Electronic Supplementary Material (ESI) for MedChemComm.
This journal is © The Royal Society of Chemistry 2015

Supplementary information

Property-based characterization of kinase-like ligand space for library design and virtual screening

Dávid Bajusz, György G. Ferenczy, György M. Keserű*

Research Centre for Natural Sciences, Hungarian Academy of Sciences, H-1117 Budapest XI., Magyar tudósok körútja 2, Hungary

*Corresponding author: György M. Keserű (keseru.gyorgy@ttk.mta.hu)

Table of contents:

Figure S1 Categorized histogram, box plot and desirability function of TPSA

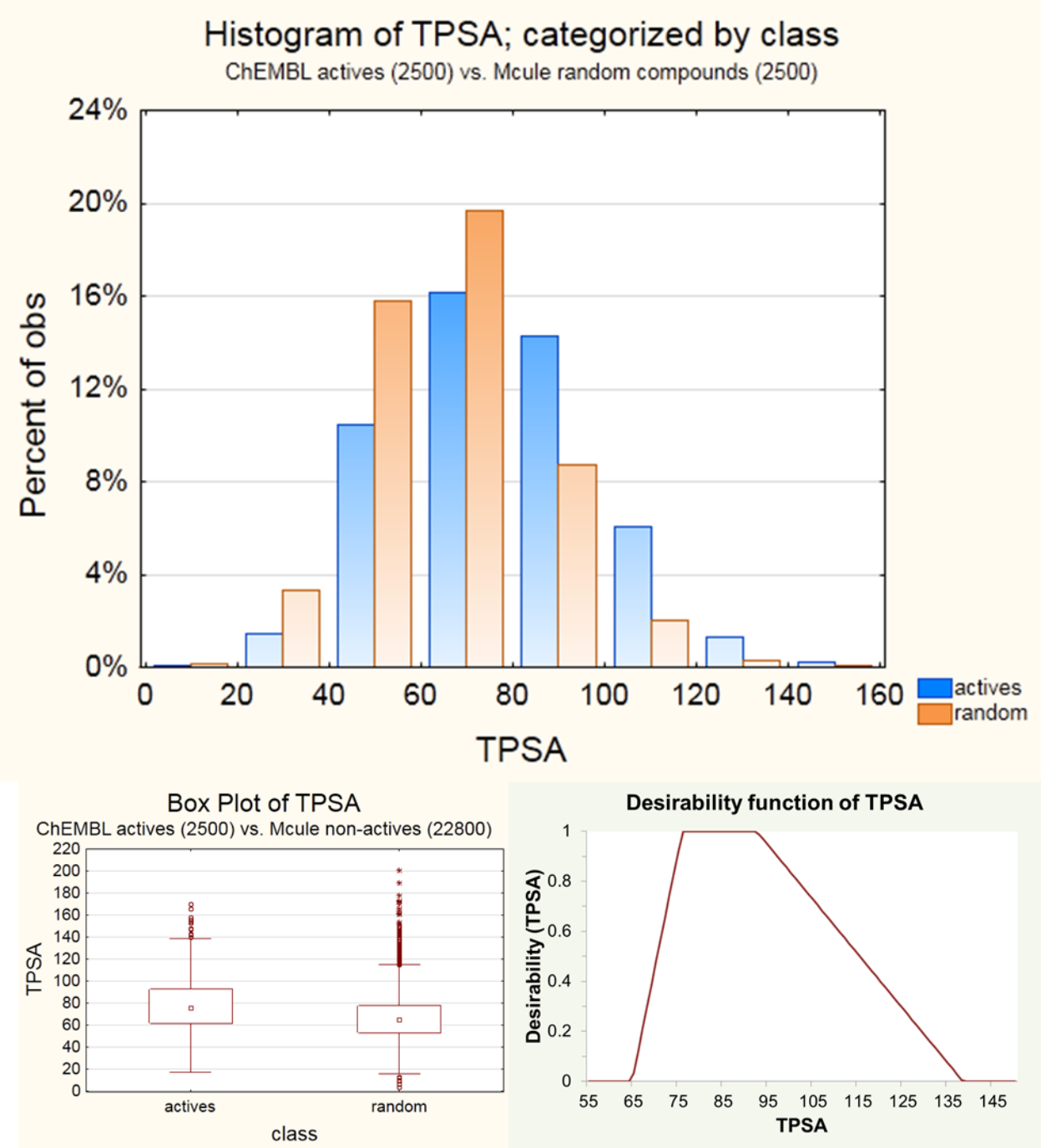

Figure S2 Categorized histogram, box plot and desirability function of the rotatable bond count

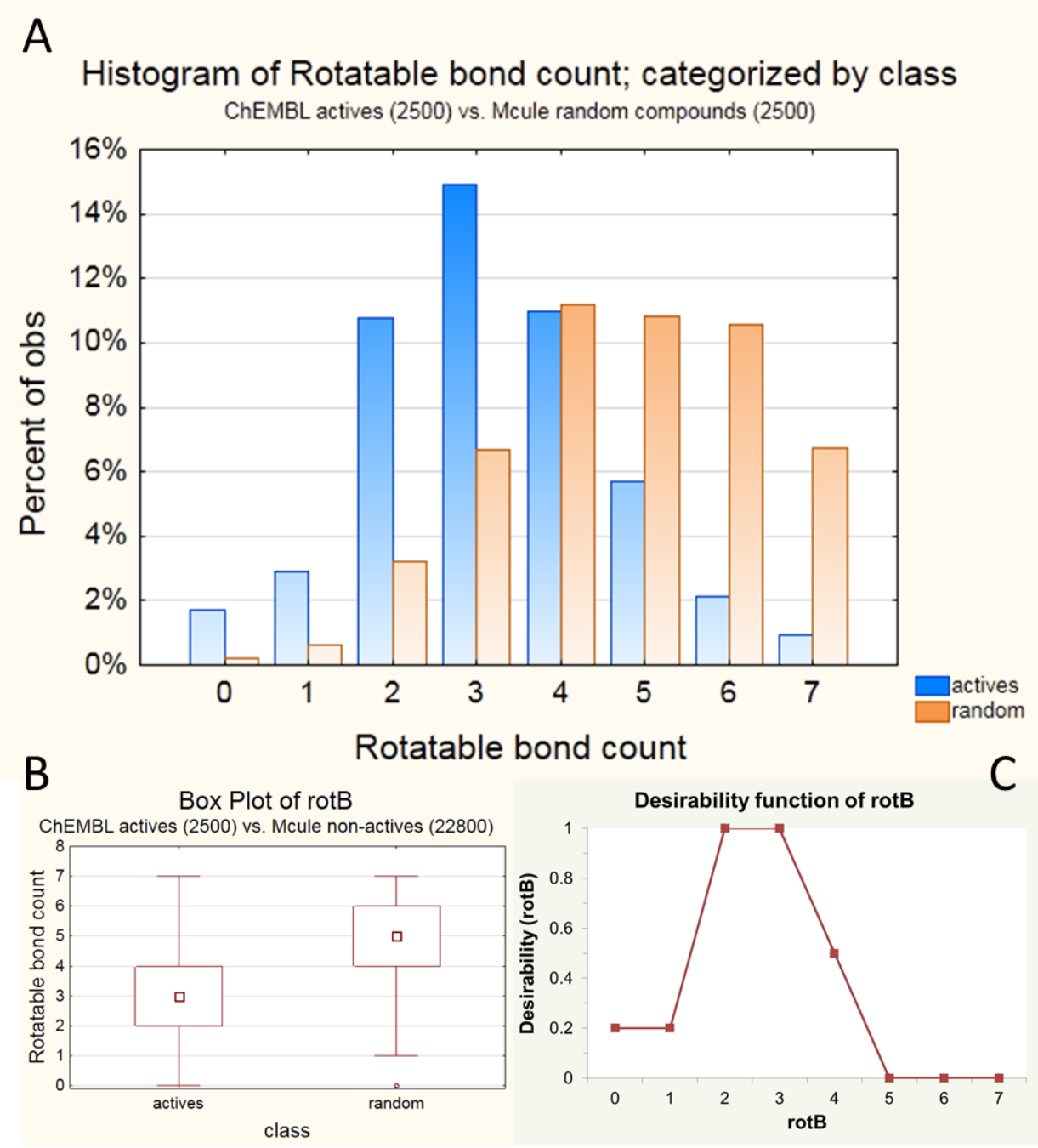

Figure S3 Categorized histogram, box plot and desirability function of the aromatic ring count

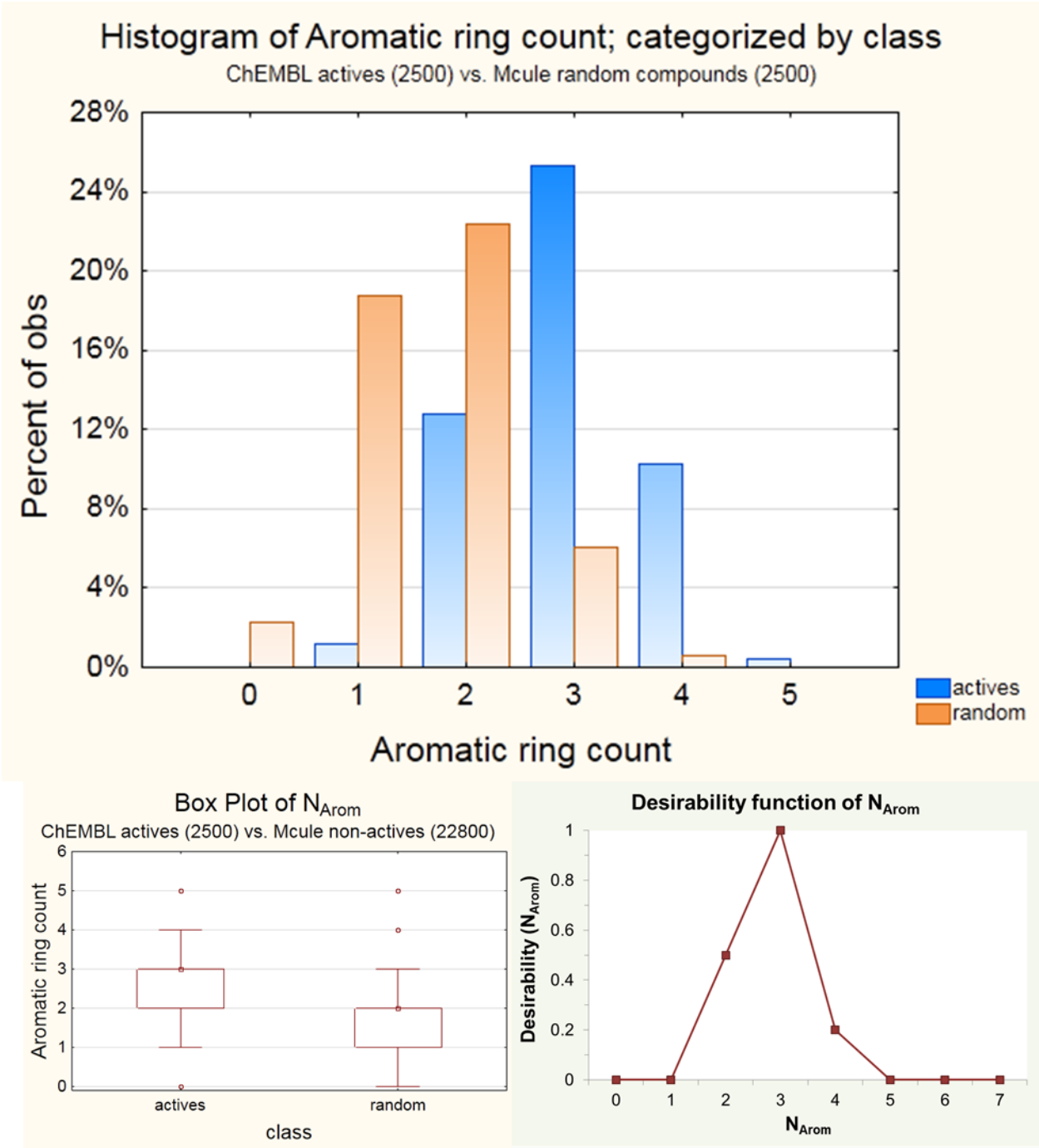

Figure S4 Categorized histogram, box plot and desirability function of the number of nitrogens

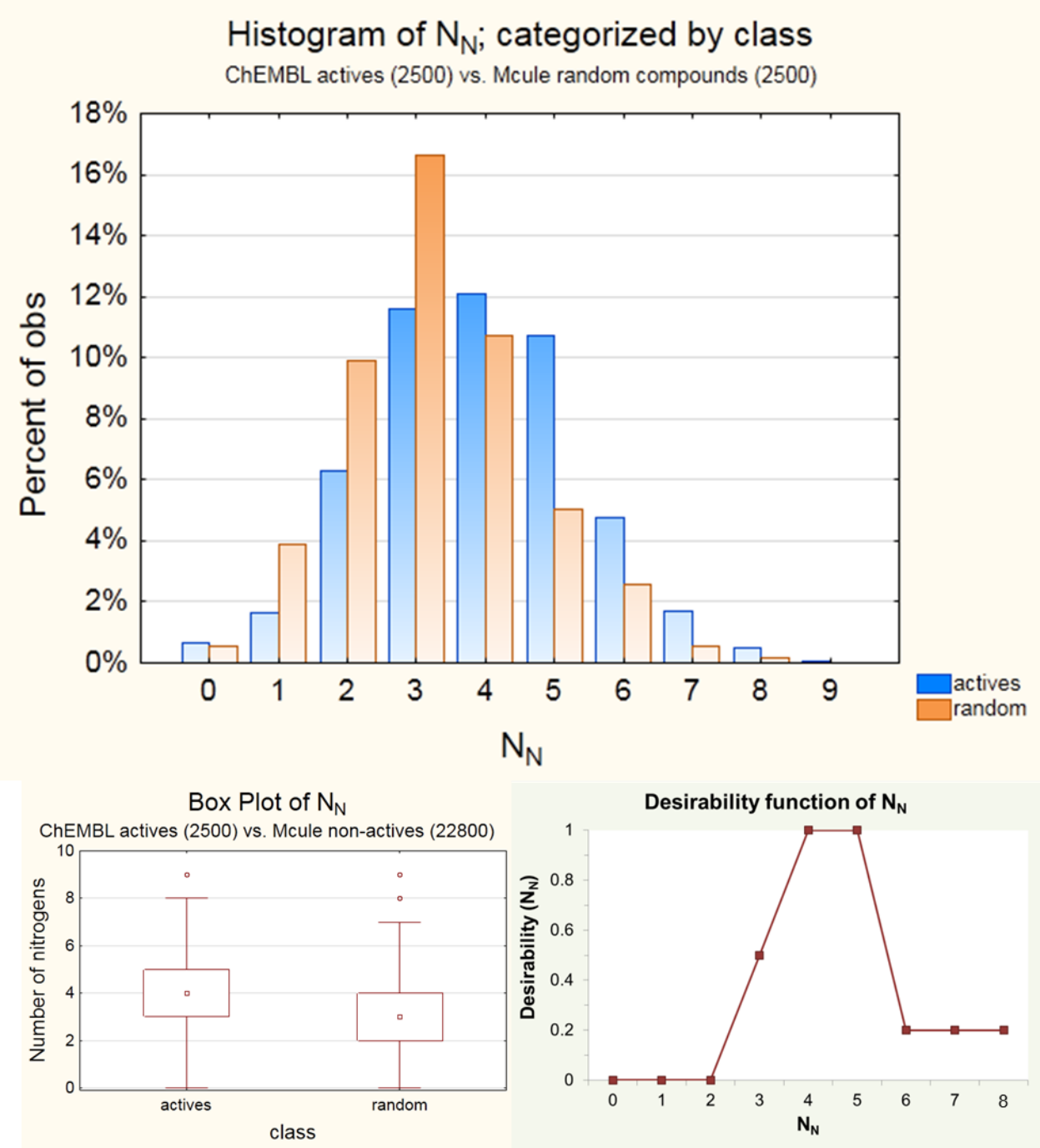

Figure S5 Categorized histogram, box plot and desirability function of the number of oxygens

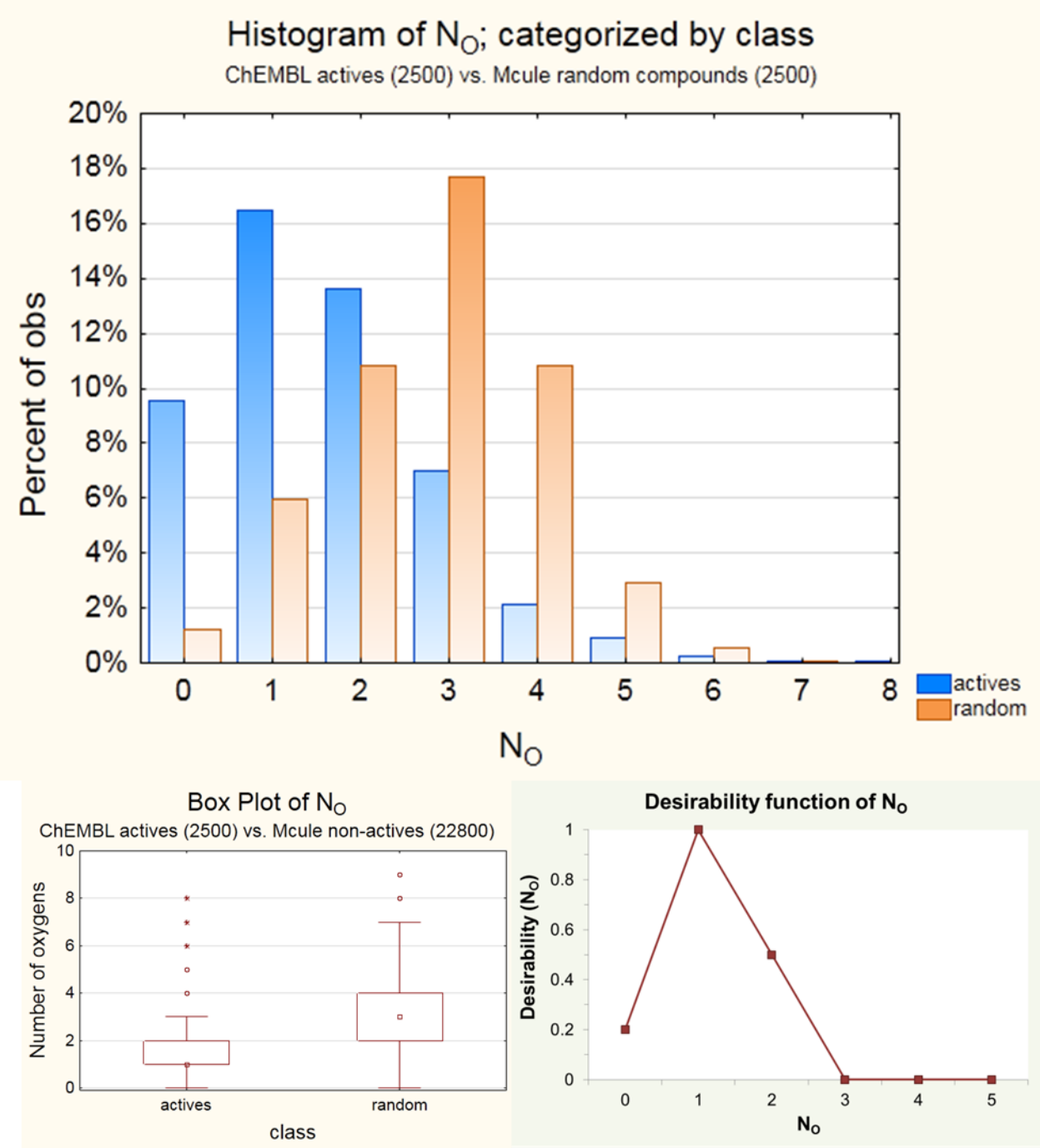

Figure S6 Categorized histogram, box plot and desirability function of the number of hydrogen bond donors

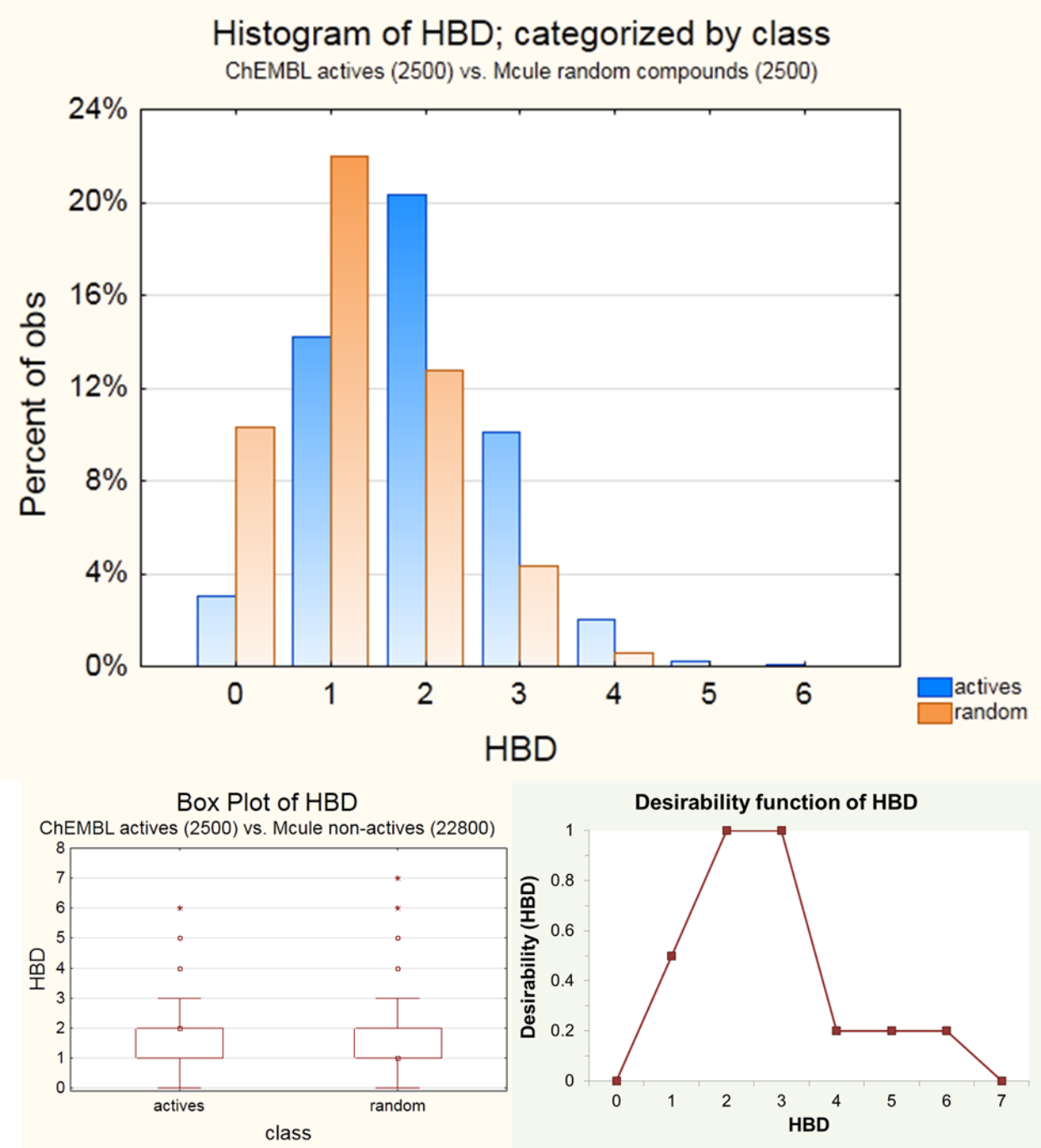

Figure S7 Categorized histogram of the KiDS distribution of the Training set

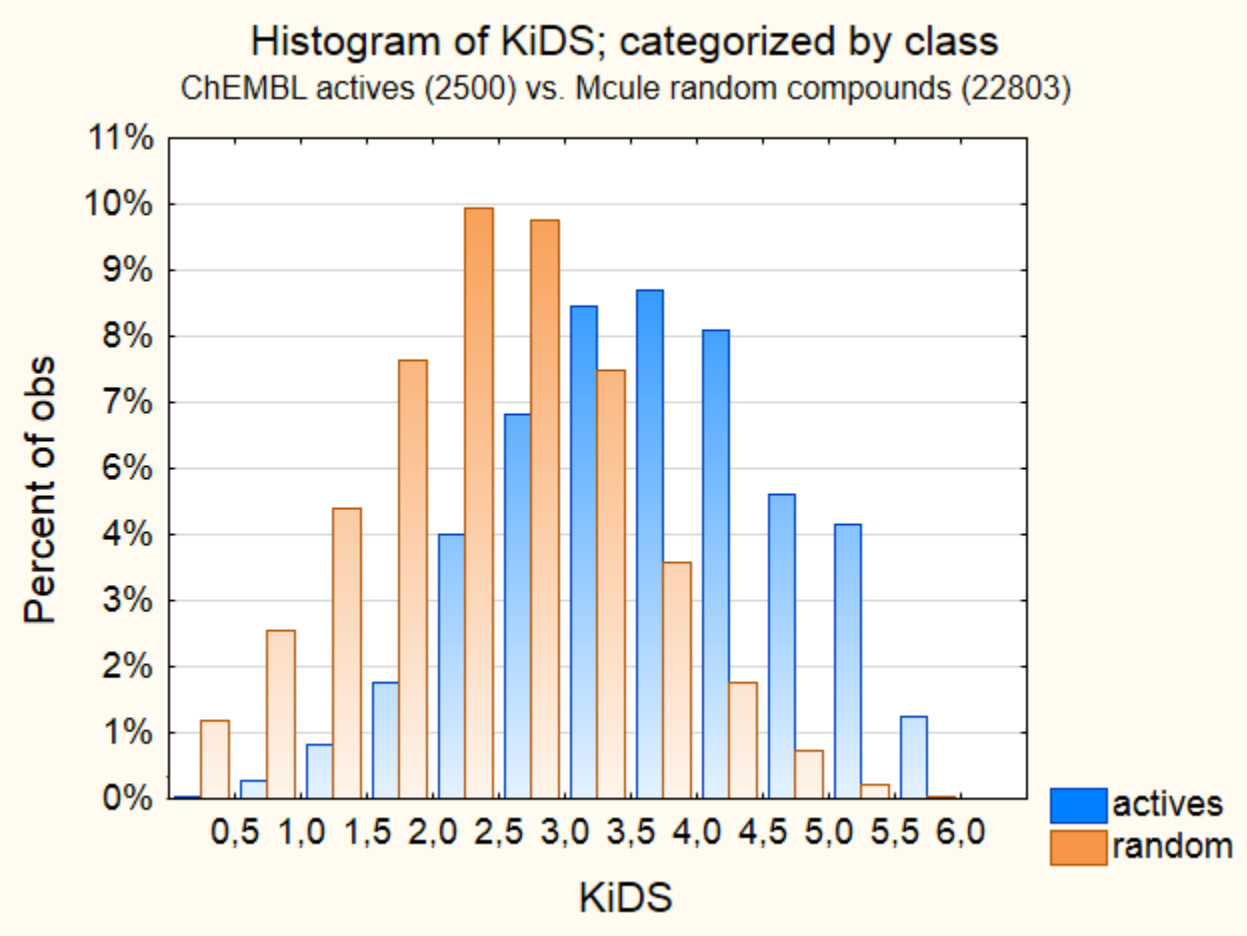

Figure S8 Categorized histogram of the KiDS distribution of Test set 1

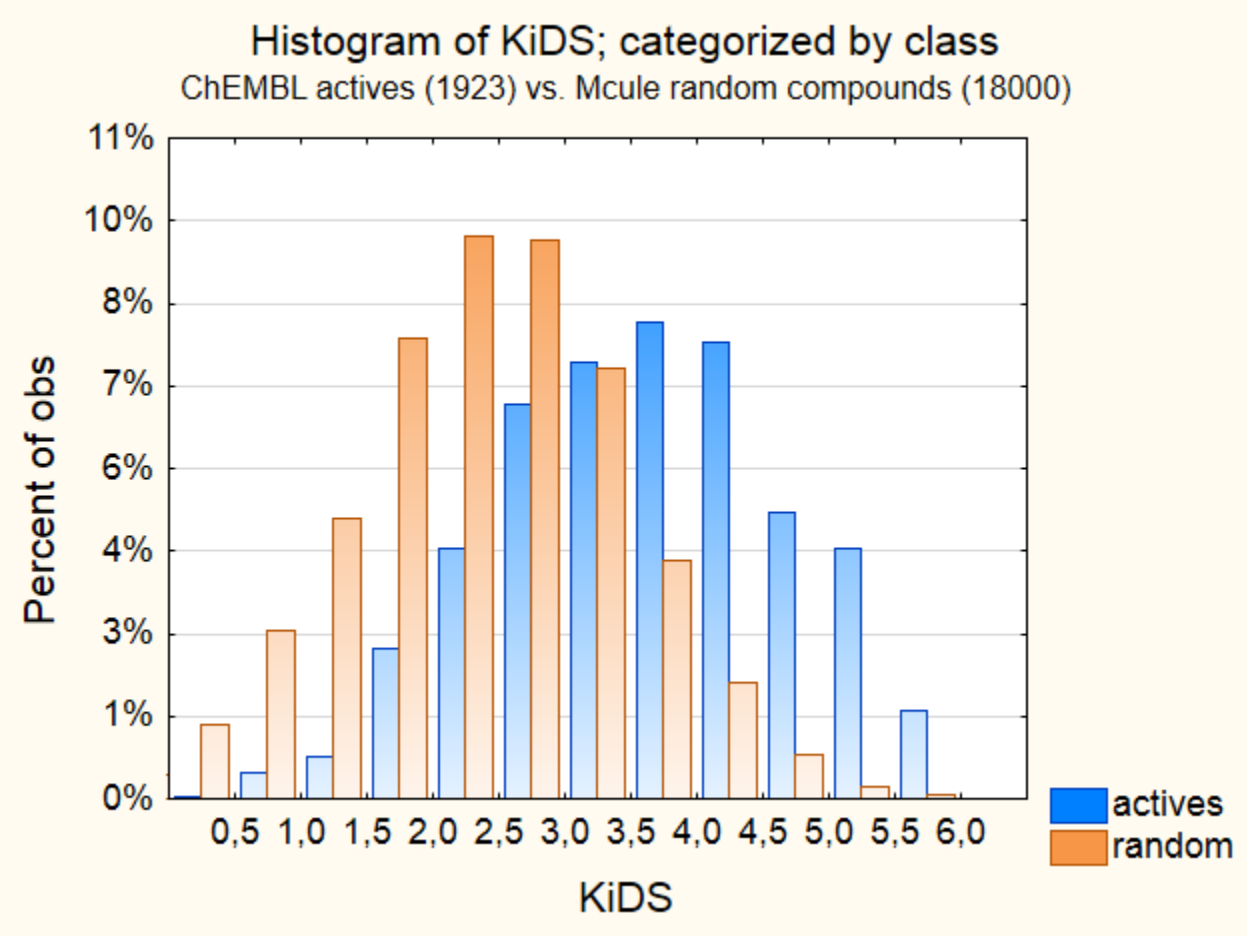

Figure S9 Categorized histogram of the KiDS distribution of Test set 2

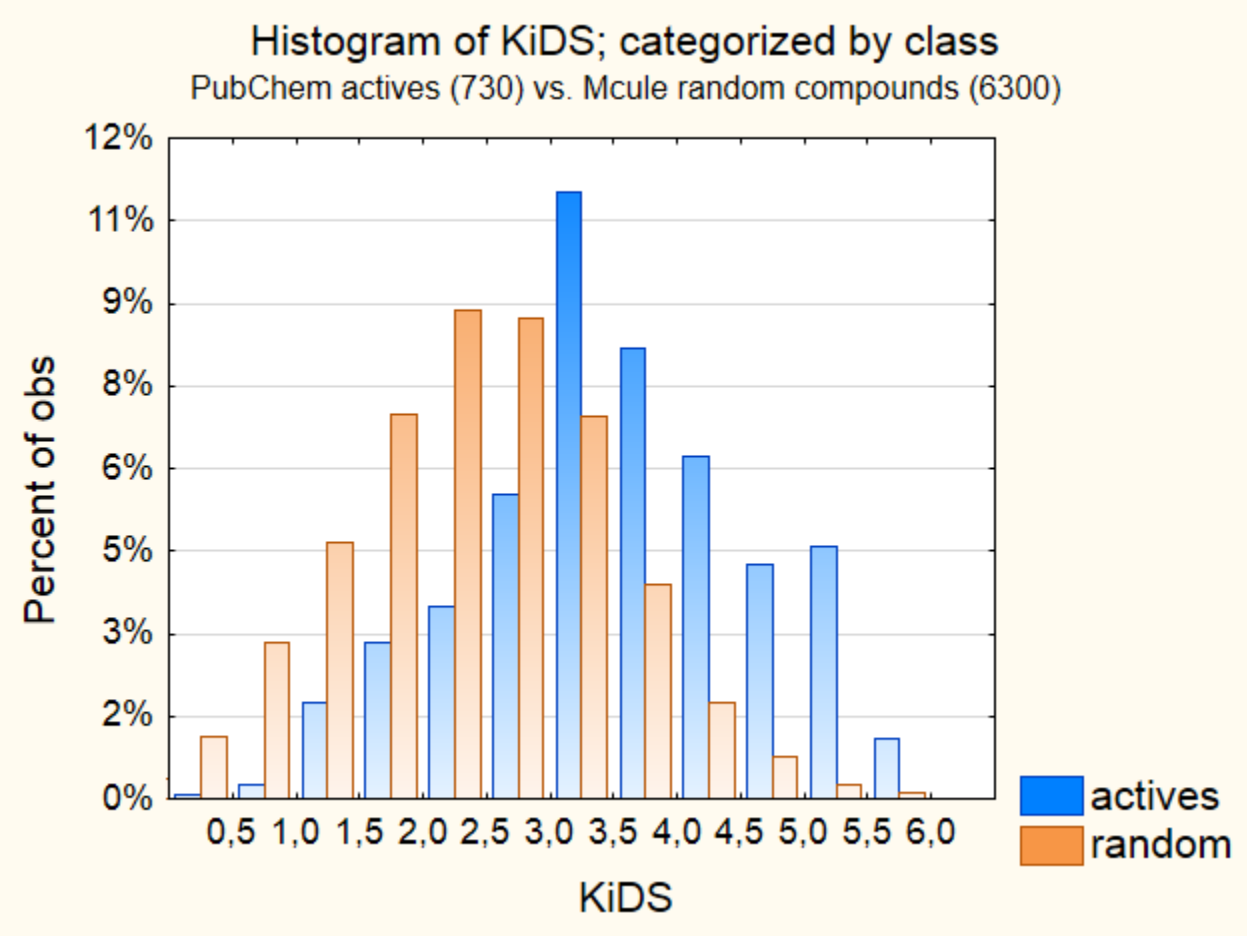

Table S1 Definitions of the desirability functions

Definitions are given in pairs of columns, where the first column (TPSA, rotB, *etc.*) contains the possible values of the descriptors, while the second column (d(TPSA), d(rotB), *etc.*) contains the score values that are assigned to the given property value. A score of 0 is assigned to any property value that is not present in the table. The overall desirability score (KiDS) is the sum of d(TPSA), d(rotB) … d(Arom).

TPSA	d(TPSA)	rotB	d(rotB)	N_N	d(N_N)	N_O	d(N_O)	HBD	d(HBD)	Arom	d(Arom)
64.63-75.85	$\frac{TPSA - 64.63}{11.23}$	0	0.2	0	0	0	0.2	0	0	0	0
75.86-92.40	1	1	0.2	1	0	1	1	1	0.5	1	0
92.41-138.3	$\frac{-(TPSA - 138.8)}{45.89}$	2	1	2	0	2	0.5	2	1	2	0.5
		3	1	3	0.5	3	0	3	1	3	1
		4	0.5	4	1	4	0	4	0.2	4	0.2
		5	0	5	1	5	0	5	0.2	5	0
		6	0	6	0.2			6	0.2	6	0
		7	0	7	0.2			7	0	7	0
				8	0.2						

Table S2 Performance evaluation of the Kinase Desirability Score: conventional enrichment factors (see definition in the Methods/Evaluation section)

Dataset	$EF_{0.5\%}$		$EF_{1\%}$		$EF_{2\%}$		$EF_{5\%}$	
	KiDS	KLS[a]	KiDS	KLS	KiDS	KLS	KiDS	KLS
Training	8.58 (2.1E-3)[b]	1.86 (3.9E-3)	7.30 (2.2E-3)	1.69 (2.8E-3)	7.20 (1.5E-3)	1.65 (2.0E-3)	5.22 (1.1E-3)	1.40 (1.1E-3)
Test 1	8.42 (3.2E-3)	1.79 (5.3E-3)	7.54 (2.8E-3)	1.31 (3.3E-3)	7.04 (2.1E-3)	1.41 (2.4E-3)	5.19 (1.5E-3)	1.32 (1.4E-3)
Test 2	7.49 (9.4E-3)	2.50 (1.6E-2)	6.80 (8.3E-3)	2.64 (1.1E-2)	6.93 (4.9E-3)	2.36 (7.4E-3)	4.76 (4.1E-3)	1.66 (4.2E-3)

[a] Performance parameters obtained for the same datasets with the KLS score of Singh *et al.* are provided as a reference.[1]

[b] 1.96σ values (corresponding to 95% confidence intervals) are given in parentheses.[2]

Table S3 External validation of KiDS with random molecules from ZINC: AUC values and early enrichment factors

Dataset	$EF_{0.5\%}$		$EF_{1\%}$		$EF_{2\%}$		$EF_{5\%}$		AUC	
	KiDS	KLS[a]	KiDS	KLS	KiDS	KLS	KiDS	KLS	KiDS	KLS
Training Z	15.8 (1.7E-2)[b]	2.80 (6.8E-3)	14.3 (1.1E-2)	2.32 (4.3E-3)	8.68 (5.5E-3)	2.02 (2.8E-3)	5.48 (2.5E-3)	1.88 (1.6E-3)	0.786 (9.1E-3)	0.627 (0.011)
Test 1Z	15.0 (1.9E-2)	2.39 (7.0E-3)	14.0 (1.2E-2)	1.87 (4.3E-3)	8.58 (6.2E-3)	1.87 (3.0E-3)	5.34 (2.8E-3)	1.82 (1.8E-3)	0.778 (0.011)	0.624 (0.013)
Test 2Z	18.4 (3.1E-2)	4.93 (1.6E-2)	14.8 (1.9E-2)	4.25 (1.1E-2)	8.22 (9.7E-3)	3.01 (6.2E-3)	4.85 (4.4E-3)	2.25 (3.3E-3)	0.759 (0.017)	0.616 (0.020)

[a] Performance parameters obtained for the same datasets with the KLS score of Singh *et al.* are provided as a reference.[17]

[b] 1.96σ values (corresponding to 95% confidence intervals) are given in parentheses.[33]

Table S4 External validation of KiDS with random molecules from ZINC: conventional enrichment factors (see definition in the Methods/Evaluation section)

Dataset	$EF_{0.5\%}$		$EF_{1\%}$		$EF_{2\%}$		$EF_{5\%}$	
	KiDS	KLS[a]	KiDS	KLS	KiDS	KLS	KiDS	KLS
Training Z	7.24 (2.7E-3)[b]	2.49 (4.7E-3)	6.05 (2.5E-3)	2.16 (3.1E-3)	5.77 (1.8E-3)	1.92 (2.1E-3)	4.64 (1.2E-3)	1.68 (1.3E-3)
Test 1Z	9.17 (3.2E-3)	2.11 (5.4E-3)	6.88 (3.6E-3)	1.78 (3.5E-3)	6.85 (2.3E-3)	1.84 (2.5E-3)	5.20 (1.5E-3)	1.71 (1.5E-3)
Test 2Z	13.6 (1.5E-2)	4.44 (1.3E-2)	10.5 (1.0E-2)	3.86 (8.7E-3)	7.52 (6.9E-3)	3.04 (5.5E-3)	4.74 (3.7E-3)	2.18 (3.0E-3)

[a] Performance parameters obtained for the same datasets with the KLS score of Singh *et al.* are provided as a reference.[1]

[b] 1.96σ values (corresponding to 95% confidence intervals) are given in parentheses.[2]

References:

1 N. Singh, H. Sun, S. Chaudhury, M. D. M. AbdulHameed, A. Wallqvist and G. Tawa, *J. Cheminform.*, 2012, **4**, 4.

2 A. Nicholls, *J. Comput. Aided. Mol. Des.*, 2014, **28**, 887–918.