JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2018, 58, 234−243

pubs.acs.org/jcim

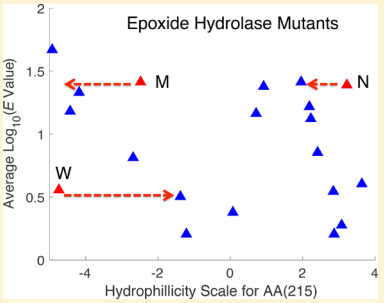
Improved Descriptors for the Quantitative Structure−Activity Relationship Modeling of Peptides and Proteins

Mark H. Barley, Nicholas J. Turner, and Royston Goodacre*

School of Chemistry, Manchester Institute of Biotechnology, University of Manchester, 131 Princess Street, Manchester, M1 7DN, U.K.

Supporting Information

ABSTRACT: The ability to model the activity of a protein using quantitative structure−activity relationships (QSAR) requires descriptors for the 20 naturally coded amino acids. In this work we show that by modifying some established descriptors we were able to model the activity data of 140 mutants of the enzyme epoxide hydrolase with improved accuracy. These new descriptors (referred to as physical descriptors) also gave very good results when tested against a series of four dipeptide data sets. The physical descriptors encode the amino acids using only two orthogonal scales: the first is strongly linked to hydrophilicity/hydrophobicity, and the second, to the volume of the amino acid residue. The use of these new amino acid descriptors should result in simpler and more readily interpretable models for the enzyme activity (and potentially other functions of interest, e.g., secondary and tertiary structure) of peptides and proteins.



INTRODUCTION

The use of quantitative structure−activity relationships (QSAR) in the prediction of the properties of small molecules (e.g., drugs) is well established. Molecular modeling of small molecules provides potentially thousands of descriptors (properties such as size, charge distribution, polarity, number of donor atoms, number of acceptor atoms, etc.) from which a model of drug activity (or solubility or any other relevant property) can be built. The ability to build similar models for the activity of peptide chains, let alone full enzymes, is much less developed. For a peptide chain or a protein the overall activity of the molecule is largely encoded and thus dependent upon the properties of the amino acids at specific sites within the chain. Changing the amino acid at a site potentially changes the way the chain of amino acids fold or form helices (secondary structure) and the way the folds and helices of the protein fit together (tertiary structure) potentially changing the activity of the molecule. However, there are no methods *a priori* of predicting the consequences of changing an amino acid at a specific site on the activity of the whole molecule, let alone predicting the effect of multiple changes at different sites. All that can be done is to collect data on the effect of changes and attempt to model the results empirically to make predictions and to gain understanding of how the amino acid residues in a folded peptide chain interact with each other. Any attempt to model the activity of a peptide or protein as a function of sequence (a method known as quantitative sequence−activity models (QSAM))[1] has to take account of these interactions by identifying those that are statistically significant from the much larger number of possible interactions—a process called variable or feature selection.

To apply QSAM methods to peptides and proteins, descriptors for the 20 amino acids coded by DNA are required. Sneath[2] introduced the idea of using principal component analysis (PCA) on a large number of physicochemical properties to group amino acids into "families". Two key descriptors that would be expected to have a big effect on the interaction of an amino acid residue with neighboring amino acids are the size of the residue and the hydrophobicity/hydrophilicity of the residue. While there is good agreement between different authors about the size of amino acid residues (as defined by the van der Waals volume)[3] there is no agreement between the various scales proposed to reflect the polarity of the amino acid residues (in this work raw data sets reflecting (potentially) the polarity/hydrophilic behavior of an amino acid will be referred to as polarity scales; once they have been processed by PCA the corresponding scale will be referred to as a hydrophilicity scale). Trinquier and Sanejouand[4] list 43 polarity scales with minimal agreement between them. Many of the suggested orders contravene physicochemical understanding of polarity and hydrogen bonding potential. For example arginine is clearly a highly polar amino acid residue with a delocalized charged group, yet two of the scales put it on the hydrophobic side of residues such as alanine, glycine and serine. Hellberg et al.[5] built upon the ideas of Sneath using PCA on 29 physicochemical properties of the 20 amino acids to produce 3 "principal properties" (PP also referred to as Z-scales) that could be used as descriptors for the amino acids. The first of their principal properties related closely to hydrophilicity or polarity of the amino acid and the second principal property to the size of the amino acid. They also showed that they could build some good partial least squares (PLS) models for the activity of several sets of peptide data using these three

Received: August 14, 2017
Published: January 16, 2018

DOI: 10.1021/acs.jcim.7b00488
J. Chem. Inf. Model. 2018, 58, 234−243

principal properties. Other authors have used similar methods to develop descriptors for amino acids (and potentially proteins). A recent paper[6] comparing sets of protein descriptors listed 13 sets of which 7 were obtained by PCA or factor analysis of a set of physicochemical properties of 20 (or more) amino acids. The minimum number of PCA derived descriptors used to represent an amino acid was 3 (as seen[5,6] for Z-scales and ProtFP) while other sets used 5 (e.g., expanded Z-scales to describe a further 67 non-natural amino acids[7]), 6 (FASGAI[8]), or 8 (VHSE[9]) PCA derived descriptors. Other descriptor sets use topological data or combinations of topological and physicochemical data as inputs to give the T_Scales[10] (5 PCs) and ST Scales[11] (8 PCs). The problem with using more descriptors to describe an amino acid is that it increases the chances of finding spurious models—particularly if many sites are being changed and interactions between these sites are likely.[12] Also the use of more descriptors to describe each amino acid greatly increases the number of terms that have to be considered when using variable selection in building QSAM models. When Hellberg et al. modeled the activity of ACE inhibitors (dipeptides) they used their descriptors (6) plus squared terms (6) and all interactions (15) (see their Table 4);[13] a total of 27 terms which they used without further selection. To create a similar model for a peptide of 16 amino acids the number of terms using the Hellberg descriptors would total 1224 and further selection to find the significant terms would be required. If only 2 descriptors describe each amino acid, this number reduces to 560; but it increases to 3320 for 5 descriptors per amino acid (e.g.: T-scales). Ideally we need a minimal number of descriptors for each amino acid, but these descriptors should accurately reflect the most important properties of the amino acids.

Despite their success in modeling peptides there are some problems with the way the Hellberg Z-scales (principal properties) have been derived. These can be summarized:

(1) The hydrophilicity scale is dominated by the nine HPLC variables that are closely correlated with each other (correlation coefficients range from 0.925 to 0.996). The authors recognized this problem and effectively reduced the weighting of these variables to a third of that of the other variables (a process they call soft-block scaling[14]). These nine variables (now equivalent to three of the other variables) still have a big and arbitrary impact on the definition of the hydrophilicity scale (Z_1).

(2) The authors do not give the proportion of variance explained by their principal properties. When seeking to reproduce their results (including the soft-block scaling referred to above) we found that the variance in the original data explained by the first three PCs was 39%, 20%, and 14%, respectively. The authors were clear that the first PC (Z_1) related to polarity/hydrophilicity and the second (Z_2) related to size but were less clear (based upon the loadings, their Table 3)[5] what the third PC (Z_3) was related to describing Z_3 as containing "information from the pK_a, pH at the isoelectric point, and ^1H NMR variables". Given the small proportion of variance explained by the third PC and the uncertainty about what it might relate to, it might be best to disregard this scale in peptide correlations.

(3) The specific order of some of the amino acids in the hydrophilicity scale contravene established rules about relative polarity of chemical functional groups. This is discussed in more detail below (section Deriving the New Hydrophilicity Scale).

Despite these issues, the first two of these principal property scales provide a starting point for the work described here. The aim of this work is to show that with some changes to the hydrophilicity scale suggested by Hellberg (mainly affecting the relative position of three amino acids) the modified scale can be used to model two properties (conversion and enantiomeric excess) of a set of mutants of a full enzyme (epoxide hydrolase). This new hydrophilicity scale, together with an orthogonal scale based upon van der Waals volumes for the amino acids, are compared to other descriptor scales: Hellberg Z-scales, MS-WHIM scores (both using three descriptors per amino acid), and T-scales. Comparisons are based upon models to describe four dipeptide data sets that are available in the public domain. In doing these comparisons the aim is not to show that one set of descriptors is statistically better than the next but rather to show that they give comparable results, i.e., which descriptor set gives the best results varies between the models considered. The new descriptors provide the possibility of modeling enzyme activity as a function of sequence using only two descriptors for each amino acid position providing simpler and more descriptive models. Two descriptors per amino acid will also reduce the search space to be explored if variable selection methods are used to identify significant interactions between amino acids in a protein or peptide.

RESULTS AND DISCUSSION

Deriving the New Hydrophilicity Scale. The data reported by Feng et al.[15] consists of experimentally determined conversion rates and enantiomeric excess (E values; average of three measurements) of 140 mutants of epoxide hydrolase. The data set is unusual (and potentially very useful for model development) as there is mutation at only two sites (AA215 and AA217) so that out of a maximum of 400 possible mutants, data for 140 (or 35%) are available which is exceptionally high coverage for a set of mutants for an enzyme. Figure 1 shows a plot of $\log_{10}(E$ values) against Hellberg Z_1 (the Hellberg

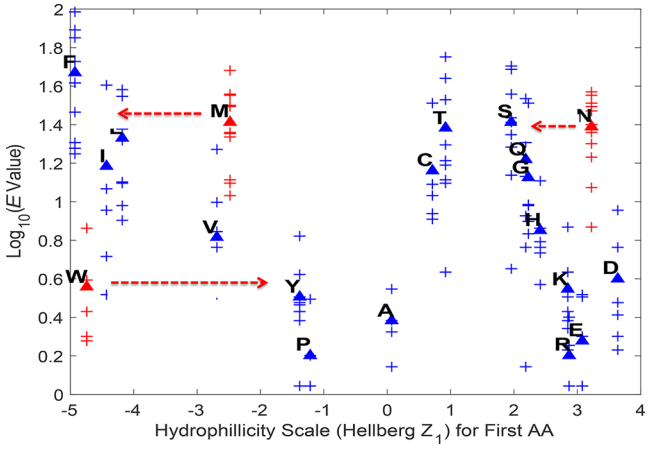


Figure 1. Plot of the $\log_{10}(E$ value) enantiomeric excess data from the work of Feng et al.[15] against the Hellberg Z_1 (hydrophilicity scale) values for the first amino acid (AA215). If the positions of three amino acids (in red) can be changed (see text) within this scale then a (distorted) cubic relationship becomes apparent. The crosses are the raw E value data, and the triangles are the averages for each amino acid.

DOI: 10.1021/acs.jcim.7b00488
J. Chem. Inf. Model. 2018, 58, 234−243

hydrophilicity scale) for the first amino acid position (AA215). The original data are shown as crosses, the average value for each amino acid is shown as a solid triangle. It can be seen that if the three amino acids in red (W, M, and N) could be moved as indicated the data would fit to a smooth curve with high *E* values associated with hydrophobic and strongly polar amino acid residues and low *E* values with weakly polar and highly charged residues. Mathematically this suggests a distorted cubic relationship between the *E* values and this modified Hellberg Z_1 scale.

From purely physical chemistry considerations, the hydrophobic/hydrophilic properties of the amino acids means they can be grouped according to similar properties. Hence F, L, I, V, A form a group of hydrophobic amino acid residues. These groups have no dipole moment and no means of interacting with strongly polar solvents such as water. There are then some weakly polar residues such as W, Y both of which have a small polar group within a large residue and then strongly polar residues such as S, T, N, Q which are all relatively small residues either with a large polar group (acid amide) or a strongly polar group (alcohol). Finally we have a set of residues that have charged groups at physiological pH: D, E, K, R; such charged groups will have strong interactions with water and other charged groups (salt bridges). They will avoid the hydrophobic groups and the interior of the protein structure. Charged groups in aqueous solution (e.g., on the outside of the protein) will be surrounded by a Debye−Huckle type "atmosphere" of counterions and orientated water molecules in a complex hydrogen bonded web of interactions.[16] These groups account for 15 of the 20 amino acids. Of the remainder, H, has an imidazole group with a pK_a (for the conjugate acid) of about 6.4 and hence may only be partially ionized at physiological pH. So this amino acid could be with either the charged group or the strongly polar group. Methionine (M) is a thioether. The corresponding ether would be considered weakly polar because of the dipole moment caused by the asymmetrical electron distribution of the carbon−oxygen bond as measured by the difference in electronegativity (3.44−2.55 according to Pauling[16]) and the geometry of the carbon−oxygen−carbon bond in an ether. However, in a thioether the difference in electronegativity is tiny (2.58−2.55) so M will have more in common with I and L than with molecules with a proper polar group. Likewise C (as a thiol) will be much less polar than S or T as the electronegativity difference between sulfur and hydrogen (2.58−2.2) is much lower than that between oxygen and hydrogen (3.44−2.2). From these arguments it can be seen that there is justification for the proposed changes in position for the three residues W, M and N. W should be moved to the right from the hydrocarbon group to the weakly polar group, M should move to the left, from the weakly polar group to the hydrocarbon group, and N should be moved from its far right position (more hydrophilic than the charged groups E, R and K) to join the strongly polar groups (S, T and Q). Using the above modifications in the order of the amino acids and assuming that both the *E* value data and the Conversion data can be modeled by a cubic relationship, it is possible to fit (with some additional constraints) a cubic curve to both sets of data using a common hydrophilicity scale (see Figure 2) that is orthogonal to a composite volume scale (see Methods for details). The final (compromise) hydrophilicity scale provided a fit to the *E* value data with $R^2 = 0.743$ and a fit to the conversion data with $R^2 = 0.749$ using 139 data points (one point was dropped as it was a clear outlier—see the Experimental Data section). The coefficients for the models are seen in Figure 2 and the fitted volume and hydrophilicity scales are shown in Table 1. The compromise polarity scale is the average of the polarity scales fitted to the two sets of data (see Methods section for more details). Within this work these new descriptors will be referred to from this point as "physical descriptors" as they closely correlate to two of the fundamental physical properties of the amino acids, namely, volume and polarity.

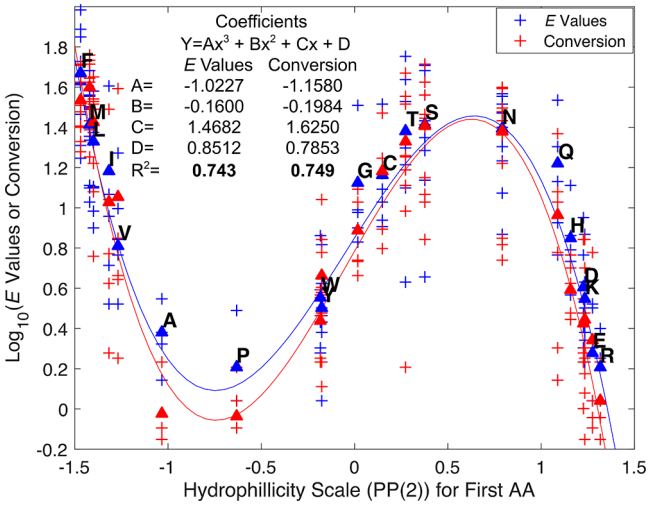


Figure 2. Cubic plots of *E* values and conversion data from the work of Feng et al.[15] to the new hydrophilicity scale for AA215 (first amino acid). The coefficients for the cubic relationships are given in the table. The correlation coefficient between $\log_{10}(E$ value) and $\log_{10}$(conversion) was 0.933.

Table 1. New Volume and Hydrophilicity Scales

		volume	hydrophilicity
	polarity scale	PP(1)[a]	PP(2)[a]
Ala	−3.11	−2.90	−1.03
Arg	3.66	2.41	1.31
Asn	1.90	−0.68	0.79
Asp	3.01	−0.92	1.23
Cys	−0.08	−1.89	0.15
Gln	2.85	0.36	1.09
Glu	3.26	0.16	1.28
Gly	−0.30	−4.04	0.01
His	3.03	0.83	1.15
Ile	−3.53	0.51	−1.32
Leu	−3.77	0.52	−1.40
Lys	3.50	0.92	1.23
Met	−4.06	0.92	−1.42
Phe	−4.06	2.22	−1.47
Pro	−1.93	−1.25	−0.64
Ser	0.70	−2.36	0.38
Thr	0.56	−1.19	0.28
Trp	−0.50	4.28	−0.18
Tyr	−0.59	2.75	−0.18
Val	−3.53	−0.65	−1.27

[a]PP(1) and PP(2) are obtained from PCA.

Testing the Physical Descriptor Scales. We report the results below using PLS regression models following the previous work on these data sets. In principle these data sets

DOI: 10.1021/acs.jcim.7b00488
J. Chem. Inf. Model. 2018, 58, 234−243

could be modeled using multilinear regression as the descriptors describing each amino acid are orthogonal and thus independent of each other (this would be principal component regression).

A note on the nomenclature around R^2 and Q^2—R^2 is a measure of the linearity of the training data (or test set data) while Q^2 is similar except that it is the sum of the measured linearity across all the independent hold out test sets. When R^2 is used to describe the quality of fit of a model to the data, it will be referred to as R^2(train). When an R^2 is quoted for an external data set not used at any stage in the modeling process, that will be R^2(test). All methods that take a portion of the training set, use the rest to build a model and then predict values for the withheld data are methods of cross validation and the predictive R^2 values obtained are called Q^2. These include Q^2 by leave one out CV (Q^2-LOO), and Q^2 by leaving a portion out (Q^2-LPO) which is the same in principle as K-fold CV. This last includes the common procedure of leave 10% out (Q^2-LPO(10%)), and Wold and co-workers[14] use a variation on this idea where the training set is split into between 5 and 10 groups; this will be referred to as Q^2-LPO(10−20%). The leave 50% out CV method described below will be referred to as Q^2-LPO(50%). Finally there is bootstrapping, which will be referred to Q^2-BS. In addition all LPO methods and bootstrapping will give a range of values (using different data splits that are generated randomly) for a given data set and model so it is important to specify the number of calculations (iterations) used to obtain a quoted mean Q^2-LPO or Q^2-BS value. In this work Q^2-LPO(10%) values are calculated using 100 iterations, Q^2-BS values are based upon 1000 iterations, and Q^2-LPO(50%) are based upon 25 or 100 iterations as described below.

Comparing PLS Models Using the Two Sets of Descriptors in a Designed Experiment. When it came to testing models using the physical descriptors and the Hellberg descriptors the combination of different numbers of descriptors with different numbers of latent variables (LVs) required a systematic method of assessing multiple models for predictive ability. To aid this, the descriptors used were split into "simple descriptors" (for dipeptides, 6 when using Hellberg descriptors, 4 with the physical descriptors); "simple + squared", 12 terms (Hellberg descriptors) or 8 terms (physical descriptors); and "full set" which includes interactions, 27 terms (Hellberg descriptors) or 14 terms (Physical descriptors). These sets are referred to as sets A, B, and C, respectively, in Figure 3 and subsequent figures. Each of these six sets of descriptors could then be run with 1−3 LVs giving the experimental design shown in Table 2. The results seen in the top panel of Figure 3 show the mean Q^2 (with 95% confidence intervals) for PLS models across multiple random 50/50 test/train splits derived from the original data set (see Methods for further detail). This method of cross validation (by analogy with the more common leave 10% out CV method) will be referred to as the leave 50% out CV method (giving Q^2-LPO(50%) values) and is potentially a much more demanding way of assessing the predictive power of a model than leave 10% out CV but one that can only be used on relatively large data sets.

For any given run in Table 2, the range of values contributing to the Q^2-LPO(50%) could be quite high. In generating random test/train splits it is always possible that a very poor split will consistently give bad results independent of which descriptors are used to model it just because (for example) all the points with high Y values end up in one set and the low Y values in the other. For example the range in the component Q^2 values for model 7 in Table 3 when analyzing the 58 ACE inhibitor dipeptides (unscreened data 25 test/train splits) was 0.27 to 0.80 highlighting that when test sets are kept aside to validate QSAR models (as is widely recommended[17]) the results can depend very much upon the specific test set chosen,

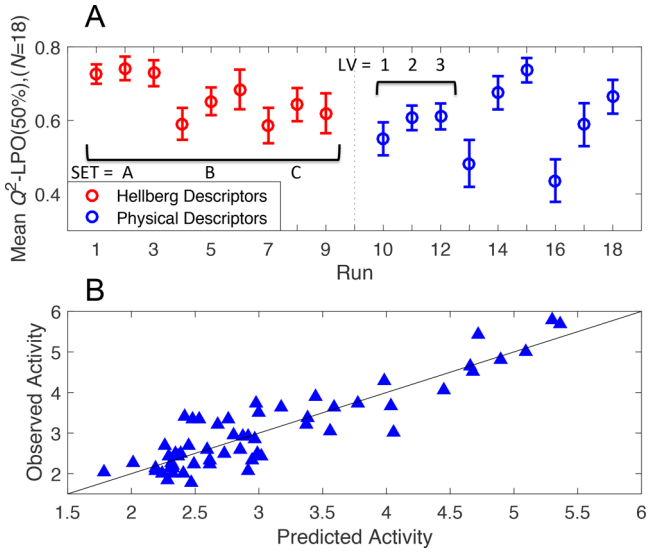


Figure 3. Analysis of the Ace inhibitor data set (58 dipeptides). (A) Assessing the different PLS models using both Hellberg and physical descriptors: (set A) simple descriptors (6 for Hellberg, 4 for physical descriptors); (set B) simple + squared terms (12 terms for Hellberg, 8 for physical descriptors); (set C) full set (simple + squared + interactions; 27 terms for Hellberg, 14 for physical descriptors); (LV) number of latent variables. The error bars show the 95% confidence limits for the mean based upon the best N 50/50 data splits. (B) Plot of observed against predicted activity for the 58 Ace inhibitor dipeptides using the model corresponding to run 15 in panel A (8 terms, physical descriptors, 3 LV). See Table 4 for details.

Table 2. Experimental Design for Assessing Alternative PLS Models

run	descriptors	set	LV[d]
1	Hellberg	A[a]	1
2	Hellberg	A	2
3	Hellberg	A	3
4	Hellberg	B[b]	1
5	Hellberg	B	2
6	Hellberg	B	3
7	Hellberg	C[c]	1
8	Hellberg	C	2
9	Hellberg	C	3
10	physical	A	1
11	physical	A	2
12	physical	A	3
13	physical	B	1
14	physical	B	2
15	physical	B	3
16	physical	C	1
17	physical	C	2
18	physical	C	3

[a]Simple descriptors (6 for Hellberg, 4 for physical). [b]Simple + squared terms (12 for Hellberg, 8 for physical). [c]Full set (simple + squared + interactions; 27 for Hellberg, 14 for physical). [d]Number of latent variables.

DOI: 10.1021/acs.jcim.7b00488
J. Chem. Inf. Model. 2018, 58, 234−243

Table 3. Results for Modeling ACE Inhibitors (N = 58) and Bitter Dipeptides (N = 48)

model	descriptors	terms	LV	R^2(train)	Q^2-CV[a,b]	R^2(test)
Ace Inhibitors Reference Model for the Whole Data Set (N = 58)						
1	Hellberg	27	3	0.839	0.67, 0.59	
2	Hellberg	12	2	0.769	0.67, 0.57	
3	Hellberg	6	1	0.774	0.74, 0.68	
4	physical	14	2	0.808	0.67, 0.57	
5	physical	14	3	0.849	0.72, 0.52	
6	physical	8	2	0.760	0.69, 0.60	
7	physical	8	3	0.814	0.75, 0.66	
8	physical	4	1	0.654	0.58, 0.57	
Fractional Factoral Design (Training Set = 9 Dipeptides, Test Set = 49)						
9	Hellberg	27	2	0.892		0.588
10	Hellberg	12	2	0.874		0.577
11	Hellberg	6	1	0.833		0.488
12	physical	14	2	0.948		0.563
13	physical	8	2	0.951		0.678
14	physical	4	1	0.664		0.680
D-Optimal Design (Training Set = 9 Dipeptides, Test Set = 49)						
15	Hellberg	27	2	0.898		0.551
16	Hellberg	12	2	0.947		0.500
17	Hellberg	6	1	0.913		0.648
18	physical	14	2	0.946		0.618
19	physical	8	2	0.912		0.721
20	physical	4	1	0.811		0.615
Bitter Dipeptides DOE (training set = 10, test set = 38)						
21	Hellberg	12	2	0.959		0.582
22	Hellberg	6	2	0.916		0.711
23	Hellberg	6	1	0.832		0.703
24	physical	8	2	0.935		0.683
25	physical	4	2	0.912		0.691
26	physical	4	1	0.858		0.694

[a]First number = Q^2-LPO(10%) 100 iterations. [b]Second number = Q^2-LPO(50%) 25 iterations.

and hence why, here, we repeat these test/train splits multiple times. In Tables 3 (top panel) and 4 we have quoted a mean Q^2-LPO(50%) based upon 25 or 100 random 50/50 test/train data splits. It should be noted that while screened data are used for the plots (to give more compact error bars) the mean values for Q^2-LPO(50%) quoted in Tables 3 and 4 use the complete set of random test/train splits without any screening.

ACE Inhibitors (Set of 58 Dipeptides).[13] A reported PLS model using the Hellberg descriptors for this data set used 6 descriptors (Z_1, Z_2, and Z_3 for two amino acid positions) and 1 latent variable (LV) and gave R^2(train) = 0.77, Q^2-LPO(10−20%) = 0.74.[14] Plots of predicted against experimental values suggested that squared terms might be important so a further model based upon 27 terms (the 6 above, plus 6 squared terms and 15 interactions) gave an improved model (3 LV), but the R^2 and Q^2 values were not reported. The first model was successfully reproduced (R^2(train) = 0.77, Q^2-LPO(10%) = 0.735), and the 27 term model with 3 LV gave R^2(train) = 0.84, Q^2-LPO(10%) = 0.67. It should be noted the authors identify the potential influence of the squared terms but then include the interaction terms as well. The high R^2 and the reduced Q^2 values suggest the data are being overfitted with the set of 27 terms.

Table 4. Best Models for Dipeptide Datasets Using the Physical Descriptors

terms	LV	R^2(train)	Q^2-LPO(10%)[a]	Q^2-LPO(50%)[a]	Q^2-BS[b]
ACE Inhibitors (N = 58)					
8	3	0.814	0.749	0.68	0.689
Bitter Dipeptides (N = 48)					
8	3	0.856	0.794	0.73	0.742
4	2	0.796	0.731	0.68	0.684
Elastase Substrate Data Set (Part) (N = 41)					
8	3	0.879	0.833	0.76	0.790
ACE Inhibitors (N = 167, One Outlier Was Dropped)					
8	3	0.727	0.692	0.68	0.667

[a]100 Iterations. [b]1000 Iterations.

Figure 3A (see section above on comparing PLS models for more information about how this figure is constructed) summarizes the predictive ability of the different models, and key data are summarized in Table 3 (top section). Figure 3 shows that the most predictive models using Hellberg descriptors are those with simple descriptors; adding in the squared terms or using the full set makes the predictive ability significantly worse. This suggests that any improvement in predictive ability from the addition of one or more squared terms (as suggested by the authors) is compromised by the fact that all the squared terms are added into the model. The lower values for the full set of terms (in Figure 3) does support the point above that the high R^2(train) value and low Q^2-LPO(10%) for the 3 LV model is due to overfitting. When the physical descriptors are employed, the best results are obtained with simple + squared terms (8 terms) with 2 or 3 LV or the full set with 3 LV(error bars overlap—see Figure 3). However, the statistics of the fits shown in Table 3 suggest that the 8 term model with 3 LV (model 7) should be used as a physical descriptor reference model for this set of dipeptides (see Tables 4 and 5 and Figure 3B).

A more demanding test of these PLS models is to split the 58 data points into a designed training set and test set. Hellberg et al. have suggested a designed training set of 9 peptides[13] and in a later publication a set of 9 peptides selected by a D-optimal algorithm.[14] A comparison of models based upon these training sets using Hellberg and physical descriptors is summarized in Table 3 (second and third panels) with the quality of the models being assessed by R^2 for the external test set, R^2(test). For the training sets containing 9 dipeptides the method of assessing possible models using Q^2-LPO (50%) values as described above is not viable and meaningful Q^2 values, even by leave-one-out (LOO) CV, could not be obtained. The plots in Figure 4 show the models for ACE inhibitors (panels A and B) using a D-optimal training set (9 dipeptides) and Bitter Dipeptides (panels C and D) using a partially designed training set of 10 dipeptides. For simplicity results are shown for models based upon the simple descriptors and 1 LV. In general the results are very good for such simple models. Note that from Table 3 it is clear that in some cases better results can be obtained by using models with squared terms included and/or more LV. So adding in squared terms into the physical descriptors model (an expansion from 4 to 8 terms) of the D-optimal training set (Figure 4B) and using 2 LV increases R^2(train) from 0.811 to 0.912 and R^2(test) from 0.615 to 0.721. A similar expansion of terms using the Hellberg descriptors (from 6 to 12) improves R^2(train) from 0.913 to 0.947, but R^2(test) is reduced from 0.648 to 0.500. The important point is that the physical descriptors (panels B and D) give very comparable models to those provided by the Hellberg descriptors (panels A and C). Although, in these plots, there

DOI: 10.1021/acs.jcim.7b00488
J. Chem. Inf. Model. 2018, 58, 234−243

Table 5. Coefficients for the Best Models using the Physical Descriptors[a]

constant	H(1)	V(1)	H(2)	V(2)	$H(1)^2$	$V(1)^2$	$H(2)^2$	$V(2)^2$
				ACE Inhibitors (N = 58)				
3.0510	−0.1880	0.0776	−0.2552	0.8328	0.1226	−0.0036	−0.3300	0.1229
				Bitter Dipeptides (N = 48)				
1.9829	−0.0597	0.2885	−0.2131	0.4460	0.0717	0.0156	−0.0036	0.1457
1.9829	−0.1054	0.3271	−0.2256	0.4414				
				Elastase Substrate (N = 41)				
0.5265	−0.0908	−0.0585	0.1184	−0.2186	−0.1856	−0.1666	−0.1785	−0.2544
				ACE Inhibitors (N = 167)				
2.0858	0.1028	0.1296	−0.2309	−0.8132	−0.0308	0.0519	0.2783	0.0150

[a]H(1) refers to the hydrophilicity scale for the first amino acid.

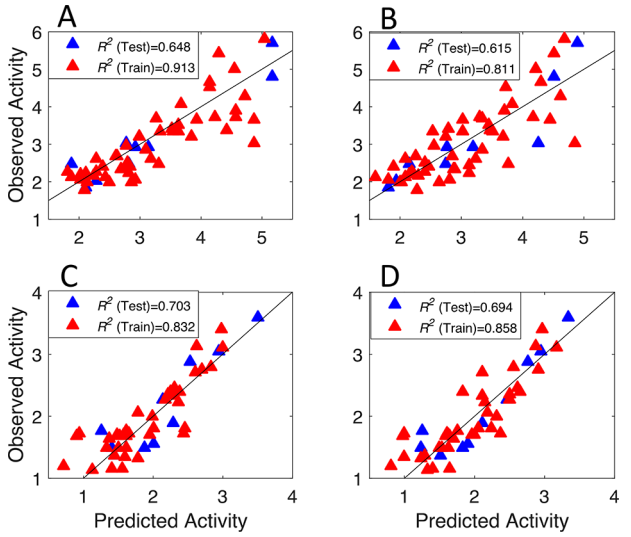


Figure 4. Prediction of a large test set from a designed training set. In each panel, the training set is shown by blue symbols, and the test set, by red symbols. The black line is $Y = X$. Details of the four models are given in Table 3. In all cases, only the simple descriptors and 1 LV have been used. (A and B) ACE inhibitor data set (58 dipeptides). The training set (9 dipeptides) was selected by the D-optimal algorithm and modeled using 6 Hellberg descriptors (panel A) or 4 physical descriptors (panel B). (C and D) Bitter dipeptides (48). Partially designed training set (10 dipeptides) modeled using 6 Hellberg descriptors (panel C) or 4 physical descriptors (panel D).

is quite a lot of random scatter of the points around the $X = Y$ line, considering the simplicity of these models and that they are based upon only 9 or 10 data points and are being used to predict 3 or 4 times that number of points, the results are generally very good.

Other authors have reported results for fitting the whole data set but not for these designed training sets. Tian et al. used a carefully selected set of 4 T-scales (out of the 10 potentially used to describe dipeptides) to model this data set and got very good results with 2 LV (R^2(train) = 0.845, Q^2-(LOO) = 0.786).[10] Mei et al. started with 16 VHSE descriptors and used stepwise multiple regression (SMR) and an external test set to select the best five which were then used in a PLS model giving (with 1 LV) R^2(train) = 0.77, Q^2-LOO = 0.745.[9] Yang et al. started with 16 ST-scale descriptors and reduced this number to 7 using SMR and got a good model (R^2(train) = 0.855, Q^2-LPO(10−20%) = 0.774), but this required a PLS model with 5 LV.[11]

Analysis of the coefficients for the three models using 8 terms based upon the physical descriptors (models 7, 13, and 19 in Table 3—see Table 5 for the coefficients of model 7) shows that the most important term is the volume of the second amino acid (always with a positive coefficient) followed by the hydrophilicity of both amino acids (both with a negative coefficients- so hydrophobic residues increase activity). For the squared terms the most important are the hydrophilicity (negative coefficient) and volume (positive coefficient) of the second amino acid. These squared terms have the same sign as the corresponding simple descriptor so reinforce the conclusion that high activity is associated with large hydrophobic amino acids in the second position and hydrophobic residues in the first position. Hellberg et al. agreed with both these points but also found that their Z_3 (electronic properties) for the first amino acid was important.[13] For the other descriptor sets it is much more difficult to relate increasing activity to specific properties of the amino acid residues. The selected descriptors for the models using T-scales and ST-scales are not reported by the authors so it is not possible to draw any comparable conclusions about how activity may be affected by amino acid properties.[10,11] In the work of Mei et al. the VHSE descriptors are related to the properties of the amino acids and the selected descriptors for the model are provided; however they do not relate well to the results above.[9] Two of the five descriptors relate to the hydrophilicity of the second amino acid which agrees with the results from Hellberg and those reported here for the physical descriptors.

They also pick out electronic properties of the first amino acid (in agreement with Hellberg) but miss completely the volume/steric effects of the second amino acid (most important term for the physical descriptors as judged by the size of the coefficients). Interpretation is more difficult for the VHSE descriptors than with the Hellberg or physical descriptors and almost impossible for T- and ST-scales.

Bitter Dipeptides (Set of 48 Dipeptides).[13,18] Hellberg et al. reported a 2 LV model based upon the 6 simple descriptors with R^2(train) = 0.83, Q^2-LPO(10−20%) = 0.76.[13,14] Figure 5A compares the models using Hellberg and physical descriptors. This plot suggests that the best model with the Hellberg descriptors would be simple descriptors with 3 LV. For the physical descriptors there are no significant differences between the various models (error bars overlap—all with mean values >0.6) except for run 13 (simple + squared descriptors, 1 LV) which gives particularly bad results. However, from the mean Q^2 values shown in panel A the best model should be either with simple descriptors with 2 LV (run 11) or simple + squared descriptors with 3 LV (run 15). Statistical data for both these models are included in Table 4, and the plot of predicted against experimental values for the simple descriptor model is shown in Figure 5B. The model using 8 terms and 3 LV has

DOI: 10.1021/acs.jcim.7b00488
J. Chem. Inf. Model. 2018, 58, 234−243

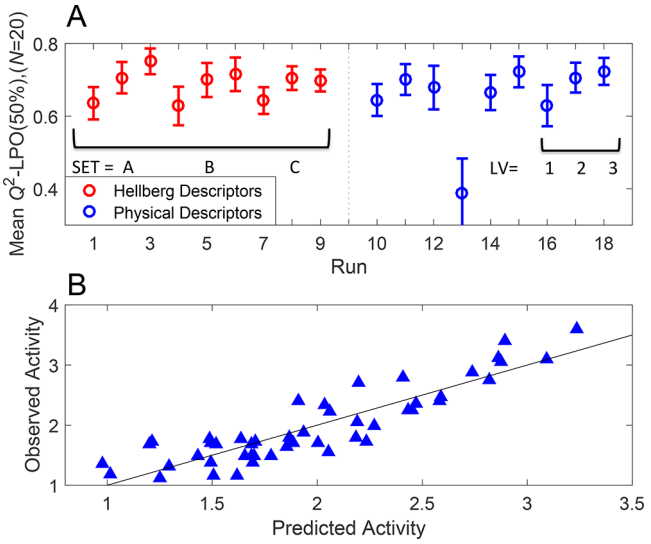


Figure 5. Analysis of the bitter dipeptides data set (48 dipeptides). (A) Assessing the different PLS models using both Hellberg and physical descriptors: (set A) simple descriptors (6 for Hellberg, 4 for physical descriptors); (set B) simple + squared terms (12 terms for Hellberg, 8 for physical descriptors); (set C) full set (simple + squared + interactions; 27 terms for Hellberg, 14 for physical descriptors); (LV) number of latent variables. The error bars show the 95% confidence limits for the mean based upon the best N 50/50 data splits. (B) Plots of observed against predicted activity for the 48 bitter dipeptides using the model corresponding to run 11 in panel A (4 new descriptors, 2 LV). Details of the model are given in Table 4.

R^2(train) = 0.856, Q^2-LPO(10%) = 0.79 while the simpler model with 4 terms and 2 LV gives R^2(train) = 0.796, Q^2-LPO(10%) = 0.73. These two models compare very favorably with the Hellberg results shown above.

Hellberg et al. attempted to use a full factorial design in Z_1 and Z_2 as a training set but could only identify 10 dipeptides (out of the required 16) that could be fitted into this design.[13] Despite this, they used this incomplete set as the training set and the remaining 38 dipeptides as a test set. They obtained a model using the simple descriptors and 2 LV with R^2(train) = 0.93, Q^2-LPO(10−20%) = 0.54 but did not report the R^2(test).[13,14] When we reproduced this model we got R^2(train) = 0.92, R^2(test) = 0.71 (see model 22 in Table 3). When using the physical descriptors, good models (based upon the R^2(test) values) were obtained with 4 or 8 terms (see models 24−26 in Table 3). A plot of predicted against experimental values for model 26 (4 simple descriptors with 1 LV; R^2(train) = 0.86, R^2(test) = 0.69) is seen in Figure 4D.

Analysis of the coefficients of the physical descriptor models for the full data set (see Table 5 for coefficients of the reference model) showed that increased bitterness was favored by having large hydrophobic amino acid residues at both positions; the improvement seen in the model with 8 terms is mainly due to the squared term for the volume of AA2. This is in good agreement with the conclusions of Hellberg and co-workers.[13]

Finally, Mei et al. obtained a very good model (R^2(train) = 0.91, Q^2-(LOO) = 0.82) for this data set using 8 selected VHSE scales and 3 LV.[9] A PLS loadings plot for the first two components (LV), explaining ~85% of the variance in the Y values (bitterness), showed that the bitterness was correlated to hydrophobicity but failed to pick up the association with large size seen with the Hellberg and physical descriptors.[9]

Elastase Substrates and Inhibitors (Set of 41 Peptides). Kinetic data[19] for 89 synthetic substrates (varied at 2 amino acid positions) for porcine pancreatic elastase have been analyzed using 5 Z-scales so that noncoded amino acids could be included. Sandberg and co-workers[7] used a set of 17 terms (selected from simple descriptors + squared terms + cross terms using Z-scales 1−4) to model the kinetic data by PLS using 4 LV and obtained R^2(train) = 0.83, Q^2-LPO(10−20%) = 0.77. In this work we analyze k_{cat} values for a subset (41 peptides) of the original data using the original Z-scales (Z_1−Z_3) and the physical descriptors. The subset of 41 peptides uses a small set of amino acids (A, F, G, I, L, P, V) which covers the range of the volume descriptor quite well but is restricted to the hydrophobic end of the hydrophilicity scale. It was found that better models were obtained for this data set by autoscaling the simple descriptors before calculating the squared and cross terms and then repeating the autoscale process on the final sets of terms. Figure 6A compares the models using Hellberg and

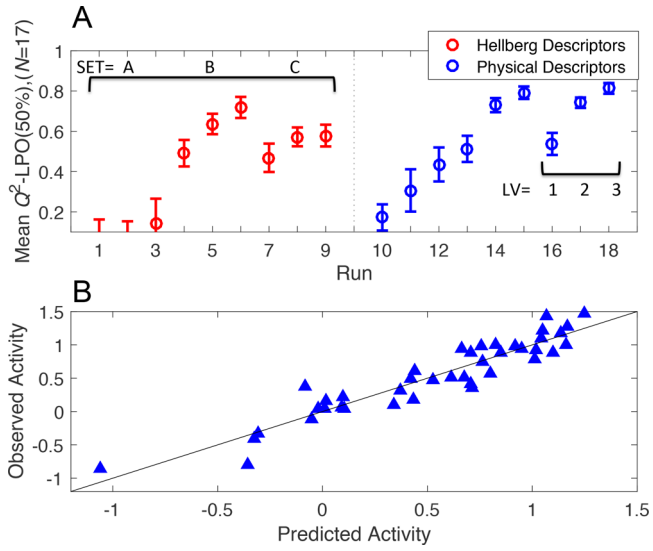


Figure 6. Analysis of the elastase substrate k_{cat} data set (41 peptides). (A) Assessing the different PLS models using both Hellberg and physical descriptors: (set A) simple descriptors (6 for Hellberg, 4 for physical descriptors); (set B) simple + squared terms (12 terms for Hellberg, 8 for physical descriptors); (set C) full set (simple + squared + interactions; 27 terms for Hellberg, 14 for physical descriptors); (LV) number of latent variables. The error bars show the 95% confidence limits for the mean based upon the best N 50/50 data splits. (B) Plots of observed against predicted k_{cat} values for the 41 elastase substrates using the model corresponding to run 15 in panel A (8 new descriptors, 3 LV). Details of the model are given in Table 4.

physical descriptors. This plot suggests that the best model with the Hellberg descriptors would be simple + squared descriptors with 3 LV. For the physical descriptors, the best model is either with the full set of terms (14) using 3 LV or with simple + squared descriptors with 3 LV (error bars overlap). The k_{cat} data were correlated using the latter model (as it is the simpler model using 8 terms) to give R^2(train) = 0.88, Q^2-LPO(10%) = 0.83 (see Table 4 and Figure 6B). Other authors[7,20] have had some success in modeling $-\log(K_m)$ as well as K_{cat} using the whole data set, but we found we were unable to get a good model for this quantity possibly because we were modeling a subset (as described above) of the original data.

DOI: 10.1021/acs.jcim.7b00488
J. Chem. Inf. Model. 2018, 58, 234−243

ACE Inhibitors (Set of 168 Dipeptides from Multiple Laboratories). This data set[21] includes the data set described previously and provides multiple values from different sources for many dipeptides. Note that in this data set the activity is given by the $\log_{10}$ of the inhibition concentration in micromolar units while in the previous data set (58 dipeptides) the activity is the negative $\log_{10}$ of the concentration in molar units. The multiple values provide some information on the error within the data set. For 19 of the dipeptides, 3 or more values are provided and from these values a combined standard deviation of 0.38 could be calculated. The mean value of the activity of these 19 dipeptides was 1.49 giving a relative standard deviation of about 25%. This shows there is a lot of scatter in the data. The worst example of this scatter was the dipeptide aspartic acid-glycine which had two values: 1.09 and 4.15. Preliminary modeling suggested that the 1.09 value was in error so this point (the only outlier) was removed from the data set. Figure 7A shows a comparison of the models using Hellberg and physical descriptors. The best model with the physical descriptors is using the simple + squared descriptors (8 terms) and 3 LV. This gave a model with R^2(train) = 0.73, Q^2-LPO(10%) = 0.69 (N = 167), and the scatter plot of observed against predicted activity is seen in Figure 7B. More statistical details for this model are given in Table 4. Wu and colleagues get a very similar model for these dipeptides after the removal of one or more outliers (it is not clear how many outliers they removed).[21] They found that high Y values (low activity) were associated with positive coefficients for hydrophilicity at both amino acid positions along with a large negative coefficient for the volume of the second amino acid. This is very similar to the results found here, and for the results reported for the first data set (58 dipeptides) although the signs of the coefficients are reversed because of the different ways the data are expressed.

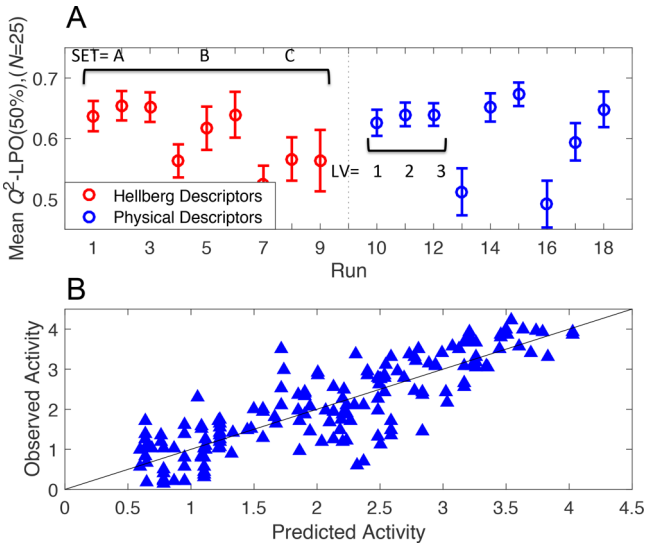


Figure 7. Analysis of the Ace inhibitor dipeptides multilaboratory data set with one outlier removed (167 dipeptides, see text). (A) Assessing the different PLS models using both Hellberg and physical descriptors: (set A) simple descriptors (6 for Hellberg, 4 for physical descriptors); (set B) simple + squared terms (12 terms for Hellberg, 8 for physical descriptors); (set C) full set (simple + squared + interactions; 27 terms for Hellberg, 14 for physical descriptors); (LV) number of latent variables. The error bars show the 95% confidence limits for the mean based upon the best N 50/50 data splits. (B) Plots of observed against predicted activity values for the 167 Ace inhibitors using the model corresponding to run 15 in panel A (8 physical descriptors, 3 LV). Details of the model are given in Table 4.

T-Test Comparison of the Hellberg and Physical Descriptors. For all four sets of models shown in Table 3, we have done a T-test (N = 9) across the experimental design shown in Table 2 based upon R^2(train) for the reference model for ACE inhibitors and R^2(test) for results using small training sets (see ACE Inhibitors (Set of 58 Dipeptides) section and Bitter Dipeptides (Set of 48 Dipeptides) section). In no case was there a significant (at the 95% level) difference between the models for the two sets of descriptors. This reflects the great variability in the R^2 values with changing number of terms in the models.

Some Results Using Some Nonphysicochemical Descriptors. In the Supporting Information, we provide plots similar to Figure 3 for two additional sets of descriptors—the MS-WHIM descriptors of Zaliani and Gancia and the T-scales of Tian et al. While both the physical descriptors and the Hellberg descriptors from which they are derived are based upon PCA of the physical and chemical properties of the amino acids, the MS-WHIM scores are based upon an analysis of steric and electrostatic properties and the T-scales are derived from structural and topological properties. The plots (Figures S1–S4) show a direct comparison between models using these alternative descriptors with models using the physical descriptors across the range of models shown in Table 2. Figures S1–S4 show that there is great variation in the cross-validation statistic (Q^2-LPO(50%)) depending upon the model being considered and the data set being analyzed; as is also seen for the comparisons between models using the Hellberg and physical descriptors presented in the main paper. In Table S3 some key results from the different models are presented (along with some results reproduced from Table 3). When the 9 ACE inhibitor dipeptides, selected by fractional factorial design, are modeled by T-scales, the R^2(test) values are surprisingly low (all below 0.6 for the three models considered- see models 36–38 in Table S1), but when a different set of 9 ACE inhibitor dipeptides (selected by a D-optimal algorithm) are used as the training set, the test set values for the same models are very much better (all above 0.6—see models 42–44 in Table S3). So based upon the fractional factorial training set, it would be concluded that the T-scales poorly model this set of ACE inhibitor dipeptides while, if the D-optimal training set was used, the conclusion would be that T-scales model the data quite well. This demonstrates how selection of specific training and test sets can have a big impact upon the apparent predictive ability of a model.

CONCLUSIONS

We present some new descriptors for amino acids with a view to using them to model the activity of dipeptides, oligopeptides and full enzymes. These physical descriptors were tested on four different peptide data sets, and all gave highly satisfactory results and have the following advantages:

(1) They have been shown to give good results when used to model the activity of 141 mutants of epoxide hydrolase (as part of their derivation).
(2) They also give good results when used to model four sets of dipeptide data.
(3) Two independent scales describe the amino acids. This is fewer than any other proposed descriptor sets (which use three or more scales to describe the amino acids) and

DOI: 10.1021/acs.jcim.7b00488
J. Chem. Inf. Model. 2018, 58, 234–243

potentially reduces the search space when using variable selection methods in building QSAM.

(4) The two independent scales are closely linked to well understood properties of amino acids: namely hydrophilicity/hydrophobicity and volume/steric properties.

Overall the use of these scales instead of alternative descriptors should provide, for a wide variety of peptides and proteins, simpler models which are relatively easy to interpret.

Although we obtained good results for the modeling of dipeptide data we had less success with the modeling of oligomeric data. One reason for this could be the relative sparseness of the data. All the dipeptide data sets considered here had a relatively high coverage of the total search space (the smallest data set, the elastase substrates, covered about 10%). However, the "curse of dimensionality" ensures that for a set of hexapeptides similar coverage would require 6.4 million measured sequences. The other reason is the problem of the correct selection of terms within a model: the variable selection method (sometimes known as feature selection), which was discussed in the Introduction and will be the subject of a future paper.

METHODS

Experimental Data. The conversion (%) and enantiomeric excess or *E* values (after 60 min of reaction) were taken from Tables SI2 and SI3 in the Supporting Information of Feng et al.[15] The values quoted are the average of three measurements, and the standard deviations provided give some indication of the accuracy of the measurements, which vary over a wide range. In general the larger *E* values are more accurate than the smaller values. As indicated above, the measured values are mainly determined by the identity of the amino acid in the first position (AA215), the AA217 amino acid having a much smaller effect on the value. On the basis of this observation, a single outlier was identified (dipeptide tyrosine−leucine, conversion = 50.7%, *E* = 120.5) and excluded from the fits as it was completely out of line with other data with tyrosine at the AA215 position (conversions 0.2−3.6%, *E* 1.1−6.6).

The dipeptide data sets used to assess the different sets of descriptors are reproduced in the Supporting Information (section S3).

Fitting Hydrophilicity and Polarity Scales to Both the *E* Values and Conversion Data of Feng et al. A polarity scale and the associated hydrophilicity scale were required to provide good fits to both of the properties measured by Feng et al.[15] An initial polarity scale based upon Hellberg's PP(1) was systematically modified in small increments using a simple genetic algorithm (GA) of our own design.[22] To the best of our knowledge no systematic studies have been performed to establish the best GA parameters to use for this type of problem. We recently reported on recommended parameters when a GA was used to direct an experimental program,[22] but the results are not applicable to this problem. For the current GA the size of the search space was limited using a granular polarity scale (smallest increment = 0.01). The initial search used 500 generations, population = 10, and the three best results go onto the next generation, Decloning (the process of systematically replacing duplicate individuals with new randomly generated individuals[23]) was not used, but each individual was subjected to two rounds of mutation. During mutation the polarity value for two randomly selected AA residues were changed by a random increment up to ±0.1 in size. The GA was run multiple times to identify a good fit to the data and the best results for the first search was used as the input for subsequent searches. The final result is a good fit to both *E* values and conversion given the constraints but not necessarily the best possible solution.

The fitness value for the GA was obtained by taking the polarity scale and three measures of volume[3,24,25] plus molecular weight[5] and applying PCA to generate five principal components of which the first two (explaining 78% and 21% of the total explained variance respectively) are a composite volume scale and the derived hydrophilicity scale. The experimental data are then fitted to this hydrophilicity scale assuming a cubic relationship and the R^2 value is returned as the fitness value for the GA. However, it soon became clear that this method was fitting some of the error. Some of the amino acids started to change relative positions and large gaps opened up around the maximum and minimum of the cubic curve. It was decided to force the amino acids to stay in a certain order (consistent with the comments on relative polarity of amino acids discussed in the Deriving the New Hydrophilicity Scale section above) by applying a penalty of −0.5 to the R^2 value if any of the rules below were broken. For the polarity scale (where hydrophobic amino acids have a value of about −4 and charged amino acids have a value of about +3.5):

(1) F ≤ L,I ≤ V ≤ A
(2) F ≤ M
(3) N,Q ≤ R,D,E,K
(4) S,T ≤ N,Q
(5) Y,W ≤ S,T

A similar penalty was imposed if a large gap was formed in the polarity scale. No information was available about what should be an acceptable maximum gap between adjacent amino acids but a good fit to the data could be obtained if the maximum gap was set at 15% of the full range of the polarity scale. If the fit was attempted using 14% then a poor fit was obtained because penalties were applied to the R^2 value as the GA evolved. A maximum gap of 16% or above generated very similar fits to the 15% case so we selected the latter value for this work, as it was more conservative.

Using these methods a good cubic fit to a different hydrophilicity scale can be made for the two Y values (*E* values and conversion). However, these two properties are quite closely correlated: a plot of $\log_{10}$(conversion) (y) against $\log_{10}$(*E* values) (x) gives a correlation with the equation of $y = 1.0832x - 0.148$ and a R^2 of 0.87. As the two polarity scales derived from the GA results are very similar, the average of the two scales gives a common polarity scale from which a common hydrophilicity scale and volume scale can be obtained by PCA as described above. The two sets of experimental data can then be refitted to this common hydrophilicity and volume scales to give the results seen in Figure 2 and Table 1. These new descriptors (physical descriptors) are compared to other sets of descriptors in the Supporting Information (Tables S1 and S2)

Comparing PLS Models Using the Two Sets of Descriptors. The upper panel in Figure 3 (and subsequent figures) shows the results for a designed set of PLS models (sets A, B, and C for Hellberg and physical descriptors using 1, 2, and 3 LVs; see Table 2) for a specific data set (in the case of Figure 3: 58 ACE inhibitor dipeptides). The data set to be modeled is randomly split 50/50 into a test and training set, and this process is repeated 25 or 100 times. Test/train splits are assessed by plotting box−whisker plots and those splits that

DOI: 10.1021/acs.jcim.7b00488
J. Chem. Inf. Model. 2018, 58, 234−243

consistently (across the design shown in Table 2) give poor results (i.e., some points 0.5 or more below the average) are removed to give reduced set of "screened" test/train splits. The PLS models specified in Table 2 are tested across these screened data and the Q^2-LPO(50%) values can be plotted as a mean with a 95% confidence limit as seen in Figure 3. Designed test/train splits in the data would be expected to give more consistent Q^2 values, but when these were generated by our preferred method (sphere exclusion method[26]) and assessed, it was found that they were very highly correlated with each other. To obtain meaningful results, it was essential that the test/train sets were independent of each other so randomly generated test/train data sets had to be used for this work.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00488.

Tables of descriptor values, results for alternative descriptors, datasets used in this paper, and references (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Phone +44-161-306-4480. E-mail roy.goodacre@manchester.ac.uk.

ORCID

Mark H. Barley: 0000-0002-4696-4206
Nicholas J. Turner: 0000-0002-8708-0781
Royston Goodacre: 0000-0003-2230-645X

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded by the Biotechnology and Biological Sciences Research Council (BBSRC) and Glaxo-SmithKline (GSK) under the Strategic Longer and Larger (sLoLa) grant initiative ref BB/K00199X/1.

■ REFERENCES

(1) Jonsson, J.; Norberg, T.; Carlsson, L.; Gustafsson, C.; Wold, S. Quantitative Sequence-Activity Models (QSAM)–Tools for Sequence Design. *Nucleic Acids Res.* **1993**, *21* (3), 733−739.

(2) Sneath, P. H. A. Relations between Chemical Structure and Biological Activity in Peptides. *J. Theor. Biol.* **1966**, *12*, 157−195.

(3) Tsai, J.; Taylor, R.; Chothia, C.; Gerstein, M. The Packing Density in Proteins: Standard Radii and Volumes. *J. Mol. Biol.* **1999**, *290* (1), 253−266.

(4) Trinquier, G.; Sanejouand, Y. H. Which Effective Property of Amino Acids Is Best Preserved by the Genetic Code? *Protein Eng., Des. Sel.* **1998**, *11* (3), 153−169.

(5) Hellberg, S.; Sjöström, M.; Skagerberg, B.; Wold, S. Peptide Quantitative Structure-Activity Relationships, a Multivariate Approach. *J. Med. Chem.* **1987**, *30* (7), 1126−1135.

(6) van Westen, G. J.; Swier, R. F.; Wegner, J. K.; Ijzerman, A. P.; van Vlijmen, H. W.; Bender, A. Benchmarking of Protein Descriptor Sets in Proteochemometric Modeling (Part 1): Comparative Study of 13 Amino Acid Descriptor Sets. *J. Cheminf.* **2013**, *5* (1), 41.

(7) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* **1998**, *41* (14), 2481−2491.

(8) Liang, G.; Li, Z. Factor Analysis Scale of Generalized Amino Acid Information as the Source of a New Set of Descriptors for Elucidating the Structure and Activity Relationships of Cationic Antimicrobial Peptides. *QSAR Comb. Sci.* **2007**, *26* (6), 754−763.

(9) Mei, H.; Liao, Z. H.; Zhou, Y.; Li, S. Z. A New Set of Amino Acid Descriptors and Its Application in Peptide QSARs. *Biopolymers* **2005**, *80* (6), 775−786.

(10) Tian, F.; Zhou, P.; Li, Z. T-Scale as a Novel Vector of Topological Descriptors for Amino Acids and Its Application in QSARs of Peptides. *J. Mol. Struct.* **2007**, *830* (1−3), 106−115.

(11) Yang, L.; Shu, M.; Ma, K.; Mei, H.; Jiang, Y.; Li, Z. ST-Scale as a Novel Amino Acid Descriptor and Its Application in QSAM of Peptides and Analogues. *Amino Acids* **2010**, *38* (3), 805−816.

(12) Seasholtz, M. B.; Kowalski, B. The Parsimony Principle Applied to Multivariate Calibration. *Anal. Chim. Acta* **1993**, *277*, 165−177.

(13) Hellberg, S.; Eriksson, L.; Jonsson, J.; Lindgren, F.; Sjöström, M.; Skagerberg, B.; Wold, S.; Andrews, P. Minimum Analogue Peptide Sets (MAPS) for Quantitative Structure-Activity Relationships. *Int. J. Pept. Protein Res.* **1991**, *37* (5), 414−424.

(14) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and Megavariate Data Analysis: Principles and Applications*; Umetrics Academy: Umea, Sweden, 2001.

(15) Feng, X.; Sanchis, J.; Reetz, M. T.; Rabitz, H. Enhancing the Efficiency of Directed Evolution in Focused Enzyme Libraries by the Adaptive Substituent Reordering Algorithm. *Chem. - Eur. J.* **2012**, *18* (18), 5646−5654.

(16) Atkins, P.; de Paula, J. *Physical Chemistry*, 10th ed.; Oxford University Press: Oxford, 2014.

(17) Golbraikh, A.; Tropsha, A. Beware of Q̂2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.

(18) Asao, M.; Iwamura, H.; Akamatsu, M.; Fujita, T. Quantitative Structure-Activity Relationships of the Bitter Thresholds of Amino Acids, Peptides, and Their Derivatives. *J. Med. Chem.* **1987**, *30* (10), 1873−1879.

(19) Nomizu, M.; Iwaki, T.; Yamashita, T.; Inagaki, Y.; Asano, K.; Akamatsu, M.; Fujita, T. Quantitative Structure-Activity Relationship (QSAR) Study of Elastase Substrates and Inhibitors. *Int. J. Pept. Protein Res.* **1993**, *42* (3), 216−226.

(20) Kimura, T.; Miyashita, Y.; Funatsu, K.; Sasaki, S. Quantitative Structure-Activity Relationships of the Synthetic Substrates for Elastase Enzyme Using Nonlinear Partial Least Squares Regression. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 185−189.

(21) Wu, J.; Aluko, R. E.; Nakai, S. Structural Requirements of Angiotensin I-Converting Enzyme Inhibitory Peptides: Quantitative Structure-Activity Relationship Study of Di- and Tripeptides. *J. Agric. Food Chem.* **2006**, *54* (3), 732−738.

(22) Barley, M. H.; Turner, N. J.; Goodacre, R. Recommendations on the Implementation of Genetic Algorithms for the Directed Evolution of Enzymes for Industrial Purposes. *ChemBioChem* **2017**, *18*, 1087−1097.

(23) Jedrzejowicz, P.; Skakovski, A. Improving Performance of the Differential Evolution Algorithm Using Cyclic Decloning and Changeable Population Size. *J. Univers. Comput. Sci.* **2016**, *22* (6), 874−893.

(24) Pontius, J.; Richelle, J.; Wodak, S. J. Deviations from Standard Atomic Volumes as a Quality Measure of Protein Crystal Structures. *J. Mol. Biol.* **1996**, *264*, 121−136.

(25) Harpaz, Y.; Gerstein, M.; Chothia, C. Volume Changes on Protein Folding. *Structure* **1994**, *2* (7), 641.

(26) Golbraikh, A.; Tropsha, A. Predictive QSAR Modeling Based on Diversity Sampling of Experimental Datasets for the Training and Test Set Selection. *J. Comput.-Aided Mol. Des.* **2002**, *16* (5−6), 357−369.