SCIENTIFIC REPORTS

OPEN

Quantitative design rules for protein-resistant surface coatings using machine learning

Received: 10 September 2018
Accepted: 23 November 2018
Published online: 22 January 2019

Tu C. Le[1], Matthew Penna[1,2], David A. Winkler[3,4,5,6] & Irene Yarovsky[1,2]

Preventing biological contamination (biofouling) is key to successful development of novel surface and nanoparticle-based technologies in the manufacturing industry and biomedicine. Protein adsorption is a crucial mediator of the interactions at the bio – nano -materials interface but is not well understood. Although general, empirical rules have been developed to guide the design of protein-resistant surface coatings, they are still largely *qualitative*. Herein we demonstrate that this knowledge gap can be addressed by using machine learning approaches to extract *quantitative* relationships between the material surface chemistry and the protein adsorption characteristics. We illustrate how robust linear and non-linear models can be constructed to accurately predict the percentage of protein adsorbed onto these surfaces using lysozyme or fibrinogen as prototype common contaminants. Our computational models could recapitulate the adsorption of proteins on functionalised surfaces in a test set with an r^2 of 0.82 and standard error of prediction of 13%. Using the same data set that enabled the development of the Whitesides rules, we discovered an extension to the original rules. We describe a workflow that can be applied to large, consistently obtained data sets covering a broad range of surface functional groups and protein types.

The behaviour of proteins on surfaces is of critical importance in a wide range of applications, particularly medical applications of nanomaterials[1], biomedical implants, artificial tissue scaffolds or industrial applications where surfaces are compromised when exposed to microbial or other biological contaminants[2–8]. Protein adsorption at solid and liquid interfaces is a common but very complex phenomenon that is not well understood despite over four decades of research[2,3]. This paucity of mechanistic information on protein adsorption limits the rational design of the next generation of bioinert materials.

Poly(ethylene glycol) (PEG) derivatives have long been the gold-standard for antifouling materials, however, there are still a number of issues with these material[9]. PEG can be oxidised into non-biodegradable products whose impact on the body is currently unknown. Furthermore, in some circumstances, repeat exposure to PEGylated particles through multiple injections results in significant decrease in blood circulation time, limiting the efficacy of PEG functionalised particles[10–12]. A host of surface functionalizations, a wide range of zwitterionic, hydroxyl acrylate, oxazoline, vinylpyrrolidone, and glycerol polymers, peptides, and peptoids, have been used to block protein adsorption across a range of applications[13], with varying degrees of success. Regardless of the system employed, the complex underlying mechanisms for, and influence of various surface chemistries on, protein adsorption are poorly understood.

Among the general, empirical rules that have been proposed to aid the design of protein repellent surfaces, the "Whitesides rules" are arguably the most widely used. They arose from a systematic study of the protein adsorption capacity of 48 types of self-assembled monolayers (SAMs)[14,15]. These were prepared by the reaction of an amine HNR'R with a SAM that displays interchain carboxylic anhydrides on its surface, and their structure is shown in Fig. 1. Table 1 summarizes the compositions of the 48 SAMS, identifying the diversity of structures and physicochemical properties within this set of materials.

[1]School of Engineering, RMIT University, GPO Box 2476, Melbourne, Victoria, 3001, Australia. [2]ARC Industrial Transformation Research Hub for Australian Steel Manufacturing, Wollongong, NSW, 2522, Australia. [3]Monash Institute of Pharmaceutical Sciences, Monash University, Parkville, Victoria, 3052, Australia. [4]La Trobe Institute for Molecular Science, La Trobe University, Bundoora, Victoria, 3084, Australia. [5]CSIRO Manufacturing, Clayton, Victoria, 3168, Australia. [6]School of Pharmacy, University of Nottingham, Nottingham, NG7 2RD, UK. Correspondence and requests for materials should be addressed to T.C.L. (email: Tu.Le@rmit.edu.au) or I.Y. (email: Irene.Yarovsky@rmit.edu.au)

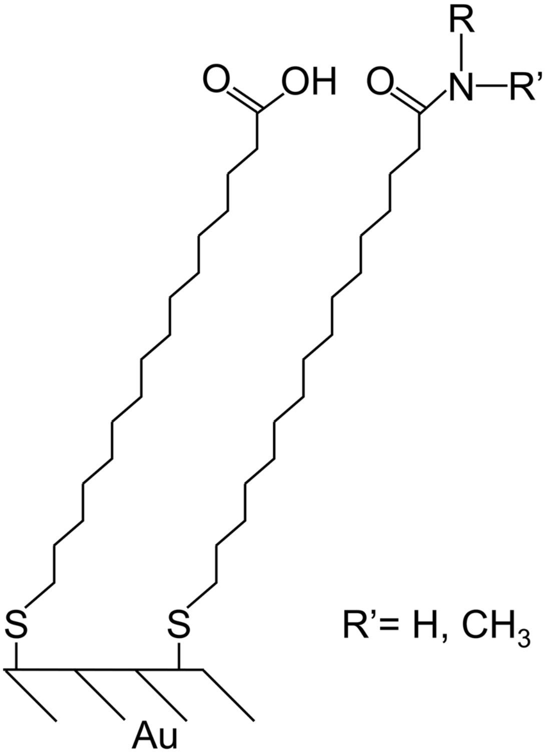


Figure 1. Chemical structure of the self-assembled monolayers (SAMs).

According to "Whitesides' rules", protein resistant surfaces should have the following characteristics:

- Polar (hydrophilic) functional groups and hydrogen bond acceptor groups.
- No hydrogen bond donor groups or net charge.

Although the rules are qualitative, they have been used to develop many types of bioinert surfaces such as oligo-/poly(ethylene glycol)s, oligo-/polyglycerols, and zwitterionic polymers[16]. The performance of alkane-SAM based coatings has now been somewhat superseded by hydrogel coatings and in many cases, coatings without the characteristics proposed by Whitesides *et al.* still have good performance. Researchers currently construct non-fouling layers underneath bioactive signalling molecules, rarely relying on alkane-SAMs but, instead, on hydrogel layers grafted to or from the surface. Unfortunately, there is a paucity of published data for hydrogels or polymer brushes that is consistently generated using the same experimental and measurement conditions and large enough to train machine learning models. There is a clear need for consistent and standardised procedures for generating experimental data on the adsorption of proteins at functionalised interfaces that can be used to generate more widely applicable machine models to aid in the understanding and design of new, efficient antifouling materials. To the best of our knowledge, the only previously reported application of machine learning to model the adsorption of proteins on material surfaces considered the adsorption of fibrinogen on polyarylate and polymethacrylate surfaces and no general design rules for these esters were reported[17–23]. Hence, this study aims to demonstrate the usefulness of statistical and machine learning techniques to identify quantitative relationships between the diverse chemistry of the material surface and the protein adsorption characteristics. These surfaces were functionalised with esters, ethers, amines, amides, sugars, nitriles and other functional groups. We use the same experimental data set from which the "Whitesides rules" were derived to illustrate that the technique can mine the data to extract established design rules *quantitatively* as well as initiate new rules.

Methods

Machine learning methods have been very successful in many areas of molecular design for generating robust, predictive models linking microscopic structure and macroscopic properties of materials[24]. They are supervised learning methods that can extract the complex structure–activity (property) relationships from reliable data sets of molecules or materials whose microscopic structures are well defined and their macroscopic properties of interest are measured. Quantitative structure–property relationship (QSPR) techniques have been applied successfully to a broad range of materials properties from physical, chemical, and biological to mechanical, electronic, and optical properties[24]. In this work, we used QSPR techniques to derive the relationships between the chemical structures and physicochemical protperties of SAMs and their protein adsorption profile.

The adsorption data, consisting of the percentage protein monolayer coverage on a mixed SAM (%ML), reported by Ostuni *et al*[15]. was used to train the models. Four functional groups, (sulfonate, phosphate, chloro and fluoro) were underrepresented in the data. Underrepresented features cannot be adequately captured by the

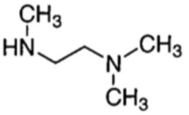
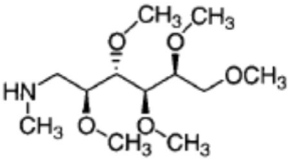
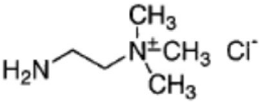
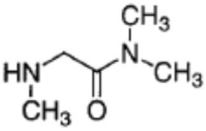
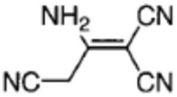
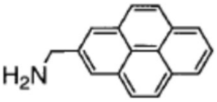
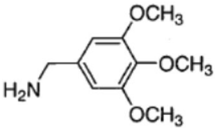
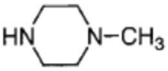
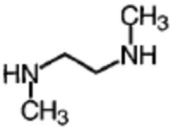
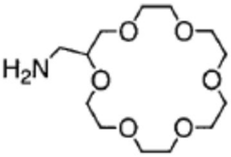
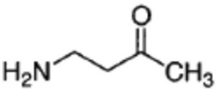
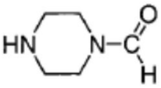
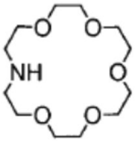
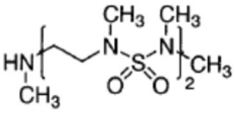
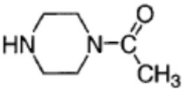
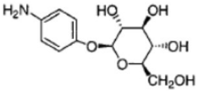
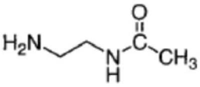
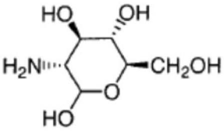
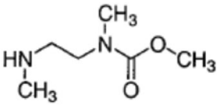
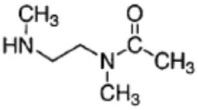
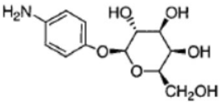
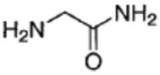
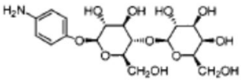
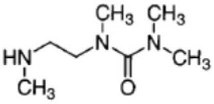
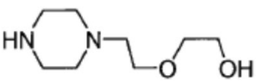
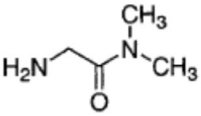
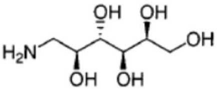
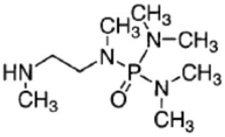
Entry	R	Entry	R	Entry	R	Entry	R
1	$H_2N(CH_2)_{10}CH_3$	13		25	$H_2N(Gly)_3N(CH_3)_2$	37	
2	$H_2NCH_2(CF_2)_6CF_3$	14		26	 $H(CH_3)N(Sar)_1N(CH_3)_2$	38	
3		15	$H_2N(CH_2CH_2O)_2CH_2CH_2NH_2$	27	$H(CH_3)N(Sar)_3N(CH_3)_2$	39	$HN(CH_2CH_2CN)_2$
4		16		28	$H(CH_3)N(Sar)_4N(CH_3)_2$	40	$HN(CH_2CN)_2$
5	$H_2NCH_2CH_2OCH_3$	17		29	$H(CH_3)N(Sar)_5N(CH_3)_2$	41	$H_2NCH_2CH_2CN$
6	$H_2NCH_2CH_2OH$	18	$HN(CH_3)_2$	30		42	
7	$HN(CH_2CH_2OCH_3)_2$	19		31		43	
8	$H_2N(CH_2CH_2O)_3CH_3$	20		32		44	$H_2NC(CH_2CH_2CH_2OH)_3$
9	$H_2N(CH_2CH_2O)_3H$	21		33		45	
10	$H_2N(CH_2CH_2O)_6CH_3$	22		34		46	$H(CH_3)NCH_2CH(OCH_3)_2$
11	$H_2N(CH_2CH_2O)_6H$	23		35		47	
12		24	 $H_2N(Gly)_1N(CH_3)_2$	36		48	

Table 1. The chemical structure of –R of the self-assembled monolayers (SAMs)[15].

models therefore SAMs containing these groups were excluded. The combined data set (176 data points) used to train the models pertained to adsorption of lysozyme and fibrinogen at 3 and 30 minutes exposure times. These prototype proteins were used because they have different properties such as size, shape, and pI. Fibrinogen is a large (340 kDa) tetrameric aggregate with a pI of 5.5. It readily adsorbs onto hydrophobic and charged surfaces. It is similar to the extracellular matrix protein fibronectin. Lysozyme is a small (MW15 kDa), ubiquitous model protein with a pI of 10.9. It is positively charged at physiological pH. Molecular descriptors (mathematically encoded properties of molecules) used in the models related to structure, partial charges, existence of particular molecular fragments or functional groups, the molecular graph, and atomic mass and were calculated using the Dragon software[25]. Indicator variables specifying the protein type (lysozyme or fibrinogen) and time scale (3 or 30 minutes) were also included as descriptors. The total size of the pool of descriptors was 67.

Data sets were divided into training (80%) and test (20%) sets using the k-means clustering algorithm. Only the training set was used to generate the models. The ability of the models to predict the protein adsorption on SAMs not included in the training set was validated using the test set. Two QSPR modeling methods were employed: sparse multiple linear regression with expectation maximization (MLREM) and non-linear Bayesian regularized artificial neural networks with Bayesian prior (BRANNGP)[26–28]. The neural networks consisted of input, hidden, and output layers. The number of nodes in the input layer was equal to the number of descriptors and the output layer had only one node corresponding to the protein adsorption value %ML. Two or three nodes in the hidden layer were found to be sufficient to build good models. It has been shown that increasing the number is unnecessary as the Bayesian regularization automatically controls the complexity of the models to optimize the test's predictive capacity[29].

The performance of the models was assessed using the coefficient of determination (r^2), the standard error of estimation (SEE), and the standard error of prediction (SEP). r^2 is the square of the correlation coefficient between the predicted and measured %ML. SEE and SEP are the root-mean-square values, adjusted for degrees of freedom, of the difference between the predicted and measured %ML for the training and test sets respectively. SEE and SEP are more robust estimates of the predictive ability of models because, unlike r^2, they do not depend on the number of data points in the training set or the number of descriptors in the model[30]. Predictive, robust models have r^2 values close to 1.0 and SEE and SEP values that are similar and close to the experimental error.

Results

Whitesides Rules.

We were interested in the degree to which the elements of "Whitesides' rules" can make *quantitative* predictions of protein adsorption behaviour for the adsorption data set. Because the charged groups were underrepresented and excluded from the modelling data set, only three factors from 'Whitesides rules' remain as the model inputs: hydrophilicity; number of hydrogen bond acceptors; number of hydrogen bond donors. The hydrophilicity was represented by the hydrophilic factor (Hy)[31] calculated using the Dragon software. Indicator variables for protein type and time at which measurements were made were also required, leading to a model containing 5 input parameters.

Figure 2(A) shows the scaled (normalized) MLR coefficients generated for the model. These coefficients offer deeper understanding of the effect of each property on the degree of protein adsorption (%ML). A positive MLR coefficient indicates that the property promotes the adsorption of a protein onto the surface while a negative coefficient indicates that the property inhibits the adsorption. As Fig. 2(A) shows, the hydrophilicity (Hy) and the presence of hydrogen bond accepting functional groups (nHAcc) were associated with low protein adsorption. Larger number of hydrogen bond donor groups (nHDon) was associated with increased adsorption. These conclusions are in good agreement with the empirically derived rules. The three Whitesides' rule properties have a much larger impact on protein adsorption than the protein type and or the time scale. Given that only two proteins were studied and the time points were reasonably short, this is not too surprising. However, the quantitative prediction ability of the model is poor, with a test set r^2 value of 0.35 and standard error of prediction (SEP) of 24%. Replacement of the hydrophilic factor (Hy) with ALogP, the log octanol-water partition coefficient calculated using Ghose-Crippen-Viswanadhan method[32–34], produces MLR coefficients again consistent with the Whiteside rules (Fig. 2(B)). The predictive power of this model improved slightly with the tests set r^2 value rising to 0.54 and SEP dropping inconsequentially to 23%. However, both protein type and time made larger contributions to the model in this case, and the contribution of hydrogen bond acceptors was reduced. There is clearly a correlation between the donor/acceptor properties and the hydrophilicity/hydrophobicity properties, and logP values are influenced by the number of hydrogen bond donors.

A possible explanation for the relatively poor predictive power of these models is that the data set did not contain enough charged moieties to include them in the model, and the input parameters are all related to the ability of the functional ligands to interact with water and thus only reflect the hydration theory of protein repulsion[35]. To account for the steric repulsion theory of protein adsorption[36–38] or a combination of both[2], inclusion of parameters that reflect the dynamic character of ligands might improve the predictive power and design utility of the model.

Comprehensive descriptor set.

We computed a more comprehensive descriptor set using the Dragon package and employed both linear (MLREM) and non-linear (BRANNGP and BRANNLP) methods to make quantitative predictions of the protein adsorption behaviour. The initial pool of 67 Dragon descriptors which includes those capturing Whitesides original rules were pruned using MLREM sparse feature selection method[27,28] to identify the most important descriptors that affect the protein adsorption. These approaches have been shown to be useful in carrying out sparse descriptor selection[39–42]. By tuning the sparsity of the MLREM progressively, the least informative descriptors were pruned out and the most relevant descriptors retained.

Figure 3 shows that by increasing the sparsity of the models and pruning out irrelevant descriptors, the predictive power of the models is enhanced and lower test set SEP values were obtained. When too many descriptors

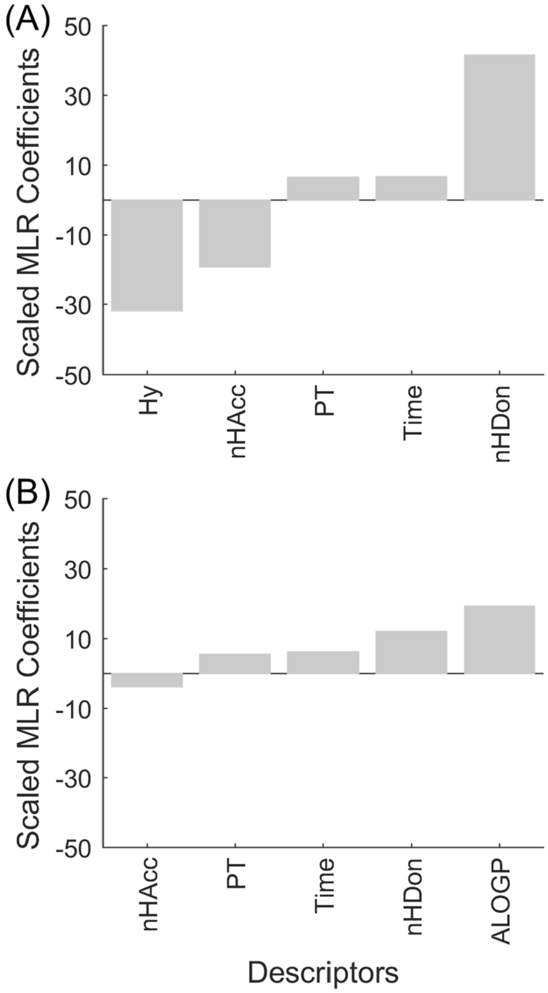


Figure 2. Scaled MLR coefficients for Whitesides rule descriptors to prevent protein adsorption. (**A**) Model using Hy parameter for hydrophilicity. (**B**) Model using AlogP for hydrophobicity.

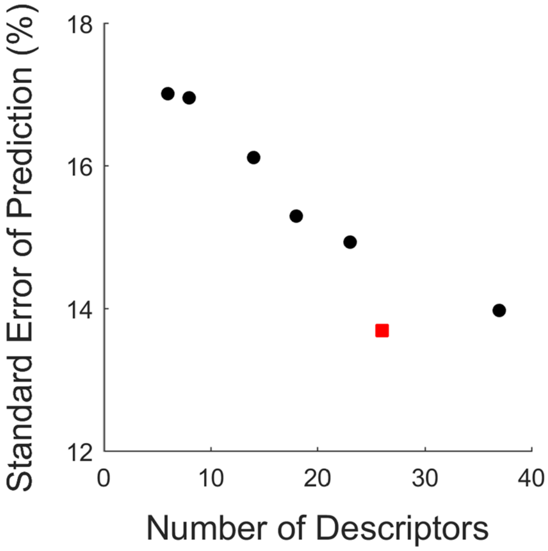


Figure 3. The dependence of the standard error of prediction (SEP) on the number of descriptors for models constructed using the MLREM approach to prune out irrelevant descriptors. The red data point indicates the best models with optimal sparsity.

Modelling technique	N_{eff}	Training set		Test set	
		r^2	SEE [%]	r^2	SEP [%]
MLREM	27	0.81	13	0.78	14
BRANNGP*	28	0.82	12	0.76	14
BRANNGP#	35	0.84	10	0.79	14

Table 2. Statistics of the optimal linear and non-linear models of protein adsorption (fibrinogen and lysozyme) on different surfaces at 3 and 30 minutes. (N_{eff} is the number of effective weights (adjustable parameters) in the model). *BRANNGP model built using the entire pool of 67 descriptors. #BRANNGP model built using 25 descriptors selected by MLREM.

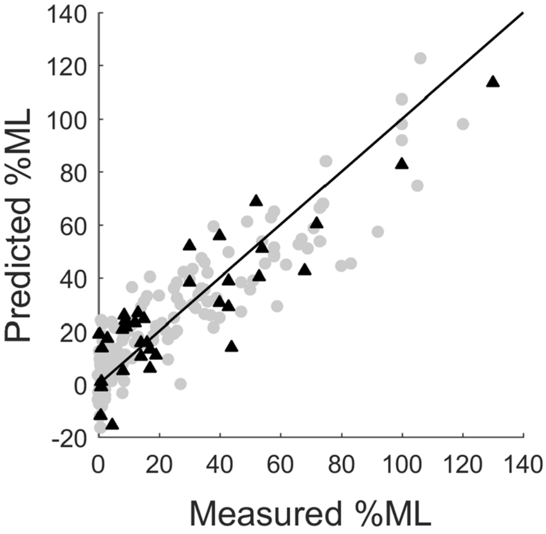


Figure 4. Prediction of the best MLREM model of percentage protein monolayer coverage on SAMs (%ML). Training set (grey circles) and test set (black triangles).

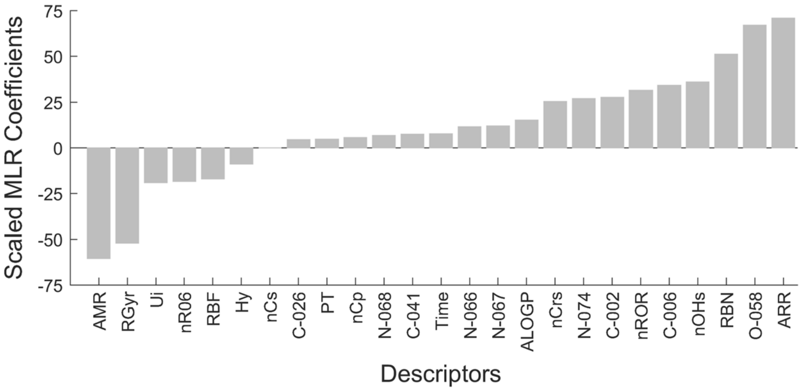


Figure 5. Scaled MLR coefficients of the most relevant descriptors selected from the pool 67 descriptors.

are removed, the performance of the models decreases and the SEP increases. The best models were those with the lowest SEP values and the lowest complexity (least number of descriptors). The performance of these models is summarized in Table 2 and Fig. 4. We also attempted to use different combinations of descriptors and the performance of the models using these descriptor sets is presented in Table S1.

As can be seen in Table 2 and Fig. 3, the best models contain 25–28 descriptors or effective parameters and further pruning of descriptors result in a significant drop in predictive performance of the models. The linear and nonlinear models had equal ability to predict the %ML for SAMs in the test sets. This means that the relationship between the adsorption of protein on functionalized surfaces (%ML) and the descriptors is complex but largely linear. An examination of relevant structural descriptors selected by the models can provide some insight into the most significant factors that affect the adsorption process. The contribution of the most important descriptors is illustrated in Fig. 5 and the details of descriptors are listed in Table 3. When different sets of descriptors were used to construct the models, the contribution of these descriptors is presented in Fig. S1.

Descriptor	Definition	Type	Contribution
Negative			
AMR	Ghose-Crippen molar refractivity	Continuous	−60
RGyr	radius of gyration (mass weighted)	Continuous	−52
Ui	unsaturation index	Continuous	−19
nR06	number of 6-membered rings	Integer	−19
RBF	rotatable bond fraction	Integer	−17
Hy	hydrophilic factor	Continuous	−9
Positive			
C-026	number of R–CX–R	Integer	5
ProteinType	protein type indicator	Integer	5
nCp	number of terminal primary C(sp3)	Integer	6
N-068	number of Al_3-N fragments	Integer	7
C-041	number of X-C(=X)-X fragments	Integer	8
Time	time scale indicator	Integer	8
N-066	number of Al-NH2 fragments	Integer	11
N-067	number of Al_2-NH fragments	Integer	12
ALOGP	Ghose-Crippen octanol-water partition coeff. (logP)	Continuous	15
nCrs	number of ring secondary C(sp3)	Integer	25
N-074	number of R≡N / R=N- fragments	Integer	27
C-002	number of CH_2R_2 fragments	Integer	28
nROR	number of ethers (aliphatic)	Integer	32
C-006	number of CH_2RX fragments	Integer	34
nOHs	number of secondary alcohols	Integer	36
RBN	number of rotatable bonds	Integer	51
O-058	number of O=	Integer	67
ARR	aromatic ratio	Integer	71

Table 3. The most relevant descriptors selected by MLREM and their contributions to the model predicting %ML. The descriptors are listed in the order of least negative to most positive.

Positive contributors to adsorption. The examination of the positive descriptors and their associated scaled MLR values (Fig. 5) shows that 5 of the top 10 predictors of protein adsorption apply only to a specific group of ligands. These descriptors can therefore be classified as primary exclusion criteria and these types of chemistries should not be considered for inclusion in antifouling ligands. For example, both aromatic groups and nitriles show high protein adsorption and this is reflected in the presence of the descriptor ARR and N-074 as positive contributors to protein adsorption. The hydrogen bond donating secondary alcohols (nOHs) and NHR_2 groups (N-066 and N-067) can be classified as promoting adsorption. The NHR_2 descriptor must be taken with some caution as all ligands contain either one N-066 and N-067 based on the covalent attachment point.

Negative contributors to adsorption. Four of the six descriptors that make negative contributors to the protein adsorption model are continuous descriptors that adopt non-zero values for all ligands. Molar refractivity (AMR), a measure of ligand size and polarizability[43], has the largest negative impact on protein adsorption. The trend of decreasing adsorption with increasing substituent size is evident from observation of, for example, linear EG derivatives or substituted amide groups.

The final two continuous descriptors relate to the conformational flexibility of the ligands: radius of gyration (RGyr) and rotatable bond fraction (RBF). Both have a relatively strong correlation (>0.84) to the number of rotatable bonds (RBN) which was a positive contributor to protein adsorption. Given the respective scaled MLR coefficients the combination of these three parameters suggests that increased ligand conformational freedom deters protein adsorption.

The remaining two descriptors relate to specific ligand chemistry. Ten ligands contain aromatic rings (non-zero nR06 descriptors, number 6 membered rings), of these, 6 have applicable primary exclusion criteria, either aromatic content or secondary hydrogen bonds. Similarly, 23 ligands have non-zero unsaturation indices (Ui), 10 of which have defined primary exclusion criteria. 11 of the remaining 13 ligands are amino acids that exhibit decreasing adsorption with increasing number of repeating units, consistent with Ui being a negative contributor to adsorption.

Reconciling QSPR predictions with Whitesides Rules and existing theories. Two theories exist regarding the underlying mechanism of protein resistance by SAM protected surfaces: steric repulsion and hydration theory. Steric repulsion rationalises protein adsorption resistance based on the conformational freedom of surface grafted ligands in good solvent conditions which present a high entropic penalty working against protein adsorption[37,38,44]. Hydration theory[35] was developed to account for high density SAMs, where ligands would have restricted dynamics, yet showed resistance to protein adsorption. It has been reported that the capacity for

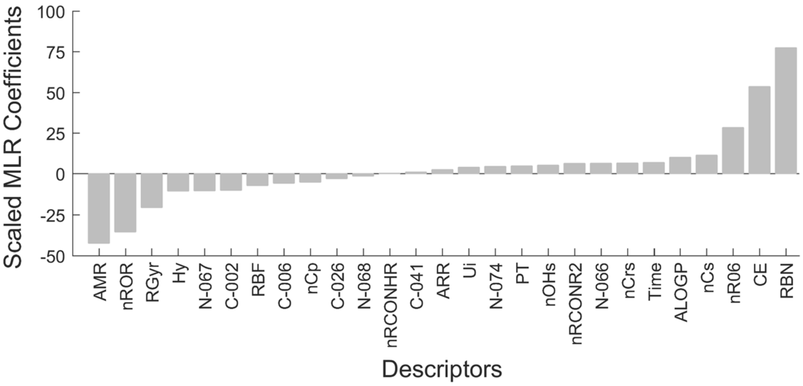


Figure 6. Scaled MLR coefficients of descriptors in the updated model.

functional ligands to coordinate water within the SAM layers and at the SAM/water interface is critical to limiting protein adsorption[11]. Hybrids of the two have been reported for PEG systems[45] and recent MD simulations support the synergistic influence of these two factors in preventing protein adsorption[2].

Ostensibly Whitesides Rules fall within the hydration theory of protein adsorption resistance as they only account for ligand chemistries related to the interaction of water with the functional ligand. Interestingly, there is a significant number of reported ligands that partially contradict these general rules and still possess protein adsorption resistance[16]. This can be explained by the fact that in the original derivation the rules were not accounting for the steric repulsion theory and the entropic contribution of the ligands to protein resistance, possibly amongst other contributing factors. Nevertheless, the rules, together with the data set from which they are derived, provide an excellent framework for refinement using the QSPR models via chemical identification along with systemic understanding to create improved models with greater predictive power.

When a broader range of properties captured by descriptors were added to the models, the quality of the prediction of protein adsorption improved markedly (test set r^2 value of 0.78 and SEP of 14% compared to 0.35 and 23% and respectively). Lipophilic properties captured by the descriptors ALOGP, nCrs, C-002, C-006, ARR were strongly associated with high protein adsorption, consistent with widely reported findings[3]. Neither the total number of hydrogen bond donors (nHDon) nor acceptors (nHAcc) in general were identified as significant for predicting the extent of protein adsorption. The former was captured in the model using the comprehensive descriptor set by the number of secondary alcohols and (to a lesser extent) primary and secondary amines (N-066, N-067, nOHs) which did promote adsorption in line with the Whitesides rules. The models also predicted that the number of oxygens with double bond (O-058) and the total number of aliphatic ether (nROR) moieties (both hydrogen bond acceptors) contributed positively to adsorption. This is inconsistent with the Whitesides' rules and, more surprisingly, the well-established role of polyethylene glycol (PEG) as the gold standard in preventing non-specific protein adsorption. To improve the models, we added an indicator variable to differentiate between the crown ether type structures which cause high protein adsorption rate and linear ethers (PEGs) leading to the low protein adsorption. We also replaced O-058 in the model with more specific descriptors for amides (nRCONH2, nRCONR2, and nRCONHR), and ketone (nRCOR) which were in the original pool of descriptors but got pruned out during the feature selection step. Although the newly obtained model only has slightly better performance with the test set r^2 value of 0.82 and standard error of prediction of 13%, it is able to differentiate between the high and low adsorption polyethylene glycols as can be seen in the highly negative MLR coefficient for nROR and highly positive coefficient for the crown ether indicator in Fig. 6. The scaled MLR coefficients of descriptors in the obtained model using crown ether indicator are reported in Table S2 of the Supplementary Information. The workflow and the corresponding evolvement of the model performance are summarized in Fig. 7.

Modifications to the QSPR model are critical to aid in interpretation of the model results, particularly when the model identifies criteria with positive contributions from hydrogen bond acceptors that directly contradict both well-established theory and empirical results. The important role of molar refractivity, radius of gyration and rotatable bond fraction in promoting low attachment was maintained with the addition of the flag for crown ether (CE), see Figs 5 and 6. These parameters are associated with ligand dynamics, size and polarizability, factors that were not captured by Whitesides original rules. By adding the first two descriptors (AMR and RGyr) to the model using the original descriptors from Whitesides rules (Hy, nHDon, nHAcc), the test set r^2 value of the model increases significantly from 0.35 to 0.53 and the SEP dropped from 23% to 20%. Moreover, since an increased predictive power was previously observed when ALogP replaced Hy in the model, we attempted to build a model using the set of AMR, RGyr, ALogP, nHDon and nHAcc. As expected, when Hy is replaced by ALogP, the performance of the model improved further, with the test set r^2 value of 0.56 and SEP of 17%.

Using the same strategy, we constructed models using different combinations of descriptors and provided the details in Table S1 and Fig. S2 of the Supporting Information. The observed importance of the dynamic factors (AMR and RGyr) suggests a synergistic effect between hydration and steric mechanisms in the ability of functional SAM to resist protein adsorption. We, therefore, propose that Whitesides' original rules be extended as follows. For self-assembled functional ligands to resist protein adsorption they should comprise:

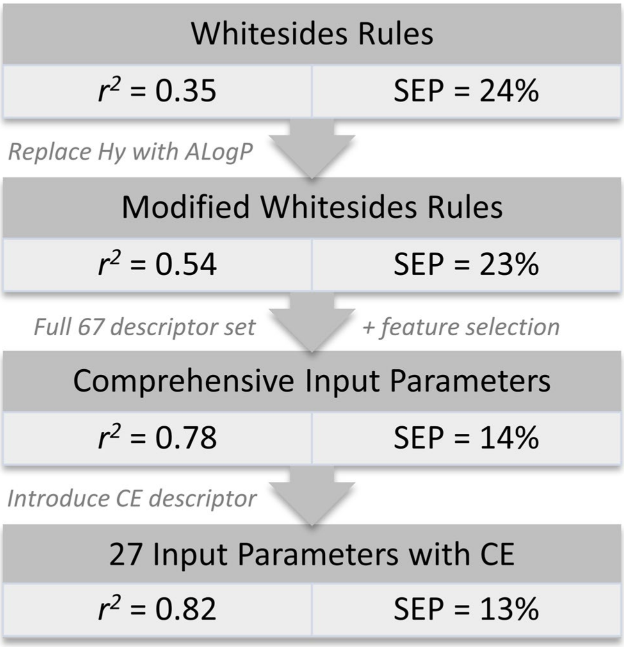


Figure 7. The workflow and the corresponding evolvement of the model performance.

- polar (hydrophilic) functional groups and hydrogen bond acceptor groups.
- no hydrogen bond donor groups or net charge.
- relatively *large, conformationally mobile and polarizable functional groups*.

Obviously, the additional properties (size, flexibility and polarizability) are highly related to the surface coating density which has been shown to play an important role in the antifouling ability of materials[10,46–50]. In the original work, the grafting density was not reported and therefore this information was not included in our models. The inclusion of grafting density or defects in the layers as inputs might improve the predictability of the models.

Limitations and Improvements. An important limitation is the under-representation of charged functional groups in the model training set. Clearly, a larger and more diverse data set will improve the predictions of the models and strengthen the validity of the empirical design rules the Whitesides group developed and we have extended. It must be noted that the predictive power of the model is within experimental error despite a number of limitations in the original data set used for this initial proof of concept utilising only calculable data for each molecule. However, the model can be further improved if the limitations are addressed. For example, while in the current model no consideration is given to the mode of the ligand attachment to the surface, it can be assumed that this region will have lesser impact on adsorption behaviour and can be accounted for in future models. However, the constraints associated with ligand attachment can have an influence on the overall dynamics of the ligand, reducing the RGyr, and, therefore, inclusion of relevant descriptors in future predictions should be considered. There are also parameters external to the ligand chemistry which will influence protein adsorption behaviour. For example, here, all ligands were treated equally with regards to grafting density (surface coverage) which was unreported in the original work, while it is known that the grafting density can influence protein adsorption behaviour[10]. Furthermore, in the original work, the authors presented data relating protein adsorption behaviour to advancing water contact angle in cyclooctane and found no correlation. While water contact angle is not a good predictor of protein adsorption in isolation it will likely be very useful when considered as one of numerous factors facilitating or preventing adsorption. Inclusion of these experimentally measurable (non-computed) properties will likely increase the predictive power of the model as it will more accurately reflect the physical system. Lastly, it has been shown that the interfacial bound water and its distinct properties influence the protein adsorption profile[51]. Hence the addition of parameters characterizing the hydration layer structure of the materials as input descriptors may improve the model performance.

Only two proteins were presented in the original data set available and it would clearly be useful to measure the attachment of a wider range of proteins with more diverse properties (size, lipophilicity, shape etc.) To this end, in our QSPR models, the protein type flag was treated as binary. In every model protein type was found to be a positive predictor of protein adsorption indicating that fibrinogen (flagged as 1) had a higher adsorption propensity than lysozyme (flagged as 0). While the use of a binary flag was sufficient for the proof of concept presented here, more comprehensive parameters to describe the adsorbing proteins should be considered in future QSPR models. A method for encoding the nature of more diverse types of proteins will most likely improve the models and the rules. Also, proteins are not passive in the adsorption process. A partial list of properties that might be considered which have previously been reported to play a role in protein adsorption are size, surface composition[52], conformational flexibility (conformational 'hardness' or 'softness')[53], surface activity or organisation capacity (i.e. clusterin[54,55]; hydrophobin[56]; and adhesins[57]).

A final challenging aspect of understanding the various influences at work remains the lack of consistency in available experimental data. A wide range of protein properties, including the amount of adsorbed protein, have been reported under different conditions. The efficacy of the models presented, based on an older and somewhat limited data set, suggests that with a standardised procedure for generating experimental data on the adsorption of proteins at functionalised interfaces QSPR models could greatly aid in the understanding and design in this space.

Conclusions

Understanding the effect of surface chemistry on protein adsorption is critical for the design of novel bioinert materials. We have shown how computational modelling using machine learning algorithms can generate a quantitative relationship between surface chemistry and protein adsorption. The models elucidate design concepts through the model weights of over 20 physical/nonphysical parameters. Such concepts can be useful for designing protein resistant as well as protein attracting surfaces. The models also highlighted the challenge of balancing related properties which deter or promote protein adsorption and supported the notion of synergy between hydration and steric effects in preventing adsorption. The model is capable of reliably predicting the degree of protein adsorbed on SAMs (within the applicability domain of the models) for new surface chemistries. Therefore, the machine learning based predictions are demonstrated to be useful to identify surfaces with the best performance for synthesis and fabrication. However, this requires more robust data sets with good quality molecular level data characterising the interface to which proteins are exposed under practically relevant conditions to allow QSPR models to become more widely applied in the diverse range of technologies reliant on mediation of protein adsorption.

Data Availability

The raw/processed data required to reproduce these findings are available to download from https://drive.google.com/drive/u/0/folders/1o4noYh7dXnYg113kaJdLOjseuTTMlNQf.

References

1. Wang, B. *et al*. Thermostability and Reversibility of Silver Nanoparticle-Protein Binding. *Phys Chem Chem Phys* **17**, 1728–1739, https://doi.org/10.1039/c4cp04996a (2015).
2. Penna, M., Ley, K., Maclaughlin, S. & Yarovsky, I. Surface Heterogeneity: a Friend or Foe of Protein Adsorption - Insights from Theoretical Simulations. *Faraday Discuss* **191**, 435–464, https://doi.org/10.1039/c6fd00050a (2016).
3. Rabe, M., Verdes, D. & Seeger, S. Understanding Protein Adsorption Phenomena at Solid Surfaces. *Adv Colloid Interfac* **162**, 87–106, https://doi.org/10.1016/j.cis.2010.12.007 (2011).
4. Banerjee, I., Pangule, R. C. & Kane, R. S. Antifouling Coatings: Recent Developments in the Design of Surfaces That Prevent Fouling by Proteins, Bacteria, and Marine Organisms. *Adv Mater* **23**, 690–718, https://doi.org/10.1002/adma.201001215 (2011).
5. Callow, J. A. & Callow, M. E. Trends in The Development of Environmentally Friendly Fouling-Resistant Marine Coatings. *Nat Commun* **2**, ARTN 24410, https://doi.org/10.1038/ncomms1251 (2011).
6. Lynch, I., Salvati, A. & Dawson, K. A. Protein-Nanoparticle Interactions What Does the Cell See? *Nat Nanotechnol* **4**, 546–547, https://doi.org/10.1038/nnano.2009.248 (2009).
7. Shemetov, A. A., Nabiev, I. & Sukhanova, A. Molecular Interaction of Proteins and Peptides with Nanoparticles. *Acs Nano* **6**, 4585–4602, https://doi.org/10.1021/nn300415x (2012).
8. Thevenot, P., Hu, W. J. & Tang, L. P. Surface Chemistry Influences Implant Biocompatibility. *Curr Top Med Chem* **8**, 270–280 (2008).
9. Lee, J. H., Lee, H. B. & Andrade, J. D. Blood Compatibility of Polyethylene Oxide Surfaces. *Prog Polym Sci* **20**, 1043–1079, https://doi.org/10.1016/0079-6700(95)00011-4 (1995).
10. Unsworth, L. D., Sheardown, H. & Brash, J. L. Protein Resistance of Surfaces Prepared by Sorption of End-Thiolated poly(ethylene glycol) to Gold: Effect of Surface Chain Density. *Langmuir* **21**, 1036–1041, https://doi.org/10.1021/la047672d (2005).
11. Herrwerth, S., Eck, W., Reinhardt, S. & Grunze, M. Factors that Determine the Protein Resistance of Oligoether Self-Assembled Monolayers - Internal Hydrophilicity, Terminal Hydrophilicity, and Lateral Packing Density. *J Am Chem Soc* **125**, 9359–9366, https://doi.org/10.1021/ja034820y (2003).
12. Ishida, T. & Kiwada, H. Accelerated Blood Clearance (ABC) Phenomenon Upon Repeated Injection of PEGylated Liposomes. *Int J Pharm* **354**, 56–62, https://doi.org/10.1016/j.ijpharm.2007.11.005 (2008).
13. Lowe, S., O'Brien-Simpson, N. M. & Connal, L. A. Antibiofouling Polymer Interfaces: Poly(Ethylene Glycol) and Other Promising Candidates. *Polym Chem-Uk* **6**, 198–212, https://doi.org/10.1039/c4py01356e (2015).
14. Chapman, R. G. *et al*. Surveying for Surfaces that Resist the Adsorption of Proteins. *J Am Chem Soc* **122**, 8303–8304, https://doi.org/10.1021/ja000774f (2000).
15. Ostuni, E., Chapman, R. G., Holmlin, R. E., Takayama, S. & Whitesides, G. M. A Survey of Structure- Property Relationships of Surfaces that Resist the Adsorption of Protein. *Langmuir* **17**, 5605–5620, https://doi.org/10.1021/la010384m (2001).
16. Wei, Q. *et al*. Protein Interactions with Polymer Coatings andBiomaterials. *Angew Chem Int Edit* **53**, 8004–8031, https://doi.org/10.1002/anie.201400546 (2014).
17. Smith, J. R. *et al*. Using Surrogate Modeling in the Prediction of Fibrinogen Adsorption Onto Polymer Surfaces. *J Chem Inf Comp Sci* **44**, 1088–1097, https://doi.org/10.1021/ci0499774 (2004).
18. Weber, N., Bolikal, D., Bourke, S. L. & Kohn, J. Small Changes in the Polymer Structure Influence the Adsorption Behavior of Fibrinogen on Polymer Surfaces: Validation of a New Rapid Screening Technique. *J Biomed Mater Res A* **68a**, 496–503, https://doi.org/10.1002/jbm.a.20086 (2004).
19. Smith, J. R., Kholodovych, V., Knight, D., Kohn, J. & Welsh, W. J. Predicting Fibrinogen Adsorption to Polymeric Surfaces in Silico: a Combined Method Approach. *Polymer* **46**, 4296–4306, https://doi.org/10.1016/j.polymer.2005.03.012 (2005).
20. Smith, J. R., Kholodovych, V., Knight, D., Welsh, W. J. & Kohn, J. QSAR models for the analysis of bioresponse data from combinatorial libraries of biomaterials. *QSAR Comb Sci* **24**, 99–113, https://doi.org/10.1002/qsar.200420062 (2005).
21. Gubskaya, A. V., Kholodovych, V., Knight, D., Kohn, J. & Welsh, W. J. Prediction of Fibrinogen Adsorption for Biodegradable Polymers: Integration of Molecular Dynamics and Surrogate Modeling. *Polymer* **48**, 5788–5801, https://doi.org/10.1016/j.polymer.2007.07.007 (2007).

22. Kholodovych, V. *et al*. Prediction of Biological Response for Large Combinatorial Libraries of Biodegradable Polymers: Polymethacrylates as a Test Case. *Polymer* **49**, 2435–2439, https://doi.org/10.1016/j.polymer.2008.03.032 (2008).
23. Costache, A. D., Ghosh, J., Knight, D. D. & Kohn, J. Computational Methods for the Development of Polymeric Biomaterials. *Adv Eng Mater* **12**, B3–B17, https://doi.org/10.1002/adem.200980020 (2010).
24. Le, T., Epa, V. C., Burden, F. R. & Winkler, D. A. Quantitative Structure-Property Relationship Modeling of Diverse Materials Properties. *Chem Rev* **112**, 2889–2919, https://doi.org/10.1021/cr200066h (2012).
25. Mauri, A., Consonni, V., Pavan, M. & Todeschini, R. Dragon Software: An Easy Approach to Molecular Descriptor Calculations. *MATCH-Commun Math Comput Chem* **56**, 237–248 (2006).
26. Burden, F. R. & Winkler, D. A. Robust QSAR Models Using Bayesian Regularized Neural Networks. *J Med Chem* **42**, 3183–3187, https://doi.org/10.1021/jm980697n (1999).
27. Burden, F. R. & Winkler, D. A. An Optimal Self-Pruning Neural Network and Nonlinear Descriptor Selection in QSAR. *QSAR Comb Sci* **28**, 1092–1097, https://doi.org/10.1002/qsar.200810202 (2009).
28. Burden, F. R. & Winkler, D. A. Optimal Sparse Descriptor Selection for QSAR Using Bayesian Methods. *QSAR Comb Sci* **28**, 645–653, https://doi.org/10.1002/qsar.200810173 (2009).
29. Polley, M. J., Winkler, D. A. & Burden, F. R. Broad-Based Quantitative Structure-Activity Relationship Modeling of Potency and Selectivity of Farnesyltransferase Inhibitors Using a Bayesian Regularized Neural Network. *J Med Chem* **47**, 6230–6238, https://doi.org/10.1021/jm049621j (2004).
30. Alexander, D. L. J., Tropsha, A. & Winkler, D. A. Beware of R-2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J Chem Inf Model* **55**, 1316–1322, https://doi.org/10.1021/acs.jcim.5b00206 (2015).
31. Todeschini, R., Vighi, M., Finizio, A. & Gramatica, P. 3D-Modelling and Prediction by WHIM Descriptors. Part 8. Toxicity and Physico-Chemical Properties of Environmental Priority Chemicals by 2D-TI and 3D-WHIM Descriptors. *SAR QSAR Environ Res* **7**, 173–193, https://doi.org/10.1080/10629369708039130 (1997).
32. Ghose, A. K. & Crippen, G. M. Atomic Physicochemical Parameters for 3-Dimensional Structure-Directed Quantitative Structure-Activity-Relationships .1. Partition-Coefficients as a Measure of Hydrophobicity. *J Comput Chem* **7**, 565–577, https://doi.org/10.1002/jcc.540070419 (1986).
33. Ghose, A. K., Viswanadhan, V. N. & Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J Phys Chem A* **102**, 3762–3772, https://doi.org/10.1021/jp980230o (1998).
34. Viswanadhan, V. N., Reddy, M. R., Bacquet, R. J. & Erion, M. D. Assessment of Methods Used for Predicting Lipophilicity - Application to Nucleosides and Nucleoside Bases. *J Comput Chem* **14**, 1019–1026, https://doi.org/10.1002/jcc.540140903 (1993).
35. Wang, R. L. C., Kreuzer, H. J. & Grunze, M. Molecular Conformation and Solvation of Oligo(Ethylene Glycol)-Terminated Self-Assembled Monolayers and Their Resistance to Protein Adsorption. *J Phys Chem B* **101**, 9767–9773, https://doi.org/10.1021/jp9716952 (1997).
36. Szleifer, I. Protein Adsorption on Tethered Polymer Layers: Effect of Polymer Chain Architecture and Composition. *Physica A* **244**, 370–388, https://doi.org/10.1016/S0378-4371(97)00293-8 (1997).
37. Jeon, S. I., Lee, J. H., Andrade, J. D. & Degennes, P. G. Protein Surface Interactions in the Presence of Polyethylene Oxide .1. Simplified Theory. *J Colloid Interf Sci* **142**, 149–158, https://doi.org/10.1016/0021-9797(91)90043-8 (1991).
38. Jeon, S. I. & Andrade, J. D. Protein Surface Interactions in the Presence of Polyethylene Oxide .2. Effect of Protein Size. *J Colloid Interf Sci* **142**, 159–166, https://doi.org/10.1016/0021-9797(91)90044-9 (1991).
39. Le, T. C. *et al*. An Experimental and Computational Approach to the Development of ZnO Nanoparticles that are Safe by Design. *Small* **12**, 3568–3577, https://doi.org/10.1002/smll.201600597 (2016).
40. Le, T. C., Yan, B. & Winkler, D. A. Robust Prediction of Personalized Cell Recognition from a Cancer Population by a Dual Targeting Nanoparticle Library. *Adv Funct Mater* **25**, 6927–6935, https://doi.org/10.1002/adfm.201502811 (2015).
41. Autefage, H. *et al*. Sparse Feature Selection Methods Identify Unexpected Global Cellular Response to Strontium-Containing Materials. *Proc Natl Acad Sci USA* **112**, 4280–4285, https://doi.org/10.1073/pnas.1419799112 (2015).
42. Yin, H. *et al*. A Comparative Study of the Physical and Chemical Properties of Nano-Sized ZnO Particles from Multiple Batches of Three Commercial Products. *J Nanopart Res* **17**, ARTN 96, https://doi.org/10.1007/s11051-014-2851-y (2015).
43. Ghose, A. K., Viswanadhan, V. N. & Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J Comb Chem* **1**, 55–68, https://doi.org/10.1021/cc9800071 (1999).
44. Szleifer, I. Protein Adsorption on Surfaces with Grafted Polymers: A theoretical Approach. *Biophys J* **72**, 595–612, https://doi.org/10.1016/S0006-3495(97)78698-3 (1997).
45. Li, L. Y., Chen, S. F., Zheng, J., Ratner, B. D. & Jiang, S. Y. Protein Adsorption on Oligo(Ethylene Glycol)- Terminated Alkanethiolate Self-Assembled Monolayers: The Molecular Basis for Nonfouling Behavior. *J Phys Chem B* **109**, 2934–2941, https://doi.org/10.1021/jp0473321 (2005).
46. Chen, W. L., Cordero, R., Tran, H. & Ober, C. K. 50th Anniversary Perspective: Polymer Brushes: Novel Surfaces for Future Materials. *Macromol* **50**, 4089–4113, https://doi.org/10.1021/acs.macromol.7b00450 (2017).
47. Emmenegger, C. R. *et al*. Interaction of Blood Plasma with Antifouling Surfaces. *Langmuir* **25**, 6328–6333, https://doi.org/10.1021/la900083s (2009).
48. Riedel, T. *et al*. Complete Identification of Proteins Responsible for Human Blood Plasma Fouling on Poly(ethylene glycol)-Based Surfaces. *Langmuir* **29**, 3388–3397, https://doi.org/10.1021/la304886r (2013).
49. Unsworth, L. D., Sheardown, H. & Brash, J. L. Polyethylene Oxide Surfaces of Variable Chain Density by Chemisorption of PEO-Thiol on Gold: Adsorption of Proteins from Plasma Studied by Radiolabelling and Immunoblotting. *Biomater* **26**, 5927–5933, https://doi.org/10.1016/j.biomaterials.2005.03.010 (2005).
50. Leckband, D., Sheth, S. & Halperin, A. Grafted Poly(Ethylene Oxide) Brushes as Nonfouling Surface Coatings. *J Biomat Sci-Polym E* **10**, 1125–1147, https://doi.org/10.1163/156856299x00720 (1999).
51. Molino, P. J. *et al*. Hydration Layer Structure of Biofouling-Resistant Nanoparticles. *ACS Nano*. https://doi.org/10.1021/acsnano.8b06856 (2018).
52. Settanni, G. *et al*. Protein Corona Composition of Poly(Ethylene Glycol)- and Poly(Phosphoester)-Coated Nanoparticles Correlates Strongly with the Amino Acid Composition of the Protein Surface. *Nanoscale* **9**, 2138–2144, https://doi.org/10.1039/c6nr07022a (2017).
53. Norde, W. My Voyage of Discovery to Proteins in Flatland and Beyond. *Colloids Surf B Biointerf* **61**, 1–9, https://doi.org/10.1016/j.colsurfb.2007.09.029 (2008).
54. Schottler, S. *et al*. Protein Adsorption is Required for Stealth Effect of Poly(Ethylene Glycol)- and Poly(Phosphoester)-Coated Nanocarriers. *Nat Nanotechnol* **11**, 372–377, https://doi.org/10.1038/Nnano.2015.330 (2016).
55. Lundqvist, M. *et al*. Nanoparticle Size and Surface Properties Determine the Protein Corona with Possible Implications for Biological Impacts. *Proc Natl Acad Sci USA* **105**, 14265–14270, https://doi.org/10.1073/pnas.0805135105 (2008).
56. Sunde, M., Kwan, A. H. Y., Templeton, M. D., Beever, R. E. & Mackay, J. P. Structural Analysis of Hydrophobins. *Micron* **39**, 773–784, https://doi.org/10.1016/j.micron.2007.08.003 (2008).
57. Heilmann, C., Hussain, M., Peters, G. & Gotz, F. Evidence for Autolysin-Mediated Primary Attachment of Staphylococcus Epidermidis to a Polystyrene Surface. *Mol Microbiol* **24**, 1013–1024, https://doi.org/10.1046/j.1365-2958.1997.4101774.x (1997).

Acknowledgements

TCL thanks RMIT University for the Vice Chancellor's Research Fellowship.

Author Contributions

Matthew Penna and Tu C. Le conceived the project and wrote the manuscript. Tu C. Le performed the modelling work. David A. Winkler and Irene Yarovsky helped with the interpretation of results. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-36597-5.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019