Supplementary Material (ESI) for Lab on a Chip

The BAD project: data mining, database and prediction of protein adsorption on surfaces

Elena N. Vasina, Ewa Paszek, Dan V. Nicolau, Jr., Dan V. Nicolau

Neural Networks-based assessment and prediction of the BAD

1. Assessment of how representative the BAD is

A subset of the BAD data was analysed by neural networks procedures, as implemented in Statistica. First, the 420 dataset described above was divided in three sets: training (50, 100, 150, 200, 250 and 300 data points), selection (or validation, 60 data points), and testing (60 data points). The magnitude of the training, selection and test sets has been automatically suggested by Statistica. For each of these combinations we run 10 separate runs, each run fitting several types of neuronal nets using the above sets. Each run uses sets of identical magnitude, but with different composition, prepared through a random process of selection from the BAD 420-subset. For each run we recorded the results, i.e., training, selection and test error as averages, standard deviations and minimum values, only for the networks that improved during neural network procedure. Each run produced an average of 84 improved networks. Second, we run the same calculation plan but with double the size of test sets (120 data points); consequently the training set could be only up to 240 data points. The quality of the prediction when using all the data in the BAD is presented in the Figure SI-NN1.

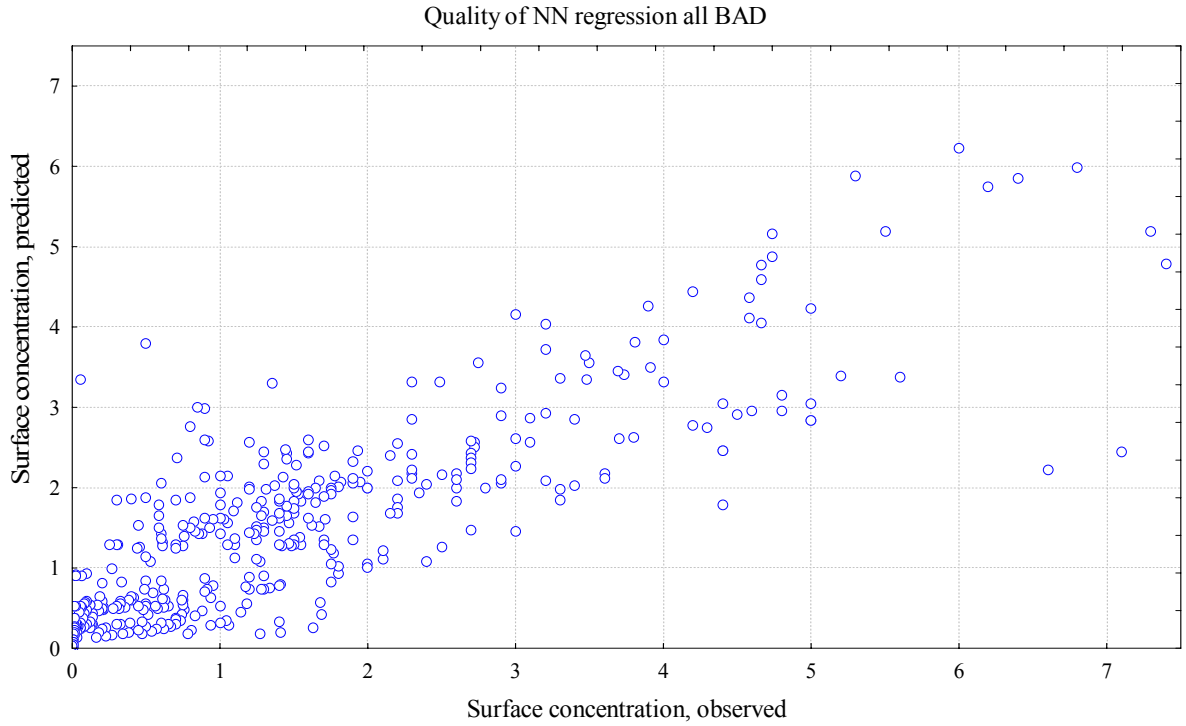

Figure SI-NN1. Predicted vs. observed for NN-based prediction all data in the BAD

Finally, we divided the 420 data set in two quasi-equal data sets (211 and 209 data points), each related to different surfaces, i.e., hydrophilic (up to 45deg) and hydrophobic. The same procedures were run as before on these two reduced data sets, but only up to 150 data points for the training set (the remainder being used for the selection and test sets). These surface-specific neural networks calculations produced an average of 50 and 80 nets per training set, for hydrophilic and hydrophobic surfaces, respectively. The quality of the prediction for hydrophobic and hydrophilic surfaces is presented in Figure SI-NN2 and SI-NN3, respectively.

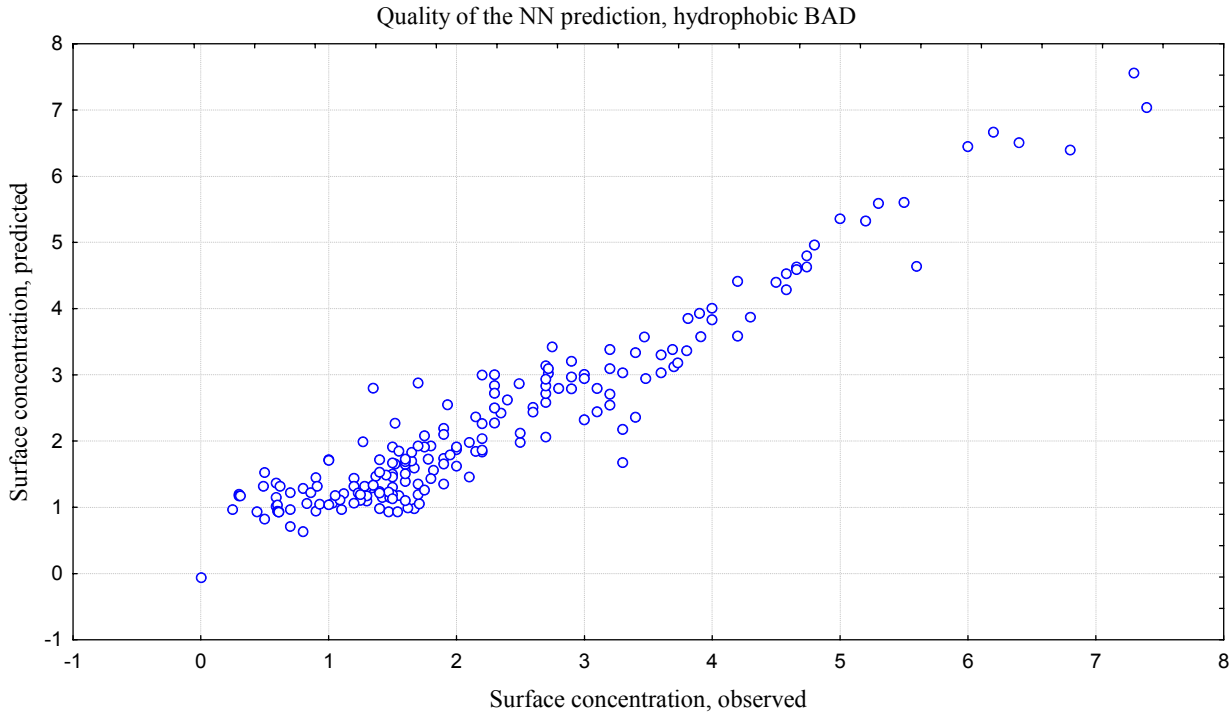

Figure SI-NN2. Predicted vs. observed for NN-based prediction for hydrophobic surfaces.

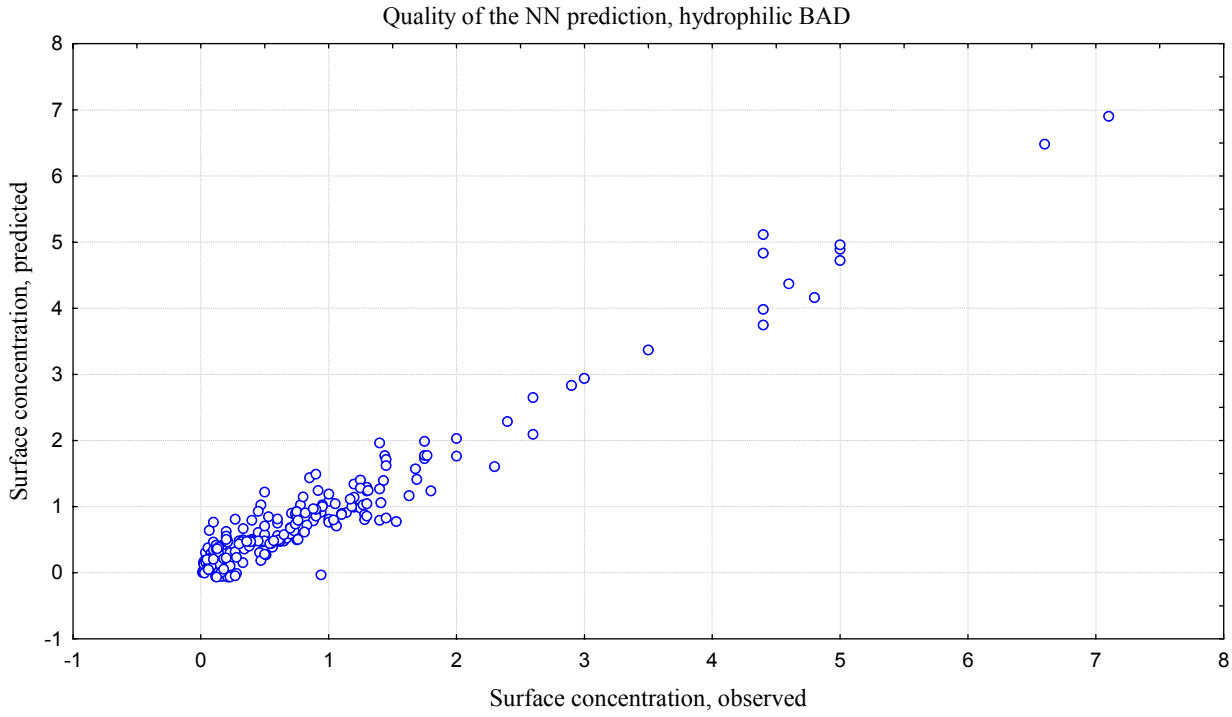

Figure SI-NN3. Predicted vs. observed for NN-based prediction for hydrophilic surfaces.

In all calculations, the protein surface concentration was the continuous output and the continuous inputs were protein concentration in solution, ionic strength, contact angle, absolute difference between the pH of buffer and the isoelectric point of the protein, protein hydrophobicity and its standard deviation. In all calculations, both the linear and logistic regression output encoding have been used. The criterion used to select the retained networks was the balance error against diversity. The types of the selected neural networks are linear, probabilistic, general regression, radial basis function and three and four layer perceptron.

The comparison of the *average* and *minimum* performance criteria, i.e. for *all* networks that improve during training, are presented in Figure SI-NN4. The minimum errors are below 10% even for small training sets, suggesting that the BAD is fully representative, even when the data is split in two for hydrophobic and hydrophilic data.

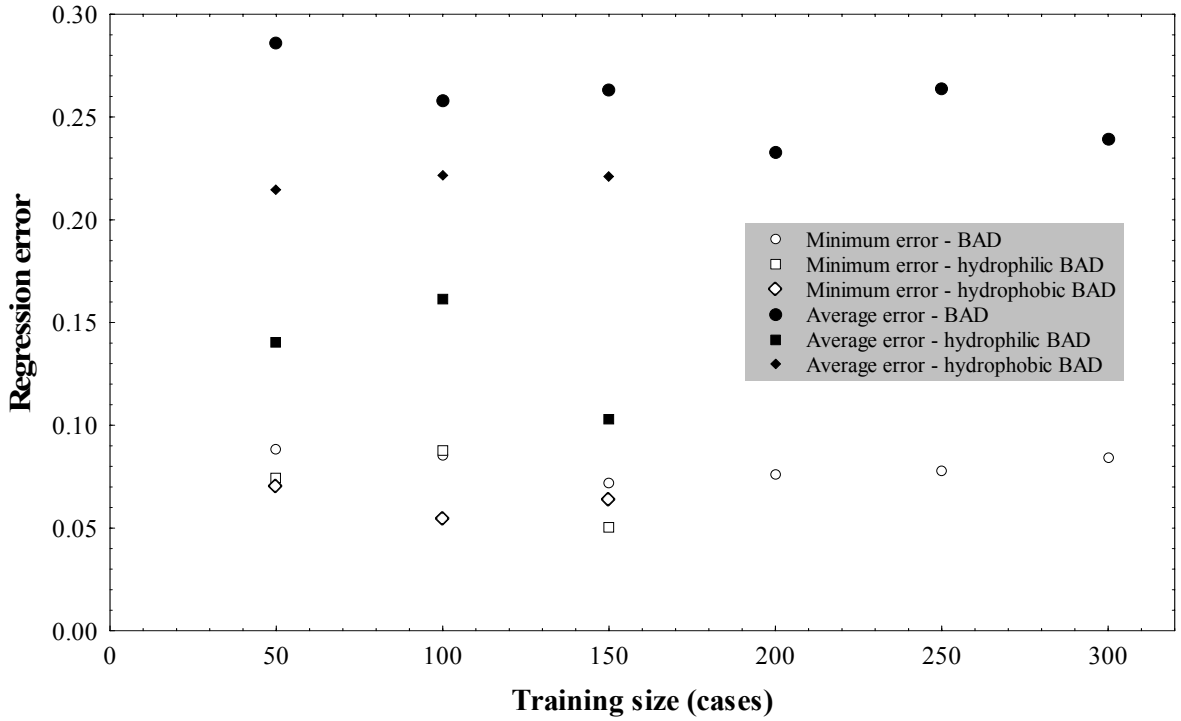

Figure SI-NN4. Comparison of the average and minimum errors for the NN trials, for all, hydrophobic and hydrophilic data in the BAD vs. training set sizes.

2. Neural networks-based prediction

Further optimisation of the neural networks allowed the identification of a set of networks that are predicting with high accuracy the amount of protein adsorbed on the surface.
Analysis of different network types and architectures using the maximum network performance and minimum error criteria has revealed that the multiple percepton neural network (MLP) give the optimal results. For the purpose of the further analysis a 3-layer MLP was chosen for fitting hydrophobic surfaces and 4-layer MLP for hydrophilic surfaces, respectively and trained using back propagation method in Statistica. The most improved network models were selected. For hydrophobic surfaces 7:10:1 architecture composed of 7 neurons in the first input layer, 10 neurons in the second hidden layer and 1 neuron in the third output layer was selected with the mean-squared error of 0.031. For the hydrophilic

surfaces 7:11:11:1 architecture was fitted with the mean squared error of 0.025. In both cases the hyperbolic function was used as an activation function in hidden layers and the logistic function in output layers, respectively.

The best neural networks, as determined by the procedures detailed above, were further analyzed in Matlab. The BAD subsets data served as an input to the networks, i.e., 191x7 matrix for hydrophobic surfaces and 201x7 matrix for hydrophilic surfaces, respectively. Figure SI-NN5 and Figure SI-NN5 illustrate the relationship between predicted and observed surface concentration [mg/m^2] for the hydrophilic and hydrophobic surfaces, respectively.

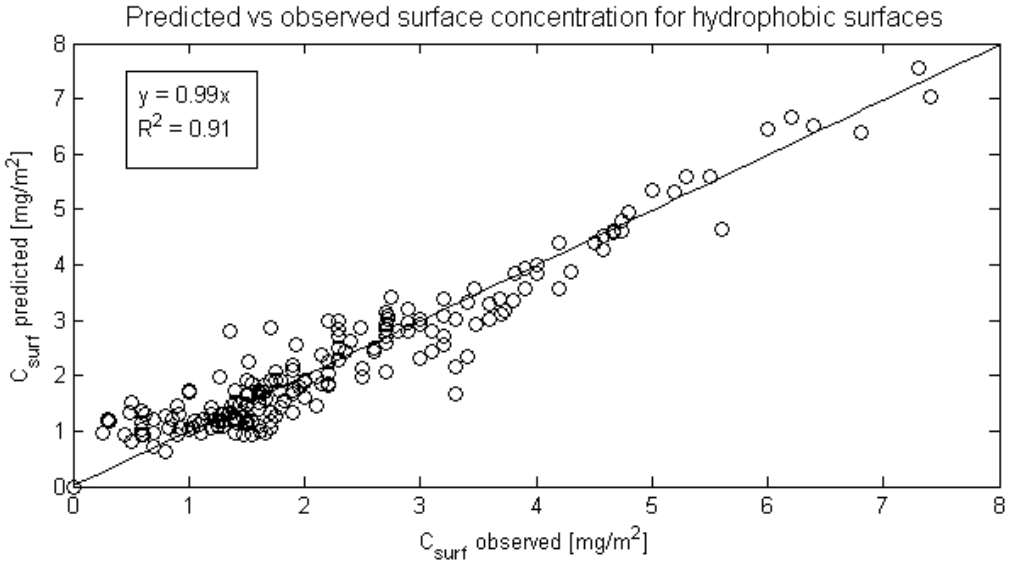

Figure SI-NN5. Predicted vs. observed for optimum NN-based prediction for hydrophobic surfaces.

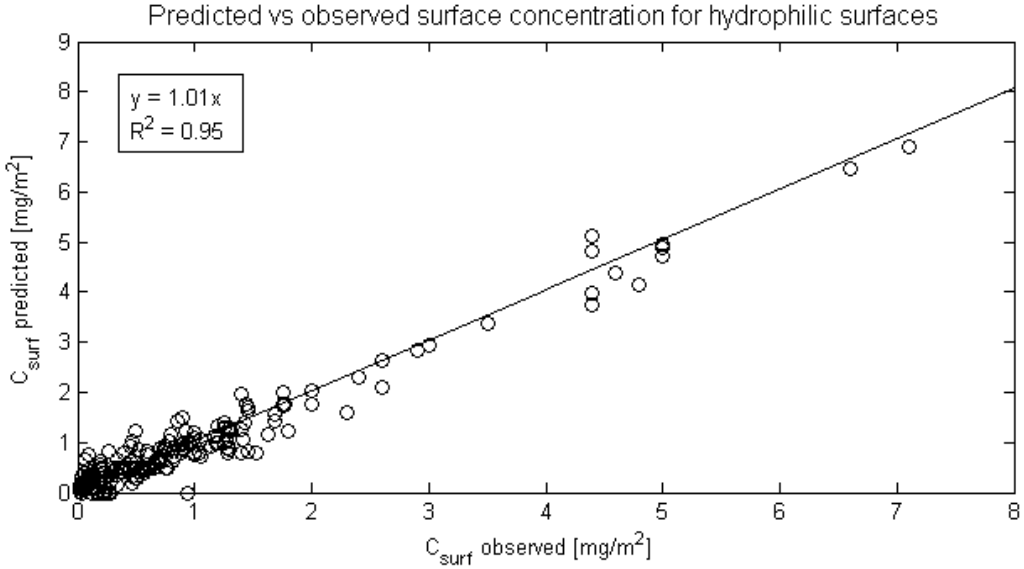

Figure SI-NN5. Predicted vs. observed for optimum NN-based prediction for hydrophilic surfaces.

The quality of NN predictions was evaluated using linear regression forced through the origin of coordinates. For the hydrophobic surfaces the correlation coefficient (R) equals to 0.95 and the regression slope (b coefficient) to 0.99. More satisfactory result was obtained for the hydrophilic surfaces: the correlation coefficient (R) equals 0.97; the regression slope equals 1.01. Detailed information concerning the regression parameters are presented in the Table SI-NN1.

Table SI-NN1. Comparison of the predictions for optimum NN for hydrophilic and hydrophobic surfaces

Parameter	**Hydrophilic**	**Hydrophobic**
Prediction error	0.031	0.025
Regression coefficient (R^2)	0.95	0.91
Correlation coefficient (R)	0.97	0.95
Estimate of the error variance	0.71	0.19
95% confidence interval	(0.98,1.03)	(0.97,1.02)
Regression slope	1.01	0.99

Neural networks are implemented in Visual Basic and embedded into BAD portal to allow the BAD users to predict online the amount of protein adsorbed to the surface. User is asked to specify protein name, solution pH, isoelectric point, water contact angle and protein concentration in solution and depending on the surface hydrophobicity one of two fitted networks is utilized to make prediction.