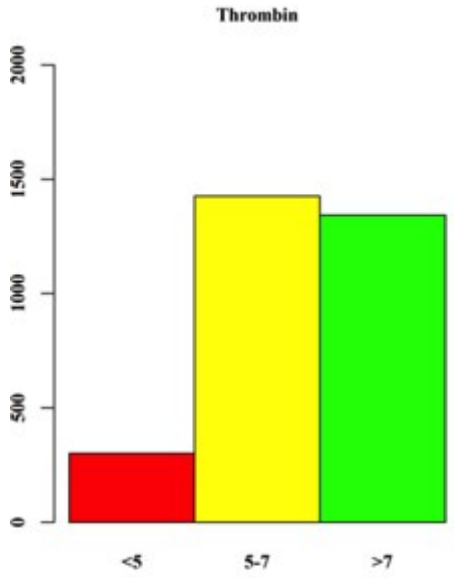
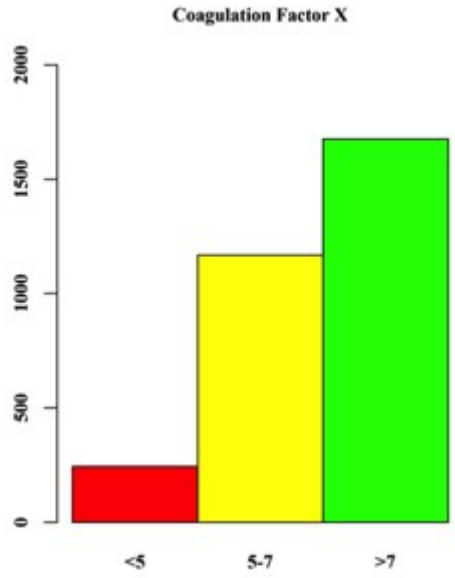
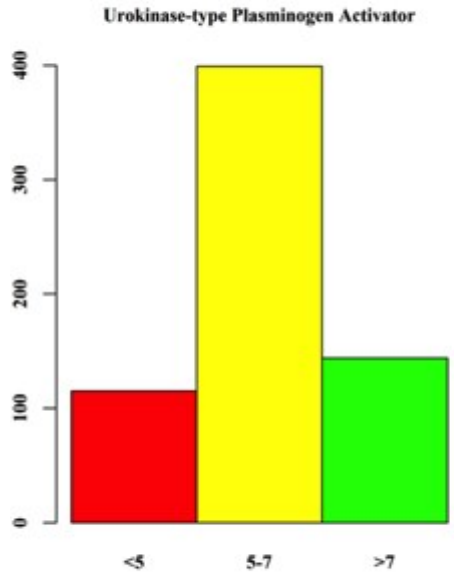
Electronic Supplementary Material (ESI) for MedChemComm.

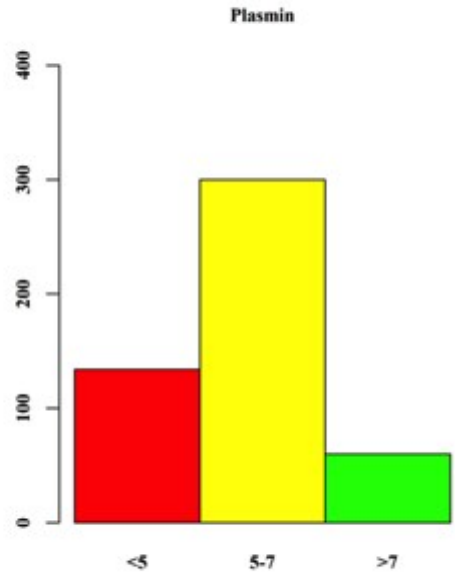
Table S1: Distribution of compounds corresponding to each target used in PCM modelling

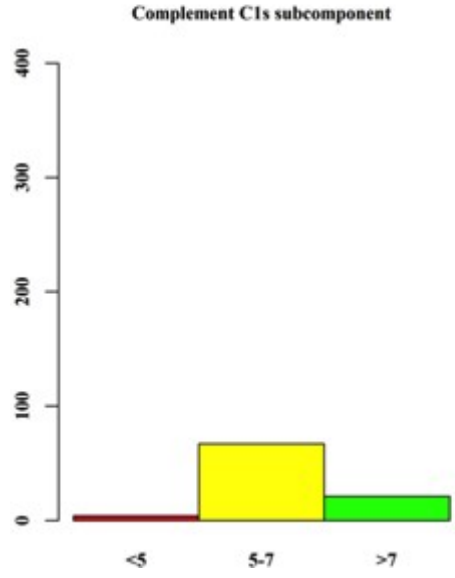
Target	Gene name	PDB	No. of compounds per target
Thrombin	F2	1QUR	2822
Coagulation Factor Xa	F10	2XBW	2779
Urokinase-type Plasminogen Activator	PLAU	1OWE	561
Plasminogen	PLG	1DDJ	415
Tissue-Type Plasminogen Activator	PLAT	1RTF	184
Suppressor of tumorigenicity 14 protein/Matriptase	ST14 / MT-SP1	1EAX	176
Coagulation Factor IXa	F9	1RFN	142
Alpha Tryptase	TPSAB1	1LTO	128
Beta-1 Tryptase	TPSAB1	2FPZ	114
Chymase	CMA1	3N7O	100
Complement C1s subcomponent	C1S	1ELV	91
Plasma Kallikrein	KLKB1	2ANY	81
Activated Protein C	PROC	1AUT	61
Coagulation Factor XIa	F11	3BG8	54
Kallikrein-1	KLK1	1SPJ	51
Kallikrein-3	KLK3	2ZCH	44
Kallikrein-7	KLK7	2QXI	27
Prostasin	PRSS8	3DFJ	23
Granzyme B	GZMB	1IAU	20
Coagulation Factor XIIa	F12	4XDE	17
Kallikrein-5	KLK5	2PSX	14
Trypsin-3	PRSS3	1H4W	2
Hepsin	HPN	1Z8G	1
Complement C1r subcomponent	C1R	1MD8	1

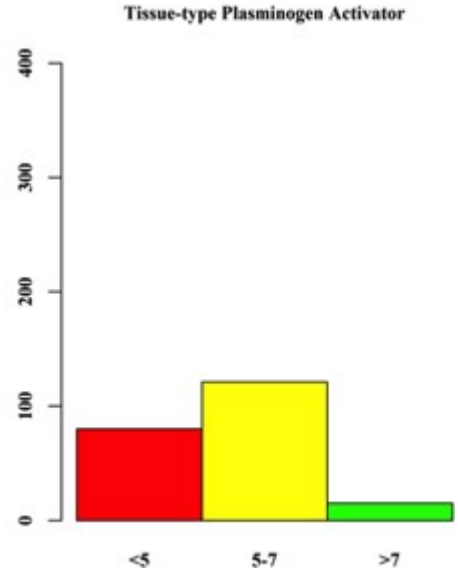

Thrombin
0
500
1000
1500
2000
<5
5-7
>7

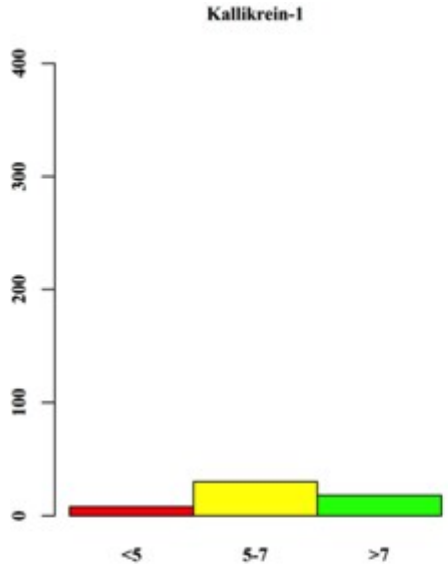

Coagulation Factor X
0
500
1000
1500
2000
<5
5-7
>7

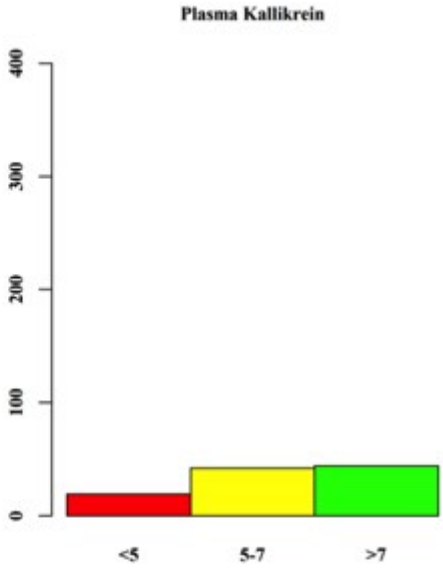

Urokinase-type Plasminogen Activator
0
100
200
300
400
<5
5-7
>7

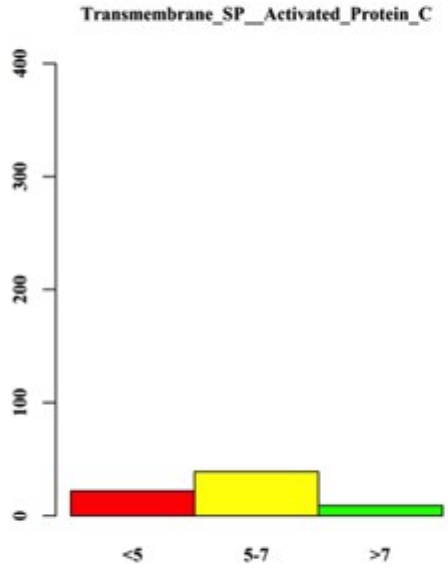

Plasmin
0
100
200
300
400
<5
5-7
>7

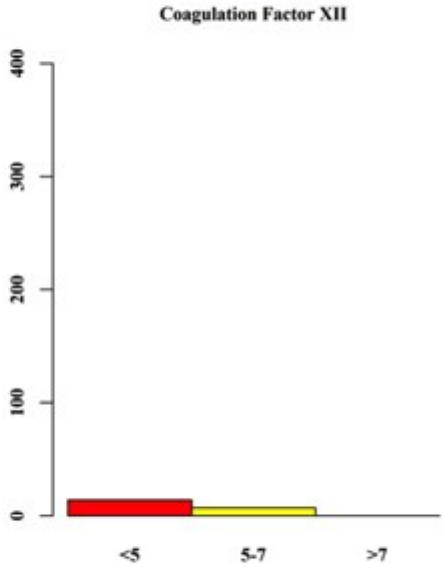

Complement C1s subcomponent
0
100
200
300
400
<5
5-7
>7

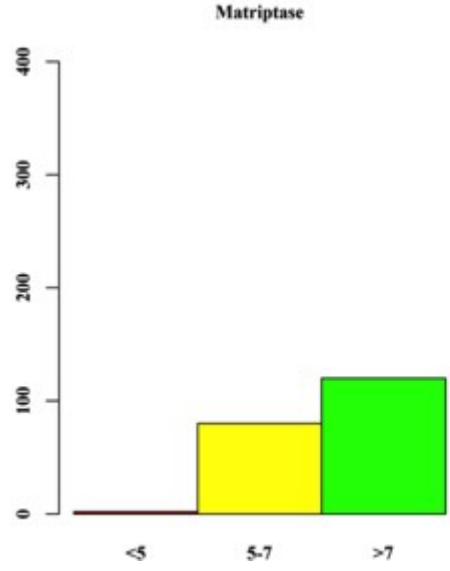

Tissue-type Plasminogen Activator
0
100
200
300
400
<5
5-7
>7

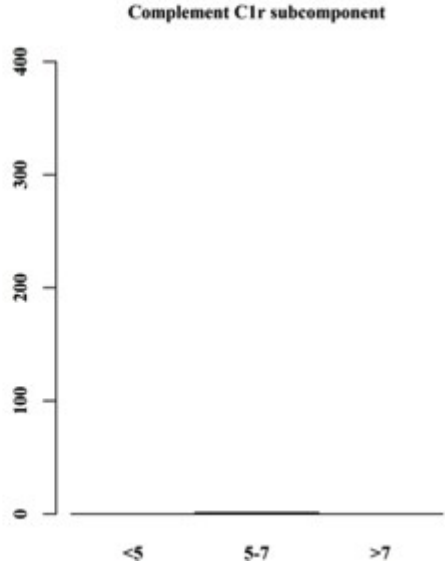

Kallikrein-1
400
300
200
100
0
<5
5-7
>7

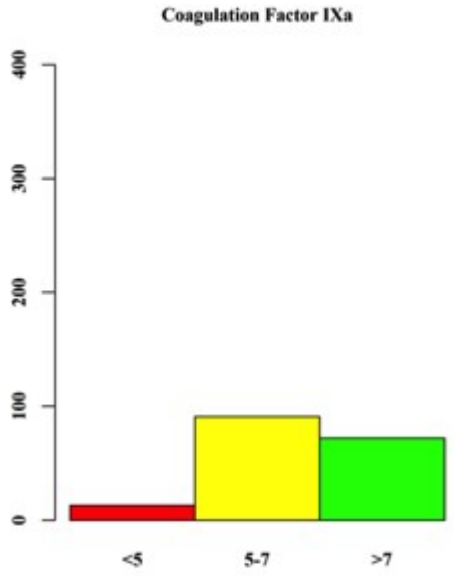

Plasma Kallikrein
400
300
200
100
0
<5
5-7
>7

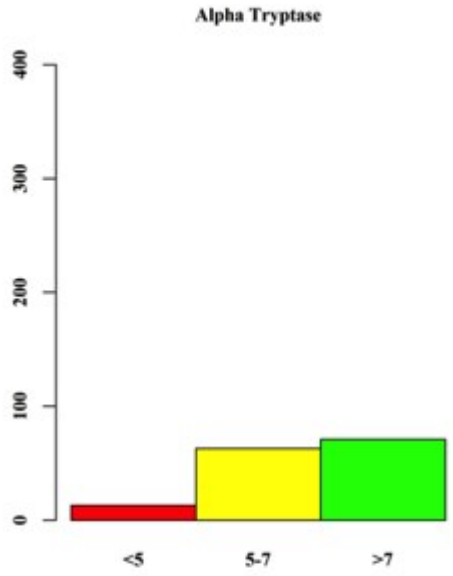

Transmembrane_SP__Activated_Protein_C
400
300
200
100
0
<5
5-7
>7

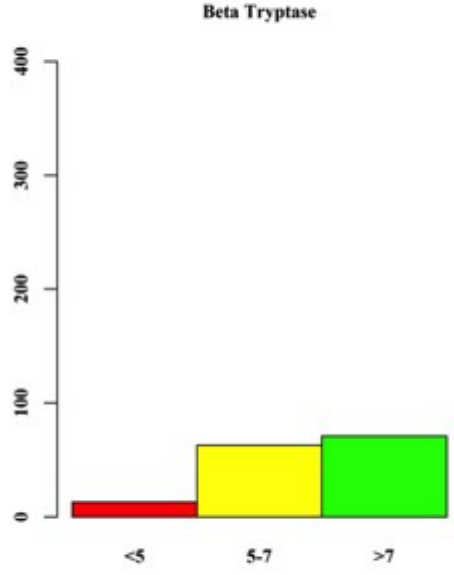

Coagulation Factor XII
400
300
200
100
0
<5
5-7
>7

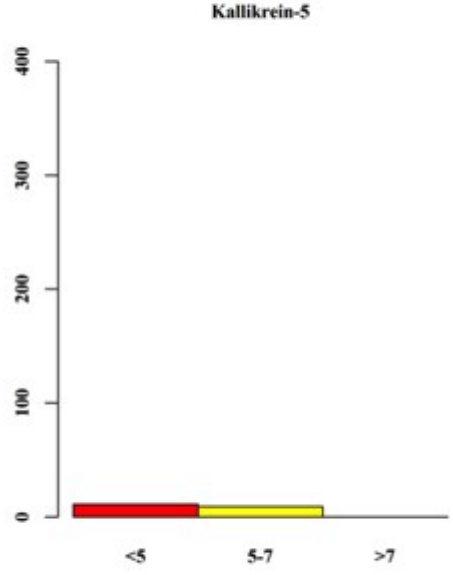

Matriptase
400
300
200
100
0
<5
5-7
>7

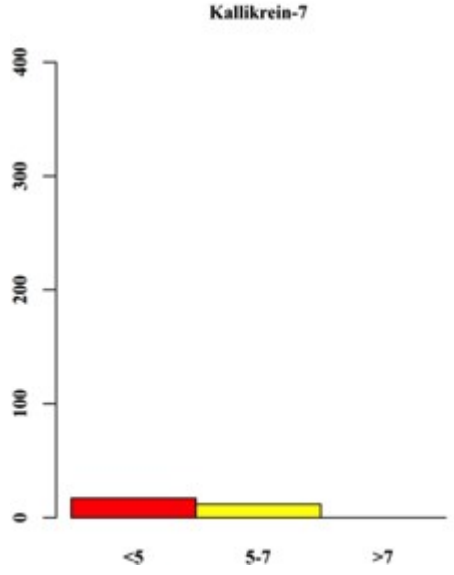

Complement C1r subcomponent
400
300
200
100
0
<5
5-7
>7

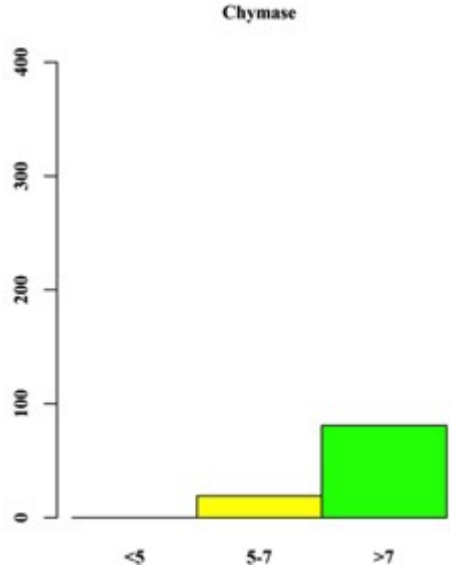

Coagulation Factor IXa
400
300
200
100
0
<5
5-7
>7


Alpha Tryptase
400
300
200
100
0
<5
5-7
>7


Beta Tryptase
400
300
200
100
0
<5
5-7
>7


Kallikrein-5
400
300
200
100
0
<5
5-7
>7


Kallikrein-7
400
300
200
100
0
<5
5-7
>7


Chymase
400
300
200
100
0
<5
5-7
>7

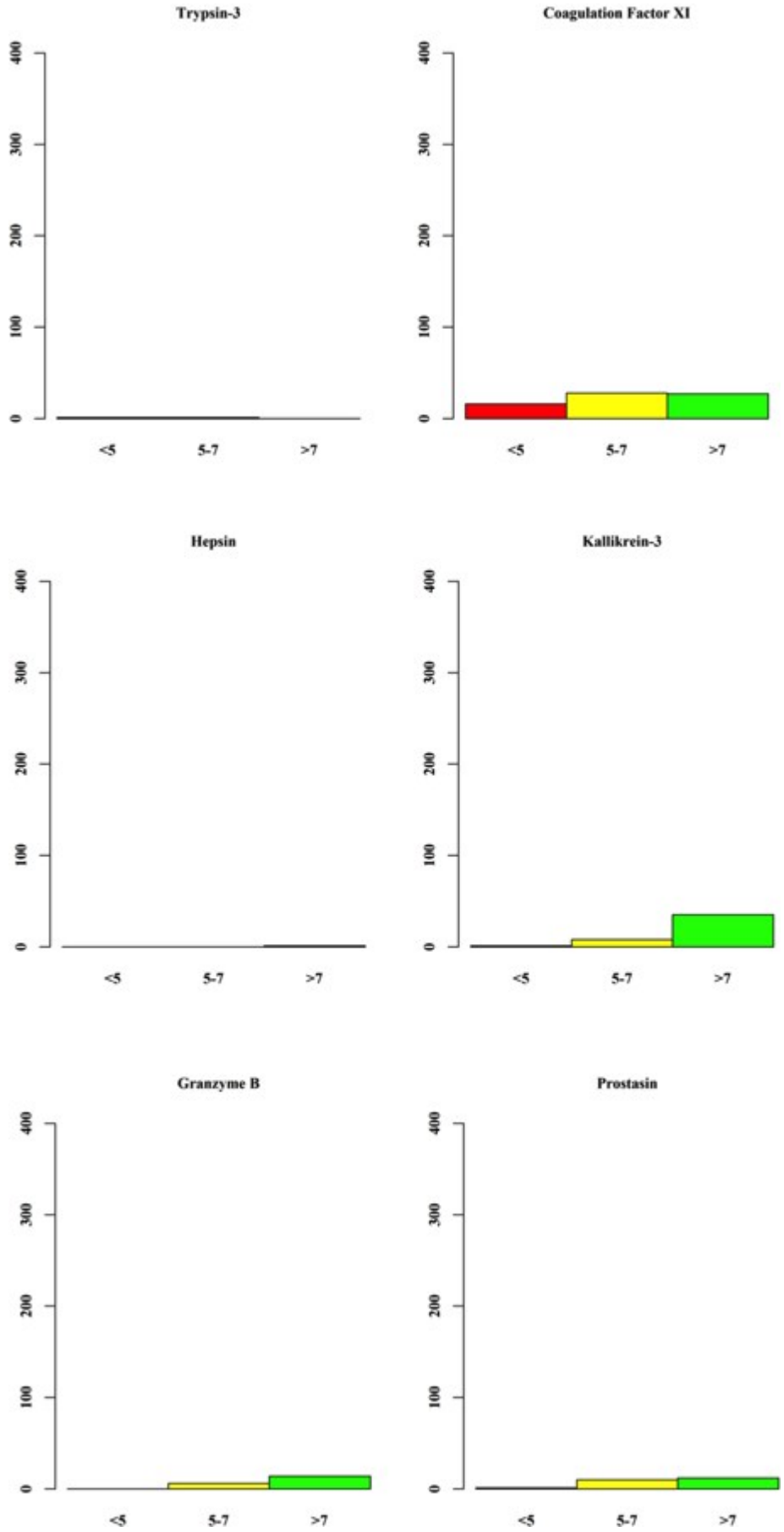


Figure S1. pK_i distributions of the 24 serine proteases

Table S2: Protein and ligand features used in PCM modelling after RFE

Protein descriptors	Ligand descriptors	No. of protein features	No. of ligand features	No. of cross-terms in PLS models
Field-based models				
Polar, lipophilic, stable and unstable water fields	RDkit	72	250	18000
Polar, lipophilic, stable and unstable water fields	MOE	90	160	14240
Polar, lipophilic, stable and unstable water fields	4-PFP	20	438	8760
Polar, lipophilic, stable and unstable water fields	Grind	10	94	940
Sequence-based models				
Amino acid and dipeptide composition	RDkit	15	167	-
Autocorrelation descriptors	RDkit	20	185	-
Composition, transition and distribution	RDkit	21	186	-
Sequence order and Pseudo amino acid composition	RDkit	15	148	-

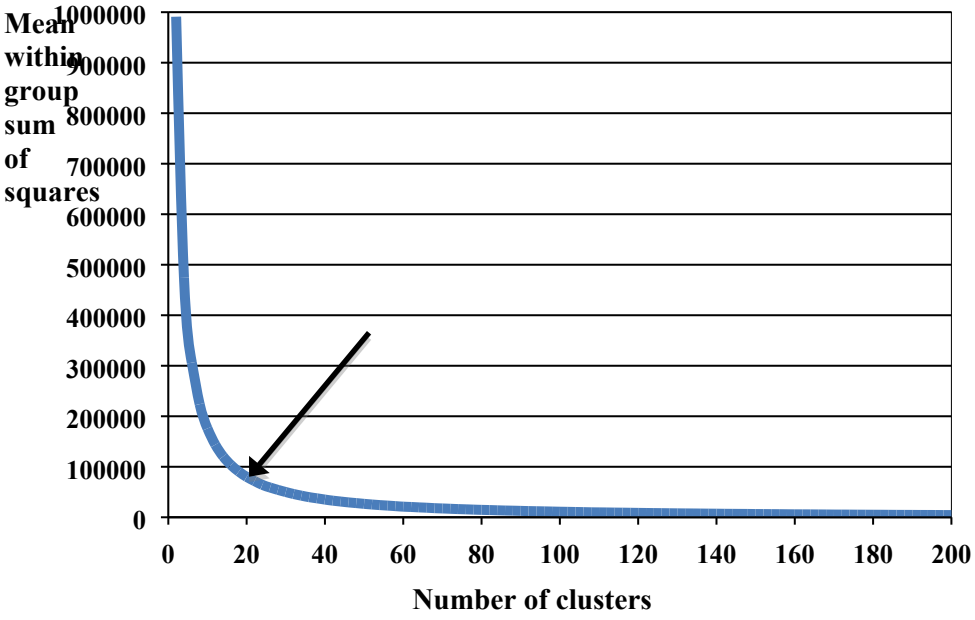


Figure S2. Plot of number of clusters against the mean within group sum of squares. Arrow pointer in the figure shows the optimal number of clusters chosen based on the Elbow method

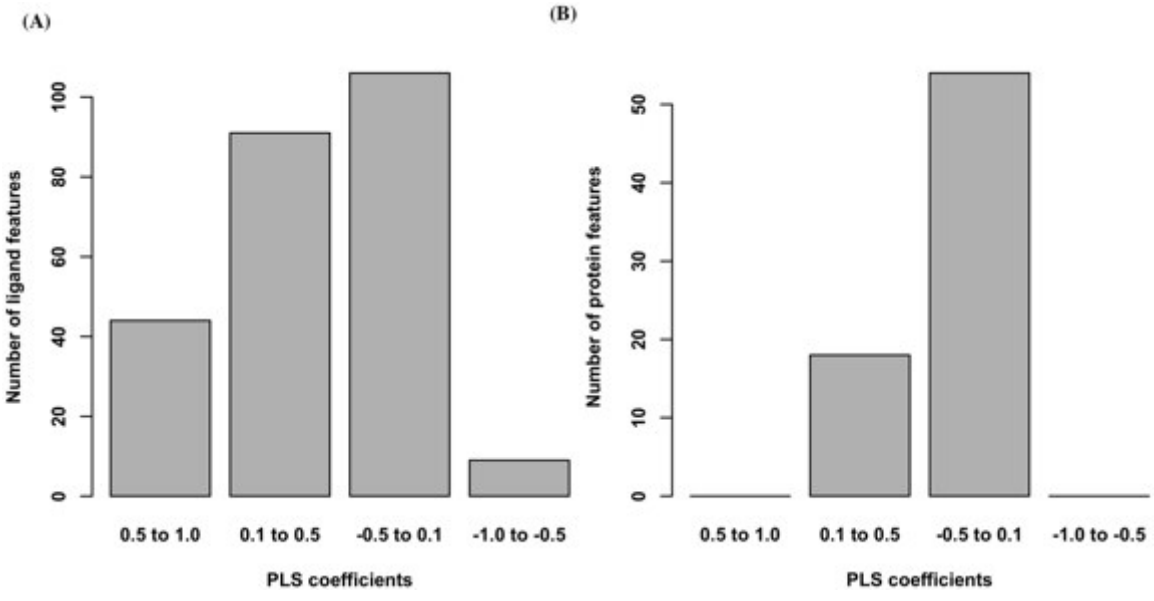


Figure S3: Distribution of the PLS coefficients of ligand (A) and protein features (B) used in PCM modelling

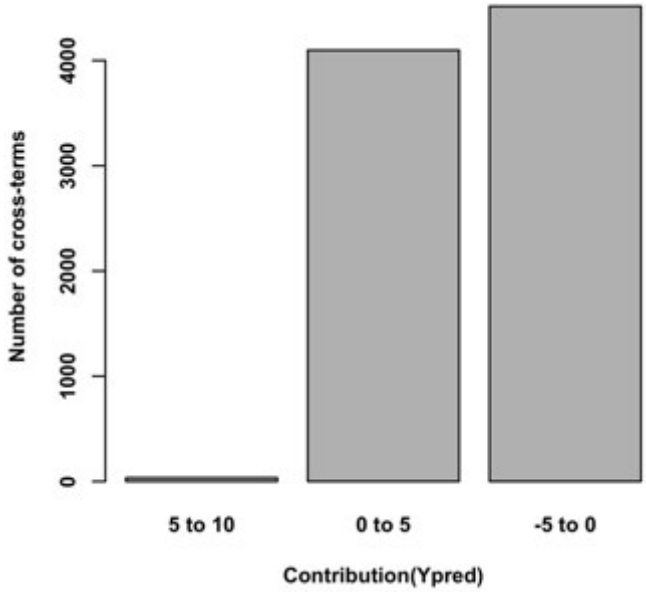


Figure S4: Distribution of the component contribution values of the cross-terms that have a positive effect on the interactions of CHEMBL73193 and coagulation factor 10

Table S3: Permutation validation results of RF and PLS models

Ligand descriptors	R^2	Q^2	R^2 intercept	Q^2 intercept	Random Permutation Models	
					Mean R^2	Mean Q^2
RF models						
RDkit	0.957	0.737	-0.094	-8.851	-0.095	-8.850
MOE	0.961	0.703	-0.116	-7.589	-0.116	-7.588
4-PFP	0.928	0.566	-0.089	-8.802	-0.089	-8.801
Grind	0.951	0.430	-0.127	-10.473	-0.127	-10.472
PLS models						
RDkit	0.671	0.588	0.162	-0.162	0.166	-0.153
MOE	0.504	0.433	0.056	-0.153	0.063	-0.141
4-PFP	0.554	0.451	0.163	-0.089	0.174	-0.092
Grind	0.311	0.264	0.033	-0.055	0.035	-0.050

Table S4. K_is of compounds CHEMBL73193 and CHEMBL315014 interacting with different serine proteases.

Serine Protease	K_i (nM)	
	CHEMBL73193	CHEMBL315014
F10	0.11	85
Thrombin	2000	>5000
Trypsin I (Bos Taurus)	280	1400

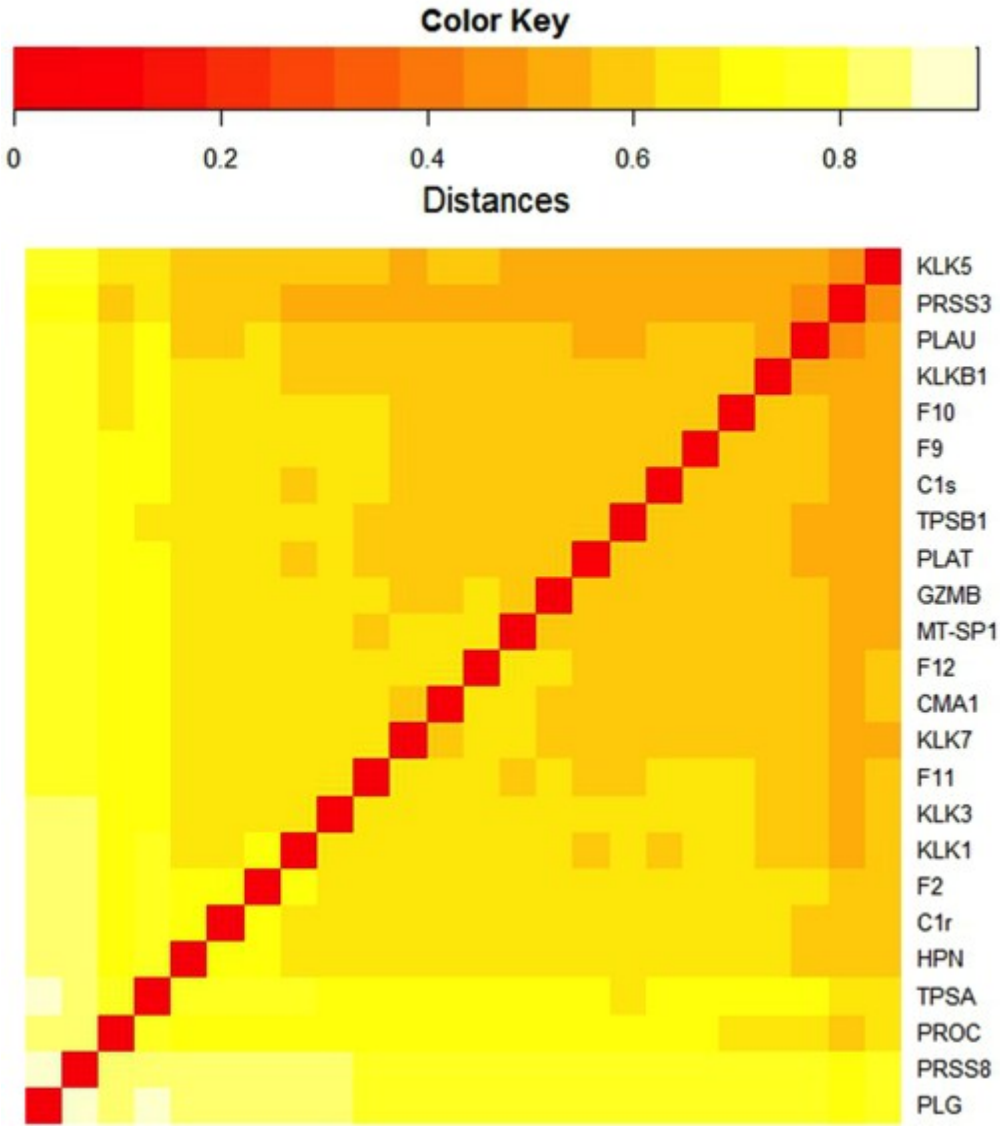


Figure S5. Distance-based heatmaps of 24 proteases based on their protein fields.

Table S5. Prediction performances of RDkit based Random Forest models after excluding one target at a time.

Excluded target	Correlation (R^2)	Predictability (Q^2)	RMSEE	$RMSEP_{test}$	R^2_{test}
Coagulation Factor XII	0.957	0.738	0.36	1.45	0.614
Kallikrein-5	0.957	0.739	0.35	1.16	0.549
Coagulation Factor XI	0.957	0.739	0.35	1.32	0.524
Prostasin	0.957	0.739	0.36	0.92	0.455
Kallikrein-7	0.957	0.737	0.36	1.3	0.452
Plasma Kallikrein	0.957	0.74	0.36	1.32	0.399
Kallikrein-1	0.957	0.738	0.36	0.9	0.34
Tissue-Type Plasminogen Activator	0.957	0.89	0.36	1.24	0.248
Activated protein C	0.957	0.738	0.35	1.09	0.226
Coagulation Factor IX	0.957	0.738	0.36	1.34	0.153
Alpha tryptase	0.958	0.742	0.35	1.31	0.144
Granzyme B	0.957	0.738	0.36	1.24	0.122
Plasmin	0.956	0.729	0.36	1.18	0.106
Complement C1s subcomponent	0.957	0.739	0.36	0.758	0.1
Beta-1 tryptase	0.958	0.74	0.35	1.33	0.068
Urokinase-type Plasminogen Activator	0.957	0.737	0.36	1.139	0.039
Thrombin	0.958	0.746	0.35	1.61	0.034
Coagulation Factor X	0.957	0.735	0.33	2.13	0.001
Kallikrein-3	0.957	0.739	0.36	1.56	0
Chymase	0.957	0.739	0.36	1.45	0
Matriptase	0.958	0.743	0.35	1.63	0
Complement C1r subcomponent	0.957	0.739	0.35	0.736	NA
Hepsin	0.957	0.739	0.35	2.65	NA
Trypsin-3	0.957	0.739	0.36	0.49	NA

NAs used for R^2_{test} of proteases that have only 1 or 2 observations in test set.

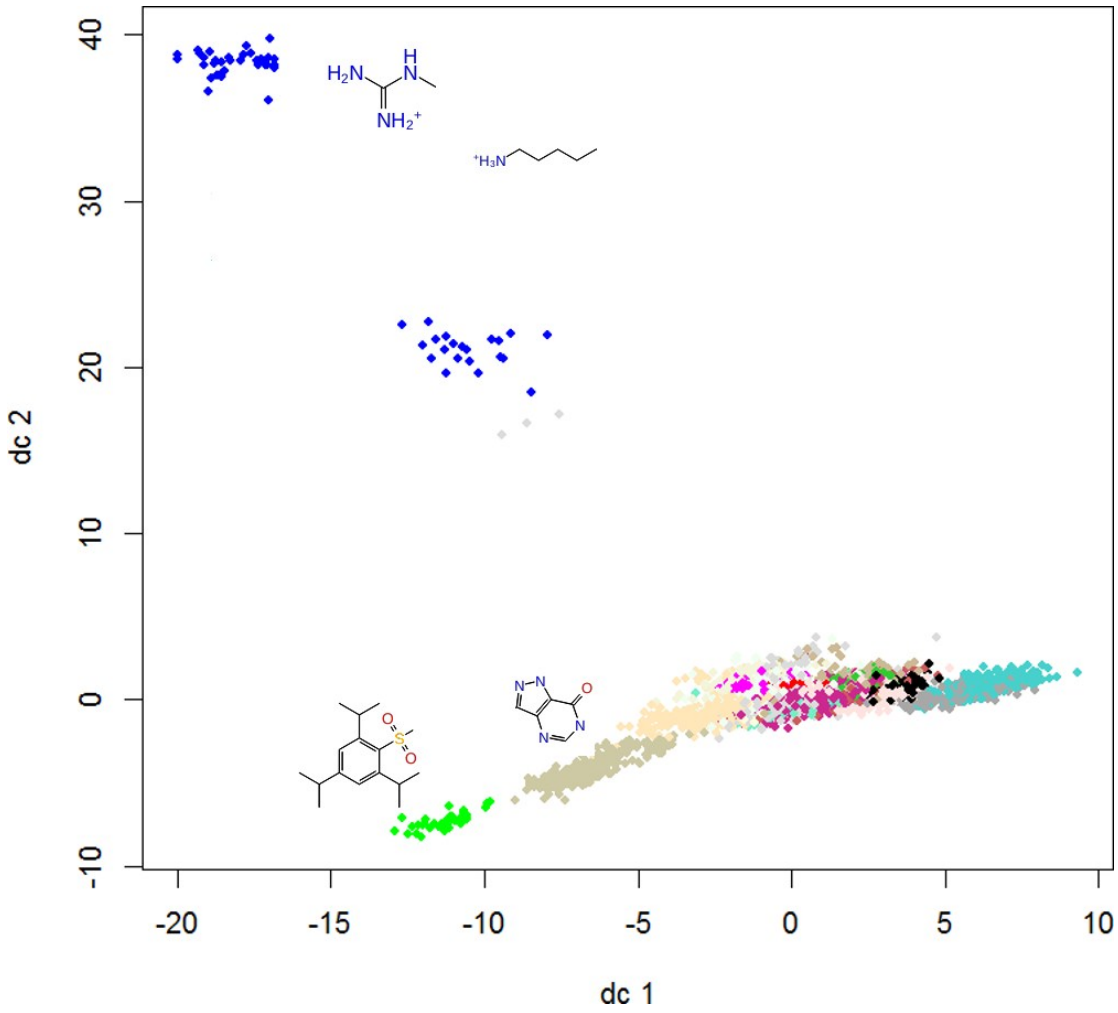


Figure S6. Distribution of the 20 compound clusters in 2 dimensions. Clusters shown here are obtained by performing k-means clustering based on the RDkit fingerprints of the compounds. Different colours represent the different clusters. C3, C4 and C17 refer to the cluster numbers and the scaffolds shown in these clusters represent the most common ones. dc1 and dc2 shown in labels correspond to discriminant coordinates.

Table S6. Prediction performances of RDkit based Random Forest models after excluding one compound cluster at a time.

Excluded compound cluster	No. of datapoints	Correlation (R^2)	Predictability (Q^2)	RMSEE	$RMSEP_{test}$	R^2_{test}
C9	113	0.498	0.263	1.32	1.08	0.017
C2	134	0.517	0.266	1.32	1.09	0.001
C10	77	0.511	0.265	1.32	1.33	0.036
C19	130	0.486	0.25	1.33	1.33	0.003
C3	112	0.497	0.261	1.32	1.35	0
C18	637	0.523	0.261	1.33	1.46	0.002

C5	426	0.494	0.255	1.33	1.47	0.013
C8	675	0.486	0.252	1.33	1.48	0.003
C16	1656	0.488	0.269	1.34	1.48	0
C7	522	0.496	0.253	1.33	1.54	0.004
C6	894	0.493	0.255	1.32	1.55	0.001
C20	289	0.517	0.274	1.31	1.61	0.009
C13	141	0.508	0.227	1.31	1.61	0.003
C1	360	0.493	0.256	1.32	1.63	0.006
C4	67	0.504	0.268	1.31	1.65	0.037
C12	253	0.505	0.265	1.31	1.67	0.003
C14	623	0.476	0.259	1.31	1.74	0.023
C15	232	0.514	0.268	1.31	1.74	0.001
C11	299	0.504	0.261	1.31	1.91	0
C17	268	0.505	0.268	1.29	2.29	0

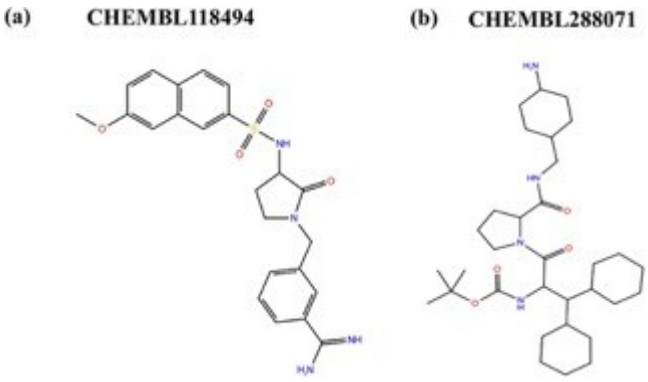


Figure S7. Chemical structures of the outlier compounds highlighted in Figure 4b