Barun Bhhatarai[1], W. Patrick Walters[2], Cornelis E. C. A. Hop[3], Guido Lanza[4] and Sean Ekins[5]*
[1]*Novartis Institutes for Biomedical Research, Cambridge, MA, USA.* [2]*Relay Therapeutics, Cambridge, MA, USA.* [3]*Genentech, South San Francisco, CA, USA.* [4]*Numerate, San Francisco, CA, USA.* [5]*Collaborations Pharmaceuticals Inc., Raleigh, NC, USA.*
**e-mail: sean@collaborationspharma.com*

Published online: 18 April 2019
https://doi.org/10.1038/s41563-019-0332-5

References

1. Gupta, R. R. et al. *Drug Metab. Dispos.* **38**, 2083–2090 (2010).
2. Ekins, S., Honeycutt, J. D. & Metz, J. T. *Drug Discov. Today* **15**, 451–460 (2010).
3. Page, K. M. *Mol. Pharm.* **13**, 609–620 (2016).
4. Webborn, P. J. H. *Future Med. Chem.* **6**, 1233–1235 (2014).
5. Zientek, M. et al. *Chem. Res. Toxicol.* **23**, 664–676 (2010).
6. Zhang, H. et al. *Toxicol. In Vitro* **23**, 134–140 (2009).
7. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. *Nature* **559**, 547–555 (2018).
8. Wang, S. et al. *Mol. Pharm.* **13**, 2855–2866 (2016).
9. Ekins, S. & Williams, A. J. *Lab Chip* **10**, 13–22 (2010).
10. Winiwarter, S. et al. *J. Comput. Aided Mol. Des.* **29**, 795–807 (2015).
11. Clark, A. M., Williams, A. J. & Ekins, S. *J. Cheminform.* **7**, 9 (2015).
12. Martin, E. J., Polyakov, V. R., Tian, L. & Perez, R. C. *J. Chem. Inf. Model.* **57**, 2077–2088 (2017).
13. Ericksen, S. S. et al. *J. Chem. Inf. Model.* **57**, 1579–1590 (2017).
14. Verras, A. et al. *J. Chem. Inf. Model.* **57**, 445–453 (2017).
15. Capuzzi, S. J. et al. *J. Chem. Inf. Model.* **57**, 105–108 (2017).
16. Sushko, I. et al. *J. Comput. Aided Mol. Des.* **25**, 533–554 (2011).
17. Russo, D. P., Zorn, K. M., Clark, A. M., Zhu, H. & Ekins, S. *Mol. Pharm.* **15**, 4361–4370 (2018).
18. Sheridan, R. P. *J. Chem. Inf. Model.* **53**, 2837–2850 (2013).
19. Roy, K., Kar, S. & Ambure, P. *Chemom. Intell. Lab. Syst.* **145**, 22–29 (2015).
20. Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E. & Svetnik, V. *J. Chem. Inf. Model.* **55**, 263–274 (2015).
21. Liu, K. et al. Preprint at https://arxiv.org/abs/1803.06236 (2018).
22. Ramsundar, B. et al. *J. Chem. Inf. Model.* **57**, 2068–2076 (2017).
23. Hop, P., Allgood, B. & Yu, J. *Mol. Pharm.* **15**, 4371–4377 (2018).
24. Rodríguez-Pérez, R. & Bajorath, J. *ACS Omega* **3**, 12033–12040 (2018).
25. Korotcov, A., Tkachenko, V., Russo, D. P. & Ekins, S. *Mol. Pharm.* **14**, 4462–4475 (2018).
26. Lane, T. et al. *Mol. Pharm.* **15**, 4346–4360 (2018).
27. Xu, Y., Ma, J., Liaw, A., Sheridan, R. P. & Svetnik, V. *J. Chem. Inf. Model.* **57**, 2490–2504 (2017).
28. Ramsundar, B. et al. Preprint at https://arxiv.org/abs/1502.02072 (2015).
29. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. *J. Comput. Aided Mol. Des.* **30**, 595–608 (2016).
30. Ekins, S. *Pharm. Res.* **33**, 2594–2603 (2016).

Acknowledgements

D. Chipman and E. Cutler are kindly acknowledged for organizing the AI Applications in Biopharma Summit. S.E. acknowledges A. Clark, J. Freundlich and A. Williams for their many discussions on machine learning and ADME/Tox models. S.E. acknowledges funding to Collaborations Pharmaceuticals Inc. from NIGMS R44 GM122196-02A1.

Avoiding common pitfalls in machine learning omic data science

This Comment describes some of the common pitfalls encountered in deriving and validating predictive statistical models from high-dimensional data. It offers a fresh perspective on some key statistical issues, providing some guidelines to avoid pitfalls, and to help unfamiliar readers better assess the reliability and significance of their results.

Andrew E. Teschendorff

Most fields of science have undergone a big data revolution[1–3]. In biology, it began two decades ago, with the emergence of the first microarray technologies[4–6] and subsequent enhancements through next-generation sequencing[7,8], which allowed molecular measurements (for example, gene expression) to be made on a truly genome-wide scale, referred to generally as 'omic' data. These technological advances have led to major breakthroughs in our fundamental understanding of cell biology, and have also begun to reshape the clinical treatment and management landscape of many diseases[9], with further significant improvements on the horizon[10–12]. However, major obstacles in translating omic data into tangible benefits for healthcare remain.

One of these obstacles relates to the process of statistical inference, a complex endeavour due mainly, but not exclusively, to the high-dimensional nature of omic data: measuring molecular properties at hundreds of thousands, if not a million or more, genomic features means that associations with phenotypes or outcomes of interest can arise purely out of random chance, an issue commonly referred to as the 'multiple-testing problem'[13]. In addition, omic data is often plagued by poorly understood, or unknown, confounding sources of data variation, which, if unaccounted for, can lead to false discoveries or mask the identification of genuine biological variation[14,15].

Statistical bioinformatics has emerged as a relatively young discipline to tackle these challenges and, arguably, because of its infancy, application of statistical methodology to complex omic datasets has not been without its problems and controversy[16]: erroneous results[17,18], retractions[19] and claims of irreproducibility have been frequent[20–22]. Indeed, a recent study in neuroscience concluded that specific errors in statistical analyses are present in approximately 50% of published papers[23]. This high frequency of errors is likely to extend to all major types of statistical analyses and research fields in biology and related disciplines. Part of the underlying problem is that most of molecular biology has traditionally been a non-quantitative discipline that did not require the need for sophisticated mathematics and statistics. The ongoing omic data revolution has, however, changed this situation very quickly and quite dramatically, with current data analysis demands requiring an ever deeper and broader understanding of the statistical issues at hand. Thus, although many biologists have addressed this problem by teaming up or contracting mathematically qualified scientists, the shortage of suitably trained professionals means that research groups often rely on untrained or inexperienced scientists to perform complex data analysis tasks[16]. Although previous commentaries[16,24] have offered recommendations and guidelines to avoid common pitfalls associated with these analysis tasks, in practice, following these guidelines has proven more difficult than expected. By bringing in a novel perspective on some of the key statistical issues, we hope that the exposition below will help data analysts and readers better assess and interpret results from studies that apply machine learning techniques to predict outcomes, not only across biomedicine[25] and

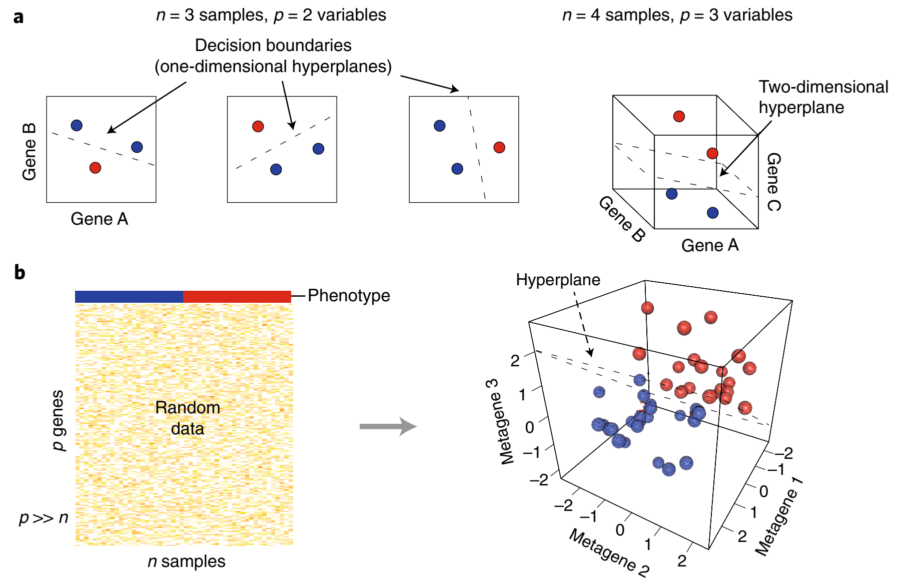

Fig. 1 | The curse of dimensionality and overfitting. a, Low-dimensional examples designed to illustrate the curse of dimensionality phenomenon. For the case of three samples and two variables (say expression of genes A and B), it is always possible to find a linear combination of the two genes' expression values, so that the resulting line (a one-dimensional hyperplane) perfectly discriminates a binary phenotype, no matter how this phenotype is specified. The line can be used to define a decision boundary for class assignment, in which case the resulting predictor would achieve 100% discrimination accuracy. Also shown is the case of four samples and three variables (expression of genes A, B and C). Here, for any random binary phenotype, it is always possible to find a plane (a two-dimensional hyperplane) that can perfectly discriminate the two phenotypes. Generally, for n samples, if the number of variables is at least as large as $n-1$, we can always find an $(n-2)$-dimensional hyperplane that can perfectly discriminate the two phenotypes. **b**, Extrapolation to a real example of an omic data matrix with 50 samples, representing two phenotypes (red and blue) and over $p = 20{,}000$ genes, where the underlying data are random (generated from a Gaussian of mean 0 and variance 1). Even when the data is random in relation to the phenotype, it is still possible to find projections of the data (called metagenes) onto a lower-dimensional subspace (here three-dimensional) so that the two phenotypes are perfectly or almost perfectly discriminated.

DNA-based material science[12], but also more generally across other big data science disciplines.

One of the machine learning tasks where pitfalls are most often encountered revolves around the problem of 'class prediction' — that is, the task of building a molecular classifier that can predict a phenotype or outcome of interest (for example, disease diagnostics, prognosis of a cancer). To derive the molecular classifier, a typical omic 'discovery' study may perform molecular measurements (say gene expression) on a fairly large number p ($p \approx 10^5$–10^7) of variables (genes, regulatory elements) and, due to cost or logistical reasons, on a relatively smaller number n of samples ($n \approx 100$–$1{,}000$). The resulting data is then analysed using machine learning methods[26,27] to build the predictor/classifier, which typically involves an optimization process where an objective function (for example, the misclassification rate) is minimized. Broadly speaking, the resulting molecular classifier will typically consist of three elements: (i) a set of molecular features (for example, genes); (ii) a set of corresponding weights that inform on the relative contribution of each feature to the classifier; and (iii) a mathematical rule that integrates these weights with the feature values of a given sample (for example, gene expression), converting them into a predictive score that allows subsequent class assignment for that sample via a decision rule that involves a choice of threshold[16]. Owing to its ease of interpretation, a popular choice of mathematical rule in biomedicine is that of a simple linear function of the molecular features and weights[28]. However, nonlinear rules (for example, neural networks) are also widely applied and are rapidly increasing in popularity thanks to breakthroughs in deep learning[29].

The first statistical challenge that emerges is commonly known in the machine learning and omic literature as the 'curse of dimensionality' (Fig. 1a). A well-known statistical theorem states that for n samples and $n-1$ variables under general conditions, there always exists an $(n-2)$-dimensional hyperplane that can perfectly discriminate any random labelling of the n samples into two phenotypes (for example, cancer and healthy)[30]. In the trivial case $n = 3$, this leads to the well-known fact that given any random assignment of three samples into two groups, these two can always be discriminated by a line (a one-dimensional hyperplane) in the plane spanned by the two variables (Fig. 1a). By extrapolation, this simple example illustrates that it is always possible to perfectly discriminate any random binary phenotype if sufficient numbers of molecular measurements are taken — that is, if $p \geq n-1$ (Fig. 1a). In practice, $p \gg n$, and therefore $p \gg n-1$, which means that many different high-dimensional hyperplanes can be found that perfectly discriminate any arbitrary binary phenotype (Fig. 1b). Thus, naive application of machine learning methods to omic data will lead to overfitting: that is, the derived predictive model fits random variation present in the data that does not represent true biological variation associated with the phenotype of interest[16].

In principle, the solution to the above problem is to include a regularization or penalty term into the objective function to be optimized[27]. The purpose of this regularization term is to penalize predictive models that contain too many features, and thus to favour simpler models defined by a relatively small set of predictive features. Thus, regularization enables a form of feature selection. However, this regularization step introduces a set of parameters, notably a penalty parameter, that is not specified a priori, and which therefore needs to be determined.
The penalty parameter directly controls the complexity of the model — that is, the number of features contributing to the predictive model — and needs to be estimated via a separate optimization procedure known as cross-validation (CV). Normally, CV involves splitting the dataset up into a training and test set, but ideally would require three subsets[27]: a training set, a model selection set and a test set, with the latter two both being blinded to the training process (Fig. 2a). The purpose of the training set is to learn the parameters (that is, features and weights) that define a predictor/model and to do this for a range of different predictors/models of variable complexity, as given, for instance, by the specific value of the penalty parameter. The purpose of the model selection set is to then identify the model with best generalization performance. In other words, by estimating the parameters of different predictors in

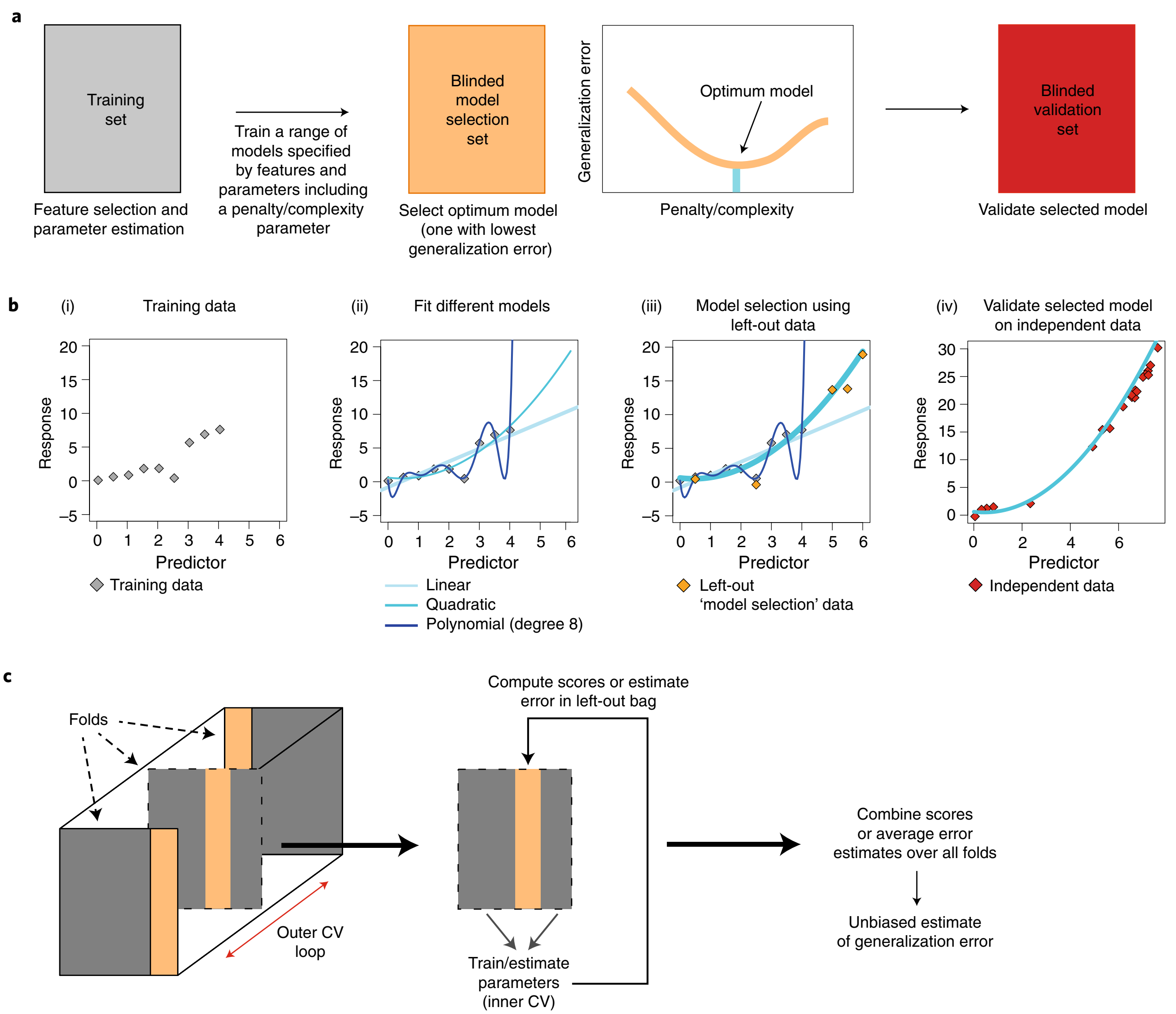
a
Training set
Feature selection and parameter estimation
Train a range of models specified by features and parameters including a penalty/complexity parameter
Blinded model selection set
Select optimum model (one with lowest generalization error)
Generalization error
Optimum model
Penalty/complexity
Blinded validation set
Validate selected model
b
(i) Training data
(ii) Fit different models
(iii) Model selection using left-out data
(iv) Validate selected model on independent data
Response
Predictor
20
15
10
5
0
−5
0 1 2 3 4 5 6
30
25
0 2 4 6
◆ Training data
Linear
Quadratic
Polynomial (degree 8)
◆ Left-out 'model selection' data
◆ Independent data
c
Folds
Outer CV loop
Compute scores or estimate error in left-out bag
Train/estimate parameters (inner CV)
Combine scores or average error estimates over all folds
Unbiased estimate of generalization error

Fig. 2 | Avoiding bias when training and evaluating molecular predictors. a, In the ideal scenario, we have enough data that we can split the samples up into a training set (grey box), a model selection set (orange box) and a validation set (red box). The training set is used to learn a number of different predictive models each characterized by a set of parameters, including a penalty/complexity parameter. The purpose of the model selection set is to then assess the predictive accuracy of each model on blinded data, so as to select the model with the lowest generalization error. This optimum model is the most likely to be true. Finally, this model selection step needs to be validated on another blinded validation set. **b**, An illustrative low-dimensional toy example (where a penalty term is not needed), consisting of nine training samples with one response (*y* axis) and one feature (*x* axis) (i). The training data was generated using a purely quadratic polynomial with Gaussian noise added (although in practice the generating function is unknown to us). Three different polynomials are fitted to the training data: a linear model, a quadratic model and a polynomial of degree eight (ii). Note that for the latter case, there are nine parameters (intercept plus eight regression parameters), and that therefore this model can be fit perfectly to the training data, leading to zero error. However, upon adding more data (left-out data, orange data points), we can see that only the quadratic model fits the new data well (iii). The quadratic model has the lowest generalization error and is therefore selected as the true model. Confirmation that this is indeed the model most likely to be true is shown by validating the quadratic model against new independent data (red data points) (iv). **c**, In the more realistic scenario, omic datasets are often underpowered, in which case the training and model selection sets are combined together in a nested cross-validation (CV) procedure. To obtain unbiased estimates of the generalization error, an inner CV loop is used to estimate the parameters within the training subset of each fold, while an outer CV loop is used to obtain predictive scores and/or error estimates in the left-out bags. Thus, this procedure separates out the process of training and estimation (inner CV loop) from the process of error estimation and model selection (outer CV loop), as required to avoid reporting biased performance estimates.

the training set, we can construct many models that can accurately predict the phenotype in the training set, but not all of these will generalize or accurately predict the phenotype in independent blinded data. In effect, the model that achieves best generalization performance in the model selection set is deemed to be the one most likely to be true. The purpose of the blind validation set is to then confirm (or 'validate') this predictive model by demonstrating that the predictive performance of the selected model is similar in the model selection and validation sets (Fig. 2a). If the latter holds, then generally speaking overfitting has been avoided.

For simplicity, the following example illustrates these key concepts in a low-

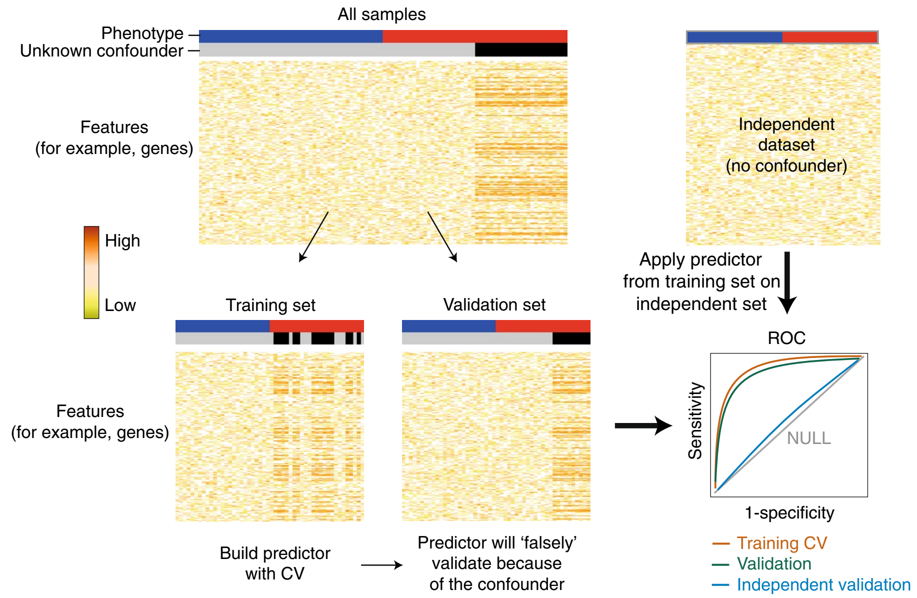

Fig. 3 | Unknown confounders and class prediction. An unrecorded or unknown study-specific confounder that affects a subset of cases (red), may lead to consistent prediction performance across a training and validation set, falsely suggesting the molecular predictor is not overfitted. However, the significant discrimination accuracy (as measured by the area under the receiver operating characteristic (ROC) curve) in both training and validation sets is driven by the unknown confounder, as splitting a dataset up into training and validation sets does not remove the effect of the confounder. If the same predictor is then applied to an independent validation set, say an equivalent cohort generated by a different lab with potentially a different experimental protocol or technology, it fails, as there is no true association between the molecular data and phenotype. While this example illustrates an 'extreme case', confounding factors are an important reason why derived molecular predictors may fail or underperform when evaluated in equivalent but independent datasets. The ROC curve describes the relation between sensitivity — that is, the fraction of true cases classified by the predictor to be a 'case' — and the false positive rate, which equals 1-specificity.

dimensional setting, but the general idea and strategy also holds for high-dimensional omic data. The example considers fitting a curve to nine training data points, each specified by a pair of values, one value for the response variable and another value for the given feature (Fig. 2b,i). The data was generated using a function (in practice, this function is unknown to us) that maps each feature value to a response value, with noise added. To determine the function that gave rise to the data, one might consider a range of different polynomial models of variable degrees (Fig. 2b,ii). In this example, the degree of the polynomial specifies the complexity of the model, since the number of parameters to be estimated increases with the degree of the polynomial. As the example shows, it is possible to fit a range of different polynomials to the training data quite well (Fig. 2b,ii), but only one of the models does well in predicting the five new data points not used in the training process (Fig. 2b,iii). Thus, we would select the polynomial with best performance in this 'model selection set' and finally confirm its validity in an independent validation set (Fig. 2b,iv). What is important to appreciate is that no matter what the underlying predictors and response variables are, the more complex a model is (that is, the more parameters it has), the better it will fit to the training data, but at the expense of doing much worse in the left-out data (that is, the model selection set) (Fig. 2a,b).

With the above procedure, a difficulty arises because omic datasets are often not large enough to afford splitting them up into three fairly large groups (training, model selection and validation). This is why the training and model selection sets are often combined together in the CV procedure[26]. It is at this CV step where most pitfalls are encountered[16,31]. In CV, the training set is typically split up into a number of equally sized bags, where initially all bags but one are used for training, with the left-out bag acting as a 'mini' model selection set. However, due to the typically small size of this left-out bag, this bag alone cannot be used for error estimation and model selection. This problem is circumvented by repeating the analysis, but now using a different bag as the mini model selection set and the rest of bags for training. Thus, assuming m bags, there are m different partitions (called 'folds') of the m bags into a training set of $m-1$ bags and one bag left out for testing. Thus, one can train predictors for each of the m possible folds, with predictor-specific scores computed for all samples in the left-out bag (Fig. 2c). Although the ranking of features and associated weights will differ between folds, the main purpose of the cross-validation is to determine the complexity parameter (for example, the number of features) that optimizes the performance in the left-out bags. This can be achieved by merging together the prediction sample scores of all left-out bags at a given complexity parameter value (Fig. 2c), and subsequently identifying the value that maximizes prediction accuracy on the merged scores. A special case of CV is leave-one-out CV (LOOCV), whereby all but one sample are used for training, and with as many folds as there are samples, a procedure that yields a (nearly) unbiased estimate of the true prediction error[16]. We note that, although the CV error estimate is obtained using all available samples in the training set, it does not result in a unique predictor, because each fold generates a different one[16]. However, a unique predictor can be constructed — for instance, by averaging the estimated weights over all m predictors, ranking the features according to their average weight and then selecting the number given by the previously inferred optimal complexity parameter[16,31,32]. Alternatively, one could learn a new model using all data but with the complexity parameter now fixed at the optimal value estimated from the previous CV procedure.

A common pitfall that arises when implementing the CV procedure is to run it on a reduced subset of features, which have previously been selected using all training set samples according to how well they correlate with the phenotype of interest. Although the feature rankings and weights are re-estimated for each CV fold, the feature space has already been preselected using all samples, and would lead to severely inflated performance values[33,34]. Likewise, performing dimensionality reduction by preselecting features based on a large number of bootstraps[35] from the training set will also lead to bias: although each bootstrap will use a different subset of training set samples to perform feature selection, if the final feature selection set is obtained by selecting those that are consistently chosen over a large number of bootstraps, this is largely equivalent to using all training set samples for feature selection, and will therefore also lead to overoptimistic performance estimates. As pointed out

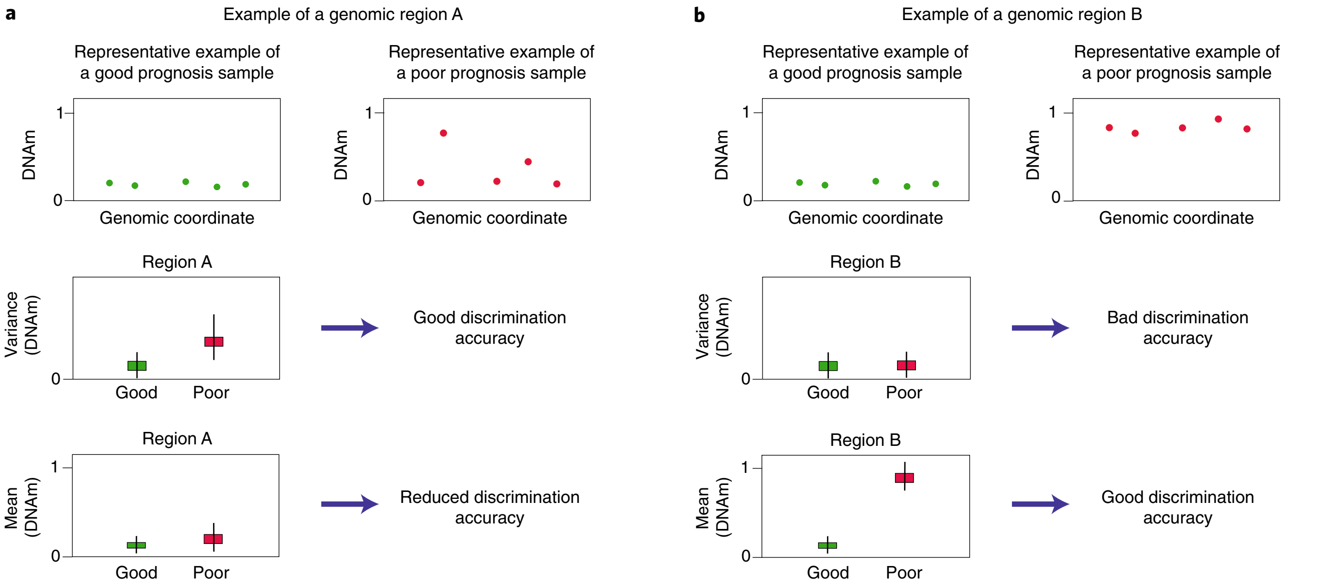

Fig. 4 | Avoiding bias when comparing feature selection methods. The figure is intended to provide a concrete example of how feature selection methods are often not compared in an unbiased fashion. **a**, For instance, we may select features (genomic regions) that differ between good and poor prognosis cancer samples based on the variation in DNA methylation (DNAm), a covalent modification of DNA that can control gene expression, within the given genomic region. Depicted is an example of one genomic region with five DNAm measurements within a given selected region A, across one representative good prognosis and one representative poor prognosis sample. Because of how the feature was selected, the DNAm variance between good and poor prognosis samples provides a good discrimination of the two phenotypes. The same selected region would, however, not show as good a discrimination in terms of its average DNAm. It would be wrong, however, to conclude that selecting features based on intraregional variability is better. **b**, As **a** but now selecting genomic regions that differ between the two phenotypes in terms of their average DNAm. Now, a given selected region B would discriminate the two phenotypes based on the mean DNAm but would not discriminate them in terms of their DNAm variance. Thus, an objective comparison of the two feature selection methods requires both methods to be applied separately, to derive potentially different sets of genomic regions, from which then separate predictors would be constructed and assessed on independent validation sets.

previously[16,31], these pitfalls can be avoided if a nested CV loop strategy is adopted, whereby an inner CV loop is used to tune the classifier parameters, while an outer CV loop is used to estimate the error or classification accuracy (Fig. 2c). This nested CV loop strategy thus separates the process of parameter estimation from the process of error estimation and model selection, which is necessary to avoid reporting biased performance estimates[31].

Another common pitfall relates to an underappreciation of the effect of confounders in omic data, which can make interpretation of results extremely tricky. Even if a predictor derived in the training set performs similarly in the CV as in the validation set, we can't be certain that no overfitting has occurred. The need for caution arises specially when training and validation sets are generated together as part of the same study, which is the most common scenario, and where an unknown confounder that is correlated with the phenotype of interest could easily drive 'consistent' performance between training and validation sets (Fig. 3). This is because randomly splitting the set into a training and test set does not necessarily remove the effect of the confounder. In such a scenario, a good class prediction performance in both sets would be driven by the correlation between confounder and phenotype. As a concrete example, cases and controls may not be matched for ethnic group or age, in which case selected features would be capturing unwanted genetic or ageing effects, and not effects associated with the disease defining case/control status. For this reason, it is now widely appreciated that a good gold-standard validation set is always one generated by an independent lab using an identical or largely equivalent technology and profiling a largely equivalent clinical cohort[36]. The significant advantage of using such a gold-standard validation set is that it will help determine if confounding by study-specific factors in the discovery set has occurred. For instance, in the example above, the gold-standard validation set may be from one ethnic group, or age would be matched between cases and controls, in which case the predictor derived in the discovery set may fail to validate, signalling a potential problem in the discovery set. While confounding by age or ethnicity can be easily spotted and in principle also easily corrected for, confounding factors are often unknown, not recorded, or not easily identifiable[14,15]. For instance, time between sample collection and storage, type of surgery or protocol used to collect blood or buccal swabs may affect molecular measurements, yet these factors may not have been recorded or are not publicly available[37]. Although advanced statistical methods have been developed to address the problem of unknown or unrecorded confounders[38], these methods also assume that the underlying statistical models are faithful representations of the underlying data distributions, an assumption that is often not satisfied and can lead to residual confounding[15]. For instance, relations between response and predictor variables may be highly nonlinear[39] and dependent on the actual features, whereas state-of-the-art statistical methods for removing confounding variation are linear, or largely assume that the functional relationship with the response variable is common to all features[40].

When predictors fail to validate in independent datasets, it is important to point out, however, that an independently generated dataset may also not constitute a good gold-standard. For instance, if the technology used in the validation only measured a fraction of the genomic features in the discovery set, not all features in the predictor may be available, and thus no objective evaluation of the predictor is possible. Alternatively, the validation set samples may derive from a different ethnic group, or may have undergone a different type of medical treatment. In practice, determining whether predictors fail because of confounding variation in the discovery set, or simply because of a suboptimal validation set, can be extremely difficult, especially if not all relevant variables are made public.

The need for a gold-standard validation set is particularly pertinent in studies that aim to demonstrate improved performance or 'added value' over the current state-of-the-art. A common pitfall that arises here is the comparison of a new omic predictor to an established one, using for evaluation only a validation set derived from the same omic study. Because of residual or unknown confounders, any such comparison is unlikely to be objective, naturally favouring the new predictor. For instance, one may derive a prognostic gene expression signature for breast cancer using a large cohort of patients with matched gene expression data[9] and a comparison of interest would be to an established prognostic index that uses orthogonal prognostic information (for example, age, tumour size, tumour stage and so on)[41]. Because the latter index was not trained on the clinical samples used to derive the gene expression based predictor, comparing the two prognostic indices on a left-out validation set from the expression study, is likely to favour the gene-expression-based predictor, due to residual confounding effects.

A related pitfall also arises when assessing the statistical significance of the improvement itself. For instance, the classification accuracy achieved by a new predictor may be higher than that of the clinically established method, yet if the accuracies of both methods are close, there may not be any statistical evidence for claiming an improvement. Because studies are often underpowered, the need also arises to compare predictors across multiple independent datasets. Thus, a significant difference between predictors may not be achieved in individual studies, but a meta-analysis over multiple studies increases power and may indicate statistical significance. In this context, and as a rule of thumb, demonstration of improvement across five to six studies may be necessary to rigorously claim improvement over the current state of the art. Indeed, under the null hypothesis that a predictor A is not better than an existing predictor B, the probability that a predictor A is found to do better than a predictor B across six independent studies is as low as 0.001 under a standard non-parametric test. While obtaining data from five to six independent but equivalent studies can be prohibitively difficult, any comparison across a smaller number of studies should therefore be interpreted with caution.

While our discussion so far has focused on supervised analyses, it is worth highlighting that similar pitfalls are also encountered in the unsupervised context. One of the most common ones is to perform unsupervised clustering of samples over features that were selected in a supervised manner using the same, or a proportion of the same, samples[10,21]. Even if the features themselves are not truly associated with the phenotype of interest — for instance, if the features have small *P*-values but don't pass genome-wide significance levels — inferred clusters will still automatically segregate samples according to the phenotype used to select them. If indeed none of the features pass genome-wide significance levels, the observed segregation is unlikely to be statistically meaningful and can't be used as supportive evidence for the feature selection procedure. Related pitfalls can also occur when benchmarking the quality of a novel feature selection method against other methods, as recently discussed[10,21]. For instance, if we were to select features using a particular method on a set of samples, then subsequent clustering of the samples over these features should yield a better segregation of the phenotype compared to features that were not selected in the same fashion. As a concrete example, one may try to predict prognosis of a cancer by using a training set to identify genomic regions where DNA methylation (DNAm) variation in good prognosis patients is low while being high in poor prognosis individuals[42] (Fig. 4a). If instead of using variability one uses the average DNA methylation level of these same regions to predict prognosis, the predictive power may be substantially reduced, but only because the original genomic regions were not selected using average DNAm as the marker (Fig. 4a). In this particular instance, an objective comparison would require a new set of genomic regions to be selected using their average DNAm levels (Fig. 4b), so that the predictive accuracies of the two feature selection methods can be meaningfully compared.

In summary, although omic data science is now over two decades old, lessons need to be finally learned to avoid the common pitfalls that continue to propagate within the field. The exposition of some of the key statistical issues presented here may help towards this goal and may apply more broadly to the whole spectrum of big data science, including material science. ❐

Andrew E. Teschendorff[1,2]

[1]*Statistical Cancer Genomics, UCL Cancer Institute and Department of Woman's Cancer, University College London, London, UK.* [2]*CAS Key Lab of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institute for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China.*
e-mail: *a.teschendorff@ucl.ac.uk*

Published online: 26 November 2018
https://doi.org/10.1038/s41563-018-0241-z

References

1. Kalinin, S. V., Sumpter, B. G. & Archibald, R. K. *Nat. Mater.* **14**, 973–980 (2015).
2. Marx, V. *Nature* **498**, 255–260 (2013).
3. Mattmann, C. A. *Nature* **493**, 473–475 (2013).
4. Fodor, S. P. et al. *Science* **251**, 767–773 (1991).
5. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. *Science* **270**, 467–470 (1995).
6. Perou, C. M. et al. *Proc. Natl Acad. Sci. USA* **96**, 9212–9217 (1999).
7. Wheeler, D. A. et al. *Nature* **452**, 872–876 (2008).
8. Nagalakshmi, U. et al. *Science* **320**, 1344–1349 (2008).
9. van 't Veer, L. J. et al. *Nature* **415**, 530–536 (2002).
10. Guo, S. et al. *Nat. Genet.* **49**, 635–642 (2017).
11. Gerlinger, M. et al. *N. Engl. J. Med.* **366**, 883–892 (2012).
12. Xu, R. H. et al. *Nat. Mater.* **16**, 1155–1161 (2017).
13. Storey, J. D. & Tibshirani, R. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
14. Leek, J. T. et al. *Nat. Rev. Genet.* **11**, 733–739 (2010).
15. Teschendorff, A. E., Zhuang, J. & Widschwendter, M. *Bioinformatics* **27**, 1496–1505 (2011).
16. Simon, R., Radmacher, M. D., Dobbin, K. & McShane, L. M. *J. Natl Cancer Inst.* **95**, 14–18 (2003).
17. Ioannidis, J. P. *PLoS Med.* **2**, e124 (2005).
18. Jager, L. R. & Leek, J. T. *Biostatistics* **15**, 1–12 (2014).
19. Sebastiani, P. et al. *Science* **333**, 404 (2011).
20. Ioannidis, J. P. et al. *Nat. Genet.* **41**, 149–155 (2009).
21. Seoighe, C., Tosh, N. J. & Greally, J. M. *Nat. Genet.* **50**, 1062–1063 (2018).
22. Jacob, L. & Speed, T. P. *Genome Biol.* **19**, 97 (2018).
23. Nieuwenhuis, S., Forstmann, B. U. & Wagenmakers, E. J. *Nat. Neurosci.* **14**, 1105–1107 (2011).
24. Qin, L. X., Huang, H. C. & Begg, C. B. *J. Clin. Oncol.* **34**, 3931–3938 (2016).
25. Ernst, J. & Kellis, M. *Nat. Biotechnol.* **33**, 364–376 (2015).
26. Vapnik, V. N. *Statistical Learning Theory* (Wiley, New York, 1998).
27. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, New York, 2006).
28. Friedman, J., Hastie, T. & Tibshirani, R. *J. Stat. Softw.* **33**, 1–22 (2010).
29. Webb, S. *Nature* **554**, 555–557 (2018).
30. Bishop, C. M. *Neural Networks for Pattern Recognition* (Oxford Univ. Press, Oxford, 1995).
31. Varma, S. & Simon, R. *BMC Bioinform.* **7**, 91 (2006).
32. Teschendorff, A. E. et al. *Genome Biol.* **7**, R101 (2006).
33. Ambroise, C. & McLachlan, G. *J. Proc. Natl Acad. Sci. USA* **99**, 6562–6566 (2002).
34. Reunanen, J. *J. Mach. Learn. Res.* **3**, 1371–1382 (2003).
35. Efron, B. & Tibshirani, R. *J. J. Am. Stat. Assoc.* **92**, 548–560 (1997).
36. Simon, R. *J. Natl Cancer Inst.* **97**, 866–867 (2005).
37. Biton, A. et al. *Cell Rep.* **9**, 1235–1245 (2014).
38. Leek, J. T. & Storey, J. D. *PLoS Genet.* **3**, 1724–1735 (2007).
39. Horvath, S. *Genome Biol.* **14**, R115 (2013).
40. Leek, J. T. & Storey, J. D. *Proc. Natl Acad. Sci. USA* **105**, 18718–18723 (2008).
41. Galea, M. H., Blamey, R. W., Elston, C. E. & Ellis, I. O. *Breast Cancer Res. Treat.* **22**, 207–219 (1992).
42. Bartlett, T. E. et al. *PLoS ONE* **10**, e0143178 (2015).