Supporting Information

Using Matched Molecular Series as a Predictive Tool to Optimize Biological Activity

Noel M. O'Boyle, Jonas Boström, Roger A. Sayle, Adrian Gill.

Table of Contents

Frequency of common R groups in series of different lengths in the dataset

Figure S1 shows the normalized frequency for the most common R groups in series of particular lengths. The steeper curves for shorter series indicates that shorter matched series are more likely to contain common R groups. In fact, the 5 most common R groups in series of length 2 cover 53% of the R groups, while this decreases to 30% for N=3, 23% for N=4, 18% for N=5 and 15% for N=6.

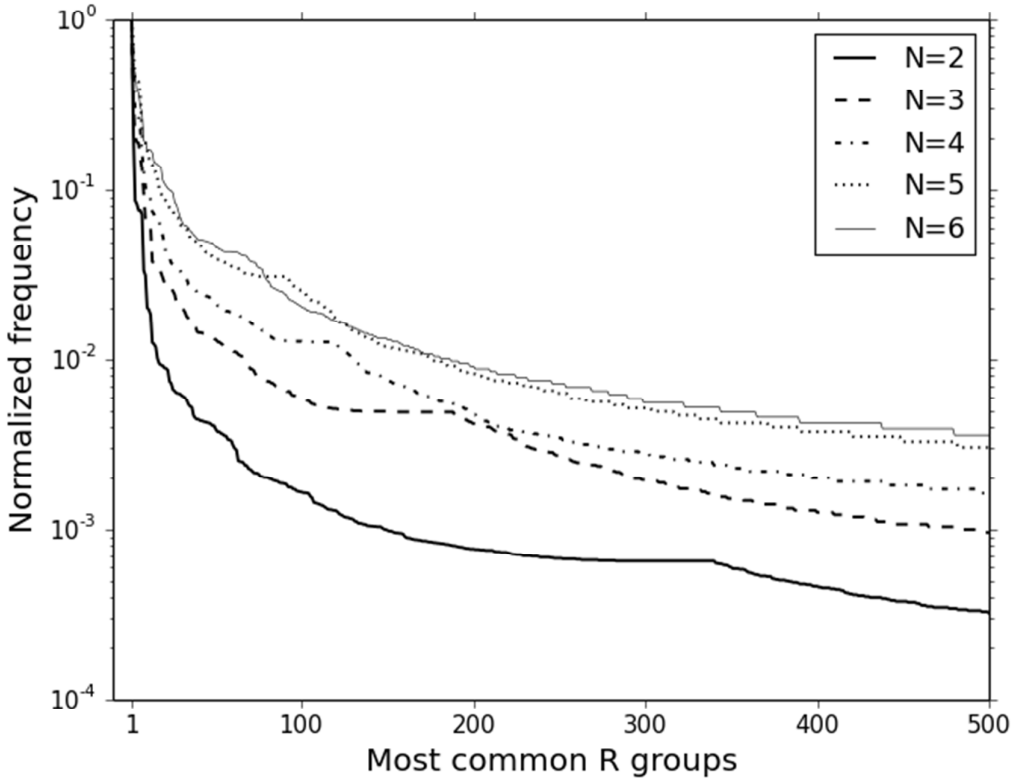


Figure S1 – Comparison of frequencies of the most common R groups in matched series of different lengths (N). To highlight differences between the frequency curves in this log plot, all frequency values for a particular series length were divided by the frequency of the most common R group for that length.

Retrospective test using a focussed subset

The training data in the main retrospective test indiscriminately combined pIC_{50} data from a wide range of targets. The fact that it still worked well indicates either that SAR data is in general transferable between different targets (due for example to similar protein environments occurring in different binding sites, or due to the same physicochemical driving forces at play), or that a sufficiently long ordered matched series is characteristic of particular targets and so in effect we are selecting those matched pairs relevant to a particular target. This leads to the question of whether a dataset targeted at a particular protein class would provide better predictions for that class. Since ChEMBL provide a list of kinase and GPCR related targets (in association with the Kinase SARfari and GPCR SARfari web portals), we could investigate this question.

The GPCR data is only available for targets up until 2011 and even then there is little 2011 data. As a result the test set was generated from the GPCR data from 2010 and 2011 (6213 matched series, reduced to 321 after removing series shorter than length 5 and duplicates), while the earlier GPCR

data was used for training (65339 matched series). To compare with results based on all of the data, predictions were made for the same test set but using training data from all ChEMBLdb assays prior to 2010 (332713 matched series). As before, 100 repetitions were used when generating the test set.

The results are shown in Figure S2. For those series where predictions were made, improved performance is seen for the GPCR subset for the series of lengths 2 and 3. This may also be the case for the series of lengths 4 and 5 but the dataset size prevents any conclusions to be drawn as the variance is too high. However, as the GPCR dataset is much smaller than the combined ChEMBLdb data, the number of series for which no predictions can be made is also larger. This suggests that a useful approach would be to use the focussed subset if the target is a GPCR and then fall back to the larger dataset if no prediction is possible.

A similar analysis can be done for the kinase subset. Kinase data is available for targets up until 2011. The test set is generated from the 2011 kinase data (2697 matched series, reduced to 461 after removing series shorter than length 5 and duplicates), while the earlier data was using for training (49364 matched series). To compare with results based on all of the data, predictions were made for the same test set but using training data from all ChEMBLdb assays prior to 2011 (377219 matched series). The results are shown in Figure S3. The results are similar to those for the GPCR data except that no improved performance is seen at N=2.

It is worth comparing these results with those of Mills *et al.*[1] who report that a matched series of length 3 ("local SAR" in the context of the paper) was a better guide to prediction than matched pair data restricted to the target of interest.

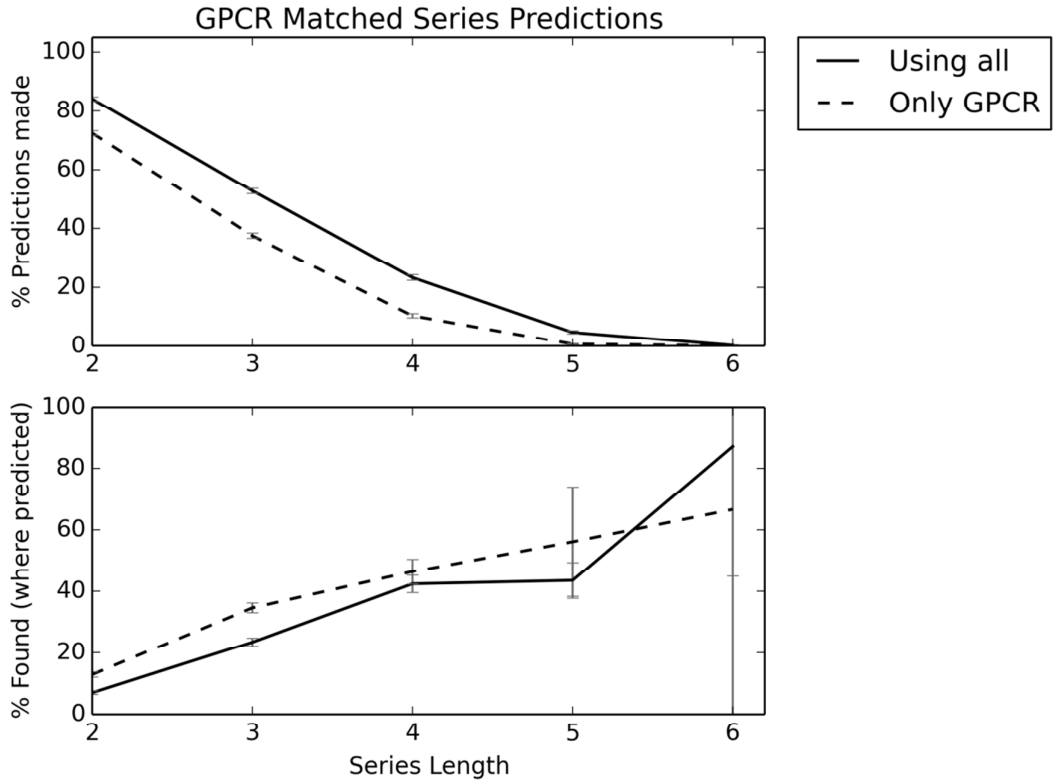


Figure S2 – Retrospective test results for the ChEMBL GPCR data.

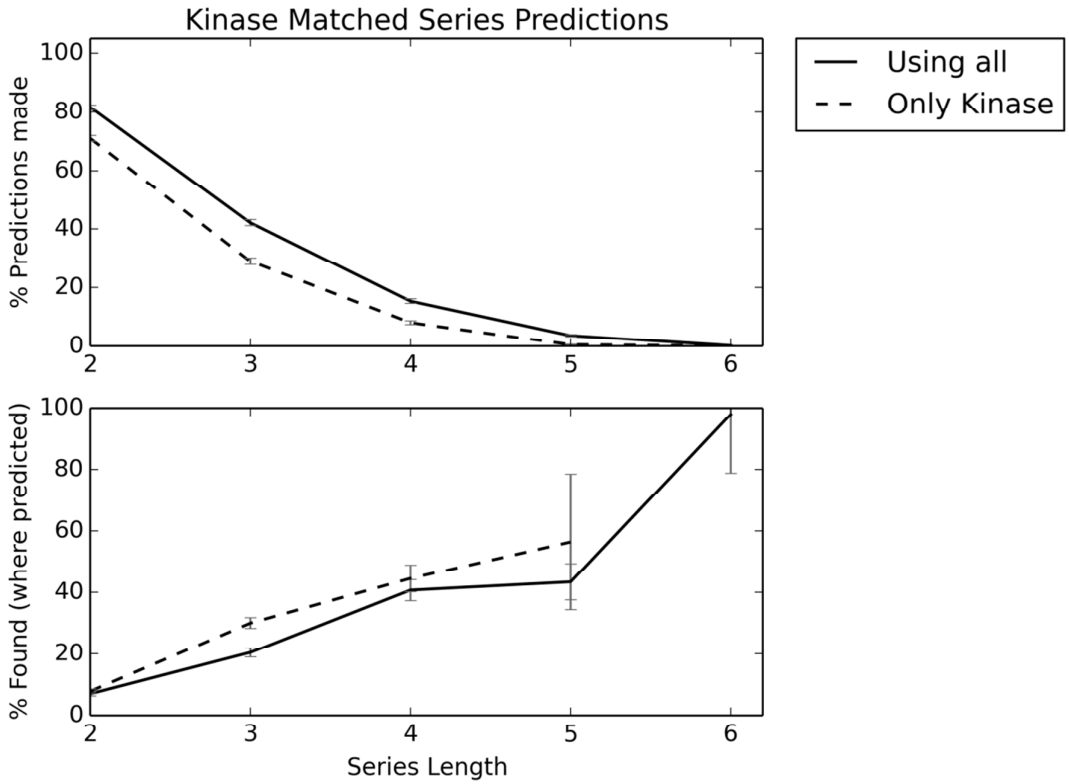


Figure S3 – Retrospective test results for the ChEMBL kinase data.

Enrichment values for predictions related to the Topliss Tree

Each substituent predicted by Matsy has an associated enrichment value for the ordered matched series consisting of itself in combination with the query. For example, given [4-Cl>H] Matsy predicts 3,4-diCl. The resulting ordered matched series [3,4-diCl>4-Cl>H] has an enrichment of 2.31 compared to other orderings of those three substituents. This enrichment has a p-value of 0.000. Table S1 lists enrichment values and their associated p-values for all of the ordered series discussed in the text as well as those shown in Figure 2.

Table S1 – Enrichments and their p-values for ordered series related to the Topliss Tree comparison

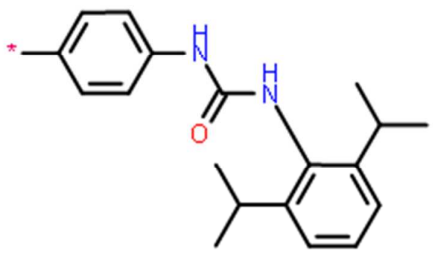
Series	Enrichment	p-value
4-Cl>H	1.12	0.000
3,4-diCl>4-Cl>H	2.31	0.000
2-naphthyl>3,4-diCl>4-Cl>H	4.14	0.000
4-NO_2>3,4-diCl>4-Cl>H	2.69	0.003
3-CF_3-4-Cl>3,4-diCl>4-Cl>H	5.05	0.007
4-Cl>3,4-diCl>H	1.09	0.415
4-Br>4-Cl>3,4-diCl>H	2.84	0.000

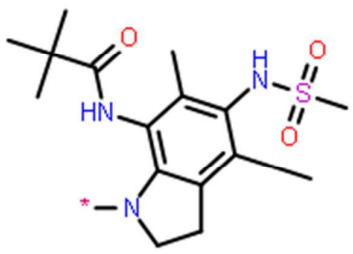
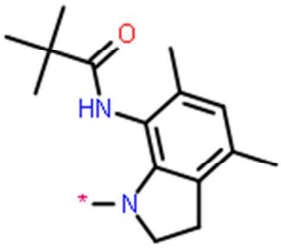
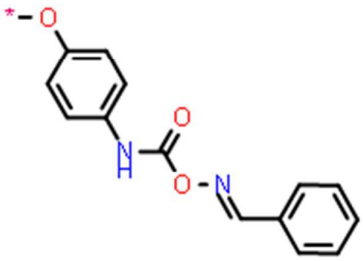
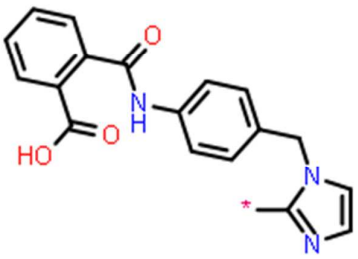
4-NO_2>4-Cl>3,4-diCl>H	2.20	0.021
4-OMe>4-Cl>3,4-diCl>H	1.55	0.078
2,4-diCl>4-Cl>3,4-diCl>H	1.20	0.613
4-Cl>H>3,4-diCl	0.69	0.003
3-Cl>4-Cl>H>3,4-diCl	0.81	0.728
4-OMe>4-Cl>H>3,4-diCl	0.78	0.631
H>4-Cl	0.88	0.000
4-OMe>H>4-Cl	0.78	0.000
4-OH>4-OMe>H>4-Cl	2.73	0.000
H>4-Cl>4-OMe	1.03	0.610
2-F>H>4-Cl>4-OMe	2.44	0.000
cyclohexyl>H>4-Cl>4-OMe	2.49	0.002
4-OH>H>4-Cl>4-OMe	2.18	0.004
3-Cl>H>4-Cl>4-OMe	1.70	0.003
4-OH>H>4-Cl	1.35	0.013
2-Cl>4-OH>H>4-Cl	2.73	0.006
3-Cl>4-OH>H>4-Cl	1.57	0.285
3-OH>4-OH>H>4-Cl	2.40	0.026
H>4-OH>4-Cl	0.74	0.070
4-OMe>H>4-OH>4-Cl	0.68	0.568
H>4-Cl>4-OH	0.81	0.188
4-F>H>4-Cl>4-OH	1.35	0.395

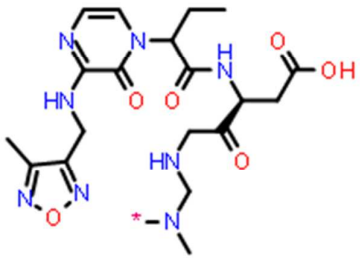
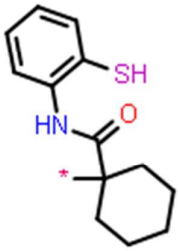
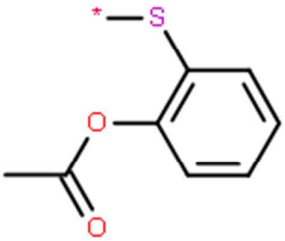
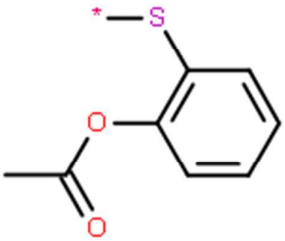
Scaffolds and targets for Matsy predictions

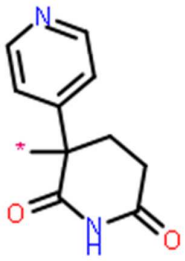
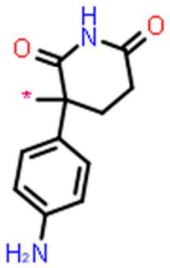
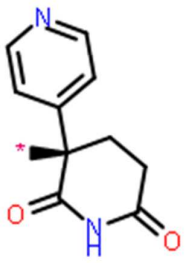
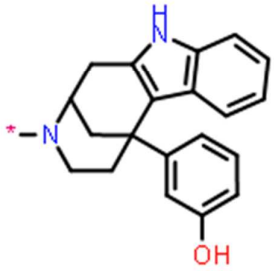
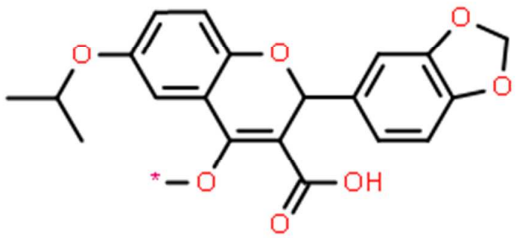
Table 3 summarizes the Matsy predictions for [CCC>CC>C]. The best prediction was n-hexyl, which improved the binding in 40 out tof 53 cases. Here we provide additional information on the 40 targets and scaffolds on which the prediction of hexyl was based.

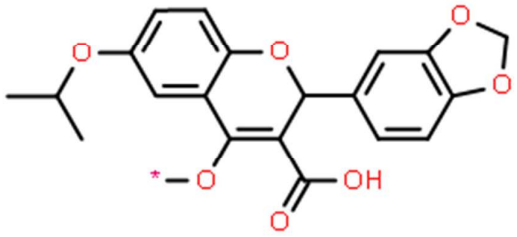
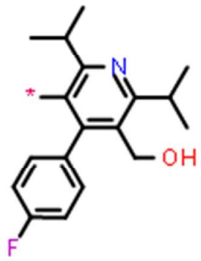
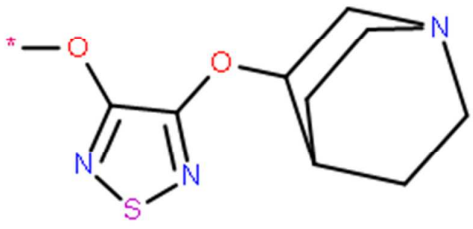
Table S2 – Scaffolds and targets for the n-hexyl prediction in Table 3

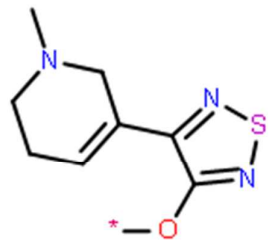
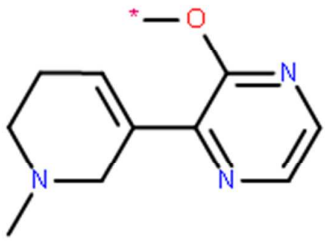
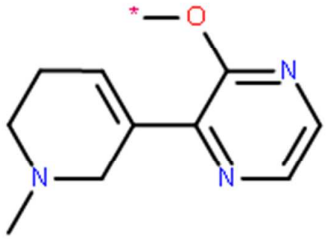
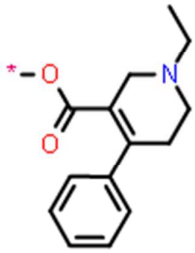
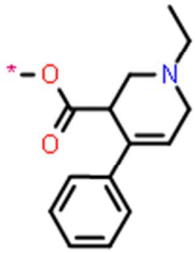
Scaffold	ChEMBL assay	Target
	CHEMBL1040228	Acyl coenzyme A:cholesterol acyltransferase 1

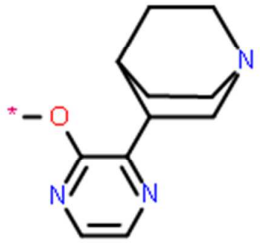
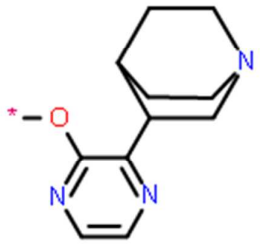
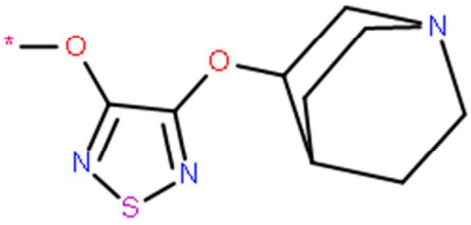
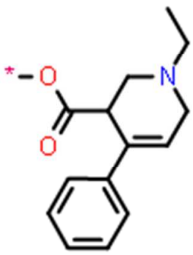
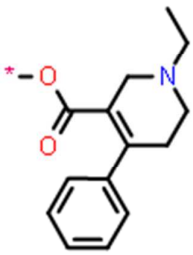
	CHEMBL1054937	Acyl-CoA:cholesterol acyltransferase
	CHEMBL1054937	Acyl-CoA:cholesterol acyltransferase
	CHEMBL1068137	Anandamide amidohydrolase
	CHEMBL2209179	Angiotensin II receptor
	CHEMBL645260	Angiotensin II receptor (AT-1) type-1

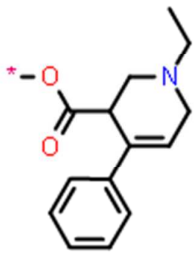
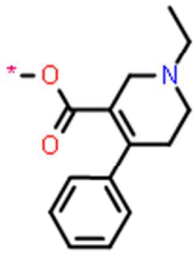
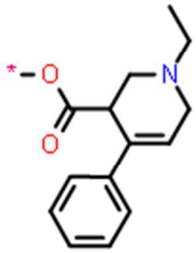
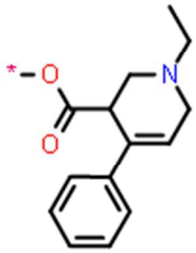
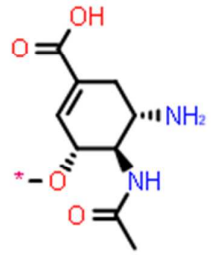
	CHEMBL646537	C-C chemokine receptor type 5
	CHEMBL649303	Caspase-3
	CHEMBL661605	Cholesteryl ester transfer protein
	CHEMBL663079	Cyclooxygenase-1
	CHEMBL665927	Cyclooxygenase-2

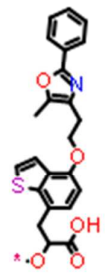
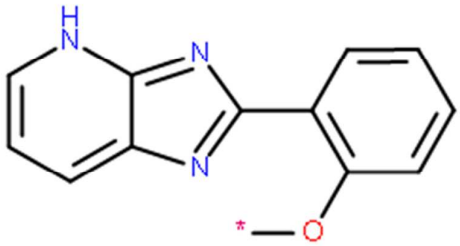
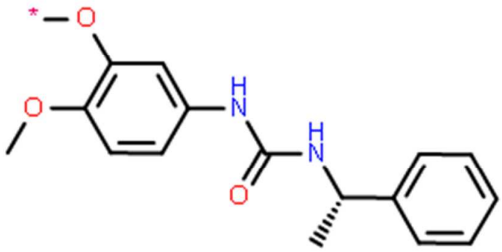
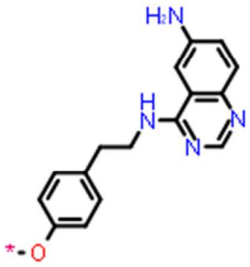
	CHEMBL666799	Cytochrome P450 19A1
	CHEMBL670726	Cytochrome P450 19A1
	CHEMBL679284	Cytochrome P450 19A1
	CHEMBL744319	Delta opioid receptor
	CHEMBL744319	Endothelin receptor ET-A

	CHEMBL746513	Endothelin receptor ET-B
	CHEMBL746513	Glucagon receptor
	CHEMBL747492	Malonyl-CoA decarboxylase
	CHEMBL747493	Malonyl-CoA decarboxylase
	CHEMBL747656	Muscarinic acetylcholine receptor M1

	CHEMBL747701	Muscarinic acetylcholine receptor M1
	CHEMBL748244	Muscarinic acetylcholine receptor M1
	CHEMBL749237	Muscarinic acetylcholine receptor M1
	CHEMBL749237	Muscarinic acetylcholine receptor M1
	CHEMBL749504	Muscarinic acetylcholine receptor M1

	CHEMBL749925	Muscarinic acetylcholine receptor M1
	CHEMBL751533	Muscarinic acetylcholine receptor M1
	CHEMBL751534	Muscarinic acetylcholine receptor M1
	CHEMBL751535	Muscarinic acetylcholine receptor M2
	CHEMBL755211	Muscarinic acetylcholine receptor M3

	CHEMBL758014	Muscarinic acetylcholine receptor M3
	CHEMBL760085	Muscarinic acetylcholine receptor M4
	CHEMBL764258	Muscarinic acetylcholine receptor M4
	CHEMBL765526	Muscarinic acetylcholine receptor M5
	CHEMBL829509	Neuraminidase

	CHEMBL859660	Peroxisome proliferator-activated receptor gamma
	CHEMBL878575	Phosphodiesterase 5A
	CHEMBL925154	Inhibition of hemolysis
	CHEMBL983195	Inhibition of NF-kappaB activation

References

(1) Mills, J. E. J.; Brown, A. D.; Ryckmans, T.; Miller, D. C.; Skerratt, S. E.; Barker, C. M.; Bunnage, M. E. *MedChemComm* **2012**, *3*, 174–178.