Journal of Computer-Aided Molecular Design **17:** 103–118, 2003.

QSAR modeling of β-lactam binding to human serum proteins

L. Mark Hall[1], Lowell H. Hall[1] & Lemont B. Kier[2]
[1]*Department of Chemistry, Eastern Nazarene College, Quincy, MA 02170, USA;* [2]*Department of Medicinal Chemistry, School of Pharmacy, Virginia Commonwealth University, Richmond, VA 23298, USA*

MS received 1 November 2002; accepted for publication 25 November 2002

Summary

The binding of beta-lactams to human serum proteins was modeled with topological descriptors of molecular structure. Experimental data was the concentration of protein-bound drug expressed as a percent of the total plasma concentration (percent fraction bound, PFB) for 87 penicillins and for 115 β-lactams. The electrotopological state indices (E-State) and the molecular connectivity chi indices were found to be the basis of two satisfactory models. A data set of 74 penicillins from a drug design series was successfully modeled with statistics: $r^2 = 0.80$, $s = 12.1$, $q^2 = 0.76$, $s_{press} = 13.4$. This model was then used to predict protein binding (PFB) for 13 commercial penicillins, resulting in a very good mean absolute error, MAE = 12.7 and correlation coefficient, $q^2 = 0.84$. A group of 28 cephalosporins were combined with the penicillin data to create a dataset of 115 beta-lactams that was successfully modeled: $r^2 = 0.82$, $s = 12.7$, $q^2 = 0.78$, $s_{press} = 13.7$. A ten-fold 10% leave-group-out (LGO) cross-validation procedure was implemented, leading to very good statistics: MAE = 10.9, $s_{press} = 14.0$, q^2 (or r^2_{press}) = 0.78. The models indicate a combination of general and specific structure features that are important for estimating protein binding in this class of antibiotics. For the β-lactams, significant factors that increase binding are presence and electron accessibility of aromatic rings, halogens, methylene groups, and =N– atoms. Significant negative influence on binding comes from amine groups and carbonyl oxygen atoms.

Introduction

Drugs in the human circulatory system often bind to components of the plasma such as albumin, acid-glycoprotein (AGP) or lipoproteins. Such binding can take place in association with single, or multiple, plasma elements. Albumin, which comprises more than half of all blood proteins, is the most significant plasma component involved in the binding of drugs. Albumin can interact with either acidic or basic drugs by van der Waals dispersion forces, hydrophobic bonding, hydrogen bonding and ionic interactions [1], AGP binds primarily with basic entities, and lipoproteins bind both basic and neutral compounds.

The binding of a drug to serum proteins in human plasma is a powerful determinant of its pharmacodynamic behavior and can affect the systemic distribution of the drug in several ways. The drug-protein complex does not permeate phospholipid membranes, including capillary membranes, glomerular membranes in the nephrons, and the blood brain barrier [1]. Oral bioavailability (%F) is directly affected by the extent to which a drug binds to plasma proteins because the fraction of drug bound to proteins is not available to the mechanisms that govern first pass metabolism. The free concentration of the drug, hence the biological activity, is thus at stake when a drug binds to available serum proteins. The reversible drug-protein complex circulates through the system and serves as a depot, making available unbound drug when the concentration of free drug has been depleted by metabolic and excretory clearance. By this mechanism, bound drug can replenish the free concentration of the drug *in vivo.* For these reasons, drugs with high protein binding activity values tend to have a greater half-life compared to those with lower values. The prolonged activity resulting from these factors may be desirable, or may promote the emergence of undesirable side effects

Estimation of this significant ADME property (absorption, distribution, metabolism and excretion) is one of the important activities carried out in the early

stages of drug design. Some understanding of the possible binding characteristics of candidate molecules is valuable information affecting the strategies of the design process. Recently available information suggests that success in drug development occurs at a rate of less than 20% and poor pharmacokinetic properties in the candidate compounds account for almost half of the reported failures [2, 3]. A goal of this present study is to create models that can predict the extent of serum protein binding for a candidate drug molecule.

Of particular interest in our investigation is the protein binding of cephalosporins and penicillins, which together constitute the β-lactam class of antibiotics. The common scaffolds for penicillins and cephalosporins are given in Figure 1. Designing antibiotics that will bind to serum protein to a predictable extent is a way of regulating the available concentration for the treatment of an infection. It is also a way of controlling the half-life to ensure a sufficient time for the availability of the drug to carry out its antimicrobial role. Since proteins are rarely excreted in the urine, the protein-bound antibiotic is available as a depot for a prolonged effect. Furthermore, the protein-bound antibiotic is much less available for metabolism that will end its utility as an antibiotic.

Additionally, information and understanding of some of the molecular structure influences on protein binding by the β-lactams may also be useful in anticipating the potencies of derivatives since the mechanism of action of this antibiotic class involves binding to the surface of bacterial cell walls, destroying their normal function. Information about binding is a potential correlate of mechanism, hence potency for the intrinsic activity of the drug. In contrast to these benefits accruing from protein binding, the prolonged presence of a significant concentration of antibiotic may expose the patient to undesirable side effects. A predictive model of serum protein binding is of value to optimize these events in the design of new drugs.

The accuracy and precision of reported percent protein binding of many drugs is not entirely clear. Data is to be found in selected articles, text appendices, drug information handbooks, and package inserts. These sources may reference seminal articles; they may also express error limits or deviations in the data, or this information may be lacking. Any attempt to study the structure-activity relationships of this property must be done with the awareness that the data is from eclectic sources and of generally unknown accuracy. Our experience from examining protein binding data of over 300 drugs from these sources leads us to expect a variation of at least ± 5 percentage points and perhaps twice that range in some cases. This problem is not a fatal flaw, denying the use of such data to create QSAR models, since the objective at the early stages of drug design is to classify the extent of this property among candidate molecules. A useful QSAR model is thus one that will permit prediction of percent protein binding within useful ranges. Indeed, Frostell-Karlsson et al. [4] have suggested the utility of a classification scheme that makes use of three levels (low, medium, high) to characterize the extent of binding.

Objectives

The principle objective of this investigation is to develop quantitative models that will predict the extent to which β-lactam compounds exhibit reversible binding to human serum proteins (HSP). The investigation makes use of the topological method, the main focus of which is the development of QSAR models based on structure information whose relation to the activity can be interpreted in a direct and straightforward manner. Rather than an attempt to simulate the binding process or to develop equations for interactions in an assumed mechanism, we use in our model information that directly represents the structure in the set of molecules in the data set. In this approach, topological descriptors represent molecular structure information. Statistical methods yield a model that captures the parallel between variation in structure features and corresponding variation in activity values. These methods do not require information based on 3D geometry, which often requires time-consuming quantum mechanical calculations, nor is it necessary to make assumptions about the mechanism of the process.

The first segment of the investigation involves the examination of a database composed of a series of structures synthesized as part of an effort to optimize protein binding in a specific class of drugs. For this we use the report of Bird and Marshall in which a series of penicillins were synthesized and assayed by an ultrafiltration method [5] to establish percent fraction bound to serum proteins (PFB). All compounds in this study were synthesized and measured for activity in the same lab, by the same method [6]. The predictive capability of the resulting QSAR model is tested on a set of known, commercially available penicillins, whose activity values are given in Goodman

and Gillman [7]. The second segment of the investigation involves the use of an expanded data set of β-lactam structures, containing both penicillins and cephalosporins. The predictive quality is revealed in a rigorous statistical treatment.

Data and methods

Data sets used for the studies of the investigation

In this investigation, three data sets were assembled and then used to develop and validate a progression of two models. The first model was developed with a training set of 74 penicillins taken from the paper of Bird and Marshall [6]. The R_2 substituents to the common scaffold for the compounds in the penicillin training set are given in Figure 2. The second dataset is composed of 13 additional penicillins obtained from Goodman and Gilman [7], and was used for external validation of the model based on the 74 structures from Bird and Marshall [6]. The R_2 substituent to the common scaffold for the compounds in the penicillin external validation set are given in Figure 3. There are a total of 79 compounds in the publication of Bird and Marshal, five of which are commercially available penicillins with a value also reported in Goodman and Gillman. Each of the five compounds found in both sources was included in the external validation test set and excluded from the training set. Comparison of the PFB values for compounds that appeared in both sources showed sufficient agreement to warrant the use of all values in a combined data set.

To conclude the investigation, a second model was constructed from a data set consisting of 115 β-lactams, created by adding the structures and PFB values for 28 commercially available cephalosporins to the data for the 87 penicillins used in the first study. The R_1 and R_2 subsitiuents for the cephalosporin additions to the β-lactam training set are given in Figure 4.

A preliminary analysis of the 74 penicillins used to develop the model in the first study revealed two observations considered to be of significance. First, the common scaffold of penicillin class compounds does not bind strongly to serum proteins. To cite examples, compound numbers 3 and 5, which are essentially the common scaffold, have PFB values of 15.0 and 18.0 respectively. These values are two of the five lowest values in the 74 examples. Secondly, initial modeling efforts, based on whole molecule description only, produced models of unsatisfactory quality. For both of these reasons, a substituent level of descriptor was created in the hope of distilling out pertinent structure information by focusing on the region of the molecule where the variation is greatest. When the cephalosporin compounds were added during the study on the β-lactam class antibiotics, substituent descriptors were created for both the R_2 fragment and the R_1 fragment, which is present only in the cephalosporins. Composite descriptors were also created to indicate E-State atom types and hydrogen E-State atom types present anywhere in the structure outside of the common scaffold (on either R_1 or R_2). For example, the $C\_S^T(dO)$ is defined for E-State values of carbonyl oxygen atoms not occurring in the common scaffold. The combination and composite descriptors appear in Figure 5 and are defined along with all others appearing in a final model. Molecular connectivity [8–11] and electrotopological indices [12–15] were used within the program MDL-QSAR [16].

The preliminary collection of descriptors in the databases for the studies was screened, resulting in the removal of all descriptors that had the same value for more than 90% of the compounds in the training set. This included descriptors whose value was zero for more than 90% of compounds. Intercorrelated pairs of descriptors were left in the descriptor pool for consideration, but no two descriptors with pairwise correlation of greater than 0.80 were used in a final model. Because of the nature of the topological method, a value for every descriptor exists for each compound. A value of zero for an atom type E-State index indicates that the atom type is absent from the molecule. Such presence/absence information is valuable in the design process as an aid in selection of groups for synthesis of new candidate structures.

Results

Study 1: Serum protein binding of penicillins

The first study in this investigation involved a database of 74 penicillins from the work of Bird and Marshall [4], see Figure 2. A five-variable QSAR model was obtained, where the variables are defined in Figure 5.

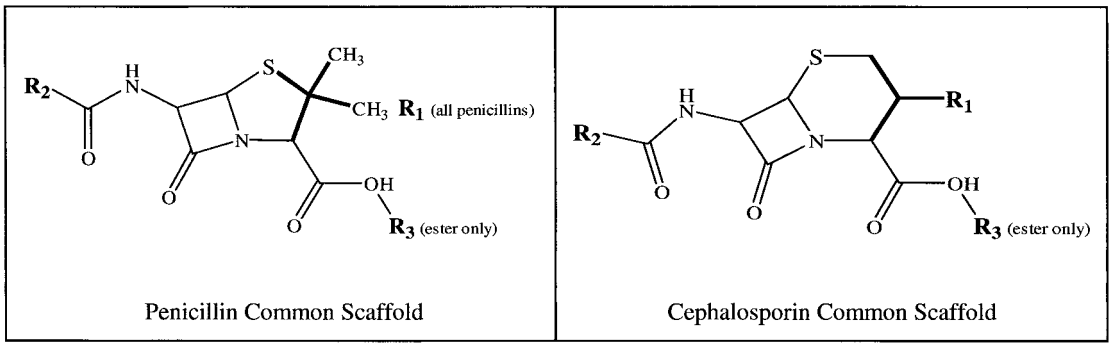


Figure 1. The β-lactam class of antibiotics. The common scaffolds of penicillins and cephalosporins with substitution points labeled R_1, R_2, and R_3.

The equation model and statistics are as follows:

$$\begin{aligned} PFB = {} & +14.60 * R_2{}^2\chi^v(\pm 1.6) \\ & +1.35 * S^T(F\_Cl)(\pm 0.25) \\ & +4.44 * R_2S^T(-CH_2-)(\pm 1.0) \\ & +2.25 * S^T(arom)(\pm 0.45) \\ & -21.1 * R_2HS^T(amine)(\pm 1.9) \\ & +13.2 \end{aligned} \quad (1)$$

$r^2 = 0.80$, $s = 12.1$, $F = 60$, $n = 74$

$q^2 = 0.76$, $s_{press} = 13.4$

Quantities in parentheses are the standard deviation of the coefficients. The statistical quantities q^2 and s_{press} are based on the leave-one-out (LOO) method. Only two residuals exceed two-sigma (but less than three sigma). The observed, calculated and residual values from the penicillin training set are given in Figure 2 along with the R_2 substituents. A plot of observed and calculated PFB values for the penicillin training set is given in Figure 6.

This analysis typifies a common procedure in the early phase of a drug development project in which candidates in a class of compounds are targeted for further synthetic modification and a model is developed using information obtained in the preliminary evaluations. In that early stage several important questions may be asked: how good is such a model in predicting the activity for molecules that are similar to those at which the design process is targeted? Are properties well predicted for compounds similar to the ones that are hoped to result from the drug design process? From this evaluation there may arise a model useful for new compound design. To address this specific question in this study, we utilize the model described above (Equation 1) to predict the protein binding of an external validation test set consisting of thirteen penicillins in clinical use. The structures and PFB values were obtained from Goodman and Gilman [7]. The following statistics were obtained from the external validation:

predicted correlation coefficient $q^2 = 0.84$
root mean square (RMS) = 16.4
mean absolute error (MAE) = 12.7

These predictions, along with the R_2 subsittuents of the compounds in the test set, are found in Figure 3. A plot of observed and calculated PFB values for the external validation test set of 13 commercial penicillins is given in Figure 7. These results give us confidence that the process is leading to a useful model.

Study 2: Serum Protein Binding of β-Lactams

An expanded study of drug protein binding examined the possibility of building a model for a more general set of antibiotics, the β-lactams, that includes both penicillins and cephalosporins. The data set used for the second study combines the PFB data on the penicillins from Bird and Marshall with protein binding data from the penicillins in the external validation set and PFB values from 28 cephalosporins taken from Goodman and Gilman [7]. The R_1 and R_2 substituents, along with the reported PFB values for the cephalosporins are included in Figure 4. This composite database contained a total of 115 compounds. An eight-variable QSAR model was obtained, where the variables are defined in Figure 5. The equation model

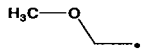
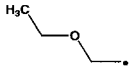
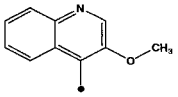
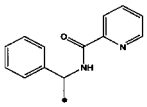
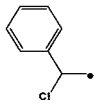
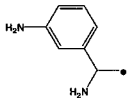
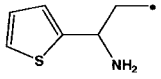
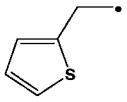
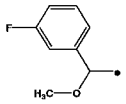
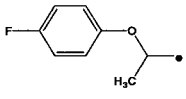
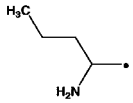
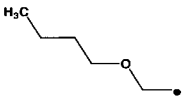
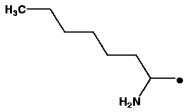
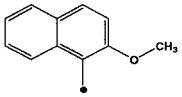
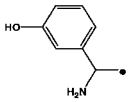
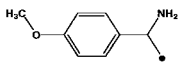
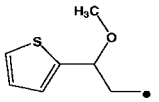
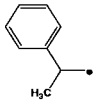
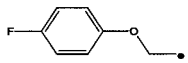
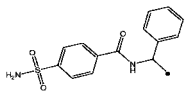
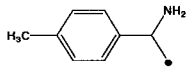
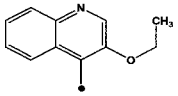
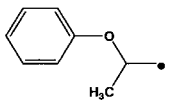
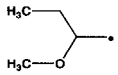
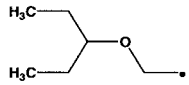
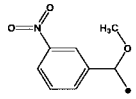
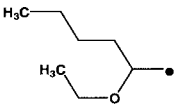
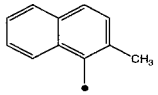
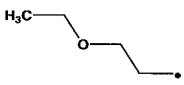
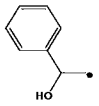
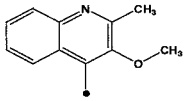
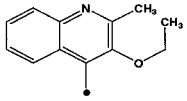
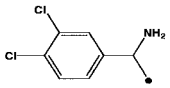
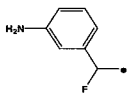
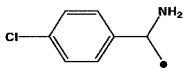
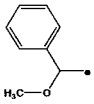
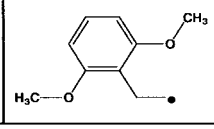
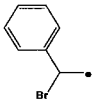
ID Name PFB Calc Res	ID Name PFB Calc Res	ID Name PFB Calc Res	ID Name PFB Calc Res	ID Name PFB Calc Res
1 Penicillin_5 7.2 26.7 -19.5 	9 Penicillin_6 28.0 33.5 -5.5 	17 Penicillin_76 57.0 61.7 -4.7 	25 Penicillin_36 63.0 66.3 -3.3 	33 Penicillin_22 78.0 83.9 -5.9 
2 Penicillin_31 12.0 4.8 7.2 	10 Penicillin_72 32.0 45.7 -13.7 	18 Penicillin_71 58.0 69.9 -11.9 	26 Penicillin_45 65.0 80.1 -15.1 	34 Penicillin_62 80.0 81.3 -1.3 
3 Penicillin_2 15.0 15.4 -0.4 	11 Penicillin_12 33.0 21.6 11.4 	19 Penicillin_7 58.8 55.7 3.1 	27 Penicillin_13 66.2 55.3 10.9 	35 Penicillin_67 80.0 70.7 9.3 
4 Penicillin_32 16.8 32.1 -15.3 	12 Penicillin_34 38.0 38.7 -0.7 	20 Penicillin_73 59.0 77.8 -18.8 	28 Penicillin_19 68.0 75.2 -7.2 	36 Penicillin_54 81.5 75.0 6.5 
5 Penicillin_1 18.0 15.4 2.6 All R_2 Values = 0	13 Penicillin_35 42.0 51.6 -9.6 	21 Penicillin_27 60.0 43.6 16.4 	29 Penicillin_78 69.7 68.3 1.4 	37 Penicillin_48 81.5 72.2 9.3 
6 Penicillin_9 20.0 39.2 -19.2 	14 Penicillin_8 47.0 54.7 -7.7 	22 Penicillin_46 60.0 64.8 -4.8 	30 Penicillin_10 74.0 68.5 5.5 	38 Penicillin_66 82.1 75.0 7.1 
7 Penicillin_11 25.0 44.7 -19.7 	15 Penicillin_37 53.2 65.5 -12.3 	23 Penicillin_77 61.7 64.6 -2.9 	31 Penicillin_79 74.5 71.1 3.4 	39 Penicillin_29 82.2 58.6 23.6 
8 Penicillin_30 26.0 47.5 -21.5 	16 Penicillin_28 55.0 49.3 5.7 	24 Penicillin_38 62.0 70.3 -8.3 	32 Penicillin_18 77.0 69.8 7.2 	40 Penicillin_21 82.5 85.0 -2.5 

Figure 2. R_2 substituents for the penicillin training set given with observed PFB value. Calculated PFB value and residual are derived from Equation 1.

ID Name PFB Calc Res	ID Name PFB Calc Res	ID Name PFB Calc Res	ID Name PFB Calc Res	ID Name PFB Calc Res
41 Penicillin_40 83.0 81.3 1.7	48 Penicillin_64 86.0 71.4 14.6	55 Penicillin_52 91.5 74.8 16.7	62 Penicillin_24 94.0 94.7 -0.7	69 Penicillin_43 96.5 101.0 -4.5
42 Penicillin_44 83.0 90.3 -7.3	49 Penicillin_50 86.1 79.8 6.3	56 Penicillin_42 92.0 91.1 0.9	63 Penicillin_69 94.7 99.4 -4.7	70 Penicillin_25 97.0 104.6 -7.6
43 Penicillin_53 83.5 75.1 8.4	50 Penicillin_14 87.0 60.8 26.2	57 Penicillin_3 92.4 87.5 4.9	64 Penicillin_56 94.8 83.1 11.7	71 Penicillin_59 97.0 94.1 2.9
44 Penicillin_75 83.6 93.1 -9.5	51 Penicillin_39 88.0 80.8 7.2	58 Penicillin_49 92.5 85.4 7.1	65 Penicillin_63 95.2 99.8 -4.6	72 Penicillin_51 97.2 111.1 -13.9
45 Penicillin_41 84.0 90.8 -6.8	52 Penicillin_68 89.3 77.5 11.8	59 Penicillin_4 93.3 107.1 -13.8	66 Penicillin_65 95.6 87.9 7.7	73 Penicillin_58 97.4 92.9 4.5
46 Penicillin_61 84.0 81.8 2.2	53 Penicillin_74 89.7 95.7 -6.0	60 Penicillin_20 93.6 71.0 22.6	67 Penicillin_55 96.0 82.5 13.5	74 Penicillin_70 97.4 83.2 14.2
47 Penicillin_60 86.0 81.9 4.1	54 Penicillin_16 91.0 102.0 -11.0	61 Penicillin_23 94.0 94.8 -0.8	68 Penicillin_57 96.0 92.7 3.3	

Figure 2. Continued.

and statistics are as follows:

$$
\begin{aligned}
\mathrm{PFB} \;=\; & +1.91 * \mathrm{S^T(arom)}(\pm 0.38) \\
& +14.4 * \mathrm{R_2^2}\chi^{\mathrm{v}}(\pm 1.4 \\
& +53.4 * \mathrm{R_1 d^0}\chi(\pm 14.0) \\
& +1.26 * \mathrm{S^T(F\_Cl)}(\pm 0.26) \\
& +3.81 * \mathrm{R_2 S^T}(-\mathrm{CH_2}-)(\pm 0.97) \\
& +5.60 * \mathrm{S^T}(=\mathrm{N}-)(\pm 0.67) \\
& -19.4 * \mathrm{R_2 HS^T(amine)}(\pm 1.8) \\
& -0.52 * \mathrm{C\_S^T}(=\mathrm{O})(\pm 0.21) \\
& -3.71 \qquad (2)
\end{aligned}
$$

$r^2 = 0.82$, s = 12.7, F = 60, n = 115

$q^2 = 0.78$, $s_{press} = 13.7$

The observed, calculated and residual values for all 115 compounds are given in Table I. Quantities in parentheses are the standard deviation of the coefficients. The statistical quantities q^2 and s_{press} are based on the leave-one-out (LOO) method. Only five residuals exceed two-sigma (but less than three sigma). A plot of observed and calculated PFB values for the penicillin training set is given in Figure 8.

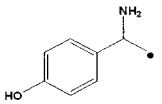
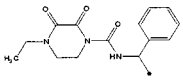
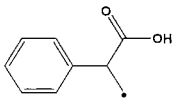
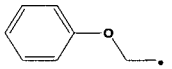
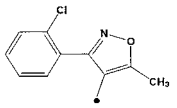
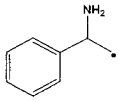
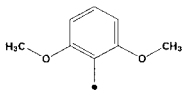
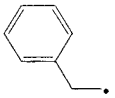
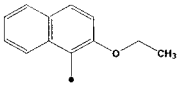
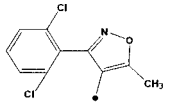
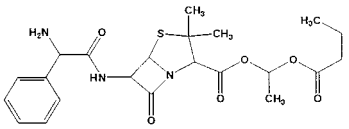
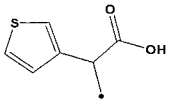
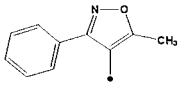
ID Name PFB Pred Res	ID Name PFB Pred Res	ID Name PFB Pred Res	ID Name PFB Pred Res	ID Name PFB Pred Res
75 Amoxicillin 18.0 32.3 -14.3 	78 Piperacillin 30.0 64.7 -34.7 	80 Carbenicillin 50.0 68.9 -18.9 	83 Penicillin V 80.0 66.5 13.5 	86 Cloxacillin 95.0 91.0 4.0 
76 Ampicillin 18.0 38.8 -20.8 	79 Methicillin 39.0 59.8 -20.8 	81 Penicillin G 60.0 71.5 -11.5 	84 Nafcillin 89.0 93.3 -4.3 	87 Dicloaxacillin 96.0 101.6 -5.6 
77 Bacampicillin 20.0 38.8 -18.8 		82 Ticarcillin 65.0 65.7 -0.7 	85 Oxacillin 92.0 80.5 11.5 	

Figure 3. R_2 substituent, or entire structure, for the external validation set of commercial penicillin used to evaluate the model given in Equation 1. Structures are given along with observed PFB value. Predicted PFB value and residual are derived from Equation 1. Structures also included in the β-lactam training set for the second study.

To reveal the strength of the model, two computational tests were used. The first test is commonly referred to as, 'y-randomization', where the PFB values were randomly scrambled and the statistics for the model re-computed. This process was carried out 100 times. The r^2 values obtained in this randomization process ranged from 0.015 to 0.16, with an average of 0.071. Based on this information, we conclude that the model (Equation 2) is significantly different from that obtained from random numbers; that is, the model encodes significant rather than random information.

The second test procedure was based on the ten-fold leave-group-out-and-predict approach. In this case, for each fold, 10% of the data was deleted; a new model was obtained for the remaining 90%; and then the deleted compounds were predicted from the new model. This process was repeated until all 115 compounds had been left out and predicted once. The entire process was then repeated ten times. For all 1,150 residuals obtained in this manner, the mean absolute error is MAE = 10.9 and the standard error is $s_{press} = 14.0$. For this multiple 10% leave method, the q^2 (or r^2_{press}) $= 0.79$. This approach is considered a very strong internal validation test of the model, significantly better than the typical leave-one-out (LOO) method, and has become known as the leave-group-out method (LGO) [17-22]. A plot of predicted and observed PFB values for the β-lactam antibiotics is given in Figure 9.

For the model based on eq. 2, the average residual for the 25 compounds with binding percentages of 0 to 30% was found to be 11.4%, while the average for 27 compounds with binding percent greater than 90% was only 8.0%. The greater accuracy of the predictions at high binding percent is desirable since small changes in this range reflect major changes in the unbound percent. For example, the change in binding from 96 to 92 percent binding corresponds to a doubling of the unbound fraction.

Interpretation of the model: β-lactams

A common occurrence of variables runs through the studies described above. One possible use of this information is the proposal of a mechanism of binding for the data modeled. This notion is not a realistic goal in view of the variety of binding sites encoded in the experimental data. An achievable goal is to define the features that are prominent in the models and then to incorporate this general information into the drug design process. In this way trends can be derived from these models that tell us whether a structure feature is important in supporting binding or is inimical to that event. Accordingly we may dissect the equations and extract this general information for use in drug design.

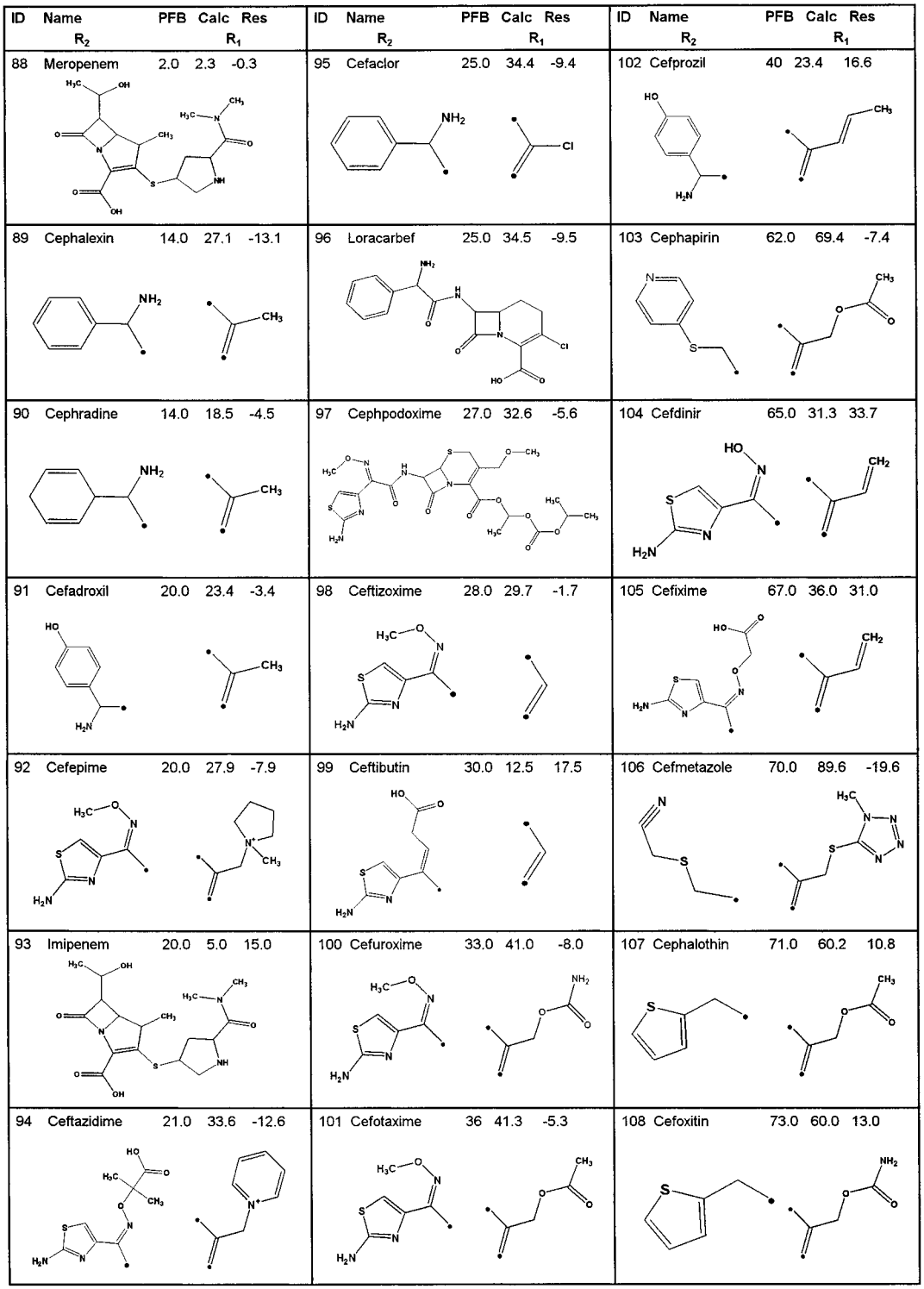


Figure 4. R_1 and R_2 substituents, or entire structure, for the cephalosporins included in the β-lactam training set. Structures are given along with observed PFB value. PFB calculated value and residual are derived from Equation 2.

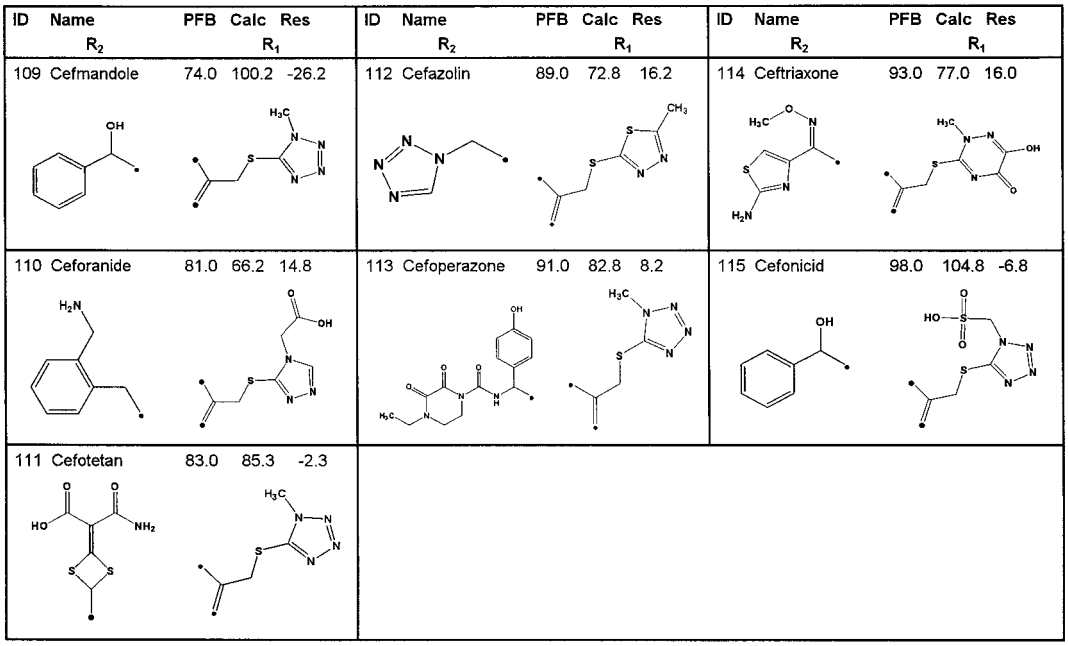


Figure 4. Continued.

Symbol	Definition	Symbol	Definition
S^T(arom)	$= S^T$ =CH= + S^T =C= + S^T =C=	R_2 S^TssCH$_2$	$= S^T$ $-CH_2-$ - for R_2 only
R_2HS^T(amine)	$= HS^T$ $-NH_2$ + HS^T $-\overset{H}{N}-$ - for R_2 only	S^TdsN	$= S^T$ =N—
R_1 d $^0\chi$	- zero order difference chi index - R_1 only	S^T(F_Cl)	$= S^T$ —Cl + S^T —F
R_2 $^2\chi^v$	- second order chi valence path two index - R_2 only	C_S^TdO	$= S^T$ =O - on either R_1 or R_2

Figure 5. Structure descriptors and their definitions. All descriptors appear in a models discussed in one of the studies in this investigation.

The definitions of the structure descriptors are given in Figure 5.

On average, two of the descriptors in Equation 2 account for 58% of the calculated PFB. These two descriptors encode more general aspects of the β-lactam structures, skeletal variation, and aromaticity: R_2 $^2\chi^v$ and S^T(arom). The other six descriptors account for more specific structure features. It should be noted that the common scaffold of these β-lactams does not contribute much to the binding. Compound 10 in Table 1 (penicillin_1) only binds 18%. Compound 6, penicillin_2, with an additional methyl group in R_2, only binds 15%.

It is significant to note that when the dataset was increased from 74 structures to 115 structures, all of the descriptors that were in the best model based on the 74 penicillins were also found in the best model for the 115 β-lactams. The three additional descriptors added to the model for the 115 beta-lactams, namely, R_1 $d^0\chi$, S^T(=N−) and C_S^T(=O), encode information for structural features that are absent in the data set of the 74 penicillins. The addition of the 28 cephalosporins and the 13 commercial penicillins resulted in a model with five of the eight descriptors being identical and the additions being structural description of new features in the data set.

Interpretation of individual descriptors

S^T(arom)

The S^T(arom) variable records the collective E-State values for aromatic carbons and CH fragments in

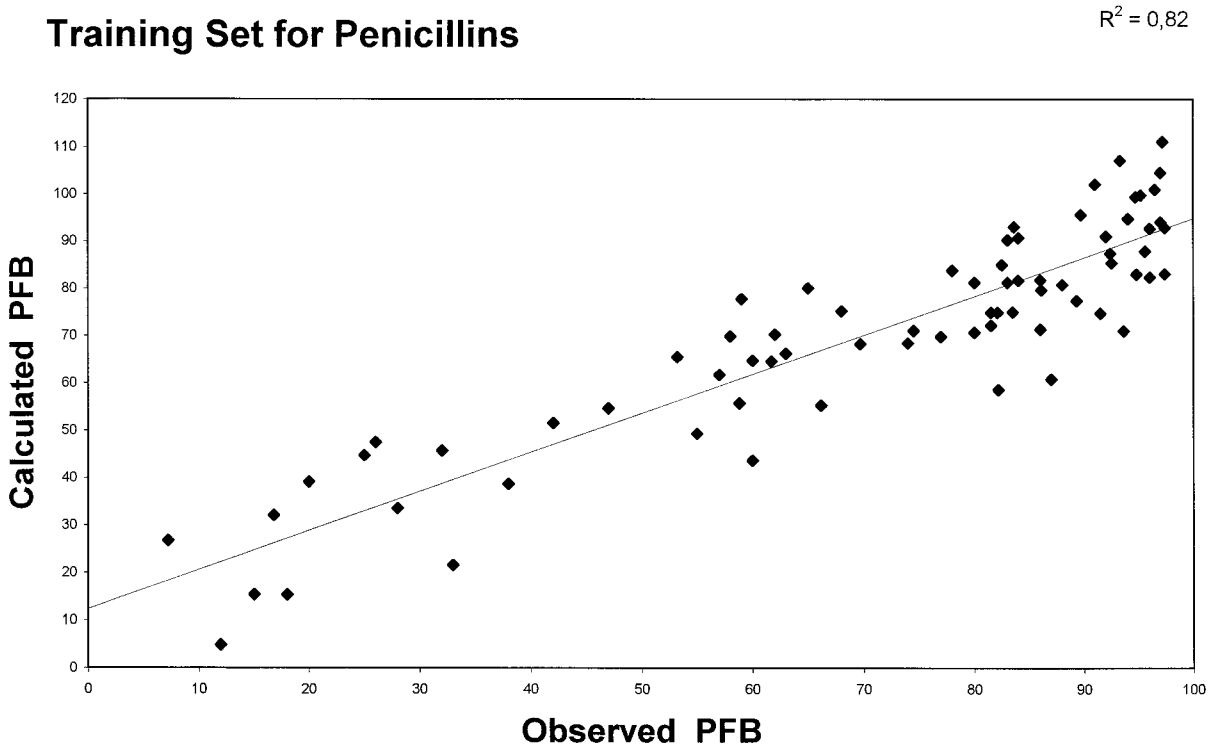


Figure 6. Plot of calculated PFB versus experimental values for the penicillin training set according to the topological model, Equation 1.

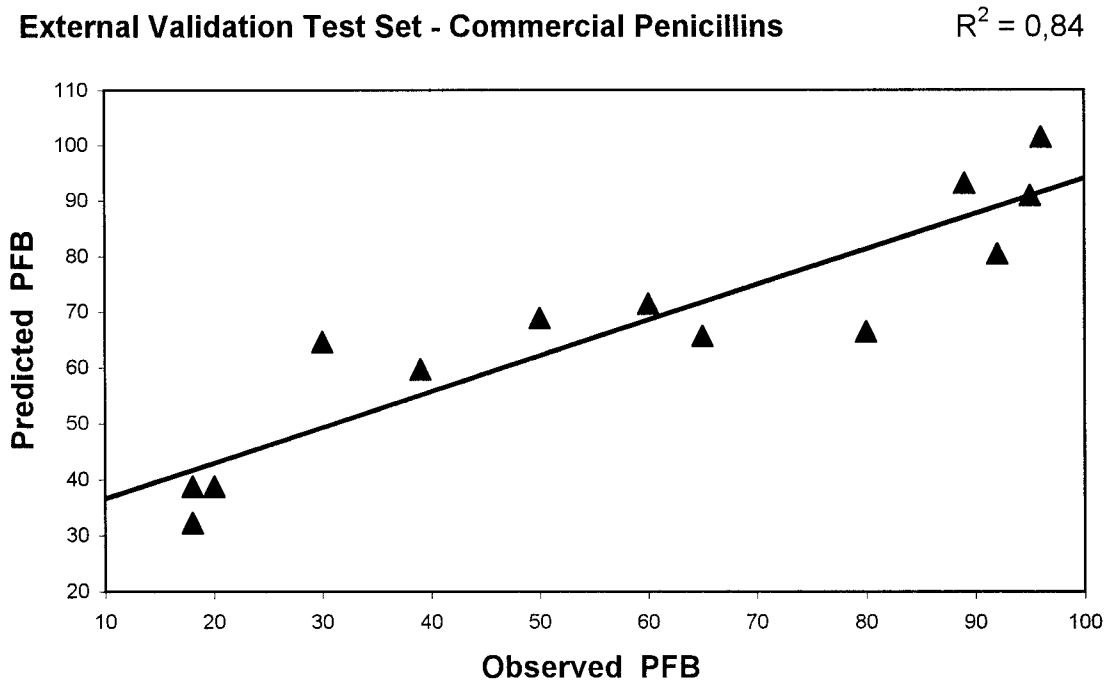


Figure 7. Plot of predicted PFB versus experimental values for the commercial penicillin external validation test set, according to the topological model, Equation 1.

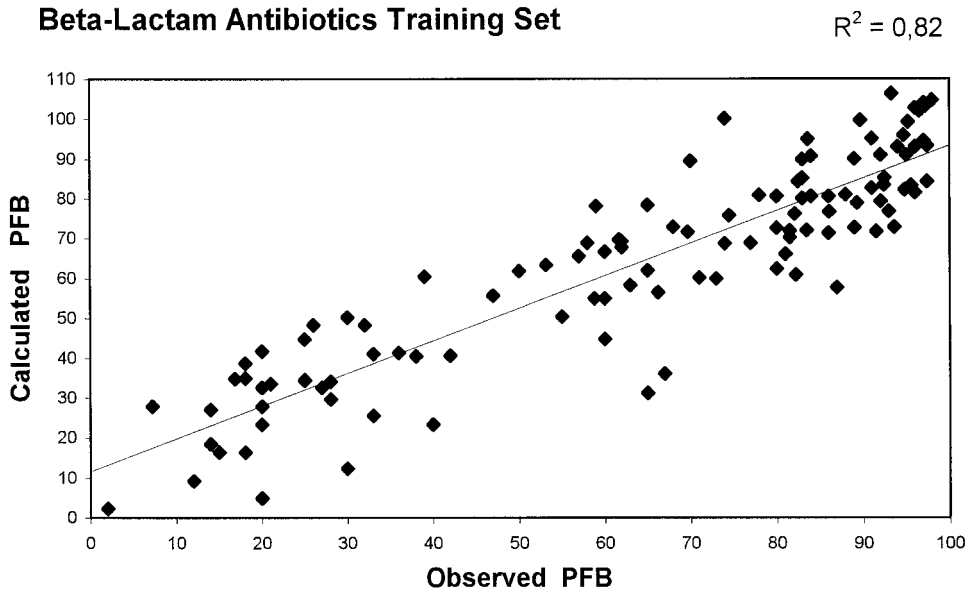


Figure 8. Plot of calculated PFB versus experimental values for β-lactam data set, according to the topological model, Equation 2.

a molecule, including phenyl groups and heteroaromatic rings. This index has been found useful in earlier studies in which contributions from the three aromatic carbon atom types were all important [23]. The magnitude of this index depends upon the presence and nature of the substituents. For example, electronegative groups decrease the value of S^T(arom). Substituted aromatic carbon atoms also tend to have smaller E-State values than the unsubstituted carbon atoms. The significance of aromatic functional groups is well illustrated by the difference in activity value between penicillin_2 (#3) and penicillin_14 (#50). Penicillin_2 is essentially the common scaffold for the penicillin class. Penicillin_2 has a methyl group for its R_2 substituent and a PFB of 15%. In penicillin_14, the methyl group is replaced with a phenyl, causing a dramatic increase in the PFB value to 87%. The highest value for the S^T(arom) index occurs for an unsubstituted phenyl ring. For the 96 compounds with aromatic rings, the S^T(arom) descriptor accounts for 14.6% of the calculated PFB values on average; ranging from 2.3 to 34.2%. The positive coefficient of this index indicates that binding is enhanced by the presence of relatively non-polar aromatic fragments. Compounds with high PFB values tend to have high S^T(arom) descriptor values. Eleven of the nineteen compounds with zero values for S^T(arom) have PFB values less than 50%.

R_2 $^2\chi^v$

This chi descriptor, the second order chi valence index ($^2\chi^v$) for the R_2 substituents only, encodes the skeletal architecture of those substituents. The more branched the substituent, the greater is the value of R_2 $^2\chi^v$. Heteroatoms beyond fluorine in the periodic table also increase the value. The positive coefficient indicates that the larger the descriptor, the larger the contribution to calculated PFB. The R_2 $^2\chi^v$ descriptor contributes on average 42.9% of the calculated, ranging from 0 to 66.3%. Forty of the fifty compounds with PFB > 80% have $R_2{}^2\chi^v$ values greater than its average value.

R_1 d^0 χ

The zero order difference simple chi index ($d^0\chi$) encodes the extent of branching in the R_1 substituent that is present only in the cephalosporins and, on average, accounts for 23.3% of the calculated PFB. The contribution varies up to 100% for the three smallest compounds: penicillin_1, penicillin_2 and imipenem. Because of the way in which the R_1 substituent fragment was created, all compounds have a value for this descriptor even though only the cephalosporins have a substituent at this position. The same value for R_1 $d^0\chi$ is assigned to all penicillins in the data set. This variable acts as a discriminant to differentiate between the two chemical classes as well as providing branching information about the R_1 substituent for the cephalosporins. The greater the degree of branching

Table 1. Experimental and calculated protein binding for β-lactam data set

ID	Name	PFB[a]	calc[b]	res[c]
1	Meropenem*	2.0	2.3	−0.3
2	Penicillin_5*	7.2	27.9	−20.7
3	Penicillin_31*	12.0	9.2	2.8
4	Cephalexin	14.0	27.1	−13.1
5	Cephradine	14.0	18.5	−4.5
6	Penicillin_2*	15.0	16.5	−1.5
7	Penicillin_32*	16.8	34.8	−18.0
8	Amoxicillin*	18.0	35.0	−17.0
9	Ampicillin*	18.0	38.7	−20.7
10	Penicillin_1*	18.0	16.5	1.5
11	Bacampicillin**	20.0	32.6	−12.6
12	Cefadroxil	20.0	23.4	−3.4
13	Cefepime	20.0	27.9	−7.9
14	Imipenem*	20.0	5.0	15.0
15	Penicillin_9*	20.0	41.7	−21.7
16	Ceftazidime	21.0	33.6	−12.6
17	Cefaclor	25.0	34.4	−9.4
18	Loracarbef	25.0	34.5	−9.5
19	Penicillin_11*	25.0	44.8	−19.8
20	Penicillin_30*	26.0	48.3	−22.3
21	Cephpodoxime	27.0	32.6	−5.6
22	Ceftizoxime	28.0	29.7	−1.7
23	Penicillin_6*	28.0	34.2	−6.2
24	Ceftibuten	30.0	12.5	17.5
25	Piperacillin**	30.0	50.3	−20.3
26	Penicillin_72*	32.0	48.3	−16.3
27	Cefuroxime	33.0	41.0	−8.0
28	Penicillin_12*	33.0	25.5	7.5
29	Cefotaxime	36.0	41.3	−5.3
30	Penicillin_34*	38.0	40.5	−2.5
31	Methicillin	39.0	60.5	−21.5
32	Cefprozil	40.0	23.4	16.6
33	Penicillin_35*	42.0	40.6	1.4
34	Penicillin_8*	47.0	55.7	−8.7
35	Carbenicillin**	50.0	61.9	−11.9
36	Penicillin_37*	53.2	63.4	−10.2
37	Penicillin_28*	55.0	50.5	4.5
38	Penicillin_76*	57.0	65.6	−8.6
39	Penicillin_71*	58.0	69.0	−11.0
40	Penicillin_7*	58.8	55.0	3.8
41	Penicillin_73*	59.0	78.2	−19.2
42	Penicillin_27*	60.0	44.8	15.2
43	Penicillin_46*	60.0	55.0	5.0
44	PenicillinG**	60.0	66.7	−6.7
45	Penicillin_77*	61.7	69.8	−8.1
46	Cephapirin	62.0	69.4	−7.4
47	Penicillin_38*	62.0	67.9	−5.9
48	Penicillin_36*	63.0	58.3	4.7
49	Cefdinir	65.0	31.3	33.7
50	Penicillin_45*	65.0	78.5	−13.5
51	Ticarcillin**	65.0	62.0	3.0
52	Penicillin_13*	66.2	56.5	9.7
53	Cefixime	67.0	36.0	31.0
54	Penicillin_19*	68.0	72.9	−4.9
55	Penicillin_78*	69.7	71.7	−2.0
56	Cefmetazole	70.0	89.6	−19.6
57	Cephalothin	71.0	60.2	10.8
58	Cefoxitin	73.0	60.0	13.0
59	Cefamandole	74.0	100.2	−26.2
60	Penicillin_10*	74.0	68.8	5.2
61	Penicillin_79*	74.5	75.9	−1.4
62	Penicillin_18*	77.0	69.0	8.0
63	Penicillin_22*	78.0	81.0	−3.0
64	Penicillin_62*	80.0	80.7	−0.7
65	Penicillin_67*	80.0	72.7	7.3
66	PenicillinV*	80.0	62.5	17.5
67	Ceforanide	81.0	66.2	14.8
68	Penicillin_54*	81.5	72.0	9.5
69	Phenethicillin*	81.5	70.3	11.2
70	Penicillin_66*	82.1	76.3	5.8
71	Penicillin_29*	82.2	61.0	21.2
72	Penicillin_21*	82.5	84.4	−1.9
73	Cefotetan	83.0	85.3	−2.3
74	Penicillin_40*	83.0	80.1	2.9
75	Penicillin_44*	83.0	89.9	−6.9
76	Penicillin_53*	83.5	72.1	11.4
77	Penicillin_75*	83.6	95.1	−11.5
78	Penicillin_41*	84.0	90.8	−6.8
79	Penicillin_61*	84.0	80.7	3.3
80	Penicillin_60*	86.0	80.7	5.3
81	Penicillin_64*	86.0	71.5	14.5
82	Propicillin*	86.1	76.9	9.2
83	Penicillin_14*	87.0	57.7	29.3
84	Penicillin_39*	88.0	81.1	6.9
85	Cefazolin	89.0	72.8	16.2
86	Nafcillin**	89.0	90.1	−1.1

Table 1 continued.

Table 1 continued.

87	Penicillin_68*	89.3	79.0	10.3
88	Penicillin_74*	89.7	99.8	−10.1
89	Cefoperazone	91.0	82.8	8.2
90	Penicillin_16*	91.0	95.2	−4.2
91	Penicillin_52*	91.5	71.8	19.7
92	Clometocillin*	92.0	91.1	0.9
93	Oxacillin**	92.0	79.4	12.6
94	Penicillin_3*	92.4	83.6	8.8
95	Penicillin_49*	92.5	85.4	7.1
96	Ceftriaxone	93.0	77.0	16.0
97	Penicillin_4*	93.3	106.4	−13.1
98	Penicillin_20*	93.6	72.9	20.7
99	Penicillin_23*	94.0	93.2	0.8
100	Penicillin_24*	94.0	93.1	0.9
101	Penicillin_69*	94.7	96.1	−1.4
102	Penicillin_56*	94.8	82.4	12.4
103	Cloxacillin*	95.0	91.1	3.9
104	Penicillin_63*	95.2	99.4	−4.2
105	Penicillin_65*	95.6	83.4	12.2
106	Dicloxacillin*	96.0	102.8	−6.8
107	Penicillin_55*	96.0	81.7	14.3
108	Penicillin_57*	96.0	93.2	2.8
109	Penicillin_43*	96.5	102.1	−5.6
110	Penicillin_25*	97.0	104.1	−7.1
111	Penicillin_59*	97.0	94.7	2.3
112	Penicillin_51*	97.2	103.3	−6.1
113	Penicillin_58*	97.4	93.4	4.0
114	Penicillin_70*	97.4	84.4	13.0
115	Cefonicid	98.0	104.8	−6.8

*From Bird and Marshall [1]. **Compounds set aside for validation set to test Equation 1, from Goodman and Gilman [4].
a. PAB: percent compound bound.
b. Calc: PFB calculated from Equation 2.
c. res: PFB – calc.

in R_1, the greater is the contribution of this term. As noted above, this term is constant for penicillins. The positive coefficient indicates that increased branching in R_1 increases the contribution to calculated PFB.

S^T (F_Cl)

This E-State index, encoding the effect of the halogens (fluorine or chlorine) on the molecule, indicates that they are favorable structure features for binding by increasing electron accessibility. This descriptor encodes a combination of dispersion and dipolar interaction information. The positive coefficient indicates that the greater the descriptor value, the greater the contribution to calculated PFB. The significance of these halogen groups is evident in the difference in activity values between penicillin G (#81) and the series; penicillin_22 (#33), penicillin_24 (#62) and penicillin_25 (#70). Penicillin G has a phenmethyl for its R_2 substituent and a PFB of 60%. In penicillin_22, penicillin_24, and penicillin_25, the phenmethyl is substituted with one, two and three chlorines respectively, showing in increase in PFB from 60% to 78.0%, 94.0% and 97.0%. For the 30 compounds with fluorines or chlorines, S^T(F_Cl) contributes 15.2% on average but ranges from 6.7 to 26.1% of the calculated PFB value. When present, −F or −Cl make a significant contribution to the fraction bound.

S^T (=N–)

The atom type E-State index for the =N– group encodes the electron accessibility of the atom in molecules where non-aromatic nitrogen-heterocycles are present as well as for isolated =N– atoms. This structure feature makes a positive contribution to the extent of protein binding. For the sixteen molecules with this feature, S^T(=N–) contributes on average 31.4% but ranges from 14.8 to 67.6%; that is, in some compounds, the =N– nitrogen atom is very important to binding. The S^T(=N–) descriptor makes its largest contributions in cefmetazole (60.1%) and cefazolin (67.6%). The nitrogen in this kind of fragment is usually not very basic; hence, the expected contribution may be a hydrogen bond acceptor as in oxazoles and imidazoles.

R_2 S^T(–CH_2–)

The R_2 S^T(–CH_2–) descriptor is a second variable that is specific for a particular single structure feature, –CH_2– groups on the R_2 substituent. This E-State value is decreased by electronegative groups nearby and increased with an increased number of methylene groups. For the 40 compounds with –CH_2– groups in R_2, this descriptor contributes only 9.0% on average but ranges from 0 to 37.1% (penicillin_3) of the calculated PFB value. The contribution to PFB can be appreciable when several groups are present, such as in compounds 57 (penicillin_3) and 59 (penicillin_4) which have 6 methylenes each and PFB values of 92.4% and 93.3% .

R_2 HS^T (amine)

In contrast to the first six variables described above, this nitrogen-containing fragment is implicated as a deterrent to protein binding. This index is computed only for the R_2 substituent. It reflects polar fragments that are likely hydrated and, hence, less available

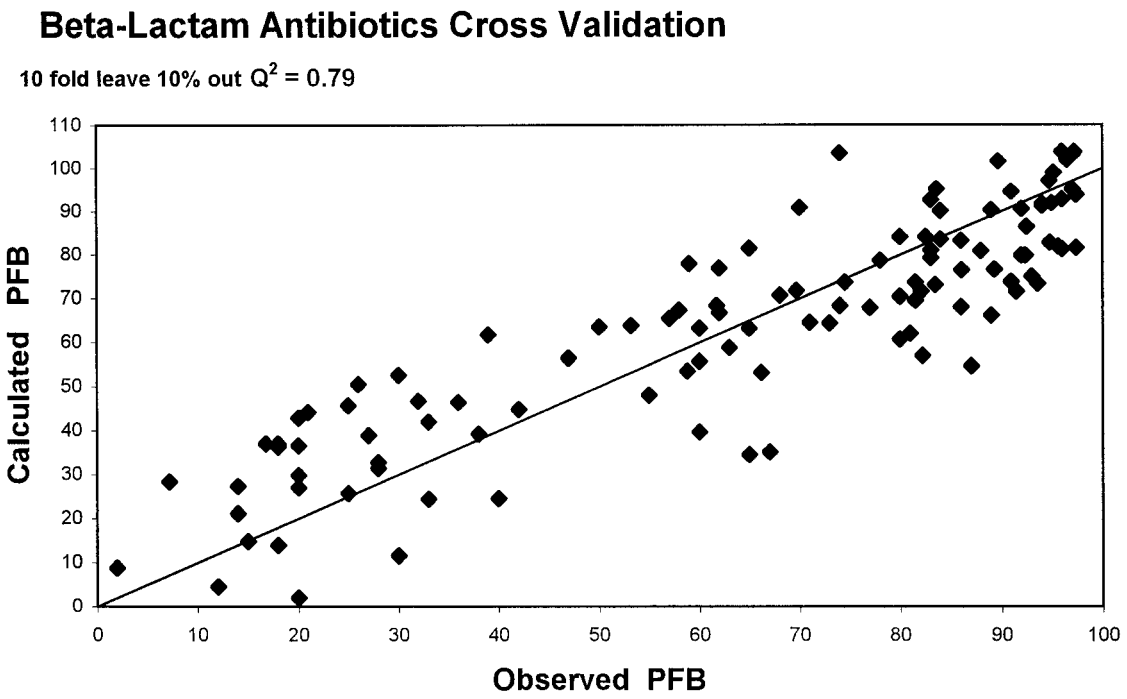


Figure 9. Plot of predicted PFB versus experimental values for the 10 fold leave-10%-out cross validation of the β-lactam data set, according to the topological model, Equation 2.

for interaction with much of the protein surface. The negative influence of amines is illustrated by the comparison of penicillin_3 (#53) and penicillin_13 (#27). The addition of a single $-NH_2$ to penicillin_3 causes a drop in activity from 92.4% to 66.2%. The effect is even more pronounced between penicillin G (#81) and ampicillin (#76) where the addition of $-NH_2$ causes a drop in observed PFB from 60% to 18%. This index contributes an average of about 8.5% of the influence on the prediction. For the 35 compounds with $-NH_2$ and /or –NH– groups in R_2, the average contribution is 28.0% and ranges from 17.8 to 45.0%, making it a significant factor in binding considerations. Of the 34 compounds with PFB $< 50\%$, 24 have non-zero values for this descriptor; none of the compounds with PFB $> 93\%$ have amine groups present.

C_S^T (=O)

The C_S^T(=O) descriptor also encodes E-State electron accessibility information on a specific atom type, =O outside the common scaffold. The descriptor value increases with number of =O atoms and decreases with electronegative atoms nearby. This descriptor is one of two with a negative coefficient, indicating that presence of carbonyl groups (=O atoms) decreases binding. The descriptor contributes the smallest amount on average, however, for the 22 compounds with =O groups (outside the common scaffold), its average contribution is 8.1% and ranges from 3.9 to 33.9%. The negative coefficient indicates that the larger the C_S^T(=O) value, the lower the calculated PFB. The C_S^T(=O) contribution is largest for meropenem (33.9%), the compound with the smallest PFB in the database, 2.0%.

Discussion

These studies on databases, ranging from a close congeneric series of penicillin molecules to a more mixed structural set of β-lactams, reveal the potential for modeling this important pharmacodynamic property with topological indices. The quality of the models is at the upper limits of expectation for a property that is measured by a variety of methods with results of modest accuracy and precision. The kind of predictions arising from this investigation are more useful than merely categorizing the predictions into three categories such as high, medium or low extent of binding [5]. It is encouraging to see that our predicted results among the high percent binding molecules are better than the median value of the standard error. That upper

range has the highest variation of free drug, a critical attribute in dose assignment.

A second benefit from these models is the finding that, in each study, most of the variables are common among the datasets. This finding implies that we have been able to reveal a number of common structural features contributing to the protein binding of these drugs. We cannot escape the reality that there is more than one known binding site on albumin and that there are multiple proteins involved in the measured percent binding. Nevertheless it is possible from these results to infer that certain structure features enhance binding while others are detrimental.

The models created here using the E-State and molecular connectivity chi indices are perhaps near the limits of accuracy for this data, in view of the varied experimental methods employed and the relative uncertainty of the results from each method. These models should be considered within the context of the use of this information for the design of new compounds. The prediction of the extent of serum protein binding is often a critical issue in accepting a candidate molecule for further study. What is usually needed is a categorical prediction of this property. A range of two standard deviations for this model, 25%, suggests that the predictions may be useful for four or five categories of PFB. That realm is the utility that we have demonstrated here.

It is always desirable to develop a QSAR model that matches a measured activity of interest with molecular structure. A model employing physical properties or other empirical parameters, statistically coupled with this activity, is an intermediate model that requires further processing, that is, the translating of the empirical parameters into molecular structure in order to guide further synthesis. The use of parameters that reflect or are abstractions of the expected three-dimensional structures are less useful when the quality of the data is only approximate, as it is here. The significance of such parameters is unclear when the mechanism and/or the target of the candidate molecules is unknown or vaguely defined.

Under these circumstances the use of structural descriptors that encode more of the content and context of fragments of molecules is of greater value. This principal of 'scaling' of parameters to the attributes of the data and the expectations of the use of the model becomes an important issue in the choice of parameters, the potential value of the information generated, and the economy of effort in generating it. We advocate the topological approach used in this report for the type of study described, recognizing the significance of scaling in modeling design.

Conclusions

Models for both data sets yield reasonable estimates of binding, commensurate with the experimental quality of the data. The equations may be used for quantitative estimates with the expectation that predictions are useful for the usual process of drug design. For the protein binding by β-lactams, the model indicates that binding is increased by the presence of aromatic groups, branching in both the R_1 and R_2 substituents, presence of =N–, as well as –F and –Cl atoms, but is decreased by the presence of amino groups and carbonyl oxygen atoms. These qualitative statements on structure influences on binding may be used to assist the drug design process. Because these models involve topological descriptors of molecular structure, creating the models is straightforward. Further, the models may used to predict protein binding or binding affinity at high speed for virtual libraries of structures. A significant conclusion from this investigation is that through the use of the topological method for QSAR modeling, data that is typically available during early phases of the drug discovery process may be used to predict the properties of commercial drugs, which are the anticipated end point of the discovery process.

References

1. Foye, W.O. *Principles of Medicinal Chemistry, 3rd Edition*, Lea and Febiger, Philadelphia (1998) 33.
2. Prentis, R.A., Lis Y., Walker, S.R., Br. J. Clin. Pharmacol., 25 (1988) 387–396
3. Kennedy, T., Drug Discov. Today, 2(10) (1997) 436–444
4. Frostell-Karlsson, A., Ramaeus, A., Roos, H., Andersson, K., Borg, P., Hamalainen, M. and Karlsson, R., J. Med. Chem., 43 (2000) 1986.
5. Rolinson, G.N. and Sutherland, R., Br. J. Pharmac. Chemother. 25 (1965) 638.
6. Bird, A.E., Marshall A.C., Biochem. Pharmacol., 16 (1967) 2275.
7. Hardman, J.G. and Limbird, L.E. (Eds.) *Goodman and Gilman's The Pharmacological Basis of Therapeutics, 9th Edition*, McGraw-Hill Publications, New York (1996) appendix
8. Kier, L.B. and Hall, L.H., *Molecular Connectivity In Chemistry and Drug Research*. Academic Press, New York (1976)
9. Kier, L.B. and Hall, L.H., *Molecular Connectivity in StructureActivity Analysis*. John Wiley Publications, London (1986)
10. Hall, L.H., Kier, L.B., The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Relations, in *Reviews of Computational Chemistry*, Chap. 9, 367–

422, eds. Donald Boyd and Ken Lipkowitz, VCH Publishers, Inc. 1991.
11. Hall, L.H., Kier, L.B., In Devillers, J. and Balaban, A.T. (Eds.), Molecular Connectivity Chi Indices for Database Analysis and Structure-Property Modeling, in *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach, Reading, UK, 1999, pp. 307–360.
12. Hall, L.H., Kier, L.B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling. In Lipkowitz, K.B. and Boyd, D.B. (Eds.) *Reviews of Computational Chemistry*, VCH Publ., N.Y., 1991.
13. Kier, L.B., Hall, L.H., In Devillers, J. and Balaban, A.T. (Eds.), The Kappa Indices for Modeling Molecular Shape and Flexibility, in *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach, Reading, UK, 1999, pp. 455–490.
14. Kier, L.B. and Hall, L.H., *Molecular Structure Description: The Electrotopological State*, Academic Press, San Diego, (1999)
15. Hall, L.H.; Kier, L.B. In Devillers J. and Balaban, A.T. (Eds.) The Electrotopological State: Structure Modeling for QSAR and Database Analysis, in *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach, Reading, UK, 1999, 491–562.
16. Hall, L.H., Mohney, B.K., Kier, L.B. Quant. Struct.-Act. Relat., 12 (1993) 44–48.
17. Kier, L.B., Hall, L.H. Med. Chem. Res., 2 (1992) 497–502.
18. MDL QSAR^{TM}, MDL-SciVision, 200 Wheeler Road, Burlington MA.
19. Maw, H.H., Hall, L.H., J. Chem. Inf. Comput. Sci., 40 (2000) 1270–1275.
20. Maw, H.H., Hall, L.H., J. Chem. Inf. Comput. Sci., 41 (2001) 1248–1254.
21. Maw, H.H.; Hall, L.H. J. Chem. Inf. Comput. Sci., 42 (2002) 290–298.
22. Hall, L.H., Kier, L.B., J. Pharm. Sci., 67 (1978) 1743–1747.
23. Hall, L.H., Kier, L.B., J. Pharm. Sci., 67 (1978) 1408–1412.
24. Kier, L.B., Hall, L.H., J. Med. Chem. 20 (1977) 1631–1636.
25. Gough, J.D., Hall, L.H., J. Chem. Inf. Comput. Sci., 39 (1999) 356–361.