

Supporting Information:

Identification of Allosteric Modulators of
Metabotropic Glutamate 7 Receptor Using
Proteochemometric Modeling

Gary Tresadern,^{a,} Andres A. Trabanco,^b Laura Pérez-Benito,^a John P. Overington,^c Herman W. T. van Vlijmen,^d Gerard J. P. van Westen,^{c,*}*

^a Computational chemistry, ^b Neuroscience Medicinal Chemistry, Janssen Research & Development, Janssen-Cilag S.A., Jarama 75A, 45007, Toledo, Spain. ^c ChEMBL Group, EMBL -EBI, Wellcome Trust Genome Campus, CB10 1SD, Hinxton, United Kingdom. ^d Computational chemistry, Janssen Research & Development, Turnhoutseweg 30, B-2340, Beerse, Belgium.

* Corresponding author phone: +32 1460 2111. E-mail: gtresade@its.jnj.com; phone: +31 71 527 3511. E-mail: gerard@lacdr.leidenuniv.nl

Contents:

• Method for selectivity assays	p3
• Table S1. Example of training and test set distribution	p4
• Table S2. Aligned Residues used.	P5
• Table S3. Euclidian distance matrix using the descriptors used in the model.	p6
• Table S4: Physicochemical properties used in addition to fingerprints	p7
• Table S5. Smiles and purity information of identified hits	p8
• Figure S1. Sequence identity between 7-TM domains of human mGlu receptors	p9
• Figure S2. Example of the sequence alignment human mGlu receptors	p10
• Figure S3: Chemical space mean distance versus target space mean distance	p11
• Figure S4. ROC curve for Out-of-Bag validation.	p12
• Figure S5. Model quality in different validation experiments	p13
• Figure S6. Properties of queries used for fingerprint analogue searches.	p14
• Figure S7. Comparing PCM highly ranked false positives to true positives	p15

Method for selectivity assays

mGlu receptor panel selectivity assays: Ca^{2+} assays with human mGlu1, 3, 5, 7, or 8 receptor expressing HEK 293 cells were performed as reported in Lavreysen et al. (2013), except for a slight change in the procedure for mGlu₅: cells expressing the human mGlu₅ receptor were seeded at 40,000 cells/well in MW384. Twenty-four hours after seeding, cells were incubated for 90 min in Ca^{2+} assay kit (Molecular Devices) dissolved in saline PBS supplemented with 5 mmol/L probenecid, pH 7.4 (f.c. 2.5 mmol/L probenecid as loading buffer was added on the cell layer without removal of medium) before measurements. Measurement of [³⁵S]GTP γ S binding to membranes from CHO cells expressing the rat mGlu₆ receptor and membranes from L929sA cells expressing the human mGlu₄ receptor were conducted also as described in Lavreysen et al. 2013.

Lavreysen H, *et al.* Pharmacological Characterization of JNJ-40068782, a New Potent, Selective, and Systemically Active Positive Allosteric Modulator of the mGlu2 Receptor and Its Radioligand [3H]JNJ-40068782. *J. Pharmacol. Exp. Ther.* **2013**, 346, 514- 527.

Supporting Table S1: Example of training and test set distribution

Receptor	Actives_Training			Inactives_Training			Actives_Testing			Inactives_Testing		
	Total	ChEMBL	Janssen	Total	ChEMBL	Janssen	Total	ChEMBL	Janssen	Total	ChEMBL	Janssen
grm1_human	391	366	25	396	1	395				3765	8	3757
grm1_mouse							15	15				
grm1_rat	28	28		26	26		288	288				
grm2_human	2234	296	1938	2231	6	2225				1481	3	1478
grm2_rat	3	3		4	4		237	237				
grm3_human	269	13	256	272	1	271				3191	4	3187
grm3_rat	5	5		3	3		24	24				
grm4_human	99	88	11	86		86				3844	11	3833
grm4_rat							32	32				
grm5_human	1422	996	426	1401	14	1387				2341	17	2324
grm5_mouse							2	2				
grm5_rat	56	56		46	46		588	588				
grm6_human	1	1		2	2					2	2	
grm6_rat	5		5	5		5				4084		4084
grm7_human	23		23	17		17				3957		3957
grm7_rat							20	20				
grm8_human	13	5	8	13		13				3734		3734
Total	4549			4502			1206			26399		
Fraction	0.50			0.50			0.04			0.96		

Supporting Table S2: Aligned Residues used. Residues marked in yellow were removed in the model due to the absence of variance over the sequences.

		GAPS				TM2				GAPS			TM3				GAPS			T M 4	G A P S	E C L 2	GAPS							TM5					GAPS				TM6				GAPS						TM7							G A P								
S e q u e n c e s		1	2	3	4	2	2	2	2	2	2	5	6	7	3	3	3	3	3	3	3	3	8	9	0	4	1	1	2	5	2	1	1	1	1	1	1	5	5	5	5	5	1	1	2	2	6	6	6	6	6	2	2	2	2	2	2	7	7	7	7	7	7	4
HUMAN	1	-	-	-	-	I	G	I	G	C	L	-	-	-	Q	R	V	G	S	S	C	Y	-	-	-	Q	-	-	C	T	-	-	-	-	-	-	V	P	L	N	I	-	-	-	-	T	I	W	F	Y	-	-	-	-	-	-	I	T	A	V	S	V	A	-
	2	-	-	-	-	L	G	V	C	M	F	-	-	-	R	R	L	G	A	F	C	Y	-	-	-	Q	-	-	C	H	-	-	-	-	-	-	M	S	L	N	I	-	-	-	-	T	I	W	F	F	-	-	-	-	-	-	T	M	S	V	S	G	V	-
	3	-	-	-	-	L	G	V	S	M	F	-	-	-	R	R	L	G	S	F	C	Y	-	-	-	Q	-	-	C	V	-	-	-	-	-	-	M	S	L	D	V	-	-	-	-	T	I	W	F	F	-	-	-	-	-	-	T	M	S	V	S	G	V	-
	4	-	-	-	-	L	G	I	C	T	M	-	-	-	R	R	L	G	G	M	S	Y	-	-	-	Q	-	-	C	I	-	-	-	-	-	-	L	L	L	S	M	-	-	-	-	T	V	W	F	F	-	-	-	-	-	-	T	L	S	V	S	A	S	-
	5	-	-	-	-	I	G	I	G	C	L	-	-	-	Q	R	I	G	S	P	S	Y	-	-	-	Q	-	-	C	T	-	-	-	-	-	-	V	P	L	N	I	-	-	-	-	T	I	W	F	Y	-	-	-	-	-	-	I	M	S	V	S	A	A	-
	6	-	-	-	-	L	G	I	I	I	M	-	-	-	R	R	L	G	G	T	S	Y	-	-	-	Q	-	-	C	M	-	-	-	-	-	-	L	C	L	S	M	-	-	-	-	T	I	W	F	F	-	-	-	-	-	-	T	L	S	L	S	A	S	-
	7	-	-	-	-	L	G	I	C	I	M	-	-	-	R	R	L	G	G	M	S	Y	-	-	-	Q	-	-	C	I	-	-	-	-	-	-	I	S	L	S	M	-	-	-	-	T	V	W	F	F	-	-	-	-	-	-	T	L	S	M	S	A	A	-
	8	-	-	-	-	L	G	I	C	I	M	-	-	-	R	R	L	G	G	M	S	Y	-	-	-	Q	-	-	C	I	-	-	-	-	-	-	L	S	L	S	M	-	-	-	-	T	I	W	F	F	-	-	-	-	-	-	T	L	S	M	S	A	S	-
RAT	1	-	-	-	-	I	G	I	G	C	L	-	-	-	Q	R	V	G	S	S	C	Y	-	-	-	Q	-	-	C	T	-	-	-	-	-	-	V	P	V	N	I	-	-	-	-	T	I	W	F	Y	-	-	-	-	-	-	I	T	A	V	S	V	A	-
	2	-	-	-	-	L	G	V	C	M	F	-	-	-	R	R	L	G	A	F	C	Y	-	-	-	Q	-	-	C	H	-	-	-	-	-	-	M	S	L	N	I	-	-	-	-	T	I	W	F	F	-	-	-	-	-	-	T	M	S	V	S	G	V	-
	3	-	-	-	-	L	G	V	S	M	F	-	-	-	R	R	L	G	S	F	C	Y	-	-	-	Q	-	-	C	V	-	-	-	-	-	-	M	S	L	D	V	-	-	-	-	T	I	W	F	F	-	-	-	-	-	-	T	M	S	V	S	G	V	-
	4	-	-	-	-	L	G	I	C	T	M	-	-	-	R	R	L	G	G	M	S	Y	-	-	-	Q	-	-	C	I	-	-	-	-	-	-	L	L	L	S	M	-	-	-	-	T	V	W	F	F	-	-	-	-	-	-	T	L	S	V	S	A	S	-
MOUSE	5	-	-	-	-	I	G	I	G	C	L	-	-	-	Q	R	I	G	S	P	S	Y	-	-	-	Q	-	-	C	T	-	-	-	-	-	-	V	P	L	N	I	-	-	-	-	T	I	W	F	Y	-	-	-	-	-	-	I	M	S	V	S	A	A	-
	8	-	-	-	-	L	G	I	C	I	M	-	-	-	R	R	L	G	G	M	S	Y	-	-	-	Q	-	-	C	I	-	-	-	-	-	-	L	S	L	S	M	-	-	-	-	T	I	W	F	F	-	-	-	-	-	-	T	L	S	M	S	A	S	-

Supporting Table S3: Euclidian distance matrix using the descriptors used in the model.

Receptor	grm1 human	grm1 mouse	grm1 rat	grm2 human	grm2 rat	grm3 human	grm3 rat	grm4 human	grm4 rat	grm5 human	grm5 mouse	grm5 rat	grm6 human	grm6 rat	grm7 human	grm7 rat	grm8 human
grm1 human		0.40	0.40	1.16	1.16	1.13	1.13	1.10	1.10	0.65	0.65	0.65	1.15	1.15	1.12	1.12	1.10
grm1 mouse	0.40		0.00	1.23	1.23	1.20	1.20	1.18	1.18	0.76	0.76	0.76	1.22	1.22	1.19	1.19	1.18
grm1 rat	0.40	0.00		1.23	1.23	1.20	1.20	1.18	1.18	0.76	0.76	0.76	1.22	1.22	1.19	1.19	1.18
grm2 human	1.16	1.23	1.23		0.00	0.57	0.57	1.01	1.01	1.01	1.01	1.01	1.02	1.02	0.94	0.94	0.92
grm2 rat	1.16	1.23	1.23	0.00		0.57	0.57	1.01	1.01	1.01	1.01	1.01	1.02	1.02	0.94	0.94	0.92
grm3 human	1.13	1.20	1.20	0.57	0.57		0.00	0.98	0.98	0.99	0.99	0.99	1.04	1.04	0.92	0.92	0.90
grm3 rat	1.13	1.20	1.20	0.57	0.57	0.00		0.98	0.98	0.99	0.99	0.99	1.04	1.04	0.92	0.92	0.90
grm4 human	1.10	1.18	1.18	1.01	1.01	0.98	0.98		0.00	0.95	0.95	0.95	0.68	0.68	0.46	0.46	0.52
grm4 rat	1.10	1.18	1.18	1.01	1.01	0.98	0.98	0.00		0.95	0.95	0.95	0.68	0.68	0.46	0.46	0.52
grm5 human	0.65	0.76	0.76	1.01	1.01	0.99	0.99	0.95	0.95		0.00	0.00	1.03	1.03	0.96	0.96	0.94
grm5 mouse	0.65	0.76	0.76	1.01	1.01	0.99	0.99	0.95	0.95	0.00		0.00	1.03	1.03	0.96	0.96	0.94
grm5 rat	0.65	0.76	0.76	1.01	1.01	0.99	0.99	0.95	0.95	0.00	0.00		1.03	1.03	0.96	0.96	0.94
grm6 human	1.15	1.22	1.22	1.02	1.02	1.04	1.04	0.68	0.68	1.03	1.03	1.03		0.00	0.64	0.64	0.54
grm6 rat	1.15	1.22	1.22	1.02	1.02	1.04	1.04	0.68	0.68	1.03	1.03	1.03	0.00		0.64	0.64	0.54
grm7 human	1.12	1.19	1.19	0.94	0.94	0.92	0.92	0.46	0.46	0.96	0.96	0.96	0.64	0.64		0.00	0.35
grm7 rat	1.12	1.19	1.19	0.94	0.94	0.92	0.92	0.46	0.46	0.96	0.96	0.96	0.64	0.64	0.00		0.35
grm8 human	1.10	1.18	1.18	0.92	0.92	0.90	0.90	0.52	0.52	0.94	0.94	0.94	0.54	0.54	0.35	0.35	

Supporting Table S4. Physicochemical Descriptors used

Descriptor	Notes
CMP_Formal_Charge	Calculated in PP after ionization at 7.4
CMP_AlogP	
CMP_Num_atoms	
CMP_Num_Bonds	
CMP_Num_Hydrogens	
CMP_Positive_Atoms	
CMP_Negative_Atoms	
CMP_Ring_Bonds	
CMP_Rotatable_Bonds	
CMP_Aromatic_Bonds	
CMP_Bridge_Bonds	
CMP_Num_Rings	
CMP_Aromatic_Rings	
CMP_Ring_Assemblies	
CMP_Num_Chains	
CMP_Chain_Assemblies	
CMP_Molecular_Weight	
CMP_H_Acceptors	
CMP_H_Donors	
CMP_SP3_Carbon_fraction	
CMP_SP2_Carbon_fraction	
CMP_SP_Carbon_fraction	
CMP_Total_Atoms	Heavy atoms and hydrogens
CMP_Aliphatic_Rings	
CMP_Aromatic_Bonds_Frac	
CMP_Bridgebonds_Frac	
CMP_Ringbonds_Frac	
CMP_Aliphatic_Ring_Bonds_frac	
CMP_Rotatable_Bonds_Frac	
CMP_Positive_Atoms_Frac	Out of heavy atoms
CMP_Negative_Atoms_Frac	Out of heavy atoms
CMP_H_Acceptors_Fraction	
CMP_H_Donors_fraction	
CMP_Rigidity_Index	$(\text{AromaticBonds_Frac} + (1 - \text{RotatableBonds_Frac}) + \text{Aliphatic_Ringbonds_Frac}) / 3$

Supporting Table S5. Smiles and purity information of identified hits

Number in Paper	Method	Purity by mass % (Multiple values are repeat measurements)	Source	mGlu7 PAM pEC50	mGlu7 EMAX (%)	Smiles
1	FPrint Active Analogues	100, 100, 100	Internal synthesis	4.90	86	<chem>COc1c(F)cccc1COc2ccc(cn2)C(=O)N[C@H](C)C(C)(C)C{A20=R}</chem>
2	FPrint Active Analogues	100, 100	Internal synthesis	4.90	103	<chem>COc1cc(F)cc(c1)COc2ccc(cn2)C(=O)N[C@H](C)C(C)(C)C{A20=R}</chem>
3	PCM	91, 94, 98, 99	Internal synthesis	5.80	76	<chem>CNC(=O)N1CCCC(CN2CCC(CC2)OC(c3cccc3)c4cccc4)C1</chem>
4	PCM	100	Internal synthesis	4.50	57	<chem>CC(C)(C)OC(=O)N1CCc2ccc(cc2CC1)n3cc(cn3)c4ccncc4</chem>
5	PCM	-	Asinex	4.80	99	<chem>COc1ccc(cc1)CN2CCC3=C(C2)N=C(NC3=O)N4CCC(C)CC4</chem>
6	PCM	86	ChemOvation	<4.50	65	<chem>CC1CCN(CC1)c2nc(nc3cnccc23)c4ccncc4</chem>

	1	2	3	4	5	6	7	8	9	10	11
1:GRM1_XRAY_...		96.2	47.6	47.2	47.2	43.9	43.0	74.7	40.8	44.0	44.0
2:GRM1_HUMAN	98.8		48.0	47.6	47.6	44.6	43.7	78.1	41.5	44.8	44.8
3:GRM2_HUMAN	49.6	48.7		99.3	99.3	74.7	52.0	50.9	48.0	49.8	50.9
4:GRM2_RAT	49.2	48.3	99.3		99.3	74.7	52.0	50.6	48.0	49.8	50.9
5:GRM2_MOUSE	49.2	48.3	99.3	99.3		74.3	52.0	50.9	48.0	50.2	51.3
6:GRM3_HUMAN	45.7	45.3	74.7	74.7	74.3		46.6	47.9	45.8	46.9	47.7
7:GRM4_HUMAN	46.1	45.7	53.5	53.5	53.5	48.0		49.1	75.5	78.3	82.7
8:GRM5_HUMAN	76.7	78.1	50.2	49.8	50.2	47.2	46.9		43.7	46.6	47.3
9:GRM6_HUMAN	43.8	43.4	49.4	49.4	49.4	47.2	75.5	45.7		72.6	76.2
10:GRM7_HUMAN	47.3	46.8	51.3	51.3	51.7	48.3	78.3	48.7	72.6		85.9
11:GRM8_HUMAN	47.3	46.8	52.4	52.4	52.8	49.1	82.7	49.4	76.2	85.9	

Figure S1. Sequence identity between 7-TM domains of mGlu receptors. Proteins are identified with their gene ID's GRM#, where #1-8 corresponds to the equivalent receptor protein. Sequence identity within mGlu receptor subgroups mGlu 1&5, 2&3, 4-6-7&8 is typically in the range 75-85% whereas between members of different groups it is typically in the range of 45-50%.

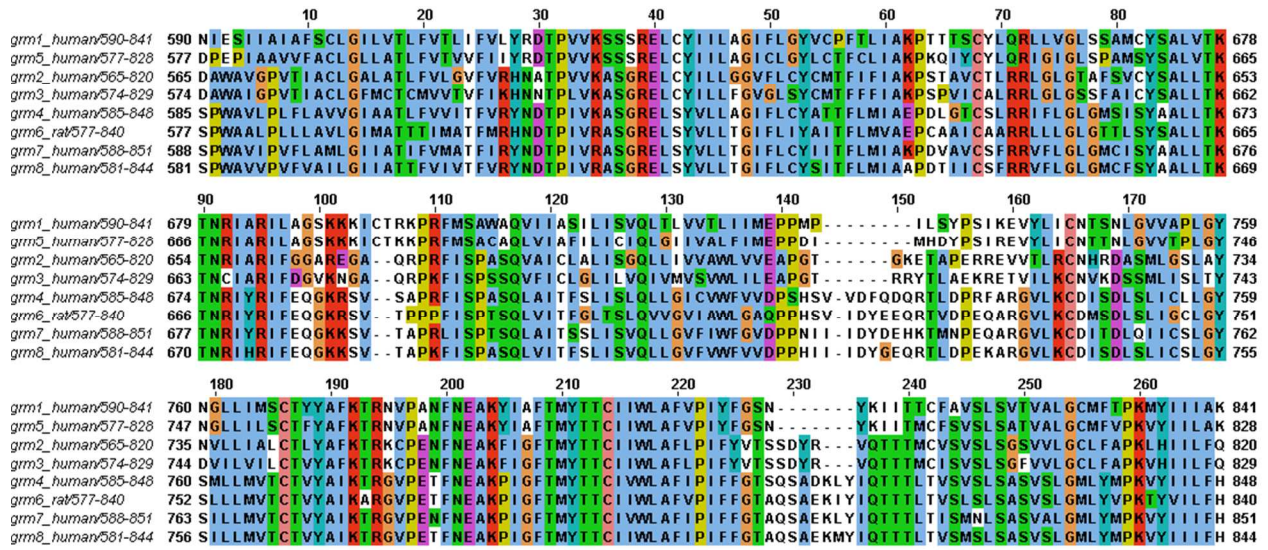
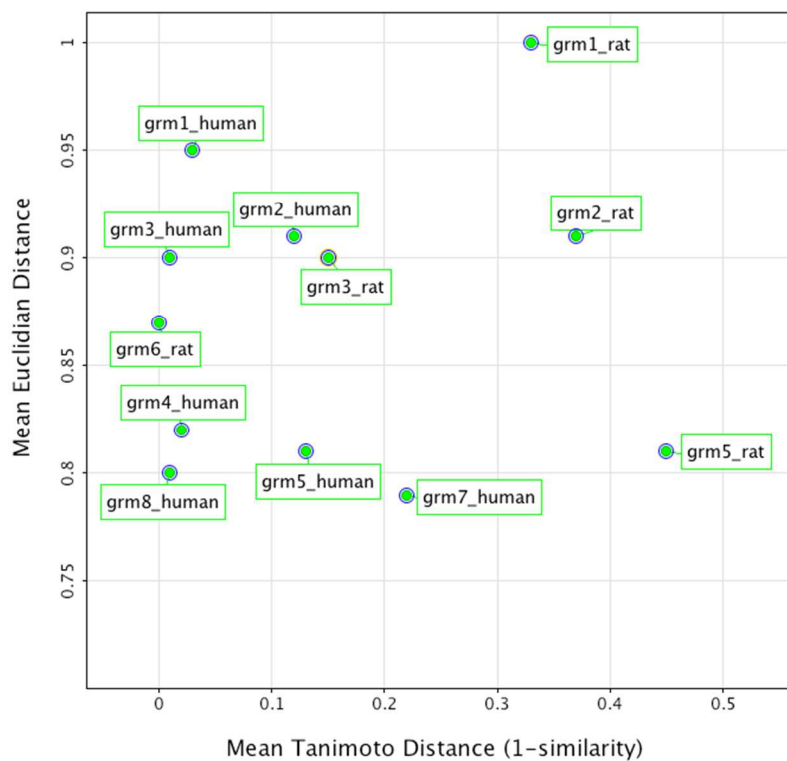


Figure S2. Example of the sequence alignment for selected amino acids in the 7-TM.



Supporting Figure S3: Mean Tanimoto distance (chemical descriptors) plotted versus the mean Euclidian distance (sequence descriptors). The worst performing receptor in the learning curve (rat mGlu5) is shown to have the largest average chemical distance to the training set, while the Euclidian distance to the training set is rather low. In fact, the distance to the human and mouse orthologs is 0. We speculate that the high chemical distance combined with the low number of actives led to the poor performance.

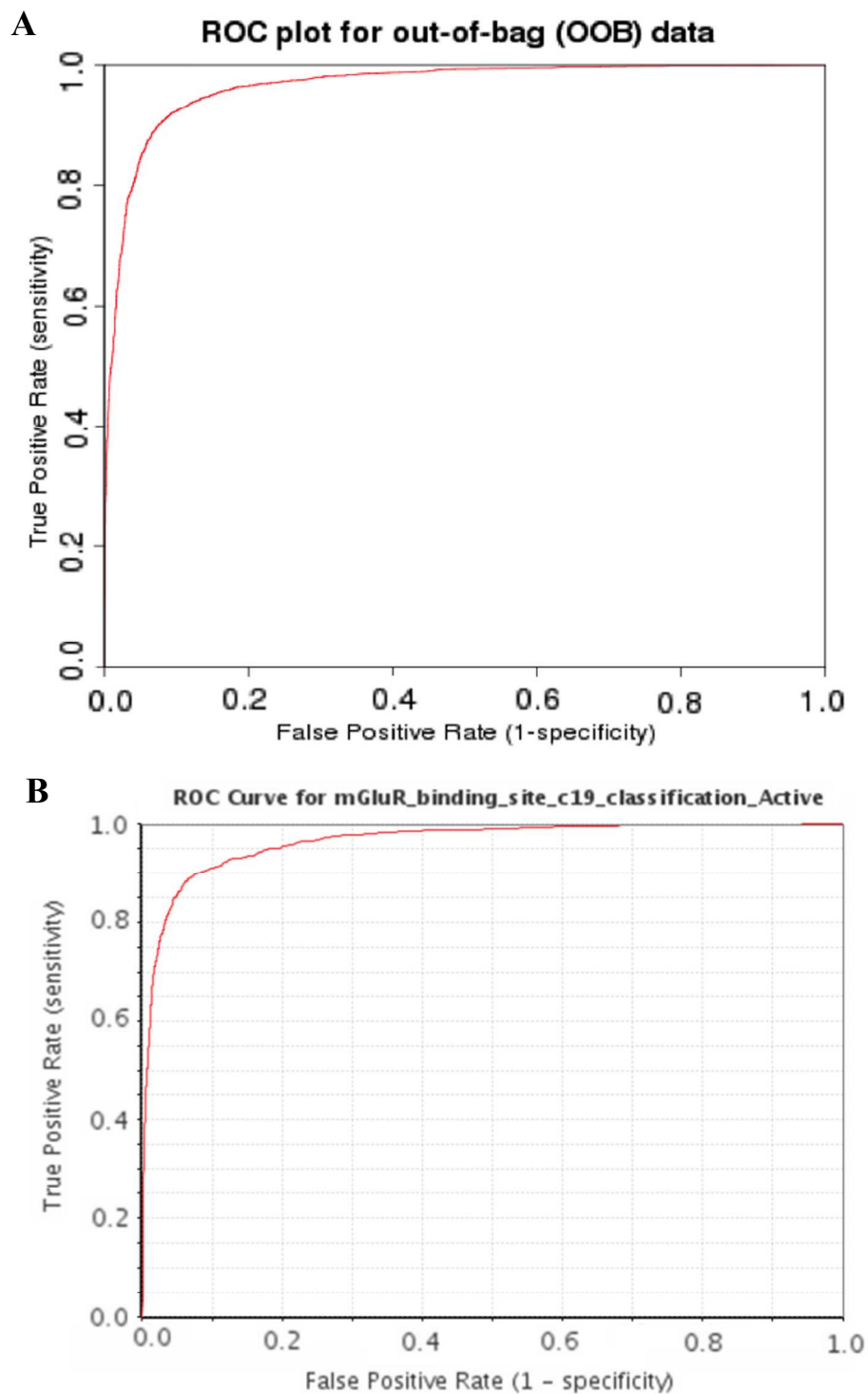


Figure S4: ROC curve for out-of-bag cross validation (A) and external validation (B). The ROC curve was generated on the data from table S3 and represents one of the 5 models used in ensemble modeling of the final predictions.

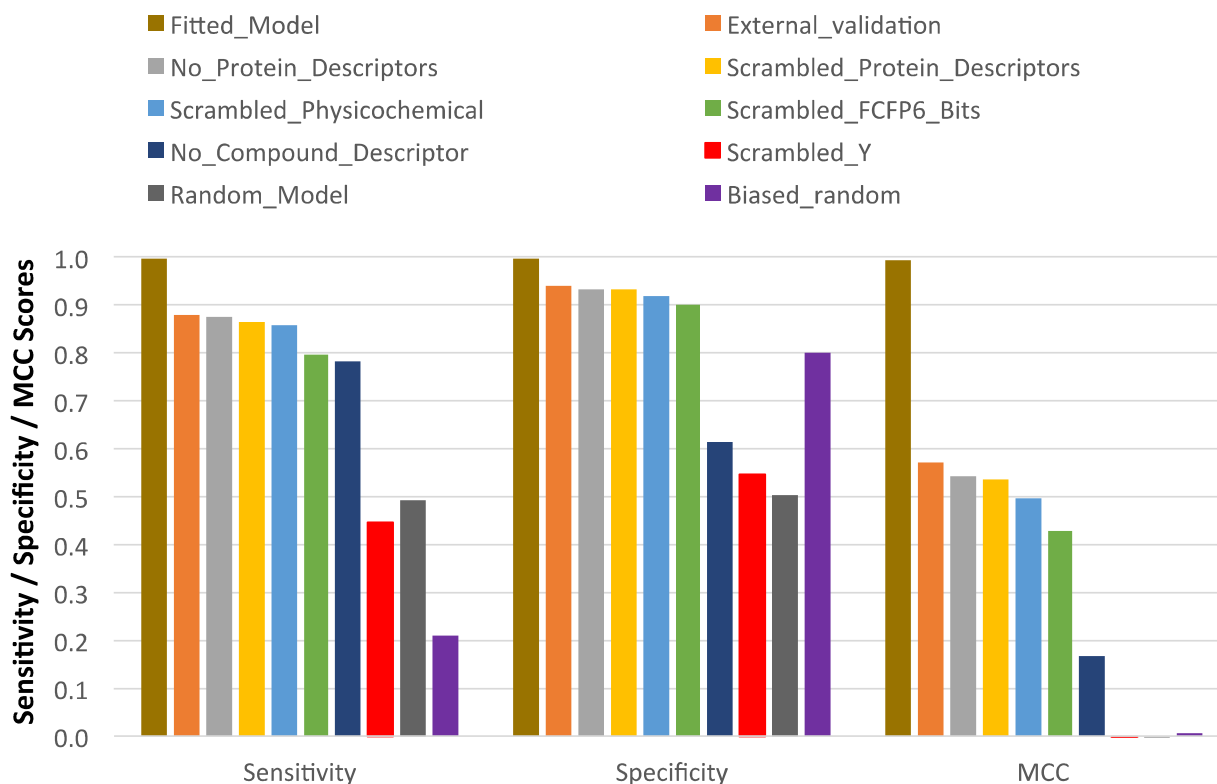


Figure S5. Model quality in different validation experiments. Shown are model quality for true data: fitted model (brown), externally validated (orange). Model quality for scrambled data: no protein descriptor (grey) protein descriptor scrambled (yellow), compounds physicochemical descriptors scrambled (light blue), compound FCFP_6 bits scrambled (green), no compound descriptor (dark blue), Y-scrambled (where the output variable (activity) has been randomized and coupled to input vectors with true descriptors, red), a random model (dark grey), and a biased random model (purple). As all scrambled experiments deteriorate model quality compared to models build on the true data, it can be said that each part of the data adds information for the model. However, the influence of the compound descriptors is much larger than the influence of the protein descriptors. This can also be explained given that a larger part of the variance in the data set is located in the compound descriptors (1000s) as opposed to the relatively low number of proteins (17).

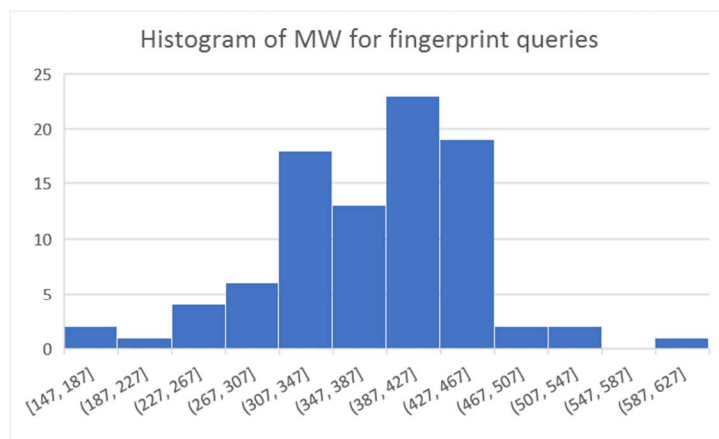
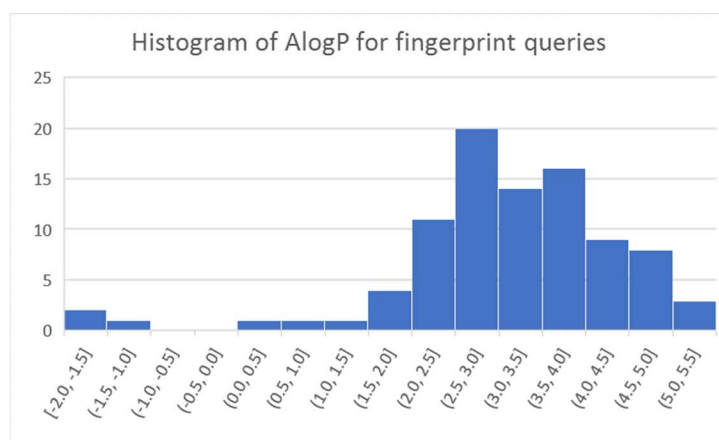
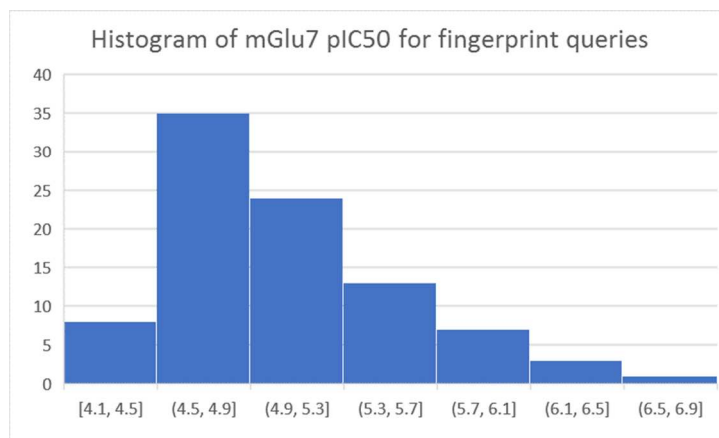
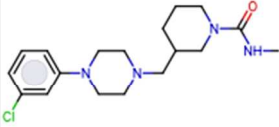
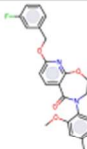
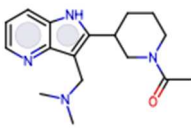


Figure S6. Properties of queries used for fingerprint analogue searches. Histogram of mGlu₇ pIC₅₀ showing the skew to low activity, also histograms of calculated AlogP and MW.

JNJ Structure	Method
	PCM
	PCM
	PCM

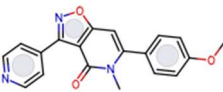
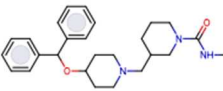
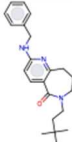
tautomeric_smiles	Receptor
	grm7_human
	grm7_human
	grm7_human

Figure S7. Comparing PCM highly ranked false positives to true positives. Left: Example of top ranking hits from the prospective virtual screen PCM model that later turned out inactive (false positives). Right: Example of known mGlu₇ receptor actives (true positives).