Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supporting Tables and Figures

Deep Learning-Generated Potential NMDA Receptor Antagonists Reveal Advantages and Limitations of Artificial Intelligence-Based Molecular Generation

Katherine J. Schultz,[†] Sean M. Colby,[†] Yasemin Yesiltepe,[†] Jamie R. Nuñez,[†] Monee Y. McGrady,[†] Ryan R. Renslow[†*]

[†] Pacific Northwest National Laboratory, Richland, WA, USA.

Table of Contents

Table S1. Chemical properties used in QSAR activity models.

Molecular weight	Atom count	Molecular polarizability
Avg molecular polarizability	principal components of polarizability tensor (axx, ayy, azz)	Water accessible surface area incl. ASA+, ASA-, ASA_H, and ASA_P
Aromatic atom count	Aromatic bond count	Aromatic ring count
Asymmetric atom count	Balaban index	Carboaromatic ring count
Carbo ring count	Chiral center count	Cyclomatic number
Dreiding energy	Fused aliphatic ring count	Fused aromatic ring count
Fused ring count	Harary index	Hyper wiener index
Maximal projection area	Maximal projection radius	Minimal projection area
Minimal projection radius	3D Van der Waals surface area	Platt index
Ring atom count	Ring bond count	Ring count
Polar surface area	Randic index	Szeged index
Aliphatic atom count	Bond count	Aliphatic bond count
Rotatable bond count	Aliphatic ring count	Hetero ring count
Heteroaliphatic ring count	Heteroaromatic ring count	Chain atom count
Chian bond count	Smallest ring size	Largest ring size
Wiener index	Wiener polarity	Stereoisomer count
Double-bond stereoisomer count	Tautomer count	Tetrahedral steroisomer count
Markush enumerated structure count	logP at range of pH values	pKa
h-bond acceptor count	h-bond donor count	h-bond acceptor site count
h-bond donor site count	Molecular refractivity	

Table S2. AutoDock Vina scores for PCP site antagonist library compounds (standard deviation in parentheses).

	Actives	**Inactives**	**Decoys**
Mean Score	-6.7 (0.9)	-6.1 (1.2)	-6.2 (0.9)
Mean Top Pose Score	-7.6 (0.8)	-6.9 (1.3)	-7.0 (0.8)

Table S3. Proportion of largest unique substructures present in over 50% of known actives found in generated compounds.

		Actives		AI Generated	
Num Heteroatoms	Substructure	Hits	%	Hits	%
10	CCCCCCC=CC=C	359	50.1	68	0.00649
10	CCCCc1ccccc1	422	58.9	36	0.00343
9	CC(c1ccccc1)N	373	52.1	2	0.000191
9	CCCCC=CC=CC	432	60.3	387	0.0369
8	CNCC=CC=CC	429	59.9	170	0.0162
8	CNCCC=CC=C	369	51.5	193	0.0184
7	CC(=CC=C)CC	427	59.6	361	0.0344
7	CC=CC(=C)CC	359	50.1	409	0.0390
7	CC=CCC(N)C	380	53.1	274	0.0261
7	CCC(=CC)C=C	475	66.3	439	0.0419
7	CCCC(=C)C=C	368	51.4	359	0.0342
7	CCCC(=CC)C	392	54.8	748	0.0713
6	CCC(CC)C	409	57.1	1831	0.175
6	CCC(CC)N	406	56.7	750	0.0715
6	CCC(NC)C	390	54.5	1268	0.121
6	CCCC(C)C	367	51.3	2148	0.205
6	CCCC(N)C	405	56.6	1284	0.123
6	CCCCCN	415	58.0	2562	0.244
6	CCCCNC	397	55.5	2624	0.250
6	CCNCCC	399	55.7	2201	0.210

Table S4. Presence of largest unique substructures found in over 50% of actives in each generated finalist compound.

Substructure	a CCC1C2CCC31C2(N=S)CC(=N3)C	b COC1CCC(N=C1)CC1C=C1SC	c N#CCC=C1SC2CCCC1(C)COC2	e CNCCCSC1(C)CC1CC1CC1C	d CCSCCCC1=NCC=NCCC1C	e CNCCCSC1(C)CC1CC1CC1C	f CNCCCCCC1CCC2=CC=C1CC2	g CCCCC=C1C2CCC1CN(C2)CC	h CCCCCSCS12CCCC1CNC2CC	i CCCNCC1C2NCCC2CC1(CC)NCC	j CCCCC1CCC#S21CCCCN(C2C)CC	k CCCNCC1C2NCCC2CC1(CCC)NCC	l OCCNCC1C2SCCC2CC1(CCC)NCC	Hits
CCCCCCC=CC=C	0	0	0	0	0	0	1	0	0	0	0	0	0	1
CCCCc1ccccc1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CC(c1ccccc1)N	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CCCCC=CC=CC	0	0	0	0	0	0	1	0	0	0	0	0	0	1
CNCC=CC=CC	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CNCCC=CC=C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CC(=CC=C)CC	0	0	0	0	0	0	1	0	0	0	0	0	0	1
CC=CC(=C)CC	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CC=CCC(N)C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CCC(=CC)C=C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CCCC(=C)C=C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CCCC(=CC)C	0	0	0	0	0	0	1	0	0	0	0	0	0	1
CCC(CC)C	1	0	0	1	1	1	1	1	0	1	0	1	1	8
CCC(CC)N	1	1	0	0	0	0	0	0	0	1	0	1	1	5
CCC(NC)C	0	0	0	0	0	0	0	0	0	1	0	1	1	3
CCCC(C)C	1	1	1	1	1	1	1	1	0	1	0	1	1	10
CCCC(N)C	1	1	0	0	0	0	0	0	0	1	0	1	1	5
CCCCCN	1	1	0	0	1	0	1	1	1	1	0	1	1	9
CCCCNC	0	0	0	0	0	0	1	1	1	1	1	1	1	7
CCNCCC	0	0	0	0	0	0	0	1	1	1	1	1	1	6

Table S5. Similarity of finalist generated compounds to training set actives.

Generated compound	L1 nearest therapeutic active	L1 distance	Tanimoto nearest active	Tanimoto distance
CNCCCCCC1CCC2=CC=C1CC2	NC12CC3CC(C2)CC(C1)C3	49.72	CC1(C)CC(CN)CC(C)(C)C1	0.51
CCCCC=C=C1C2CCC1CN(C2)CC	CC12CC3CC(C1)(C)CC(C2)(C3)N	35.66	C[C@H]1CCC[C@](C)(N)C1	0.52
CCCNCC1C2NCCC2CC1(CCC)NCC	COc1ccc2c(c1)[C@@]13CCCC[C@@H]3[C@H](C2)N(CC1)C	45.27	C1CCC(C2(N3CCCCC3)CCCCC2)CC1	0.59
CCCNCC1C2NCCC2CC1(CC)NCC	COc1ccc2c(c1)[C@@]13CCCC[C@@H]3[C@H](C2)N(CC1)C	45.10	C1CCC(C2(N3CCCCC3)CCCCC2)CC1	0.59
CCC1C2CCC31C2(N=S)CC(=N3)C	CC12CC3CC(C1)(C)CC(C2)(C3)N	32.57	C[C@H]1CCC[C@](C)(N)C1	0.51
CCCCCSCS12CCCC1CNC2CC	COc1ccc2c(c1)[C@@]13CCCC[C@@H]3[C@H](C2)N(CC1)C	43.21	C[C@H]1C[C@@H](C)C[C@@](C)(N)C1	0.51
OCCNCC1C2SCCC2CC1(CCC)NCC	COc1ccc2c(c1)[C@@]13CCCC[C@@H]3[C@H](C2)N(CC1)C	47.28	O=C1CCCN1CC#CCN1CCCC1	0.51
COC1CCC(N=C1)CC1C=C1SC	CC12CC3CC(C1)(C)CC(C2)(C3)N	29.82	C=CC1(N)CC(C)(C)CC(C)(C)C1	0.51
CCSCCCC1=NCC=NCCC1C	CC12CC3CC(C1)(C)CC(C2)(C3)N	34.99	C[C@H]1C[C@@H](C)C[C@@](C)(N)C1	0.53
CNCCCSC1(C)CC1CC1CC1C	CC12CC3CC(C1)(C)CC(C2)(C3)N	40.17	C[C@H]1C[C@@H](C)C[C@@](C)(N)C1	0.49
N#CCC=C1SC2CCCC1(C)COC2	CC12CC3CC(C1)(C)CC(C2)(C3)N	37.62	O=C1CCCN1CC#CCN1CCCC1	0.50
CCCCC1CCC#S21CCCCN(C2C)CC	COc1ccc2c(c1)[C@@]13CCCC[C@@H]3[C@H](C2)N(CC1)C	43.73	C[C@H]1C[C@@H](C)C[C@@](C)(N)C1	0.51

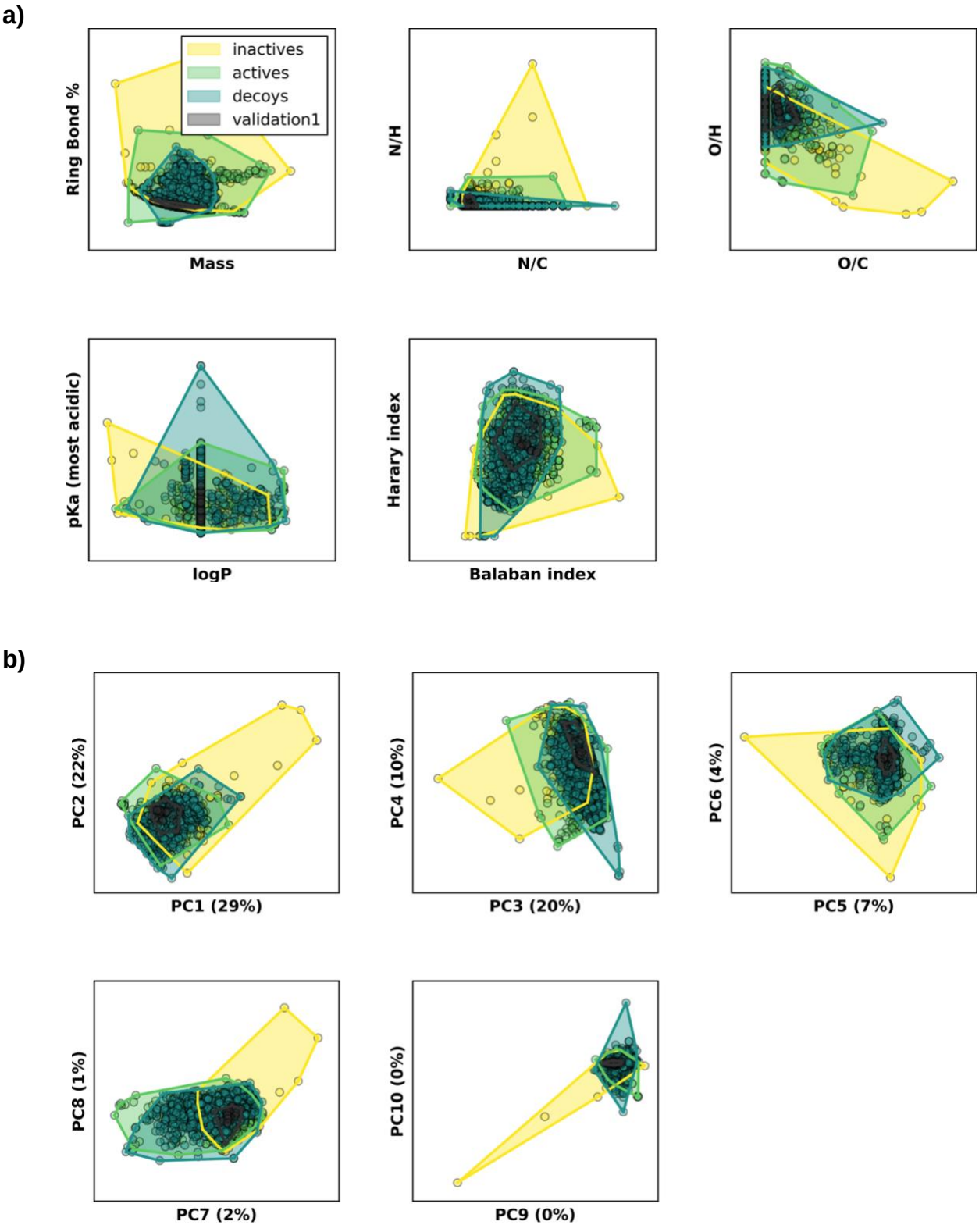
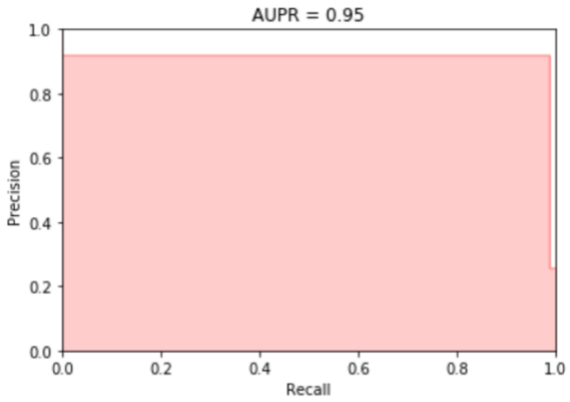


Figure S1. Location of NMDAR PCP site library antagonist actives, inactives, decoys, and validation set (i.e. the set used as the basis for decoy generation with DUD-E) in DarkChem's (a) property space and (b) PCA space

a)

	precision	recall	f1-score	support
0	1.00	0.97	0.98	517
1	0.92	0.99	0.95	180
micro avg	0.97	0.97	0.97	697
macro avg	0.96	0.98	0.97	697
weighted avg	0.98	0.97	0.97	697

b)



c)

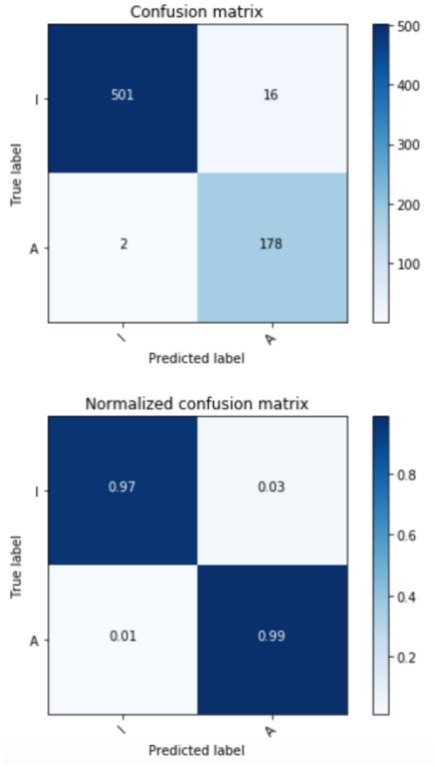


Figure S2. SVM evaluation: (a) precision, recall, and f1; (b) AUPR; (c) confusion matrices, where I = inactive and A = active

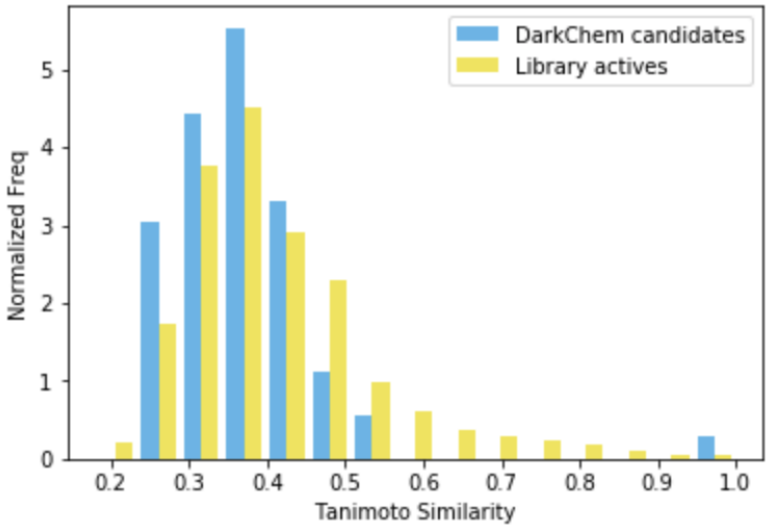


Figure S3. Tanimoto score distribution of finalist generated candidates and PCP site library actives