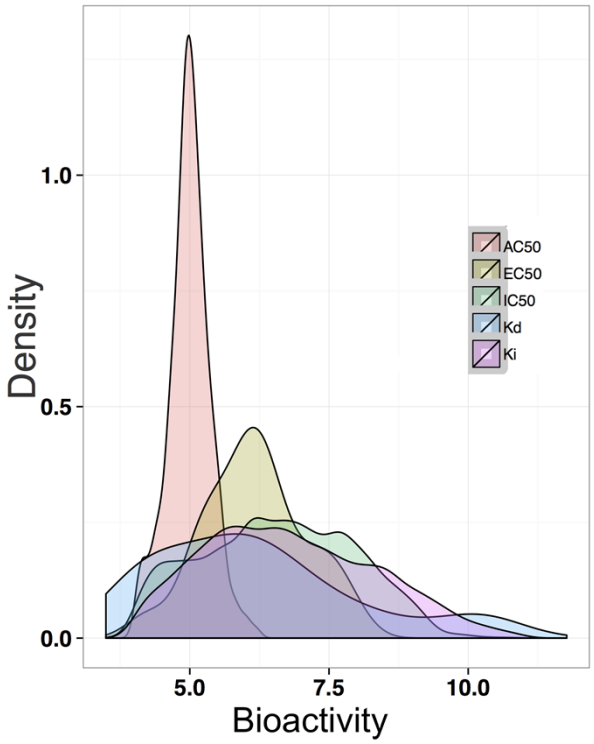
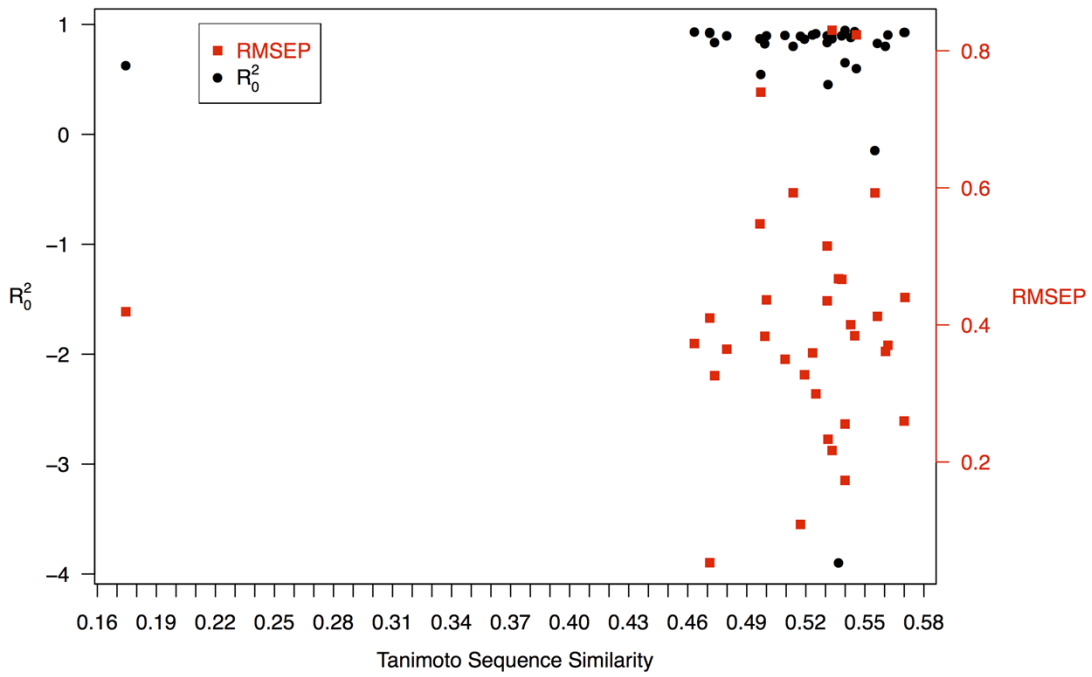
Electronic Supplementary Material (ESI) for Integrative Biology.

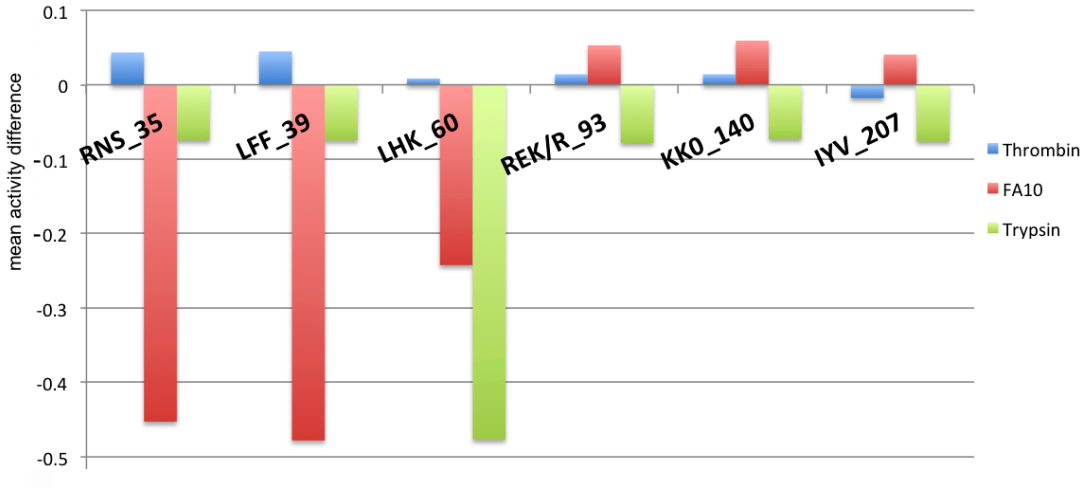
Supplementary Information



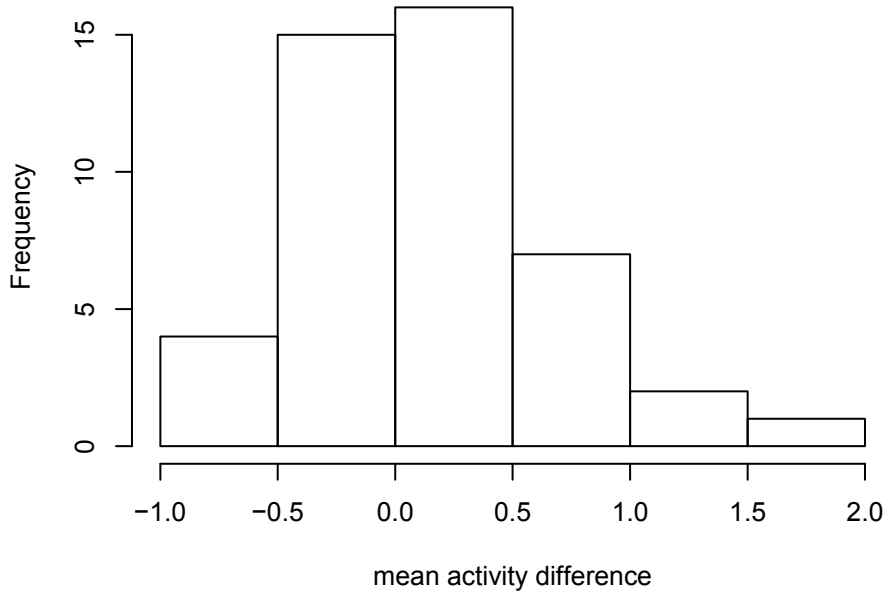
Supplementary Figure SF 1: Distribution of activity values in the serine protease dataset. A normal distribution is found for Ki and IC50 affinity values. The other three activity measures have less data points in the complete set and a narrow peak is observed.



Supplementary Figure SF 2: Applicability domain of target space of PCM model. All 67 targets are equally similar exhibiting an average similarity of 46%. Likely for this reason, we do not see an improvement of predictive correlation (R^2) or decrease in predicted error (RMSE) with increase in target Tanimoto similarity. Only two most diverse targets (C1r_human and FA10_dog) were observed to have a poor prediction ($R^2<0$; RMSE>0.5)



Supplementary Figure SF 3: The protease PCM model predicted amino acids at position 35, 39, 60, 93, 140, and 207 as the key residues impacting the binding activity of compounds. Positively charged arginine and histidine in FA10 and lysine in TRY at position 35 and 60 were predicted to decrease the binding affinity on average by 0.45, 0.25 and 0.48 log units. The amino acids arginine-35, leucine-39, arginine/lysine-93 and tyrosine-207 were predicted to be selectively affecting the binding activity against THRB, TRY and FA10 respectively.



Supplementary Figure SF 4: Distribution of mean activity difference of features employed in feature analysis. 244 structural features of compounds were employed for feature analysis. The distribution of mean activity difference calculated for each of these 244 features is shown here. The average of mean activity difference was calculated to be 0.21 with a standard deviation of 0.54.

Supplementary Table ST1: Number of compounds per target. All compounds were retrieved from the following databases: ChEMBL-17 database, Directory of Useful Decoys (DUD), ZINC and Binding DB. For all compounds, IC_{50}, Ki, $AC_{50,}$ EC_{50} and Kd were kept and combined into pChEMBL units. pChEMBL values range between 3.4 and 11.7. All these protease targets used in this study lie within a similarity range of 43% to 58%, however, there are significant differences in binding pockets of these drug targets which are visible in biological space plot (Figure 2).

Protein Target	**Paralogs**	**Orthologs**	**Number of datapoints**	**UniProt ID**	**PDB ID**
Thrombin (THRB)		THRB_Human, THRB_Mouse, THRB_Rat, THRB_Bovin	6719	P00734, P19221, P18292, P00735	1ype, 2ocv, 1a0h
Coagulation Factor (FA)	FA10_Human, FA11_Human, FA9_Human, FA7_Human, FA12_Human	FA10_Rat, FA10_Dog, FA10_Rabbit	6957	P00742, Q63209, P19540, O19045, P03951, P00740, P08709, P00748	3kcg, 1iod, 3kl6, 4bdw, 1xx9, 1o5d
Trypsin (TRY)	TRY1_Human, TRY2_Human, TRY3_Human	TRY1_Bovin, TRY2_Rat	2679	P07477, P00760, P07478, P00763, P35030	1fxy, 1h4w, 1amh, 1aq7, 1avw
Plasminogen (PLMN)		PLMN_Human, PLMN_Mouse,	731	P00747, P20918, Q01177, P06868	1ki0

		PLMN_Rat, PLMN_Bovin			
Cathepsin-G (CATG)	CATG_Human, GRAB_Human	CATG_Mouse, MCPT9_Mouse	201	P08311, P28293, P10144, O35164	1kyn
Chymotrypsin (CTRL)	CTRL_Human, CTRB1_Human	CTRA1_Bovin, CTRB_Bovin	303	P40313, P17538, P00766, P00767	3gch, 1cbw
Tryptase (TRYB)	TRYB1_Human, PRSS31_Human	TRYB_Rat, TRYB_Mouse, TRYB_Dog	172	Q15661, P27435, Q02844, P15944, Q9NRR2	1lto
Elastase (CEL)		CELA1_Human, CEL2A_Rat, CEL2A_Pig, CELA1_Pig	131	Q9UNI1, P00774, P08419, P00772	1okx, 1bru
Urokinase (UROK)		UROK_Human, UROK_Mouse, UROK_Rat, UROK_Pig	936	P00749, P06869, P29598, P04185	1c5w, 3laq
Kallikrein (KLK)	KLK1_Human, KLK2_Human, KLK3_Human, KLK4_Human, KLK5_Human, KLK6_Human, KLK7_Human, KLK8_Human, KLKB1_Human, PRTN3_Human		101	P06870, P20151, P07288, Q9Y5K2, Q9Y337, Q92876, P49862, O60259, P03952, P24158	2psx, 3vfe, 1spj, 3bsq, 4k1e, 2zch
Transmembrance Serine	TMPS2_Human,		828	O15393, Q9NRS4, Q8IU80, P00750,	1dfp, 1fuj, 3dfj,

Protease (TMPS)	TMPS4_Human, TMPS6_Human, TPA_Human, TM11D_Human, PROC_Human, ST14_Human, HEPS_Human, PRSS8_Human, TRYG1_Human, CFAD_Human			O60235, P04070, Q9Y5Y6, P05981, Q16651, Q9NRR2, P00746	1o5e, 3f6u, 2gv6
Chymase (CMA)	CMA1_Human		366	P23946	1pjp
Complement C1r (C1R)	C1R_Human		89	P00736	1apq

Supplementary Table ST2: Aligned binding site amino acids of 67 targets employed in the study.

Targets		35	39	40	41
Q15661_TRYB1_HUMAN	V	F	C	G	
P15944_TRYT_DOG	L	I	C	G	
P27435_TRYB1_RAT	V	F	C	G	
Q02844_TRYB1_MOUSE	A	F	C	G	
Q9NRR2_TRYG1_HUMAN	L	V	C	G	
Q8IU80_TMPS6_HUMAN	V	I	C	G	
Q9Y5Y6_ST14_HUMAN	A	I	C	G	
O60235_TM11D_HUMAN	L	H	C	G	
O15393_TMPS2_HUMAN	V	V	C	G	

All the alignment dependent descriptors were calculated based on this information. The position of each amino acid is given in the headers of the file. A separate excel sheet has also been added as supplementary information.

Supplementary Table ST3. Leave One Target Out validation of 67 Serine Protease Targets. R^2 and RMSE of each target was calculated individually and averaged together as mean R^2 and mean RMSE in the main text for each sub-family.

Sr. No.	Target	R^2_{test}	$RMSE_{test}$ (log units)	Target Sub family
1.	O19045_FA10_RABIT	0.51	2.37	FA10
2.	P00740_FA9_HUMAN	0.22	1.03	
3.	P00742_FA10_HUMAN	0.13	1.67	
4.	P00743_FA10_BOVIN	0.53	1.83	
5.	P00748_FA12_HUMAN	0.38	0.81	

6.	P03951_FA11_HUMAN	0.34	1.03	
7.	P08709_FA7_HUMAN	0.02	1.43	
8.	Q63207_FA10_RAT	0.70	0.93	
9.	O77669_FA10_DOG	0.01	1	
10.	P00734_THRB_HUMAN	0.08	1.68	THRB
11.	P00735_THRB_BOVIN	0.23	1.27	
12.	P18292_THRB_RAT	0.63	1.25	
13.	P19221_THRB_MOUSE	0.90	0.60	
14.	P00749_UROK_HUMAN	0.20	1.05	UROK
15.	P04185_UROK_PIG	0.94	0.42	
16.	P06869_UROK_MOUSE	0.96	1.41	
17.	P29598_UROK_RAT	0.99	0.82	
18.	O60259_KLK8_HUMAN	0.40	0.18	KLK
19.	P03952_KLKB1_HUMAN	0.49	1.05	
20.	P06870_KLK1_HUMAN	0.34	1.01	

21.	P07288_KLK3_HUMAN	0.02	1.92	
22.	P20151_KLK2_HUMAN	0.99	0.90	
23.	P24158_PRTN3_HUMAN	0.71	0.36	
24.	P49862_KLK7_HUMAN	0.30	0.73	
25.	Q92876_KLK6_HUMAN	0.007	0.37	
26.	Q9Y337_KLK5_HUMAN	0.20	1.04	
27.	Q9Y5K2_KLK4_HUMAN	0.50	0.27	
28.	P00760_TRY1_BOVIN	0.46	1.01	TRY
29.	P00761_TRYP_PIG	0.58	0.75	
30.	P00763_TRY2_RAT	0.33	0.53	
31.	P07477_TRY1_HUMAN	0.29	0.89	
32.	P07478_TRY2_HUMAN	0.39	1.09	
33.	P35030_TRY3_HUMAN	0.80	1.45	
34.	Q29463_TRY2_BOVIN	0.40	0.72	
35.	P00747_PLMN_HUMAN	0.34	0.94	PLMN

36.	P06868_PLMN_BOVIN	0.69	0.30	
37.	P20918_PLMN_MOUSE	0.98	0.57	
38.	Q01177_PLMN_RAT	0.40	0.78	
39.	P15944_TRYT_DOG	0.31	0.99	TRYB
40.	P27435_TRYB1_RAT	0.09	0.87	
41.	Q02844_TRYB1_MOUSE	0.22	1.26	
42.	Q15661_TRYB1_HUMAN	0.13	1.27	
43.	Q9NRR2_TRYG1_HUMAN	0.23	2.14	TMPS
44.	O15393_TMPS2_HUMAN	0.93	0.36	
45.	O60235_TM11D_HUMAN	0.2	0.7	
46.	P00750_TPA_HUMAN	0.05	1.09	
47.	P04070_PROC_HUMAN	0.40	1.24	
48.	Q16651_PRSS8_HUMAN	0.40	1.83	
49.	Q8IU80_TMPS6_HUMAN	0.80	0.91	

50.	Q9NRS4_TMPS4_HUMAN	0.20	0.56	
51.	Q9Y5Y6_ST14_HUMAN	0.20	1.5	
52.	P00746_CFAD_HUMAN	0.09	2.34	
53.	P05981_HEPS_HUMAN	0.1	1.54	
54.	P00772_CELA1_PIG	0.56	1.04	CELA
55.	P00774_CEL2A_RAT	0.97	0.60	
56.	P08419_CEL2A_PIG	0.01	1.30	
57.	Q9UNI1_CELA1_HUMAN	0.13	1.30	
58.	P00766_CTRA_BOVIN	0.46	0.99	CTRL
59.	P00767_CTRB_BOVIN	0.30	1.30	
60.	P17538_CTRB1_HUMAN	0.31	0.99	
61.	P40313_CTRL_HUMAN	0.31	2.2	
62.	O35164_MCPT9_MOUSE	0.99	0.49	CATG
63.	P00736_C1R_HUMAN	0.10	0.90	
64.	P08311_CATG_HUMAN	0.40	1.1	

65.	P10144_GRAB_HUMAN	0	0.001	
66.	P23946_CMA1_HUMAN	0.2	1.3	
67.	P28293_CATG_MOUSE	0.1	0.87	

The statistical correlations and error were defined as following:

$$R^2_{test} = 1 - \frac{\sum_{j=1}^{N}\left(y_j - \tilde{y}_j\right)^2}{\sum_{j=1}^{N}\left(y_j - \overline{y}_{test}\right)^2}$$

$$RMSE_{test} = \sqrt{\sum_{j=1}^{N}\left(y_j - \tilde{y}_j\right)^2 \Big/ N}$$

Where $N, y_j, \tilde{y}_j, \overline{y}_{test}$ represents the size of training set, the observed, the predicted and the mean of the response variables comprising the test set.[1,2]

Supplementary References

1. A. Golbraikh and A. Tropsha, *J. Mol. Graph. Model.*, 2002, **20**, 269–276.

2. A. Tropsha, P. Gramatica, and V. Gombar, *QSAR Comb. Sci.*, 2003, **22**, 69–77.