Proper reporting of predictor performance

To the Editor: In many fields, including the study of genetic variation, prediction methods are essential for interpreting experimental data, and it is important to present their performance in a systematic way. Recently, Kumar *et al.*[1] published a Correspondence about the use of evolutionary information to predict the consequences of amino acid substitutions. The authors claimed that machine-learning classifiers would benefit from training separately at different amino acid conservation levels in order to better predict harmful protein variants.

The approach might be useful, but it is difficult to judge as its performance is reported in a defective and partly misleading way. Several measures are needed to fully capture method performance[2,3]. In the Correspondence[1] some of those measures were used, but a number of important details were omitted. The greatest problem relates to the use of the Matthews correlation coefficient (MCC), one of the most widely used measures for binary predictor performance. The MCC is based on true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values in a contingency table, with the accepted definition expressed as:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

In contrast, Kumar *et al.*[1] used ratios of the four values in their formulation. They also converted the incorrectly calculated MCC values to percentages, but only for the positive half of the values, thereby not considering their full range from –1 (perfect disagreement) to 1 (perfect agreement). The correct values are listed in **Table 1** and affect the conclusions of the work in ref. 1. When the results are combined for the conservation classes ('total'; **Table 1**), it is evident that EvoD is overall the poorest of the tested methods.

Table 1 | Corrected MCC values

Method	Evolutionary conservation	Ratio[a]	Original MCC[b]	Corrected MCC[c]
EvoD	Ultra	0.10	39%	0.24
	Well	0.65	45%	0.45
	Less	5.38	41%	0.30
	Total	0.91	NR	0.42
Condel	Ultra	0.10	21%	0.20
	Well	0.65	38%	0.40
	Less	5.38	30%	0.22
	Total	0.86	NR	0.51
PolyPhen-2	Ultra	0.10	26%	0.20
	Well	0.68	45%	0.45
	Less	5.71	31%	0.28
	Total	0.86	NR	0.63

[a]Ratio of positive to neutral variants in the test set. Ratios deviating from 1 indicate an imbalance. [b]Original MCC from ref. 1. [c]MCC calculated without correcting for class imbalance as it is a very robust measure and can be applied except to extremely biased distributions. NR, not reported.

The use of erroneous and misleading performance parameters prevents readers from obtaining a true idea of the qualities of a method. Evaluation of machine-learning methods has three prerequisites[2]: (i) there have to be sufficient numbers of known positive and negative cases available, for example, in the VariBench database for variation benchmark datasets[4]; (ii) proper measures have to be used for method assessment, and the class imbalance (difference in the number of positive and negative cases), if present, needs to be corrected; and (iii) training and test datasets should be disjoint.

Kumar *et al.*[1] did not address class imbalance, and did not report whether data used for training their EvoD method were also used for testing. Thus, the performance data they cite may actually indicate how well the EvoD method learned the training data rather than how well it will perform on independent test data. Condel and PolyPhen2 have been trained with the same cases that are now used for testing the performance. In their analysis, the authors also did not include methods that have been shown in a systematic comparison to have superior performance[5].

Sequence conservation is known to be an important feature for variation predictors. The results in **Table 1** show, contrary to the conclusion of the Correspondence[1], that variations at ultra-conserved and less conserved sites are considerably less reliably predicted than those at well conserved sites by all the three tested methods.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

Mauno Vihinen

Department of Experimental Medical Science, Lund University, Lund, Sweden.
e-mail: mauno.vihinen@med.lu.se

1. Kumar, S. *et al. Nat. Methods* **9**, 855–856 (2012).
2. Vihinen, M. *BMC Genomics* **13** (suppl. 4), S2 (2012).
3. Vihinen, M. *Hum. Mutat.* **34**, 275–282 (2013).
4. Nair, P.S. & Vihinen, M. *Hum. Mutat.* **34**, 42–49 (2013).
5. Thusberg, J., Olatubosun, A. & Vihinen, M. *Hum. Mutat.* **32**, 358–368 (2011).

Kumar *et al.* reply: We disagree with Vihinen's[1] suggestion that the performance of the EvoD[2] method based on evolutionary stratification of prediction models was not evaluated correctly, and we affirm the importance of the method. The need for evolutionary stratification arose because we discovered that existing methods exhibited a very high rate of false positive diagnoses for variants occurring at the most highly conserved positions (ref. 2 Table 1). We had observed a high rate of false negatives for variants found in positions that have evolved the fastest[2]. These discoveries established the biological pitfalls of existing approaches, all of which fit a single prediction model that is agnostic to differences in evolutionary conservation among positions.

By considering ultra-, well- and less-conserved positions separately[2], the variant prediction models become biologically

realistic and overcome the high error rates; thus, classifier performance combined across all classes is not relevant. Vihinen[1] may disagree with this approach, but does not provide a reason for why one should fit a single model to positions, when such a model has been shown to suffer from a high degree of misdiagnosis for subsets of variants that can be determined a priori using the long-term evolutionary conservation of positions. In fact, our results provide strong evidence that machine-learning classifiers must incorporate all known biological knowledge and, in the current case, embrace differences in the level of functional importance of positions when assembling training and testing datasets[2,3].

In developing classifiers for each evolutionary conservation category, we selected equal numbers of positive and negative controls (balanced datasets) and performed model construction and evaluation with disjoint training and testing data (ref. 2 Supplementary Methods). For these balanced datasets, the Matthews correlation coefficient (MCC) values calculated on absolute counts showed that EvoD performs better than other methods. This is consistent with recent findings that prediction models achieve the best results on the combined use of balanced training sets and balanced testing sets[4]. Indeed, the presence of unequal numbers of positive and negative controls in the training data leads to over-optimizing of true positive rate at ultraconserved sites and over-optimizing of true negative rate at less-conserved sites when a single model is generated for all positions[3], which is remedied through the development of evolutionary stratified prediction models[2,3]. It is important to note that the MCC should be calculated on balanced datasets for each model[4–6], so our use of ratios in the MCC formula is equivalent to computing the MCC using random sampling of 100 positive and 100 negative cases from the complete dataset. Therefore, Vihinen's concerns[1] about EvoD performance are unfounded.

Both the ratios and absolute counts (ref. 1 Table 1) show superior performance for EvoD in every conservation category. Furthermore, the ratio-based MCC is more consistent across categories (0.39–0.45) than the count-based MCC that varies by twofold (0.24–0.45). These observations demonstrate the advantage of ratio-based MCC when it is applied to highly unbalanced test datasets[3,5] and invalidate Vihinen's conclusion[1] that the accuracy of the prediction of ultraconserved and less-conserved sites are lower.

ACKNOWLEDGMENT

This work is supported by research grants from the US National Institutes of Health (HG002096-12 and LM010834-04).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Sudhir Kumar[1,2], Jieping Ye[2,3] & Li Liu[1,2]

[1]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, Pennsylvania, USA. [2]Center for Evolutionary Medicine and Informatics, Arizona State University, Tempe, Arizona, USA. [3]School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, Arizona, USA
e-mail: s.kumar@temple.edu

1. Vihinen, M. Proper reporting of predictor performance. *Nat. Methods* **11**, 781 (2014).
2. Kumar, S. *et al. Nat. Methods*, **9**, 855–856 (2012).
3. Liu, L. & Kumar, S. *Mol. Biol. Evol.* **30**, 1252–1257 (2013).
4. Wei, Q. & Dunbrack, R.L. *PLoS ONE* **8**, e67863 (2013).
5. Eiland, E.E. & Liebrock, L.M. *Adv. Artificial Intelligence* **2013**, 427958 (2013).
6. Obayashi, T. & Kinoshita, K. *DNA Res.* **16**, 249–260 (2009).

Predictor performance with stratified data and imbalanced classes

To the Editor: The disagreement between Vihinen[1] and Kumar *et al.*[2] over the presentation of EvoD[3] raises important issues relevant to any binary classifier, including the problems of class imbalance, what constitutes an appropriate performance metric and the legitimacy of training on stratified data. Researchers need to be aware that competing strategies to calibrate and evaluate a classifier may lead to differing perceptions of performance.

A binary classifier assigns observations to one of two categories; class imbalance occurs when one category contains many more observations than the other, and it can distort performance metrics. For example, EvoD attempts to classify HumVar variants in ultraconserved positions, 91.2% of which have known disease associations. Thus, a simple rule that predicts all such variants to be disease-associated achieves 91.2% accuracy, which is greater than that of EvoD, Condel and PolyPhen-2. But is greater accuracy synonymous with outperformance? Consider a diagnostic test that always identifies an individual affected by a rare disease. Even if the false positive rate is low, say 2%, for a disease with 1% prevalence, a more accurate 'test' would predict all individuals to be disease-free. Thus, accuracy alone is insufficient to capture the utility of a binary classifier, particularly for imbalanced classes. Vihinen[1] and Kumar *et al.*[2] disagree on what would be sufficient in its stead.

Binary classifier performance depends on four counts: the numbers of true negatives (TN), false positives (FP), false negatives (FN) and true positives (TP), which Kumar *et al.*[3] report in Table 1. Vihinen[1] used these data to calculate his chosen metric, the Matthews correlation coefficient (MCC), and objects to the nonstandard version of the MCC (balanced MCC; BMCC[4]) used by Kumar *et al.*[3]. The BMCC effectively reweights observations depending on their class. To see this, consider the ratio (r) of neutral to disease-associated variants reported by Vihinen[1] in his Table 1 ($r = (TN + FP)/(FN + TP)$). The BMCC reduces to the MCC if the number of neutral variants is multiplied by $(1 + 1/r)/2$ while disease variants are multiplied by $(1 + r)/2$ (**Supplementary Note**). As compared to the MCC, the BMCC exaggerates the degree to which EvoD outperforms its competitors, but both metrics still favor EvoD. The choice of metric is ultimately subjective, and the BMCC is not an unreasonable measure of performance, but its use invites criticism. For ultraconserved positions, $r = 0.1$, meaning that the BMCC upweights predictions on neutral variants (the underrepresented class) by an order of magnitude. As the MCC is already robust to class imbalance, the implicit reweighting of the BMCC might be deemed an unnecessary manipulation, and it should have been discussed in the original paper[3].

Vihinen[1] further questions how Kumar *et al.*[3] treat stratification. Both sides acknowledge the importance of positional conservation to classifying molecular variants; Kumar *et al.*[3] leverage this by calibrating and evaluating their method independently for each of three conservation levels that they define. Positional conservation is a quantitative score, leaving room to debate whether and how it should be stratified; nevertheless, once stratification is established, Kumar *et al.*[3] treat the data appropriately by conditioning on positional conservation, effectively creating three independent versions of EvoD. Positional conservation is known in most scenarios so that