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