Widespread data leakage inflates performance estimates in cancer drug response prediction

Amir Asiaee[1*], Jared Strauch[2], Leila Azinfar[1], Samhita Pal[1], Heather H. Pua[3], James P. Long[4], Kevin R. Coombes[5]

[1*]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, 37232, USA.
[2]Department of Biostatistics, Vanderbilt University, Nashville, TN, 37232, USA.
[3]Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN, 37232, USA.
[4]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA.
[5]Department of Population Health Sciences, Georgia Cancer Center, Augusta University, Augusta, GA, 30912, USA.

*Corresponding author(s). E-mail(s): amir.asiaeetaheri@vumc.org;

Abstract

Drug response prediction models are widely used to nominate biomarkers and guide preclinical drug prioritization. However, their reported performance hinges on rigorous separation of training and test data during cross-validation (CV). Here we show that a commonly used pattern—supervised feature screening performed on the full dataset before CV—introduces data leakage that systematically underestimates prediction error. Analyzing 265 drugs across 1,462 cancer cell lines, we find that leakage-free CV increases mean squared error (MSE) by 16.6%, with low feature-set overlap between leaked and leakage-free pipelines (mean Jaccard 0.18). A manual audit of 12 recent deep learning and classical methods found confirmed leakage in 10. Such inflated performance estimates likely contribute to computational predictions that fail during independent validation or experimental follow-up. We provide an audit guide and reference implementation to prevent leakage, and introduce a tissue-aware Data Shared Elastic Net (DSEN) that, under correct evaluation, improves prediction for 65.7% of drugs while yielding sparser, more targeted biomarker sets.

Keywords: data leakage, cross-validation, drug response prediction, pharmacogenomics, elastic net, reproducibility, tissue specificity

1 Introduction

Genomic markers that predict drug sensitivity in cancer cells are routinely nominated using machine learning models trained on large-scale pharmacogenomic screens such as the Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC) [1–4]. These models aim to identify molecular features—gene expression, mutations, copy number alterations—that stratify cell lines by drug response and may ultimately guide patient selection in clinical trials.

Cross-validation (CV) is the standard approach for estimating prediction performance in these high-dimensional settings, where the number of features often exceeds the number of samples by orders of magnitude. However, subtle errors in CV implementation can introduce data leakage—the inadvertent use of test-set information during model training—leading to optimistic performance estimates that do not reflect true generalization [5–8].

Here we systematically quantify one prevalent form of leakage in drug response prediction: supervised feature screening performed on the full dataset before CV splits are defined. We demonstrate that this pattern inflates performance estimates by 16.6% on average, destabilizes selected biomarker sets, and is present in 10 of 12 recently published methods we audited. We provide a practical audit guide and reference implementation for leakage-free evaluation, and introduce a tissue-aware modeling approach that yields genuine improvements under correct assessment.

2 Results

2.1 A widely used "feature screening then cross-validation" pattern leaks test-set information

Drug response prediction in pharmacogenomics is typically high dimensional: tens of thousands of molecular features are used to predict continuous sensitivity measurements across hundreds to thousands of cell lines for each compound, as in widely used resources such as CCLE and GDSC [1–4]. To make this tractable, many pipelines apply supervised feature screening (for example, selecting features correlated with response) and then report cross-validation (CV) performance of the downstream model. When screening is computed once using all samples, it implicitly uses test-fold labels (and often test-fold feature statistics), violating the CV assumption that the test fold is held out and biasing error estimation [5–7].

We quantified the impact of this pattern using GDSC/Sanger release 6.0 drug sensitivity for 265 compounds, matched to CCLE molecular features for 1,462 cell lines spanning 31 tissue lineages—the most commonly used dataset pairing across the methods we audited (Online Methods). For each drug, we compared two elastic-net pipelines: (i) an **incorrect** implementation that applies variance filtering, correlation

screening and scaling to the full dataset once and then runs CV; and (ii) a **correct** implementation that repeats screening and scaling independently within each CV fold using training data only (Fig. 1a; Online Methods). Both pipelines used identical hyperparameter grids and performance metrics.

Leakage-free CV increased the estimated prediction error for most compounds. Across 265 drugs, the corrected pipeline increased mean MSE by 16.6% (median 14.0%), and 83.0% of drugs showed inflated performance under the leaked pipeline (Fig. 1b–c). The effect was heterogeneous: 35.8% of drugs showed $\geq$20% inflation and 10.9% showed $\geq$40% inflation, reaching 70.3% for the most affected compound.

2.2 Leakage distorts biomarker discovery by producing non-replicable feature sets

Beyond performance, drug response modeling is often interpreted through the lens of selected genomic features as candidate biomarkers or mechanistic hypotheses. Because leaked screening uses information from the test fold, it can stabilize or enrich feature sets that do not generalize.

We evaluated the overlap between the feature sets selected by the incorrect and correct pipelines using the Jaccard index at a stringent stability threshold (features selected in $\geq$80% of bootstrap resamples; Online Methods). Overlap was low (mean Jaccard 0.18; median 0.04), and 36.3% of drugs had **zero** overlap (Fig. 1d). The leaked pipeline also produced much larger "stable" signatures: on average, 18.1 features met the $\geq$80% threshold under the incorrect pipeline versus 2.2 under the correct pipeline—an eight-fold inflation that could generate many false-positive candidates for experimental follow-up. Together, these results indicate that leakage can simultaneously inflate apparent accuracy and overstate confidence in derived biomarker lists.

2.3 Forensic audit of recent methods reveals recurring leakage modes

To assess how broadly leakage risks extend beyond a single elastic-net baseline, we conducted a manual, code-level audit of 12 drug response prediction methods spanning classical machine learning and deep learning (Supplementary Note 1). Methods were selected based on high citation counts and availability of public code repositories, and include DeepCDR [9], DrugCell [10] and MOLI [11]. We categorized leakage into six recurring modes, including preprocessing fit on the full dataset, test data used for early stopping or model selection, and evaluation splits that mix cell lines or drugs across train and test despite claims of cold-start generalization.

Across the audited methods, 10/12 contained confirmed leakage by direct inspection of released code (Supplementary Table 1). Importantly, several leakage modes (for example, monitoring test performance during training) can bias results even when a nominal train/test split exists. These observations underscore that leakage is not confined to a single modeling family, and motivate providing reusable evaluation templates and audit checklists.

2.4 A tissue-aware Data Shared Elastic Net improves prediction under leakage-free evaluation

Correct evaluation often reduces headline performance, but it also clarifies where methodological innovations yield genuine gains. Motivated by the hierarchical structure of pharmacogenomic panels (cell lines grouped by tissue lineage), we applied the **Data Shared Elastic Net (DSEN)** [12], which decomposes each regression coefficient into a component shared across tissues and a tissue-specific deviation (Online Methods). DSEN adapts the data-shared lasso framework [13] to elastic-net regularization [14, 15] for large-scale drug-by-drug modeling.

Across 265 drugs evaluated with leakage-free CV and tissue-stratified folds, DSEN improved MSE relative to a standard elastic net for 65.7% of compounds, with mean improvement 1.66% (median 0.71%; Fig. 2). Improvements were concentrated in a subset of compounds (95th percentile 8.94%), with a maximum improvement of 23.6% for Imatinib, consistent with strongly context-dependent drug activity (Fig. 2).

DSEN also reduced model complexity. In bootstrap feature-selection analyses, DSEN selected fewer features than the standard elastic net for 93.2% of drugs, reducing the number of ever-selected features by 39% on average (mean 658 features for elastic net versus 382 for DSEN; Supplementary Fig. 1; Online Methods). For 8.3% of drugs, the learned model contained no shared features, suggesting predominantly tissue-specific predictive structure (Supplementary Fig. 2).

2.5 DSEN retains biological signal as assessed by recovery of known drug targets

To test whether corrected evaluation and tissue-aware modeling preserve meaningful biological signal, we curated 464 drug–target pairs (262 unique drugs) and mapped target gene names to features in our molecular matrix (87.3% of curated pairs mapped to a gene represented in the feature space; Online Methods). Using a stability threshold of $\geq$50% of bootstrap resamples (relaxed from the 80% used for Jaccard analyses to improve sensitivity for this validation task), known targets were recovered infrequently by both models, as expected under strong regularization and highly polygenic response. Nevertheless, DSEN showed modestly higher target recovery than the standard elastic net: among 220 drugs with at least one target present in the feature set, DSEN recovered any target for 5.5% of drugs versus 4.5% for elastic net, and full target sets for 4.5% versus 3.2% (Fig. 4).

Case studies highlighted biologically coherent recoveries, including BRAF for BRAF inhibitors and MDM2 for Nutlin-3a (Supplementary Table 2), supporting that leakage-free evaluation does not eliminate signal but provides more conservative and reproducible estimates.

3 Discussion

Drug response prediction sits at the intersection of high-dimensional biology and methodological innovation, making it particularly susceptible to subtle evaluation errors. Our results demonstrate that supervised feature screening and normalization

performed once on the full dataset—an often implicit "preprocessing" step—can produce substantial optimism in reported accuracy, even for relatively simple elastic-net baselines. Importantly, the same pattern can propagate into biomarker discovery by amplifying the apparent stability and size of selected feature sets.

The practical implication is not that prior biological conclusions are necessarily invalid, but that many published performance numbers and feature rankings are likely overconfident. This matters for comparative benchmarking, for claims of improvement over baselines, and for downstream tasks such as drug repurposing where inflated accuracy can translate into misleading prioritization. Because leakage can arise in many forms—especially in deep learning workflows that repeatedly evaluate on a "test" split during training—robust evaluation requires systematic safeguards and transparent reporting.

We therefore provide two concrete deliverables intended to improve practice. First, we release an audit taxonomy and a reproducible reference implementation of leakage-free CV for drug response prediction, designed to be directly reusable by the community. Second, we introduce DSEN as a tissue-aware baseline that yields consistent (though modest) gains under correct evaluation while reducing model complexity and maintaining biological signal. DSEN's improvements are not universal, highlighting that tissue-dependent structure is drug specific and motivating future work on adaptive sharing across lineages or molecular subtypes.

Our study has limitations. Quantification of performance inflation focused on a widely used elastic-net pipeline and one benchmark dataset; other leakage types identified in the audit (for example, test-set early stopping or pair-level splits inconsistent with cold-start claims) could yield different magnitudes of bias. In addition, our manual audit emphasizes code-level evidence; re-running each audited method with corrected splits is an important but computationally intensive direction for follow-up. Despite these limits, the convergence of quantitative inflation, feature instability and widespread audit findings suggests that leakage is a field-wide reproducibility risk.

4 Online Methods

4.1 Data sources

We analyzed drug sensitivity measurements from GDSC/Sanger release 6.0 and molecular features derived from CCLE/DepMap releases distributed with the CCLE update by Ghandi et al. [2–4]. The predictor matrix contained 86,546 features spanning gene expression, copy number, mutation and protein measurements for 1,462 cell lines, and drug response covered 265 compounds (AUC/activity-area values). Data provenance is documented in the accompanying code repository.

4.2 Correct and incorrect cross-validation pipelines

For each drug, we fit elastic-net regression models using `glmnet` [16] to predict response from molecular features. The **incorrect** pipeline applied the full preprocessing stack once using all samples: (i) variance filtering, (ii) correlation screening with

the response, (iii) duplicate-feature removal, and (iv) scaling (excluding binary mutation features), and then performed 5-fold CV on the filtered matrix. The **correct** pipeline repeated the same steps independently within each fold, fitting preprocessing on the training split and applying it to the held-out split.

Both approaches used the same hyperparameter grid over the elastic-net mixing parameter (α) and selected λ by minimum CV error along the regularization path. We report mean squared error (MSE) on held-out folds and compute standard errors as the fold-to-fold standard deviation divided by $\sqrt{K}$ (fixing a common SD-versus-SE reporting mistake).

4.3 Data Shared Elastic Net (DSEN)

DSEN models responses across tissues as related tasks. For tissue t, coefficients are parameterized as $\beta_{\text{shared}} + \beta_{\text{tissue},t}$. We construct a sparse block matrix Z that couples shared and tissue-specific coefficients and fit elastic-net models with penalty factors that control the ratio between shared and tissue-specific regularization and optionally weight tissue-specific penalties by tissue sample size. Cross-validation used stratified folds over tissue labels with per-fold feature screening and scaling performed using training data only.

4.4 Bootstrap feature selection analyses

To quantify feature selection characteristics, we performed bootstrap resampling (100 replicates) and counted how often each feature received a non-zero coefficient. We report counts of selected features, distributions of selection frequencies, and overlap metrics (Jaccard index) where applicable.

4.5 Drug–target curation and target recovery

We curated drug–target relationships from GDSC metadata and standardized gene names using a curated alias-to-HGNC mapping. Target recovery was evaluated by checking whether each drug's target gene(s) appeared among selected model features (selected in $\geq$50% of bootstrap replicates) for elastic net and DSEN.

4.6 Use of large language models

Large language models were used to assist with drafting and editing the manuscript text. All scientific claims, quantitative results and citations were verified against the repository source code and result files by the authors, who take full responsibility for the content.

4.7 Reporting summary

No human participants or animal experiments were involved.

Data availability

Processed result tables and figures will be made available in an accompanying code repository upon publication. The underlying CCLE/DepMap molecular data and GDSC drug response data are available from their original sources.

Code availability

All analysis code, including leakage-free CV implementations, DSEN, and an audit guide, will be made available upon publication.

Author contributions

[To be completed.]

Competing interests

The authors declare no competing interests.

Figures

References

[1] Barretina, J. *et al.* The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012). URL http://dx.doi.org/10.1038/nature11003.

[2] Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012). URL http://dx.doi.org/10.1038/nature11005.

[3] Iorio, F. *et al.* A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016). URL http://dx.doi.org/10.1016/j.cell.2016.06.017.

[4] Ghandi, M. *et al.* Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569**, 503–508 (2019). URL http://dx.doi.org/10.1038/s41586-019-1186-3.

[5] Ambroise, C. & McLachlan, G. J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences* **99**, 6562–6566 (2002). URL http://dx.doi.org/10.1073/pnas.102102699.

[6] Varma, S. & Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7** (2006). URL http://dx.doi.org/10.1186/1471-2105-7-91.

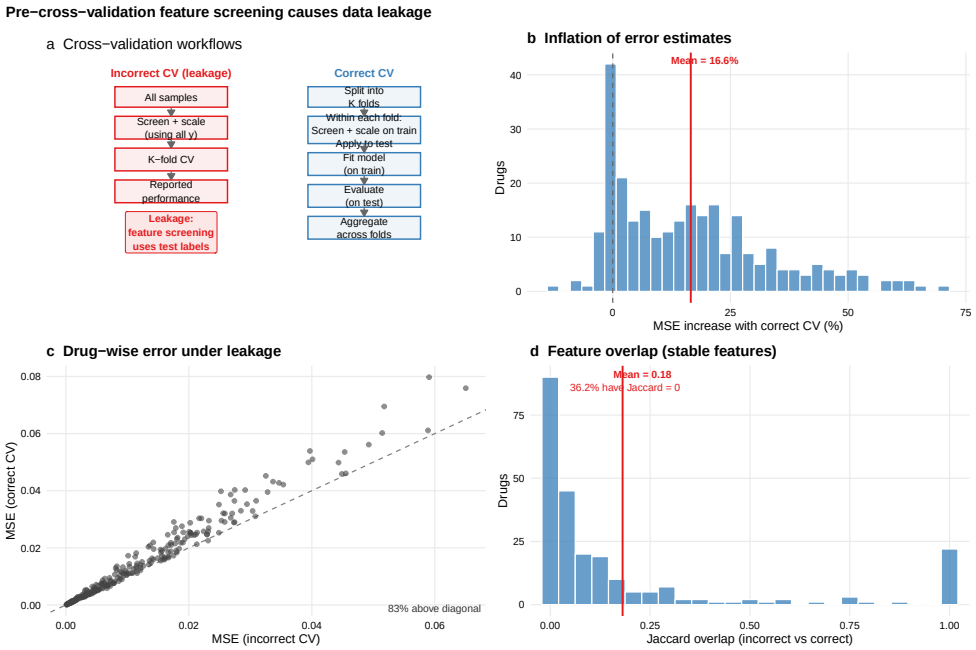


Fig. 1 Data leakage from pre-CV feature screening inflates performance and destabilizes feature sets. a, Schematic contrasting an incorrect pipeline (supervised screening and scaling performed once on all samples before CV) with a correct pipeline (all supervised screening and scaling performed within each fold using training samples only). **b,** Distribution of MSE inflation (%) when switching from incorrect to correct CV across 265 drugs. **c,** Scatter of drug-wise MSE estimates under incorrect versus correct CV; points above the diagonal indicate underestimated error under the incorrect pipeline. **d,** Jaccard overlap between stable feature sets obtained from incorrect versus correct pipelines (features selected in $\geq$80% of bootstrap resamples), highlighting poor reproducibility and frequent zero overlap.

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

[8] Whalen, S., Schreiber, J., Noble, W. S. & Pollard, K. S. Navigating the pitfalls of applying machine learning in genomics. *Nature Reviews Genetics* **23**, 169–181 (2022). URL https://doi.org/10.1038/s41576-021-00434-9.

[9] Liu, Q., Hu, Z., Jiang, R. & Zhou, M. Deepcdr: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics* **36**, i911–i918 (2020). URL http://dx.doi.org/10.1093/bioinformatics/btaa822.

[10] Kuenzi, B. M. *et al.* Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* **38**, 672–684.e6 (2020). URL http://dx.doi.org/10.1016/j.ccell.2020.09.014.

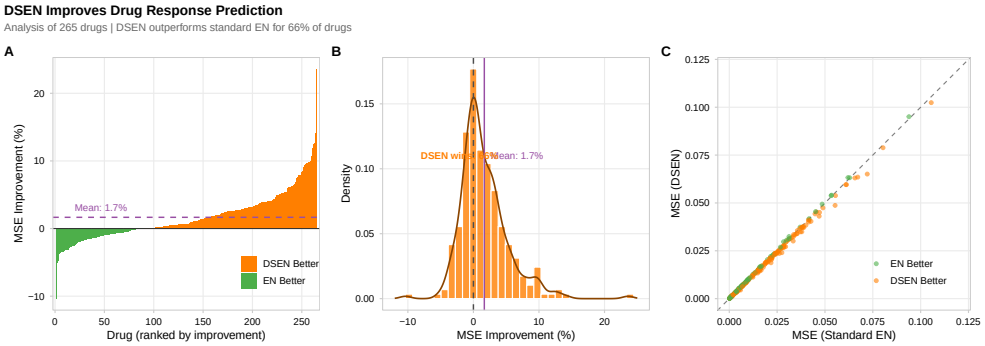


Fig. 2 DSEN improves leakage-free drug response prediction across 265 drugs. Waterfall plot of per-drug MSE improvement (%) for DSEN relative to a standard elastic net under leakage-free CV, accompanied by the distribution of improvements and representative concordance between model errors. Positive values indicate improved performance with DSEN.

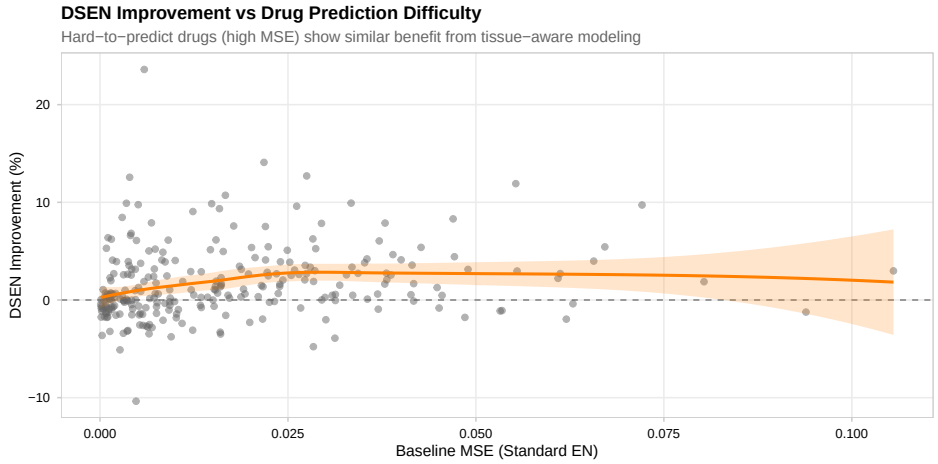
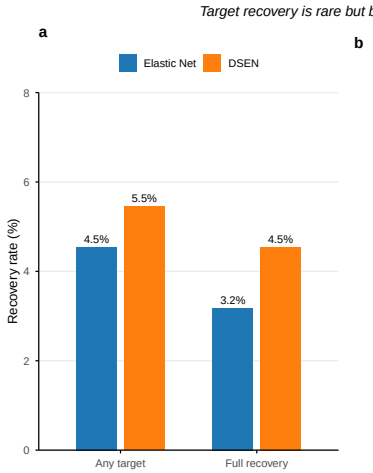


Fig. 3 DSEN benefits are drug dependent and correlate with baseline prediction difficulty. Relationship between DSEN improvement and baseline error for each drug, illustrating that DSEN gains concentrate in a subset of difficult-to-predict compounds.

[11] Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C. & Ester, M. Moli: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* **35**, i501–i509 (2019). URL http://dx.doi.org/10.1093/bioinformatics/btz318.

[12] Strauch, J. & Asiaee, A. Improving drug sensitivity prediction and inference by multi-task learning. *bioRxiv* (2024). URL https://doi.org/10.1101/2024.05.09.593186. Preprint.

[13] Gross, S. M. & Tibshirani, R. Data shared lasso: A novel tool to discover uplift. *Computational Statistics & Data Analysis* **101**, 226–235 (2016). URL http://dx.



b

Drug	Mechanism	Target(s)	EN	DSEN
PLX−4720	BRAF V600E inhibitor	BRAF	Full	Full
SB590885	BRAF inhibitor	BRAF	Full	Full
Dabrafenib	BRAF inhibitor	BRAF	Full	Full
AZ628	BRAF inhibitor	BRAF	Full	Full
Cetuximab	EGFR antibody	EGFR	–	Full
Gefitinib	EGFR TKI	EGFR	Full	Full
Afatinib	Pan−HER TKI	EGFR;ERBB2	Partial	Full
Nutlin−3a	MDM2 antagonist	MDM2	Full	Full
Pazopanib	Multi−TKI	PDGFRA	Partial	–
Axitinib	Multi−TKI	PDGFRA	–	Partial
Pictilisib	PI3K inhibitor	PIK3CA	–	Full

Fig. 4 Target recovery is rare under strong regularization but biologically coherent when it occurs. a, Summary of target recovery rates for elastic net and DSEN under a stringent stability threshold (selected in $\geq$50% of bootstrap resamples) among 220 drugs with at least one target gene represented in the feature space. **b,** Case studies showing drugs where known targets were identified, grouped by mechanism of action. Recovery is marked as "Full" when all annotated targets were found or "Partial" when a subset was recovered. Notable patterns include consistent BRAF recovery for BRAF inhibitors and EGFR/ERBB2 recovery for receptor tyrosine kinase inhibitors.

doi.org/10.1016/j.csda.2016.02.015.

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

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

[16] Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33** (2010). URL http://dx.doi.org/10.18637/jss.v033.i01.