Robust Prediction of Personalized Cell Recognition from a Cancer Population by a Dual Targeting Nanoparticle Library

*Tu C. Le, Bing Yan, and David A. Winkler**

Nanomaterials are used increasingly in diagnostics and therapeutics, particularly for malignancies. Efficient targeting of nanoparticles to specific cells is an important requirement for the development of successful nanoparticle-based theranostics and personalized medicines. Gold nanoparticles are surface modified using a library of small organic molecules, and optionally folate, to investigate their ability to target four cell lines from common cancers, three having high levels of folate receptors expression. Uptake of these nanoparticles varies widely with surface chemistriy and cell lines. Sparse machine learning methods are used to computationally model surface chemistry–uptake relationships, to make quantitative predictions of uptake for new nanoparticle surface chemistries, and to elucidate molecular aspects of the interactions. The combination of combinatorial surface chemistry modification and machine learning models will facilitate the rapid development of targeted theranostics.

1. Introduction

Understanding how nanoconstructs recognize diseased cells is essential for the development of effective and selective theranostics. It is clear that, while markers such as folate receptors[1] and other proteins on cancer cells,[2] or aldehyde dehydrogenase receptors on cancer stem cells[3] can provide some selective targeting of therapeutics or diagnostics, a single protein marker will not be enough to achieve the selectivity required. A better understanding of the types surface marker populations in specific types of cells is necessary to generate the selectivity required to generate maximum efficacy against cancers, and to avoid effects on healthy bystander cells that also express a subset of these markers. Personalized cancer therapy requires medicine specifically formulated and delivered to each patient. However, personalized drug delivery is still a very substantial challenge. We and others have shown that surface modification of nanoparticles by libraries of small organic molecules can markedly alter the uptake in different types of cells, and that the modulation of uptake by surface chemistry can be modeled and predicted quantitatively.[4–6] We hypothesized that dual targeting nanoconstructs, i.e., one ligand targets a common receptor overexpressed in a type of cancer and another ligand targets a receptor related only to an individual's genetic background, may get us closer to the goal of personalized delivery. Biochemical studies of diverse markers on cells are complex and potentially time-consuming, and high throughput materials science such as nanocombinatorial chemistry[7] can provide rapid results that are synergistic to these traditional and more detailed biochemical studies. In particular, nanoparticles, increasingly used to deliver therapies to cells or in disease diagnostics, can also play an important role in understanding how to selectively target cells. Cell uptake of nanoparticles can depend on a number of factors, such as size, shape, surface chemistry and charge, surface coatings, etc.[8]

We have developed a library of nanoparticles that display a diverse array of surface chemistries and have used these to probe the interaction of functionalized nanoparticles with cells, proteins, and enzymes.[4,6,9] To tackle the complex issue of understanding how to specifically target multiple markers on cells we have devised experiments employing a bespoke array of surface functionalized nanoparticles, where the well-known folate targeting moiety (folic acid, FA) is combined with a combinatorial mixture of different surface chemistries previously shown to be taken up selectively by different types of cells.[10] Herein, we describe how these data may be used to generate quantitative numerical models that predict the uptake of nanoparticles by several types of cells derived from common tumors, identify potential synergistic interactions of folate with other types of surface chemistry in cell-specific targeting, and describe how different surface chemistries relate to the specific uptake. These are data-driven machine-learning models that provide quantitative predictions of nanoparticle uptake in complex environments, albeit with some sacrifice of mechanistic insight.

2. Results and Discussion

It is clear from inspection of **Figure 1** and **Table 1** that the surface chemistry has a marked effect on the degree of uptake

Dr. T. C. Le, Prof. D. A. Winkler
CSIRO Manufacturing
Clayton 3169, Australia
E-mail: dave.winkler@csiro.au
Prof. B. Yan
School of Chemistry and Chemical Engineering
Shandong University
Jinan 250100, China
Prof. D. A. Winkler
Monash Institute of Pharmaceutical Sciences
Parkville 3052, Australia
Prof. D. A. Winkler
Latrobe Institute for Molecular Science
Bundoora 3083, Australia
Prof. D. A. Winkler
School of Chemical and Physical Sciences
Flinders University
Bedford Park 5042, Australia

DOI: 10.1002/adfm.201502811

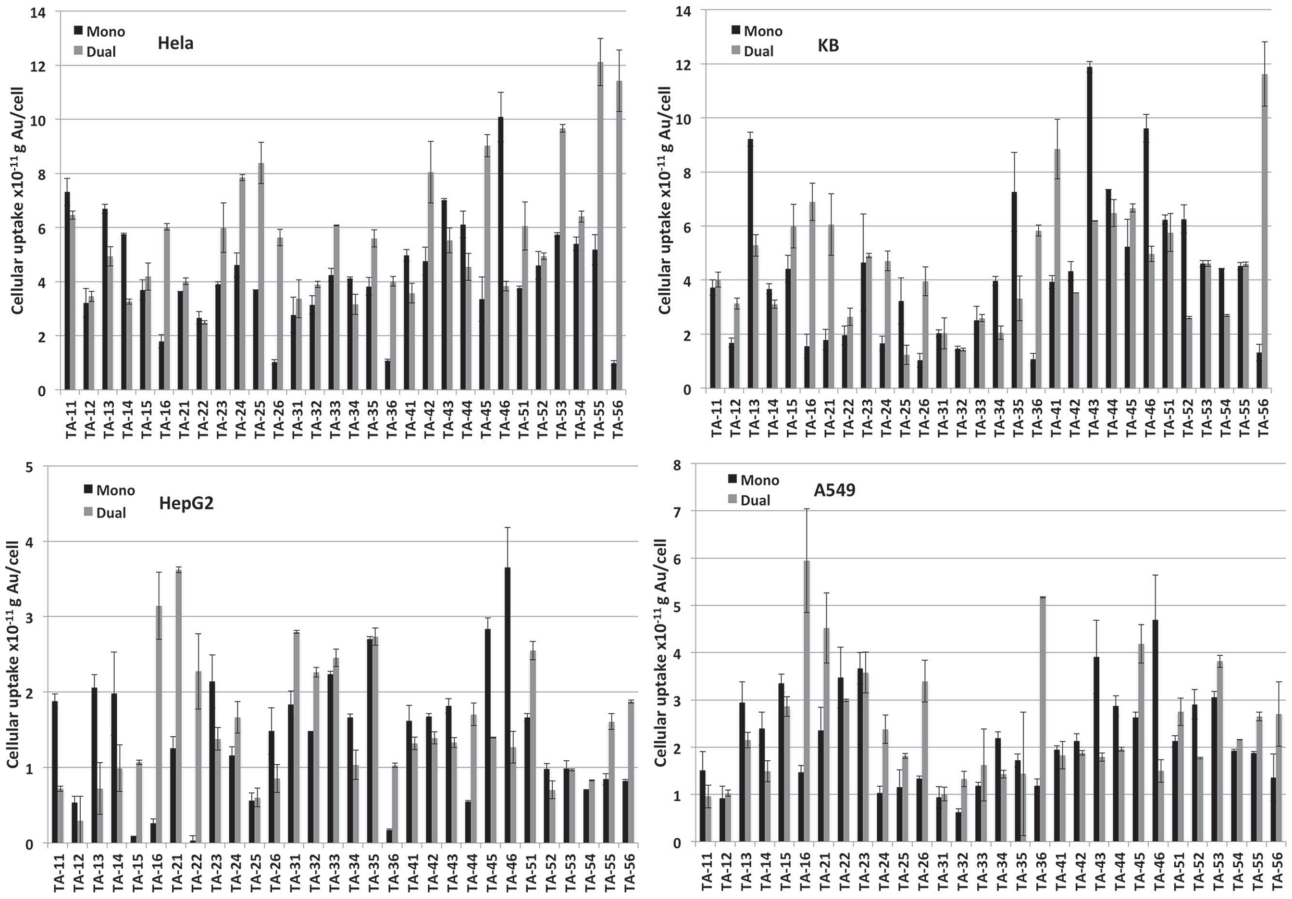


Figure 1. Histograms showing the variation of cellular uptake of modified nanoparticles as a function of surface chemistry.

of the nanoparticles. There is considerable variation in the response of different cancer cell lines to the nanoparticle surface chemistry. There is no correlation between the uptake of mono- and dual-ligand nanoparticles for the same cell line (r^2 = 0 for all cell lines), and low-to-moderate correlation between the responses of different cell lines to the same library of surface-modified nanoparticles (r^2 between 0.22 and 0.54 for the monoligand library and between 0.0 and 0.18 for the dual-ligand library).

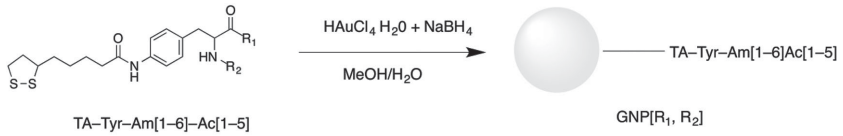
2.1. Dual-Ligand Nanoparticle Cellular Uptake

2.1.1. HeLa Cells

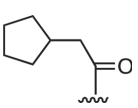
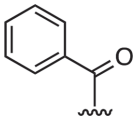
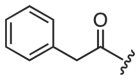
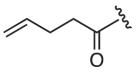
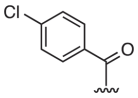
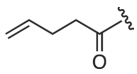
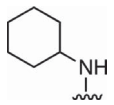
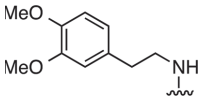
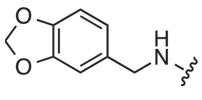
Both linear and nonlinear models were constructed to predict the uptake of the nanoparticles with organic ligands plus FA attached to their surfaces in cells derived from cervical cancer. The sparsity of the models was gradually increased to reduce the number of descriptors from an initial pool of 482 descriptors. The quality of the models degraded significantly when the number of descriptors in the model was reduced to below 13, indicating that relevant information was being removed from the models. The performance of the optimally sparse models is shown in **Table 2**. These models were constructed using 13 descriptors listed in Table S1 (Supporting Information) and their abilities to predict the training and test sets are illustrated in **Figure 2**. As can be seen, both linear MLREM and nonlinear BRANNGP models could predict very well the uptake of nanoparticles by HeLa cell, accounting for more than 90% of the variance in the data. The linear model had an r^2 value of 0.98 and a standard error of estimation (SEE) of 0.58×10^{-11} g Au cell^{-1} for the training set and an r^2 value of 0.93 and a standard error of prediction (SEP) of 0.81×10^{-11} g Au cell^{-1} for the test set. The nonlinear neural network model achieved better predictivity, with an SEE of only 0.28×10^{-11} g Au cell^{-1} and an SEP of 0.76×10^{-11} g Au cell^{-1}, suggesting some nonlinearity in the relationship between surface chemistry and uptake.

The effects of the surface ligands encoded by the descriptors on the cervical cancer cellular uptake of dual-ligand nanoparticles are shown in Figure S1 in the Supporting Information. A positive value of the multiple linear regression (MLR) coefficient for a descriptor indicates that the descriptor promotes cellular uptake whereas a negative value of the MLR coefficient indicates that the descriptor inhibits the cellular uptake. For example, the sum of geometrical distances between nitrogen atoms in the nanoparticles has a negative coefficient. Larger distances between nitrogen atoms result in a smaller HeLa cell uptake of nanoparticles carrying this functionality, and the converse situation applies. In contrast, the sum of geometrical distances between oxygen and chlorine atoms has a positive coefficient suggesting that larger distances between these

Table 1. Surface chemistry and uptake of fGNPs in four cancer cell lines (units: 10^{-11} g Au cell^{-1}). TA is thioctic acid (1,2-dithiolan-3-yl)pentanecarboxylic). The first uptake number is for organic ligand alone (mono), the second figure is for organic ligand plus folate (dual).

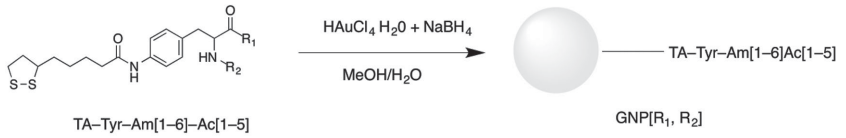


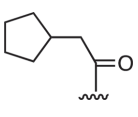
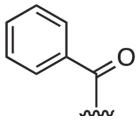
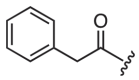
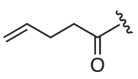
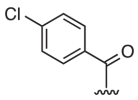
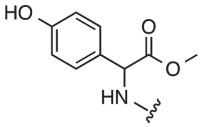
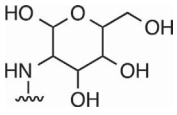
TA–Tyr–Am[1–6]–Ac[1–5]

					
	11 HeLa 7.32 ± 0.51 6.46 ± 0.15 KB 3.72 ± 0.3 4.02 ± 0.3 HepG2 1.88 ± 0.10 0.72 ± 0.03 A549 1.51 ± 0.39 0.96 ± 0.24	21 HeLa 3.64 ± 0.00 4.01 ± 0.13 KB 1.79 ± 0.4 6.05 ± 1.1 HepG2 1.26 ± 0.15 3.62 ± 0.04 A549 2.35 ± 0.49 4.52 ± 0.74	31 HeLa 2.77 ± 0.66 3.36 ± 0.70 KB 2.04 ± 0.1 2.03 ± 0.6 HepG2 1.84 ± 0.18 2.80 ± 0.02 A549 0.94 ± 0.23 1.01 ± 0.15	41 HeLa 4.97 ± 0.22 3.57 ± 0.37 KB 3.93 ± 0.2 8.84 ± 1.1 HepG2 1.62 ± 0.20 1.32 ± 0.08 A549 1.94 ± 0.08 1.83 ± 0.29	51 HeLa 3.75 ± 0.07 6.05 ± 0.89 KB 6.24 ± 0.2 5.76 ± 0.7 HepG2 1.66 ± 0.05 2.55 ± 0.12 A549 2.13 ± 0.12 2.75 ± 0.29
	12 HeLa 3.22 ± 0.53 3.45 ± 0.19 KB 1.69 ± 0.2 3.13 ± 0.2 HepG2 0.53 ± 0.09 0.30 ± 0.32 A549 0.92 ± 0.25 1.02 ± 0.06	22 HeLa 2.65 ± 0.24 2.48 ± 0.07 KB 1.96 ± 0.4 2.64 ± 0.3 HepG2 0.03 ± 0.06 2.28 ± 0.50 A549 3.47 ± 0.64 2.99 ± 0.02	32 HeLa 3.13 ± 0.34 3.89 ± 0.12 KB 1.47 ± 0.1 1.44 ± 0.1 HepG2 1.48 ± 0.00 2.26 ± 0.06 A549 0.63 ± 0.07 1.32 ± 0.17	42 HeLa 4.75 ± 0.52 8.04 ± 1.14 KB 4.33 ± 0.4 3.53 ± 0.0 HepG2 1.68 ± 0.03 1.39 ± 0.08 A549 2.13 ± 0.15 1.88 ± 0.05	52 HeLa 4.60 ± 0.51 4.93 ± 0.12 KB 6.25 ± 0.5 2.61 ± 0.0 HepG2 0.99 ± 0.06 0.70 ± 0.12 A549 2.90 ± 0.31 1.77 ± 0.02
	13 HeLa 6.69 ± 0.17 4.93 ± 0.36 KB 9.21 ± 0.3 5.28 ± 0.4 HepG2 2.06 ± 0.18 0.72 ± 0.34 A549 2.95 ± 0.44 2.15 ± 0.16	23 HeLa 3.90 ± 0.09 5.99 ± 0.92 KB 4.66 ± 1.8 4.90 ± 0.1 HepG2 2.14 ± 0.35 1.38 ± 0.15 A549 3.67 ± 0.33 3.58 ± 0.43	33 HeLa 4.25 ± 0.24 6.07 ± 0.02 KB 2.52 ± 0.5 2.59 ± 0.1 HepG2 2.24 ± 0.04 2.45 ± 0.12 A549 1.19 ± 0.07 1.62 ± 0.76	43 HeLa 7.02 ± 0.05 5.52 ± 0.46 KB 11.88 ± 0.2 6.19 ± 0.0 HepG2 1.82 ± 0.10 1.33 ± 0.06 A549 3.91 ± 0.78 1.79 ± 0.09	53 HeLa 5.74 ± 0.07 9.66 ± 0.15 KB 4.61 ± 0.1 4.61 ± 0.1 HepG2 0.99 ± 0.11 0.97 ± 0.01 A549 3.06 ± 0.11 3.82 ± 0.13
	14 HeLa 5.75 ± 0.04 3.26 ± 0.10 KB 3.68 ± 0.2 3.10 ± 0.2 HepG2 1.98 ± 0.55 0.99 ± 0.31 A549 2.39 ± 0.35 1.49 ± 0.22	24 HeLa 4.61 ± 0.44 7.85 ± 0.11 KB 1.66 ± 0.3 4.71 ± 0.4 HepG2 1.16 ± 0.12 1.67 ± 0.21 A549 1.03 ± 0.15 2.38 ± 0.30	34 HeLa 4.11 ± 0.05 3.16 ± 0.37 KB 3.97 ± 0.2 2.06 ± 0.2 HepG2 1.67 ± 0.05 1.04 ± 0.20 A549 2.19 ± 0.13 1.43 ± 0.08	44 HeLa 6.11 ± 0.50 4.54 ± 0.50 KB 7.34 ± 0.0 6.48 ± 0.5 HepG2 0.55 ± 0.02 1.71 ± 0.15 A549 2.87 ± 0.22 1.96 ± 0.04	54 HeLa 5.39 ± 0.25 6.40 ± 0.20 KB 4.41 ± 0.0 2.69 ± 0.0 HepG2 0.70 ± 0.01 0.83 ± 0.01 A549 1.92 ± 0.04 2.16 ± 0.00

Continued

Table 1. Continued



					
	15	25	35	45	55
	HeLa	HeLa	HeLa	HeLa	HeLa
	3.69 ± 0.38 4.19 ± 0.51	3.70 ± 0.00 8.38 ± 0.76	3.82 ± 0.33 5.59 ± 0.31	3.35 ± 0.82 9.02 ± 0.41	5.18 ± 0.57 12.11 ± 0.87
	KB	KB	KB	KB	KB
	4.42 ± 0.5 5.99 ± 0.8	3.23 ± 0.8 1.24 ± 0.4	7.26 ± 1.5 3.32 ± 0.8	5.24 ± 1.0 6.66 ± 0.2	4.52 ± 0.1 4.59 ± 0.1
	HepG2	HepG2	HepG2	HepG2	HepG2
	0.09 ± 0.00 1.07 ± 0.03	0.56 ± 0.10 0.60 ± 0.12	2.71 ± 0.03 2.74 ± 0.11	2.83 ± 0.15 1.40 ± 0.01	0.85 ± 0.07 1.61 ± 0.10
	A549	A549	A549	A549	A549
	3.35 ± 0.19 2.86 ± 0.21	1.15 ± 0.37 1.81 ± 0.05	1.72 ± 0.14 1.44 ± 1.31	2.63 ± 0.11 4.19 ± 0.41	1.88 ± 0.02 2.65 ± 0.10
	16	26	36	46	56
	HeLa	HeLa	HeLa	HeLa	HeLa
	1.79 ± 0.25 6.03 ± 0.12	1.03 ± 0.09 5.63 ± 0.30	1.08 ± 0.04 4.01 ± 0.18	10.08 ± 0.92 3.83 ± 0.18	0.99 ± 0.10 11.42 ± 1.14
	KB	KB	KB	KB	KB
	1.55 ± 0.4 6.89 ± 0.7	1.04 ± 0.3 3.95 ± 0.5	1.08 ± 0.2 5.83 ± 0.2	9.61 ± 0.5 4.97 ± 0.3	1.32 ± 0.3 11.62 ± 1.2
	HepG2	HepG2	HepG2	HepG2	HepG2
	0.26 ± 0.06 3.14 ± 0.45	1.49 ± 0.31 0.86 ± 0.19	0.17 ± 0.02 1.03 ± 0.03	3.65 ± 0.53 1.27 ± 0.21	0.83 ± 0.01 1.87 ± 0.02
	A549	A549	A549	A549	A549
	1.47 ± 0.14 5.95 ± 1.10	1.34 ± 0.05 3.40 ± 0.44	1.19 ± 0.14 5.17 ± 0.01	4.69 ± 0.95 1.49 ± 0.24	1.35 ± 0.50 2.70 ± 0.69
Folic acid alone	HeLa	KB	HepG2	A549	
	3.49 ± 0.15	3.86 ± 0.20	0.77 ± 0.27	1.75 ± 0.02	

atoms in substituents promotes HeLa cell uptake. Among the descriptors used in the models, Mor06u, Mor11m, Mor19m, Mor18e, and Mor30p, MoRSE descriptors have the most significant effects on the cellular uptake. These descriptors measure the sinusoidal radial distribution of atoms in the molecules unweighted or weighted by atomic mass, van der Waals volume, polarizability, or electronegativity. Mass-weighted descriptors are most significant for heavy atoms such as P, S, Cl, Br, and I, while volume-weighted descriptors are more influenced by Si, P, Br, and I. Polarizability-weighted descriptors are similar to volume-weighted ones but better describe the effect of elements with more deformable electron clouds. Electronegativity-weighted descriptors describe the contributions of F, O, and Cl. Most MoRSE descriptors make a negative contribution to the HeLa uptake models, with Mor19m having the most significant effect on cellular uptake. Indeed, nanoparticle TA-12 that has the highest Mor19m value has one the lowest value of HeLa cellular uptake of $<4 \times 10^{-11}$ g Au cell^{-1} whereas TA-55 with the lowest Mor19m value has the highest HeLa uptake of $>12 \times 10^{-11}$ g Au cell^{-1}. A more detailed explanation of MoRSE descriptors and their interpretation could be found elsewhere.[11]

2.1.2. KB Cell Uptake

Uptake of nanoparticles by cells derived from epidermal cancers (KB cells) was also modeled using both linear (MLREM) and nonlinear (BRANNGP) methods. The same pool of 482 descriptors calculated by the DRAGON software was used to construct the models of cellular uptake of nanoparticles. By progressively increasing the model sparsity, we found the optimal number of descriptors in the model was 8 (7 plus the intercept) as can be seen in Table 2. The statistical significance of these models was not as high as the HeLa cell uptake models, but the overall prediction of the training and test set data was good, accounting for about 70% of the variance in the data. The best linear model with eight effective weights had an r^2 value of 0.80 and SEE of 1.32×10^{-11} g Au cell^{-1} for the training set and r^2 value of 0.71 and SEP of 1.78×10^{-11} g Au cell^{-1} for the test set. The best nonlinear model performed noticeably better, with an SEE of 1.04×10^{-11} g Au cell^{-1} and SEP of 1.50×10^{-11} g Au cell^{-1}. The descriptors used for these models and their descriptions are listed in Table S2 (Supporting Information) and the performance of the models in predicting the

Table 2. Statistical results of the best MLREM and BRANNGP models for the uptake of single- and dual-ligand nanoparticles (N_{eff} is number of effective weights (adjustable parameters) in the model).

Cell line	Method	N_{eff}	Training set		Test set	
			r^2	SEE $\times 10^{-11}$ g Au cell^{-1}	r^2	SEP $\times 10^{-11}$ g Au cell^{-1}
HeLa (dual)	MLREM	14	0.98	0.58	0.93	0.81
	BRANNGP	15	0.96	0.28	0.94	0.76
KB (dual)	MLREM	8	0.80	1.32	0.71	1.78
	BRANNGP	8	0.71	1.04	0.67	1.50
HepG2 (dual)	MLREM	14	0.88	0.43	0.72	0.53
	BRANNGP	12	0.76	0.28	0.71	0.56
A549 (dual)	MLREM	16	0.99	0.20	0.93	0.55
	BRANNGP	16	0.96	0.12	0.93	0.55
HeLa (mono)	MLREM	14	0.88	1.00	0.82	1.25
	BRANNLP	14	0.87	0.53	0.85	1.36

KB cellular uptake for the training and test sets illustrated in **Figure 3**.

Figure S2 (Supporting Information) shows the effect of all seven descriptors used in the QSPR models of the uptake of dual-ligand nanoparticles by cells from epidermal cancer (KB). The descriptors contributing most strongly to the models are G(N···O) and Mor18p. The positive coefficients for these descriptors mean that they promote uptake by KB cells. Collectively, nanoparticles with functionality having the largest nitrogen–oxygen atom distances and high Mor18p descriptors are will have the best KB cellular uptake. However, in the data set considered here, no nanoparticle has surface functionality with the highest values of the G(N···O) and Mor18p descriptors. Nanoparticle TA-36 has the highest G(N···O) descriptor value has a relatively small Mor18p whereas nanoparticle TA-11 with the highest Mor18p descriptor has only a modest G(N···O) descriptor magnitude. Nanoparticle TA-56 with the highest uptake in KB cells has a G(N···O) descriptor in the top range but a modest Mor18p value. The nonlinear uptake model being substantially better than the linear model, suggests there must be substantial nonlinearity and interaction between the molecular properties, so optimization of surface chemistry in this case will be more complex than for structure-uptake models that are essentially linear.

2.1.3. HepG2 Cell Uptake

We also attempted to model the uptake of dual-ligand nanoparticles in cells derived from hepatocellular carcinoma (HepG2 cells) using MLREM and BRANNGP methods. As Table 2 shows, both linear and nonlinear models could recapitulate the HepG2 cellular uptake of the training and test sets with good fidelity. The optimally sparse linear model using 14 descriptors had an r^2 value of 0.88 and SEE of 0.43×10^{-11} g Au cell^{-1} for the training set and an r^2 value of 0.72 and SEP of 0.53×10^{-11} g Au cell^{-1} for the test set. The nonlinear model with SEP of 0.56×10^{-11} g Au cell^{-1} clearly did not perform any better than the linear one. **Figure 4** illustrates the abilities of these models to predict the HepG2 cellular uptake for both training and test sets.

Table S3 (Supporting Information) shows the important descriptors for the HepG2 cellular uptake of dual-ligand nanoparticles. The E2s WHIM index was the most important

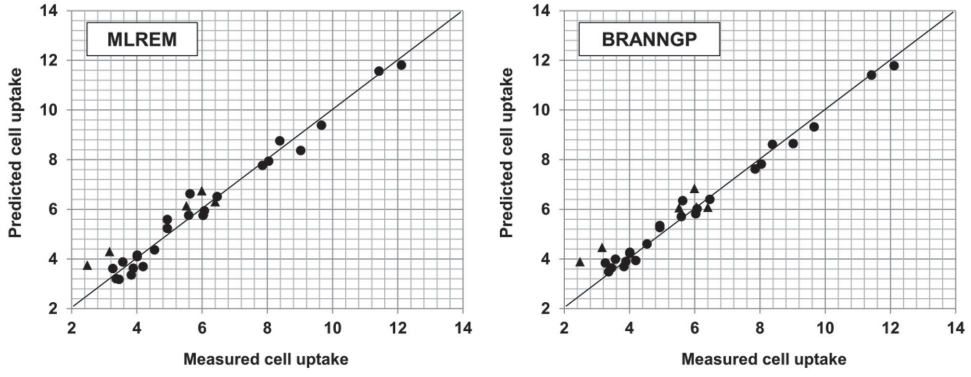


Figure 2. Measured and predicted HeLa cell uptake ($\times 10^{-11}$ g Au cell^{-1}) of nanoparticles with dual ligands on their surfaces. The left panel is the linear model and the right panel is the nonlinear model. The training set is denoted by circles, and the test set by triangles.

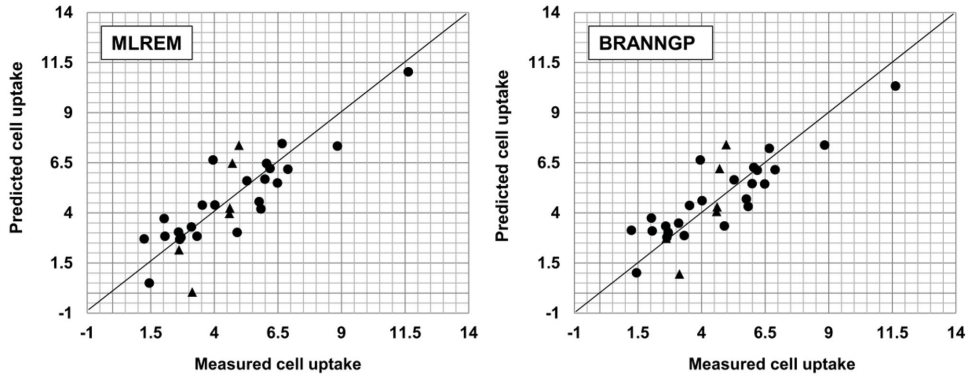


Figure 3. Measured and predicted KB cell uptake ($\times 10^{-11}$ g Au cell^{-1}) of nanoparticles with dual ligands on their surfaces. The left panel is the linear model and the right panel is the nonlinear model. The training set is denoted by circles, and the test set by triangles.

descriptors (Figure S3, Supporting Information) in the models. WHIM descriptors, developed by Todeschini et al.,[12] are statistical indices that measure the projections of molecular properties along principal axes of the molecule. They capture the 3D distribution of molecular size, shape, symmetry, and atom distribution with respect to invariant reference frames. The large negative coefficient for E2s means that nanoparticles with surface chemistries with large E2s descriptors will have lower HepG2 cellular uptakes. The lowest HepG2 cellular uptake in the data set is 0.3×10^{-11} g Au cell^{-1} for nanoparticle TA-12 that has E2s of value of 0.583, very close to the maximum value of 0.636. Clearly, the cellular uptake of nanoparticles is a result of the contribution of all descriptors but E2s makes an important contribution.

2.1.4. A549 Cell Uptake

The uptake data for A549 calls generated models with the best performance of the four cell lines, in spite of this cell line reportedly having a lower expression of folate receptors. The linear model using 15 descriptors and an intercept had an r^2 value of 0.99 and SEE of 0.20×10^{-11} g Au cell^{-1} for the training set and an excellent r^2 value of 0.93 and SEP of 0.55×10^{-11} g Au cell^{-1} for the test set (Table 2). The performance of the nonlinear BRANNGP model was essentially the same as that of the linear model indicating a linear relationship between nanoparticle surface chemistry and uptake by A549 cells. The SEP value was 0.55×10^{-11} g Au cell^{-1} for both MLREM and BRANNGP models, representing about 10% of the range of the A549 nanoparticle uptake. The performance of the models in predicting the A549 cell uptake of dual-ligand nanoparticles is illustrated in **Figure 5**.

Descriptors used in the models are listed in Table S4 (Supporting Information) and their significant contributions to the A549 cellular uptake of dual-ligand nanoparticles are shown in Figure S4, Supporting Information. One of the MoRSE descriptors, Mor21e that describes the distribution of electronegativities in the molecule with respect to an angular (sinusoidal) scattering function, was the most relevant positive contributor to the cellular uptake models. 3D MoRSE descriptors (3D Molecule Representation of Structures based on Electron diffraction), proposed by Gasteiger and Schuur,[13] are derived from infrared spectra simulations. Nanoparticles

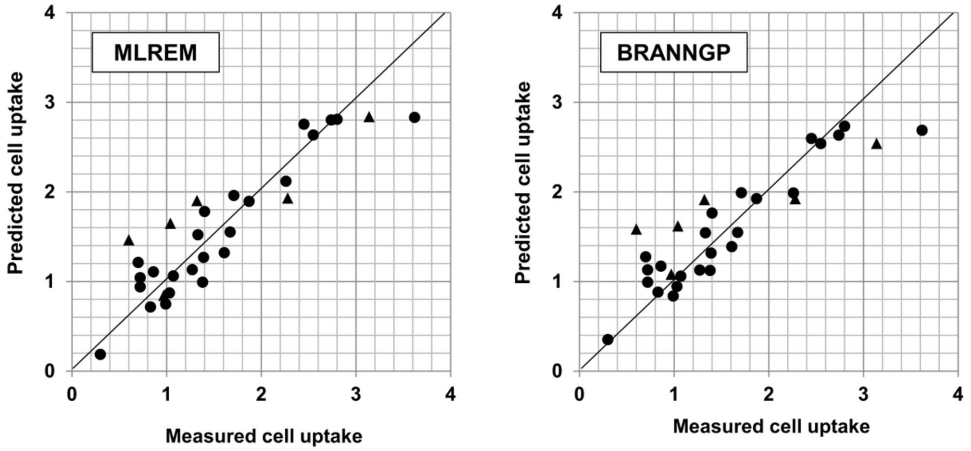


Figure 4. Measured and predicted HepG2 cell uptake ($\times 10^{-11}$ g Au cell^{-1}) of nanoparticles with dual ligands on their surfaces. The left panel is the linear model and the right panel is the nonlinear model. The training set is denoted by circles, and the test set by triangles.

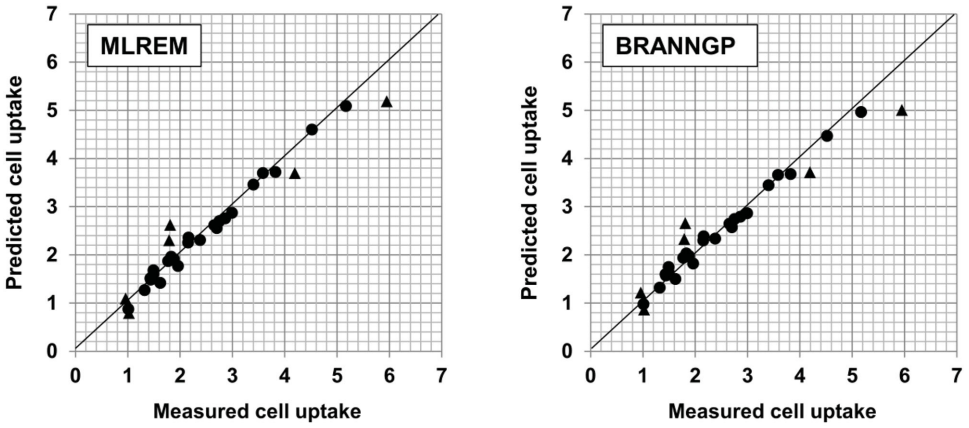


Figure 5. Measured and predicted A549 cell uptake ($\times 10^{-11}$ g Au cell^{-1}) of nanoparticles with dual ligands on their surfaces. The left panel is the linear model and the right panel is the nonlinear model. The training set is denoted by circles, and the test set by triangles.

with high value of Mor21e have higher uptake and vice versa. However, there are a number of other descriptors with similar significance such as E2m, G3v, PJI3, G3p, and RDF115p. Therefore, nanoparticle TA-45 with the highest Mor21e value does not have the highest A549 cellular uptake of 6×10^{-11} g Au cell^{-1} because the values for other descriptors are not optimal. However, its cellular uptake of 4.19×10^{-11} g Au cell^{-1} is still the top four in the thirty nanoparticle list.

2.2. Single Ligand Nanoparticle Cellular Uptake

As we noted before, there was no correlation between the uptake of nanoparticles with or without FA on the surface for any cell lines. We also attempted to model the cellular uptake of monoligand nanoparticles. However, the original pool of 482 descriptors did not give good models for these data sets. We therefore employed a larger pool of 699 descriptors calculated by the DRAGON software. With these challenging data sets, sparse feature selection using linear methods did not result in models with good statistical significance for most cancer cell lines. Nonlinear sparse feature selection using BRANNLP could only generate good models for the cervical cancer cellular uptake data and significantly ill-posed models for A549 cellular uptake, which we considered not sufficiently reliable to be useful. No satisfactory models were built for the HepG2 and KB cellular uptakes of modified gold nanoparticles. This suggests that the organic ligands in our library had relatively small cell targeting abilities on their own for these two cell lines, but were able to synergize with FA to target all four cancer cell lines. In some cases the organic ligands generated a substantially higher uptake than FA alone (e.g., 11.4×10^{-11} g Au cell^{-1} vs 3.5×10^{-11} g Au cell^{-1} for HeLa, 11.6×10^{-11} g Au cell^{-1} vs 3.9×10^{-11} g Au cell^{-1} for KB cells, 3.6×10^{-11} g Au cell^{-1} vs 0.8×10^{-11} g Au cell^{-1} for HepG2 cells, and 6.0×10^{-11} g Au cell^{-1} vs 1.8×10^{-11} g Au cell^{-1} for A549 cells). Other surface chemistries appeared to antagonize the uptake compared to FA alone (e.g., 2.5×10^{-11} g Au cell^{-1} vs 3.5×10^{-11} g Au cell^{-1} for HeLa cells, 1.2×10^{-11} g Au cell^{-1} vs 3.9×10^{-11} g Au cell^{-1} for KB cells, 0.3×10^{-11} g Au cell^{-1} vs 0.8×10^{-11} g Au cell^{-1} for HepG2 cells, and 1.0×10^{-11} g Au cell^{-1} vs 1.8×10^{-11} g Au cell^{-1} for A549 cells).

2.2.1. HeLa Cell Uptake

The nonlinear sparse feature selection pruned out irrelevant descriptors and kept the 13 most relevant descriptors to build the models predicting the HeLa uptake of nanoparticles with diverse organic ligands on their surface, but no FA. Using these descriptors, the linear models had an r^2 value of 0.88 and SEE of 1.00×10^{-11} g Au cell^{-1} for the training set and an r^2 value of 0.82 and SEP of 1.25×10^{-11} g Au cell^{-1} for the test set (Table 2). The nonlinear model had similar performance with r^2 value of 0.87 and SEE of 0.53×10^{-11} g Au cell^{-1} for the training set and r^2 value of 0.85 and SEP of 1.36×10^{-11} g Au cell^{-1} for the test set. **Figure 6** illustrates the abilities of both models in predicting the HeLa uptake of monoligand nanoparticles for the training and test sets.

Figure S5 (Supporting Information) shows the effects of 13 different descriptors selected by BRANNLP model on the cervical cancer cellular (HeLa) uptake of monoligand nanoparticles. The MoRSE descriptors, Mor20e (promoting) and Mor20u (inhibiting), were the most significant contributors to the uptake models (Table S5, Supporting Information). Nanoparticles surface chemistries characterized by high Mor20e values have higher cellular uptakes while those with high Mor20u values have lower HeLa uptake. Among the 30 nanoparticles in the data set, nanoparticle TA-13 has not only the largest Mor20e descriptor but also the highest Mor20u descriptor. However, the influence of the Mor20e descriptor dominates and the nanoparticle TA-13 is taken up by HeLa cells at 6.69×10^{-11} g Au cell^{-1}, one of the four highest uptakes in this cell line. Conversely, nanoparticle TA-32 has the smallest Mor20e and Mor20u descriptors and its uptake of 3.13×10^{-11} g Au cell^{-1} was among the lowest.

2.2.2. A549 Cell Uptake

The A549 cellular uptake data set for 30 monoligand nanoparticles was very difficult to model, possibly because it has a more diverse set of surface markers than the other cell lines. Inspection of Table 1 reveals that the uptake of nanoparticles with small organic ligand surfaces but not folate is substantially lower

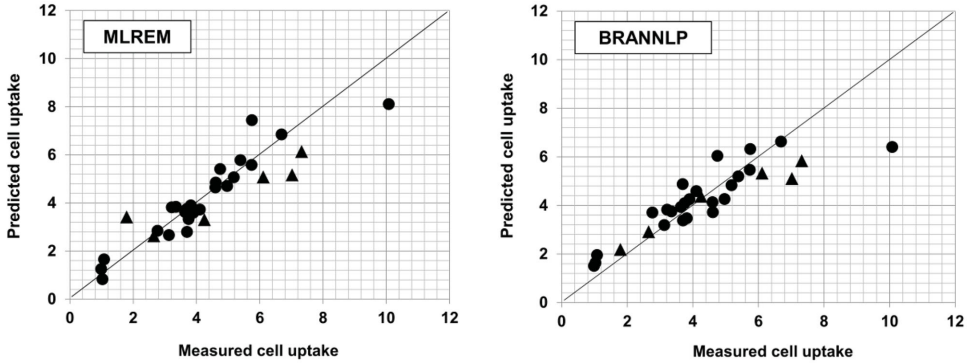


Figure 6. Measured and predicted HeLa cell uptake ($\times 10^{-11}$ g Au cell^{-1}) of nanoparticles with single ligands on their surfaces. The left panel is the linear model and the right panel is the nonlinear model. The training set is denoted by circles, and the test set by triangles.

for A549 cells than for HeLa and KB cells and comparable to HepG2 cells. Both linear and nonlinear sparse feature selection failed to generate statistically significant models. We then tried different combinations of descriptors suggested by a number of very sparse nonlinear BRANNLP models with reasonably low SEP. Although, this led to the generation of one or two models that had apparently acceptable statistics, generation of multiple models of similar sparsity suggested that the models were ill-posed and unstable. Therefore, we have not reported them here.

3. Conclusions

Understanding the effect of surface chemistry on higher affinity and personalized uptake of nanoparticles is critical for the design of new theranostics, but also to ensure nanoparticles do not have significant side effects. We have shown how computational modeling of limited biological data on nanoparticle uptake can elucidate the relationship between surface chemistry and uptake, and allow predictions of uptake to be made reliably (within the applicability domain of the models) for new surface chemistries. Models with high predictivity for cancer cell uptake were generated, providing an important proof of concept for computational approaches based on sparse machine learning methods. The models identified the relative cell targeting provided by small organic ligands and a well-known cancer cell targeting ligand, folic acid. The models will be useful for designing nanoparticle surfaces that can be optimized for uptake in particular cells. Although some descriptors used are arcane, making interpretation in terms of surface chemistry more difficult, the easiest way to use the models is to generate a large virtual library of potential nanoparticle surface ligands and use the models to select those with the best predicted uptake in the cell of choice. If models also exist for multiple cell types, it should be possible to "design in" selectivity as well as uptake efficacy, and get closer to the ideal theranostics nanoparticle with broadly tuneable cell recognition.

4. Experimental Section

We chose for this study cell lines from cervical (HeLa), epidermal (KB), and hepatocellular (HepG2) cancers that are known to overexpress the folate receptor, and adenocarcinomic human alveolar basal epithelial (A549) cells that express very low levels of folate receptor.[14] They should have different surface receptors than HeLa cells because of their different organ origin. The biological data and a description of the nanoparticle library were reported in Zhou et al.[10] In this experimental study, nanoparticles were generated with diverse surface chemistries and the experiments involved assessing uptake of these nanoparticles in the four cancer cell lines alone, or with the addition of the FA ligand. The average number of organic ligands per nanoparticle was 290 ± 60 and the average number of FA groups on each nanoparticle was 30 ± 5. The ratio of the secondary ligand to FA was between 7 and 12 per nanoparticle. All surface-modified nanoparticles had similar ζ potential values from −38 to −40 mV in water, suggesting that they do not agglomerate. This is supported by the size distributions from TEM showing they had average sizes of 7.4–9.9 nm with a standard deviation of 1.2 nm in water and $13–18 \pm 2.2$ nm in phosphate buffered saline (PBS).[9]

Two sets of data corresponding to amide ligands (Table 1) alone (monoligand set) or the amide ligands plus FA (dual-ligand set) on the surface of the nanoparticles were used to build the QSPR models. The data set was divided into eight 30 data point subsets corresponding to cellular uptake in the four cancer cell lines (Hela, KB, HepG2, and A549) with one or two (with FA) types of ligands on the nanoparticle surface. The structures of the 30 different types of surface chemistry (Table 1) plus that of FA were used to generate molecular descriptors that capture their biologically relevant properties. The DRAGON software[15] was used to calculate these descriptors that encoded mathematically properties such as geometry, partial charges, existence of molecular fragments, or distribution of atoms and atomic mass. The data sets were divided into training and test sets which consisted of 80% and 20% of the data, respectively, using *k*-means clustering algorithm. The models were generated using only the training sets and their ability to predict new data that the model has not seen before was validated using the test sets. Three sparse machine-learning methods were employed: multiple linear regression with expectation maximization (MLREM),[16] nonlinear Bayesian regularized artificial neural networks with Gaussian prior (BRANNGP), and nonlinear Bayesian regularized artificial neural networks with Laplacian prior (BRANNLP).[16] The neural networks consisted of input, hidden, and output layers. The number of nodes in the input layer was equal to the number of molecular descriptors whereas the output layer had only one node corresponding to the cellular uptake. Two or three nodes in the hidden layer were found to be sufficient to build good models, and increasing the number is unnecessary as the Bayesian regularization automatically controls the complexity of the models to optimize predictivity.[17] The number of effective weights derived from the neural network models tended to a constant value as the number of hidden layer nodes increased. Details of the Bayesian regularization applied to feed forward neural networks can be found elsewhere.[16,18] The most important descriptors that control the cellular uptakes were identified using sparse feature selection methods: MLREM and BRANNLP. These approaches have been shown to be useful in carrying out sparse linear and nonlinear

descriptor selection. By tuning progressively the sparsity of MLREM and BRANNLP, the least informative descriptors could be pruned out and the most relevant descriptors retained.

The performance of the models obtained was assessed using the coefficient of determination (r^2), the standard error of estimation (SEE), and the standard error of prediction (SEP). The r^2 is the square of the correlation coefficient (r) between the predicted and measured values of the cellular uptake. SEE and SEP are the root-mean-square values, adjusted for degrees of freedom, of the difference between the predicted and measured cellular uptake values for the training and test sets, respectively. They are more robust estimates of the predictive ability of QSPR 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.[4,19]

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This work was supported by the Natural Science Foundation of China (21137002 to B.Y.), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB14030401 to B.Y.), and the Advanced Materials Transformational Capability Platform in CSIRO. We thank Dr. Hongyu Zhou for the experimental data used in the modelling.

Received: July 9, 2015
Revised: September 13, 2015
Published online: October 20, 2015

[1] A. R. Hilgenbrink, P. S. Low, *J. Pharm. Sci.* **2005**, *94*, 2135.
[2] a) J. Mercer, A. Helenius, *Science* **2008**, *320*, 531; b) S. Ran, A. Downes, P. E. Thorpe, *Cancer Res.* **2002**, *62*, 6132; c) A. J. Surman, G. D. Kenny, D. K. Kumar, J. D. Bell, D. R. Casey, R. Vilar, *Chem. Commun.* **2011**, *47*, 10245; d) T. Zhang, R. F. Lan, C. F. Chan, G. L. Law, W. K. Wong, K. L. Wong, *Proc. Natl. Acad. Sci. USA* **2014**, *111*, E5492.
[3] F. Ismail, D. A. Winkler, *ChemMedChem* **2014**, *9*, 885.
[4] V. C. Epa, F. R. Burden, C. Tassa, R. Weissleder, S. Shaw, D. A. Winkler, *Nano Lett.* **2012**, *12*, 5808.
[5] R. Weissleder, K. Kelly, E. Y. Sun, T. Shtatland, L. Josephson, *Nat. Biotechnol.* **2005**, *23*, 1418.
[6] D. A. Winkler, F. R. Burden, B. Yan, R. Weissleder, C. Tassa, S. Shaw, V. C. Epa, *SAR QSAR Environ. Res.* **2014**, *25*, 161.
[7] H. Y. Zhou, Q. X. Mu, N. N. Gao, A. F. Liu, Y. H. Xing, S. L. Gao, Q. Zhang, G. B. Qu, Y. Y. Chen, G. Liu, B. Zhang, B. Yan, *Nano Lett.* **2008**, *8*, 859.
[8] N. Oh, J. H. Park, *Int. J. Nanomed.* **2014**, *9*, 51.
[9] Y. Y. Liu, D. A. Winkler, V. C. Epa, B. Zhang, B. Yan, *Nano Res.* **2015**, *8*, 1293.
[10] H. Y. Zhou, P. F. Jiao, L. Yang, X. Li, B. Yan, *J. Am. Chem. Soc.* **2011**, *133*, 680.
[11] O. Devinyak, D. Havrylyuk, R. Lesyk, *J. Mol. Graph. Model.* **2014**, *54*, 194.
[12] R. Todeschini, V. Consonni, P. Gramatica, in *Comprehensive Chemometrics*, Vol. *4* (Eds: S. Brown, R. Tauler, R. Walczak), Elsevier, Oxford, UK **2009**, 129.
[13] J. H. Schuur, P. Selzer, J. Gasteiger, *J. Chem. Inf. Comp. Sci.* **1996**, *36*, 334.
[14] a) N. Parker, M. J. Turk, E. Westrick, J. D. Lewis, P. S. Low, C. P. Leamon, *Anal. Biochem.* **2005**, *338*, 284; b) H. H. Cai, J. Pi, X. Lin, B. Li, A. Li, P. H. Yang, J. Cai, *Biosens. Bioelectron.* **2015**, *74*, 165; c) C. J. Weber, S. Muller, S. A. Safley, K. B. Gordon, P. Amancha, F. Villinger, V. M. Camp, M. Lipowska, J. Sharma, C. Muller, R. Schibli, P. S. Low, C. P. Leamon, R. K. Halkar, *Surgery* **2013**, *154*, 1385.
[15] A. Mauri, V. Consonni, M. Pavan, R. Todeschini, *MATCH Commun. Math. Comput. Chem.* **2006**, *56*, 237.
[16] F. R. Burden, D. A. Winkler, *QSAR Comb. Sci.* **2009**, *28*, 645.
[17] M. J. Polley, D. A. Winkler, F. R. Burden, *J. Med. Chem.* **2004**, *47*, 6230.
[18] a) F. R. Burden, D. A. Winkler, *J. Med. Chem.* **1999**, *42*, 3183; b) D. A. Winkler, F. R. Burden, *Mol. Simul.* **2000**, *24*, 243.
[19] D. L. J. Alexander, A. Tropsha, D. A. Winkler, *J. Chem. Inf. Mod.* **2015**, *55*, 1316.