Available online at www.sciencedirect.com
SCIENCE DIRECT®
Polymer 45 (2004) 7367–7379
polymer
www.elsevier.com/locate/polymer

Feature Article

Accurate predictions of cellular response using QSPR: a feasibility test of rational design of polymeric biomaterials

Vladyslav Kholodovych[a,*], Jack R. Smith[b], Doyle Knight[c], Sascha Abramson[b], Joachim Kohn[b], William J. Welsh[a,*]

[a]*Department of Pharmacology, University of Medicine and Dentistry of New Jersey (UMDNJ), Robert Wood Johnson Medical School and the Informatics Institute of UMDNJ, 675 Hoes Lane, Piscataway, NJ 08854, USA*
[b]*Department of Chemistry and Chemical Biology and the New Jersey Center for Biomaterials, Rutgers, The State University of New Jersey, New Brunswick, NJ 08854, USA*
[c]*Department of Mechanical and Aerospace Engineering, Rutgers, The State University of New Jersey, New Brunswick, NJ 08903, USA*

Received 28 July 2004; received in revised form 31 August 2004; accepted 1 September 2004
Available online 15 September 2004

Abstract

We present a Surrogate (semi-empirical) model for prediction of cellular response to the surfaces of biodegradable polymers that have been designed for tissue engineering applications. The predictions of our model, when tested against experimental results, show a high degree of accuracy that is sufficient for rational design of polymeric materials for biomedical applications. The model was determined by fitting experimental data for a series of 62 polyarylates to a small number of polymer structure-based ‘molecular descriptors’ using the technique of partial least squares (PLS) regression. While PLS is commonly applied in quantitative structure activity relationship (QSAR) analysis employed in the pharmaceutical industry, this study marks the first time the technique has been extended to the problem of biomaterials discovery/design. Quantitative predictions of cellular response to six polymers (untested prior to model building) concurred with experiment within 15.8% on average. This performance compares quite favorably with the overall variation in experimental values for the library of polyarylates. Examination of the PLS ‘loadings’ reveals those structure-based features most associated with variations in the polymer performance properties, thereby providing direct guidance to the synthetic chemist in biomaterials design.

Keywords: Biomaterials; Tyrosine degradable polyarylates; QSPR model

1. Introduction

We have generated and tested a semi-empirical model for the prediction of cellular response to polymer surfaces for a combinatorial library of tyrosine-derived biodegradable polymers (polyarylates). The procedure employed is based on quantitative structure activity relationship (QSAR) model protocols developed by researchers in the pharmaceutical industry for designing small-molecule compounds with optimized bioactivity. Briefly, the process consists of associating calculated molecular structure-based features known as ‘molecular descriptors’ with experimentally measured properties in a phenomenological manner. While there are many advantages in applying this strategy to biomaterials development, doing so poses several specific challenges. Both the advantages and the challenges will be treated in detail in this introduction.

1.1. Structure/property correlations in the development of polymeric materials

While correlations between chemical structure and macroscopic polymer properties have been explored since the 1930s (when the macromolecular structure of polymers was first recognized), this has typically not been done systematically. Traditionally, a given property is studied for a collection of structurally unrelated materials (e.g. a group of test materials consisting of polyethylene, Teflon, Dacron, etc). It is impossible to draw global conclusions from such

* Corresponding authors. Tel.: +1 732 235 3234; fax: +1 732 235 3475.
E-mail address: welshwj@umdnj.edu (W.J. Welsh).

0032-3861/$ - see front matter © 2004 Published by Elsevier Ltd.
doi:10.1016/j.polymer.2004.09.002

studies; observed structure/property relationships may well not be general and cannot be extrapolated to other sets of test materials. This type of sequential or ad hoc materials development has certainly been one of the major limitations to new materials discovery in biomedical materials science [1]. Here the inherent complexity of biological responses to materials and the lack of a unified theory to explain such interactions severely limit the scope of conclusions that can be drawn from any series of experiments involving only a handful of disparate candidate materials [1]. To date, most biomaterials development consists of studies involving various surface modifications of the same off-the-shelf materials (PEG, PP, PS, PLA, etc.) that have been chosen for their availability rather than their suitability for any particular application [1,2]. A more rewarding approach is to study sets of polymers that share common structural features and use the information gained from such analysis in the design of new materials tailored for specific applications [3–6].

In this context, we define a ‘library’ as a group of test materials in which all members share common properties and common structural features to some degree. These shared properties and features facilitate the construction of quantitative structure-performance relationship (QSPR) models that can predict properties of untested polymers and can guide the rational design of novel polymers within the same family. Such a strategy, successfully employed, yields optimized materials while saving an enormous amount of resources and labor in biomaterials development (particularly in the synthesis and testing stages).

1.2. The need for a combinatorial and systematic approach to biomaterials design

Combinatorial approaches, which lend themselves to QSAR protocols, have profoundly altered the process by which potential new drugs are identified [7]. These approaches, often collectively known as Combinatorial Chemistry (‘CombiChem’), involve the automated synthesis of tens of thousands to millions of compounds as randomly distributed moieties within a single reaction vessel followed by the identification of potentially active compounds in a selective bioassay. This basic methodology has also been successfully implemented in the design of catalytically active polymers [3].

A combinatorial approach is most effective when discernible correlations between the basic design variables (e.g. biomaterial chemistry and structure) and the performance of the product material are not available. This is clearly the case for materials being designed for tissue engineering scaffolds where the relationship between molecular structure and performance (i.e. cell–biomaterial interactions) is largely unknown [1]. Likewise, the use of combinatorial methods may be the most cost-effective and rapid approach whenever interdependent requirements and a large number of parameters result in unacceptably complex experimental designs. Again, biomaterials design with its many requirements and parameters is an appropriate area for the consideration of combinatorial methods—especially as an initial screening technique to identify promising polymer structures for further study.

1.3. The challenge of applying ‘CombiChem’ methods to biomaterials design

The major problem in applying the traditional CombiChem approaches to polymeric materials discovery stems from the fact that it would generate thousands or millions of polymers within a single test tube. The mixture created by this process would be a blend of polymers that could not be resolved into individual, homogeneous materials for testing. A more feasible combinatorial approach is to employ parallel synthesis of a system of monomers such that each resulting polymer is obtained in pure form in its own reaction vessel. This strategy permits the measurement of biomaterial properties of each homogeneous polymeric material in the library [8].

An additional complication arises with regard to testing or evaluating candidate biomaterials in a combinatorial fashion. While it is possible to devise simple high-throughput assays that allow one to test millions of compounds for some specific biological activity, there is no simple bioassay that can identify a suitable biomaterial within a group of polymers. This is simply because the relationship between materials structure and biological response is currently unknown. Therefore, combinatorial biomaterials design requires innovative fundamental research to identify the best predictors of biocompatibility. This, in turn, requires simple test assays that can be performed rapidly and inexpensively on a large number of test specimens. Several examples of these (e.g. protein adsorption and cellular response) have been under development by the Kohn group [9,10]. The availability of combinatorial libraries of candidate materials, together with high-throughput procedures for measuring biorelevant properties, offer a superior platform on which to develop important correlations between design and performance/function for biomedical applications.

1.4. Computational modeling of virtual polymer libraries

The use of computational methodologies has been sparse in the field of biomaterials, and computations involving molecular-level properties are virtually non-existent [11, 12]. Most of this is due, of course, to the inherent complexity involved in modeling bioresponse phenomena. However, considering the significant contributions computational methodologies have made to virtually all fields of research and development as well as the advent of relatively inexpensive, high-performance computing hardware and software, the introduction of these methodologies to biomaterials science is a timely endeavor.

Virtual polymer libraries constructed and evaluated entirely in silico provide an extraordinary tool to explore a wide range of new polymer compositions in a rapid and cost-effective manner. Briefly, virtual polymer libraries are large numbers of polymer structures that are created on a computer using various molecular modeling environments. Computational models evaluate members of virtual libraries using 'molecular descriptors' (i.e. quantifications of some aspect of molecular structure) to predict polymer properties such as biological response. Ultimately, this allows the rational selection of a smaller subset of these virtual polymers for actual synthesis and exploration. Although this approach is common and effective in drug discovery, as yet it has not been applied widely as a tool in biomaterials design.

1.5. QSAR models for materials property prediction/design

Computational techniques intended to build, screen, and mine virtual libraries of compounds have evolved rapidly in recent years as an efficient strategy for molecular discovery and optimization. In the field of drug discovery, QSAR models are constructed to correlate the experimentally determined properties of this subset of compounds with their calculated molecular descriptors. The QSAR models enable prediction of target properties for the full library of compounds. Then experimentation can be used to confirm predictions, particularly for those compounds found by the model as optimal for the desired application. One of the most successful approaches is an iterative process that cycles through prediction and experiment several times, with each cycle yielding improved agreement between prediction and experiment. Each cycle represents an enrichment process that culminates in compounds (e.g. biomaterials) with optimal performance properties. Experimental testing is reserved only for those compounds predicted to exhibit optimal performance.

Preliminary results showing the feasibility of the present approach have already been reported. Using Kohn's library of polyarylates, Reynolds [13] used similar techniques to predict several polymer properties without extensive (i.e. expensive) experimentation [10,14]. Reynolds first created a virtual polymer library, then employed similarity–diversity analysis and a genetic algorithm-driven QSPR model to design diverse and focused libraries of copolymers. He found that the same concepts of molecular similarity and diversity, so useful in the pharmaceutical industry to discover new drug candidates, to be highly amenable to synthetic polymers.

2. Methodology/background

2.1. The library of tyrosine-derived polyarylates

The Kohn laboratory has used combinatorial chemistry techniques to prepare a series of structurally related polyarylates derived from monomers consisting of a tyrosine-derived diphenol and a diacid (Fig. 1). In the combinatorial approach, 'AB' copolymers are synthesized from a set of *x* structural variations of 'A' and *y* structural variations of 'B'. The 'A' monomer template in the polyarylate library is the DTR diphenol shown in Figs. 1 and 2 while the 'B' monomer template is a dicarboxylic acid shown in Fig. 1. All possible combinations with the available 14 diphenols and 8 diacids yield 112 structurally distinct, but closely related, polymers. These can be prepared within one week in a custom designed parallel synthesis reactor [10,14].

2.2. Combinatorial libraries and phenomenological property prediction

A familiar example illustrates the utility of such a combinatorial polymer design and molecular descriptors in the phenomenological prediction of polymer properties. With regard to the measured glass transition temperature T_g for each of the 112 polyarylates in the library, it is possible to sort the polymers such that there is a gradual progression from low to high T_g (Fig. 3). However, the relationship between the sorting scheme and the chemical structures of the polymers is unclear and highly non-intuitive. By introducing an exceedingly simple polymer structural descriptor called the 'total flexibility index' (TFI), a useful exponential relationship between T_g and TFI emerges (Fig. 4). TFI is defined as the number of carbon and oxygen atoms in the variable portions of the backbone and pendent chain. In fact, it has been shown [13] that measurements of T_g performed on a representative subset of 17 polyarylates make it possible to predict T_g for each of the remaining 72 with a relatively high degree of accuracy. The use of structural descriptors thus makes experimentation beyond the representative subset unnecessary, yielding considerable savings of time and resources.

The present study was inspired by the search for molecular descriptors that can be correlated with the performance properties of polyarylates in biomedical applications (e.g. protein adsorption and cell response/proliferation). Though the relationship between polymer structure and biological response is likely far more complicated than that of physico-mechanical properties such as T_g, computational molecular modeling techniques make this approach possible. In addition, the polyarylate library is ideal for testing QSPR models of biological response based on polymer structure. Despite their structural similarity, the polyarylates show an impressive (and reproducible) variation in fibrinogen adsorption of over 360% and in cell response of $\sim$201% [9,10] that can be directly related to changes in structure (viz. Section 2.4).

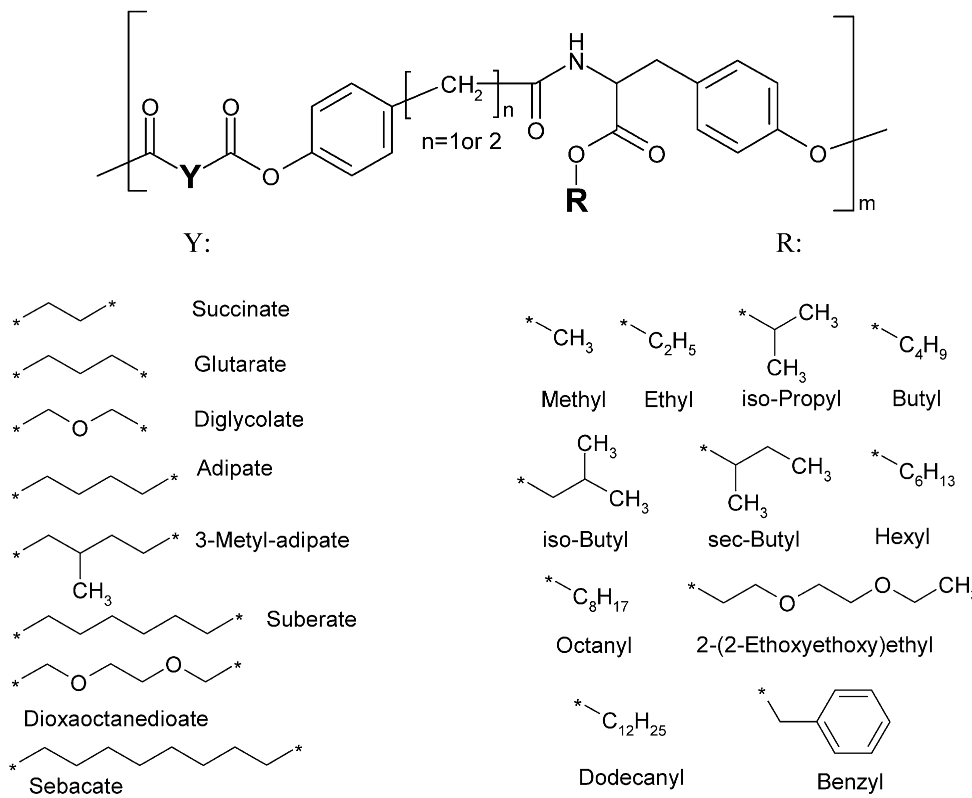


Fig. 1. Library of 112 polyarylates obtained from 14 tyrosine-derived diphenols and eight diacids. Polymers are strictly alternating copolymers consisting of a diacid (DA) and a diphenol (DP) component varied at Y and R, respectively. The number of methyl groups in the DP component is also variable.

2.3. Tyrosine-derived diphenols as monomers

The basic structure of desaminotyrosyl–tyrosine alkyl esters (DTR, Fig. 2) consists of a unit of 'desaminotyrosine' and a unit of L-tyrosine alkyl ester, linked together via a regular peptide bond. DTR is a derivative of naturally occurring tyrosine dipeptide with the important structural modification that the N terminus of the peptide was replaced by a hydrogen atom and the C terminus of the peptide is protected by an alkyl ester chain of variable length and structure. This particular design gives rise to a versatile diphenolic monomer that can be used in numerous other polymer systems [14–17].

2.4. Experimental data: cell response studies

Studies of the response of fetal rat lung fibroblasts

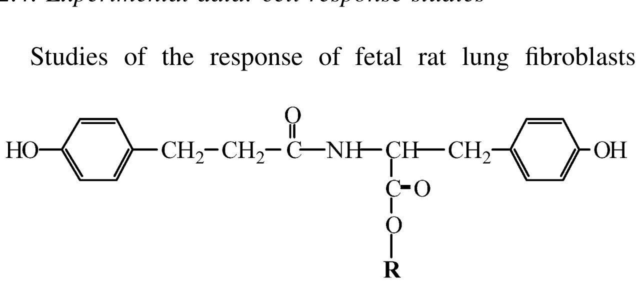


Fig. 2. Chemical structure of DTR diphenols. Note that the monomers form a homologous series, differing only in the length of their respective pendent chain (R). Commonly used pendent chains are ethyl (E), butyl (B), hexyl (H), octyl (O), and dodecyl (D) esters.

(FRLF) to polymeric substrates were performed as follows. Polyarylates were spin coated onto glass cover slips that were inserted into the bottom of wells in 24-well polystyrene plates [10]. Four samples of each polyarylate composition were created in order to provide adequate statistics. 10^4 cells/cm^2 in Dulbecco's modified Eagle's media supplemented with 10% heat-inactivated fetal bovine serum were seeded using the drop culture technique into each of the polyarylate-coated wells. A separate tissue culture polystyrene (TCPS) 24-well plate was used as a control. Wells were then incubated for 1 h at 30 °C. Subsequently, the wells were washed with PBS, replenished with media, and then incubated again at 37 °C. After seven days of incubation, the metabolic activity of remaining cells in each well was measured using a commercially available MTS colorimetric assay (Promega, Madison, WI). Cellular response to each polyarylate sample was then quantified as 'normalized metabolic activity' (NMA), which was its average measured metabolic activity given as a percentage of the average measured value for the TCPS wells. The average standard deviation over these measurements was 8.54% (NMA). The average percent standard deviation, where percent standard deviation is defined as the standard deviation expressed in terms of a percent of the mean, was 23.1%. Of the possible 112 polyarylates, 62 compositions were selected randomly for testing prior to modeling.

Following model building, six compositions that did not

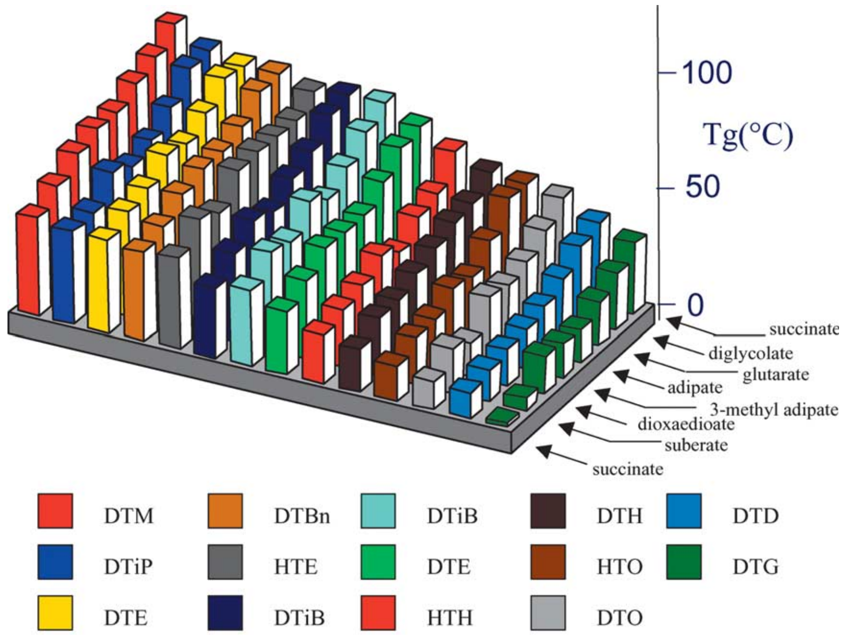


Fig. 3. Bar graph illustrating the complexity of the relationship between glass transition temperatures of 112 polymers in the polyarylate library and structure. While it is possible to sort the polymers such that there is a gradual progression from low to high T_g, the relationship between the sorting scheme and the chemical structures of the polymers is unclear and non-intuitive.

belong to the 62 polymer training set were tested for cellular response in separate experiments (same protocol). These additional six polymers were chosen to probe the accuracy of model predictions over their full range. Two of them were predicted to yield high values of FRLF NMA, two were predicted to yield middling values and two very low values. The test was 'blind' in the sense that these six samples had not been cell culture tested—prior to model building. Such a test provides a clear illustration of the predictive capability of QSPR models, thereby decreasing the experimental burden through in silico selection, rational design and evaluation of candidate polymers.

2.5. Molecular descriptors

Construction of the virtual polymer libraries and the QSPR models depends on the generation of molecular descriptors. It is possible to calculate thousands of these directly from the structure of any particular polymer using widely available molecular modeling tools [18]. Although the number of 'molecular descriptors' ranges in the hundreds and perhaps thousands, they can be divided into the several general categories (Table 1). For the present example, we deliberately selected only a relatively small number of molecular descriptors that satisfied two criteria:

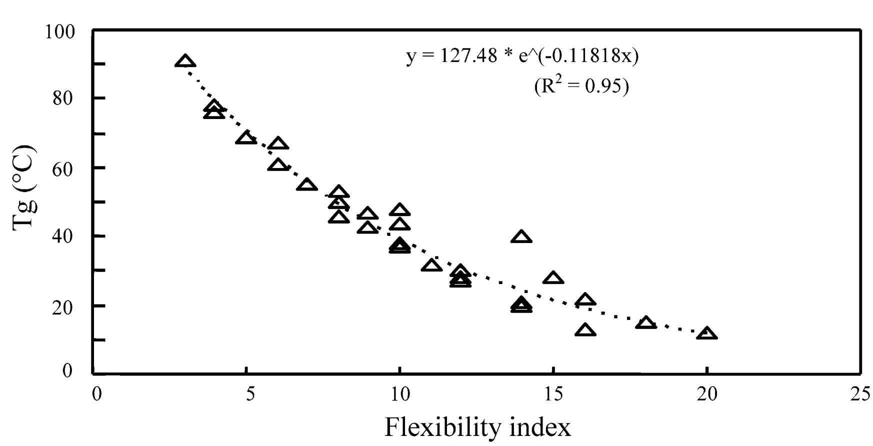


Fig. 4. Exponential correlation between the glass transition temperature of individual polymers contained within the library of polyarylates and the 'total flexibility index' (x), a descriptor that describes the chemical structure of the polymers. This illustrates the utility of the phenomenological approach. Using the equation provided [$y = 127.48e^{(-0.11818x)}$], the glass transition temperature (y) of every one of the thousands of theoretical polymer structures contained within the library can be predicted.

Table 1
Categories and examples of molecular descriptors

Category	Requirement	Example
Constitutional	Molecular composition	M_w, number of atoms/bonds, number of H-bond donors/acceptors
Topological	2-D structural formula	Kier–Hall indices, extent of branching
Geometrical	3-D structure of molecule	Molecular volume, solvent accessible surface area, polar and non-polar surface area
Electrostatic	Charge distribution	Atomic partial charges, electronegativities
Quantum mechanical	Electronic structure	HOMO–LUMO energies, band gap, dipole moment

(i) they could be calculated easily (i.e. no quantum mechanical descriptors); and (ii) they were physically intuitive and directly related to gross structural and/or molecular bio-physical properties. The rationale for this decision was that the present study was not meant to be exhaustive but, rather, as a demonstration of the general approach.

This selection process yielded the set of 15 descriptors shown in Table 2. The unsaturation index refers to the counting of unsaturated (double, triple, aromatic) bonds; the hydrophilic factor is a measure of the number of hydrophilic functional groups; and the aromatic ratio is the ratio between the number of aromatic bonds and total bonds in the H-depleted molecule. The molar refractivity is a classical measure of molecular volume and dispersion (London, van der Waals) forces that govern non-polar intermolecular interactions; and log P, defined as the logarithm of the octanol–water partition coefficient, provides a measure of the hydrophobicity of the molecule.

Values of these molecular descriptors were calculated for all of the 112 polyarylates, where each polymer was represented as a defect-free, linear chain of three repeat units in length. Subsequently, the descriptors were used to generate the QSPR models.

2.6. QSPR models

The complete list of polymers whose experimental cell response data was used to build the QSPR models appears in Table 3. Generally speaking, QSPR model building involves the application of statistical regression methods that quantify the relationship between changes in structure and changes in the target property.

At this stage, our primary objectives were three-fold: (i) to build statistically robust QSPR models in which the cell growth (as measured by NMA) was correlated with theoretically calculated molecular descriptors rather than with experimental data (T_g, air–water contact angle, etc.); (ii) to identify those particular molecular descriptors that are most important in explaining the observed variations in NMA among these polyarylates; and (iii) to apply the QSPR models for predicting the NMA for the remaining 50 polyarylates (i.e. those not used for model building) and for selecting individual polyarylates from this subset that are predicted to exhibit high NMA values.

2.7. Partial least-squares (PLS) regression and principal component analysis

The present QSPR models attempted to correlate the target property (FRLF NMA) with molecular descriptors using partial least-squares (PLS) regression [19–22]. Here we provide only a summary of these methods as the details appear elsewhere [23–28].

PLS is a popular and powerful computational method that expresses a dependent variable (target property) in terms of linear combinations of the independent variables (molecular descriptors) commonly known as principal components (PCs) [29,30]. While the descriptors themselves may be interdependent (covariant), the PCs so generated are independent (orthogonal). The calculation is briefly described as follows. First, a (square) matrix is

Table 2
Summary of 11 molecular descriptors employed to build initial QSPR models

Descriptor type	Name
Functional groups	Number of primary carbons (sp3)
	Number of secondary carbons (sp3)
	Number of tertiary carbons (sp3)[a]
	Number of substituted aromatic carbons (sp2)
	Number of branches in pendent chain (aliphatic)[a]
Empirical descriptors	Unsaturation index
	Hydrophilic factor
	Aromatic ratio
Molecular properties	Molar refractivity (MR)[a]
	Polar surface area (PSA)[a]
	Logarithm of the octanol–water partition coefficient (log P)[a]

[a] Five descriptors that contributed to the best model.

Table 3
List of 62 polyarylates used to build the QSPR model, together with values of biological activity (FRLF NMA) and % standard deviation across four independent measurements

No.	Pend	Diacid	FRLF NMA (%TCPS)	STDEV (%TCPS)
1	DTB	Adipate	32.29	30.02
2	DTB	Diglycolate	75.89	2.51
3	DTB	Dioxaoctanedioate	82.81	9.87
4	DTB	Methyl adipate	35.12	4.29
5	DTB	Sebacate	58.05	4.32
6	DTB	Suberate	76.34	9.63
7	DTB	Succinate	75.82	9.63
8	DTBn	Adipate	52.48	18.18
9	DTBn	Diglycolate	73.77	7.43
10	DTBn	Dioxaoctanedioate	69.93	10.24
11	DTBn	Glutarate	71.49	4.75
12	DTBn	Methyl adipate	32.01	23.05
13	DTBn	Sebacate	66.53	1.90
14	DTBn	Suberate	67.24	10.79
15	DTBn	Succinate	77.77	8.87
16	DTD	Adipate	2.00	1.98
17	DTD	Diglycolate	67.65	7.63
18	DTD	Dioxaoctanedioate	66.31	4.64
19	DTD	Glutarate	18.83	14.02
20	DTD	Methyl adipate	20.83	5.44
21	DTD	Sebacate	7.92	6.97
22	DTD	Suberate	31.48	3.34
23	DTE	Adipate	75.66	8.32
24	DTE	Diglycolate	82.00	1.03
25	DTE	Dioxaoctanedioate	77.69	12.45
26	DTE	Glutarate	78.47	7.34
27	DTE	Methyl adipate	38.55	15.15
28	DTE	Sebacate	68.40	3.32
29	DTE	Suberate	69.51	10.03
30	DTE	Succinate	97.59	5.20
31	DTH	Adipate	16.48	22.77
32	DTH	Diglycolate	69.50	6.38
33	DTH	Dioxaoctanedioate	59.09	9.06
34	DTH	Glutarate	52.80	3.88
35	DTH	Methyl adipate	25.48	7.05
36	DTH	Sebacate	50.62	8.73
37	DTH	Suberate	63.64	5.89
38	DTH	Succinate	30.18	13.46
39	DTiP	Adipate	62.36	14.05
40	DTiP	Diglycolate	81.44	11.47
41	DTiP	Dioxaoctanedioate	70.89	6.22
42	DTiP	Glutarate	79.79	11.94
43	DTiP	Methyl adipate	85.01	13.04
44	DTiP	Sebacate	78.30	6.09
45	DTiP	Suberate	70.44	5.63
46	DTiP	Succinate	77.07	14.67
47	DTM	Adipate	86.22	16.24
48	DTM	Diglycolate	88.33	1.95
49	DTM	Dioxaoctanedioate	84.61	0.93
50	DTM	Glutarate	94.60	9.03
51	DTM	Methyl adipate	78.00	8.17
52	DTM	Sebacate	87.85	5.32
53	DTM	Suberate	80.71	14.16
54	DTM	Succinate	114.66	2.72
55	DTO	Adipate	4.80	4.94
56	DTO	Diglycolate	66.69	0.69
57	DTO	Dioxaoctanedioate	71.67	4.32
58	DTO	Glutarate	43.18	4.13
59	DTO	Methyl adipate	40.88	7.50
60	DTO	Sebacate	17.58	6.71
61	DTO	Suberate	47.38	5.52
62	DTO	Succinate	25.57	14.15

created to represent the relationship between the descriptors and the experimental NMA data. Then, the eigenvalues and eigenvectors of that matrix are determined. The eigenvector associated with the largest eigenvalue has the same direction as the first principal component (PC_1) and the eigenvector associated with the second largest eigenvalue determines the direction of the second principal component. Similarly, the second principal component (PC_2) has the same direction as the eigenvector associated with the second largest eigenvalue, and so on. PC_1 explains the greatest amount of variance in the target property, PC_2 is next best, etc. The PLS equation, then, assumes the following form for the case of '*n*' molecular descriptors:

$$\text{PLS}: \text{ target property} = a_0 + a_1(\text{PC}_1) + a_2(\text{PC}_2) + a_3(\text{PC}_3) + \ldots + a_n(\text{PC}_n) \quad (1)$$

Each PC can be decomposed into its 'loadings' which reveals the individual contributions from the original set of molecular descriptors. The loading of an individual molecular descriptor indicates how much this variable participates in defining the PC (the squares of the loadings indicate their percentage in the PC). This information is extremely valuable since the leading PCs (esp. PC_1 and PC_2) embody those descriptors that correlate most strongly with the target property and, thus, provide clues to achieving optimal polymer design. Knowledge of these key molecular descriptors often provides insights into the fundamental mechanism of action, and, indeed, may suggest new design strategies and synthetic targets beyond the original library.

2.8. Validation

The final QSPR model was first validated internally using 'leave one out' (LOO) cross-validation. The LOO procedure provides a quantitative assessment of the capability of the model to provide predictions for polymers outside the training set. In LOO, *n* different models are constructed (where *n* is the number of training set elements). In this case, $n=62$ for the 62 polyarylates represented in Table 3. In each of the QSPR models, a different training set member is 'left out'; i.e. each model has a training set of exactly $n-1$ elements. Each model is then used to predict the biological response of the training set element that is 'left out'. The success of the models is evaluated using the cross-validated correlation coefficient (q^2) defined as:

$$q^2 = \frac{\text{SSD} - \text{PRESS}}{\text{SSD}} \quad (2)$$

where SSD is the sum-of-squared deviations of the experimentally measured NMA values of the polymers around the mean value, and PRESS represents the sum of the squared differences between the predicted and actual target property values for every compound.

The model is generally considered internally predictive if $q^2 > 0.5$ (where q^2 can vary from $-\infty$ to 1.0). The optimal number of components corresponds to that which yields either the smallest rms error or the largest q^2 value. A final PLS analysis was performed inclusive of all compounds in the data set, yielding a conventional (correlation coefficient) r^2 value which provides a measure of the internal consistency (goodness of fit) of the model.

Following this internal validation, the model was evaluated externally using a test set of polymers with experimentally measured properties that were not used to build the QSPR model. Following this, the QSPR model is considered suitable for making accurate predictions outside of the training set of polymers.

3. Results and discussion

The optimum model achieves the highest accuracy using the minimum number of descriptors. Therefore, the procedure described above was attempted with all possible combinations of the descriptors in Table 2. It was determined that the presence of the five descriptors denoted by an 'a' in that table was necessary for any model to meet the internal validation accuracy measure described in the previous section. However, it was also discovered that adding the remaining descriptors from Table 2 did not improve the accuracy in any of the models. Therefore, only the five best descriptors [i.e. number of tertiary carbons (sp3); number of branches in pendent chain (aliphatic); molar refractivity (MR); polar surface area (PSA); and logarithm of the octanol–water partition coefficient (log *P*)] were included.

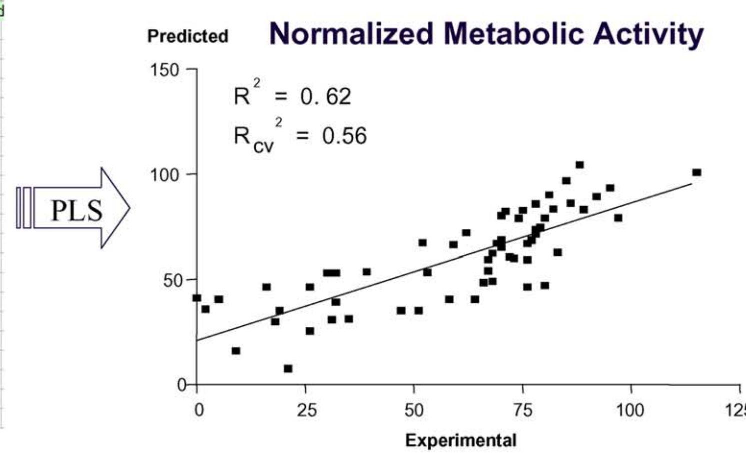
The results are summarized by plotting the QSPR-predicted versus experimentally observed NMA values for the set of 62 polyarylates in the training set (Fig. 5B). Using only five PCs (PC = 5), the resulting model was statistically significant ($r^2=0.62$; $r=0.79$). The cross-validated r^2 value ($q^2=0.56$) satisfied the generally accepted condition that $q^2 \geq 0.50$ for an internally consistent and predictive regression model [21,22,31,32].

The loadings of the initial PCs (viz. PC1, PC2, PC3) give the relative importance of each of the descriptors to the model. The loadings for PC1 were of greatest interest to us, and the five descriptors that make the largest contribution to PC1–PC5 are shown in Fig. 6. We note that this list includes each of the three molecular property descriptors listed in Table 2. Such molecular property descriptors are derived computationally and, therefore, would not be recognized merely from visual inspection of the polymers' structure. This result illustrates the importance of including theoretically calculated descriptors that encode information about both the structures and bio-physical properties of the polymers.

Interestingly, a previous study by Welsh and co-workers [33] showed that similar molecular property descriptors

1	Name	nHAcc	Ui	Hy	ARR	MR	PSA	MLOGP	Normalized
2	AA_DTB.mol	11	4.17	0.14	0.273	156.183	154.53	3.312	32
3	AA_DTBn.mol	11	4.585	0.107	0.375	166.975	154.53	3.919	52
4	AA_DTD.mol	11	4.17	0.069	0.235	188.39	154.53	4.566	2
5	AA_DTE.mol	11	4.17	0.164	0.286	146.674	154.53	2.934	75
6	AA_DTH.mol	11	4.17	0.118	0.261	165.385	154.53	3.68	16
7	AA_DTiP.mol	11	4.17	0.152	0.279	151.092	154.53	3.124	62
8	AA_DTM.mol	11	4.17	0.177	0.293	141.484	154.53	3.147	86
9	AA_DTO.mol	11	4.17	0.097	0.25	174.587	154.53	4.041	5
10	DDA_DTB.mol	13	4.17	0.162	0.261	159.67	172.99	1.884	78
11	DDA_DTBn.mo	13	4.585	0.129	0.36	170.462	172.99	2.491	74
12	DDA_DTD.mol	13	4.17	0.091	0.226	191.877	172.99	3.138	67
13	DDA_DTE.mol	13	4.17	0.186	0.273	150.16	172.99	1.506	82
14	DDA_DTH.mol	13	4.17	0.14	0.25	168.872	172.99	2.252	70
15	DDA_DTiP.mo	13	4.17	0.174	0.267	154.579	172.99	1.696	81
16	DDA_DTM.mol	13	4.17	0.199	0.279	144.97	172.99	1.719	88
17	DDA_DTO.mol	13	4.17	0.119	0.24	178.074	172.99	2.613	67
18	DGA_DTB.mol	12	4.17	0.175	0.279	148.626	163.76	2.217	83
19	DGA_DTBn.mo	12	4.585	0.14	0.383	159.418	163.76	2.838	70
20	DGA_DTD.mol	12	4.17	0.099	0.24	180.833	163.76	3.501	66
21	DGA_DTE.mol	12	4.17	0.202	0.293	139.117	163.76	1.829	78
22	DGA_DTH.mol	12	4.17	0.151	0.267	157.828	163.76	2.595	59

A



B

Fig. 5. Summary of results from QSPR model: (A) 'snapshot' of the list of 62 polyarylates used for model building. The polymers are designated on the left by code name, while values of the descriptors are depicted in successive columns. AA—adipate, DDA—dioxaoctanedioate, DGA—diglycolate etc. The whole list of polyarylate abbreviations can be found elsewhere [13]. (B) Plot of the QSPR-predicted versus experimentally observed NMA values for the set of 62 polyarylates in the training set.

figured prominently in QSAR models constructed to explain the observed inhibitor–receptor binding affinities for a series of HIV-1 protease inhibitors. The descriptors in this previous study were closely related to PSA, MR, and bioavailability (log *P*). We believe that the similarity of descriptors found in the present study is no coincidence. Indeed, this result points to the fundamental importance of such molecular property-based descriptors in constructing a rational basis for explaining the in vitro properties and, perhaps, the in vivo properties of molecules intended for biomedical applications.

Predictions of FRLF NMA were generated for all 50

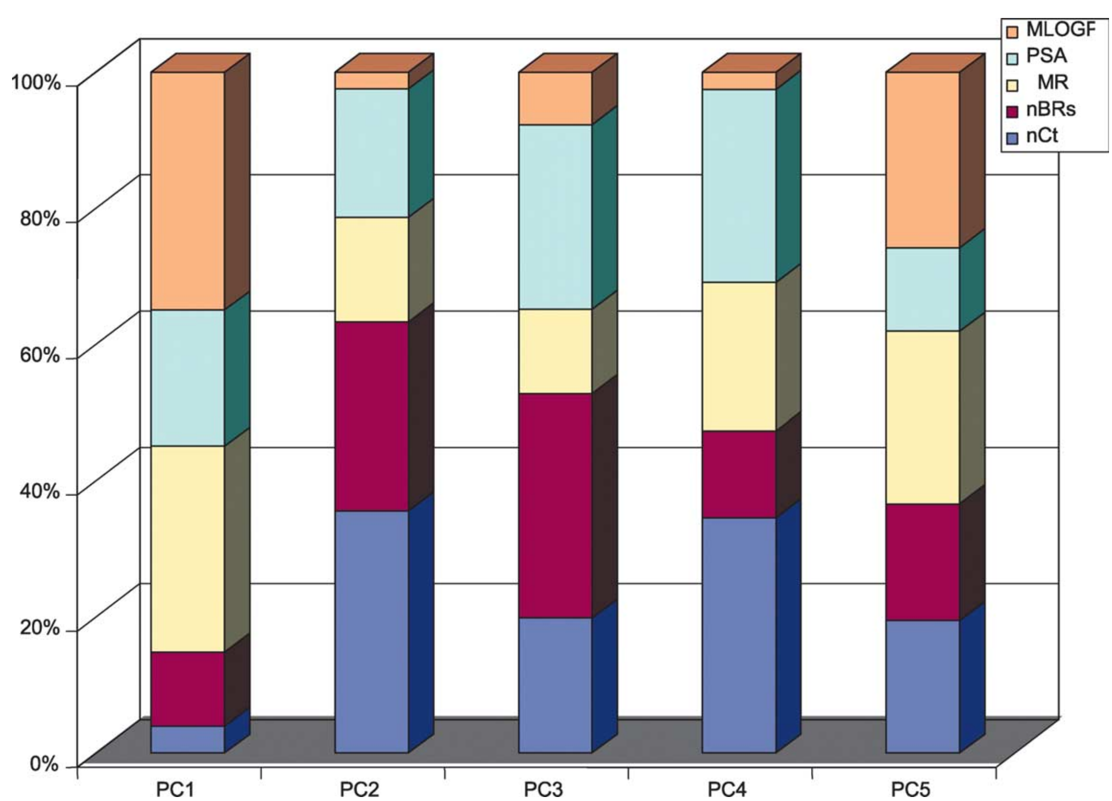


Fig. 6. Loadings for five principal components (PC1-5) extracted from QSPR model based on NMA values for 62 polyarylates. The five molecular descriptors making the largest contribution to PCs are listed to the right of the figure. It is noteworthy that three of these five descriptors are of the 'molecular property' descriptor type (see Table 2). MLOGP-Moriguchi octanol–water partition coefficient log *P*. PSA-fragment based polar surface area. MR-Ghose-Crippen molar refractivity. *n*BRs-number of branches in pendent chain (aliphatic). *n*Ct-number of tertiary C (sp3).

polymers in the library for which experimental testing was not performed prior to model building. Six of these polymers were chosen for experimental verification of these predictions. The six appear in Fig. 7 along with both the predicted values of FRLF NMA generated by the model, their monomer structures and the experimentally measured NMA values. Again, this particular set was chosen because it provides two examples of polymers predicted to be low, middle and high performers.

A graphical comparison of the predictions versus experimentally measured results appears in Fig. 8. The error bars indicate the average percent standard deviation (coefficient of variation) of the measurement for the 62 measurements included in the training set. The 45° line in the figure represents the case of perfect agreement between experiment and model. Perfect agreement, of course, is impossible given the significance of the experimental error. However, we do note that all of the predictions fall within experimental error of the measured results except for the lowest measured case (HTH Methyladipate). The average percent error of prediction is 15.8%. This is considerably less than the experimental average percent standard deviation (23.1%) and is nearly an order of magnitude less than the total variation in the set of six test polymers (149%). Then the model has certainly discriminated between the highest and lowest performers to well within the level of experimental error and, indeed, has done so prior to any experimentation. While the model is somewhat less adept at distinguishing between mid and lower performing polymers unambiguously, in most applications identification of optimal performers is sufficient. For example, these predictions could have been used to reduce the set of polymers for testing from the original six to the two top performers if high FRLF NMA was the target property. This corresponds to a decrease in the experimental burden of approximately 67% which, in the case with larger sample volume, would represent a substantial savings in both time and resources.

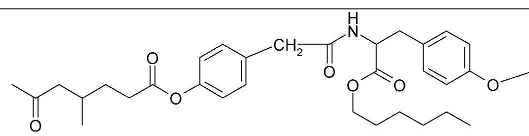
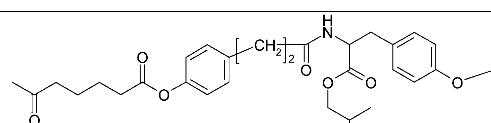
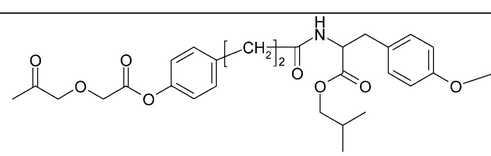
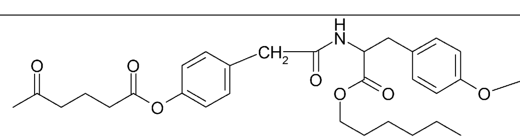
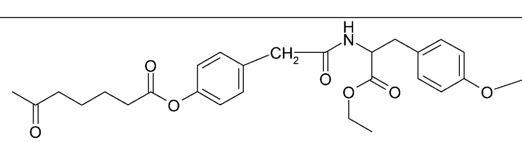
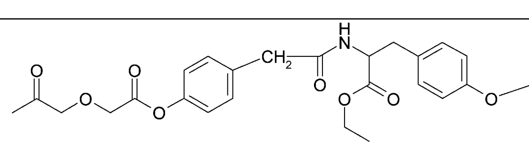
Name	Structure	Predicted NMA (% TCPS)	Measured NMA (% TCPS)
HTH Methyladipate		33.67	63.7± 12.1
DTiB adipate		40.91	41.4 ± 7.9
DTiB diglycolate		54.96	62.6 ± 11.9
HTH glutarate		59.45	53.2 ± 10.1
HTE adipate		69.68	67.1 ± 12.7
HTE Diglycolate		82.62	101.5 ± 19.3

Fig. 7. The structures, QSPR model FRLF NMA predictions and measured values for the six polyarylates in the 'blind' test set.

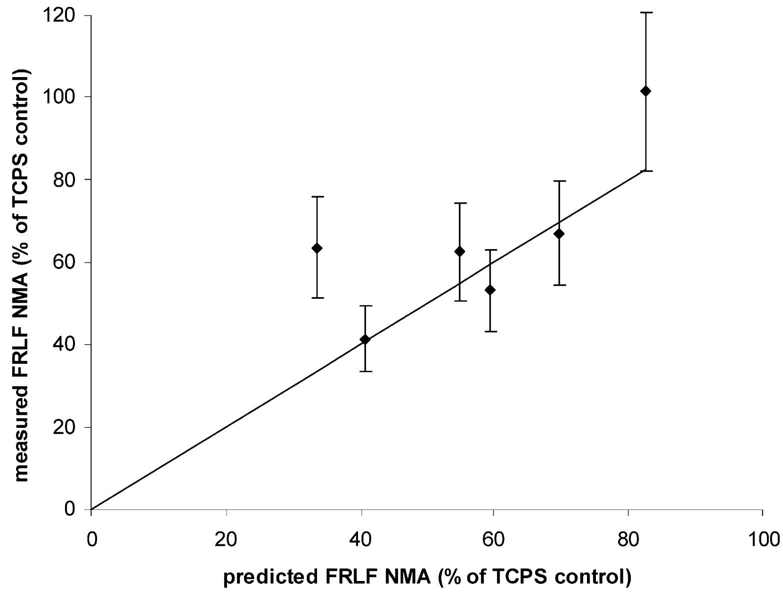


Fig. 8. QSPR predictions versus experimental results for the 'blind' test set. Error bars represent the standard deviation of the experimental measurements.

More generally, we note that the average error of prediction is even smaller when compared to the overall experimental variation (from min 2 to max 114 in terms of normalized cellular metabolic activity) in all 68 samples. The model can be used to distinguish the highest performers from the lowest. In other words, the predictions generated by this model can easily be used to test the performance of candidate structures entirely in silico, or prior to experimentation. Taken together, then, the results from our computational models give confidence in the validity of the proposed strategy. The predictions generated by the QSPR model might well be used in an iterative procedure to design a polyarylate surface that maximizes FRLF NMA. In this process, candidate structures would be proposed, evaluated using QSPR models trained on a subset of experimental data, then the structures could be refined (perhaps using any one of a number of design algorithms [34–36]) and then retested. In fact, the development of such an algorithm for iterative structural refinement and design of polymeric biomaterials is currently underway in our laboratory.

4. Conclusions and future work

We have shown that computational strategies that have demonstrated success in the rational design of new therapeutics in pharmaceutical discovery can be employed to offer guidance and direction in the design, selection, and optimization of novel biorelevant materials. This procedure works quite well for the example case involving a combinatorial library of polymers and the 'target' property of response of fibroblast cells. However, the methodology employed here is sufficiently general to treat many other problems in biological response. These include protein adsorption, immunoresponse (in the form of macrophage genotypic expression) as well as the proliferation and growth of many different cell types in the presence of these polymers. In principle, a surrogate model can be built for any measure of cell response. However, experimental data for FRLF cell response to the polyarylate library is only available for the MTS colorimetric assay.

Since computed polymer descriptors are less expensive to obtain than in vitro or in vivo measurements, the use of a computational modeling approach can significantly reduce the costs and labor associated with identifying high-performance biomaterials for specific applications. Further, with the application of other methods of design optimization already well developed in other fields of research, such a procedure may lead to great advances in biomaterials development.

Acknowledgements

This work was supported by seed funds provided by Rutgers University, by NIH grant R01 EB00286, and by 'RESBIO-The National Resource for Polymeric Biomaterials' funded under NIH grant P41 EB001046-01A1. Financial support was also provided by NIH Grant HL-60416 and the New Jersey Center for Biomaterials.

References

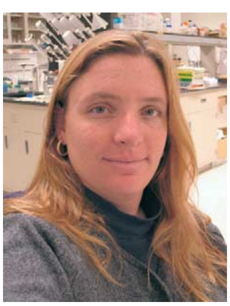
[1] Castner DG, Ratner BD. Surf Sci 2002;500:28–50.
[2] Smith J, Seyda A, Weber N, Knight D, Abramson S, Kohn J. Macromol Rapid Commun 2004;25:127–40.
[3] Menger FM, Eliseev AV, Migulin VA. J Org Chem 1995;60:6666.
[4] Meredith JC, Sormana JL, Keselowsky BG, Garcia A, Garcha A, Tona A, Karim A, Amis EJ. J Biomed Mat Res A 2003;66A(3): 483–90.

[5] Ertel SI, Kohn J. J Biomed Mater Res 1994;28:919–30.
[6] James K, Levene H, Parsons J, Kohn J. Biomaterials 1999;20(22/23): 2203–12.
[7] Lowe G. JCS Rev 1995;309–17.
[8] Brocchini S, James K, Tangpasuthadol V, Penharkar S, Tong X, Kohn J. Preliminary studies exploring a combinatorial approach toward the development of new biomaterials, New Orleans, LA; 1997.
[9] Weber N, Bolikal D, Bourke SL, Kohn J. J Biomed Mater Res 2004; 68A:496–503.
[10] Brocchini J, James K, Tangpasuthadol V, Kohn J. J Biomed Mater Res 1998;42:66–75.
[11] Perez-Luna V, Horbett T, Ratner B. J Biomed Mater Res 1994;28(10): 1111–26.
[12] Chilkoti A, Schmierer A, Perez-Luna V, Ratner B. Anal Chem 1995; 67(17):2883–91.
[13] Reynolds C. J Comb Chem 1999;1(4):297–306.
[14] Brocchini S, James K, Tangpasuthadol V. J Am Chem Soc 1997;119: 4553–4.
[15] Yu C, Kohn J. Biomaterials 1999;20(3):253–64.
[16] Yu C, Mielewczyck S, Breslauer K, Kohn J. Biomaterials 1999;20(3): 265–72.
[17] Ryan P, Foty R, Kohn J, Steinberg M. Proc Natl Acad Sci 2001;98(8): 4323–7.
[18] Todeschini R, Consonni V, Mauri A, Pavan M. Dragon Web version 3.0, Milano, Italy; 2003.
[19] Draper N, Smith H. Applied regression analysis, 2nd ed. New York: Wiley; 1981.
[20] Anderson T. An introduction to multivariate statistical analysis, 3rd ed. New York: Wiley; 2003.
[21] Wold S, Albano C, Dunn III WJ, Edlund U, Esbensen K, Geladi P, Hellberg S, Johansson E, Lindberg W, Sjstrm M. Multivariate data analysis in chemistry. In: Kowalski BR, editor. Chemometrics: mathematics and statistics in chemistry. Dordrecht, The Netherlands: Reidel; 1984. p. 17–95.
[22] Wold S. PLS for multivariate linear modelling. In: deWaterbeemd Hv, editor. Methods and principles in medicinal chemistry. Weinheim, Germany: Verlag; 1995.
[23] Tong W, Perkins R, Xing L, Welsh WJ, Sheehan DM. Endocrinology 1997;138:4022–5.
[24] Tong W, Perkins R, Chen Y, Welsh WJ, Lowis DR, Goddette DW, Heritage TW, Sheehan DM. J Chem Inf Comp Sci 1998;38:669–77.
[25] Jayatilleke P, Nair A, Zauhar R, Welsh WJ. J Med Chem 2000;43: 4446.
[26] Puri S, Chickos J, Welsh WJ. J Chem Inf Comp Sci 2002;42:109–16.
[27] Puri S, Chickos J, Welsh WJ. J Chem Inf Comp Sci 2002;42:209–14.
[28] Yu SJ, Keenan SM, Tong W, Welsh WJ. Chem Res Toxicol 2002; 15(10):1229–34.
[29] Joliffe I. Principal component analysis. New York: Springer Verlag; 1986.
[30] Malinowski EH, Howery DG. Factor analysis in chemistry. New York: Wiley; 1980.
[31] Harnett D, Murphy J. Introductory statistical analysis. Reading, MA: Addison-Wesley; 1980.
[32] Wold S, Hellberg S, Dunn WJ. Acta Pharmocol Toxicol 1983;52(2): 158–89.
[33] Nair A, Jayatilleke P, Wang X, Miertus S, Welsh W. J Med Chem 2002;45(4):973–83.
[34] Knight D. Data driven design optimization methodology—a dynamic data driven application system. New York: Springer Verlag; 2003 p. 329–336.
[35] Knight D, Elliott G, Jaluria Y, Langrana N, Rasheed K. AIAA Paper No. 2002–5636; 2002.
[36] Lefantzi S, Knight D. AIAA Paper No. 2001–3294; 2001.

Vladyslav Kholodovych. Vladyslav Kholodovych obtained his Master of Science degree in Chemistry from the Kiev State University (Ukraine) in 1993. Working under the supervision of Profs. Vladimir Kibirev and Vladimir Khilya, he developed new fluorescent substrates for blood coagulation enzymes and optimized peptide synthesis. He earned a PhD degree in Chemistry from the Institute of Bioorganic Chemistry and Petrochemistry (Kiev, Ukraine) working on physiologically active compounds under the direction of Prof. Alexander Luik. His research comprised the computer modeling, organic synthesis and biological validation of the mechanism of action of the natural immunomodulator tuftsin and its analogues. For this work Vladyslav was granted the Stipendiary Award of the Ukrainian National Academy of Sciences in 1999 and 2000 and was named leader of the Computational Group for the Institute. In 2001 he joined Prof. William Welsh as a research associate, first at the University of Missouri-St. Louis and more recently at the University of Medicine and Dentistry of New Jersey (UMDNJ), Robert Wood Johnson Medical School. He has been active in various research projects involving rational drug design, pattern recognition, computational biology, nuclear hormone receptors and biodegradable materials. He was awarded the Young Investigator Award (2002) and the Research Achievement Award (2002 and 2003) at UMDNJ. He is a member of the Informatics Institute of UMDNJ and the New Jersey Center for Biomaterials.

Jack Smith. Jack R. Smith completed his undergraduate thesis in physics in the group of Stuart Field (University of Michigan, USA) by modeling the dynamics of chaotic systems. Subsequently, he worked as an intern developing data acquisition algorithms for the Babar project in the group of Michael Levi (Lawrence Berkeley National Laboratory, USA). His PhD work was performed under the supervision of Dawn Bonnell and Roger French (University of Pennsylvania, USA) and focused on using experiment and modeling to establish correlations between bonding, atomic-scale structure and optical properties in thin, amorphous films of CrOCN. These films are used in DUV phase-shifting lithography for the printing of DRAM chips. For his thesis work, Jack was awarded the University's Stein Prize. In 2001, he was awarded the Chateaubriand Fellowship for postdoctoral study in France. In the group of Joel Chevrier at the European Synchrotron Radiation Facility (Grenoble), he studied the oxidation of AlPdMn quasicrystalline surfaces using primarily scanning probe techniques. Currently, he is a research associate at the New Jersey Center for Biomaterials working closely with Drs Knight, Welsh and Kohn on the computational modeling of biological response to polymeric materials/surfaces.

Doyle D. Knight. Doyle Knight received his PhD in Aeronautics from the California Institute of Technology in 1974. Following two years service in the United States Air Force as an Aeronautical Engineer, and one year Postdoctoral Fellowship in Applied Mathematics at the California Institute of Technology, he joined the faculty of the Department of Mechanical and Aerospace Engineering at Rutgers—The State University of New Jersey. His research interests include gas dynamics, design optimization and biomaterials. His research in gas dynamics has extended to a broad range of topics including shock wave turbulent boundary layer interaction, incipient separation on pitching airfoils, turbulence model development, high speed inlet unstart and effects of unsteady energy deposition in supersonic flows. His research activity in design optimization focuses on the application of computational fluid dynamics to the automated optimal design of high speed air vehicles (especially, missiles). His research interests in biomaterials have led to the development of surrogate models for prediction of cellular proliferation in the presence of biomaterials. His research interests have led to his significant activity in parallel processing on PC clusters under Linux including the development of Linux clusters in his research laboratory.



Sascha Abramson. Dr Sascha Abramson completed her undergraduate degree in English from the University of Pennsylvania in 1996. In 2002 she received her doctorate in Biomedical Engineering from Rutgers University for her work with tyrosine-derived polycarbonates. Dr Abramson extensively characterized the thermal changes that occurred during the degradation of the tyrosine-derived polycarbonates in addition to evaluating the in vivo osteoblastic response and in vitro bone response to the polymers. Dr Abramson joined the New jersey Center for Biomaterials and the Rutgers Center for Computational Design for a joint postdoctoral position in which she developed biological rapid screening protocols for large numbers of degradable polymers. Sascha currently works in the biomaterials research and development of Celgene Cellular Therapeutics.

Joachim Kohn. Professor Kohn received his PhD in Chemistry from the Weizmann Institute of Science. He received a fellowship for postgraduate studies on biomaterials and drug delivery at the Children's Hospital in Boston (jointly with a postdoctoral appointment at the Massachusetts Institute of Technology). He is a Board of Governors Professor of Chemistry at Rutgers—The State University of New Jersey and an Adjunct Associate Professor of Orthopaedics at the New Jersey Medical School. He serves as the Director of the New Jersey Center for Biomaterials, a three-university consortium providing a comprehensive research and development program for academic and industrial scientists. He is principal investigator of an NIH-funded biomedical technology resource center for polymeric biomaterials. He is also principal investigator of the DoD-funded Center for Military Biomaterials Research. He is a Fellow of both the American Institute for Medical and Biological Engineering and the International Union of Societies for Biomaterials Science and Engineering. Professor Kohn's research interests focus on the design of new biomaterials (with a special emphasis on the development of combinatorial approaches for the design of new polymers), the elucidation of cell–polymer interactions, and the development of polymeric scaffolds for controlled sustained release and tissue engineering. He has published over 200 scientific manuscripts and review and holds more than 34 patents. Professor Kohn is currently on the editorial boards of six leading journals in the field of biomaterials science.

William J. Welsh. Dr William J. (Bill) Welsh holds the Norman H. Edelman Endowed Professorship in Bioinformatics and Computer-Aided Molecular Design in the Department of Pharmacology at the Robert Wood Johnson Medical School (RWJMS) in Piscataway NJ, University of Medicine and Dentistry of New Jersey (UMDNJ). Concurrently, he serves as Director of the UMDNJ Informatics Institute (http://informatics.umdnj.edu) and Director of the UMDNJ Graduate Program in Bioinformatics. Dr Welsh earned a B.S. degree (magna cum laude) in Chemistry from St. Joseph's University (Phila., PA) and a Ph.D. degree in Theoretical Physical Chemistry from the University of Pennsylvania (Phila., PA). He pursued postgraduate studies in the laboratory of Prof. James E. Mark, Distinguished Professor of Polymer Science, at the University of Cincinnati (Cinti., OH). In 1985, Dr Welsh joined the University of Missouri (St. Louis) as an Associate Professor of Chemistry and rose through the ranks to Distinguished Professor in 1998. In 2001, Dr Welsh joined UMDNJ-Robert Wood Johnson Medical School to assume his present role. Dr Welsh's research interests cover a broad spectrum of applications in computer-aided molecular modeling and design including drug discovery, computational structural biology, and the design of high-performance polymers and biomaterials. He has published over 300 scientific articles and holds 20 patents in the drug discovery field. He is the recipient of numerous awards and honors, and he serves on the advisory boards of several scientific journals.