Electronic Supplementary Material (ESI) for RSC Advances

Contents

Appendix 1: List of available nanoparticles datasets

No.	Dataset	Description	Sample Size	Endpoint	Available QNAR models
1	Shaw, S.Y. et. al[1]	4 metal cores	51	Apoptosis	Model classification accuracy of 73%
2	Weissleder, R. et. al[2]	1 metal core	146	Cellular uptake by pancreatic cancer cells	$R^2 = 0.8$
3	Zhou, H.Y. et. al[3]	Carbon NanoTubes	84	Protein binding, cytotoxicity, immune responses	Nil
4	Lanone, S. et. al[4]	MNPs	24	TC_{50} of 2 human pulmonary cell lines: A549 and THP-1	Nil
5	Tassa, C. et.al[5]	MNPs	20	Ka, Kd of aurora A kinase	Nil
6	Puzyn, T. et. Al[6]	Metal oxide nanoparticles	17	EC_{50} of *E.Coli*	$R^2 = 0.85$ $Q^2_{cv} = 0.77$ $Q^2_{ext} = 0.83$
7	Liu, R. et.al[7]	Oxide nanoparticles	9	Cytotoxicity	100% classification accuracy (internal & external validations)

Appendix 2: List of compounds conjugated to nanoparticles and its corresponding cellular uptake

Fourches et.al.[8] ID No	Weissleder et.al.[2] ID No	SMILES Strings	Cellular Uptake (no. of nanoparticles per cell)	Prediction by final consensus model[a]
1	261-13-1	FC(F)(F)C(=O)OC(=O)C(F)(F)F	14791	
2	261-13-2	FC(F)(Cl)C(=O)OC(=O)C(F)(F)Cl	8913	
3	261-13-3	FC(F)(F)C(F)(F)C(=O)OC(=O)C(F)(F)C(F)(F)F	12023	
4	261-13-4	CC1(C)CC(=O)OC1=O	12882	
5	261-13-5	O=C1OC(=O)C=C1	9550	
6	261-13-6	CC1=CC(=O)OC1=O	3802	FP
7	261-13-7	CC1=C(C)C(=O)OC1=O	3020	FP
8	261-13-8	CCCCCC(=O)OC(=O)CCCCC	4467	FP
9	261-13-9	CC1CC(=O)OC1=O	4365	FP
10	261-14-1	O=C1OC(=O)c2cc(ccc12)C(=O)c1ccc2C(=O)OC(=O)c2c1	3236	
11	261-14-10	O=C1OC(=O)c2cc(ccc12)N(=O)=O	1862	FP
12	261-14-11	Brc1ccc2C(=O)OC(=O)c3cccc1c23	4266	FP
13	261-14-12	O=C1OC(=O)c2ccc3C(=O)OC(=O)c4ccc1c2c34	4677	Out of the Applicability Domain
14	261-14-13	Fc1c(F)c(F)c2C(=O)OC(=O)c2c1F	6761	
15	261-14-14	O=C1OC(=O)c2cc(cc3cccc1c23)N(=O)=O	12882	
16	261-14-15	Oc1cccc2C(=O)OC(=O)c12	9333	
17	Unable to match: Fourches et.al's structure showed a double bond in the bicyclo (2,2,2) oct-5-ene group but this was not observed in the 2D structure of compound 261-14-16 given in Weissleder et.al. paper.	O=C1OC(=O)C2C3CCC(C=C3)C12	7943	
18	261-14-17	Clc1ccc2NC(=O)OC(=O)c2c1	15136	
19	261-14-18	O=C1OS(=O)(=O)c2ccccc12	7586	
20	261-14-19	ClC1=C(Cl)C(=O)OC1=O	6918	
21	261-14-2	CC(=O)SC1CC(=O)OC1=O	3890	
22	261-14-20	Clc1cc2C(=O)OC(=O)c2cc1Cl	13183	
23	261-14-21	O=C1OC(=O)C2C3OC(C=C3)C12	6607	
24	261-14-22	O=C1OC(=O)C2C3C=CC(C12)C1C3C(=O)OC1=O	4266	
25	261-14-23	O=C1OC(=O)C2CC=CCC12	7762	
26	261-14-24	O=C1OC(=O)c2ccccc2-c2ccccc12	5888	
27	261-14-3	O=C1OC(=O)c2ccc(c3cccc1c23)N(=O)=O	8511	
28	261-14-5	O=C1OC(=O)C2C1C1C2C(=O)OC1=O	5888	
29	261-14-6	CCCCCCCCCCCC(=O)OC(=O)CCCCCCCCCCC	6607	
30	261-14-7	OC(=O)c1ccc2C(=O)OC(=O)c2c1	3548	
31	261-14-8	Cc1ccc2C(=O)OC(=O)c2c1	9550	
32	261-14-9	O=C1OC(=O)c2c1cccc2N(=O)=O	3162	FP
33	261-15-25	O=C1Cc2ccccc2C(=O)O1	6026	

34	261-15-26	O=C1CCCC(=O)O1	11749	
35	261-15-27	O=C1CN(CCN2CC(=O)OC(=O)C2)CC(=O)O1	8511	
36	261-15-28	O=C1Nc2ccccc2C(=O)O1	27542	
37	261-15-29	CN1C(=O)OC(=O)c2ccccc12	2291	FP
38	261-15-30	CC1CC(=O)OC(=O)C1	8128	
39	261-15-31	O=C1OC(=O)C2=C1CCCC2	5370	
40	261-15-32	CC(=O)OC1C(OC(C)=O)C(=O)OC1=O	8128	
41	261-16-1	Brc1c(Br)c(Br)c2C(=O)OC(=O)c2c1Br	6310	
42	261-16-3	O=C1OC(=O)C2CCCCC12	8511	
43	261-16-11	O=C1OC(=O)C2=C1CCC2	4898	FP
44	261-16-12	ICC(=O)OC(=O)CI	2630	FP
45	261-16-13	ClCC(=O)OC(=O)CCl	4266	Out of the Applicability Domain
46	261-16-14	ClC1=C(Cl)C2(Cl)C3C(C(=O)OC3=O)C1(Cl)C2(Cl)Cl	2951	
47	Both papers showed palmitic anhydride (261-16-15), however the 2D chemical structure given in Weissleder et.al. supplementary material differs from its chemical name.	CCCCCCCCCCCCCCCC(=O)OC(=O)CCCCCCCCCCCCCC	3548	
48	261-16-16	Nc1ccc2C(=O)OC(=O)c3cccc1c23	4365	
49	261-16-2	CCCCCCCCCC(=O)OC(=O)CCCCCCCCC	10715	
50	261-16-10	O=C1CC2(CCCC2)CC(=O)O1	11482	
51	Unable to match: Fourches et.al. structure (261-16-4) does not have a double bond at the norborne group	O=C1OC(=O)C2C3CCC(C3)C12	8710	
52	261-16-5	O=C1OC(=O)c2cccc3cccc1c23	9120	
53	261-16-6	O=C1CCC(C(=O)O1)c1ccccc1	10471	
54	261-16-7	Clc1c(Cl)c(Cl)c2C(=O)OC(=O)c2c1Cl	6761	
55	261-16-8	Clc1ccc(Cl)c2C(=O)OC(=O)c12	7943	
56	261-16-9	CC1(C)CCC(=O)OC1=O	8710	
57	261-44-1	CCCCCN	6026	
58	261-44-10	CC(C)CC(C)N	7079	
59	261-44-11	NC1C(O)CC(CO)C(O)C1O	2291	
60	261-44-2	CCCCCCN	5623	
61	261-44-4	CC(C)(C)N	7244	
62	261-44-5	CC(C)CN	5248	
63	Unable to find compound 63 in Weissleder et.al.'s paper	CC(C)(C)CN	5623	
64	261-44-7	CC(C)CCN	6761	
65	261-44-8	CCC(N)CC	6457	
66	261-44-9	CCC(C)(C)N	11749	
67	261-45-1	NCCN	2884	
68	261-45-10	CCCCCCCCCCCCCCCN	11482	
69	261-45-2	NCCCN	3090	
70	261-45-3	NCCCCN	3020	
71	261-45-4	NCCCCCCN	4169	
72	261-45-5	CCCCC(CC)CN	8913	

73	261-45-6	CCCCCCCCCCCCCCCCN	9333	
74	261-45-7	CCCCCC(C)N	4266	FP
75	261-45-8	CCCCCCCCCCCCCCCN	18621	
76	261-46-1	NCCNCCN	5888	FN
77	261-46-10	NCC12CC3CC(CC(C3)C1)C2	692	
78	261-46-11	NCCc1ccc(O)c(O)c1	339	
79	261-46-12	NCCc1ccc(O)cc1	589	
80	261-46-5	NCCCNCCCCNCCN	257	
81	261-46-6	NCCNCCCNCCN	170	
82	261-46-7	NCCNCCNCCNCCNCCN	347	
83	261-46-8	NC12CC3CC(CC1C3)C2	1318	
84	261-46-9	NC1C2CC3CC(C2)CC1C3	1514	
85	261-47-1	NCC(O)=O	372	
86	261-47-10	COC(=O)C(N)Cc1ccccc1	2455	
87	261-47-11	NC(CO)C(O)=O	2291	
88	261-47-12	CC(O)C(N)C(O)=O	1622	
89	261-47-13	NC(Cc1c[nH]c2ccccc12)C(O)=O	1549	
90	261-57-14	NC(Cc1ccc(O)cc1)C(O)=O	1175	
91	261-47-15	CC(C)C(N)C(O)=O	1862	
92	261-47-16	NCCCCC(N)C(O)=O	1778	
93	261-47-17	NC(C(O)=O)c1ccc(Cl)cc1	1148	
94	261-47-2	CC(N)C(O)=O	794	
95	261-47-3	NC(CCCNC(N)=N)C(O)=O	1413	
96	261-47-4	NC(CC(O)=O)C(O)=O	1950	
97	261-47-5	NC(CCC(N)=O)C(O)=O	2089	
98	261-47-6	NC(CCC(O)=O)C(O)=O	2512	
99	261-47-7	NC(Cc1c[nH]cn1)C(O)=O	2399	
100	261-47-8	CSCCC(N)C(O)=O	1698	
101	261-47-9	NC(Cc1ccccc1)C(O)=O	1950	
102	261-9-1	O=C1CCC(=O)O1	17378	
103	261-9-2	CC(=O)OC(C)=O	11220	
104	261-9-3	C=C1CC(=O)OC1=O	10965	
105	261-9-4	O=C1COCC(=O)O1	9772	
106	261-9-5	O=C1OC(=O)c2ccccc12	7943	
107	261-9-6	OC(=O)CC1CC(=O)OC1=O	10715	
108	261-9-7	Fc1ccc(F)c2C(=O)OC(=O)c12	8128	
109	261-9-8	OC(=O)CN(CCN1CC(=O)OC(=O)C1)CC N1CC(=O)OC(=O)C1	12589	
	261-14-4	CCCCCCCCCC=CCC1CC(=O)OC1=O	-	N.A.
	261-17-1	CCC(=O)OC(=O)CC	-	N.A.
	261-17-2	CC(C)(C)C(=O)OC(=O)C(C)(C)C	-	N.A.
	261-17-3	C(=O)(C(Cl)(Cl)Cl)OC(=O)C(Cl)(Cl)Cl	-	N.A.
	261-17-4	CC1CCC2C(C1)C(=O)OC2=O	-	N.A.
	261-17-5	CC=CC(=O)OC(=O)C=CC	-	N.A.
	261-17-6	C(=O)(C(C(C(F)(F)F)(F)F)(F)F)OC(=O)C(C(C(F)(F)F)(F)F)(F)F	-	N.A.
	261-17-7	O1C(=O)C2C3CC(C2C1=O)C=C3C	-	N.A.
	261-17-7	O1C(=O)C2[C@@H](C3CC2C=C3C)C1=	-	N.A.

	O		
261-17-7	O1C(=O)[C@H]2[C@H](C3C=C(C2C3)C)C1=O	-	N.A.
261-17-7	O1C(=O)[C@@H]2[C@H](C3C=C(C2C3)C)C1=O	-	N.A.
261-17-8	C(C(=O)OC(=O)C(Cl)Cl)(Cl)Cl	-	N.A.
261-17-9	CCCCC(=O)OC(=O)CCCC	-	N.A.
261-17-10	CCCC(=O)OC(=O)CCC	-	N.A.
261-17-11	CC(C)C(=O)OC(=O)C(C)C	-	N.A.
261-17-12	C1(=O)C(C(C(C(=O)O1)(F)F)(F)F)(F)F	-	N.A.
261-17-13	C1=CC=C(C=C1)C(=O)OC(=O)C2=CC=CC=C2	-	N.A.
261-17-14	O1C(=O)c2c3c(ccc4c3c(c3c5c6c(cc3)C(OC(=O)c6ccc45)=O)cc2)C1=O	-	N.A.
261-44-3	CCC(C)N	-	N.A.
261-44-6	CCCN(C)C	-	N.A.
261-45-9	CCCCCCCCCCCN	-	N.A.
261-46-2	C(CCCNCCCCCCN)CCN	-	N.A.
261-46-3	C(CN)CNCCNCCCN	-	N.A.
261-46-4	C(CCNCCCN)CN	-	N.A.
261-50-1	CC(=O)NC(CCCN=C(N)N)C(=O)O	-	N.A.
261-50-2	CC(C)(C)OC(=O)NC(CCCN=C(N)N)C(=O)O	-	N.A.
261-50-3	C1=CC(=C(C=C1[N+](=O)[O-])[N+](=O)[O-])NC(CCCN=C(N)N)C(=O)O	-	N.A.
261-50-4	OC(=O)CN1CCNC1=N	-	N.A.
261-50-5	OC(=O)C(N)CCCNC(N)=N	-	N.A.
261-50-6	C(CCN=C(N)N)CN.OS(=O)(=O)O	-	N.A.
261-50-7	C(CC(C(=O)N)N)CN=C(N)N.Cl.Cl	-	N.A.
261-50-8	C(CC(C(=O)O)N)CN=C(N)N	-	N.A.
261-50-9	O(C(=O)C(N)CCCNC(N)=N)CC	-	N.A.
261-50-10	O=C(NO)C(N)CCCNC(N)=N	-	N.A.
261-50-11	O(C)c1cc(NC(=O)C(N)CCCNC(N)=N)cc2c1cccc2	-	N.A.
261-50-12	C(CC(C(=O)NC(CCN)C(=O)N)N)CN=C(N)N[N+](=O)[O-]	-	N.A.
261-53-1	CCN(CCO)C1=CC=C(C=C1)N=NC2=CC=C(C=C2)[N+](=O)[O-]	-	N.A.
261-53-2	SCC(C(=O)C1CCCC1C(O)=O)C	-	N.A.

[a] FN: False Negative, FP: False Positive, N.A.: Not used in this study. Those with no entries were correctly predicted.

Appendix 3: Results for the rigorous validation process

1. Training and Validation Sets

Approximately 80% of the dataset were grouped into the training set while the remaining 20% were grouped into the validation set. All training sets and validation sets had almost equal proportion of nanoparticles with good and poor cellular uptake (Table A3.1).

Table A3.1: Details of the distribution of nanoparticles in the training and validation sets

Runs	Training Set		Validation Set	
	Pos	Neg	Pos	Neg
1	43	39	13	10
2	44	35	12	14
3	46	40	10	9
4	46	40	10	9
5	45	42	11	7

2. Model Selections of the 5 Runs

Details on the percentage of suitable kNN, SVM, Logistic Regression (LR) and Naïve Bayes (NB) candidate models identified using the 2 criteria are as follows:

Table A3.2: Percentage of kNN, SVM, LR and NB models selected through the 2 criteria

Runs	kNN (%)	SVM (%)	LR (%)	NB (%)
1	5.3	7.7	5.0	3.0
2	17.8	14.6	12.0	23.0
3	13.0	10.3	10.0	10.0
4	9.4	9.2	0.0	21.0
5	5.3	3.0	3.0	2.0

Approximately 10% of the candidate models (ranging from 83 to 341) were identified as suitable models for consensus modelling. Among the top 5 candidate models used for consensus modelling for each of the 5 runs, 21 of them were kNN models, 3 were SVM models and 1 was LR model.

References

1. S. Y. Shaw, E. C. Westly, M. J. Pittet, A. Subramanian, S. L. Schreiber and R. Weissleder, *Proc Natl Acad Sci U S A*, 2008, **105**, 7387-7392.
2. R. Weissleder, K. Kelly, E. Y. Sun, T. Shtatland and L. Josephson, *Nat Biotechnol*, 2005, **23**, 1418-1423.
3. H. Zhou, *Nano Letters 2008*, 2008, **8**, 859-865.
4. C. Sayes and I. Ivanov, *Risk Anal*, 2010, **30**, 1723-1734.
5. T. Shaw S.Y, C., *Bioconjugate Chem. 2010*, 2010, **21**, 14-19.
6. T. Puzyn, *Nat Nanotechnol*, 2011, **6**.
7. R. Liu, R. Rallo, S. George, Z. Ji, S. Nair, A. E. Nel and Y. Cohen, *Small*, 2011, **7**, 1118-1126.
8. D. Fourches, D. Pu, C. Tassa, R. Weissleder, S. Y. Shaw, R. J. Mumper and A. Tropsha, *ACS Nano*, 2010, **4**, 5703-5712.