Supporting Information

'Improved Descriptors for the QSAR Modeling of Peptides and Proteins'

*Mark H Barley, Nicolas J Turner and Royston Goodacre**

S1:- Tables of Descriptor values

Table S1:- Hellberg and MS-WHIM Descriptors

	Hellberg Descriptors			MS-WHIM Descriptors		
	Z(1)	Z(2)	Z(3)	Ist	2nd	3rd
Ala	0.07	-1.73	0.09	-0.73	0.20	-0.62
Arg	2.88	2.52	-3.44	-0.22	0.27	1.00
Asn	3.22	1.45	0.84	0.14	0.20	-0.66
Asp	3.64	1.13	2.36	0.11	-1.00	-0.96
Cys	0.71	-0.97	4.13	-0.66	0.26	-0.27
Gln	2.18	0.53	-1.14	0.30	1.00	-0.30
Glu	3.08	0.39	-0.07	0.24	-0.39	-0.04
Gly	2.23	-5.36	0.3	-0.31	-0.28	-0.75
His	2.41	1.74	1.11	0.84	0.67	-0.78
Ile	-4.44	-1.68	-1.03	-0.91	0.83	-0.25
Leu	-4.19	-1.03	-0.98	-0.74	0.72	-0.16
Lys	2.84	1.41	-3.14	-0.51	0.08	0.60
Met	-2.49	-0.27	-0.41	-0.70	1.00	-0.32
Phe	-4.92	1.3	0.45	0.76	0.85	-0.34
Pro	-1.22	0.88	2.23	-0.43	0.73	-0.60
Ser	1.96	-1.63	0.57	-0.80	0.61	-1.00
Thr	0.92	-2.09	-1.4	-0.58	0.85	-0.89
Trp	-4.75	3.65	0.85	1.00	0.98	-0.47
Tyr	-1.39	2.32	0.01	0.97	0.66	-0.16
Val	-2.69	-2.53	-1.29	-1.00	0.79	-0.58

References:- Hellberg Descriptors[1], MS-WHIM Descriptors[2]

Table S2:- Physical Descriptors and T-Scales

Physical Descriptors			T-SCALES				
	Vol	Hydro	T_1	T_2	T3	T4	T5
Ala	-2.90	-1.03	-9.11	-1.63	0.63	1.04	2.26
Arg	2.41	1.31	0.23	3.89	-1.16	-0.39	-0.06
Asn	-0.68	0.79	-4.62	0.66	1.16	-0.22	0.93
Asp	-0.92	1.23	-4.65	0.75	1.39	-0.4	1.05
Cys	-1.89	0.15	-7.35	-0.86	-0.33	0.8	0.98
Gln	0.36	1.09	-3	1.72	0.28	-0.39	0.33
Glu	0.16	1.28	-3.03	1.82	0.51	-0.58	0.43
Gly	-4.04	0.01	-10.61	-1.21	-0.12	0.75	3.25
His	0.83	1.15	-1.01	-1.31	0.01	-1.81	-0.21
Ile	0.51	-1.32	-4.25	-0.28	-0.15	1.4	-0.21
Leu	0.52	-1.40	-4.38	0.28	-0.49	1.45	0.02
Lys	0.92	1.23	-2.59	2.34	-1.69	0.41	-0.21
Met	0.92	-1.42	-4.08	0.98	-2.34	1.64	-0.79
Phe	2.22	-1.47	0.49	-0.94	-0.63	-1.27	-0.44
Pro	-1.25	-0.64	-5.11	-3.54	-0.53	-0.36	-0.29
Ser	-2.36	0.38	-7.44	-0.65	0.68	-0.17	1.58
Thr	-1.19	0.28	-5.97	-0.62	1.11	0.31	0.95
Trp	4.28	-0.18	5.73	-2.67	-0.07	-1.96	-0.54
Tyr	2.75	-0.18	2.08	-0.47	0.07	-1.67	-0.35
Val	-0.65	-1.27	-5.87	-0.94	0.28	1.1	0.48

References:-Physical Descriptors, This work; T-Scales[3]

S2:- Results for Alternative Descriptors.

In the main paper the Physical Descriptors are compared to the Hellberg Descriptors[1] using four literature dipeptide datasets. In this section comparisons are made between the results obtained with the Physical Descriptors and two alternative sets of descriptors:- the MS-WHIM descriptors of Zaliani and Gancia;[2] and the T-Scales of Tian et al.[3]

S2.1:- Statistical plots and data for models using the Physical Descriptors and MS-WHIM or T-Scale descriptors.

The construction of plots such as Fig. 3A in the paper (where the quality of a set of models is assessed using Q^2-LPO(50%) values) is described in the section "Comparing PLS Models Using the Two Sets of Descriptors in a Designed Experiment" and in the Methods section within the paper. Similar plots to Fig. 3A comparing the results for MS-WHIM descriptors against those for the Physical Descriptors are seen in Figures S1 and S2 while comparisons between the T-scale descriptors and the Physical Descriptors are shown in Figures S3 and S4. Table S3 then extends the statistical data (provided for the Hellberg and Physical Descriptors in Table 3 in the paper) to these alternative descriptors.

S2.1.1:- Q^2-LPO(50%) plots using MS-WHIM descriptors.

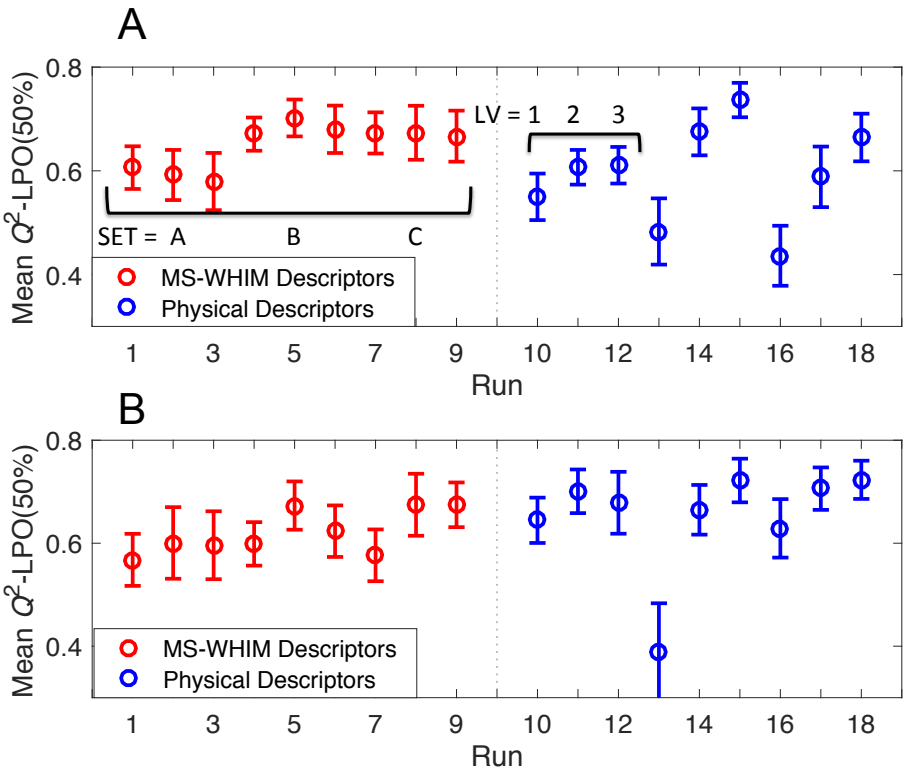

Figure S1:- Assessing the different PLS models using both MS-WHIM and the Physical Descriptors. Set A= Simple descriptors (6 for MS-WHIM, 4 for Physical Descriptors); Set B=Simple + Squared terms (12 for MS-WHIM, 8 for Physical Descriptors); and Set C=Full Set (Simple + Squared + Interactions; 27 for MS-WHIM, 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. Panel A:- ACE Inhibitor dipeptides (58 data points); *N*=20 for MS-WHIM models; *N*=18 for Physical Descriptor models. Panel B:- Bitter Dipeptides (48 data points) *N*=20 for both MS-WHIM and Physical Descriptor models.

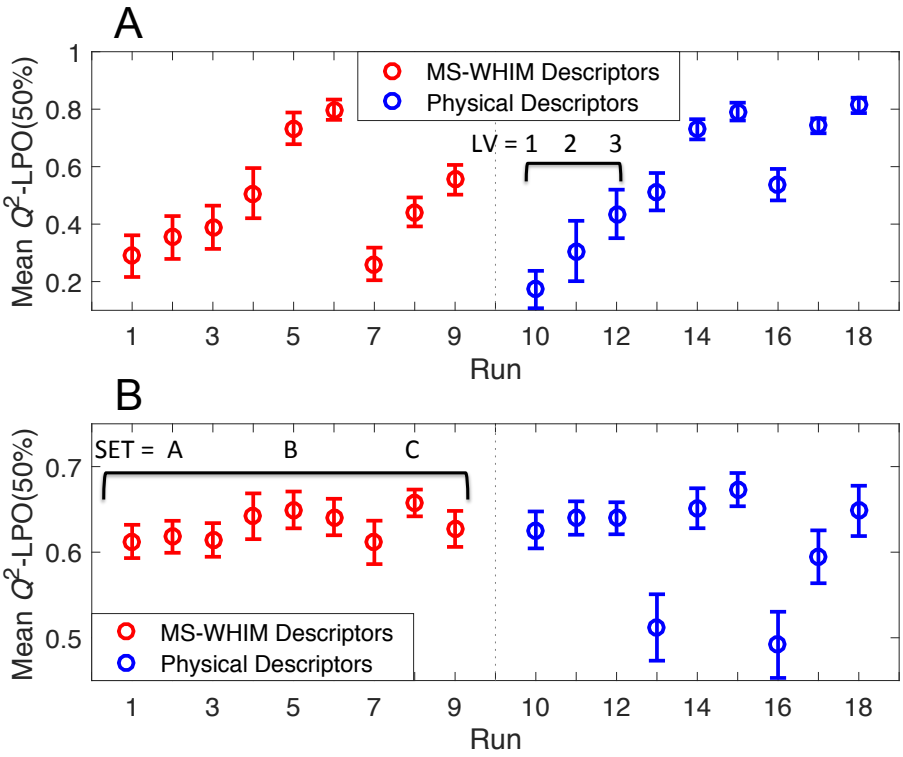

Figure S2:- Assessing the different PLS models using both MS-WHIM and the Physical Descriptors. Set A= Simple descriptors (6 for MS-WHIM, 4 for Physical Descriptors); Set B=Simple + Squared terms (12 for MS-WHIM, 8 for Physical Descriptors); and Set C=Full Set (Simple + Squared + Interactions; 27 for MS-WHIM, 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. Panel A:- Elastase Substrate k_{cat} dataset (41 Peptides); *N*=20 for MS-WHIM models; *N*=17 for Physical Descriptor models. Panel B:- Ace Inhibitor Dipeptides- Multi-laboratory dataset with one outlier removed (167 Dipeptides). *N*=25 for both MS-WHIM and Physical Descriptor models.

S2.1.2:- Q^2-LPO(50%) plots using T-Scale descriptors.

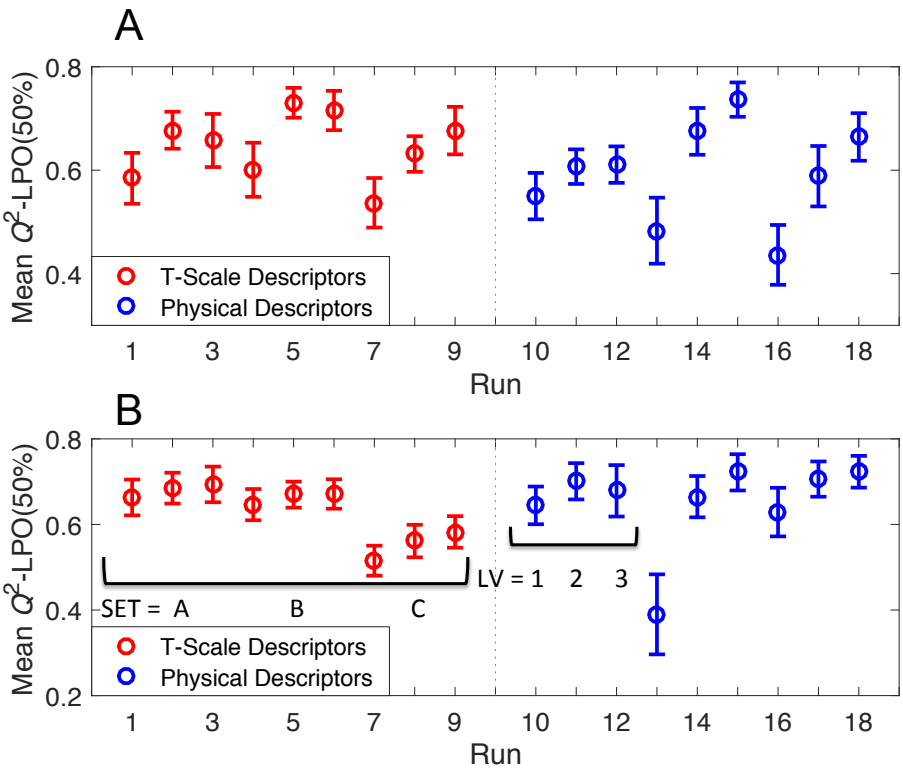

Figure S3:- Assessing the different PLS models using both T-Scale Descriptors and the Physical Descriptors. Set A= Simple descriptors (6 for T-Scales, 4 for Physical Descriptors); Set B=Simple + Squared terms (12 for T-Scales, 8 for Physical Descriptors); and Set C=Full Set (Simple + Squared + Interactions; 27 for T-Scales, 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. Panel A:- ACE Inhibitor dipeptides (58 data points); *N*=24 for T-Scale models; *N*=18 for Physical Descriptor models. Panel B:- Bitter Dipeptides (48 data points) *N*=22 for T-Scale models; *N*=20 for Physical Descriptor models.

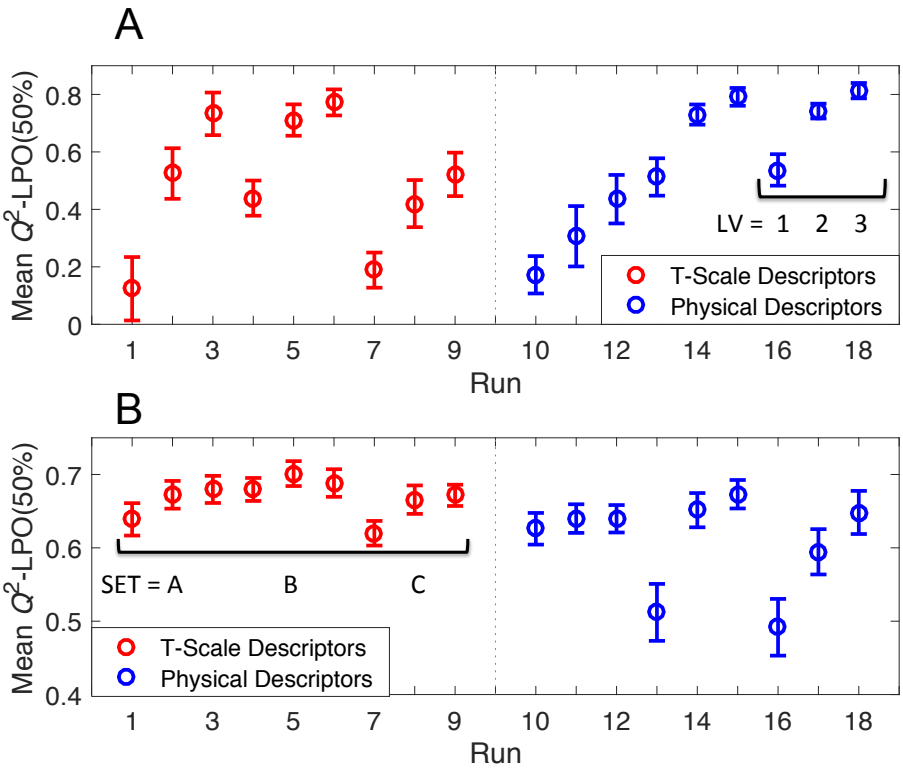

Figure S4:- Assessing the different PLS models using both T-Scale Descriptors and the Physical Descriptors. Set A= Simple descriptors (6 for T-Scales, 4 for Physical Descriptors); Set B=Simple + Squared terms (12 for T-Scales, 8 for Physical Descriptors); and Set C=Full Set (Simple + Squared + Interactions; 27 for T-Scales, 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. Panel A:- Elastase Substrate k_{cat} dataset (41 Peptides); *N*=18 for T-Scale models; *N*=17 for Physical Descriptor models. Panel B:- Ace Inhibitor Dipeptides- Multi-laboratory dataset with one outlier removed (167 Dipeptides). *N*=25 for both T-Scale and Physical Descriptor models.

S2.1.3:- Summary of Selected Results for the T-Scale and MS-WHIM Descriptors.

Table S3 Results for Modelling ACE Inhibitors (*N*=58) and Bitter Dipeptides (*N*=48)

Model	Descriptors	Terms	LV	R^2 (Train)	Q^2-CV[a,b]	R^2 (Test)
	Reference Model for the whole dataset (*N*=58)					
27	MS_WHIM	27	1	0.748	0.69, 0.59	
28	MS_WHIM	12	2	0.780	0.72, 0.70	
29	MS_WHIM	6	1	0.687	0.64, 0.58	
30	T-Scales	65	2	0.785	0.69, 0.61	
31	T-Scales	20	2	0.794	0.74, 0.69	
32	T-Scales	10	2	0.762	0.71, 0.70	
4	New	14	2	0.808	0.67, 0.57	
5	New	14	3	0.849	0.72, 0.52	
6	New	8	2	0.760	0.69, 0.60	
7	New	8	3	0.814	0.75, 0.66	
8	New	4	1	0.654	0.58, 0.57	
	Fractional Factoral Design (Training set =9 Dipeptides, Testset=49)					
33	MS_WHIM	27	1	0.959		0.462
34	MS_WHIM	12	2	0.956		0.535
35	MS_WHIM	6	1	0.875		0.215
36	T-Scales	65	2	0.981		0.298
37	T-Scales	20	2	0.950		0.565
38	T-Scales	10	2	0.907		0.373
12	New	14	2	0.948		0.563
13	New	8	2	0.951		0.678
14	New	4	1	0.664		0.680
	D-Optimal Design (Training set =9 Dipeptides, Testset=49)					
39	MS_WHIM	27	1	0.977		0.619
40	MS_WHIM	12	2	0.964		0.620
41	MS_WHIM	6	1	0.884		0.580
42	T-Scales	65	2	0.987		0.608
43	T-Scales	20	2	0.977		0.735
44	T-Scales	10	2	0.948		0.626
18	New	14	2	0.946		0.618
19	New	8	2	0.912		0.721
20	New	4	1	0.811		0.615
	Bitter Dipeptides DOE (Training set =10, Testset=38)					
21	MS_WHIM	27	2	0.955		0.700
22	MS_WHIM	12	2	0.906		0.609
23	MS_WHIM	6	1	0.851		0.531
45	T-Scales	65	2	0.759		0.621
46	T-Scales	20	2	0.702		0.703
47	T-Scales	10	2	0.700		0.653
24	New	8	2	0.935		0.683
25	New	4	2	0.912		0.691
26	New	4	1	0.858		0.694

Notes:-

a) First number=Q^2-LPO(10%):- 100 Iterations.

b) Second number=Q^2-LPO(50%):- 25 Iterations .

S3:- Datasets used in this Paper

Table A:- Data for ACE Inhibitor Dipeptides

Number	Peptide	Activity	Number	Peptide	Activity
1	VW	5.80	30	KG	2.49
2	IW	5.70	31	FG	2.43
3	IY	5.43	32	GS	2.42
4	AW	5.00	33	GV	2.34
5	RW	4.80	34	MG	2.32
6	VY	4.66	35	GK	2.27
7	GW	4.52	36	GE	2.27
8	VF	4.28	37	GT	2.24
9	AY	4.06	38	WG	2.23
10	IP	3.89	39	HG	2.20
11	RP	3.74	40	GQ	2.15
12	AF	3.72	41	GG	2.14
13	GY	3.68	42	QG	2.13
14	AP	3.64	43	SG	2.07
15	RF	3.64	44	LG	2.06
16	VP	3.38	45	GD	2.04
17	GP	3.35	46	TG	2.00
18	GF	3.20	47	EG	2.00
19	IF	3.03	48	DG	1.85
20	VG	2.96	49	PG	1.77
21	IG	2.92	50	LA	3.51
22	GI	2.92	51	KA	3.42
23	GM	2.85	52	RA	3.34
24	GA	2.70	53	YA	3.34
25	YG	2.70	54	AA	3.21
26	GL	2.60	55	FR	3.04
27	AG	2.60	56	HL	2.49
28	GH	2.51	57	DA	2.42
29	GR	2.49	58	EA	2.00

Activities given in negative $\log_{10}$ of the concentration in molar units. Data from Hellberg et al.[4] and Cushman et al.(original source)[5]

Table B:- Data for Bitter Dipeptides

Number	Peptide	Activity	Number	Peptide	Activity
1	GV	1.13	25	II	2.26
2	GL	1.68	26	IP	2.40
3	GI	1.70	27	IW	3.05
4	GP	1.35	28	IN	1.49
5	GF	1.80	29	ID	1.37
6	GW	1.89	30	IQ	1.49
7	GY	1.77	31	IE	1.37
8	AV	1.16	32	IK	1.65
9	AL	1.70	33	IS	1.49
10	AF	1.72	34	IT	1.49
11	VG	1.19	35	PA	1.32
12	VA	1.16	36	PL	2.22
13	VV	1.71	37	PI	2.33
14	VL	2.00	38	PY	1.80
15	LG	1.72	39	PF	2.80
16	LA	1.72	40	FG	1.77
17	LL	2.35	41	FL	2.87
18	LF	2.75	42	FP	2.70
19	LW	3.40	43	FF	3.10
20	LY	2.46	44	FY	3.13
21	IG	1.68	45	WE	1.56
22	IA	1.68	46	WW	3.60
23	IV	2.05	47	YL	2.40
24	IL	2.26	48	SL	1.49

Activities given as $\log_{10}$ of the reciprocal threshold molar concentration. Data from Hellberg et al.[4] and Asao et al.[6](original source)

Table C:- Data for Elastase Inhibitors

Number	Peptide	log(1/Km)	log(Kcat)	Number	Peptide	log(1/Km)	log(Kcat)
1	GA	-0.354	0.049	21	LG	-0.559	0.188
2	GV	-0.231	-0.105	22	LA	-0.326	1.107
3	GL	0.125	-0.796	23	LV	-0.185	0.780
4	GI	0.036	-0.337	24	LL	-0.340	0.412
5	GP	-0.322	0.068	25	LI	-0.452	0.991
6	GF	-0.260	-0.854	26	LP	0.142	0.996
7	AG	-0.686	0.107	27	LF	-0.152	0.033
8	AA	-0.225	0.936	28	IG	-0.708	0.607
9	AV	-0.049	0.979	29	IA	-0.371	1.210
10	AL	-0.193	0.515	30	IV	-0.252	0.927
11	AI	-0.083	0.951	31	IL	-0.312	0.358
12	AP	-0.207	1.436	32	II	-0.152	0.738
13	AF	-0.072	0.375	33	IP	0.071	1.274
14	VG	-0.493	0.467	34	IF	-0.188	0.158
15	VA	-0.215	1.185	35	FG	-0.812	0.037
16	VV	0.041	0.876	36	FA	-0.425	0.890
17	VL	-0.243	0.561	37	FV	-0.207	0.511
18	VI	-0.170	0.880	38	FL	-0.344	0.310
19	VP	0.033	1.476	39	FI	-0.255	0.490
20	VF	0.018	0.210	40	FP	-0.107	1.000
				41	FF	0.013	-0.409

Activities given as a rate constant:- k_{cat} in $\mu M^{-1}.sec^{-1}$. Data from Sandberg et al.[7] and Nomizu et al.[8](original source)

Table D:- Multi-laboratory set of results for ACE Inhibitors

Number	Peptide	Activity	Number	Peptide	Activity	Number	Peptide	Activity
1	AF	1.18	57	GY	2.32	113	RG	3.08
2	AF	1.88	58	HG	3.80	114	RL	3.39
3	AF	2.28	59	HY	1.42	115	RP	1.32
4	AG	3.40	60	IA	2.18	116	RP	1.96
5	AP	1.46	61	IF	2.97	117	RP	2.26
6	AP	2.36	62	IG	3.08	118	RW	1.20
7	AP	2.43	63	IL	1.74	119	RW	1.34
8	AW	1.08	64	IP	2.11	120	RY	1.71
9	AW	1.27	65	IR	2.84	121	RY	1.02
10	AW	1.00	66	IR	2.92	122	SF	2.11
11	AY	1.28	67	IW	0.18	123	SG	3.93
12	AY	2.00	68	IW	1.09	124	SY	1.82
13	AY	1.94	69	IW	0.67	125	TF	1.25
14	DF	2.56	70	IY	0.57	126	TF	1.95
15	DG	1.09	71	IY	0.32	127	TG	4.00
16	DG	4.15	72	IY	0.38	128	TK	3.21
17	DL	3.30	73	IY	0.36	129	TP	2.46
18	DM	2.78	74	IY	1.02	130	TP	3.32
19	DY	2.00	75	IY	0.79	131	VF	0.96
20	EG	3.87	76	IY	0.30	132	VF	1.72
21	FG	3.57	77	IY	0.43	133	VG	3.04
22	FL	1.20	78	KF	2.06	134	VK	1.11
23	FP	2.50	79	KF	1.45	135	VP	2.76
24	FQ	1.71	80	KG	3.51	136	VP	2.62
25	FY	1.40	81	KP	1.71	137	VQ	3.11
26	FY	1.63	82	KP	1.34	138	VW	0.15
27	FY	0.22	83	KP	1.48	139	VW	0.20
28	FY	0.57	84	KW	0.21	140	VW	0.52
29	FY	0.81	85	KW	1.03	141	VW	1.03
30	GA	3.30	86	KY	0.89	142	VW	0.40
31	GD	3.96	87	LF	2.54	143	VW	0.20
32	GE	3.85	88	LF	2.10	144	VW	0.23
33	GF	2.44	89	LF	3.52	145	VY	1.20
34	GF	2.85	90	LG	3.94	146	VY	1.41
35	GF	2.80	91	LW	0.83	147	VY	1.25
36	GG	3.94	92	LW	1.70	148	VY	1.55
37	GG	3.86	93	LW	1.37	149	VY	1.76
38	GH	3.49	94	LW	1.24	150	VY	1.64
39	GI	3.11	95	LY	0.83	151	VY	1.05
40	GI	3.08	96	LY	1.59	152	VY	1.34
41	GK	3.73	97	LY	0.81	153	WA	2.44
42	GL	3.40	98	LY	1.51	154	WG	3.77
43	GM	3.15	99	MF	1.65	155	WL	1.48
44	GP	2.56	100	MG	3.68	156	WM	1.98
45	GP	3.08	101	MW	1.00	157	YE	2.80
46	GP	2.65	102	MW	0.58	158	YG	3.04
47	GQ	3.73	103	MY	2.29	159	YG	3.18
48	GQ	3.75	104	NF	1.67	160	YG	3.30
49	GR	3.51	105	NP	3.36	161	YH	0.71
50	GS	3.58	106	NY	1.51	162	YL	1.21
51	GT	3.76	107	PG	4.23	163	YL	2.09
52	GV	3.66	108	PR	0.61	164	YL	1.91
53	GW	1.48	109	QG	4.00	165	YP	2.95
54	GY	2.41	110	QK	2.95	166	YP	2.86
55	GY	1.86	111	RF	1.97	167	YV	2.76
56	GY	2.42	112	RF	2.36	168	YW	1.02

Activities given in $\log_{10}$ of the inhibition concentration in micromolar units. Data from Wu et al.[9]

References

(1) Hellberg, S.; Sjostrom, M.; Skagerberg, B.; Wold, S. Peptide Quantitative Structure-Activity Relationships, a Multivariate Approach. *J. Med. Chem.* **1987**, *30* (7), 1126–1135 DOI: 10.1021/jm00390a003.

(2) Zaliani, A.; Gancia, E. MS-WHIM Scores for Amino Acids: A New 3D-Description for Peptide QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 525–533.

(3) 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 DOI: 10.1016/j.molstruc.2006.07.004.

(4) 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.

(5) Cushman, D. W.; Cheung, H.-S.; Sabo, E. F.; Ondetti, M. A. No Title. In *Proceedings of the A M Richards Symposium, May 8-9,1980.*; Horowitz, Z. P., Ed.; Urban & Schwarzenberg: Baltimore, MD, USA, 1981; pp 3–25.

(6) 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.

(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 DOI: 10.1021/jm9700575.

(8) 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.* **2009**, *42* (3), 216–226 DOI: 10.1111/j.1399-3011.1993.tb00135.x.

(9) 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 DOI: 10.1016/j.jmaa.2005.05.084.