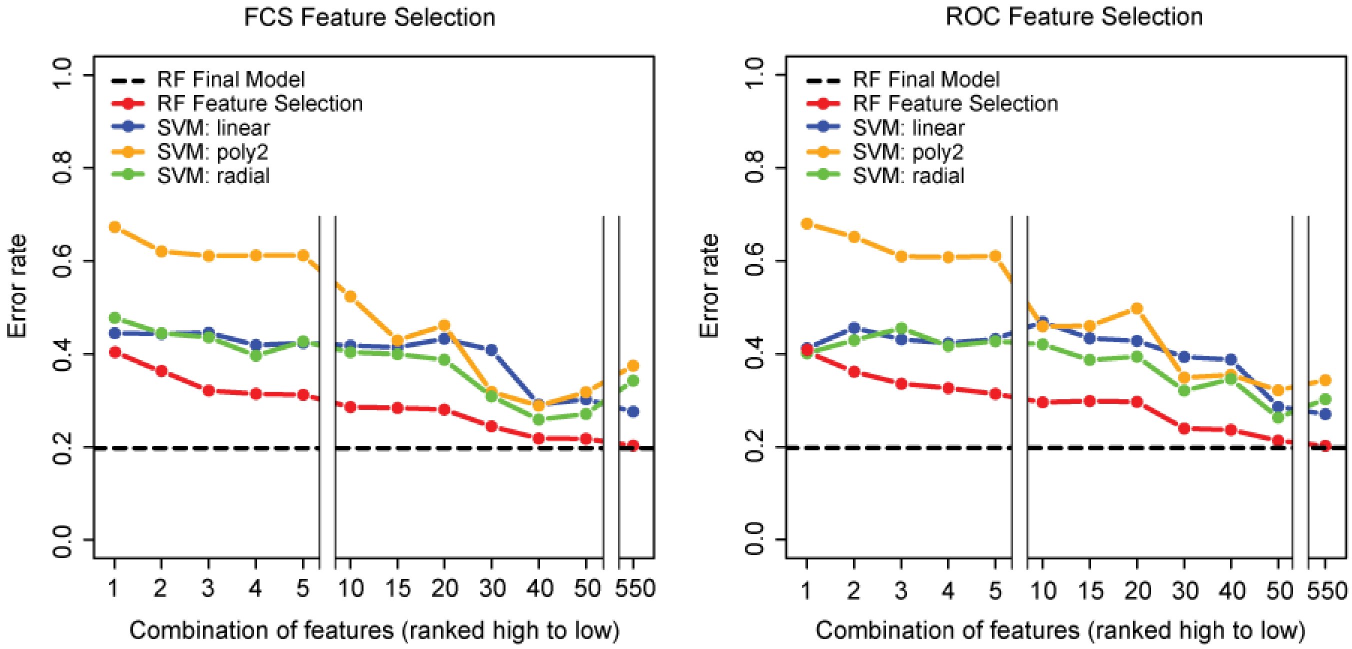
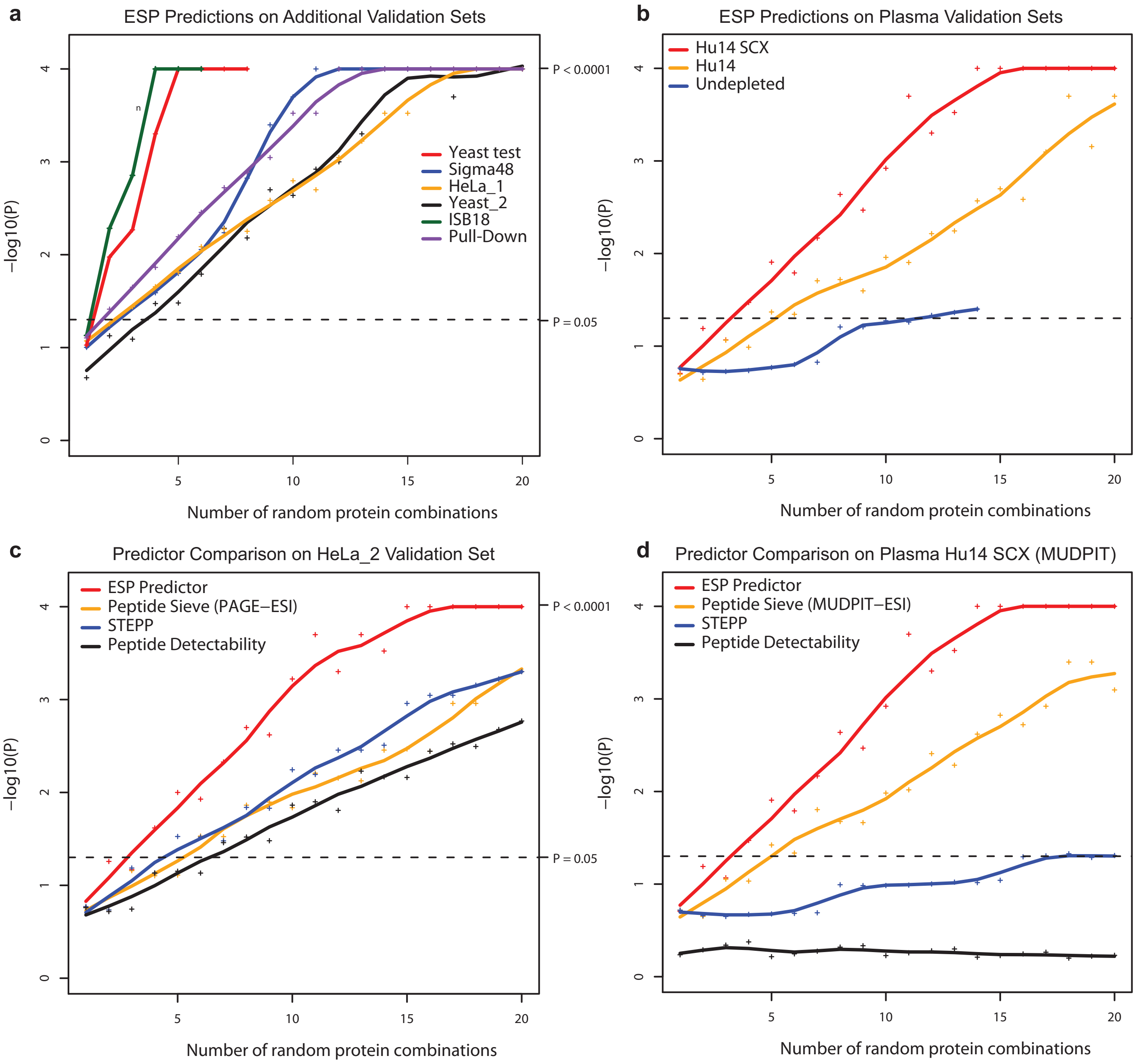
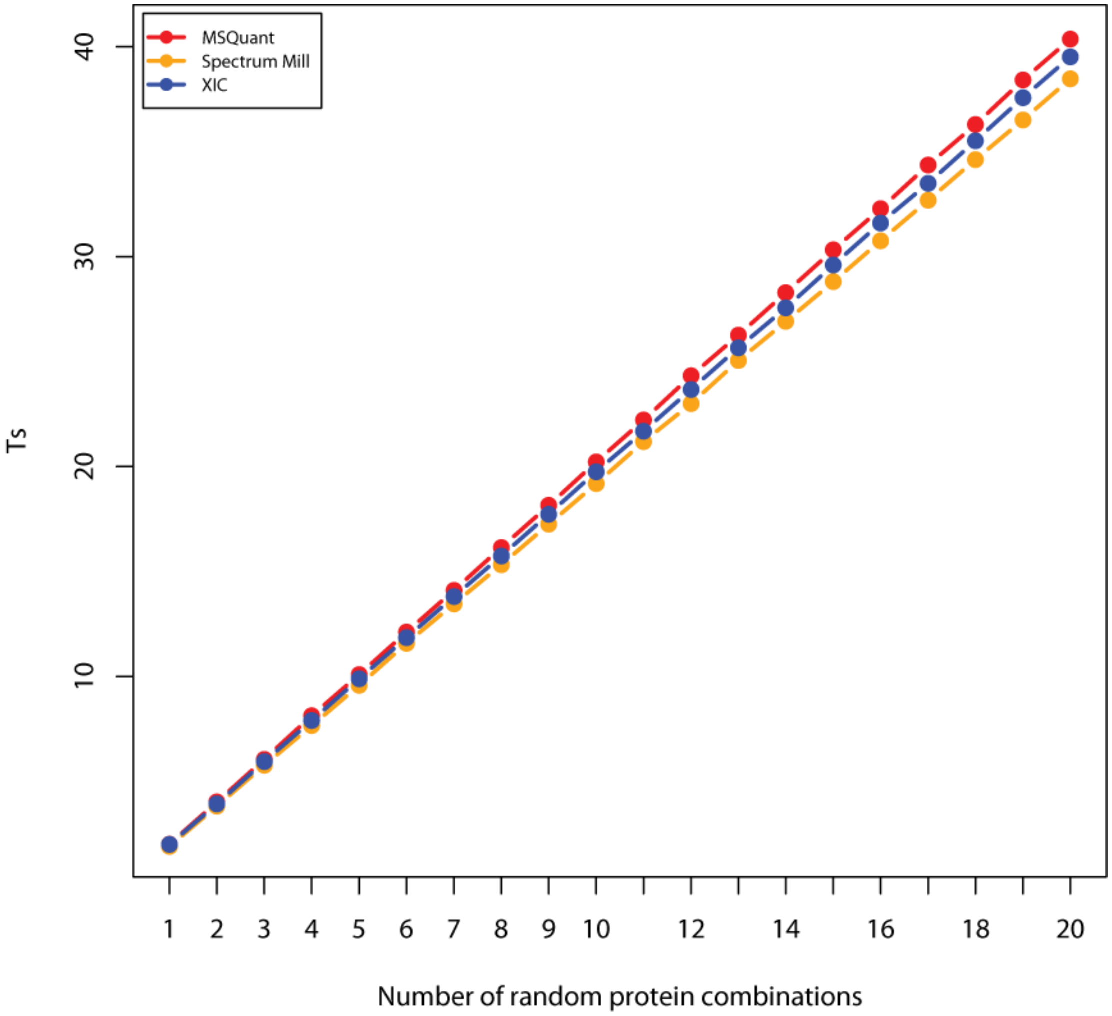


Supplementary Figure 1 Random Forest (RF) and SVM comparison with feature selection. We performed two types of feature selection to determine if we could reduce the number of features in the model. First, we used the parametric Fisher Criterion Score (Webb-Robertson et al., 2008). This allowed us to rank the features based on the mean distance between class distributions from largest to smallest distance. Second, we used the non-parametric receiver operator curve (ROC) to rank features based on the area under the curve (AUC) from largest to smallest AUC. We then compared the performance of SVM and RF using increasing numbers of the best features. For both classifiers, we selected balanced class sizes equal to the number of *high* responding peptides (n = 623). For the SVMs, we built three classifiers with different kernals – linear, polynomial degree 2, and radial and measured the error rate using 5-fold cross validation. For RF, we used 1,000 trees and the default mtry (square root of the number of features) and measured the error rate using the out-of-bag data. The dotted line represents the performance of the RF model used in the ESP predictor (main paper). RF performs considerably better than all SVM kernels and exhibits the best performance when all 550 features are included in building the model. This further demonstrates that RF is able to handle many features without impacting the error rate.

Nature Biotechnology: doi:10.1038/nbt.1524

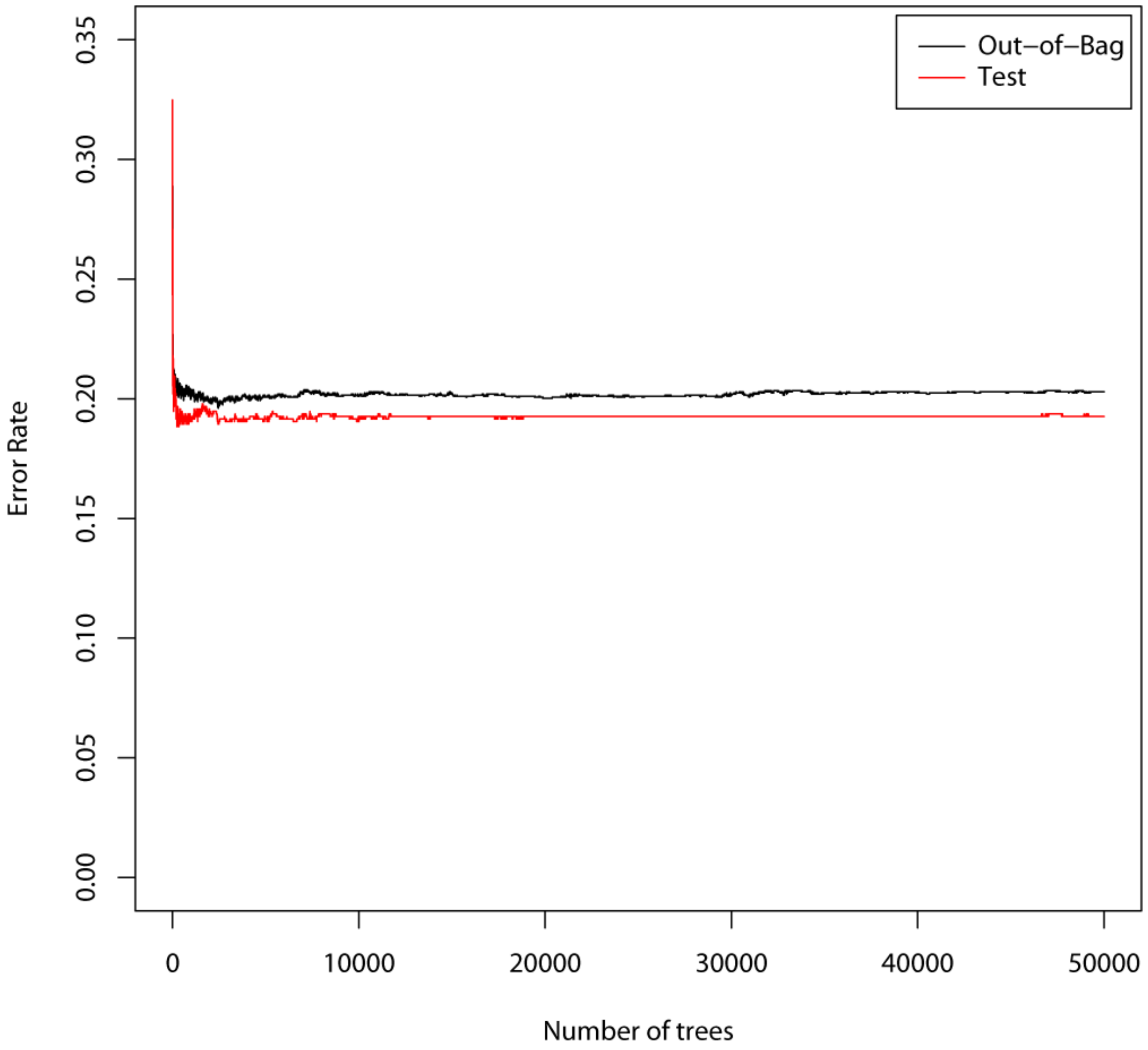


Supplementary Figure 2 ESP predictor method comparison and model validation significance plots. We statistically confirmed the hypothesis that that the ESP predictions are better than random predictions and existing methods by calculating the cumulative probability using a permutation test. For each validation set, we calculated Ts, the cumulative number of correctly predicted peptides out of the five experimentally derived highest responding peptides, for increasing numbers of proteins (chosen at random). The null distribution for P-value calculation is derived using a predictor that selects the top five high responding peptides for a protein at random. We repeated this process 10,000 times for random combinations of 1 - 20 proteins. To calculate the P-value, we compared the fraction of values from the null distribution that were greater than or equal to the mean Ts for each random protein combination (1 – 20). We applied a loess regression model to fit curves to the points. We considered predictions significant if the p-value was less than 0.05 (dashed line). (**a**) ESP predictor performance on multiple validation sets. (**b**) ESP predictions on plasma validation sets. The samples represent undepleted plasma, top 14 most abundant proteins depleted, and depleted and then fractionated using SCX (also referred to as MUDPIT). (**c**) Comparison between the ESP predictor and existing methods on a HeLa GeLC-MS cell lysate. (**d**) Comparison between the ESP predictor and existing methods on a depleted and fractionated plasma sample (also referred to as MUDPIT). This is the sample type most commonly used for MRM biomarker verification. (c – d) Refer to **Table 2** for more details. (a – d) Refer to **Table 1** for more details.

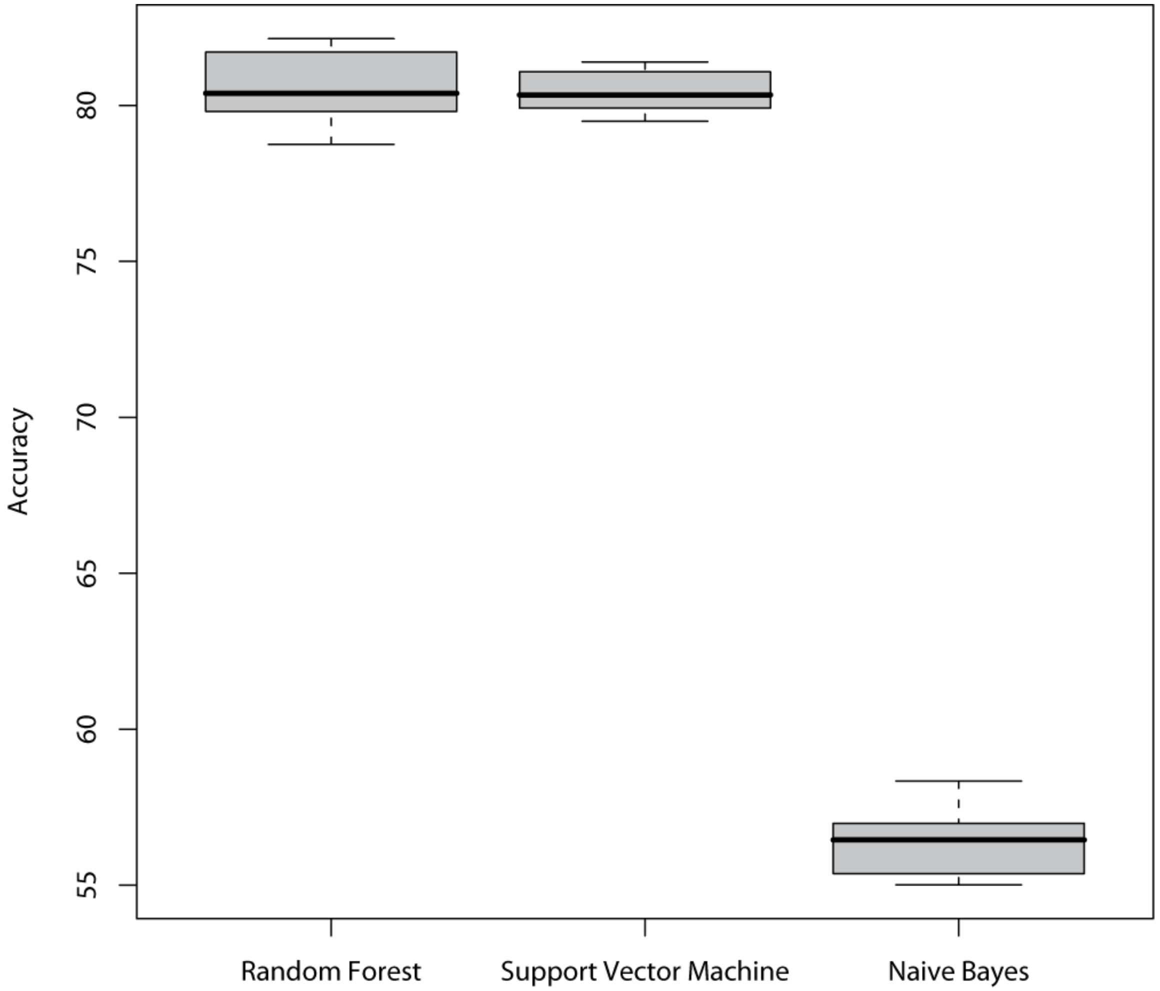


Supplementary Figure 3 Comparison of different peptide quantitation methods. The raw files from the HeLa_1 (in-solution digest of HeLa cell lysate) was searched using Mascot and Spectrum Mill to determine peptide identities. We quantified the peptides from the Mascot search using MSQuant. We used the Spectrum Mill peptide intensity for identified peptides. The Spectrum Mill intensity is calculated using a narrow m/z window around all of the isotopes theoretically expected to have intensity >17% of the most abundant isotope. Finally, we calculated the XIC of the monoisotopic peak using an in-house developed program to automate XIC calculation. The results are very consistent and demonstrate the ESP predictor performs well using different quantitation methods, as long as it is consistent.

Nature Biotechnology: doi:10.1038/nbt.1524



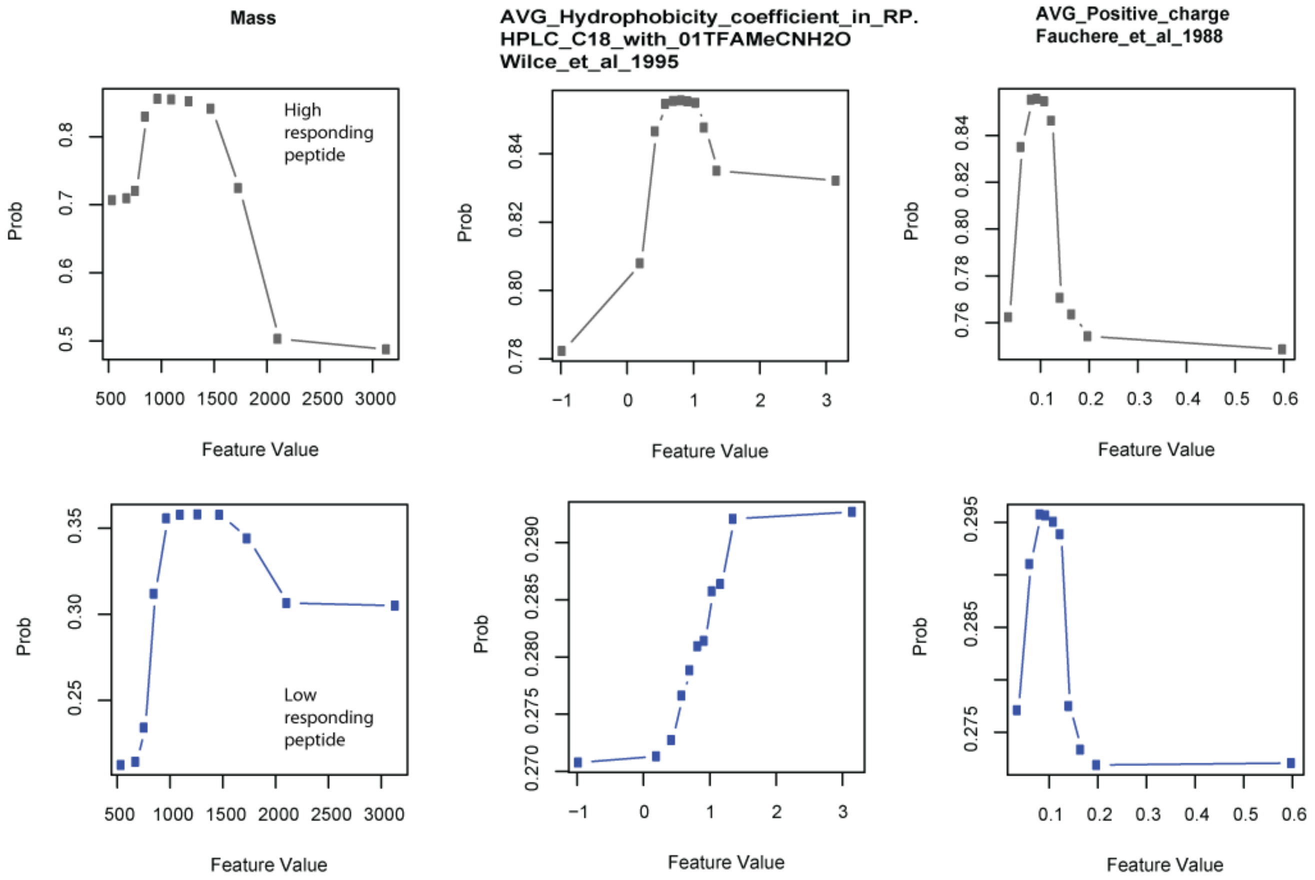
Supplementary Figure 4 Random Forest (RF) does not overfit with an increase in the number of trees contained in the forest. Using the training data, we randomly divided the data into train (70%) and test (30%) sets. The out-of-bag (OOB) is the error associated with the training data and is similar to 5-fold cross validation. The test error is the error rate associated with the test set. Since the test error remains constant as we increase the number of trees we can conclude the model is not overfit. In RF, it is typical to see the OOB error rate higher than the test error rate. This is because the OOB error is estimated using ~1/3 of the trees in the model while the test error uses all trees.



Supplementary Figure 5 Comparison between Random Forest (RF), committee SVM and Naïve Bayes (NB). RF is a committee machine because the final prediction is based on the ensemble of decision trees in the forest (each decision tree is a separate classifier). SVM and NB report predictions based on a single classifier. This difference makes it difficult to compare classifiers fairly. For example, **Supplementary Figure 1a** shows the RF performs better than the single SVM classifier. In order to make a fair comparison, we created a committee of SVM and NB classifiers. Using the original training set with all 550 properties, we split the data into 10 training (70%) and test (30%) sets. Thus, we built 10 models to show a distribution for each classifier. For RF, we used the procedure described in the main paper. For SVM and NB we used the following procedure to calculate the test set accuracy:

1. Begin with train/test set 1
2. Create equal class sizes by randomly sampling from the majority class (low/not detected) to the size of the minority class (high).
3. Build a model, using all the features in the data set, and record the test set error.
4. Repeat steps two and three 20 times
5. Average the errors from step 4 and record that error for train/test set 1
6. Repeat entire process for train/test set 2 through 10.

The committee SVM performs equally well compared to the RF while NB is considerably worse (80.5 ± 1.2%, 80.4 ± 0.6%, 56.5 ± 1.1%, (mean ± SD) respectively). The "out of the box" performance is better for RF (**Supplementary Figure 1a),** although it is possible to create an equally good model using a committee of SVMs. Since the RF classifier comes with all the required machinery out-of-the-box (including calculating feature importance during model building), with supporting theory (Breiman, 2001), we have opted to use RF for predicting high response peptides. Furthermore, the use of multiple trees built using a random subset of the features results in robust RF predictions in the presence of noisy features that do not contribute to improving prediction. We believe the lower SD for SVM is due to the multiple averaging from steps 4 and 5 which is known to reduce the variance.



Supplementary Figure 6 Random Forest (RF) sensitivity analysis. We attempted to understand how the RF is using the most important features by performing a sensitivity analysis. We start with an n-dimensional vector of peptide properties. For each of the top 35 properties (**Fig. 4** in the main paper) we sequentially vary the value of one property, while holding all other properties constant, and record the model output (predicted probability of high response, y-axis). For each property, we varied its value over its quantile range (from 0 – 100%) in steps of 10% based on the property distribution from the yeast training set. We then performed this analysis using a high (top row) and low (bottom row) responding peptide for each of the top 35 properties. We have illustrated above the response for only three of these 35 properties.

The sensitivity analysis indicates that some properties have a significant effect on the output (e.g., Mass). There are also properties that have a more subtle effect on the output, and the plot of feature value against output probability can be either similar (e.g., AVG_Positive_charge) or different (e.g., AVG_Hydophobocity_coefficient_in_RP) for low and high responding peptides. These results indicate that the context established by the values of all the other properties determine the actual output level (e.g., for high responding peptide, prob=0.76 when AVG_positive_charge=0.2, but for the same property value, prob~0.26 for low responding peptides). This context dependence is clearly brought out in all the 3 plots shown above. All this points to the fact that the actual relationships between properties captured by the RF in order to accurately predict high responding peptides is highly non-linear, and cannot be easily dissected and explained. The RF is able to exploit complex interactions between the properties in order achieve accurate prediction, and simplistic representations of the importance or magnitude/effect of individual properties can be misleading.

Fisher Criterion Score Top 35 Properties

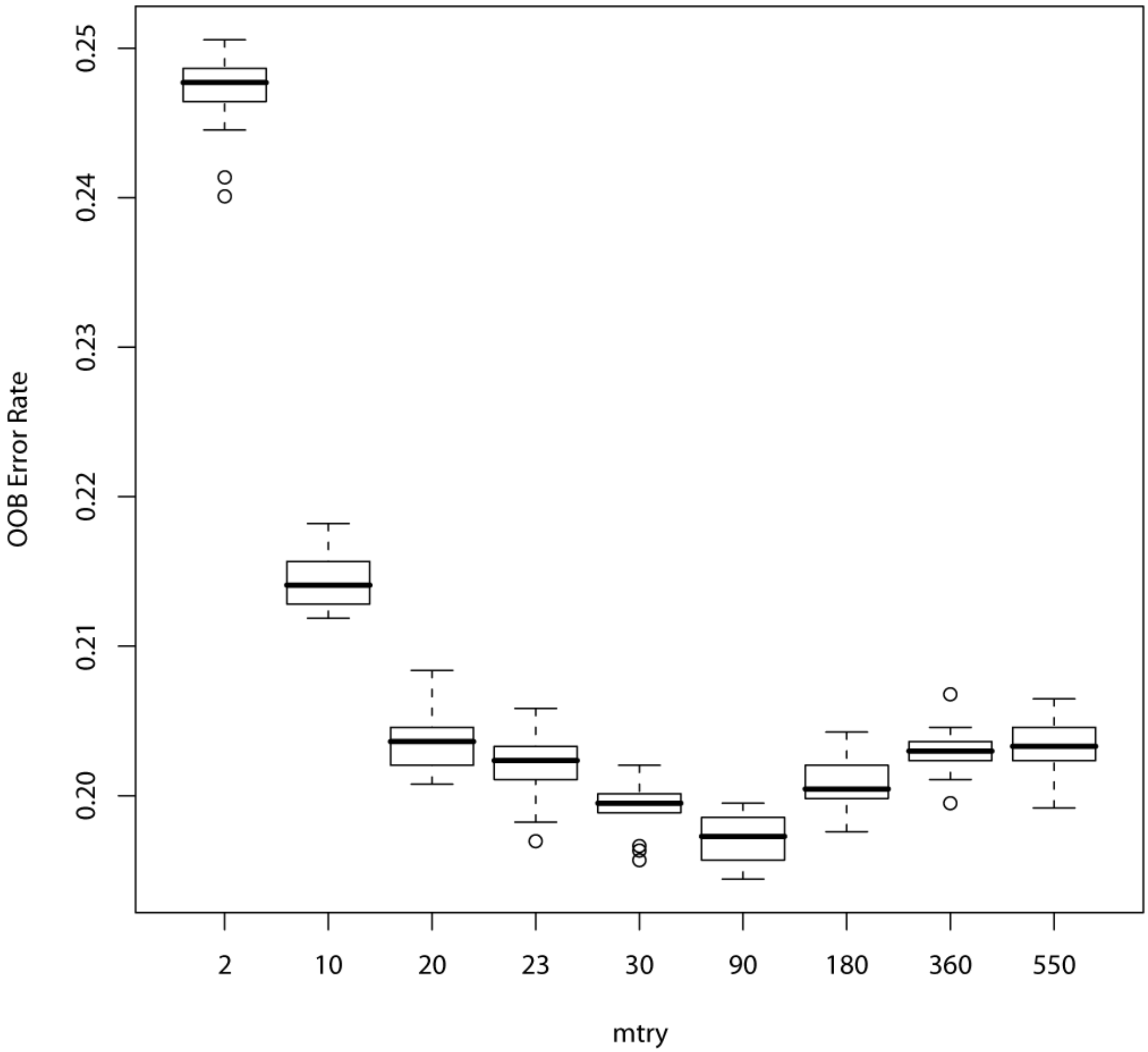
[1] Sequence frequency Jungck 1978
[2] Accessible surface area Radzicka.Wolfenden 1988
[3] Residue accessible surface area in tripeptide Chothia 1976
[4] Molecular weight Fasman 1976
[5] Average volumes of residues Pontius et al 1996
[6] Average volume of buried residue Chothia 1975
[7] Volume Grantham 1974
[8] Mean volumes of residues buried in protein interiors Harpaz et al 1994
[9] Volumes not including the crystallographic waters using the ProtOr Tsai et al 1999
[10] Entropy of formation Hutchens 1970
[11] Volumes including the crystallographic waters using the ProtOr Tsai et al 1999
[12] Residue volume Goldsack.Chalifoux 1973
[13] Side chain volume Krigbaum.Komoriya 1979
[14] Mean area buried on transfer Rose et al 1985
[15] Hydropathy scale based on self.information values in the two.state model 9 accessibility Naderi.Manesh et al 2001
[16] Residue volume Bigelow 1967
[17] Hydropathy scale based on self.information values in the two.state model 50 accessibility Naderi.Manesh et al 2001
[18] Side chain torsion angle phiAAAR Levitt 1976
[19] Absolute entropy Hutchens 1970
[20] Hydropathy scale based on self.information values in the two.state model 5 accessibility Naderi.Manesh et al 2001
[21] Polarity Zimmerman et al 1968
[22] Refractivity McMeekin et al 1964 Cited by Jones 1975
[23] Average accessible surface area Janin et al 1978
[24] Residue accessible surface area in folded protein Chothia 1976
[25] Heat capacity Hutchens 1970
[26] Percentage of exposed residues Janin et al 1978
[27] Percentage of buried residues Janin et al 1978
[28] Gas phase basicity
[29] Hydrophobicity Prabhakaran 1990
[30] Hydropathy scale based on self.information values in the two.state model 16 accessibility Naderi.Manesh et al 2001
[31] mass
[32] Transfer free energy from vap to oct Radzicka.Wolfenden 1988
[33] Side chain angle thetaAAR Levitt 1976
[34] Relative mutability Jones et al 1992
[35] Relative mutability Dayhoff et al 1978b

ROC Top 35 Properties

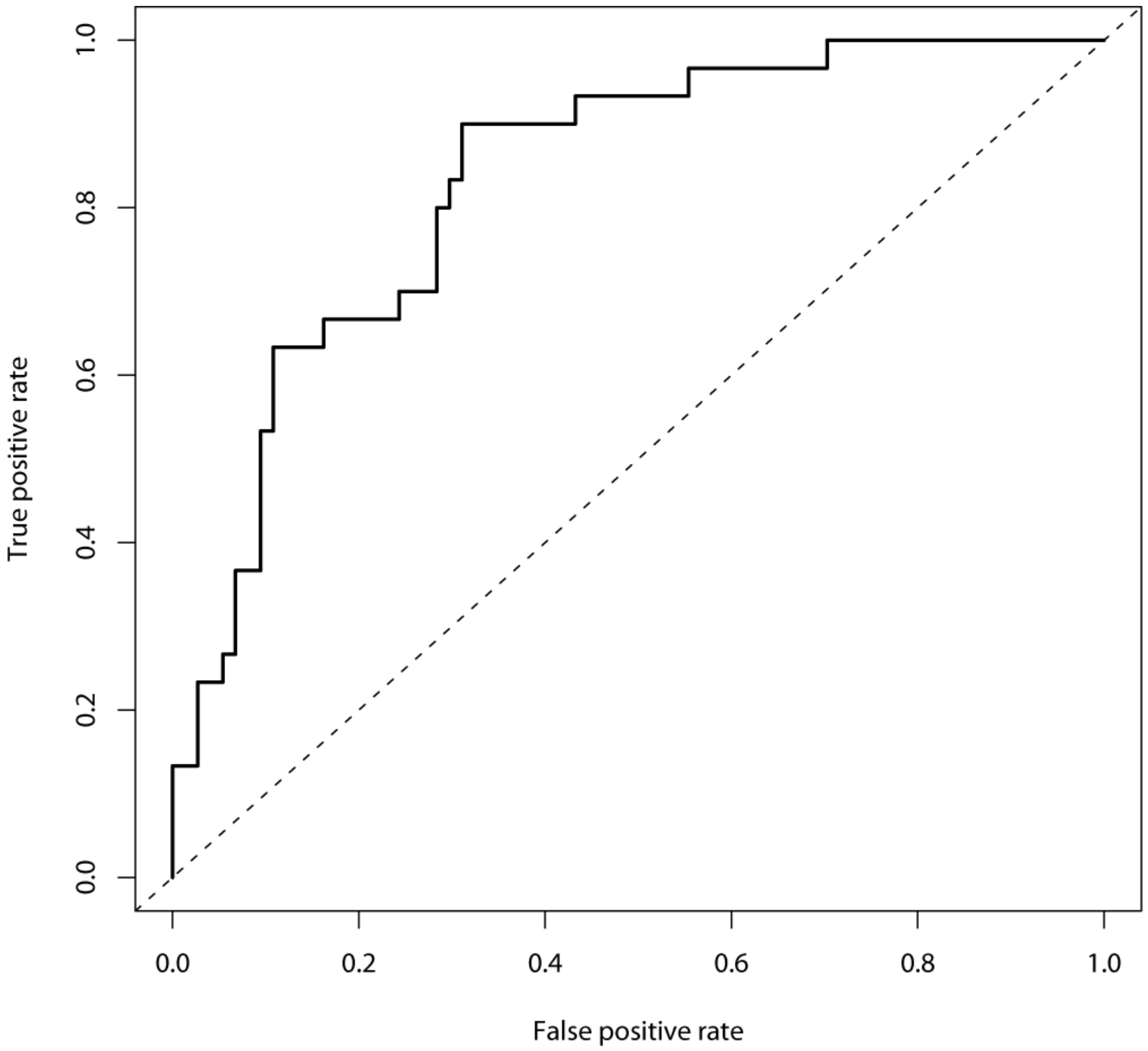
[1] Transfer free energy from vap to chx Radzicka.Wolfenden 1988
[2] Short and medium range non.bonded energy per residue Oobatake.Ooi 1977
[3] Entire chain composition of amino acids in extracellular proteins of mesophiles percent Fukuchi.Nishikawa 2001
[4] Amino acid composition Dayhoff et al 1978a
[5] Distribution of amino acid residues in the 18 non.redundant families of mesophilic proteins Kumar et al 2000
[6] Sequence frequency Jungck 1978
[7] Average non.bonded energy per residue Oobatake.Ooi 1977
[8] Interior composition of amino acids in extracellular proteins of mesophiles percent Fukuchi.Nishikawa 2001
[9] SD of AA composition of total proteins Nakashima et al 1990
[10] AA composition of total proteins Nakashima et al 1990
[11] Relative frequency of occurrence Jones et al 1992
[12] Distribution of amino acid residues in the 18 non.redundant families of thermophilic proteins Kumar et al 2000
[13] Composition of amino acids in extracellular proteins percent Cedano et al 1997
[14] AA composition of membrane proteins Nakashima et al 1990
[15] Amino acid distribution Jukes et al 1975
[16] Interior composition of amino acids in intracellular proteins of mesophiles percent Fukuchi.Nishikawa 2001
[17] Composition of amino acids in anchored proteins percent Cedano et al 1997
[18] Composition of amino acids in membrane proteins percent Cedano et al 1997
[19] Interior composition of amino acids in intracellular proteins of thermophiles percent Fukuchi.Nishikawa 2001
[20] pK.N Fasman 1976
[21] Composition of amino acids in intracellular proteins percent Cedano et al 1997
[22] Entire chain composition of amino acids in intracellular proteins of mesophiles percent Fukuchi.Nishikawa 2001
[23] Transfer free energy from vap to oct Radzicka.Wolfenden 1988
[24] Principal component II Sneath 1966
[25] AA composition of EXT of multi.spanning proteins Nakashima.Nishikawa 1992
[26] Gas phase basicity
[27] Hydrophobicity index Wolfenden et al 1979
[28] Side chain torsion angle phiAAAR Levitt 1976
[29] Composition of amino acids in nuclear proteins percent Cedano et al 1997
[30] Proportion of residues 100 buried Chothia 1976
[31] Hydration potential Wolfenden et al 1981
[32] AA composition of MEM of single.spanning proteins Nakashima.Nishikawa 1992
[33] Energy transfer from out to in95buried Radzicka.Wolfenden 1988
[34] AA composition of EXT of single.spanning proteins Nakashima.Nishikawa 1992
[35] Transmembrane regions of non.mt.proteins Nakashima et al 1990

Supplementary Figure 7 Top 35 physicochemical properties as ranked by two different feature selection methods. The properties highlighted in red overlap with the most important 35 properties derived from the RF trained with all 550 physicochemical properties (**Fig. 4c** in the main paper). The minimal overlap of the property sets indicates that different learning algorithms extract and use (possibly correlated) information in different ways, making it harder to perform direct comparisons of methods, beyond evaluating the accuracy of predicting high responding peptides.

Nature Biotechnology: doi:10.1038/nbt.1524



Supplementary Figure 8 Boxplot of Random Forest (RF) m_{try} (referred to as *num_features* in the main text) optimization. At each m_{try} step 25 RF models using 1,000 trees were calculated to produce an error distribution. The optimal mtry value of 90 was selected based on the median value with the lowest out-of-bag (OOB) error. The out-of-bag (OOB) is the error associated with the training data and is similar to 5-fold cross validation.



Supplementary Figure 9 Receiver operating characteristic (ROC) curve for the yeast test set. The ROC plots the true positive rate (sensitivity) against the false positive rate (1 – specificity) while varying a threshold (probability of high response). The area under the curve (AUC) is a standard measure of classifier performance. An AUC of 100% would indicate a perfect classifier while an AUC of 50% (dotted line) would indicate random predictions. The AUC for the yeast test set was 83% (P = 9.4e-9) indicating the performance is better than random.