



Chain melting temperature estimation for phosphatidyl cholines by quantum mechanically derived quantitative structure property relationships

Andrew J. Holder^{1,*}, David M. Yourtee^{2,*}, Derek A. White¹, Alan G. Glaros³ & Robert Smith²
¹University of Missouri-Kansas City, Department of Chemistry, Kansas City, MO 64110, USA; ²University of Missouri-Kansas City, Schools of Pharmacy and Medicine, Kansas City, MO 64110, USA; ³University of Missouri-Kansas City, School of Dentistry, Kansas City, MO 64110, USA

MS Submitted 1 November 2002; Accepted for publication 15 November 2002

Key words: AM1, chain melting temperature, phosphatidyl choline, QSPR, quantum mechanical, semiempirical, T_m

Summary

Geometries for 62 phosphatidylcholines (PC) were optimized using the AM1 semiempirical quantum mechanical method. Results obtained from these calculations were used to calculate 463 descriptors for each molecule. Quantitative Structure Property Relationships (QSPR) were developed from these descriptors to predict chain melting temperatures (T_m) for the 41 PCs in the training set. After screening each QSPR for statistical validity, the T_m values predicted by each statistically valid QSPR were compared to corresponding T_m values extracted from the literature. The most predictive, chemically meaningful QSPR provided T_m values which agreed with literature values to within experimental error. This QSPR was used to predict T_m values for the remaining 21 PCs to provide external validation for the model. These values also agreed with literature values to within experimental error. The descriptor developed by the final QSPR was the second order average information content, a topological information-theoretical descriptor.

Introduction/Background

This effort is a peripheral outgrowth of current research in the area of non-shrinking dental restorative. In particular, the ultimate goal of this research is to develop a biocompatibility test system capable of directly measuring the biochemical impact on and/or the disruption of the molecular interactions essential for cellular membrane structure, organization, and function by a biomaterial of interest.

The proposed test methodology combines the use of cell membrane constructs, created from supra-molecular liposomes of the types developed by Luisi and Walde [1, 2, 3, 4], with quantum mechanically derived QSPRs to systematically link experimental results with appropriate molecular understanding.

Current biocompatibility tests can identify many potential toxicities from these biomaterials; however, these methods are unable to adequately describe the chemical interactions occurring at the cellular membrane level which may initiate damage. Consequently, there is a paucity of data with respect to the mechanism of the toxicity at this first line of cell defense.

To examine these mechanisms means reaching into cell structure in a way that avoids coincident reactions, that allows exploration of the effects on the specific binding forces that hold cells and their reactive molecules together, and that facilitates interpretation of the relative abilities of the biomaterials to affect these forces so as to select the least damaging based upon rational identification of chemical moieties in the structures that initiate disruptive effects.

It is well recognized that eukaryotic cells are separated from the external environment by a fluid mosaic

*To whom correspondence should be addressed. E-mail: yourtee@umkc.edu

membrane [5]. The essential structural repeating unit is the phospholipid, arranged in a bilayer about 5 nm thick. The fluidity of the membrane is a basic parameter that affects its ability to perform its biological functions. This article reflects the results of our study of the chain melting temperature (T_m) which is an important measure of membrane fluidity and an indicator of the temperature at which multilamellar liposomes are formed. This knowledge is central to developing a measurement that will define alterations in liposomal structure exerted by a biomaterial.

The T_m is perhaps more appropriately referred to as the gel/liquid-crystalline thermotropic phase transition temperature. This phase transition is accompanied by the onset of a number of highly cooperative intra- and intermolecular motions of phospholipids in the two-dimensional plane of the bilayer. In fact, enclosed multilamellar liposomes are well known to form only after the phospholipids in excess water are heated above the T_m . Consequently, it is necessary to know the T_m value of the phospholipid dispersion as a guide for designing experiments [6].

The phospholipids modeled in this study were phosphatidylcholines (PC) with various length *sn*-1 and *sn*-2 acyl chains and containing a range of degrees of unsaturation in the *sn*-2 chain. Prior to this study, models were either based on empirical data or on limited computational methods primarily confined to molecular mechanics methods [7, 8] because the molecules consisted of a prohibitively large number of atoms with respect to *ab initio*, semiempirical, or density functional quantum mechanical methods.

Advances in hardware and software permit application of the AM1 [9] semiempirical quantum mechanical method to molecules of this size for the first time. By coupling the results of the calculations with a heuristic multilinear regression algorithm, we were able to develop a quantum-mechanically derived quantitative structure property relationship (QSPR) to predict the chain melting temperature (T_m) of 65 PCs, each consisting of approximately 150 atoms.

The naming convention in this paper, PC(X):(Y) Δ^n , corresponds to the following: PC is a phosphatidylcholine with X carbons in the *sn*-1 acyl chain, carbon #1 being the carbonyl carbon, Y carbons in the *sn*-2 acyl chain, carbon #1 being the carbonyl carbon, Δ indicates *cis*- double bond(s) in the acyl chain, and *n* indicating the location of the lower numbered carbon in the double bond, i.e., in a double bond between C(9) and C(10) *n* is 9. A typical PC, PC18:24 Δ [15] in our nomenclature system can be seen in Figure 1.

Testing the QSPR against 21 unique PCs not included in the original training set provided the external validation for the correlation. The study we describe was undertaken to test the efficacy of our model and to provide a new pathway by which T_m (and potentially many other properties difficult to obtain experimentally) for phospholipids could be predicted.

Procedure

The experimental T_m data reported was obtained from literature sources that are identified in the appropriate tabulated data. Values were determined by high-resolution differential scanning calorimetry (DSC). The experimental error was $\pm 5 \text{ kcal}\cdot\text{mol}^{-1}\cdot^\circ\text{C}^{-1}$.

With respect to the computational chemistry, first the ground state minimum energy conformations were identified and optimized for each molecule using the AM1 semiempirical quantum mechanical method as implemented in AMPAC 7.0 with Graphical User Interface [10]. This proved to be a nontrivial exercise because the gas phase minimum energy conformations for most of these systems do not correspond to the geometry taken on by the PCs as membrane components. To aid in our conformational analysis, a solvent model, COSMO [11, 12, 13], was employed using water as the solvent. This model essentially minimizes or maximizes the solvent accessible molecular surface area based on the molecule's hydrophobicity or hydrophilicity respectively. Once the solvent model minimum energy conformation was identified, the calculation was repeated without the solvent model because the QSPR software is not currently capable of deconvoluting solvation energies from heats of formation. As might be expected, the solvent model minimum energy conformation corresponded to a gas phase energy minimum, albeit not the global minimum. Each fully optimized molecule was then characterized using the frequency analysis described in earlier work [14] with the requirement that the geometries be stationary points, i.e., no negative eigenvalues were present. Output suitable for use in developing the multilinear regressions using algorithms in the QSAR/QSPR program CODESSA [15] was generated from these fully optimized and characterized molecules.

After loading the training set molecules, more than 500 various types of descriptors, including topological, electrostatic, quantum mechanical, constitutional, and geometrical, were calculated for each molecule. Pre-selection of the most pertinent descriptors

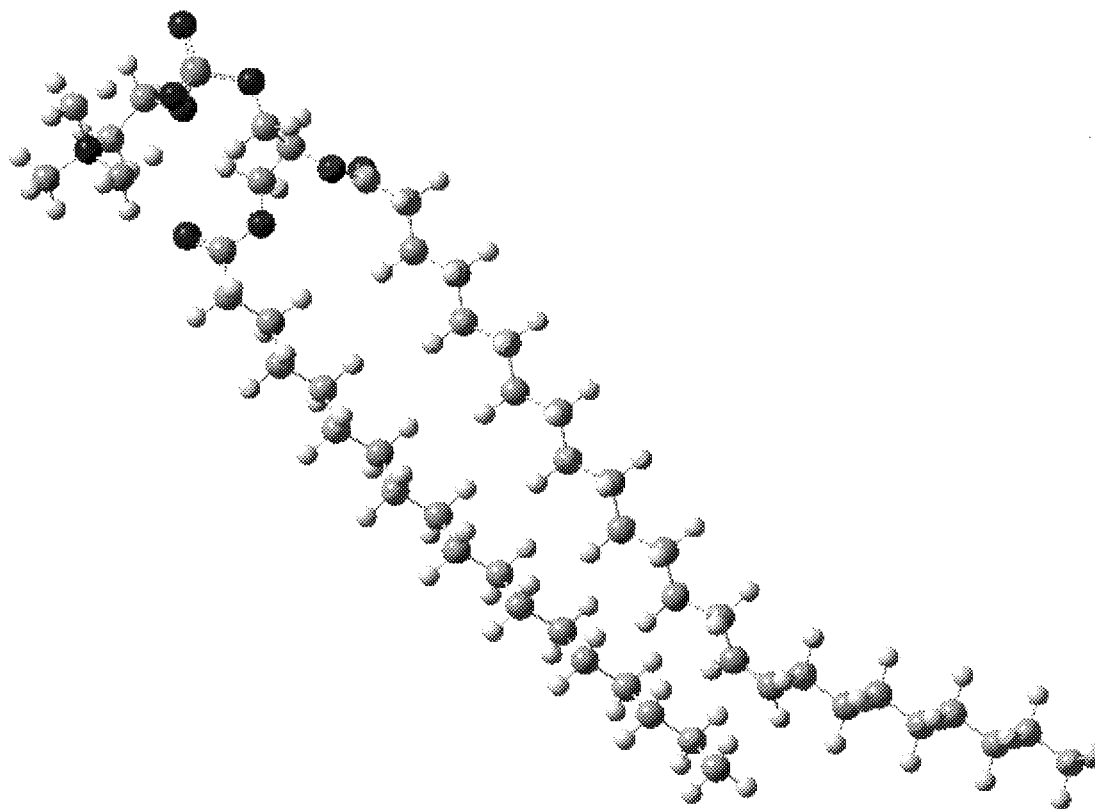


Figure 1. Phosphatidylcholine PC18:24 Δ^{15} *Note: The naming convention in this paper, PC(X):(Y) Δ^n , corresponds to the following: PC is a phosphatidylcholine with X carbons in the *sn*-1 acyl chain, carbon #1 being the carbonyl carbon, Y carbons in the *sn*-2 acyl chain, carbon #1 being the carbonyl carbon, Δ indicates *cis*- double bond(s) in the acyl chain, and *n* indicating the location of the lower numbered carbon in the double bond, i.e., in a double bond between C(9) and C(10) *n* is 9.

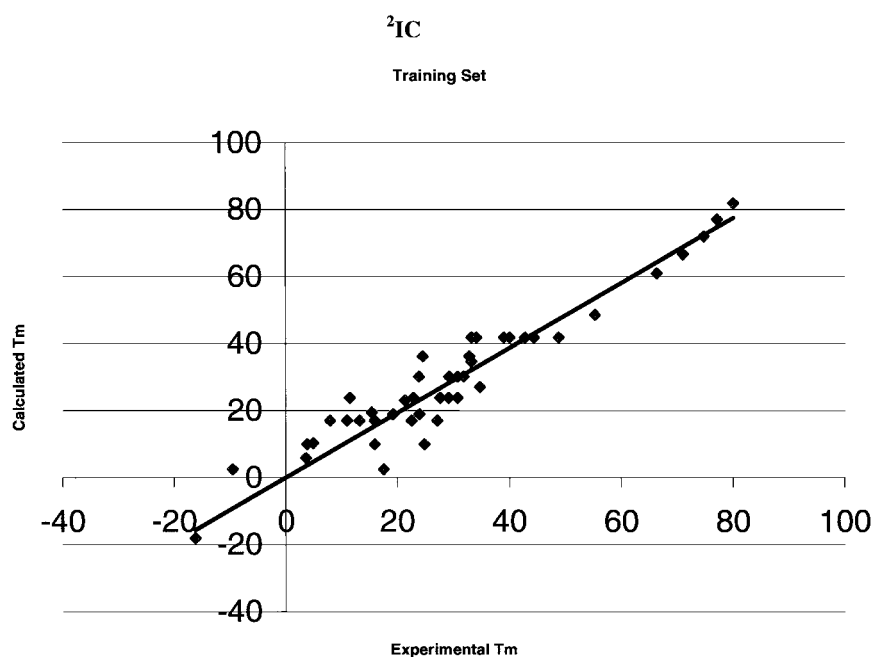


Figure 2. Calculated versus experimental chain melting temperatures for training set molecules.

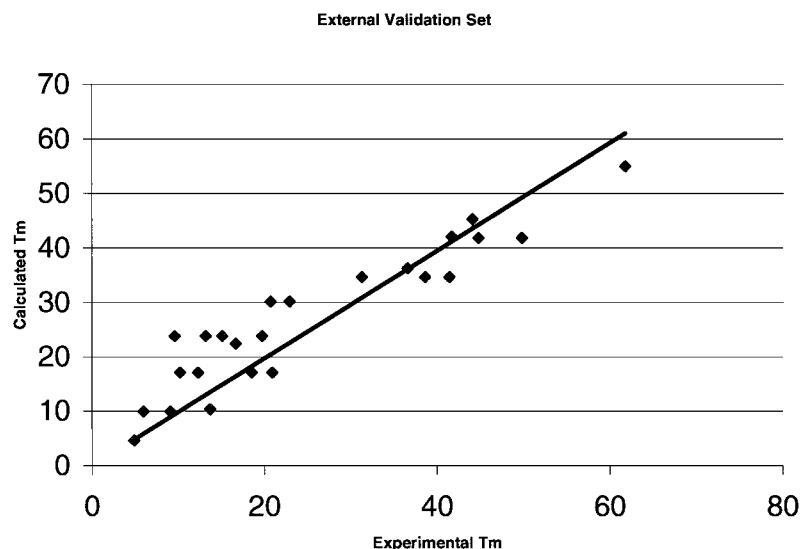


Figure 3. Calculated versus experimental chain melting temperatures for external validation set molecules

from this total was required to reduce the number to a manageable working set. These analyses were performed using CODESSA by limiting the initial pool of descriptors to include only those descriptors that existed for all molecules in the training set. Further, those descriptors whose values did not vary throughout the training set were eliminated from consideration. The remaining descriptors composed the subset used initially for regression analysis. The investigator in practice started with one descriptor and increased the number of descriptors until either there was no improvement in the quality of the correlation or the new descriptor exhibited no statistical significance. The number of descriptors was always constrained to be no greater than one third of the number of molecules in the data set with the additional caveat that the fewest number of descriptors be used consistent with providing acceptable predictive ability. Experience has shown that the number of descriptors seldom exceeds five. A heuristic method was then employed to determine the final set of descriptors developed in the multilinear regressions. In this algorithm, the initial subset of pre-selected descriptors was further reduced by eliminating a descriptor if the following user definable criteria were met: one-descriptor minimum F ratio criterion, minimum coefficient of determination, and/or significant collinearity between one or more descriptors. From the working subset of descriptors multiple multilinear regressions were developed. The ten best correlations, as determined by F ratio, were retained.

It must be noted that regression techniques are not guaranteed to locate the 'best' or 'correct' correlation equation from the descriptors and data provided. Selection of the final equation depends not just on the quality of the statistical indicators but an understanding of how the descriptors relate to chemical/physical phenomena. In this manner the equation performs its primary functions of explaining the chemical import and of providing insight into potential structural modification corresponding to the fine-tuning of molecular properties and activities.

After obtaining the ten correlations, they were subjected to a statistical analysis. Primarily, they were screened for 'goodness of fit' using adjusted R^2 values, F ratio values, variance inflation factors, and p -values all corresponding to a 95 percent confidence level. The adjusted R^2 value was calculated using the following formula:

$$R_{adj}^2 = 1 - \left[\left(\frac{n-1}{n-m-1} \right) \left(1 - \frac{\sum_i^n (\text{Prop}_{i\text{calc}} - \text{Prop}_{\text{mean}})^2}{\sum_i^n (\text{Prop}_{i\text{exp}} - \text{Prop}_{\text{mean}})^2} \right) \right]$$

where: n \equiv number of members of training set

m \equiv number of descriptors

while,

$$R^2 = \frac{\sum_i^n (\text{Prop}_{i\text{calc}} - \text{Prop}_{\text{mean}})^2}{\sum_i^n (\text{Prop}_{i\text{exp}} - \text{Prop}_{\text{mean}})^2}$$

Table 1.

Descriptor	Adjusted R ²	F Ratio	p-value	VIF
² TC	0.907	427.89	p<0.001	1

The adjusted R² value is a better measure of the proportion of variance in the data explained by the correlation than R² (especially for correlations developed using small training sets) because R² is somewhat sensitive to changes in *n* and *m*. In particular, in small samples, if *m* is large relative to *n*, there is a tendency for R² to be artificially high, i.e., for the correlation to fit the data very well. In the extreme case if *n*=(*m*+1) the correlation will fit the data exactly, i.e., R² = 1 [16]. The adjusted R² corrects for the artificiality introduced when *m* approaches *n* through the use of a penalty function which scales the result. (See Table 1) All correlations considered have an adjusted R² value > 0.89. This indicates very good agreement between the correlation and the variation in the data.

The F ratio, calculated below, is the ratio of the variance in the observed property across the training set predicted by the overall model to the variance in the observed property across the training set not accounted for by the model, e.g., variance introduced via experimental error or error in the correlation.

$$F = \frac{\left[\sum_i^n (\text{Prop}_{i\text{calc}} - \text{Prop}_{\text{mean}})^2 / m \right]}{\left[\sum_i^n (\text{Prop}_{i\text{exp}} - \text{Prop}_{i\text{calc}})^2 / n - m - 1 \right]}$$

where *n* and *m* are defined as above.

In general, the larger the magnitude of the F ratio, the better the overall model predicts the property (or activity) values in the training set. (See Table 1) All correlations considered had F ratio values of >100 which indicates that each model does an excellent job of predicting the T_m values of the training set molecules.

Variance Inflation Factors (VIF) were calculated to measure the degree of linear independence of each descriptor with respect to the other descriptors in the correlation. They were calculated using the following formula: $VIF = \frac{1}{1 - R_{\text{desc}}^2}$ where R_{desc}² corresponds to a correlation developed by setting one set of descriptor values as a property and performing a multilinear regression using the other set(s) of descriptors. This provides a quantitative basis for evaluating how the

values of each descriptor correlate to the values of the remaining descriptors. We considered a VIF<5 to be sufficient to reject linear dependence within the correlation descriptor set. (See Table 1) The VIFs of all correlations were between 1 and 3 indicating that there is less than 60 percent correlation between any two descriptors in a given correlation.

The *p*-value is defined as the smallest level of significance that would lead to rejection of the null hypothesis. We used it to determine the statistical significance of each descriptor with respect to the overall correlation. In this study, we required *p* ≤ 0.01 to be confident with a descriptor's significance and we preferred *p* ≤ 0.001. (See Table 1) None of the considered correlations had a descriptor with a *p*-value > 0.005. This insured each descriptor in a correlation made a statistically significant contribution to that correlation. Statistical analyses were performed using SPSS 10.0.5 for Windows [17].

Each remaining correlation was further evaluated by means of calculating a cross-validated R², (R_{CV}²) wherein a member of the training set was removed, the coefficients of the linear regression recalculated, and the new regression used to predict the property for the excluded member. We have found through experience that the R_{CV}² is of limited utility with a small training set, which we generally defined as less than 15–16 molecules. That is to say that when confronted with a small training set, a R_{CV}² > 0.90 did not necessarily correspond to quality of the model. This is not unusual given the relationship between the number of descriptors and the number of molecules in the training set as discussed earlier. Although the size of the training set is arbitrary with respect to the quality of the R_{CV}² value, empirical results in previous studies [18] support this conclusion. Consequently, external validation is of critical importance. The working model was selected on the basis of predictive ability, as determined by the mean signed and unsigned residual error, and on the basis of the chemical significance/meaning of the descriptor used in the correlation. A more detailed discussion of the chemical significance of the model is provided below.

Results

The QSPR that was superior in predicting the PC T_m values contained only one descriptor, the second order average information content,²IC. [19] (See Equa-

Table 2. Training set molecules

Structure	Calc. T _m	Exp. T _m	Diff.	Ref. No.
PC1321	41.8	34.1	7.7	22
PC1414	19	23.9	-4.9	23
PC1420	41.8	40	1.8	22
PC1424Δ ¹⁵	17.1	22.5	-5.4	24
PC1515	27.1	34.7	-7.6	23
PC1610	10.3	4.9	5.4	25
PC1618	41.8	48.8	-7	26
PC1618Δ ⁶	2.5	17.5	-15	24
PC1624Δ ¹⁵	23.8	27.6	-3.8	24
PC169	5.8	3.6	2.2	25
PC1810	19	19.2	-0.2	25
PC1811	23.1	21.3	1.8	25
PC1816	41.8	44.4	-2.6	25
PC1818	48.6	55.3	-6.7	23
PC1818Δ ¹¹	10	3.8	6.2	24
PC1818Δ ¹³	10	15.9	-5.9	24
PC1818Δ ⁶	10	24.8	-14.8	24
PC1818Δ ^{9,12}	-18	-16.2	-1.8	27
PC1820Δ ¹¹	17.1	13.2	3.9	24
PC1824Δ ¹⁵	30.2	31.8	-1.6	24
PC18Δ ⁹ 16	2.5	-9.5	12	28
PC18Δ ⁹ 20	17.1	15.9	1.2	29
PC1915	41.8	39	2.8	22
PC2012	34.7	33.2	1.5	25
PC2014	41.8	33.2	8.6	22
PC2018Δ ¹¹	17.1	7.9	9.2	28
PC2018Δ ¹³	17.1	15.9	1.2	28
PC2018Δ ⁶	17.1	27.1	-10	24
PC2018Δ ⁹	17.1	11	6.1	24
PC2020	61	66.4	-5.4	23
PC2020Δ ¹³	23.8	22.8	1	29
PC2020Δ ⁸	23.8	30.7	-6.9	29
PC2022Δ ¹³	30.2	29.2	1	24
PC2121	66.6	71.1	-4.5	23
PC2212	41.8	42.8	-1	25
PC2218Δ ¹¹	23.8	11.5	12.3	29
PC2218Δ ¹³	19.5	15.4	4.1	29
PC2218Δ ⁶	23.8	29.1	-5.3	24
PC2220Δ ¹³	30.2	23.8	6.4	29
PC2222	72	74.8	-2.8	23
PC2222Δ ¹³	36.3	32.8	3.5	24
PC2323	77.1	77.2	-0.1	23
PC2418Δ ⁶	30.2	30.7	-0.5	24
PC2420Δ ¹¹	36.3	24.5	11.8	24
PC2424	81.9	80.1	1.8	23

Table 3. External validation set molecules.

Structure	Calc. T _m	Exp. T _m	Diff.	Ref. No.
PC1313	10.3	13.7	-3.4	23
PC1418	34.7	38.6	-3.9	30
PC1519	41.8	44.8	-3	22
PC1616	34.7	41.4	-6.7	23
PC1622Δ ¹³	17.1	12.3	4.8	24
PC1717	41.8	49.8	-8	23
PC1814	34.7	31.3	3.4	30
PC1818Δ ¹²	10	9.1	0.9	24
PC1820Δ ¹³	17.1	18.5	-1.4	29
PC18Δ ⁹ 18	10	6	4	28
PC1919	54.9	61.8	-6.9	23
PC2018Δ ¹²	17.1	10.2	6.9	28
PC2018Δ ⁷	17.1	20.9	-3.8	29
PC2020Δ ¹¹	23.8	19.7	4.1	29
PC2024Δ ¹⁵	36.3	36.6	-0.3	24
PC2213	45.3	44.1	1.2	25
PC2218Δ ¹²	23.8	13.2	10.6	29
PC2218Δ ⁹	23.8	15.1	8.7	24
PC2220Δ ¹¹	30.2	22.9	7.3	24
PC2224Δ ¹⁵	42.1	41.7	0.4	24
PC2418Δ ⁹	30.2	20.7	9.5	24

tion 1)

$$T_m = -9.8119_2\overline{IC} + 10.6601 \quad (1)$$

The topological descriptor, ²IC, is one of a series of molecular complexity indices defined on the basis of Shannon's information theory. The order, in this case 2, corresponds to the size of the coordination sphere near a given atom. Although physical interpretation of topological indices is difficult due to their formal nature, these indices are commonly and successfully used to describe boiling points [20, 21]. Consequently, the appearance of this descriptor is reasonable. It should be noted that this index is not conformationally dependent, i.e., ²IC is a 2D descriptor. As such, quantum mechanical calculations were not required to generate the information for the descriptor used in the regression being reported. However, additional electronic descriptors, while not significantly improving the correlation statistically, were developed and should provide additional molecular design information as research progresses in this area.

When evaluating models for chemical significance the tendency is to add more descriptors in order to glean more information about the system of interest. Unfortunately, as discussed earlier, additional

descriptors do not necessarily increase the predictive ability of the model. In this instance, we developed models with one, two, three, and four descriptors. However, every model we developed contained the descriptor listed above and the addition of a second, third and fourth descriptor did not improve the predictive ability of the model in any meaningful way. In addition, each descriptor over one had very little statistical significance, i.e., p -values outside our acceptable range, $p \leq 0.01$. Consequently, we believe this single descriptor model is the most predictive model with respect to the T_m of the phosphatidylcholines in the study.

The T_m values calculated using our QSPR for the training set, listed in Table 2, are in agreement with experimental results, corresponding to an average unsigned residual error of 4.2°C . This error is within the range of experimental error, $\pm 5 \text{ kcal}\cdot\text{mol}^{-1}\cdot^\circ\text{C}^{-1}$, found with determining T_m from DSC.

That there was a good distribution of T_m values in the training set can be seen in Figure 2. The points are plotted with respect to the $x = y$ line.

The calculated T_m values for the PCs in the external validation set, listed in Table 3, are in good agreement with experimental results as well. The average signed residual error is 1.1°C and the average unsigned residual error is 4.7°C .

The 21 PCs in the external validation set are depicted on the plot in Figure 3. The R^2 for the least squares fit to the $x = y$ line is 0.8744. That the value for R^2 in the external validation set is close to the value of R^2 for the training set is a qualitative indication that the correlation is predictive although there is no direct quantitative relationship.

Subsequent investigation has revealed two shortcomings in this correlation. First, this correlation does not differentiate between molecules which differ only in the location of the point of unsaturation, i.e., PC1818 Δ^{11} vs. PC1818 Δ^{13} . Second, this correlation does not distinguish consistently between molecules having different chain lengths but the same total number of carbons in the two acyl chains, i.e., PC1420 vs. PC1618 vs. PC1816. While this limits applicability, the quality of the model for these species is still quite good.

Conclusions

This study has given us confidence that the AM1 semiempirical quantum mechanical model can be used

to successfully optimize and characterize molecules of approximately 150 atoms in a timely manner. The use of solvent models in conjunction with the AM1 method is a valuable tool in identifying solution phase geometries when those are the chemically significant conformations. The QSPRs developed using AM1 output files can be used to predict phosphatidylcholine T_m to within experimental error. This provides an experimental design tool and a springboard into the study of additional biologically significant membrane properties like critical micelle concentrations and flip-flop rates.

Acknowledgements

University of Missouri - Kansas City
Semichem, Inc.

References

1. Luisi, P.L., Bachmann, P.A., Walde, P., Lang, J., J. Am. Chem. Soc. 113 (1991) 8204.
2. Walde, P. In *Self-Production of Supramolecular Structures*; Fleischaker, G.R. et al. Eds.; Kluwer Academic Publishers: Netherlands, 1994; pp. 209–216.
3. Luisi, P.L., Walde, P., Blochli, E., Blocher, M., J. Phys. Chem. B 102 (1998) 10383.
4. Luisi, P.L., Walde, P., Blocher, M., Liu, D., Chimia 54 (2000) 52.
5. Singer, S.J., Nicholson, G.L., Science 175 (1972) 720.
6. Huang, C., Biochemistry 30 (1991) 26.
7. Gaber, B.P., Nagumo, M., Light, W.R., Chandrasekhar, I., Pattabiraman, N., Adv. Exp. Med. Biol. 238 (1988) 1.
8. Esposito, D., Zloh, M., Gibbons, W.A., Biochem. Soc. Trans. 26 (1998) S35.
9. Dewar, M.J.S., Ziebesch, E.G., Healy, E.F., Stewart, J.J.P., J. Am. Chem. Soc. 107 (1985) 3902.
10. AMPAC with Graphic User Interface, Version 7.0 Semichem, Web www.semichem.com, Shawnee Mission, KS.
11. Klampert, A., J. Chem. Soc. Perkin 2 (1993) 799.
12. Klampert, A., J. Phys. Chem. 99 (1995) 2224.
13. Klampert, A., J. Phys. Chem. 100 (1996) 3349.
14. Holder, A.J., White, D.A., Harris, C.D., Eick, J.D., Chappelow, C.C., Theochem 541 (2001) 159.
15. CODESSA Version 2.63 Semichem, Web www.semichem.com, Shawnee Mission, KS.
16. Jobson, J.D. Applied Multivariate Data Analysis, Vol. 1: Regression and Experimental Design, Springer-Verlag, New York, NY, 1991.
17. SPSS Version 10.0.5 for Windows SPSS, Inc. Web www.spss.com/spss10/, Chicago, IL.
18. Yourtee, D.M., Holder, A.J., Smith, R., Morrill, J.A., Kostoryz, E., Brockman, W., Glaros, A., Chappelow, C.C., Eick, J.D., J. Biomat. Sci. Poly. Ed. 12 (2001) 89.
19. Basak, S.C., Harris, D.K., Magnuson, V.R., J. Pharm. Sci. 73 (1984) 429.

20. Mihalic, Z., Nikilic, S., Trinajstić, N., *J. Chem. Inf. Comput. Sci.* 32 (1992) 28.
21. Needham, D.E., Wei, I.C., Seybold, P.G., *J. Am. Chem. Soc.* 110 (1988) 4186.
22. Huang, C., *Biochemistry* 30 (1991) 26.
23. Lewis, R.N.A.H., Mak, N., McElhaney, R.N., *Biochemistry* 26 (1987) 6118.
24. Wang, Z., Lin, H., Li, S., Huang, C., *J. Biol. Chem.* 270 (1995) 2014.
25. Xu, H., Huang, C., *Biochemistry* 26 (1987) 1036.
26. Wang, Z., Lin, H., Huang, C., *Biochemistry* 29 (1990) 7063.
27. Coolbear, K., Berde, C., Keough, K., *Biochemistry* 22 (1983) 1466.
28. Davis, P., Fleming, B., Coolbear, K., Keough, K., *Biochemistry* 20 (1981) 3633.
29. Davis, P., Keough, K., *Biochim. Biophys. Acta.* 778 (1984) 305.