Prediction of Peptide Ion Collision Cross Sections from Topological Molecular Structure and Amino Acid Parameters

Philip D. Mosier, Anne E. Counterman, and Peter C. Jurs*

Department of Chemistry, 152 Davey Laboratory, The Pennsylvania State University, University Park, Pennsylvania 16802

David E. Clemmer

Department of Chemistry, Indiana University, Bloomington, Indiana 47405

Quantitative structure−property relationships (QSPRs) have been developed to predict the ion mobility spectrometry (IMS) collision cross sections of singly protonated lysine-terminated peptides using information derived from topological molecular structure and various amino acid parameters. The primary amino acid sequence alone is sufficient to accurately predict the collision cross section. The models were built using multiple linear regression (MLR) and computational neural networks (CNNs). The best MLR model found contains six descriptors and predicts 94 of 113 peptides (83%) to within 2% of their experimentally determined values. The best CNN model using the same six descriptors predicts 105 of the 113 peptides (93%) to within 2% of their experimentally determined values. The best overall CNN model, using a different set of six descriptors, predicts 109 of the 113 peptides (96%) to within 2% of their experimentally determined values. In addition, this model can discriminate among peptides having identical amino acid composition, but differing in primary amino acid sequence. This represents a capability not found in previously described models. The descriptors used in the models presented may provide some insight into the nature of peptide ion folding in the gas phase.

INTRODUCTION AND THEORY

Ion mobility spectrometry (IMS)[1] is an analytical technique that is used extensively for the detection of trace amounts of analytes in the gas phase. Portable IMS instruments that can continuously monitor the surrounding air are commonly used to detect explosives at airports, chemical warfare agents on the battlefield, and toxins in the atmosphere. With the advent of gentler new ionization techniques, such as matrix-assisted laser desorption and ionization (MALDI)[2−4] and electrospray ionization (ESI),[5] IMS is no longer limited to the analysis of small or inherently gaseous molecules. Recent improvements in instrument sensitivity have made the analysis of complex mixtures commonplace.[6,7] One such example[8] discussed in this paper couples IMS with time-of-flight mass spectrometry (TOFMS) to separate large peptide libraries and protein digests and simultaneously identify the individual constituents. This method has been used to successfully identify synthetic failures in combinatorial libraries[9,10] and to examine polyalanine peptide conformations.[11]

In ion mobility spectrometry, ions are pulsed into one end of a drift tube of length, L, that is filled with an inert buffer gas and across which has been applied an electric field, E. An ion will move at an apparent constant velocity through the tube and arrive at the detector at the opposite end of the tube in drift time t_D. Ions separate in the drift tube according to size, shape, and charge state. The ratio of the drift velocity, v_D, to the applied electric field is known as the ion mobility, K, shown in eq 1.

$$K = \frac{L}{t_D E} = \frac{v_D}{E} \qquad (1)$$

It is more common to report a standardized form of the ion mobility known as the reduced ion mobility, K_0. In eq 2, P and T are the pressure and temperature of the buffer gas.

$$K_0 = K \frac{P}{760} \frac{273.3}{T} \qquad (2)$$

Since this paper addresses the role that molecular structure plays in ion mobility spectrometry, a link between the ion mobility, K, and molecular structure is sought. Revercomb and Mason[12] have

[1] Hill, H. H.; Siems, W. F.; St. Louis, R. H.; McMinn, D. G. *Anal. Chem.* **1990**, *62*, 1201A−1209A.
[2] Karas, M.; Bachmann, D.; Bahr, U.; Hillenkamp, F. *Int. J. Mass Spectrom. Ion Processes* **1987**, *78*, 53−68.
[3] Karas, M.; Hillenkamp, F. *Anal. Chem.* **1988**, *60*, 2299−2301.
[4] Tanaka, K.; Waki, H.; Ido, Y.; Akita, S.; Yoshida, Y.; Yoshida, T. *Rapid Commun. Mass Spectrom.* **1988**, *2*, 151−153.
[5] Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* **1989**, *246*, 64−71.
[6] Hoaglund, C. S.; Valentine, S. J.; Sporleder, C. R.; Reilly, J. P.; Clemmer, D. E. *Anal. Chem.* **1998**, *70*, 2236−2242.
[7] Srebalus, C. A.; Li, J.; Marshall, W. S.; Clemmer, D. E. *Anal. Chem.* **1999**, *71*, 3918−3927.
[8] Valentine, S. J.; Counterman, A. E.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 1188−1211.
[9] Srebalus, C. A.; Li, J.; Marshall, W. S.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **1999**, *11*, 352−355.
[10] Srebalus-Barnes, C. A.; Clemmer, D. E. *Anal. Chem.* **2001**, *73*, 424−433.
[11] Counterman, A. E.; Clemmer, D. E. *J. Am. Chem. Soc.* **2001**, *123*, 1490−1498.

10.1021/ac0112059 CCC: $22.00
Published on Web 02/14/2002

derived the needed relationship. In eq 3, K is the ion mobility, k_b is the Boltzmann constant, z is the charge of the ion, e is the electronic charge, m_I and m_B are the mases of the ion and the buffer gas, N is the neutral number density, and Ω is the collision cross section.

$$K = \frac{(18\pi)^{1/2}}{16}\frac{ze}{(k_bT)^{1/2}}\left(\frac{1}{m_I} + \frac{1}{m_B}\right)^{1/2}\frac{1}{N}\frac{1}{\Omega} \qquad (3)$$

It is the collision cross section that represents the ion's structural features in eq 3 and in ion mobility spectrometry in general. Substituting eqs 1 and 2 into eq 3 and solving for Ω will produce a formula that allows the experimental determination of collision cross sections. This is shown in eq 4.

$$\Omega = \frac{(18\pi)^{1/2}}{16}\frac{ze}{(k_bT)^{1/2}}\left(\frac{1}{m_I} + \frac{1}{m_B}\right)^{1/2}\frac{t_DE}{L}\frac{760}{P}\frac{T}{273.3}\frac{1}{N} \qquad (4)$$

All of the parameters, E, L, P, T, and t_D, can be precisely measured, resulting in excellent reproducibility of the measured collision cross sections.[6]

This paper presents several quantitative structure−property (QSPR) models that predict the collision cross sections of a set of 113 singly protonated lysine-terminated peptides. Certain features of the peptides' molecular structure, represented by descriptors that encode the topology of the peptide and various properties of the constituent amino acids, were found to correlate extremely well with the experimentally determined collision cross sections. These descriptors give us clues regarding the nature of peptide ion folding in the gas phase. Previous studies from this group to model the ion mobility of small organic compounds[13,14] have shown that the QSAR methodology presented here is effective in determining parameters related to IMS. Shvartsburg et al.[15] also studied prediction of ion mobility and collision cross section using methods such as projection approximation, exact hard-spheres scattering, trajectory calculations, and scattering on electron density isosurfaces with and without trajectory calculations. These give good quality results. However, they require three-dimensional modeling of the analyte, a time-consuming and error-prone process for peptide ions. An advantage of the models presented here is that no explicit three-dimensional information regarding the conformation of the peptides is needed. Recently, predictions of collision cross sections have been attempted for peptide ions using intrinsic size parameters for individual amino acids.[16,17] These also give good results but lack the ability to distinguish between peptide sequence isomers. Some of the topological descriptors used in the models presented here do not have this limitation, and thus, the resulting models are able to discriminate among sequence isomers.

EXPERIMENTAL METHOD

The 113 singly protonated, lysine-terminated peptides used in this study were selected from a database of 660 peptides whose ion mobility collision cross sections have been reported.[8] The peptides were selected using several criteria. Only singly protonated peptides were chosen, because there was only a very loose correlation between the collision cross sections of singly protonated peptides and the collision cross sections of the same peptides doubly protonated. There was no linear transformation that could be applied that would convert a singly protonated collision cross section into the corresponding doubly protonated collision cross section. Lysine-terminated peptides were chosen to maintain consistency. It was expected that the positive charge would reside on the basic lysine residue, and to help ensure positive charge localization, peptides containing other residues, such as histidine and arginine, were removed from consideration. Finally, multiple collision cross section measurements were recorded for each selected peptide so that a measure of variability for each peptide's cross section was available. A list of these peptides, along with their associated experimental cross sections, is presented in Table 1. The collision cross sections of the 113 peptides range from 140.18 to 267.51 Å^2. The median standard deviation of the measurements is 0.96 Å^2. All but one of the peptides (TVGGK) have experimental errors that are less than 2% of their experimentally determined value. The peptides used in this study were generated from a tryptic digest of some common proteins, and their associated collision cross sections were determined using an injected-ion geometry IMS-TOFMS instrument and software that is described in detail elsewhere.[6,18−20]

For the development of linear models, the set of 113 peptides was divided into a 100-member training set and a 13-member prediction set. The training set is used to build candidate models, and the prediction set is used to validate the models by ensuring that they have good predictive ability. The compounds representing the highest and lowest values of the dependent variable and each of the independent variables were included in the training set. For the purposes of this study, a peptide was considered to be an outlier if it had a predicted value that differed by more than 2% of its measured collision cross section. Linear models were selected primarily on the basis of the number of outliers and secondarily on root-mean-square errors (RMSE).

Nonlinear models were built using computational neural networks (CNNs). A 10-member cross-validation set, used to prevent overtraining of the network, was removed from the training set. This set up a 90-member training set, a 10-member cross-validation set, and a 13-member prediction set.

All computations were performed on a DEC model 500au Personal Workstation running the OSF/1 UNIX operating system. Software written in-house, including the Automated Data Analysis and Pattern Recognition Toolkit (ADAPT)[21,22] and code imple-

[12] Revercomb, H. E.; Mason, E. A. *Anal. Chem.* **1975**, *47*, 970−983.

[13] Wessel, M. D.; Jurs, P. C. *Anal. Chem.* **1994**, *66*, 2480−2487.

[14] Wessel, M. D.; Sutter, J. M.; Jurs, P. C. *Anal. Chem.* **1996**, *68*, 4237−4243.

[15] Shvartsburg, A. A.; Siu, K. W. M.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **2001**, *12*, 885−888.

[16] Counterman, A. E.; Clemmer, D. E. *J. Am. Chem. Soc.* **1999**, *121*, 4031−4039.

[17] Valentine, S. J.; Counterman, A. E.; Hoaglund-Hyzer, C. S.; Clemmer, D. E. *J. Phys. Chem. B* **1999**, *103*, 1203−1207.

[18] Hoaglund, C. S.; Valentine, S. J.; Clemmer, D. E. *Anal. Chem.* **1997**, *69*, 4156−4161.

[19] Valentine, S. J.; Counterman, A. E.; Hoaglund, C. S.; Reilly, J. P.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 1213−1216.

[20] Henderson, S. C.; Valentine, S. J.; Counterman, A. E.; Clemmer, D. E. *Anal. Chem.* **1999**, *71*, 291−301.

[21] Stuper, A. J.; Brügger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; John Wiley & Sons: New York, 1979.

[22] Jurs, P. C.; Chow, J. T.; Yuan, M. In *Computer-Assisted Drug Design*; Olson, E. C., Christoffersen, R. E., Eds.; The American Chemical Society: Washington, DC, 1979; Vol. 112, pp 103−129.

Table 1. 113 Singly Protonated, Lysine-Terminated Peptides Used in This Study

		experimental, (Å^2)[c]		prediction, (Å^2)[c]		
ID	sequence[a]	cross section[b]	error	type I	type II	type III
1	AAWGK	157.36	(1.02)	157.96	157.68	155.97
2	Ac-GDVEK	163.18	(1.29)	160.15	159.93	161.64
3	ADLAK	159.31	(1.06)	157.47	156.67	156.36
4	AFDEK	168.36	(0.44)	167.31	165.53	167.18
5	AIAEK	160.73	(1.84)	161.24	161.75	160.06
6	APNAK	147.31	(0.98)	151.63*	148.09	149.35
7	AWGGK	152.16	(2.53)	152.75	152.92	152.66
8	DIAAK	155.37	(2.98)	157.03	156.28	155.12
9	DLLFK	183.11	(3.14)	182.39	183.01	183.98
10	FFSDK	172.73	(2.52)	175.83	172.49	171.48
11	GGNMK	147.44	(1.24)	146.93	146.76	147.42
12	GITWK	169.34	(0.58)	172.15	172.23	170.24
13	GTFAK^{CV}	153.97	(0.91)	156.92	155.37	156.26
14	IFVQK	181.96	(0.51)	182.17	182.28	184.41
15	IIAEK	172.54	(0.50)	171.64	173.08	174.24
16	LDALK	172.36	(1.48)	168.30*	169.29	169.34
17	NLNEK	167.97	(0.94)	170.38	170.98	169.27
18	NTYEK	175.90	(2.66)	172.77	171.26*	175.27
19	TAWEK^{P}	170.03	(1.25)	173.56*	172.42	172.79
20	TGQIK	157.62	(0.28)	160.75	160.35	156.89
21	TLTGK	157.34	(2.11)	157.46	157.36	158.53
22	TPGSK	145.50	(1.17)	146.88	144.16	147.58
23	TVGGK	140.18	(3.52)	143.52*	143.92*	139.88
24	YYPLK	187.31	(0.38)	185.69	188.20	189.01
25	AAAAEK	160.38	(1.06)	163.76*	161.24	161.63
26	ANIDVK	176.78	(0.76)	180.20	179.15	179.90
27	ASEDLK	175.16	(0.58)	177.98	175.69	178.64
28	EAMAPK^{P}	176.19	(0.88)	175.54	175.66	176.25
29	EMPFPK	193.26	(1.43)	192.65	192.02	195.28
30	IEEIFK^{P}	197.00	(2.97)	205.23*	204.28*	203.58*
31	IVAPGK	173.33	(1.17)	171.28	171.44	173.01
32	LIFAGK^{P}	186.10	(1.12)	187.28	188.88	188.42
33	LVEDLK^{CV}	192.17	(1.05)	193.70	192.58	192.43
34	MQIFVK	203.90	(0.59)	203.84	204.12	203.58
35	NDIAAK	173.78	(0.84)	174.29	173.31	172.76
36	NLDNLK^{CV}	192.40	(0.69)	189.91	192.03	190.15
37	NVPLYK	195.27	(0.90)	195.07	197.79	196.25
38	NYQEAK^{CV}	191.16	(0.92)	189.41	191.13	190.04
39	TEAEMK^{CV}	182.54	(2.25)	182.89	183.67	185.80
40	TPVSEK	175.98	(0.73)	177.50	176.99	180.10*
41	YLTTLK	197.94	(0.70)	200.21	201.06	197.35
42	Ac-SIPETQK	205.42	(1.31)	208.67	208.44	208.39
43	APVDAFK^{CV}	189.05	(0.76)	195.47*	197.69*	192.58
44	ATDEQLK^{CV}	205.89	(0.16)	201.18*	202.79	202.09
45	ATEEQLK	206.40	(1.81)	205.39	205.58	207.09
46	DGADFAK	185.83	(0.90)	183.70	185.42	182.92
47	DSAIMLK	203.77	(1.32)	203.47	203.77	199.77
48	ELTEFAK	209.40	(0.69)	211.97	210.62	209.34
49	EVTEFAK^{P}	202.91	(0.29)	207.04*	206.83	206.47
50	FNDLGEK^{P}	206.73	(1.10)	203.10	203.81	203.36
51	GDVAFVK	200.45	(1.77)	193.14*	195.23*	195.23*
52	GGVVGIK	175.87	(0.62)	176.64	175.58	175.40
53	IATAIEK	202.92	(0.37)	201.25	202.52	202.93
54	ILLSSAK	202.97	(0.75)	204.19	204.33	203.51
55	IVTDLAK	204.75	(1.40)	203.38	203.75	202.69
56	IVTDLTK	207.02	(0.28)	207.17	206.29	206.97
57	LVTDLTK^{CV}	205.76	(0.98)	207.61	206.62	207.12
58	MIFAGIK^{CV}	207.13	(0.39)	208.51	208.65	205.73
59	MLTAEEK	209.77	(0.59)	206.63	206.30	205.80
60	NPDPWAK	198.09	(0.38)	201.78	200.66	199.82
61	VAAALTK	190.32	(0.73)	190.61	194.09	192.78
62	VADALTK	194.69	(1.73)	192.99	195.96	195.83
63	VDPVNFK	208.72	(0.91)	205.73	206.48	206.90
64	VLAAVYK	206.91	(1.08)	206.51	206.11	205.33
65	VLPVPQK	206.94	(0.57)	206.29	207.57	206.03
66	VLSAADK	193.03	(1.28)	190.22	191.49	192.79
67	VLSPADK	196.51	(1.02)	193.24	196.11	195.38
68	VLTSAAK	194.37	(0.13)	191.63	192.76	192.78
69	VSEALTK^{P}	198.59	(2.27)	198.22	198.35	200.75
70	VVTDLTK^{CV}	202.35	(0.99)	202.68	202.72	204.15
71	WNMQNGK	206.26	(0.19)	208.76	207.68	205.71
72	AADALLLK	223.86	(2.78)	217.87*	223.93	222.89

Table 1. (Continued)

ID	sequence[a]	experimental, (Å^2)[c] cross section[b]	experimental, (Å^2)[c] error	prediction, (Å^2)[c] type I	prediction, (Å^2)[c] type II	prediction, (Å^2)[c] type III
73	ADFAEISK	218.06	(0.55)	217.46	216.31	216.80
74	ADFAEVSK	214.17	(0.67)	212.97	211.99	213.12
75	ADFTDVTK	214.56	(1.13)	215.32	215.34	215.65
76	ADFTEISK	219.61	(0.21)	221.25	220.44	220.13
77	ALQASALK	218.92	(0.62)	213.20*	218.41	216.34
78	DIVGAVLK	206.07	(0.54)	213.19*	210.88*	212.60*
79	DLGEENFK	223.98	(1.41)	222.59	220.14	222.43
80	DLGEQYFK	232.22	(0.30)	232.97	234.63	232.28
81	DSADGFLK	209.30	(1.09)	208.47	208.01	209.49
82	EYEATLEK	229.28	(0.93)	231.16	231.28	229.68
83	FGVNGSEK^{P}	201.76	(0.59)	204.51	204.55	205.44
84	GASIVEDK	205.22	(1.78)	204.06	203.81	206.59
85	IDALNENK	225.09	(1.45)	221.89	225.51	228.40
86	IGDYAGIK	210.60	(0.63)	209.52	209.72	211.63
87	LIVTQTMK	243.91	(0.71)	233.99*	234.71*	241.07
88	TYETTLEK	239.55	(1.30)	232.16*	237.90	238.01
89	VLTPDLYK	230.73	(1.13)	233.54	234.40	234.78
90	YLGEEYVK	238.79	(1.26)	235.26	239.28	239.96
91	AAVTAFWGK	237.76	(0.53)	233.39	235.90	235.59
92	AAVTGFWGK^{P}	233.00	(0.18)	228.18*	233.70	233.87
93	ANELLINVK	249.69	(0.86)	249.16	247.43	246.44
94	EAVLGLWGK	235.37	(1.14)	238.55	237.48	234.24
95	FMMFESQNK^{P}	259.37	(4.27)	261.07	257.29	257.42
96	FQPLVDEPK	255.89	(0.74)	249.26*	251.14	253.85
97	MFLGFPTTK	250.02	(1.45)	249.06	249.36	251.41
98	MFLSFPTTK^{P}	255.16	(0.70)	255.30	253.04	255.00
99	QSALAELVK^{P}	234.39	(1.35)	238.60	237.49	234.46
100	QTALVELLK	245.01	(1.53)	252.21*	248.34	246.41
101	QTALVELVK	242.43	(0.50)	247.28	243.62	242.54
102	SAVTALWGK	231.09	(1.09)	231.16	231.71	232.28
103	SLVSGLWGK	236.29	(0.69)	234.01	235.76	236.17
104	TFQSFPTTK	245.22	(0.25)	247.02	249.27	248.36
105	AQSDFGVDTK	241.43	(1.36)	240.67	243.59	242.79
106	DGAGDVAFVK	229.25	(0.61)	229.03	228.92	230.18
107	LVNELTEFAK	267.51	(1.38)	271.79	263.23	263.40
108	LVNEVTEFAK	262.66	(0.85)	266.86	260.64	261.10
109	SEEEYPDLSK	257.78	(0.49)	260.64	258.09	256.48
110	TAAYVNAIEK	246.52	(2.11)	253.63*	252.48*	250.13
111	VLDSFSNGMK^{P}	252.25	(1.41)	251.95	253.19	252.68
112	VLNSFSDGLK	255.16	(0.96)	254.03	253.47	253.24
113	VLQSFSDGLK	255.95	(1.41)	257.20	256.08	255.67

[a] An Ac- prefix indicates an acetylated N-terminus. A CV superscript indicates that the peptide was chosen to be a member of the cross-validation set. A P superscript indicates that the peptide was chosen to be a member of the prediction set. [b] Values listed here are taken from Table 1 in Valentine et al.[8] [c] An asterisked value indicates that the prediction differs by more than 2% of the experimentally determined value.

menting CNNs, simulated annealing, and genetic algorithms, was used in the development of the QSAR models described here. The development cycle for the QSAR models consisted of four steps: (1) structure entry and optimization, (2) descriptor generation and objective feature selection, (3) linear model formation, and (4) nonlinear model formation.

Structure Entry and Optimization. Molecular models for each of the peptides were generated in HyperChem 3.0 (Hypercube, Inc., Waterloo, ON) on a Pentium PC using the HyperChem scripting language and building each peptide from its constituent amino acids. The peptides were modeled as neutral molecules and assigned a rough three-dimensional conformation using the model-building routine in HyperChem. This provided the required information about atom types and bond connectivity within the peptide molecules. No effort was made to assign a plausible three-dimensional conformation to the peptides, because the topological descriptors and amino acid parameters used as independent variables and as network inputs are not dependent upon the positions of the atoms for their calculation. That is, all descriptors used in this study are independent of peptide conformation. Instead, it was hypothesized that these descriptors could implicitly model the three-dimensional characteristics of the ensemble of conformations that are ultimately responsible for the collision cross section of any given peptide.

Descriptor Generation and Objective Feature Selection. Two primary classes of descriptors were generated in this study: (1) topological descriptors and (2) amino acid-based descriptors. Topological descriptors are based on graph theory and encode information about the types of atoms and bonds in a molecule and the nature of their connections. Examples of topological descriptors include counts of atom and bond types and indexes that encode the size, shape, and types of branching in a molecule.[23] Amino acid-based parameters encode features pertaining to and properties of individual amino acids. These are widely varied; examples from the literature include the NMR chemical shift of

[23] Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000.

the α carbon,[24] electronic charge index,[25] and frequency of occurrence in well-defined secondary protein structure,[26] amino acid topology,[27] and principal components derived from yet other measured or calculated properties of the amino acids.[28−31] In the present study, descriptors based on amino acid parameters were generated by simply summing the individual amino acid parameters for all of the amino acids in a given peptide. A total of 131 topological[32−40] and 68 amino acid-based[24−31] descriptors were calculated for each peptide.

To eliminate descriptors that contained little or redundant information in the set of 199 descriptors, objective feature selection was performed. Objective feature selection is carried out using only the descriptors; the dependent variable is not used. First, descriptors whose values were identical for at least 90% of the training set members were eliminated due to insufficient information content. In addition, one of two descriptors whose pairwise correlation coefficient exceeded 0.90 for the training set peptides was also removed to eliminate redundant information. The final reduced pool of descriptors contained 67 topological and amino acid-based descriptors for each peptide.

Linear Feature Selection and Model Formation (Type I). A simulated annealing feature selection algorithm was used to select subsets of descriptors from the reduced pool. The descriptor subsets were then evaluated using a multiple linear regression fitness evaluator. The fitness evaluator selected the best models primarily on the basis of the number of outliers and then by the RMSE of the training set peptides. Fewer outliers and lower RMSE values were favored. In this study, an outlier was defined as any peptide whose predicted cross section differed from its experimental cross section by more than 2%. This value was chosen because the experimental errors for all but one of the 113 peptides (TVGGK) used in the present study are 2% or less. It was hypothesized that more-predictive models could be obtained by defining a fitness evaluator whose primary objective was to predict as many training set observations as possible to within a certain tolerance and then selecting models with the lowest training set RMSE in the case of equal numbers of outliers. This method was successful. Each of the descriptors included in the MLR models was required to have a *t*-value of 2.0 of greater, ensuring that the descriptors in the model contribute significantly useful information at the 95% confidence level. MLR models were generated with varying numbers of descriptors. Preference was given to models containing fewer descriptors that did not significantly increase the outlier number or RMSE.

Using the criteria previously described, an optimal model was sought from all type I models generated. Descriptors in candidate models were tested for correlation among themselves. Models were rejected if the multiple correlation coefficient for any of its descriptors was greater than 0.90. To test the predictive power of the models, the collision cross sections were calculated for the peptides of the prediction set and were compared to their corresponding experimentally determined cross sections.

Linear Feature Selection and Nonlinear Model Formation (Type II). The descriptors that were selected for inclusion in the best type I model were used as inputs to a three-layer, fully connected, feed-forward computational neural network. The three layers in the network consist of an input layer, a hidden layer, and an output layer. The number of neurons in each layer specified the network architecture. There are as many neurons in the input layer as there are input descriptors. The number of neurons in the hidden layer may vary. In general, networks with too few hidden layer neurons tend to overgeneralize when making predictions, and too many hidden neurons tend to make the network memorize peculiarities of the training set. The output layer consists of a single neuron representing the predicted value, in this case, the collision cross section. The networks used in the type II models were trained using a BFGS (Broyden−Fletcher−Goldfarb−Shanno)[41−44] quasi-Newton optimization method coupled with an early stopping algorithm. This involved splitting the set of 100 training set members into a 90-member training set and a 10-member cross-validation set. The cross-validation set prevents the network from overtraining by allowing the training process to be stopped when the network begins to memorize peculiarities of the training set. Associated with the network is a set of weights and biases collectively known as adjustable parameters. An additional precautionary measure taken to prevent overtraining was that the number of adjustable parameters in the network was limited to half the number of peptides in the training set. The models were then validated using the peptides of the prediction set. Several network architectures were tested to determine the optimal configuration of the CNN. The best models were defined to be those with the fewest outliers, and in the case of equal numbers of outliers, the smallest difference in the training set and prediction set RMSE.

Nonlinear Feature Selection and Model Formation (Type III). It is unreasonable to expect that the descriptors chosen using a linear feature selection routine will be the optimal descriptors to use as inputs to a nonlinear computational neural network. Therefore, a genetic algorithm-driven, computational neural network-based feature selection routine was used to select an optimal set of descriptors as inputs to the CNN. The 67 descriptors of the reduced pool described earlier were considered. The fitness evaluator used here selects descriptors based only on the magnitude of the training set RMSE. The adjustable parameters in the CNN were then optimized in the same manner as described

[24] Fauchère, J.-L.; Charton, M.; Kier, L. B.; Verloop, A.; Pliska, V. *Int. J. Peptide Protein Res.* **1988**, *32*, 269−278.
[25] Collantes, E. R.; Dunn, W. J. *J. Med. Chem.* **1995**, *38*, 2705−2713.
[26] Chou, P. Y.; Fasman, G. D. *Annu. Rev. Biochem.* **1978**, *47*, 251−276.
[27] Raychaudhury, C.; Banerjee, A.; Bag, P.; Roy, S. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 248−254.
[28] Hellberg, S.; Sjöström, M.; Skagerberg, B.; Wold, S. *J. Med. Chem.* **1987**, *30*, 1126−1135.
[29] Norinder, U. *Peptides* **1991**, *12*, 1223−1227.
[30] Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. *J. Med. Chem.* **1998**, *41*, 2481−2491.
[31] Zaliani, A.; Gancia, E. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 525−533.
[32] Randic, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
[33] Kier, L. B.; Hall, L. H.; Murray, W. J.; Randic, M. *J. Pharm. Sci.* **1975**, *64*, 1971−1973.
[34] Randic, M. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164−175.
[35] Kier, L. B. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109−116.
[36] Kier, L. B.; Hall, L. H. *Pharm. Res.* **1990**, *7*, 801−807.
[37] Sharma, V.; Goswami, R.; Madan, A. K. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 273−282.
[38] Liu, S.; Cao, C.; Li, Z. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387−394.
[39] Hansch, C.; Leo, A.; Unger, S. H.; Kim, K. H.; Nikaitani, D.; Lien, E. J. *J. Med. Chem.* **1973**, *16*, 1207−1216.
[40] Miller, K. J.; Savchik, J. A. *J. Am. Chem. Soc.* **1979**, *101*, 7206−7213.

[41] Broyden, C. G. *J. Inst. Math. Its Appl.* **1970**, *6*, 76−90.
[42] Fletcher, R. *Comput. J.* **1970**, *13*, 317−322.
[43] Goldfarb, D. *Math. Comput.* **1970**, *24*, 23−26.
[44] Shanno, D. F. *Math. Comput.* **1970**, *24*, 647−656.

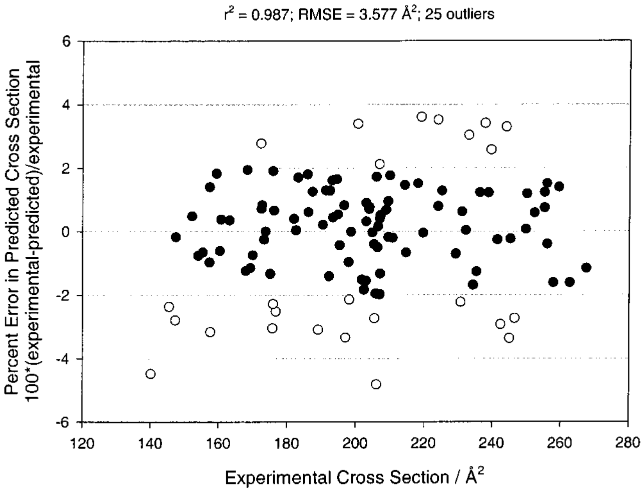

Figure 1. Plot of percent error in predicted collision cross section versus the experimentally determined collision cross section for the linear model containing only one descriptor, the number of atoms. Closed circles represent peptides whose predicted cross section was within 2% of their experimentally determined cross section. Open circles represent peptides whose predicted cross section was greater than 2% of their experimentally determined cross section.

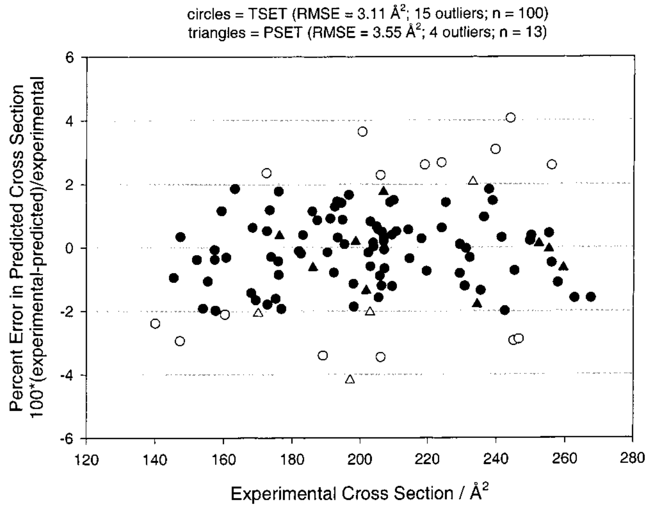

Figure 2. Plot of percent error in predicted collision cross section versus the experimentally determined collision cross section for the best type I model found. Circles represent training set members and triangles represent prediction set members. Closed shapes represent peptides whose predicted cross section was within 2% of their experimentally determined cross section. Open shapes represent peptides whose predicted cross section was greater than 2% of their experimentally determined cross section.

for type II models. Once again, the training and cross-validation sets were used to construct the models, and the prediction set was used to validate the predictive ability of the models.

RESULTS AND DISCUSSION

Dependence of Collision Cross Section on Size and Mass. To a first approximation, the collision cross section depends on the overall size or mass of the peptide. It was found for the set of 113 peptides used in this study that the collision cross section correlated extremely well with both the molecular weight of the peptide ion[8] ($r^2 = 0.965$, RMSE = 5.78 Å^2) and the number of atoms in the peptide ion ($r^2 = 0.987$, RMSE = 3.577 Å^2). The number of atoms used here includes all heavy and hydrogen atoms except the extra proton responsible for the +1 charge of the peptide ions. A simple linear model was constructed using only the number of atoms and is presented in Figure 1. This trivial linear model was able to predict 88 of 113 peptides (78%) to within 2% of their experimentally determined collision cross sections. Shvartsburg et al.[15] found that by summing the projection area contributions from each atom and dividing by the total mass of the peptide, they could fairly accurately predict the collision cross sections for the same set of 113 peptides. This is mathematically equivalent to using the number of atoms in the molecule when all atomic radii and masses are equivalent. A similar dependence upon size was found when predicting the ion mobilities of small organic molecules.[13,14] Thus, any descriptors included in a model in addition to the measure of overall size represent the more subtle conformational aspects of the gas-phase peptide ions (which interact with themselves and with the surrounding buffer gas), and in which most of our interest lies.

Type I Models. Models containing from 2 to 10 descriptors were examined for overall quality. Only descriptor subsets whose *t*-values were all above 2.0 were considered. The six-descriptor subset that was selected for further model development is presented in Table 2. A type I model was built using these six descriptors, and it was found that no descriptor had a multiple correlation coefficient of greater than 0.90, ensuring that the redundant information being introduced into the model via combinations of other descriptors was kept to an acceptable level. This model is presented in Figure 2 and has a training set RMSE of 3.11 Å^2 (15 of 100 were outliers) and a prediction set RMSE of 3.55 Å^2 (4 of 13 were outliers). The prediction that showed the worst agreement with experiment was obtained for the IEEIFK peptide which was overpredicted by 4.18%. Pairwise correlation coefficients among the six descriptors ranged from 0.016 to 0.858

Table 2. Descriptors Selected for the Type I MLR and Type II CNN Models

descriptor	type[a]	coefficient	error	range	explanation[b]
na_*-0	T	1.249	2.803×10^{-2}	68−168	no. atoms, including hydrogen
3SP3-1	T	−2.309	0.4852	1−4	count of 3° sp^3-hybridizded carbons
AA_S-17	AA	1.566	0.6697	0−2	Count of serine residues
EXTR-105	AA	−14.03	3.179	0.316−1.10	Chou−Fasman f(i + 1)
EXTR-108	AA	1.504×10^{-2}	4.541×10^{-3}	292−1030	Collantes−Dunn ISA
EXTR-143	AA	0.4150	0.1754	−2.69−7.48	Sandberg et al. z_5
constant		62.57	1.609		

[a] T, topological; AA, amino acid parameter-based. [b] na_*-0, number of heavy and hydrogen atoms; 3SP3-1, count of sp^3-hybridized carbon atoms bonded to three other carbon atoms; EXTR-105, frequency of amino acid in the (i + 1)st position of a β-turn;[26] EXTR-108, Collantes−Dunn isotropic surface area (ISA);[25] EXTR-143, fifth Sandberg et al. z-index.[30]

with a mean of 0.325. Table 1, column 5 gives the individual calculated collision cross section values generated by the type I model.

Two of the descriptors in the model are topological, and four are amino acid-based. One of the topological descriptors is the number of atoms, as described in the previous section. The other topological descriptor, 3SP3-1, encodes the number of sp^3-hybridized carbon atoms that are connected to three other carbon atoms. There are no carbons of this type in the peptide backbone, and only three naturally occurring amino acid side chains possess such a carbon atom: leucine, isoleucine, and valine, with one apiece. This descriptor, then, is equivalent to the count of leucine, isoleucine, and valine residues in the peptide. This is significant, because Valentine et al.[17] found that these three nonpolar aliphatic amino acids possessed the highest intrinsic size parameters (ISP) of any amino acid found in the set of 113 peptides (except for lysine, which is always at the C-terminus of the peptide). These Clemmer ISPs are amino acid parameters that represent a particular amino acid's contribution to the total collision cross section. They were determined by solving a set of linear equations relating the numbers and types of amino acids in a set of peptides to their collision cross sections. Additionally, Shvartsburg et al.[15] found that amino acid side chain density is a factor in determining the collision cross section. They define an ISP for an amino acid as the sum of the projection area contributions of all the atoms divided by the masses of the atoms in the amino acid. This area-per-mass measure can be seen as being inversely related to density. Thus, amino acids whose side chains contain many light atoms (nonpolar aliphatic residues in particular) will be the least dense and have the highest Shvartsburg ISPs. Leucine, isoleucine, and valine have the highest nonlysine Shvartsburg ISPs. Thus, the inclusion of the 3SP3-1 descriptor in this type I model incorporates information about side chain density.

In addition, four amino acid parameters were included in the model. AA_S-17 is the count of serine residues, and it may have been included in the model because of serine's atypically small but polar side chain. EXTR-105 is the Chou-Fasman f(i + 1) index.[26] It is the frequency of occurrence of an amino acid found in the (i + 1)st position of a β turn, a common secondary structure found in proteins and peptides. β turns are compact structures, and proline, with its sterically hindered conformation that is conducive to β turns, is most often found in this position. This descriptor may thus be encoding the ability of the peptide to fold into a compact conformation. The sign and magnitude of the coefficient associated with this descriptor relative to its actual value indicates that there will be a significant reduction in the collision cross section if several of these turn-friendly amino acids are included in the peptide. EXTR-108 is the Collantes−Dunn isotropic surface area (ISA).[25] ISA is defined as the surface area of an amino acid available for nonspecific solvent interactions and is calculated by first solvating the amino acid with water molecules at specific hydrogen bonding sites and then calculating the surface area of the amino acid that remains accessible to the solvent. This amino acid parameter was developed for use in biologically oriented quantitative structure−activity relationships (QSAR), in which an aqueous environment is commonly encountered. However, in the present study, it may be interpreted differently. The side chain functional groups of many amino acids are electron-rich and contain lone pair electrons and aromatic pi systems that will solvate a positive charge as well as be solvated by water molecules. Interactions between the positive charge and these functional groups will tend to contract the peptide into a more compact conformation, leaving the isotropic surface area exposed. The final descriptor in the model, EXTR-143 is z_5, one of a set of principal components developed by Sandberg et al.[30] The five Sandberg et al. principal components z_1 through z_5 are derived from chromatographic retention times, NMR chemical shifts, and other calculated electronic properties of both natural and unnatural amino acids. z_4 and z_5 were reported as being related to electronegativity, heat of formation, electrophilicity, hardness, and other measures of electronic character. This descriptor may be encoding the ability of the peptide ion to solvate its positive charge.

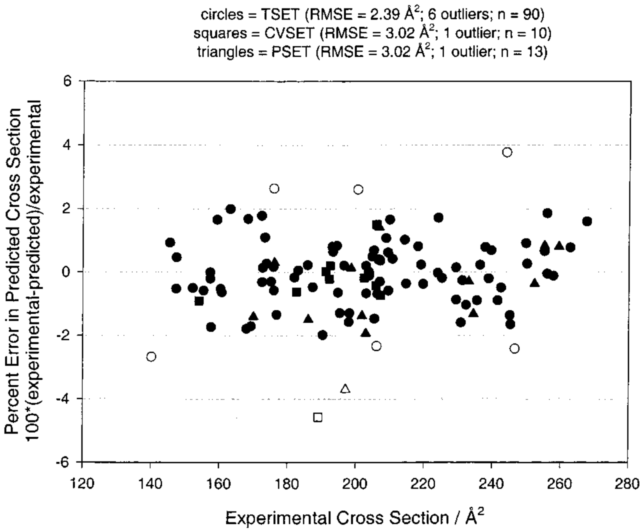

Figure 3. Plot of percent error in predicted collision cross section versus the experimentally determined collision cross section for the best type II model found. Circles represent training set members, squares represent cross-validation set members, and triangles represent prediction set members. Closed shapes represent peptides whose predicted cross section was within 2% of their experimentally determined cross section. Open shapes represent peptides whose predicted cross section was greater than 2% of their experimentally determined cross section.

Type II Models. The descriptors that were found to work well in the best type I model were used as inputs to a type II CNN model. All possible architectures from 6-2-1 (17 adjustable parameters) to 6-5-1 (41 adjustable parameters) were examined to find the model with the optimal training and prediction set error. The best model found was a 6-5-1 CNN. A plot of the percent error in the collision cross section versus the experimentally determined collision cross section is presented in Figure 3. As expected, there was a significant improvement in model quality as a result of the added nonlinear characteristics of the neural network. The training set error dropped from 3.11 to 2.39 Å^2 (6 of 90 were outliers) and the prediction set error dropped from 3.55 to 3.02 Å^2 (1 of 13 was an outlier). The cross-validation set error was also 3.02 Å^2 (1 of 10 was an outlier). The prediction that showed the worst agreement with experiment was obtained for the APVDAFK peptide, which was overpredicted by 4.57%. Table 1, column 6 gives the individual calculated collision cross section values generated by the type II model.

Type III Models. It was hypothesized that the best overall model would come about as a result of the use of a nonlinear feature selection routine coupled with a nonlinear neural network, and this was indeed the case. The optimal descriptors found for the best type III model are presented in Table 3. As before, a 6-5-1 network architecture was found to be optimal. A plot of the percent error in the collision cross section versus the experimentally determined collision cross section is presented in Figure 4. The errors associated with all three sets were lower than those in the best type II model. The training set error was 2.05 Å^2 (3 of 90 were outliers), the cross-validation set error was 2.37 Å^2 (0 of 10 were outliers), and the prediction set error was 2.82 Å^2 (1 of 13 was an outlier). The prediction that showed the worst agreement with experiment was obtained for the IEEIFK peptide, which was overpredicted by 3.34%. Table 1, column 7 gives the individual calculated collision cross section values generated by the type III model.

The descriptors included in the best type III model include four topological descriptors and two additional amino acid parameter-based descriptors. The number of atoms appears, as expected. In addition, the number of nitrogens, NN-4, appears. This descriptor includes the nitrogens of the peptide backbone, the lysine side chain nitrogen, and the nitrogen in the asparagine, glutamine, and tryptophan side chains. MDE-12 is a molecular distance-edge descriptor[38] that encodes the through-bond distances between all sp^3-hybridized carbon atoms connected to one and two other carbon atoms. With the exception of threonine, the only amino acids that possess this type of primary carbon are the nonpolar aliphatic amino acids alanine, valine, isoleucine and leucine. Hence, this descriptor could be contributing in much the same way as 3SP3-1 does in the type I and II models. EMIN-1 is the lowest electrotopological state index[36] of any heavy (non-hydrogen) atom in the peptide. Electrotopological state (e-state) indexes are calculated for each heavy atom and encode the number of valence electrons and the degree of branching at each atom. Lower e-state indexes are assigned to atoms that have fewer valence electrons and that are farther away from the periphery of the molecule (that is, they are more highly branched). For the set of 113 peptides used in this study, the atom that corresponds to the lowest e-state is almost always the α carbon of the residue possessing the most polar side chain. This makes sense for several reasons. First, the carbon atom has the lowest number of valence electrons of any heavy atom found in peptides. Second, the α carbon is as highly branched as any atom found in a peptide and is buried in the peptide backbone. Third, the e-state of an atom can be reduced significantly by neighboring atoms that have many valence electrons or are on the periphery of the molecule. Thus, side chains of increasing polarity result in corresponding α carbons with decreasing e-states. For peptides that contain no polar side chains, the lowest e-state atom is either the terminal lysine α carbon or the carbonyl carbon of the C-terminus. Thus, the EMIN-1 descriptor is indirectly and inversely related to the polarity of the most polar side chain in the peptide, and hence, the ability of the peptide to solvate the positive charge. The two remaining descriptors are the Sandberg et al. z_4 and z_5 principal components[30] that are both related to various calculated electronic properties and most likely have a role similar to the z_5 descriptor in the best type II model.

Randomizing Experiments and Correlation with Experimental Error. To show that the results obtained by the models were not due to chance correlations, a randomizing experiment was performed. The first part of the experiment involved randomly scrambling the dependent variable, in this case the collision cross section. The second part of the experiment was an attempt to construct a type III model using the same methodology as was used to build the actual type III model but using the scrambled dependent variable data. The training set error in this experiment was 18.90 Å^2 (60 of 90 were outliers), the cross-validation set error was 34.22 Å^2 (9 of 10 were outliers), and the prediction set error was 45.22 Å^2 (11 of 13 were outliers). These results show that the best models were extremely unlikely to have been found due to chance correlation effects.

Since each of the 113 peptides included in this study was measured multiple times, it was possible to calculate a standard deviation of measurement for each one, and this value was defined as the experimental error. The pairwise correlation coefficient between the experimental error and the absolute value of the prediction error was found to be 5.97×10^{-3} for the best type I model, 5.73×10^{-3} for the best type II model, and 5.29×10^{-4} for the best type III model. Hence, no significant correlation was found between experimental and prediction errors in any of the models presented.

Sequence Isomer Discrimination. The best type III model has the ability to discriminate among sequence isomers, because two of the input descriptors, MDE−12 and EMIN-1, are sequence-dependent. To examine this capability more closely, two virtual peptide libraries were created, one containing pentapeptides and the other containing hexapeptides. The pentapeptide peptide library contained all 24 permutations ($4! = 24$) of the lysine-terminated peptide GIWS(K) and the hexapeptide library contained all possible 120 permutations ($5! = 120$) of the lysine-terminated peptide FAQDM(K). The amino acid composition of the peptides was selected so as to include a variety of amino acid types (nonpolar aliphatic, aromatic, polar aliphatic, etc.). The peptides were then modeled in exactly the same fashion as those used to construct the previously described models. The peptide libraries were then submitted to the best type III model for prediction. Table 4, column 3 gives the individual calculated collision cross section values generated by the type III model for the pentapeptide library in order of increasing collision cross section. The pentapeptides predicted to have the highest and lowest cross sections are ISGWK (167.83 Å^2) and GIWSK (165.24 Å^2) respectively, representing a difference of 2.59 Å^2. In general, the smallest cross sections occur when glycine (G) is in the first (N-terminal) position and the largest cross sections occur when serine (S) is in the first position. Table 5, column 3 gives the individual calculated collision cross section values generated by the type III model for the hexapeptide library in order of increasing collision cross section. The hexapeptides predicted to have the highest and lowest cross sections are DAMFQK (188.28 Å^2) and FAQDMK (180.34 Å^2), respectively, representing a difference of 7.94 Å^2. For the hexapeptide library, the smallest cross sections occur when alanine (A) or phenylalanine (F) is in the first position and the largest cross sections occur when aspartic acid (D) is in the first position.

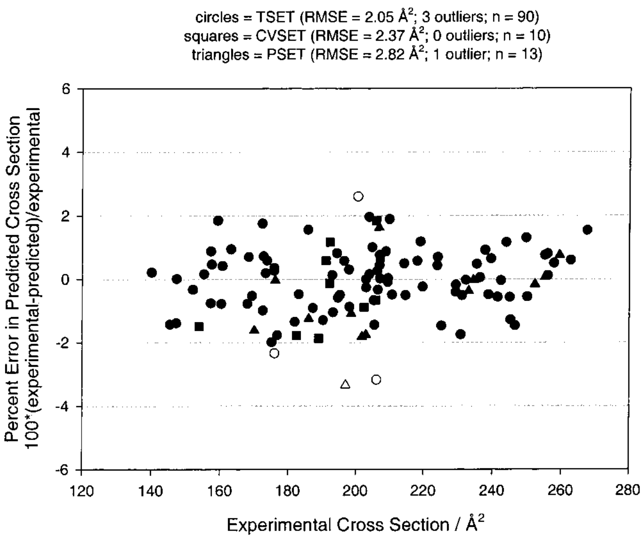

Figure 4. Plot of percent error in predicted collision cross section versus the experimentally determined collision cross section for the best type III model found. Circles represent training set members, squares represent cross-validation set members, and triangles represent prediction set members. Closed shapes represent peptides whose predicted cross section was within 2% of their experimentally determined cross section. Open shapes represent peptides whose predicted cross section was greater than 2% of their experimentally determined cross section.

Table 4. Predicted Collision Cross Sections for the Virtual Pentapeptide Library Using the Best Type III Model

ID	sequence	prediction type III (Å^2)
1	GIWSK	165.24
2	GWSIK	165.33
3	GWISK	165.50
4	IGWSK	165.70
5	GISWK	165.77
6	WGISK	165.80
7	WGSIK	166.01
8	WIGSK	166.29
9	IGSWK	166.35
10	IWSGK	166.36
11	IWGSK	166.45
12	WISGK	166.46
13	WSGIK	166.82
14	GSWIK	166.85
15	WSIGK	166.86
16	GSIWK	167.27
17	SGWIK	167.36
18	SWGIK	167.44
19	ISWGK	167.46
20	SGIWK	167.50
21	SWIGK	167.61
22	SIGWK	167.74
23	SIWGK	167.81
24	ISGWK	167.83

An important trend may be observed in both libraries by identifying the amino acid in each library that is best able to act as a counterion and to stabilize the +1 charge located on the lysine residue. In the pentapeptide library, this is serine (S), and in the hexapeptide library, it is aspartic acid (D). In the pentapeptide library, the smallest cross sections are predicted when the serine residue is one to two amino acid positions away from the terminal lysine residue. Increasingly larger cross sections are predicted as more of the nonpolar or weakly polar amino acids are interposed between the lysine and serine residues. A similar trend holds for the hexapeptide library. In this case, the smallest cross sections are predicted when the aspartic acid residue is two amino acid positions away from the terminal lysine residue, and slightly larger cross sections are predicted when the aspartic acid residue is one or three positions away from the terminal lysine residue. Again, more intervening nonpolar or weakly polar amino acids increased the predicted cross section.

These results seem reasonable; however, no sequence isomers were included in the training set used to build the model, and it is thus unclear whether the variations in the predicted cross sections are truly meaningful for sets of such compounds. Studies are currently underway to determine the validity of these results.

General Considerations. It is interesting to note that many of the same outliers were found in all three models. Seven of the eight peptides predicted to be outliers in the type II model are outliers in the type I model. Three of the four peptides predicted as outliers in the best-performing type III model (which uses a set of descriptors that is different from the type I and type II models) are outliers in the type II and type I models as well. The reason for this is not apparent upon casual examination of the peptides. Since the outliers are by definition not predicted as well as the other compounds, these can be identified as sequences whose structures differ most notably from the majority of the other sequences' structures in the training set. The poorer predictions for these peptides are probably due in part to some unique structural feature particular to these peptides that is not being encoded by the selected set of descriptors. There is no correlation between the prediction and experimental errors for these outliers.

Finally, it must be emphasized that the models presented here represent correlations between calculated (and in many cases, quite artificial) descriptors and collision cross sections of a

Table 3. Descriptors Selected for the Type III CNN Model

descriptor	type[a]	range	explanation[b]
na_*-0	T	68−168	no. atoms, including hydrogen
NN-4	T	6−12	no. nitrogen atoms
MDE-12	T	0.00−14.87	molecular distance-edge 1°-2° carbons
EMIN-1	T	−1.98 to −1.18	minimum electrotopological state atom
EXTR-142	AA	−1.29−3.83	Sandberg et al. z_4
EXTR-143	AA	−2.69−7.48	Sandberg et al. z_5

[a] T, topological; AA, amino acid parameter-based. [b] na_*-0, number of heavy and hydrogen atoms; MDE-12, molecular distance-edge connectivity between all primary and secondary carbons with C−C bonds;[38] EMIN-1, minimum electrotopological state atom;[36] EXTR-142, fourth Sandberg et al. *z*-index;[30] EXTR-143, fifth Sandberg et al. *z*-index.[30]

Table 5. Predicted Collision Cross Sections for the Virtual Hexapeptide Library Using the Best Type III Model

ID	sequence	prediction, type III (Å^2)
1	FAQDMK	180.34
2	AFQDMK	180.40
3	AMQDFK	180.50
4	FAMDQK	180.51
5	AFMDQK	180.54
6	AMFDQK	180.60
7	MAQDFK	180.61
8	FQDMAK	180.63
9	FMQDAK	180.65
10	MAFDQK	180.72
11	FMDQAK	180.76
12	MFQDAK	180.82
13	FMADQK	180.85
14	FAMQDK	180.88
15	MFDQAK	180.90
16	MQDFAK	180.92
17	MFADQK	181.00
18	AFMQDK	181.00
19	AQDFMK	181.01
20	FMAQDK	181.02
21	AFDQMK	181.03
22	FQDAMK	181.04
23	AQDMFK	181.07
24	AMFQDK	181.09
25	AQFDMK	181.10
26	MAFQDK	181.14
27	AQMDFK	181.14
28	MFAQDK	181.19
29	FADQMK	181.25
30	AMDQFK	181.30
31	MQDAFK	181.40
32	FQMDAK	181.41
33	FQADMK	181.43
34	QAFDMK	181.49
35	QFDMAK	181.49
36	QAMDFK	181.55
37	MQFDAK	181.62
38	FAQMDK	181.63
39	QMDFAK	181.65
40	MADQFK	181.66
41	MQADFK	181.68
42	AFQMDK	181.77
43	AMQFDK	181.83
44	QFMDAK	181.85
45	QFADMK	181.88
46	MAQFDK	181.89
47	QMFDAK	181.89
48	QFDAMK	181.95
49	AFDMQK	181.96
50	QMADFK	181.97
51	AQMFDK	182.11
52	FQAMDK	182.12
53	AQFMDK	182.14
54	QADFMK	182.14
55	FMDAQK	182.16
56	AMDFQK	182.17
57	QMDAFK	182.19
58	QADMFK	182.21
59	FMQADK	182.23
60	FADMQK	182.24
61	QAMFDK	182.26
62	QAFMDK	182.28
63	MFDAQK	182.30
64	MQAFDK	182.35
65	QFAMDK	182.42
66	MFQADK	182.42
67	QMAFDK	182.46
68	FQMADK	182.58
69	MADFQK	182.58
70	FDQMAK	182.78
71	MQFADK	182.84
72	QFMADK	182.87
73	QDFMAK	182.93
74	QMFADK	182.94
75	FDQAMK	183.01
76	QDMFAK	183.06
77	QDFAMK	183.16
78	MDQFAK	183.30
79	QDMAFK	183.34
80	MDQAFK	183.57
81	QDAFMK	183.78
82	FDMQAK	183.82
83	QDAMFK	183.83
84	ADQFMK	184.11
85	ADQMFK	184.15
86	MDFQAK	184.21
87	FDAQMK	184.55
88	FDMAQK	184.58
89	MDFAQK	184.94
90	ADFQMK	185.02
91	FDAMQK	185.06
92	MDAQFK	185.13
93	ADMQFK	185.22
94	ADFMQK	185.50
95	MDAFQK	185.57
96	ADMFQK	185.65
97	DQFMAK	186.72
98	DQMFAK	186.79
99	DQFAMK	186.81
100	DQMAFK	186.90
101	DQAFMK	187.03
102	DQAMFK	187.05
103	DFQMAK	187.22
104	DFQAMK	187.29
105	DMQFAK	187.48
106	DFMQAK	187.49
107	DMQAFK	187.55
108	DMFQAK	187.67
109	DFAQMK	187.68
110	DFMAQK	187.73
111	DFAMQK	187.85
112	DMFAQK	187.89
113	DAQFMK	187.90
114	DAFQMK	187.90
115	DAQMFK	187.91
116	DMAQFK	187.96
117	DAMQFK	187.99
118	DMAFQK	188.10
119	DAFMQK	188.19
120	DAMFQK	188.28

relatively small set of peptide ions. The true cause-and-effect relationships at work may be well-described by the models or hidden to some extent by any number of factors. Factors could include data set representation, experimental error, pairwise correlations among the descriptors originally calculated, intercorrelation among descriptors that appear in the final model, and insufficient information content in the calculated descriptors. An inadequate representation of the set of compounds for which the model will ultimately be used to make predictions (i.e., the training set) will lead to models that are incorrect to some degree. Even though the chemical class represented in this study (peptides) is fairly restricted in the kinds of atoms that can be included and how these atoms may be connected to one another, a training set of 90 or 100 peptides may still be too small to adequately represent the billions of possible peptides the models are intended for. Experimental error becomes a factor when the predictive power of the model approaches the resolution of the instrument used to collect the data. This is the case for this study. Additionally, it must be remembered that peptides, as most other polymers, generally do not exist as a single conformation, but instead as an ensemble of different conformations. The contributions of these many conformations will cause the standard deviations of the collision cross sections to increase, introducing more uncertainty into the measurement. Pairwise correlation between the descriptors originally calculated can be a factor, since one of the two is arbitrarily discarded when the correlation between the two is strong enough. This can mean that an important descriptor is excluded from further consideration at an early stage of model development. Intercorrelation among descriptors that appear in the final model can confound model interpretation, since a linear or nonlinear combination of these may represent the true cause of the observation being modeled. Although a model using multiple descriptors may give excellent results, it may hide a simpler explanation. Insufficient information content in the calculated descriptors may prevent the optimal model from being produced because no combination of these descriptors is able to adequately model the features that are ultimately responsible for the effects being modeled.

CONCLUSIONS

Predictive QSPRs have been presented that link topological molecular structure and derived amino acid parameters with the ion mobility spectrometry collision cross sections of a set of 113 singly protonated, lysine-terminated peptides from a tryptic digest of common proteins. No three-dimensional information about peptide conformation is explicitly included in the models. The models produced may give useful insights into the possible mechanisms responsible for the folding of peptide ions in the gas phase. A trivial linear model using only the number of atoms as an independent variable was able to predict 88 of 113 peptide collision cross sections (78%) to within 2% of their experimentally determined value. The best MLR model obtained contained six descriptors and was able to predict 94 of 113 peptide collision cross sections (83%) to within 2% of the experimentally determined value. Using the same six descriptors with a 6-5-1 CNN, the results improved to 105 of 113 peptides (93%) predicted to within 2% of experiment. Finally, an optimal set of six descriptors was chosen for use in a CNN, and this 6-5-1 model predicted 109 of 113 peptide collision cross sections (96%) to within 2% of experiment. This model was shown to have the ability to discriminate among sequence isomers, representing an additional capability not found in previously described group contribution methods.

ACKNOWLEDGMENT

We would like to thank Greg Kauffman for his work in the initial stages of this project.

Received for review November 21, 2001. Accepted January 9, 2002.

AC0112059