numerical robustness*: with few samples, some estimation procedures failed sporadically, e.g. the EB estimation of xtune for unstandardized outcomes, which we standardize. *Runtime*: the reference implementation of iEN rescales the features with a dense $p \times p$ matrix and needed 146 CPU hours in the main benchmark, our `glmnet` engine, which fits the same model for a fixed configuration, 1.2 hours. *Settings that change the fitted model*: in a few cases, the result depended on settings such as the order of the fold ids or the start of the λ path; for example, the default λ path of

`sparsegl` starts too low under strong penalty factors, which can yield empty models, so we compute the path from the exact $\lambda_{\max}$. The informative horseshoe [11] could not be run in our environment and is therefore not included.

4. Experiments and Results

4.1. Does the prior help on real data?

On the real tasks, the gain from the prior cannot be verified. Figure 2 compares each learner with the true and the permuted prior, and with the EN without prior (all values in Figure 7 and Table 5). On none of the six real tasks does any learner gain significantly from the true over the permuted prior (smallest adjusted $p = 0.06$, on colon). There are trends: on colon, 15 of 17 learners with a prior gain 0.05 to 0.23 AUC, and on CHP the prior-weighted EN, iEN (`glmnet` engine) and IPF-lasso gain 0.09 to 0.13. On LTP, however, the permuted prior gives higher means than the true prior for most methods that use the scores directly. No learner is significantly worse than the best one on any real task, so with 28 to 133 samples the benchmark cannot rank the methods on real data. The original studies of these datasets reported improvements from the prior with other protocols and baselines [16, 57, 2, 34]; our results do not contradict them, but show that gains of this size cannot be separated from the variability of the evaluation with so few samples.

With an informative prior, methods that use the scores directly win clearly. On the synthetic tasks, the prior-weighted EN, both iEN engines, ecpc and the sparse group lasso reach $\rho = 0.52$ to 0.57 and AUC 0.87 to 0.89, compared with 0.08 and 0.64 for the EN without prior; their gain over the permuted prior is significant (adjusted $p \leq 0.002$). In regression, Know-GRRF, xtune, the prior-weighted RF, varbvs and fwelnet gain significantly as well, in classification only as a trend. xrnet does not profit on any task: it models the mean of the coefficients as a function of the scores, but importance scores carry the size and not the sign of an effect (subsection 2.2). The group methods GRridge, gren, squeezy and graper barely profit on the synthetic tasks (see subsection 4.2). The reference and re-implemented engines of iEN and Know-GRRF do not differ significantly on any task.

4.2. How informative must the prior be?

To control the quality of the prior, we mix the oracle scores of the synthetic tasks with random scores, from informativeness 0 (random) to 1 (oracle), and evaluate with 3×3-fold CV (Figure 3; classification: Figure 8). The methods that use the scores directly improve steadily and reach most of the oracle's benefit at informativeness 0.5 (prior-weighted EN: $\rho = 0.50$ at 0.5 and 0.55 with the oracle). The group methods stay flat up to 0.75 and jump only with the oracle (GRridge: $\rho = 0.11$ at 0.75, 0.49 with the oracle). The reason is the conversion of the scores into five quantile groups of 200 features: unless the prior is perfect, the 20 features with an effect share the top group with at least 180

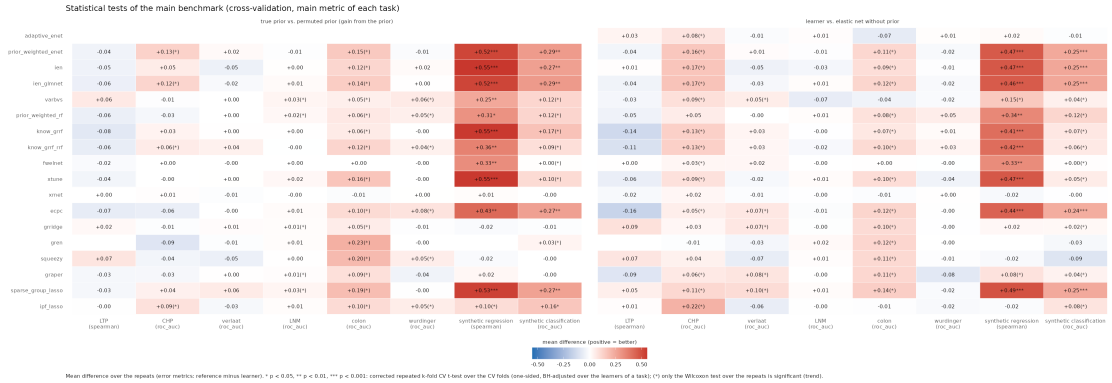


Figure 2: Main benchmark: mean difference in the main metric (Spearman ρ for LTP and synthetic regression, ROC AUC otherwise) over 10 repeats of 5-fold CV. Left: true minus permuted prior (gain from the prior); right: each learner minus the EN without prior. Stars: corrected repeated k-fold CV t-test, one-sided, BH-adjusted (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$); (*): significant only by the Wilcoxon test over the repeats (trend).

features without one, whereas the oracle puts all 980 features without an effect into one group. Group methods therefore need groups that reflect the knowledge, not the quantiles of a score.

4.3. When does the prior pay off?

We vary the number of samples $n \in \{50, 100, 300\}$ and features $p \in \{100, 1000, 5000\}$ in nested subsets of one simulated population, so that the cells differ only in n and p, and evaluate every CV model on an independent test set of 2 000 samples (Figure 4; classification: Figure 9).

- *The prior makes the dimension irrelevant.* With the true prior, the prior-weighted EN, iEN, ecpc and the sparse group lasso reach the same performance at $p = 100$ and $p = 5000$ (regression, $n = 300$: $\rho = 0.65$ to 0.66), whereas the EN without prior drops from 0.59 to 0.25, so the gain from the prior grows with p.

- *With few samples, the prior decides.* At $n = 50$ and $p \geq 1000$, the EN without prior is at chance level ($\rho \approx 0.04$, AUC 0.51), while ecpc and the sparse group lasso reach $\rho = 0.29$ to 0.35 and AUC 0.80 to 0.81; ecpc, which estimates the prior strength by EB, has the highest mean in most cells with $n = 50$.

- *Group methods profit only in low dimensions.* At $p = 100$ they are competitive, but at $p \geq 1000$ and $n \leq 100$ GRridge, gren, graper and squeezy barely use the prior, for the reason given in subsection 4.2.

- *An uninformative prior is mostly harmless.* With a permuted or random prior, the linear methods that use the scores directly mostly tune the prior away and

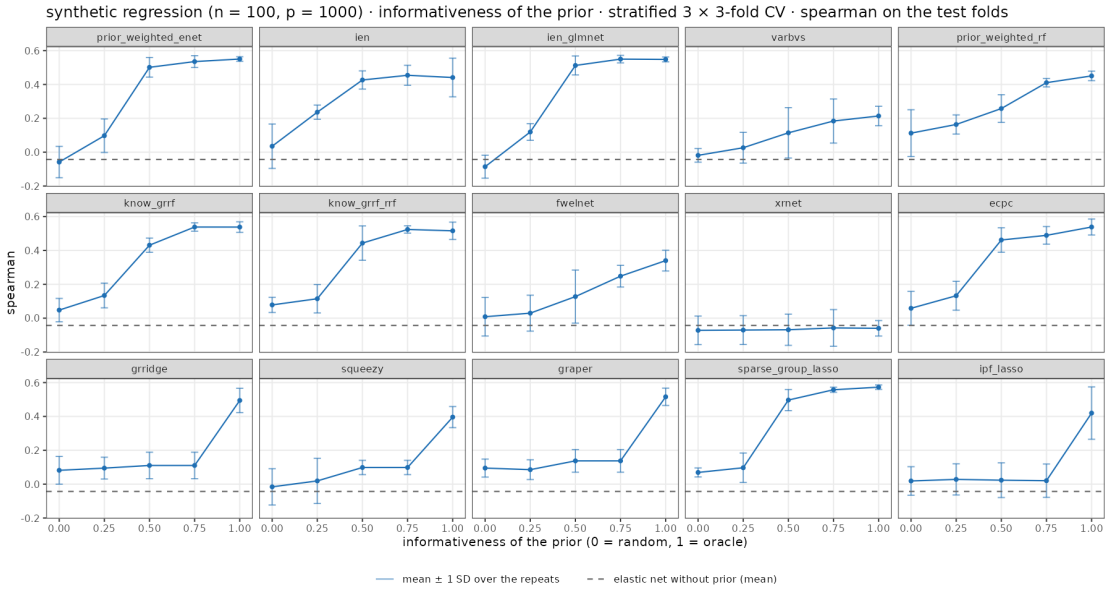


Figure 3: Informativeness of the prior (synthetic regression, $n = 100$, $p = 1000$): Spearman ρ on the test folds when the prior scores are mixed between random scores (0) and the oracle (1); mean ± 1 SD over 3 repeats of 3-fold CV. Dashed: EN without prior.

perform like the EN without prior; Know-GRRF, however, loses on average 0.14 (up to 0.40) compared with the EN.

The `glmnet`-based methods need at most 5 s per fit including tuning even at $n = 300$, $p = 5000$, Know-GRRF up to 52 minutes.

4.4. Do expert response styles matter?

Raters differ in how they use a rating scale. Prior scores elicited from experts are ratings, and ratings are subject to the biases known from survey research. Respondents differ in their *response styles*, systematic tendencies to use a rating scale independently of the content of the items [1, 47]: acquiescent respondents tend towards the upper end of the scale, disacquiescent respondents towards the lower end, extreme respondents towards the end points and midpoint respondents towards the middle, mild respondents avoid the end points, and noncontingent respondents answer independently of the item (Table 3). Response styles distort the scale of the ratings and, partly, the ranking of the items.

The experts agree on the ranking, but not on the scale. The five immunologists who scored the features of the two mass-cytometry tasks agree moderately on the ranking of the features (Kendall's $W = 0.62$ for LTP, 0.63 for CHP [24]), but much less on the

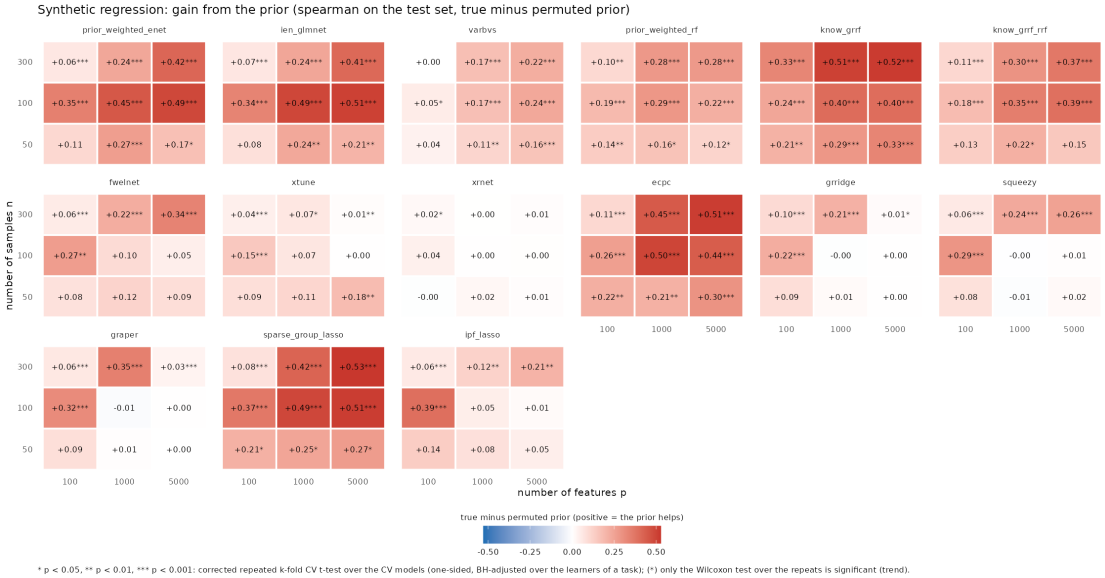


Figure 4: Gain from the prior (true minus permuted prior, Spearman ρ on an independent test set) by the number of samples n and features p, synthetic regression; mean over 15 CV models. Stars as in Figure 2.

values: the intraclass correlation for absolute agreement (0.36, 0.32) is far below that for consistency (0.55, 0.47) [32]. Their score distributions show distinct response styles (Figure 5): expert EG is *disacquiescent* (75–84% of the scores ≤ 0.1), IS is *acquiescent* (mean score 0.70–0.74), and XH is an *extreme* responder (89–92% of the scores ≤ 0.1 or ≥ 0.9), while AT and BG use the whole scale.

We simulate response styles as perturbations of the prior. Since the real experts come without a ground truth, we took a closer look by rating the prior in each response style ourselves (Table 3; illustrated in Figure 11). The strength x of a style is the fraction of the features rated in it, from 0 (the prior) to 1 (all features); the features are taken in the order in which the style affects them, e.g. the highest-scored ones first for acquiescence. In addition, an unreliable rater mixes all scores with uniform noise of weight x (signal-to-noise ratio 1 at $x = 0.5$). We evaluate all learners with the perturbed priors of the synthetic tasks, CHP, LTP and colon by 3 repeats of 5-fold CV and compare them with the unperturbed prior.

Some response styles destroy an informative prior. Table 4 summarizes the results for synthetic regression (all styles and tasks: Figures 12 and 13):

- Noncontingent, acquiescent and midpoint ratings of half of the features cost the score-based methods 0.32 to 0.52 in ρ on synthetic regression (0.17 to 0.28 AUC in classification), because they destroy the ranking among the important features. Disacquiescent ratings, which only threshold unimportant features, cost nothing.

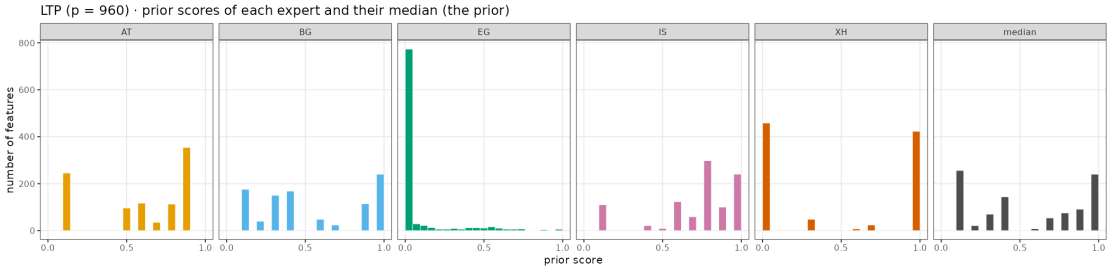


Figure 5: Score distributions of the five immunologists for the 960 features of LTP and their median, the prior of the benchmark (CHP: Figure 10). EG is disacquiescent, IS acquiescent and XH an extreme responder.

Table 3: Response styles [1, 47] and how we simulate them: the perturbation applied to a fraction x of the features. Scores are on the observed scale of the prior.

Style	Rater tends to	Perturbation of a fraction x of the features
Noncontingent	answer independently of the item	random features get uniform random scores
Acquiescent	use the upper end	the highest scores are set to the maximum
Disacquiescent	use the lower end	the lowest scores are set to the minimum
Extreme	use the end points	the scores farthest from the midpoint move to the nearest end ($x = 1$: binary scores)
Midpoint	use the middle	the scores closest to the midpoint are set to it
Mild	avoid the end points	the most extreme scores are clamped to the inner half of the scale
Noise	rate with random error	all scores are mixed with uniform noise of weight x

- Noise is tolerated up to a signal-to-noise ratio of 1, but pure noise costs as much as random scores.
- Styles that create large blocks of tied scores (extreme, mild) *help* the group methods (+0.46 to +0.52 in ρ with binary scores), because the score groups then separate the features with and without an effect.
- On CHP, the same pattern appears as a trend (prior-weighted EN and iEN: −0.10 to −0.11 AUC at noise 0.5); on LTP, where the prior does not help, no style hurts.

4.5. Can multiple experts express uncertainty?

Expert disagreement could tell a model where to trust the prior and where the data. We therefore gave the prior-weighted EN and the methods that learn the mapping of the prior (fwelnet, xtune, xrnet, ecpc) each single expert, one column of $\mathbf{Z}$ per expert, or the experts' mean and standard deviation as two columns, and compared them with the median (Figure 14). No setting is significantly better than the median for any method

Table 4: Response styles on synthetic regression: change of Spearman's ρ when 50% of the features are rated in a style (extreme: 100%, i.e. binary scores; noise: weight of the noise), compared with the unperturbed prior; mean over 3 repeats of 5-fold CV. Stars: loss significant (corrected t-test, one-sided, BH-adjusted; gains were not tested).

	Synthetic regression (change of ρ)						
Learner	random	acq.	disacq.	midp.	extreme	noise .5	noise 1
Prior-weighted EN	-0.52^{**}	-0.50^{**}	0.00	-0.32^{*}	-0.01	-0.02	-0.59^{***}
iEN (glmnet)	-0.52^{***}	-0.45^{**}	0.00	-0.32^{*}	-0.01	-0.01	-0.58^{***}
ecpc	-0.37^{**}	-0.50^{**}	0.00	-0.41^{*}	$+0.07$	0.00	-0.50^{***}
Sparse group lasso	-0.45^{**}	-0.46^{**}	$+0.01$	-0.33^{*}	$+0.02$	-0.02	-0.60^{***}
Know-GRRF	-0.36^{*}	-0.36^{*}	0.00	-0.27	-0.02	-0.02	-0.46^{**}
fwelnet	-0.37^{**}	-0.29^{*}	$+0.02$	-0.27^{*}	$+0.10$	-0.12	-0.44^{***}
xtune	-0.16^{*}	-0.16^{*}	-0.04	-0.15^{*}	0.00	-0.08	-0.16^{*}
GRridge	-0.02	-0.01	-0.01	$+0.06$	$+0.51$	$+0.07$	-0.02
graper	-0.06	-0.08	$+0.07$	$+0.15$	$+0.46$	$+0.06$	-0.05
IPF-lasso	-0.07	-0.05	$+0.05$	$+0.14$	$+0.49$	$+0.03$	-0.08

and task (smallest adjusted $p = 0.38$). On CHP, the extreme expert XH (AUC 0.63) and the acquiescent IS (0.62) give slightly higher means than the median (0.61) for the prior-weighted EN, and the disacquiescent EG the lowest (0.45); on LTP, all settings of the prior-weighted EN lie within $\rho = 0.49$ to 0.53, around the EN without prior (0.52). On the synthetic tasks, the expert who reports cubed scores costs fwelnet and xtune 0.13 in ρ, but hardly affects the prior-weighted EN and ecpc: methods that learn a parametric function of the prior depend on its scale.

5. Discussion

Should you listen to your experts? Prior knowledge on features, whether elicited from experts or taken from external studies, did not lead to verifiably better predictions on the real datasets of our benchmark. This does not prove that the experts were wrong: with 28 to 133 samples, even the gains of 0.1 to 0.2 AUC observed on CHP and colon are within the uncertainty of the evaluation, and in high dimensions the correctness of an importance score can hardly be checked otherwise. The synthetic experiments show what is at stake: an informative prior turns a model at chance level into a useful one when samples are few and features many, but only if the method uses the scores directly, and only if the scores preserve the ranking of the important features. Human experts introduce a further difficulty that is known from survey research: their response styles differ, so that the number of features they deem relevant and the weight they give them vary considerably, even when they agree on the ranking. Finally, the methods are distributed as research software whose integration required a number of workarounds, so that applying them correctly takes more effort than their interfaces suggest. Taken

together, this casts doubt on whether adaptive shrinkage guided by expert knowledge on features can be recommended in practice, unless its benefit is verified for the task at hand.

Recommendations for practitioners.

- *Always run a permuted-prior control.* A gain over an uninformed model can come from the method rather than from the knowledge; only a gain over the same method with permuted scores shows that the knowledge is used.
- *Use the scores directly.* The prior-weighted EN, iEN, ecpc and the sparse group lasso use graded scores best and mostly tune an uninformative prior away. The prior-weighted EN takes a few lines of `glmnet` code and performs on par with iEN in all our experiments.
- *Groups must carry the knowledge.* Group methods work if the groups separate relevant from irrelevant features, e.g. thresholds or curated feature sets, but not with quantiles of a score.
- *Preserve the ranking when eliciting and aggregating scores.* Acquiescent ratings, which set many features to the maximum, and random ratings remove the ranking among the important features. The median of five experts was not significantly beaten by any single expert, and their disagreement did not help as a measure of uncertainty.
- *Validate the method in your setting.* Test a method with an oracle and a random prior before trusting it, e.g. on simulated data of the same size, and prefer implementations that accept predefined folds, so that tuning respects the structure of the data.

Limitations. The real datasets are small, so the benchmark has little power to detect moderate gains or to rank the methods on real data. The synthetic tasks are linear and sparse, which favours the linear methods, and the simulated experts differ in quality but not in response style. Group methods received quantile groups of the scores, which is not their intended use; with groups from curated knowledge, they may perform better. Three packages draw their own inner folds, and two-stage methods (adaptive EN, gren) tune optimistically on folds that already influenced their weights.

Outlook. The difficulties of eliciting subjective importance scores shift the focus towards knowledge extracted from data. Foundation models such as scGPT [15] or UCE [41] encode prior knowledge as large pretraining corpora instead of feature scores, and large language models have been proposed as a source of feature scores [63]; `pkbench` can evaluate such priors with the same controls.

6. Conclusion

We reviewed methods for adaptive shrinkage guided by prior knowledge on features and compared 15 of them in `pkbench`, an open benchmark built on their reference implementations. Prior knowledge on features pays off when it is informative and used directly as scores, and most when samples are few and features many. On real biomedical data with expert scores, however, the benefit could not be verified, and the experts' response styles differ markedly. This casts doubt on recommending adaptive shrinkage with expert knowledge in practice without a control: listening to your experts can help, but only a permuted-prior control tells whether it did.

Disclosure of AI use. Generative AI was used for creating this report. Specifically, the employed AI model was Claude Opus 5.5 by Anthropic. The experiments were initially designed by a human author, while AI was used for assistance during implementation. Similarly, the initial structure of the manuscript and a draft were first provided by a human author, before AI was used for formulation and generation of TikZ figures. This report is still work in progress, and review of AI artifacts is currently ongoing.

Acknowledgments

This work was funded by the German Federal Ministry of Research, Technology and Space (BMFTR) within the junior research group THEMIS (grant numbers 01IS22077 and 16IS22077).

References

[1] Hans Baumgartner and Jan-Benedict E. M. Steenkamp. "Response Styles in Marketing Research: A Cross-National Investigation". In: *Journal of Marketing Research* 38.2 (2001), pp. 143–156. DOI: 10.1509/jmkr.38.2.143.18840.

[2] Dennis E. te Beest et al. "Improved High-Dimensional Prediction with Random Forests by the Use of Co-Data". In: *BMC Bioinformatics* 18 (2017), p. 584. DOI: 10.1186/s12859-017-1993-1.

[3] Sean C. Bendall et al. "Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum". In: *Science* 332.6030 (2011), pp. 687–696. DOI: 10.1126/science.1198704.

[4] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (1995), pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

[5] Linn Cecilie Bergersen, Ingrid K. Glad, and Heidi Lyng. "Weighted lasso with data integration". In: *Statistical Applications in Genetics and Molecular Biology* 10.1 (Aug. 29, 2011), /j/sagmb.2011.10.issue–1/sagmb.2011.10.1.1703/sagmb.2011.10.1.1703.xml. ISSN: 1544-6115. DOI: 10.2202/1544-6115.1703.

[6] Myron G. Best et al. "RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics". In: *Cancer Cell* 28.5 (2015), pp. 666–676. DOI: 10.1016/j.ccell.2015.09.018.

[7] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877. DOI: 10.1080/01621459.2017.1285773.

[8] Remco R. Bouckaert and Eibe Frank. "Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms". In: *Advances in Knowledge Discovery and Data Mining (PAKDD 2004)*. Vol. 3056. Lecture Notes in Computer Science. Springer, 2004, pp. 3–12. DOI: 10.1007/978-3-540-24775-3_3.

[9] Anne-Laure Boulesteix, Sabine Lauer, and Manuel J. A. Eugster. "A Plea for Neutral Comparison Studies in Computational Sciences". In: *PLoS ONE* 8.4 (2013), e61562. DOI: 10.1371/journal.pone.0061562.

[10] Anne-Laure Boulesteix et al. "IPF-LASSO: Integrative L1-Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data". In: *Computational and Mathematical Methods in Medicine* 2017 (2017), p. 7691937. ISSN: 1748-6718. DOI: 10.1155/2017/7691937.

[11] Claudio Busatto and Mark A. van de Wiel. "Informative Co-Data Learning for High-Dimensional Horseshoe Regression". In: *Biometrical Journal* 68.1 (2026), e70105. DOI: 10.1002/bimj.70105.

[12] Peter Carbonetto and Matthew Stephens. "Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies". In: *Bayesian Analysis* 7.1 (2012), pp. 73–108. DOI: 10.1214/12-BA703.

[13] C. M. Carvalho, N. G. Polson, and J. G. Scott. "The horseshoe estimator for sparse signals". In: *Biometrika* 97.2 (2010), pp. 465–480. DOI: 10.1093/biomet/asq017.

[14] George Casella. "An Introduction to Empirical Bayes Data Analysis". In: *The American Statistician* 39.2 (1985), pp. 83–87. DOI: 10.1080/00031305.1985.10479400.

[15] Haotian Cui et al. "scGPT: toward building a foundation model for single-cell multi-omics using generative AI". In: *Nature Methods* 21.8 (Aug. 2024), pp. 1470–1480. ISSN: 1548-7105. DOI: 10.1038/s41592-024-02201-0. URL: https://www.nature.com/articles/s41592-024-02201-0 (visited on 07/15/2026).

[16] Anthony Culos et al. "Integration of mechanistic immunological knowledge into a machine learning pipeline improves predictions". In: *Nature Machine Intelligence* 2.10 (Oct. 2020), pp. 619–628. ISSN: 2522-5839. DOI: 10.1038/s42256-020-00232-8. URL: https://www.nature.com/articles/s42256-020-00232-8 (visited on 01/11/2024).

[17] Houtao Deng and George Runger. "Gene Selection with Guided Regularized Random Forest". In: *Pattern Recognition* 46.12 (2013), pp. 3483–3489. DOI: 10.1016/j.patcog.2013.05.018.

[18] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. DOI: 10.18637/jss.v033.i01.

[19] Edward I. George and Robert E. McCulloch. "Variable Selection via Gibbs Sampling". In: *Journal of the American Statistical Association* 88.423 (1993), pp. 881–889. DOI: 10.1080/01621459.1993.10476353.

[20] Xin Guan and Li Liu. "Know-GRRF: Domain-Knowledge Informed Biomarker Discovery with Random Forests". In: *Bioinformatics and Biomedical Engineering (IWBBIO 2018)*. Ed. by Ignacio Rojas and Francisco Ortuño. Vol. 10814. Lecture Notes in Computer Science. Springer, 2018, pp. 3–14. DOI: 10.1007/978-3-319-78759-6_1.

[21] Arthur E. Hoerl and Robert W. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems". In: *Technometrics* 12.1 (1970), pp. 55–67. DOI: 10.1080/00401706.1970.10488634.

[22] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. "Group lasso with overlap and graph lasso". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009, pp. 433–440. DOI: 10.1145/1553374.1553431.

[23] Yuan Jiang, Yunxiao He, and Heping Zhang. "Variable Selection with Prior Information for Generalized Linear Models via the Prior LASSO Method". In: *Journal of the American Statistical Association* 111.513 (2016), pp. 355–376. ISSN: 0162-1459. DOI: 10.1080/01621459.2015.1008363.

[24] M. G. Kendall and B. Babington Smith. "The Problem of m Rankings". In: *The Annals of Mathematical Statistics* 10.3 (1939), pp. 275–287. DOI: 10.1214/aoms/1177732186.

[25] Simon Klau et al. "Priority-Lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data". In: *BMC bioinformatics* 19.1 (Sept. 12, 2018), p. 322. ISSN: 1471-2105. DOI: 10.1186/s12859-018-2344-6.

[26] Max Kuhn and Hadley Wickham. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles*. 2020. URL: https://www.tidymodels.org.

[27] Peter W. Laird. "Principles and challenges of genome-wide DNA methylation analysis". In: *Nature Reviews Genetics* 11.3 (2010), pp. 191–203. DOI: 10.1038/nrg2732.

[28] William Michael Landau. "The targets R Package: A Dynamic Make-like Function-Oriented Pipeline Toolkit for Reproducibility and High-Performance Computing". In: *Journal of Open Source Software* 6.57 (2021), p. 2959. DOI: 10.21105/joss.02959.

[29] Caiyan Li and Hongzhe Li. "Network-constrained regularization and variable selection for analysis of genomic data". In: *Bioinformatics* 24.9 (2008), pp. 1175–1182. DOI: 10.1093/bioinformatics/btn081.

[30] Xiaoxuan Liang et al. "sparsegl: An R Package for Estimating Sparse Group Lasso". In: *Journal of Statistical Software* 110.6 (2024), pp. 1–23. DOI: 10.18637/jss.v110.i06.

[31] David J. C. MacKay. "Bayesian Interpolation". In: *Neural Computation* 4.3 (1992), pp. 415–447. DOI: 10.1162/neco.1992.4.3.415.

[32] Kenneth O. McGraw and S. P. Wong. "Forming Inferences about Some Intraclass Correlation Coefficients". In: *Psychological Methods* 1.1 (1996), pp. 30–46. DOI: 10.1037/1082-989X.1.1.30.

[33] T. J. Mitchell and J. J. Beauchamp. "Bayesian Variable Selection in Linear Regression". In: *Journal of the American Statistical Association* 83.404 (1988), pp. 1023–1032. DOI: 10.1080/01621459.1988.10478694.

[34] Magnus M. Münch et al. "Adaptive group-regularized logistic elastic net regression". In: *Biostatistics* 22.4 (Oct. 13, 2021), pp. 723–737. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxz062.

[35] Claude Nadeau and Yoshua Bengio. "Inference for the Generalization Error". In: *Machine Learning* 52.3 (2003), pp. 239–281. DOI: 10.1023/A:1024068626366.

[36] Mirrelijn M. van Nee, Tim van de Brug, and Mark A. van de Wiel. "Fast Marginal Likelihood Estimation of Penalties for Group-Adaptive Elastic Net". In: *Journal of Computational and Graphical Statistics* 32.3 (July 3, 2023), pp. 950–960. ISSN: 1061-8600. DOI: 10.1080/10618600.2022.2128809. URL: https://doi.org/10.1080/10618600.2022.2128809 (visited on 10/06/2026).

[37] Mirrelijn M. van Nee, Lodewyk F.A. Wessels, and Mark A. van de Wiel. "Flexible co-data learning for high-dimensional prediction". In: *Statistics in Medicine* 40.26 (2021), pp. 5910–5925. ISSN: 1097-0258. DOI: 10.1002/sim.9162. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9162 (visited on 10/06/2026).

[38] Trevor Park and George Casella. "The Bayesian Lasso". In: *Journal of the American Statistical Association* 103.482 (2008), pp. 681–686. DOI: 10.1198/016214508000000337.

[39] M. A. Quintana and D. V. Conti. "Integrative variable selection via Bayesian model uncertainty". In: *Statistics in Medicine* 32.28 (2013), pp. 4938–4953. ISSN: 1097-0258. DOI: 10.1002/sim.5888. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5888 (visited on 10/06/2026).

[40] Armin Rauschenberger et al. "Penalized regression with multiple sources of prior effects". In: *Bioinformatics* 39.12 (2023), btad680. DOI: 10.1093/bioinformatics/btad680.

[41] Yanay Rosen et al. "Universal cell embedding provides a foundation model for cell biology". In: *Nature* (July 8, 2026), pp. 1–9. ISSN: 1476-4687. DOI: 10.1038/s41586-026-10689-z. URL: https://www.nature.com/articles/s41586-026-10689-z (visited on 07/15/2026).

[42] Noah Simon et al. "A Sparse-Group Lasso". In: *Journal of Computational and Graphical Statistics* 22.2 (2013), pp. 231–245. DOI: 10.1080/10618600.2012.681250.

[43] M. Stone. "Cross-Validatory Choice and Assessment of Statistical Predictions". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2 (1974), pp. 111–133. DOI: 10.1111/j.2517-6161.1974.tb00994.x.

[44] Feng Tai and Wei Pan. "Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms". In: *Bioinformatics* 23.14 (July 15, 2007), pp. 1775–1782. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btm234.

[45] J. Kenneth Tay et al. "Feature-Weighted Elastic Net: Using "Features of Features" for "Better Prediction"". In: *Statistica Sinica* 33.1 (2023), pp. 259–279. ISSN: 1017-0405. URL: https://www.jstor.org/stable/27250039 (visited on 10/06/2026).

[46] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

[47] Y. Van Vaerenbergh and T. D. Thomas. "Response Styles in Survey Research: A Literature Review of Antecedents, Consequences, and Remedies". In: *International Journal of Public Opinion Research* 25.2 (June 1, 2013), pp. 195–217. ISSN: 0954-2892, 1471-6909. DOI: 10.1093/ijpor/eds021. URL: https://academic.oup.com/ijpor/article-lookup/doi/10.1093/ijpor/eds021 (visited on 10/06/2026).

[48] Sudhir Varma and Richard Simon. "Bias in error estimation when using cross-validation for model selection". In: *BMC Bioinformatics* 7.1 (2006), p. 91. DOI: 10.1186/1471-2105-7-91.

[49] Britta Velten and Wolfgang Huber. "Adaptive penalization in high-dimensional regression and classification with external covariates using variational Bayes". In: *Biostatistics* 22.2 (Apr. 1, 2021), pp. 348–364. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxz034. URL: https://doi.org/10.1093/biostatistics/kxz034 (visited on 10/06/2026).

[50] Laura Von Rueden et al. "Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems". In: *IEEE Transactions on Knowledge and Data Engineering* (2021), pp. 1–1. ISSN: 1041-4347, 1558-2191, 2326-3865. DOI: 10.1109/TKDE.2021.3079836. URL: https://ieeexplore.ieee.org/document/9429985/ (visited on 04/25/2024).

[51] Hansheng Wang and Chenlei Leng. "A note on adaptive group lasso". In: *Computational Statistics & Data Analysis* 52.12 (Aug. 15, 2008), pp. 5277–5286. ISSN: 0167-9473. DOI: 10.1016/j.csda.2008.05.006. URL: https://www.sciencedirect.com/science/article/pii/S0167947308002582 (visited on 10/06/2026).

[52] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature Reviews Genetics* 10.1 (2009), pp. 57–63. DOI: 10.1038/nrg2484.

[53] Garrett M. Weaver and Juan Pablo Lewinger. "xrnet: Hierarchical Regularized Regression to Incorporate External Data". In: *Journal of Open Source Software* 4.44 (Dec. 6, 2019), p. 1761. ISSN: 2475-9066. DOI: 10.21105/joss.01761. URL: https://joss.theoj.org/papers/10.21105/joss.01761 (visited on 10/06/2026).

[54] Lukas M. Weber et al. "Essential guidelines for computational method benchmarking". In: *Genome Biology* 20.1 (2019), p. 125. DOI: 10.1186/s13059-019-1738-8.

[55] Mark A. van de Wiel, Mirrelijn M. van Nee, and Armin Rauschenberger. "Fast Cross-validation for Multi-penalty High-dimensional Ridge Regression". In: *Journal of Computational and Graphical Statistics* 30.4 (Oct. 2, 2021), pp. 835–847. ISSN: 1061-8600. DOI: 10.1080/10618600.2021.1904962. URL: https://doi.org/10.1080/10618600.2021.1904962 (visited on 10/06/2026).

[56] Mark A. van de Wiel and Wessel N. van Wieringen. "Leveraging external information by guided adaptive shrinkage to improve variable selection in high-dimensional regression settings". In: *The International Journal of Biostatistics* 21.2 (2025), pp. 271–283. DOI: 10.1515/ijb-2024-0108.

[57] Mark A. van de Wiel et al. "Better prediction by use of co-data: adaptive group-regularized ridge regression". In: *Statistics in Medicine* 35.3 (Feb. 10, 2016), pp. 368–381. ISSN: 1097-0258. DOI: 10.1002/sim.6732.

[58] Frank Wilcoxon. "Individual Comparisons by Ranking Methods". In: *Biometrics Bulletin* 1.6 (1945), pp. 80–83. DOI: 10.2307/3001968.

[59] J. S. Witte et al. "Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer". In: *Epidemiology* 5.6 (Nov. 1994), pp. 612–621. ISSN: 1044-3983. DOI: 10.1097/00001648-199411000-00009.

[60] Marvin N. Wright and Andreas Ziegler. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". In: *Journal of Statistical Software* 77.1 (2017), pp. 1–17. DOI: 10.18637/jss.v077.i01.

[61] Ming Yuan and Yi Lin. "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1 (2006), pp. 49–67. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2005.00532.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00532.x (visited on 10/06/2026).

[62] Chubing Zeng, Duncan Campbell Thomas, and Juan Pablo Lewinger. "Incorporating prior knowledge into regularized regression". In: *Bioinformatics* 37.4 (Feb. 1, 2021), pp. 514–521. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa776. URL: https://doi.org/10.1093/bioinformatics/btaa776 (visited on 10/06/2026).

[63] Erica Zhang et al. *LLM-Lasso: A Robust Framework for Domain-Informed Feature Selection and Regularization*. Issue: arXiv:2502.10648. Aug. 12, 2025. DOI: 10.48550/arXiv.2502.10648. arXiv: 2502.10648[cs.LG]. URL: http://arxiv.org/abs/2502.10648 (visited on 10/06/2026).

[64] Hui Zou. "The Adaptive Lasso and Its Oracle Properties". In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1418–1429. DOI: 10.1198/016214506000000735.

[65] Hui Zou and Trevor Hastie. "Regularization and Variable Selection via the Elastic Net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.

[66] Hui Zou and Hao Helen Zhang. "On the adaptive elastic-net with a diverging number of parameters". In: *The Annals of Statistics* 37.4 (Aug. 1, 2009). ISSN: 0090-5364. DOI: 10.1214/08-AOS625. arXiv: 0908.1836[math.ST]. URL: http://arxiv.org/abs/0908.1836 (visited on 10/06/2026).

A. Supplementary Material

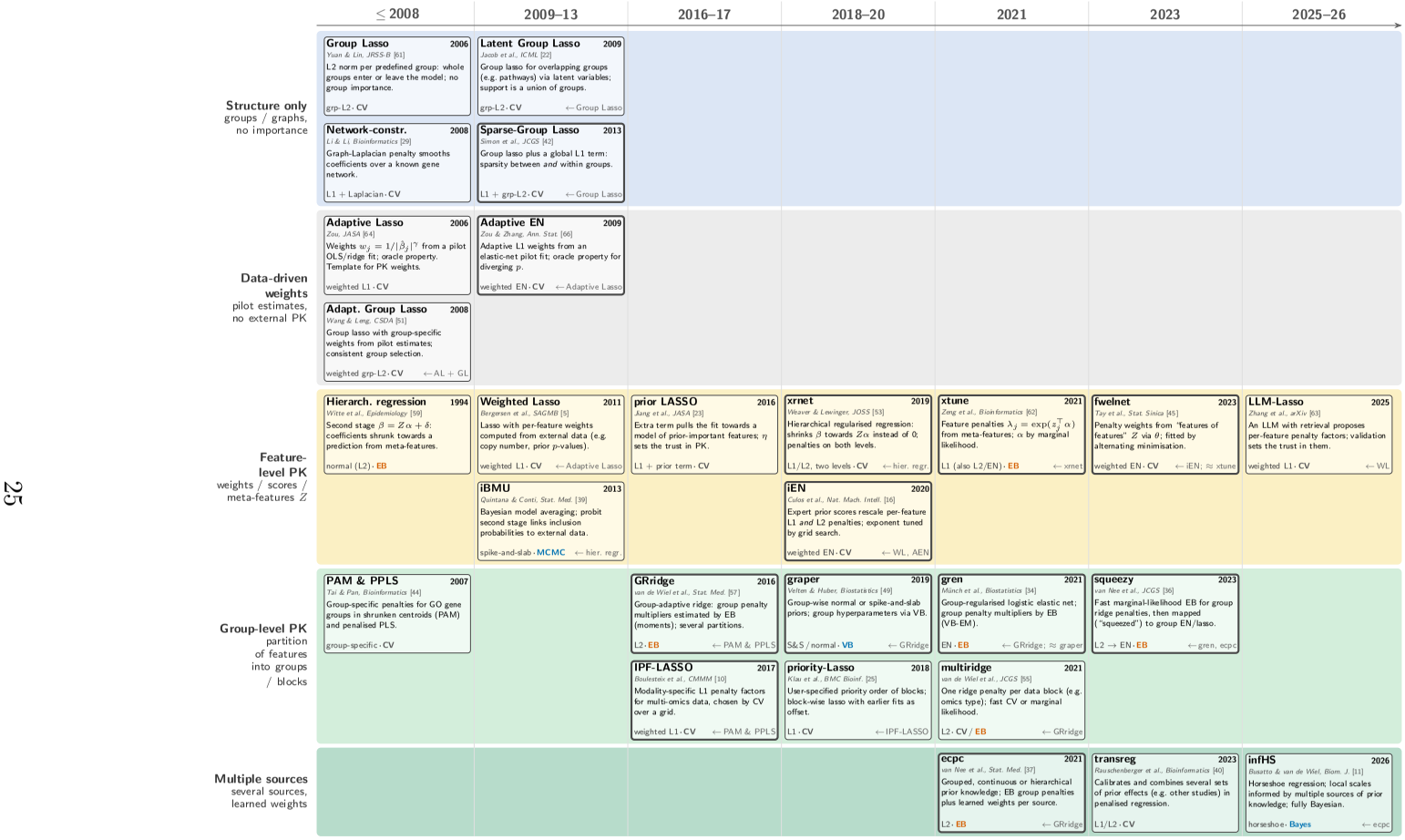


Figure 6: Detailed overview of methods for adaptive shrinkage guided by prior knowledge on features, with a short description of each method (cf. Figure 1).

Table 5: Main benchmark: main metric (LTP and synthetic regression: Spearman ρ; otherwise ROC AUC), mean over 10 repeats of 5-fold CV with the true / the permuted prior. Stars: true prior better than the permuted prior (corrected t-test, one-sided, BH-adjusted; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$); $^{(*)}$: significant only by the Wilcoxon test over the repeats (trend). gren supports classification only.

Learner	LTP	CHP	verlaat	LNM	colon	wurdinger	synth. reg.	synth. class.
EN	0.52	0.45	0.60	0.69	0.60	0.81	0.08	0.64
Adaptive EN	0.55	0.52	0.58	0.70	0.53	0.82	0.09	0.63
Prior-weighted EN	0.49/0.53	0.61/0.48$^{(*)}$	0.61/0.59	0.69/0.69	0.72/0.56$^{(*)}$	0.79/0.80	0.55/0.03***	0.88/0.60**
iEN (iEN)	0.53/0.57	0.61/0.56	0.55/0.60	0.66/0.66	0.70/0.58$^{(*)}$	0.80/0.79	0.55/0.00***	0.89/0.61**
iEN (glmnet)	0.48/0.54	0.62/0.50$^{(*)}$	0.57/0.59	0.70/0.69	0.72/0.58$^{(*)}$	0.79/0.79	0.54/0.02***	0.89/0.60**
varbvs	0.49/0.43	0.54/0.54	0.65/0.65	0.63/0.60$^{(*)}$	0.56/0.52$^{(*)}$	0.79/0.73$^{(*)}$	0.23/$-$0.02**	0.68/0.56$^{(*)}$
Prior-weighted RF	0.48/0.53	0.50/0.53	0.60/0.60	0.70/0.68$^{(*)}$	0.69/0.62$^{(*)}$	0.86/0.82$^{(*)}$	0.42/0.11*	0.75/0.64$^{(*)}$
Know-GRRF (KnowGRRF)	0.39/0.47	0.58/0.55	0.62/0.62	0.69/0.69	0.67/0.61$^{(*)}$	0.82/0.82	0.49/$-$0.05***	0.70/0.53$^{(*)}$
Know-GRRF (RRF)	0.41/0.48	0.58/0.53$^{(*)}$	0.62/0.58	0.67/0.67	0.71/0.59$^{(*)}$	0.84/0.80$^{(*)}$	0.50/0.14**	0.70/0.61$^{(*)}$
fwelnet	0.52/0.54	0.48/0.47	0.62/0.62	0.69/0.69	0.61/0.61	0.81/0.81	0.41/0.08**	0.64/0.64$^{(*)}$
xtune	0.47/0.50	0.53/0.53	0.57/0.57	0.70/0.68	0.70/0.54$^{(*)}$	0.77/0.77	0.55/0.00***	0.69/0.59$^{(*)}$
xrnet	0.50/0.50	0.47/0.47	0.59/0.59	0.69/0.69	0.59/0.60	0.81/0.81	0.06/0.05	0.63/0.63
ecpc	0.37/0.43	0.50/0.56	0.66/0.66	0.68/0.67	0.73/0.62$^{(*)}$	0.81/0.73$^{(*)}$	0.52/0.09**	0.87/0.60**
GRridge	0.61/0.59	0.47/0.48	0.67/0.65	0.69/0.68$^{(*)}$	0.71/0.65$^{(*)}$	0.81/0.81	0.10/0.12	0.66/0.67
gren	–	0.44/0.53	0.57/0.57	0.71/0.70	0.72/0.49$^{(*)}$	0.81/0.81	–	0.61/0.58$^{(*)}$
squeezy	0.60/0.52	0.48/0.51	0.53/0.59	0.70/0.70	0.71/0.51$^{(*)}$	0.80/0.75$^{(*)}$	0.06/0.08	0.54/0.54
graper	0.43/0.46	0.51/0.54	0.67/0.67	0.69/0.68$^{(*)}$	0.72/0.63$^{(*)}$	0.73/0.77	0.16/0.14	0.67/0.67
Sparse group lasso	0.57/0.60	0.56/0.52	0.70/0.64	0.70/0.68$^{(*)}$	0.74/0.56$^{(*)}$	0.79/0.80	0.57/0.04***	0.88/0.61**
IPF-lasso	0.53/0.53	0.67/0.58$^{(*)}$	0.53/0.56	0.69/0.68	0.60/0.50$^{(*)}$	0.79/0.74$^{(*)}$	0.06/$-$0.03$^{(*)}$	0.71/0.55*

Table 6: Issues of the method packages that we encountered while integrating them into `pkbench`, their effect on the benchmark, and our workaround. Package versions as pinned in the repository, where further details are documented.

Package	Issue	Effect	Workaround
iEN	Hyperparameters tuned per fold are stored by the order in which the fold ids first appear, but read by their value.	Unless the fold ids appear in order, fold models use hyperparameters tuned on another fold (198 of 200 random fold assignments).	Fold ids recoded by first appearance.
iEN	Features rescaled with a dense $p \times p$ matrix; one `glmnet` call per λ.	Up to 8 h and 4.3 GB per fit on the large tasks.	Re-implementation on `glmnet` (same model).
iEN	Standardizes the features on all samples before its CV; Spearman criterion not usable; grids of length 1 fail.	Test folds would enter the standardization.	Standardization on the training data; RMSE or ROC AUC as criterion.
sparsegl	The default λ path ignores the penalty factors and stores a zero solution for its first value.	With strong prior weights, empty models can be selected.	Explicit λ path from the exact $\lambda_{\max}$.
xtune	EB estimation depends on the scale of the outcome and on a randomized noise estimate; nearly duplicated columns are rejected.	About 10% failed fits on LTP.	Standardized outcome; noise variance from CV; duplicates removed.
GRridge	`predict()` fails for matrices on R $\geq$ 4.2; single-feature or single groups fail; no fold ids.	Random inner folds for the global penalty.	Own prediction code; groups merged; global penalty tuned on the shared folds.
gren	`predict()` fails without feature selection; the package is in a subdirectory of its repository.	Installation from GitHub fails.	Group penalty multipliers used as `glmnet` penalty factors.
ecpc, squeezy	Archived on CRAN; no fold ids; squeezy has no `predict()`.	Own inner folds.	CRAN archive versions; own prediction code.
ipflasso	For binary outcomes, the deviance is replaced by the misclassification rate; no fold ids.	Own inner folds.	Tuned on the misclassification rate.
KnowGRRF	Archived on CRAN; needs data frames; prior strength selected by an AIC on the training data.	Tuning differs from the other methods.	Second engine on `RRF`, tuned by inner CV.
xrnet	Models the mean of the coefficients as a function of the prior.	Unsigned scores cannot help when effects have both signs.	None (method property).
infHS	Does not install as published; the variational algorithm returns `NaN` in our environment.	Cannot be run.	Not included.

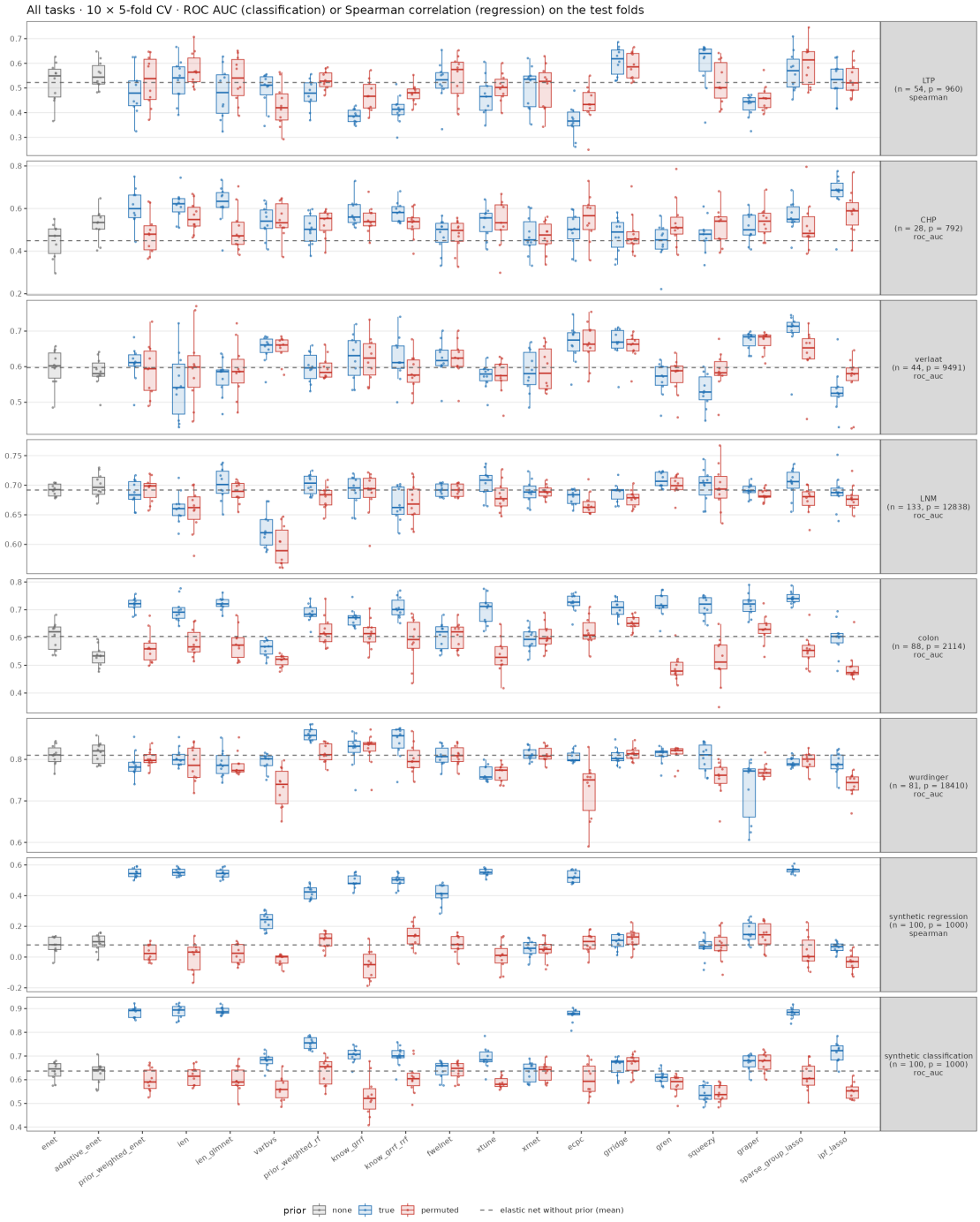


Figure 7: Main benchmark on all tasks: main metric of each learner over 10 repeats of 5-fold CV, without prior (grey), with the true prior (blue) and with the permuted prior (red). Dashed: mean of the EN without prior.

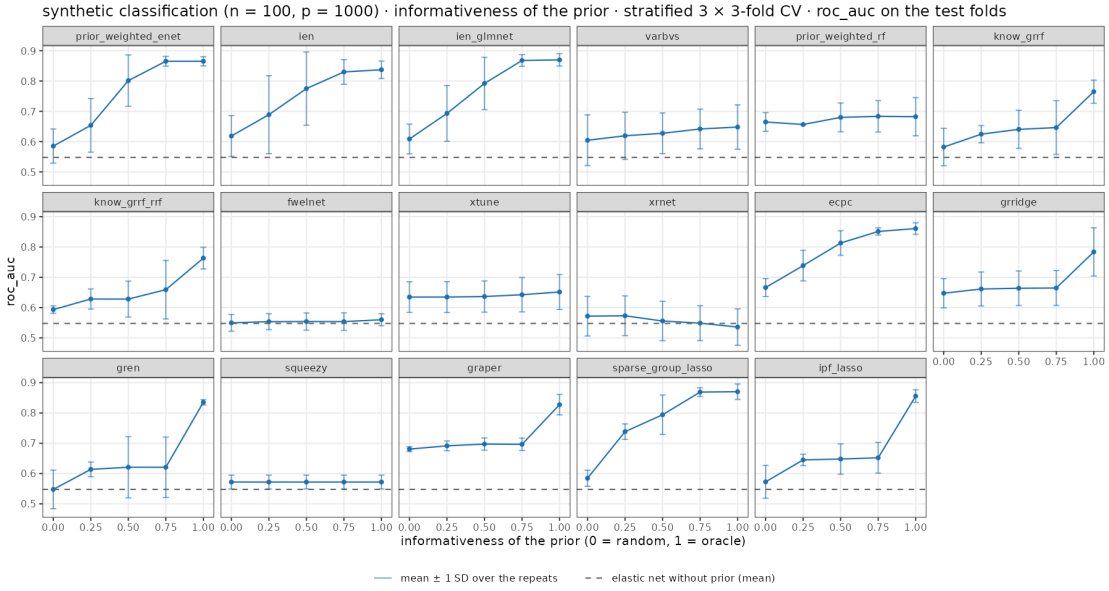


Figure 8: Informativeness of the prior, synthetic classification (ROC AUC); cf. Figure 3.

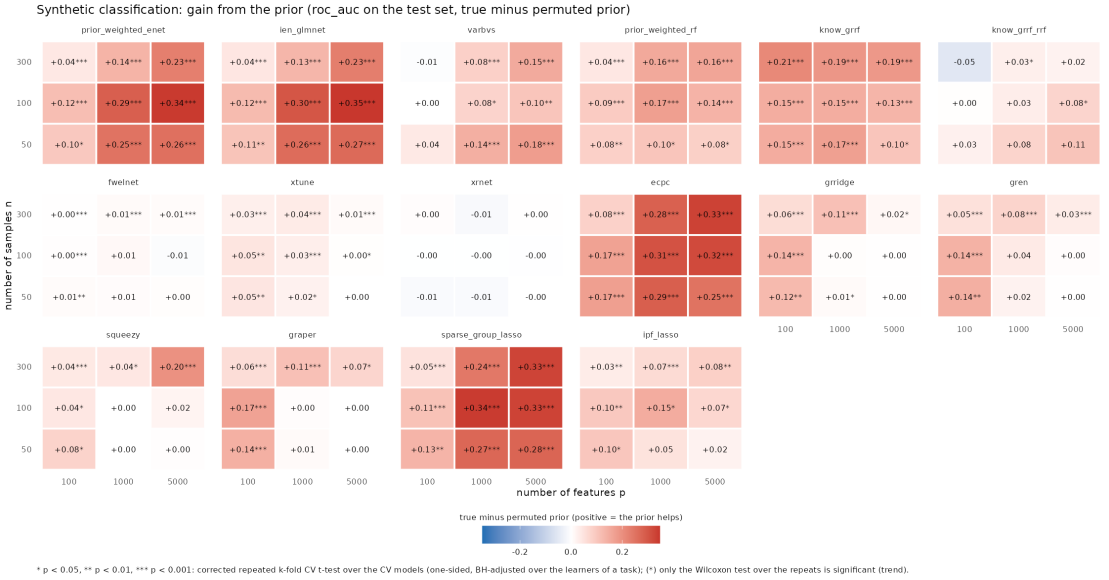


Figure 9: Gain from the prior by n and p, synthetic classification (ROC AUC); cf. Figure 4.

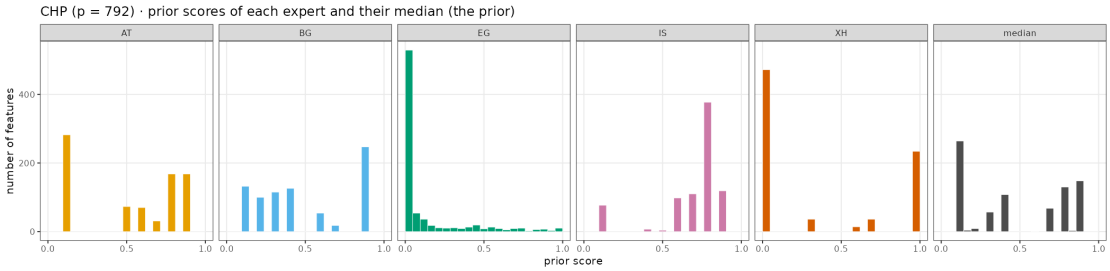


Figure 10: Score distributions of the five immunologists for the 792 features of CHP and their median; cf. Figure 5.

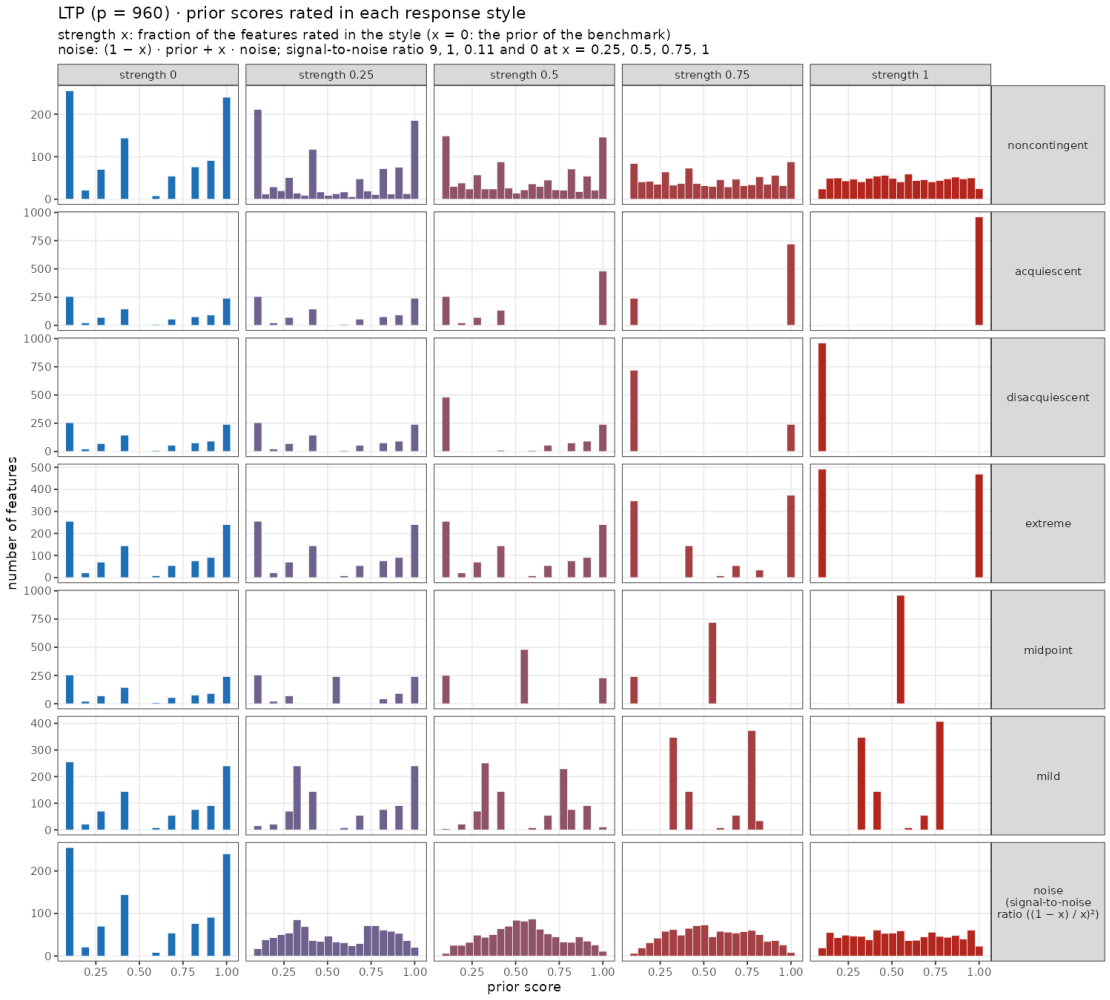


Figure 11: Simulated response styles (Table 3): distribution of the LTP prior when a fraction x of the features (strength, columns) is rated in each style (rows); noise: weight of the uniform noise in the mix. Strength 0 is the prior of the benchmark.

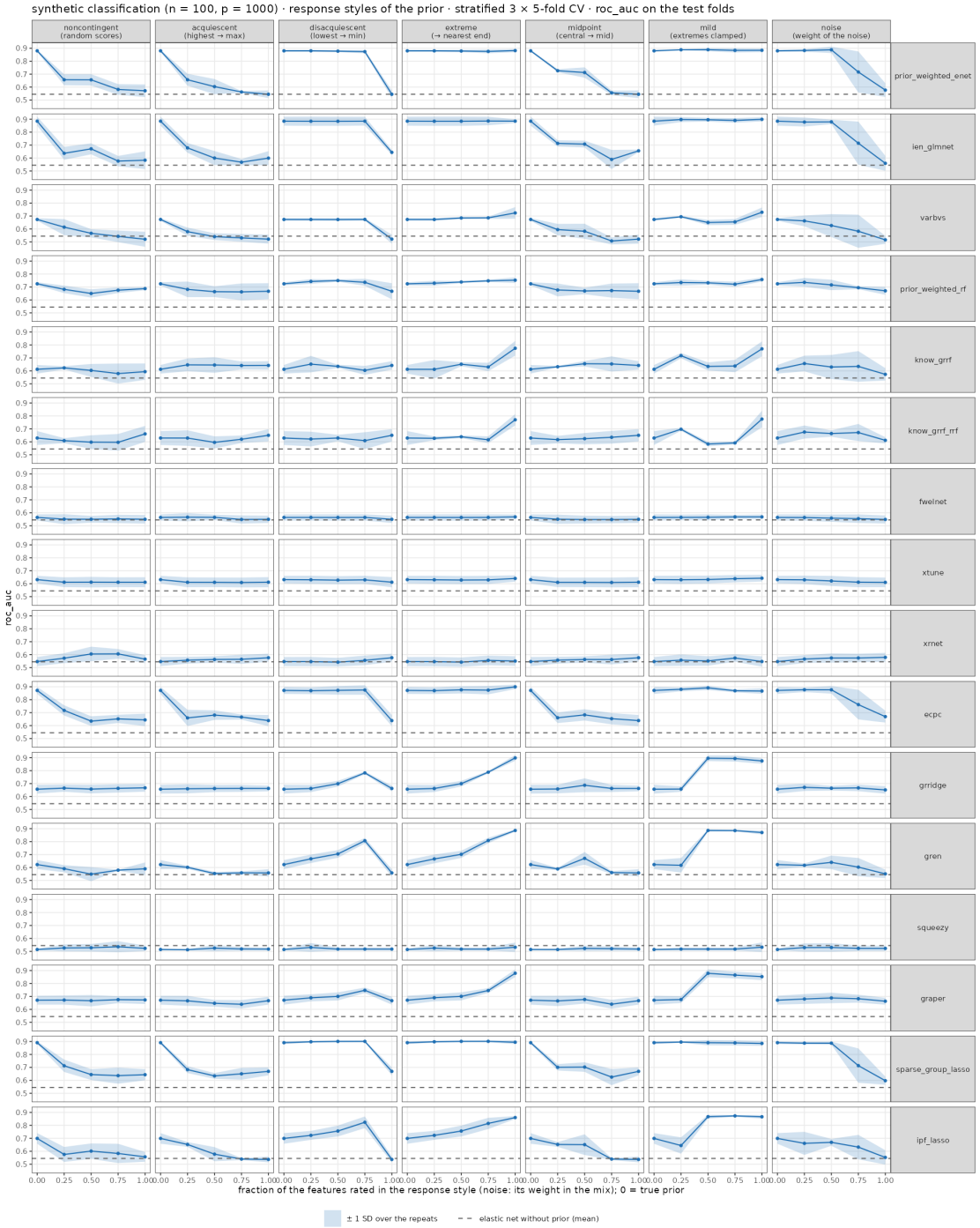


Figure 12: Response styles, synthetic classification: ROC AUC against the fraction of the features rated in each style (noise: its weight in the mix; 0 = true prior); mean ± 1 SD over 3 repeats of 5-fold CV. If the prior is informative and used, the performance should decrease with the strength of a style.

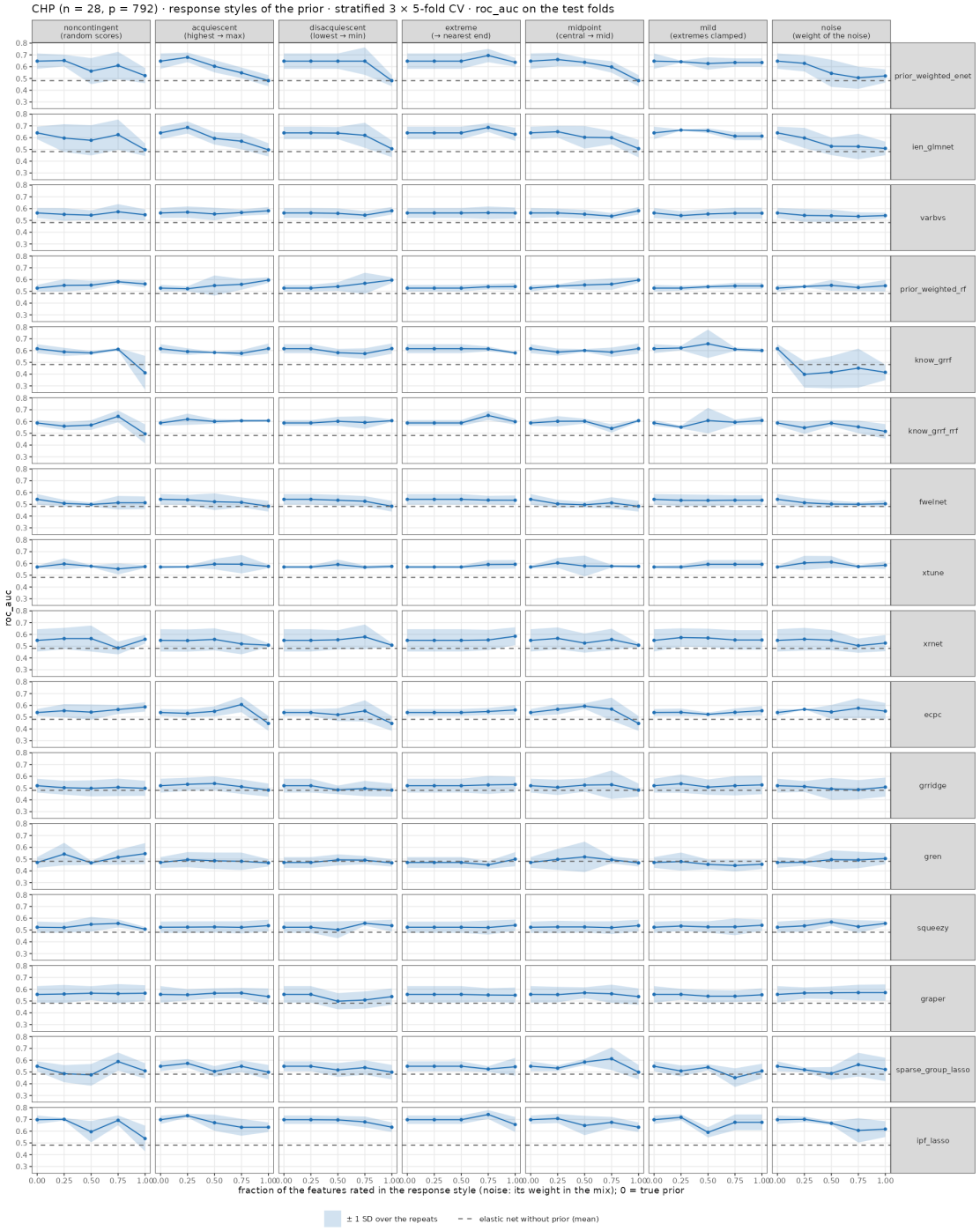


Figure 13: Response styles on CHP (ROC AUC); cf. Figure 12. No loss is significant, as the prior does not help significantly on this task.

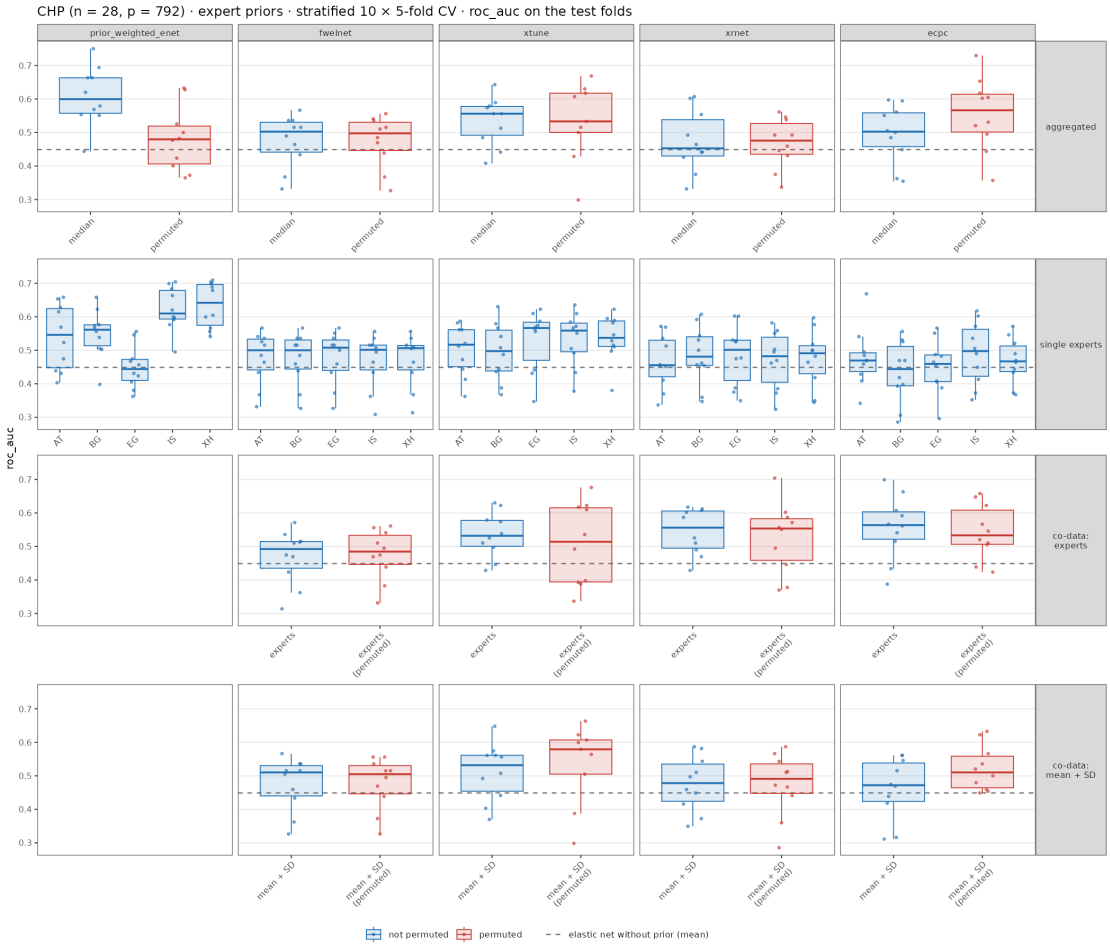


Figure 14: Multiple experts on CHP (ROC AUC, 10 repeats of 5-fold CV): the median of the experts and each single expert as the prior (rows 1–2), and, for the methods that learn the mapping of the prior, all experts as five columns of $\mathbf{Z}$ or their mean and standard deviation as two columns (rows 3–4); true (blue) and permuted (red) prior. No setting is significantly better than the median.