SUPPORTING INFORMATION

Multi-Step Protocol for Automatic Evaluation of Docking Results Based on Machine Learning Methods – a Case Study of Serotonin Receptors 5-HT_6 and 5-HT_7

*Sabina Smusz[1,2], Stefan Mordalski[1], Jagna Witek[1], Krzysztof Rataj[1], Rafał Kafel,[1] Andrzej J. Bojarski[1]**

[1]Department of Medicinal Chemistry, Institute of Pharmacology, Polish Academy of Sciences, Smętna 12, 31-343 Kraków, Poland

[2]Faculty of Chemistry, Jagiellonian University, R. Ingardena 3, 30-060 Kraków, Poland

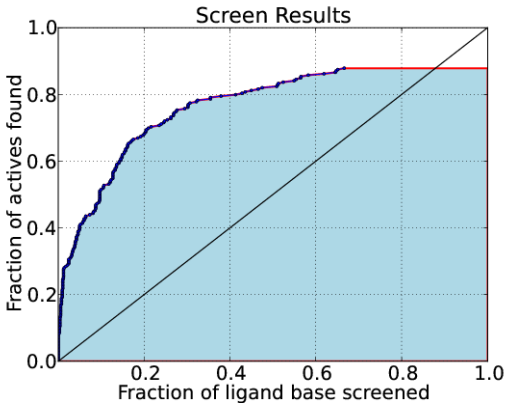
***e-mail: bojarski@if-pan.krakow.pl**

Table S1. Crystal structures used for homology modeling purposes.

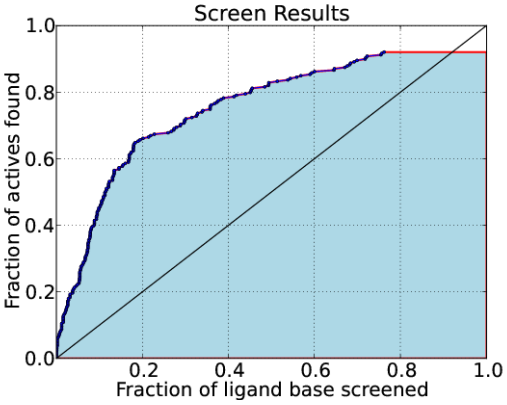
Template	PDB ID	Resolution [Å]
$5\text{-}HT_{1B}$	4IAR	2.70
$5\text{-}HT_{2B}$	4IB4	2.70
A2A	3QAK	2.71
Beta1	2Y00	2.50
Beta2	3P0G	3.50
CXCR4	3OE0	2.90
D_3	3PBL	2.89
H_1	3RZE	3.10
M_2	3UON	3.00
M_3	4DAJ	3.40

Figure S1. ROC curves obtained for the selected homology models.

5-HT$_6$_template_Beta2

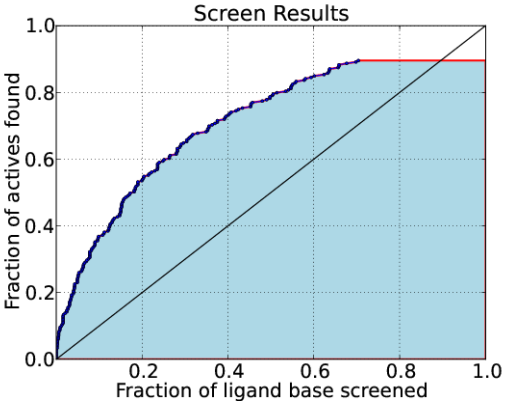


5-HT$_6$_template_CXC

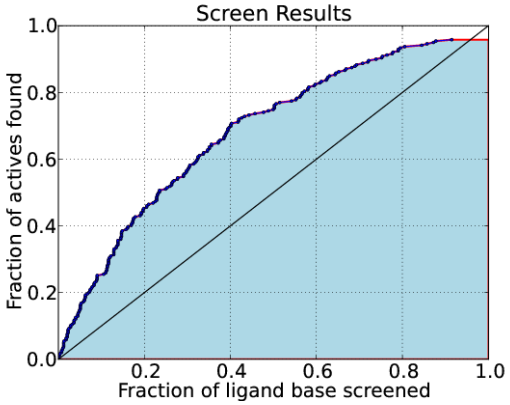


5-HT$_6$_template_A2A

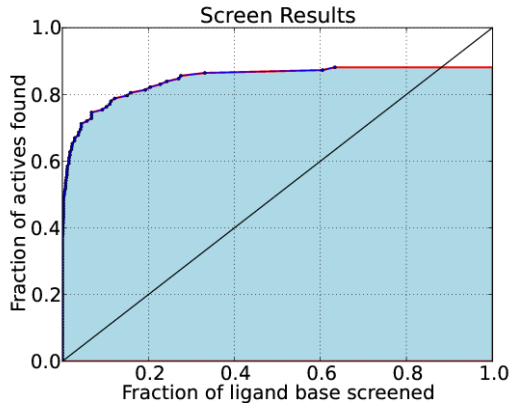
5-HT$_6$_template_D3



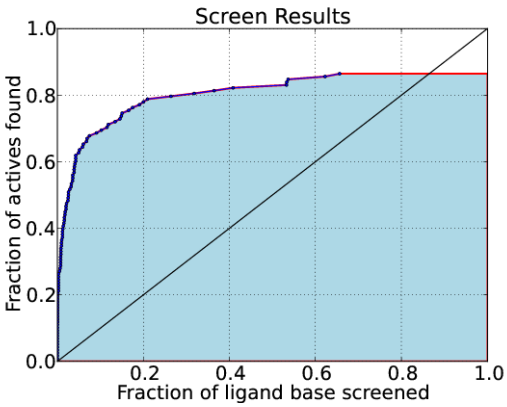
5-HT$_6$_template_M3



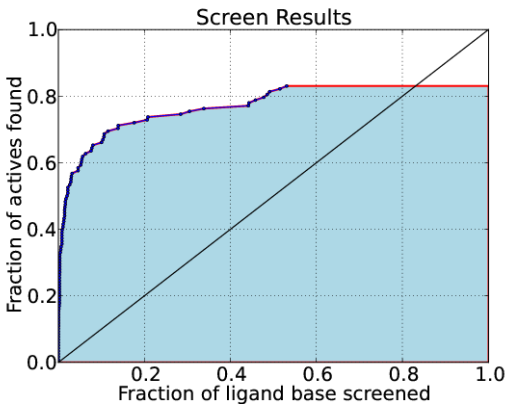
5-HT$_7$_template_H1



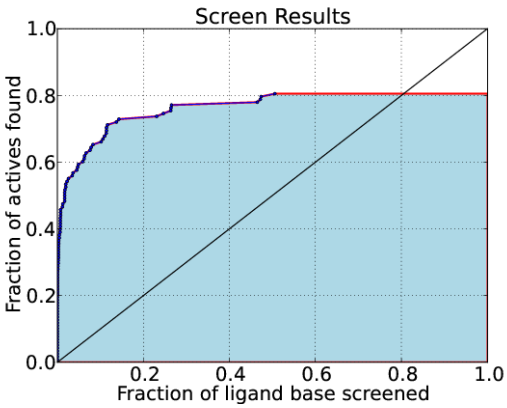
5-HT$_7$_template_Beta1



5-HT$_7$_template_D3



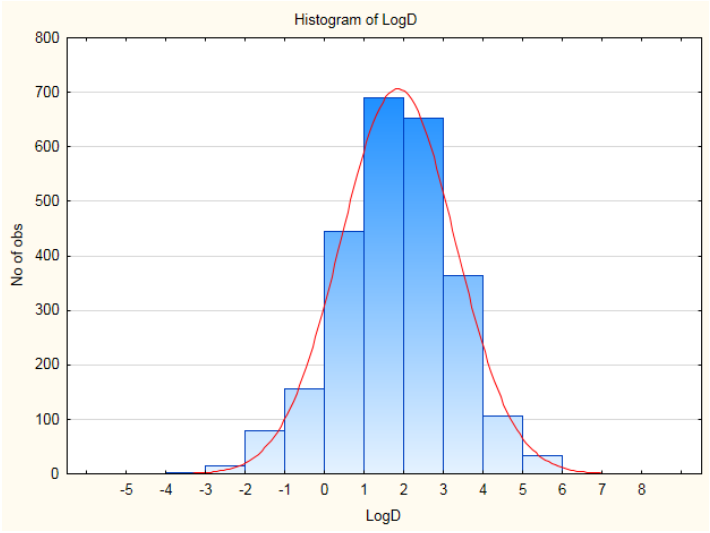
5-HT$_7$_template_Beta2

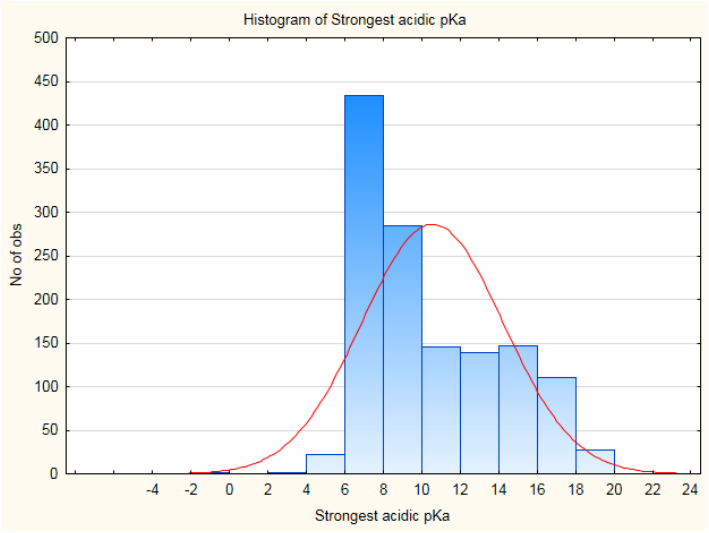


5-HT$_7$_template_M3

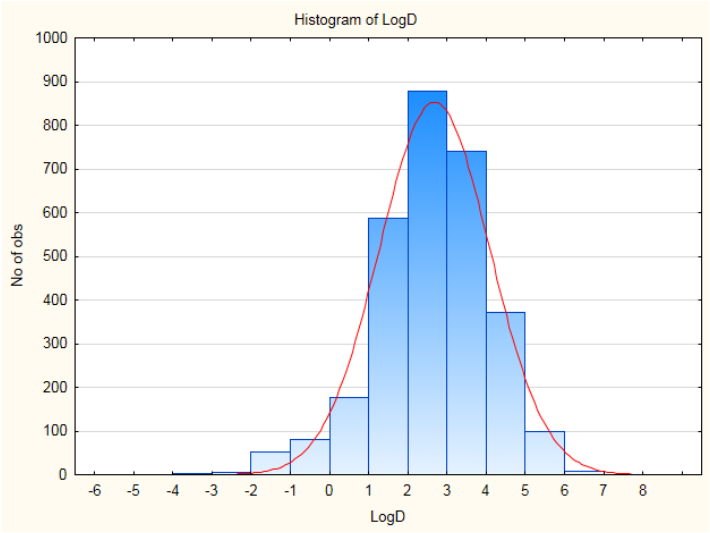
Figure S2. Distribution of logD and pKa value for 5-HT_6 (a) and 5-HT_7 (b) oriented sets of compounds

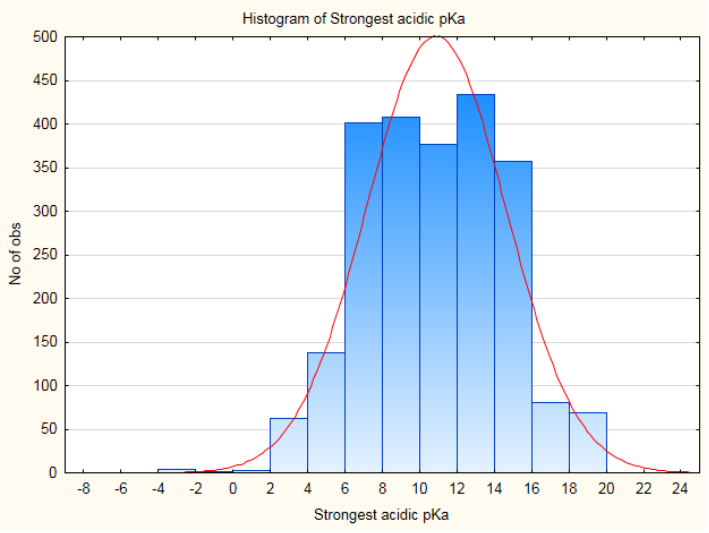
5-HT_6 - Actives



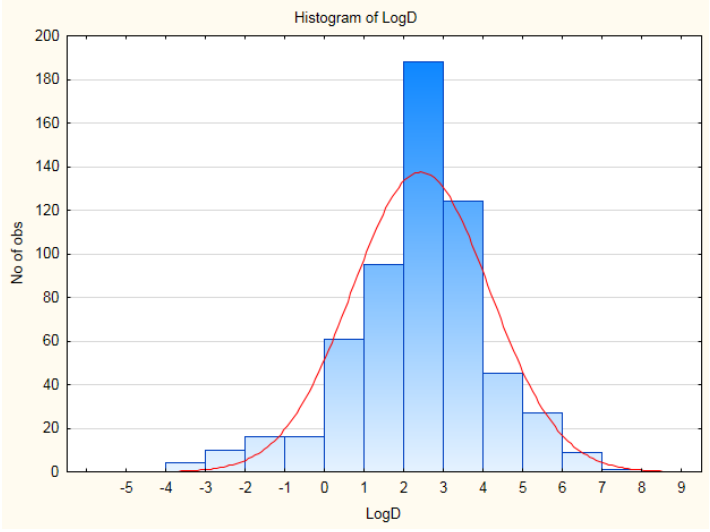


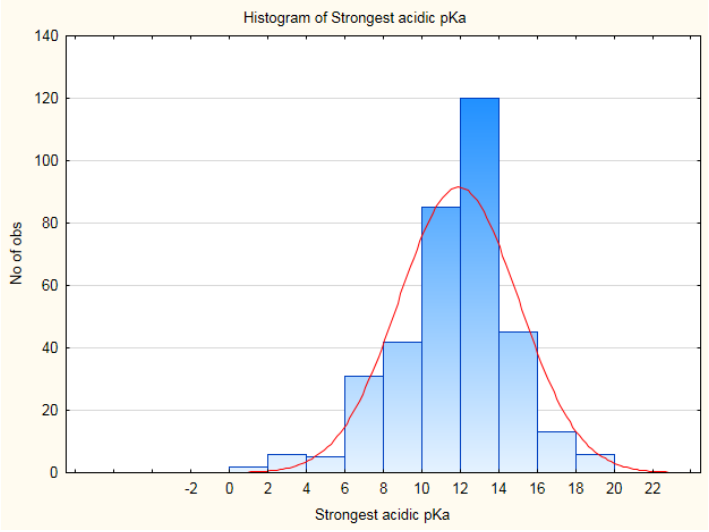
5-HT_6 - DUD





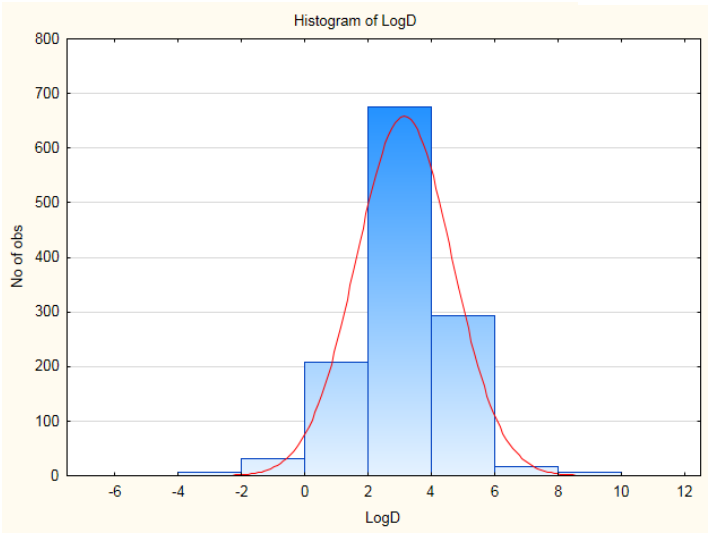
5-HT_6 - True inactives

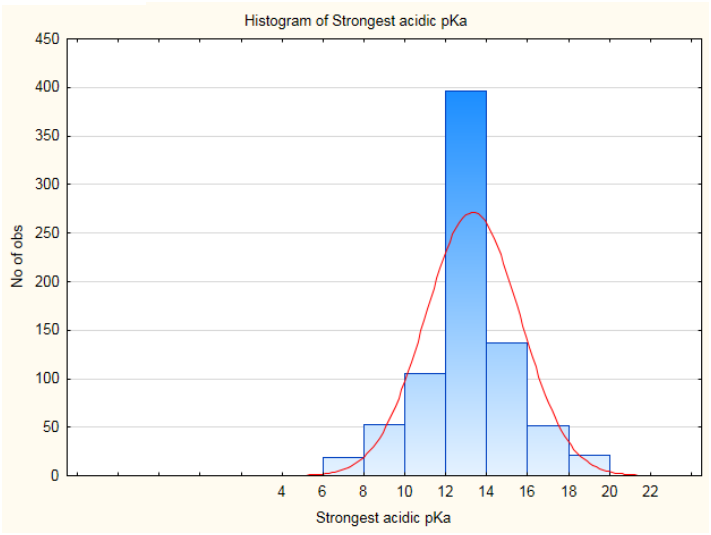




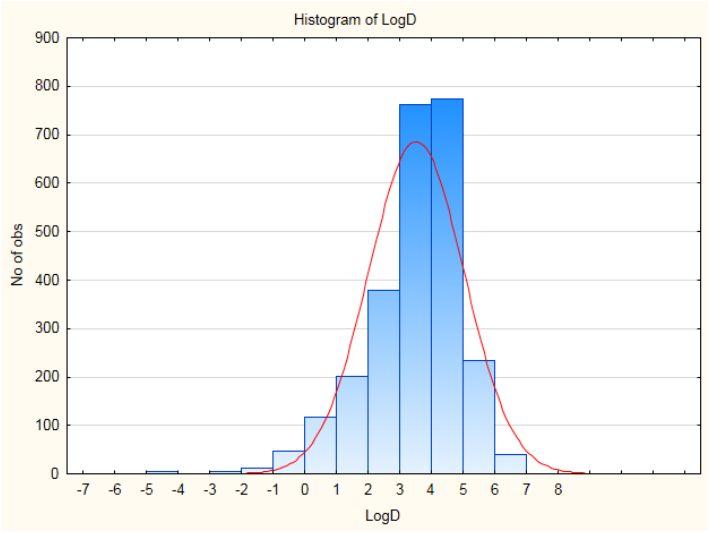
a)

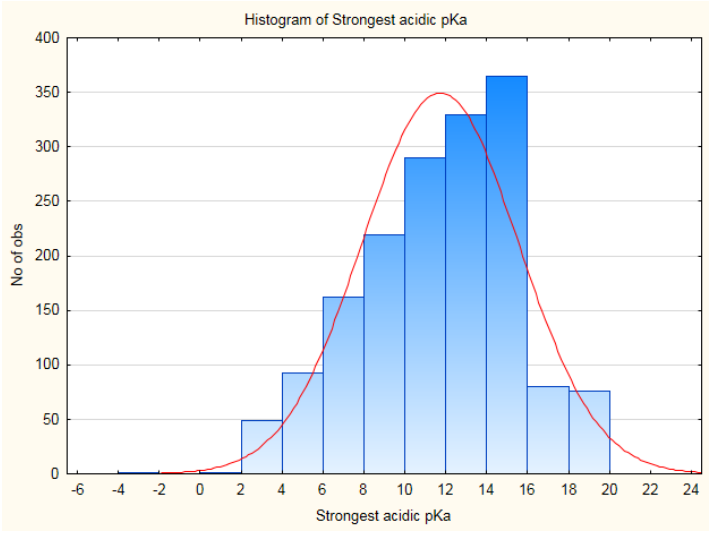
5-HT$_7$ - Actives



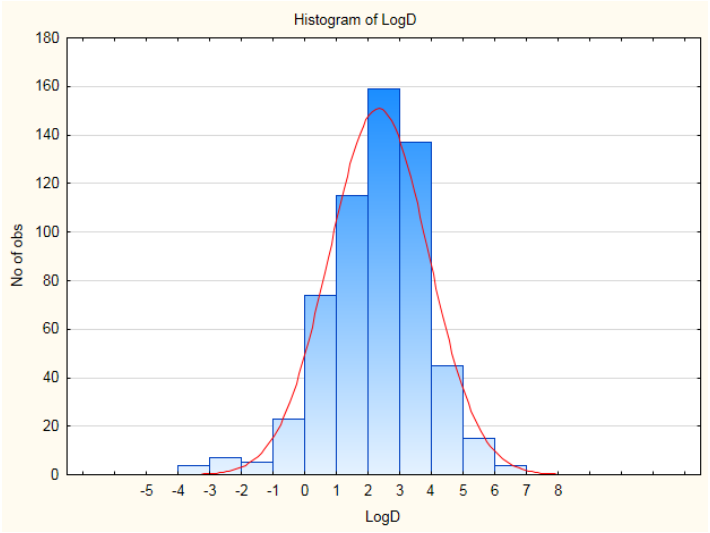


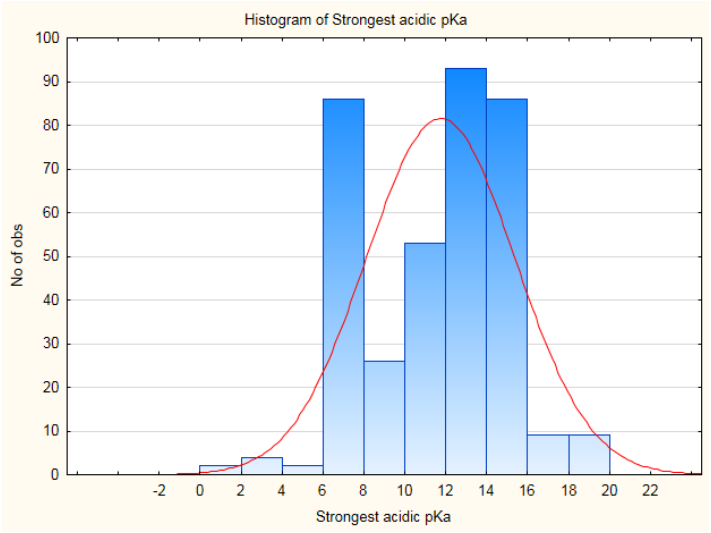
5-HT$_7$ - DUD





5-HT$_7$ - True inactives

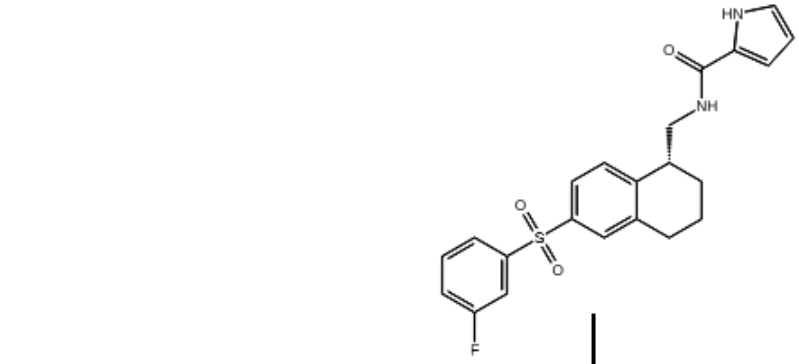




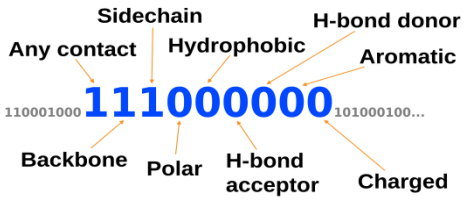
b)

Figure S3. Example of SIFts and Spectrophores generation.

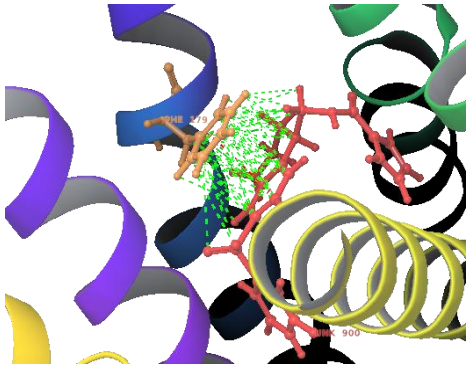
Compound active towards 5-HT$_6$ receptor: K$_i$ = 5.01 nM



SIFts

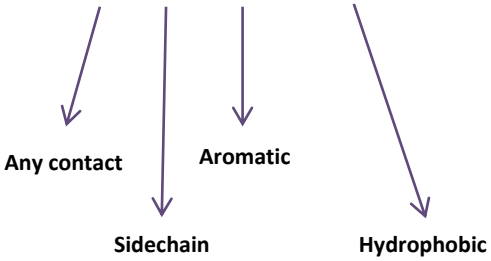


General scheme



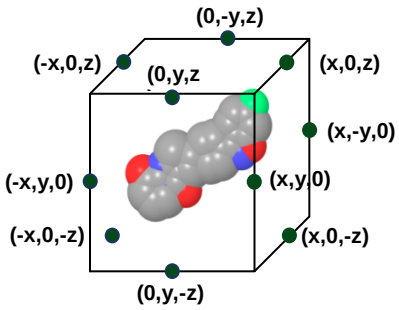
Docked ligand

...101010010...



SIFt fragment describing ligand interactions with Phe 179

Spectrophores



General scheme

6.92593; 4.69471; 6.28958;
8.57898; 11.2911; 8.91104;
8.46942; 10.4937; 11.2948;
11.5109; 13.3622; 10.5366;

Values calculated from atomic partial charges

9.68665; 11.2825; 19.8132;
21.4979; 23.4751; 18.7107;
19.0882; 29.435; 31.9205;
30.5689; 30.6511; 25.3839;

Values calculated from atomic lipophilicity properties

388.12; 597.596; 596.249;
1148.96; 1466.3; 1444.56;
1013.16; 1095.02; 1350;
1826.13; 1952.6; 1439.14;

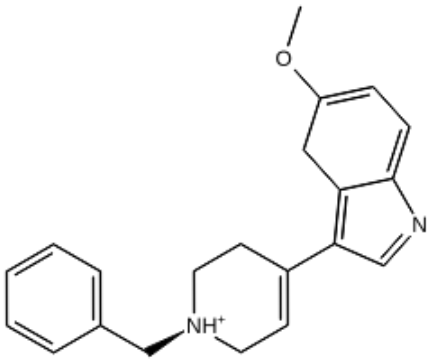
Values calculated from atomic shape deviations

2.28233; 3.09556; 3.62346;
3.99749; 3.23315; 3.77505;
1.94061; 4.94089; 4.33196;
3.61812; 3.56182; 4.97617

Values calculated from atomic electrophilicity properties

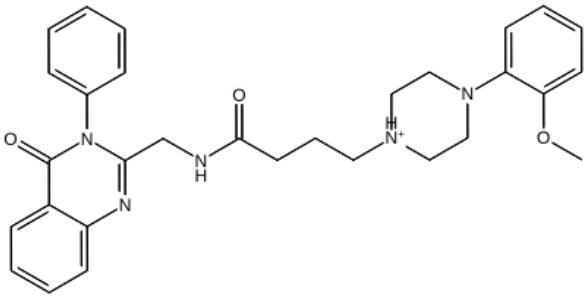
Figure S4. Exemplary analysis of activity of two compounds by docking evaluation protocol.

Compound inactive towards 5-HT$_7$ receptor: K_i = 1100 nM



template (AUROC)	ML method	ML method weight	ML predictions - conf. 1	ML predictions - conf. 2	conformations consensus (glide score weights)	template consensus (AUROC weights)	final class assignment
H1 (0.828)	IBk	1	-1	-1	-1	-0.782	-1
	J48	1	-1	-1			
	NB	0.998	-1	-1			
	RF	1	-1	-1			
	SMO	1	-1	-1			
	ML methods consensus		-1	-1			
	Glide Score		-9.1	-5.7			
Beta1 (0.786)	IBk	1	-1	-1	-1		
	J48	1	-1	-1			
	NB	1	-1	-1			
	RF	0.982	-1	-1			
	SMO	1	-1	-1			
	ML methods consensus		-1	-1			
	Glide Score		-6.9	-4.9			
D3 (0.764)	IBk	0.935	-1	+1	-0.49		
	J48	1	-1	-1			
	NB	0.973	-1	-1			
	RF	0.812	+1	+1			
	SMO	1	-1	-1			
	ML methods consensus		-0.656	-0.260			
	Glide Score		-6.6	-4.9			
Beta2 (0.757)	IBk	0.953	-1	-	-0.65		
	J48	1	-1	-			
	NB	0.966	-1	-			
	RF	0.826	+1	-			
	SMO	1	-1	-			
	ML methods consensus		-0.652	-			
	Glide Score		-4.7	-			
M3 (0.749)	IBk	0.911	-1	-1	-0.81		
	J48	1	-1	-1			
	NB	0.977	-1	-1			
	RF	0.880	+1	-1			
	SMO	1	-1	-1			
	ML methods consensus		-0.631	-1			
	Glide Score		-8.4	-8.6			

Compound active towards 5-HT$_7$ receptor: K$_i$ = 10.5 nM



template (AUROC)	ML method	ML method weight	ML predictions - conf. 1	conformations consensus (glide score weights)	template consensus (AUROC weights)	final class assignment
H1 (0.828)	IBk	1	-	0	0.389	+1
	J48	1	-			
	NB	0.998	-			
	RF	1	-			
	SMO	1	-			
	ML methods consensus		-			
	Glide Score		undocked			
Beta1 (0.786)	IBk	1	-	0		
	J48	1	-			
	NB	1	-			
	RF	0.982	-			
	SMO	1	-			
	ML methods consensus		-			
	Glide Score		undocked			
D3 (0.764)	IBk	0.935	+1	+1		
	J48	1	+1			
	NB	0.973	+1			
	RF	0.812	+1			
	SMO	1	+1			
	ML methods consensus		+1			
	Glide Score		-3.6			
Beta2 (0.757)	IBk	0.953	-	0		
	J48	1	-			
	NB	0.966	-			
	RF	0.826	-			
	SMO	1	-			
	ML methods consensus		-			
	Glide Score		undocked			
M3 (0.749)	IBk	0.911	+1	+1		
	J48	1	+1			
	NB	0.977	+1			
	RF	0.880	+1			
	SMO	1	+1			
	ML methods consensus		+1			
	Glide Score		8.0			

Table S2. The docking statistics for beta-2 adrenergic receptor crystal structures.

The initial number of compounds			Number of compounds after LigPrep			Crystal structure	Number of docked compounds		
Actives	True inactives	DUDs	Actives	True inactives	DUDs		Actives	True inactives	DUDs
271	324	2000	550	601	2526	2RH1[1]	531	574	2382
						3D4S[2]	511	539	2376
						3NY8[3]	539	582	2398
						3NY9[3]	519	564	2388
						3NYA[3]	498	521	2385

(1) Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G.; Thian, F. S.; Kobilka, T. S.; Choi, H. J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor, *Science* 2007, 318, 1258–1265.

(2) Hanson, M. A.; Cherezov, V.; Griffith, M. T.; Roth, C. B.; Jaakola, V. P.; Chien, E. Y.; Velasquez, J.; Kuhn, P.; Stevens, R. C. A specific cholesterol binding site is established by the 2.8 A structure of the human beta2-adrenergic receptor, *Structure* **2008**, 16, 897–905.

(3) Wacker, D.; Fenalti, G.; Brown, M. A.; Katritch, V.; Abagyan, R.; Cherezov, V.; Stevens, R. C. Conserved binding mode of human beta2 adrenergic receptor inverse agonists and antagonist revealed by X-ray crystallography, *J. Am. Chem. Soc.* **2010**, 132, 11443–11445.

Figure S5. The comparison between MCC values calculated for the raw docking results for 5 best models and those obtained after the application of docking results analysis protocol in 3-fold CV for 5-HT_6 (a) and 5-HT_7 (b) ligands.

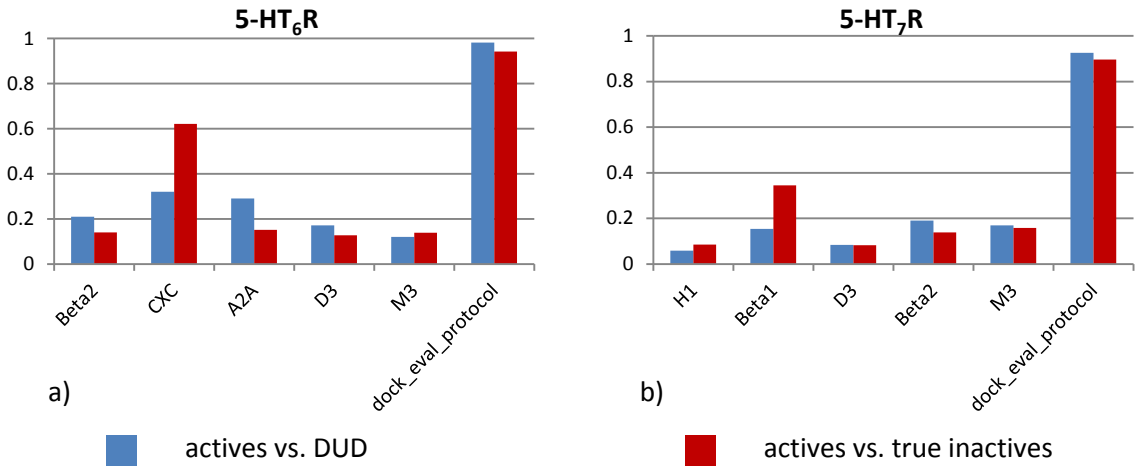


Figure S6. MCC values obtained after applying the docking results analysis protocol for various representations of ligand-protein complexes for 5-HT_6 (a) and 5-HT_7 (b) ligands for 3-fold CV.

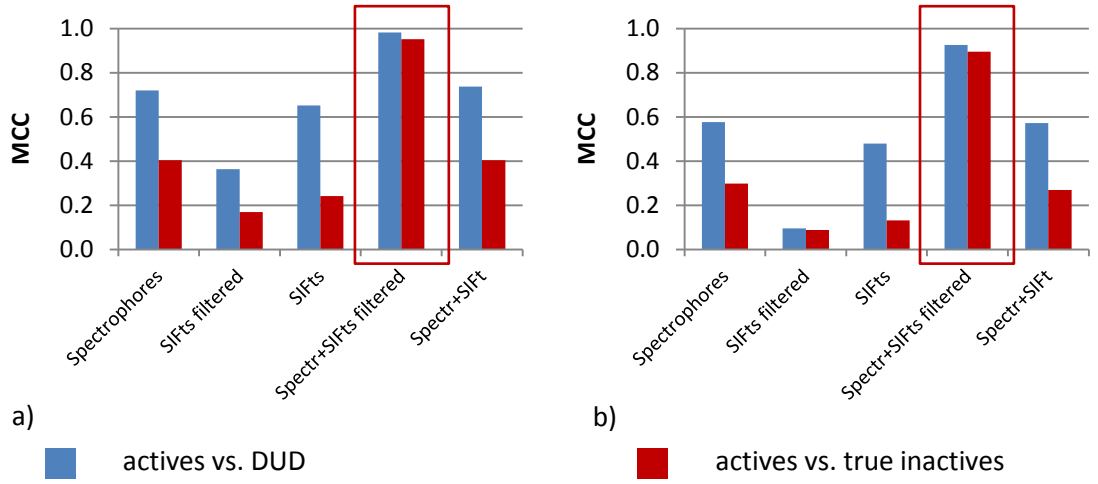
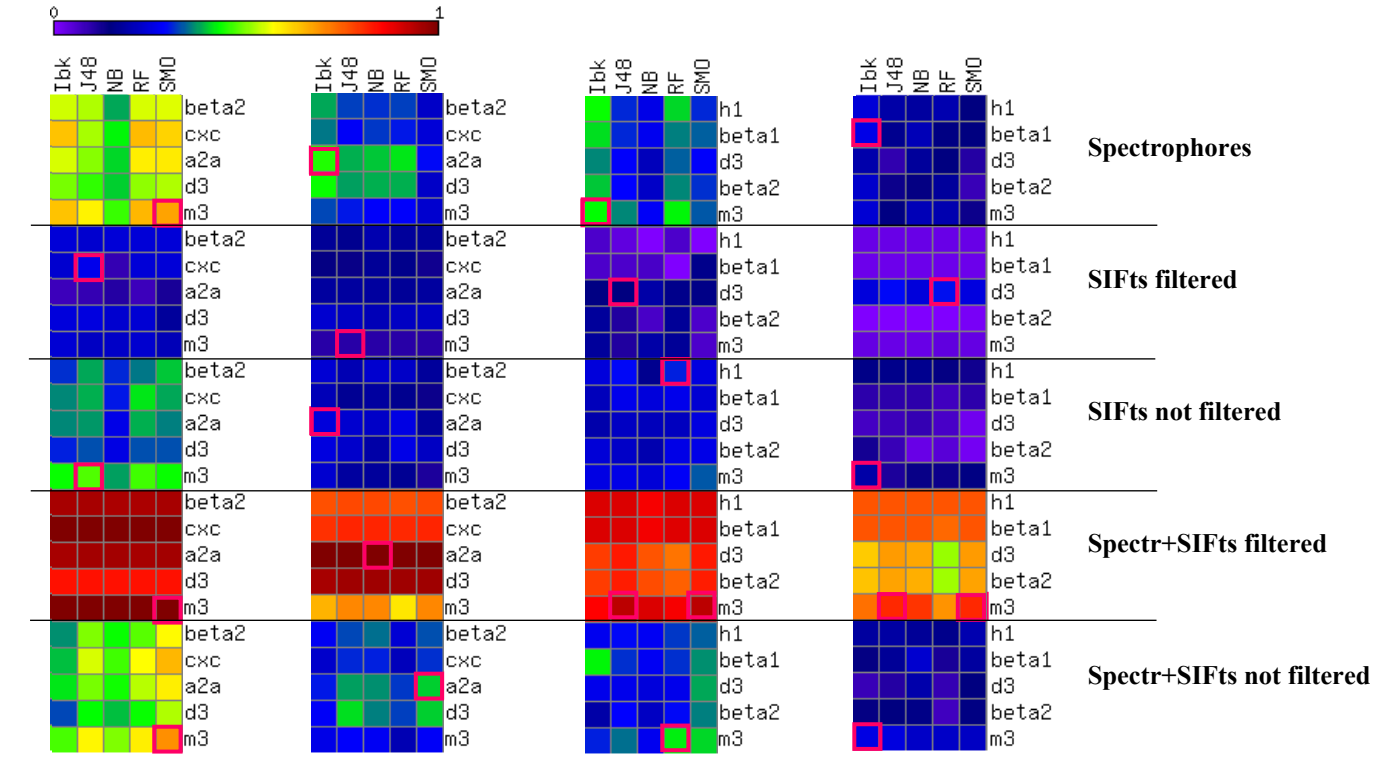


Figure S7. Comparison of the developed protocol with the performance of the individual ML experiment providing the highest classification efficiency in 3-fold CV.



5-HT_6 act vs. DUD 5-HT_6 act vs. trueinact 5-HT_7 act vs. DUD 5-HT_7 act vs. trueinact

Figure S8. Comparison of the developed protocol with the performance of the individual ML experiment providing the highest classification efficiency in 3-fold CV.

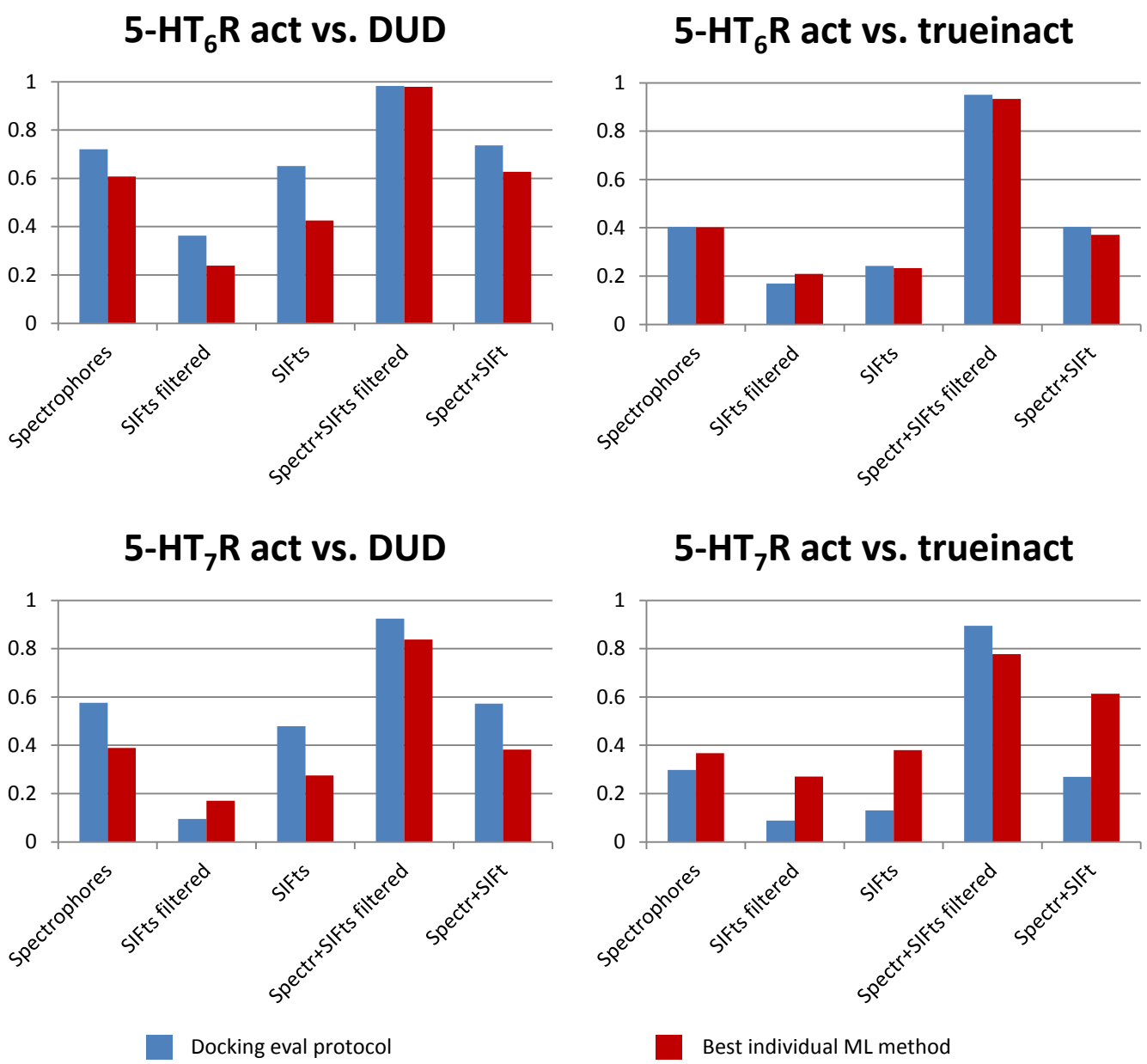


Table S3. Numerical values of MCC parameters obtained for various representations of ligand-protein complexes for 3-fold CV

Target	Type of experiment/MCC value				
5-HT_6	**actives vs DUDs**				
	Spectrophores	SIFts filtered	SIFts	Spectr+ SIFts filtered	Spectr+ SIFts
	0.720	0.363	0.651	**0.982**	0.737
	actives vs true inactives				
	Spectrophores	SIFts filtered	SIFts	Spectr+ SIFts filtered	Spectr+ SIFts
	0.404	0.169	0.242	**0.951**	0.404
5-HT_7	**actives vs DUDs**				
	Spectrophores	SIFts filtered	SIFts	Spectr+ SIFts filtered	Spectr+ SIFts
	0.576	0.096	0.479	**0.925**	0.572
	actives vs true inactives				
	Spectrophores	SIFts filtered	SIFts	Spectr+ SIFts filtered	Spectr+ SIFts
	0.298	0.088	0.131	**0.895**	0.270

Figure S9. The comparison between the MCC values obtained for individual beta-2 adrenergic receptor crystal structures (the best ML method) and with the use of docking evaluation protocol.

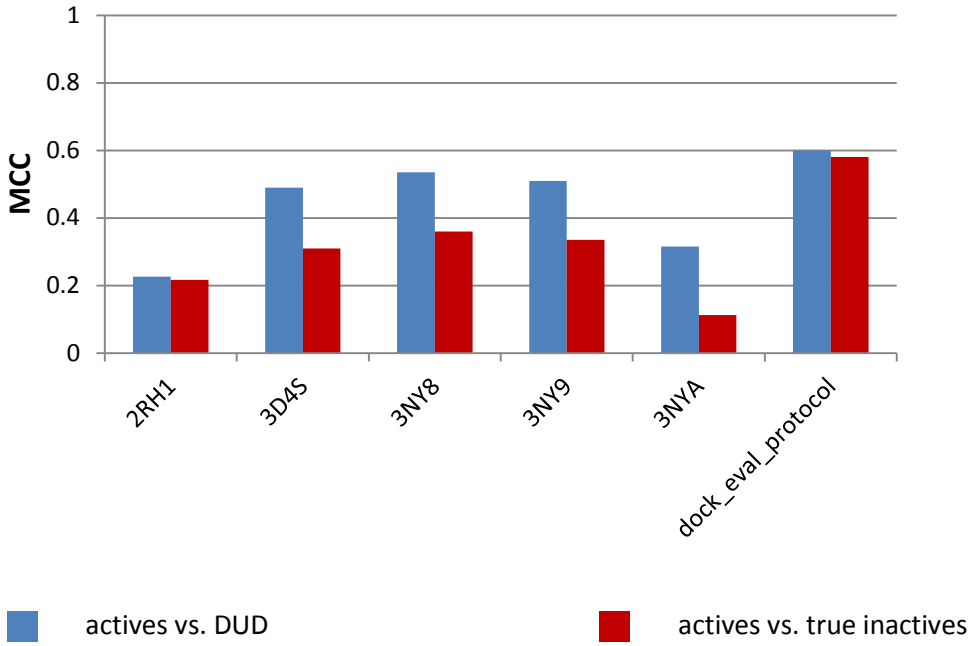


Table S4. Optimal values and standard deviation of the weights assigned to undocked compounds providing the best performance of the whole evaluation protocol

5-HT$_6$		**actives vs DUDs**				
		Spectrophores	SIFts filtered	SIFts not filtered	Spectr+ SIFts filtered	Spectr+ SIFts not filtered
	Optimal weight value	**-0.4**	**-0.4**	**-0.2**	**-0.1**	**-0.1**
	SD of MCC over various weight values	0.045	0.062	0.054	0.038	0.049
		actives vs true inactives				
		Spectrophores	SIFts filtered	SIFts not filtered	Spectr+ SIFts filtered	Spectr+ SIFts not filtered
	Optimal weight value	**-0.5**	**-0.5**	**-0.4**	**0**	**-0.4**
	SD of MCC over various weight values	0.025	0.083	0.046	0.012	0.032
5-HT$_7$		**actives vs DUDs**				
		Spectrophores	SIFts filtered	SIFts not filtered	Spectr+ SIFts filtered	Spectr+ SIFts not filtered
	Optimal weight value	**-0.2**	**-0.5**	**-0.2**	**-0.1**	**-0.3**
	SD of MCC over various weight values	0.094	0.067	0.11	0.112	0.105
		actives vs true inactives				
		Spectrophores	SIFts filtered	SIFts not filtered	Spectr+ SIFts filtered	Spectr+ SIFts not filtered
	Optimal weight value	**0**	**-0.5**	**0**	**0**	**0**
	SD of MCC over various weight values	0.006	0.047	0.012	0.03	0.013

Figure S10. Analysis of the influence of threshold changes on ML protocol performance for Spectrophores+SIFTs representation for a given template.