Electronic Supplementary Material (ESI) for RSC Advances.

Supplementary Information

PepBio: Predicting the bioactivity of host defense peptides

Saw Simeon[1,2], Hao Li[1], Thet Su Win[1], Aijaz Ahmad Malik[1], Abdul Hafeez Kandhro[1,3], Theeraphon Piacham[4], Watshara Shoombuatong[1], Pornlada Nuchnoi[3,5], Jarl E. S. Wikberg[6], M. Paul Gleeson[7], and Chanin Nantasenamat[*1]

[1] *Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand*

[2] *Interdisciplinary Graduate Program in Bioscience, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand*

[3] *Center for Research and Innovation, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand*

[4] *Department of Clinical Microbiology and Applied Technology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand*

[5] *Department of Clinical Microscopy, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand*

[6] *Department of Pharmaceutical Biosciences, Uppsala University, Uppsala 751 24, Sweden*

[7] *Department of Biomedical Engineering, Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand*

[*]Corresponding author. E-mail: chanin.nan@mahidol.edu

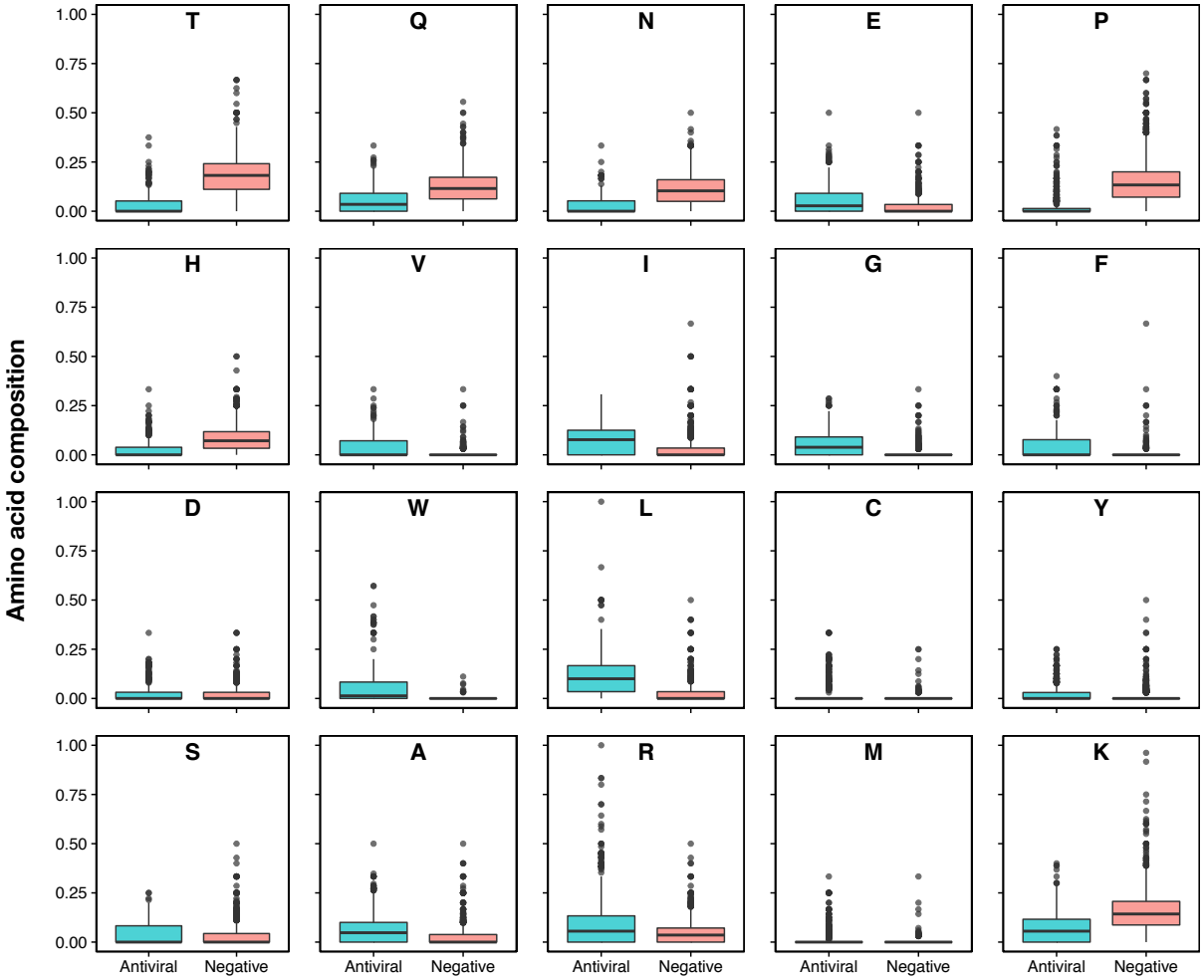

Figure S1: Box plots showing comparative analysis of anti–bacterial HDPs versus the negative set of peptides.

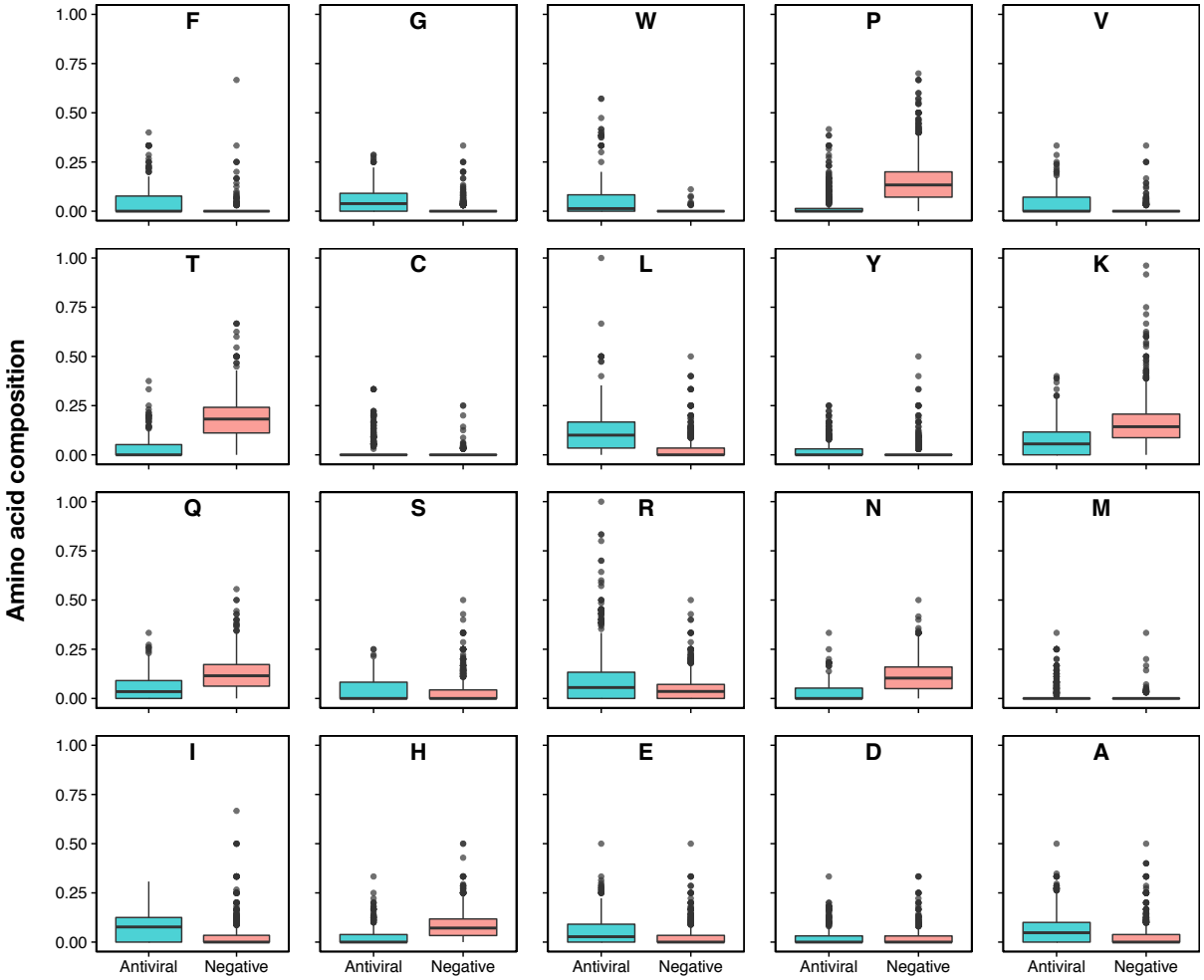

Figure S2: Box plots showing comparative analysis of anti–cancer HDPs versus the negative set of peptides.

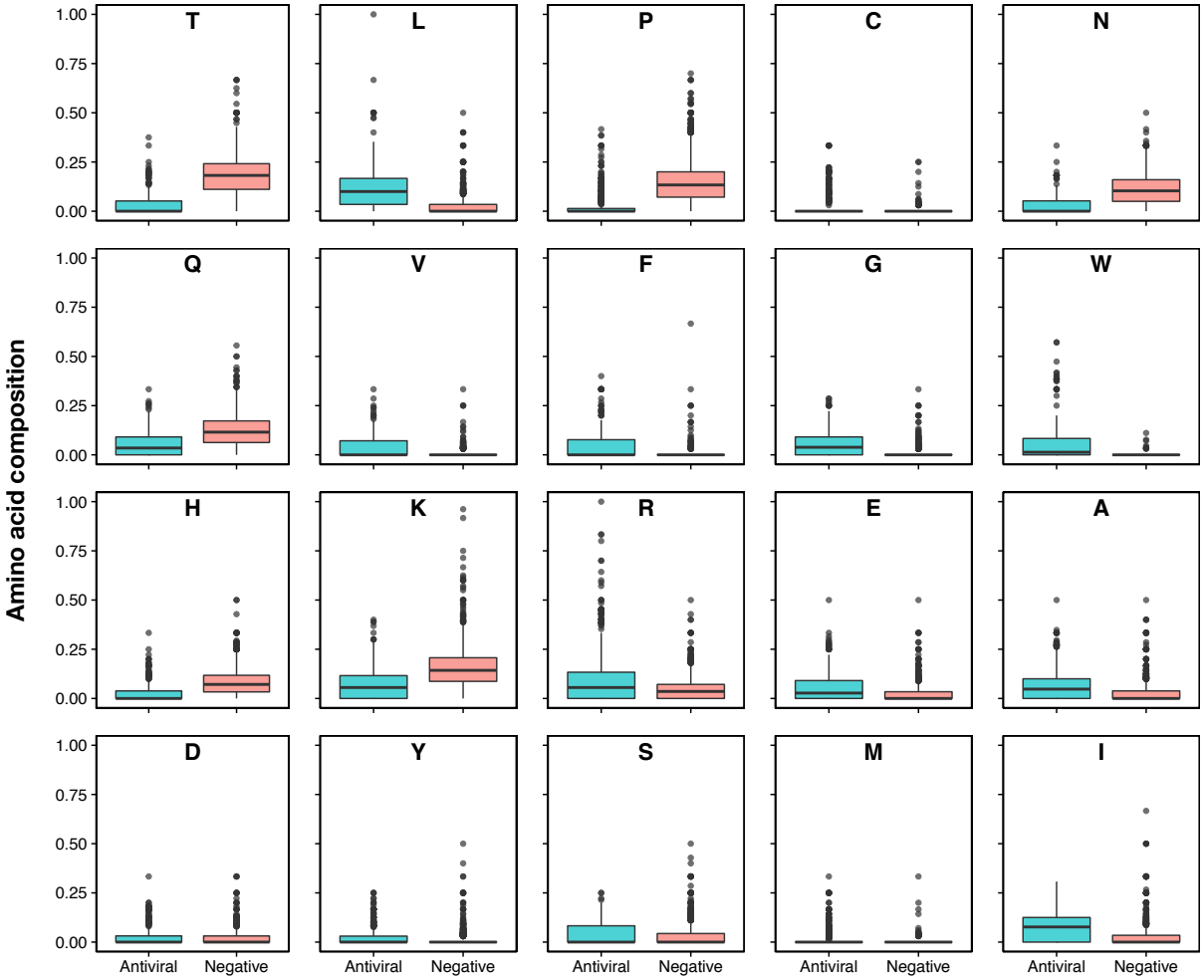

Figure S3: Box plots showing comparative analysis of anti–fungal HDPs versus the negative set of peptides.

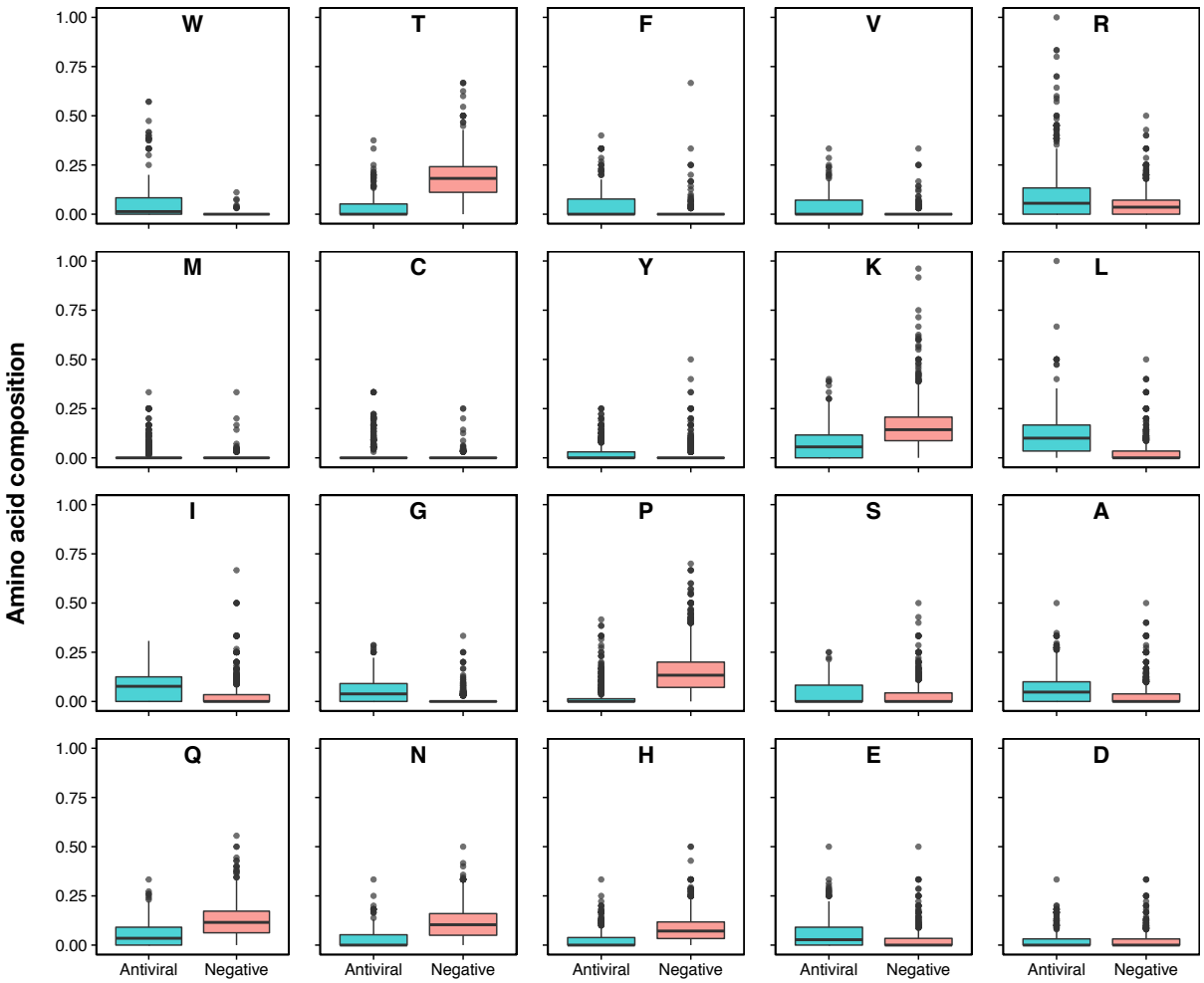

Figure S4: Box plots showing comparative analysis of anti–viral HDPs versus the negative set of peptides.

Table S1

Performance summary for predicting the bioactivity of HDPs using the DT classifier as a function of the combined set of AAC + DPC descriptors. It should be noted that the unit for accuracy, sensitivity and specificity is represented in percentage.

Model type	**Classes**	**Training set**				**10-fold CV**				**External set**			
		Ac (%)	**Sn (%)**	**Sp (%)**	**MCC**	**Ac (%)**	**Sn (%)**	**Sp (%)**	**MCC**	**Ac (%)**	**Sn (%)**	**Sp (%)**	**MCC**
Binary class	Antibacterial	99.08±0.14	99.35±0.17	98.87±0.22	0.98±0.01	96.37±0.28	96.43±0.36	96.32±0.34	0.93±0.01	96.50±0.58	96.54±1.01	96.48±0.86	0.93±0.01
	Anticancer	99.35±0.17	99.53±0.16	98.70±0.61	0.98±0.01	97.40±0.34	98.45±0.28	93.61±1.00	0.92±0.01	97.57±0.67	98.52±0.63	94.15±2.43	0.93±0.02
	Antifungal	99.07±0.15	99.35±0.17	98.86±0.24	0.98±0.00	96.40±0.30	96.48±0.42	96.34±0.37	0.93±0.01	96.46±0.65	96.61±1.09	96.36±0.91	0.93±0.01
	Antiviral	99.12±0.19	98.31±0.55	99.36±0.23	0.98±0.01	95.98±0.45	90.99±1.12	97.51±0.36	0.89±0.01	96.01±0.91	90.91±2.99	97.60±0.81	0.89±0.02
	HDPs	99.39±0.07	99.43±0.09	99.17±0.21	0.98±0.00	97.97±0.13	98.36±0.11	95.96±0.49	0.93±0.01	98.11±0.30	98.50±0.27	96.11±1.12	0.93±0.01
Multi–class	Overall	96.64±0.19	96.66±0.20	96.32±1.23	0.77±0.01	93.29±0.26	98.35±0.21	43.85±1.89	0.53±0.02	93.58±0.56	94.63±0.45	75.36±5.71	0.55±0.05
	Antibacterial	96.62±0.19	96.64±0.21	96.27±1.21	0.77±0.01	93.31±0.26	98.37±0.22	43.86±1.88	0.54±0.02	93.47±0.48	94.57±0.44	74.36±5.11	0.54±0.04
	Anticancer	96.63±0.19	96.64±0.20	96.55±1.24	0.77±0.01	93.31±0.26	98.37±0.21	43.93±1.74	0.54±0.02	93.72±0.55	94.71±0.48	76.27±5.01	0.55±0.05
	Antifungal	96.66±0.18	96.69±0.21	96.30±1.24	0.77±0.01	93.25±0.28	98.31±0.21	43.80±2.11	0.53±0.02	93.55±0.59	94.58±0.44	75.50±6.51	0.54±0.05
	Antiviral	96.65±0.20	96.68±0.20	96.14±1.33	0.77±0.02	93.27±0.24	98.33±0.21	43.80±1.82	0.53±0.02	93.60±0.61	94.67±0.45	75.33±6.23	0.55±0.05

Table S2

Performance summary for predicting the bioactivity of HDPs using the DT classifier as a function of CC descriptors. It should be noted that the unit for accuracy, sensitivity and specificity is represented in percentage.

Model type	**Classes**	**Training set**				**10-fold CV**				**External set**			
		Ac (%)	**Sn (%)**	**Sp (%)**	**MCC**	**Ac (%)**	**Sn (%)**	**Sp (%)**	**MCC**	**Ac (%)**	**Sn (%)**	**Sp (%)**	**MCC**
Binary class	Antibacterial	99.10±0.17	99.35±0.19	98.91±0.26	0.98±0.00	96.79±0.27	96.76±0.41	96.82±0.29	0.93±0.01	96.84±0.59	96.82±1.00	96.88±0.74	0.94±0.01
	Anticancer	99.32±0.16	99.42±0.15	98.49±0.58	0.98±0.01	97.74±0.27	98.58±0.24	94.96±0.75	0.93±0.01	97.82±0.68	98.51±0.59	95.38±2.29	0.94±0.02
	Antifungal	99.12±0.16	99.35±0.19	98.95±0.25	0.98±0.00	96.78±0.27	96.75±0.41	96.81±0.34	0.93±0.01	96.81±0.57	96.78±1.05	96.85±0.81	0.94±0.01
	Antiviral	98.90±0.25	97.71±0.82	99.26±0.28	0.97±0.01	95.81±0.39	90.97±1.03	97.27±0.30	0.88±0.01	95.98±0.89	91.03±2.47	97.49±0.90	0.89±0.03
	HDPs	99.39±0.08	99.45±0.09	99.10±0.21	0.98±0.00	98.12±0.15	98.55±0.10	95.95±0.58	0.93±0.01	98.15±0.31	98.59±0.29	95.91±1.23	0.93±0.01
Multi–class	Overall	82.09±0.30	86.95±1.34	72.67±3.50	0.34±0.02	92.45±0.27	97.82±0.22	38.52±2.24	0.46±0.02	79.32±0.61	77.68±1.80	47.40±6.88	0.17±0.04
	Antibacterial	67.10±0.51	85.80±4.33	65.99±0.50	0.26±0.01	92.45±0.30	97.80±0.22	38.76±2.40	0.46±0.02	63.31±0.68	53.01±6.21	64.06±0.33	0.08±0.03
	Anticancer	92.78±0.14	93.24±0.36	66.25±5.97	0.27±0.04	92.43±0.26	97.83±0.22	38.23±2.06	0.46±0.21	91.49±0.45	92.45±0.32	37.63±12.25	0.14±0.06
	Antifungal	72.56±0.27	72.73±0.38	66.25±5.61	0.12±0.02	92.45±0.24	97.83±0.21	38.56±2.14	0.46±0.02	69.84±0.68	71.08±0.24	22.49±8.26	-0.02±0.03
	Antiviral	95.83±0.27	96.05±0.28	92.13±1.92	0.71±0.02	92.45±0.29	97.82±0.23	38.53±2.36	0.46±0.02	92.64±0.63	94.19±0.45	65.43±0.66	0.46±0.05

Table S3

Performance summary for predicting the bioactivity of HDPs using the RF classifier as a function of AAC descriptors. It should be noted that the unit for accuracy, sensitivity and specificity is represented in percentage.

Model type	Classes	Training set				10-fold CV				External set			
		Ac (%)	Sn (%)	Sp (%)	MCC	Ac (%)	Sn (%)	Sp (%)	MCC	Ac (%)	Sn (%)	Sp (%)	MCC
Binary class	Antibacterial	99.99±0.02	99.98±0.03	99.99±0.02	1.00±0.00	98.77±0.10	98.15±0.20	99.26±0.12	0.98±0.00	98.77±0.36	98.09±0.72	99.31±0.41	0.98±0.01
	Anticancer	100.00±0.00	100.00±0.00	100.00±0.00	1.00±0.00	99.00±0.01	99.48±0.12	97.31±0.45	0.97±0.00	99.01±0.54	99.45±0.42	97.40±1.75	0.97±0.02
	Antifungal	99.99±0.02	99.98±0.03	99.99±0.02	1.00±0.00	98.80±0.12	98.18±0.18	99.28±0.15	0.98±0.00	98.84±0.31	98.20±0.64	99.34±0.35	0.98±0.01
	Antiviral	100.00±0.00	100.00±0.00	100.00±0.00	1.00±0.00	98.64±0.14	96.46±0.45	99.30±0.12	0.96±0.00	98.66±0.47	96.62±1.57	99.27±0.42	0.96±0.01
	HDPs	99.96±0.02	100.00±0.00	99.74±0.09	1.00±0.00	99.17±0.05	99.84±0.03	95.87±0.25	0.97±0.00	99.17±0.18	99.84±0.10	95.90±1.04	0.97±0.01
Multi–class	Overall	87.06±0.17	88.12±0.45	76.88±1.51	0.57±0.01	70.17±0.37	68.27±0.42	42.74±1.67	0.08±0.02	69.06±0.82	67.30±0.74	41.60±3.00	0.06±0.03
	Antibacterial	76.33±0.27	76.49±0.95	76.29±0.43	0.47±0.01	48.90±0.45	22.66±0.83	58.60±0.33	-0.17±0.01	47.02±0.79	20.33±1.36	57.40±0.53	-0.21±0.02
	Anticancer	94.58±0.12	95.81±0.43	68.25±3.75	0.51±0.03	86.31±0.36	91.06±0.41	23.36±3.63	0.12±0.04	85.65±0.86	90.66±0.59	21.65±5.32	0.11±0.06
	Antifungal	78.82±0.21	81.51±0.32	66.46±0.77	0.41±0.01	52.95±0.40	65.75±0.26	9.68±0.78	-0.23±0.01	51.22±1.00	64.80±0.52	7.66±1.20	-0.26±0.02
	Antiviral	98.50±0.10	98.65±0.09	96.51±1.09	0.89±0.01	92.53±0.26	93.60±0.18	79.31±1.93	0.59±0.02	92.34±0.63	93.41±0.50	79.69±4.98	0.59±0.04

Table S4

Performance summary for predicting the bioactivity of HDPs using the RF classifier as a function of DPC descriptors. It should be noted that the unit for accuracy, sensitivity and specificity is represented in percentage.

Model type	Classes	Training set				10-fold CV				External set			
		Ac (%)	Sn (%)	Sp (%)	MCC	Ac (%)	Sn (%)	Sp (%)	MCC	Ac (%)	Sn (%)	Sp (%)	MCC
Binary class	Antibacterial	99.99±0.02	100.00±0.00	99.98±0.03	0.99±0.00	97.22±0.17	97.30±0.24	97.16±0.22	0.94±0.00	97.40±0.50	97.70±0.78	97.16±0.78	0.95±0.01
	Anticancer	100.00±0.1	100.00±0.00	99.98±0.07	1.00±0.00	98.56±0.16	99.57±0.11	94.81±0.60	0.96±0.01	98.68±0.52	99.52±0.31	95.54±2.27	0.96±0.02
	Antifungal	99.99±0.01	100.00±0.00	99.99±0.02	1.00±0.00	97.27±0.17	97.30±0.24	97.24±0.23	0.94±0.00	97.41±0.50	97.40±0.79	97.41±0.64	0.95±0.01
	Antiviral	99.99±0.02	100.00±0.00	99.99±0.03	1.00±0.00	97.79±0.21	92.04±0.78	99.52±0.11	0.94±0.01	97.97±0.64	92.62±2.47	99.56±0.35	0.94±0.02
	HDPs	99.97±0.01	99.99±0.01	99.82±0.07	1.00±0.00	98.82±0.07	99.43±0.08	95.79±0.32	0.96±0.00	98.90±0.21	99.47±0.19	96.11±1.03	0.96±0.01
Multi–class	Overall	86.98±0.16	88.13±0.31	78.04±1.27	0.56±0.01	70.99±0.35	68.93±0.35	43.27±1.47	0.09±0.01	69.92±0.82	67.84±0.64	42.51±2.87	0.07±0.03
	Antibacterial	76.13±0.25	77.23±0.68	75.79±0.36	0.47±0.01	49.86±0.38	24.08±0.75	59.17±0.27	-0.15±0.01	47.98±0.91	21.35±1.20	57.91±0.56	-0.19±0.02
	Anticancer	94.64±0.12	95.43±0.21	72.95±3.02	0.49±0.02	86.89±0.36	90.95±0.20	21.83±2.80	0.10±0.02	86.48±0.81	90.67±0.37	20.98±4.73	0.09±0.04
	Antifungal	78.73±0.21	81.32±0.30	66.57±0.90	0.41±0.01	53.93±0.40	66.40±0.24	10.80±0.86	-0.21±0.01	52.10±0.85	65.33±0.45	8.13±1.28	-0.25±0.01
	Antiviral	98.43±0.08	98.55±0.08	96.86±0.81	0.89±0.01	93.27±0.28	94.30±0.20	81.30±1.94	0.63±0.02	93.14±0.71	94.00±0.53	83.03±4.93	0.63±0.04

Table S5

Performance summary for predicting the bioactivity of HDPs using the RF classifier as a function of the combined set of AAC + DPC descriptors. It should be noted that the unit for accuracy, sensitivity and specificity is represented in percentage.

Model type	Classes	Training set				10-fold CV				External set			
		Ac (%)	Sn (%)	Sp (%)	MCC	Ac (%)	Sn (%)	Sp (%)	MCC	Ac (%)	Sn (%)	Sp (%)	MCC
Binary class	Antibacterial	99.99±0.02	100.00±0.00	99.98±0.03	1.00±0.00	98.87±0.11	98.42±0.16	99.22±0.13	0.98±0.00	98.88±0.32	98.46±0.63	99.20±0.38	0.98±0.01
	Anticancer	99.94±0.03	100.00±0.00	99.72±0.16	1.00±0.00	99.04±0.13	99.67±0.11	96.70±0.49	0.97±0.00	99.12±0.44	99.66±0.31	97.11±1.69	0.97±0.01
	Antifungal	99.99±0.02	100.00±0.00	99.98±0.03	1.00±0.00	98.86±0.11	98.40±0.14	99.22±0.15	0.98±0.00	98.92±0.31	98.46±0.61	99.27±0.37	0.98±0.01
	Antiviral	100.00±0.00	100.00±0.00	100.00±0.00	1.00±0.00	98.83±0.16	96.16±0.60	99.64±0.09	0.97±0.01	98.78±0.47	96.08±1.75	99.59±0.32	0.97±0.01
	HDPs	99.97±0.01	100.00±0.00	99.81±0.08	1.00±0.00	99.23±0.06	99.85±0.04	96.20±0.29	0.97±0.00	99.30±0.18	99.85±0.10	96.63±1.01	0.98±0.01
Multi–class	Overall	87.10±0.17	88.19±0.38	77.86±1.31	0.57±0.01	70.49±0.36	68.52±0.35	42.99±1.53	0.08±0.02	69.39±0.89	67.55±0.68	41.91±2.90	0.06±0.03
	Antibacterial	76.37±0.24	77.04±0.80	76.16±0.35	0.47±0.01	49.18±0.41	23.21±0.69	58.75±0.29	-0.17±0.01	47.33±0.95	20.90±1.25	57.57±0.62	-0.20±0.02
	Anticancer	94.69±0.14	95.59±0.32	71.78±3.04	0.50±0.02	86.51±0.40	90.93±0.28	21.66±3.18	0.10±0.03	86.11±0.86	90.60±0.34	20.83±4.76	0.09±0.04
	Antifungal	78.87±0.21	81.53±0.32	66.63±0.78	0.41±0.01	53.31±0.36	66.02±0.23	10.21±0.80	-0.22±0.01	51.33±0.94	64.91±0.48	8.16±1.25	-0.25±0.01
	Antiviral	98.49±0.09	98.61±0.09	96.87±1.08	0.89±0.01	92.96±0.28	93.92±0.19	81.33±1.84	0.62±0.02	92.77±0.78	93.79±0.63	81.07±4.98	0.62±0.05

Table S6

Performance summary for predicting the bioactivity of HDPs using the RF classifier as a function of CC descriptors. It should be noted that the unit for accuracy, sensitivity and specificity is represented in percentage.

Model type	Classes	Training set				10-fold CV				External set			
		Ac (%)	Sn (%)	Sp (%)	MCC	Ac (%)	Sn (%)	Sp (%)	MCC	Ac (%)	Sn (%)	Sp (%)	MCC
Binary class	Antibacterial	99.97±0.01	99.94±0.03	100.00±0.00	1.00±0.00	97.77±0.16	97.13±0.25	98.27±0.17	0.95±0.00	97.82±0.51	97.16±0.88	98.34±0.64	0.96±0.01
	Anticancer	99.96±0.03	99.95±0.04	100.00±0.00	0.99±0.00	98.57±0.16	99.22±0.12	96.17±0.51	0.96±0.01	98.76±0.52	99.35±0.46	96.62±2.15	0.96±0.02
	Antifungal	99.97±0.01	99.94±0.03	100.00±0.00	0.99±0.00	97.79±0.14	97.17±0.23	98.27±0.17	0.96±0.00	97.87±0.50	97.16±0.91	98.43±0.58	0.96±0.01
	Antiviral	99.95±0.02	99.80±0.09	100.00±0.00	0.99±0.00	96.89±0.25	92.24±0.71	98.29±0.20	0.91±0.01	96.87±0.76	92.34±2.62	98.22±0.70	0.91±0.02
	HDPs	99.97±0.01	99.99±0.01	99.85±0.06	0.99±0.00	98.70±0.07	99.50±0.06	94.76±0.28	0.95±0.00	98.75±0.27	99.56±0.18	94.72±1.26	0.96±0.01
Multi–class	Overall	86.93±0.16	87.99±0.43	76.38±1.43	0.57±0.01	69.71±0.36	67.62±0.42	40.87±1.80	0.06±0.02	68.68±0.86	66.74±0.74	40.21±2.99	0.04±0.03
	Antibacterial	76.15±0.21	76.39±0.90	76.09±0.39	0.47±0.01	48.37±0.43	21.12±0.82	58.19±0.32	-0.19±0.01	46.72±0.83	19.04±1.37	57.14±0.53	-0.22±0.02
	Anticancer	94.48±0.12	95.75±0.45	67.23±3.37	0.50±0.03	86.06±0.33	90.93±0.41	22.41±3.87	0.12±0.04	85.59±0.90	90.58±0.53	21.55±5.04	0.10±0.05
	Antifungal	78.65±0.21	81.17±0.26	66.60±0.72	0.41±0.01	52.81±0.41	65.48±0.26	9.15±0.85	-0.23±0.01	50.88±1.04	64.47±0.50	7.22±1.07	-0.27±0.01
	Antiviral	98.43±0.10	98.64±0.10	95.59±1.24	0.89±0.01	91.59±0.26	92.97±0.17	73.72±2.16	0.53±0.02	68.68±0.86	66.74±0.74	40.21±2.99	0.04±0.03

Table S7

Frequency count for the amino acid class composition for the twenty most informative DPC descriptors

Bioactivity	**Hydrophobicity**	**van der Waals volume**	**Polarity**	**Polarizability**	**Charge**	**Secondary structure**	**Solvent accessibility**
Antimicrobial							
11	1	2	2	1	0	1	5
12	2	4	1	3	2	3	3
13	3	3	2	3	0	1	0
21	4	3	3	4	3	3	3
22	3	2	2	3	14	3	1
23	2	3	3	3	0	1	2
31	2	0	3	0	1	4	0
32	1	2	2	2	0	2	3
33	2	1	2	1	0	2	3
Anticancer							
11	5	2	2	1	2	6	5
12	1	3	1	2	3	1	2
13	1	2	3	2	0	2	1
21	5	0	1	1	4	6	2
22	1	3	1	4	11	0	5
23	1	2	4	2	0	2	0
31	2	1	1	1	0	2	0
32	2	4	1	4	0	0	5
33	2	3	6	3	0	1	0
Antifungal							
11	2	3	3	3	2	5	8
12	1	3	1	3	6	4	1
13	5	1	1	0	0	1	0
21	1	1	3	1	1	3	6
22	3	1	3	1	11	2	2
23	3	0	1	1	0	1	0
31	1	3	5	1	0	1	1
32	1	5	1	7	0	2	1
33	3	3	2	3	0	1	1
Antiviral							
11	3	3	4	1	1	4	6
12	3	2	1	2	1	3	2
13	1	1	2	1	0	1	0
21	1	4	2	4	3	3	1
22	3	4	3	6	14	1	3
23	2	2	1	2	1	2	3
31	2	1	1	1	0	2	1
32	1	2	3	2	0	3	1
33	4	1	3	1	0	1	3

Table S8

Amino acid attributes and the division of the amino acids into three groups for each attribute

	Group 1	**Group 2**	**Group 3**
Hydrophobicity	Polar	Neutral	Hydrophobicity
	R, K, E, D, Q, N	G, A, S, T, P, H, Y	C, L, V, I, M, F, W
Normalized van der Waals volume	0-2.78	2.95-4.0	4.03-8.08
	G, A, S, T, P, D, C	N, V, E, Q, I, L	M, H, K, F, R, Y, W
Polarity	4.9-6.2	8.0-9.2	10.4-13.0
	L, I, F, W, C, M, V, Y	P, A, T, G, S	H, Q, R, K, N, E, D
Polarizability	0-1.08	0.128-0.186	0.219-0.409
	G, A, S, D, T	C, P, N, V, E, Q, I, L	K, M, H, F, R, Y, W
Charge	Positive	Neutral	Negative
	K, R	A, N, C, Q, G, H, I, L	D, E
		M, F, P, S, T, W, Y, V	
Secondary structure	Helix	Strand	Coil
	E, A, L, M, Q, K, R, H	V, I, Y, C, W, F, T	G, N, P, S, D
Solvent accessibility	Buried	Exposed	Intermediate
	A, L, F, C, G, I, V, W	R, K, Q, E, N, D	M, S, P, T, H, Y