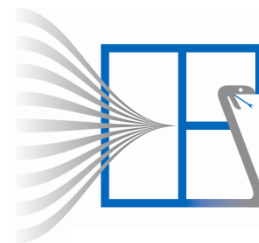


Real World Applications of Proteochemometric Modeling

*The Design of Enzyme Inhibitors and
Ligands of G-Protein Coupled Receptors*



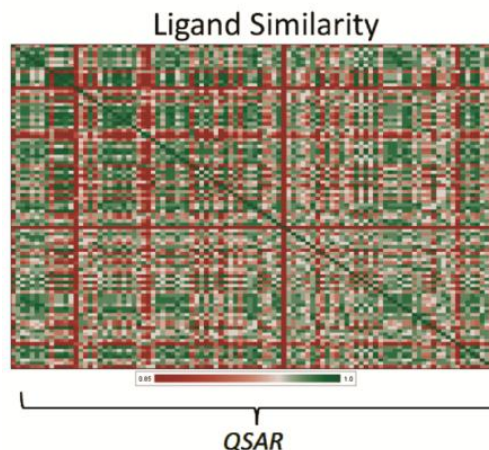
Leiden / Amsterdam
Center for Drug Research

Contents

- Our current approach to Proteochemometric Modeling
- Part I: PCM applied to non-nucleoside reverse transcriptase inhibitors and HIV mutants
- Part II: PCM applied to small molecules and the Adenosine receptors
- Conclusions

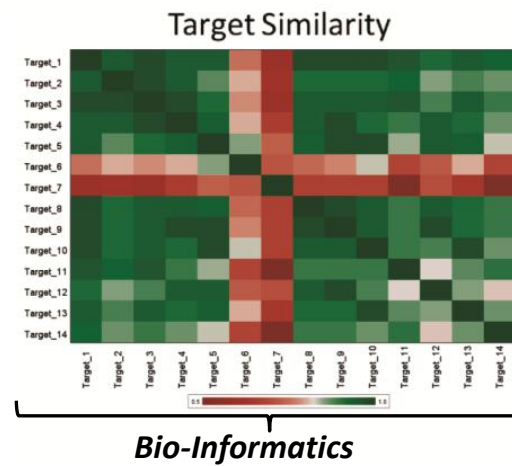
What is PCM ?

- Proteochemometric modeling needs both a ligand descriptor and a target descriptor
- Descriptors need to be compatible with each other and need to be compatible with machine learning technique...



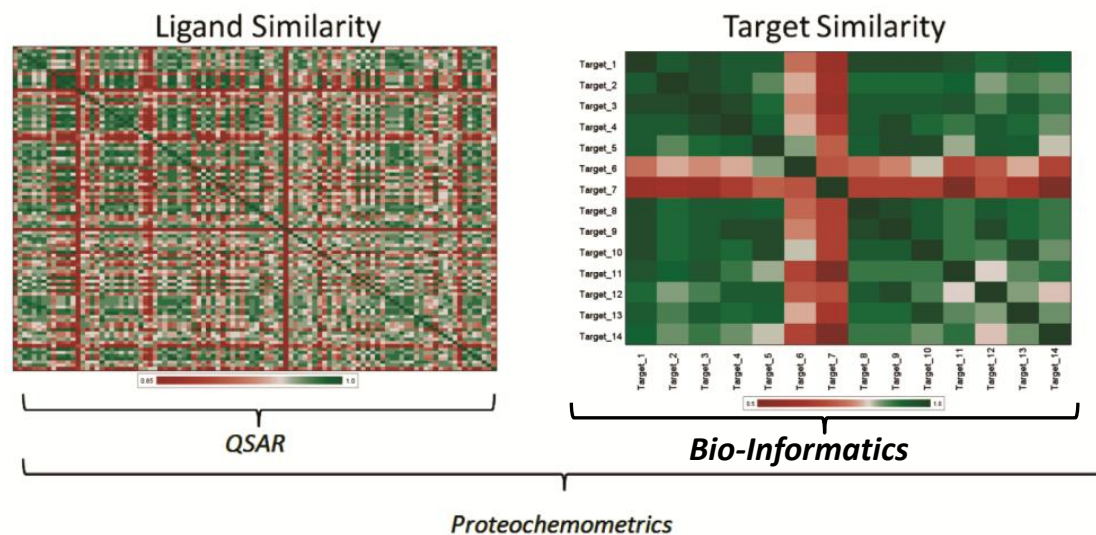
What is PCM ?

- Proteochemometric modeling needs both a ligand descriptor and a target descriptor
- Descriptors need to be compatible with each other and need to be compatible with machine learning technique...



What is PCM ?

- Proteochemometric modeling needs both a ligand descriptor and a target descriptor
- Descriptors need to be compatible with each other and need to be compatible with machine learning technique...



Ligand Descriptors

- Scitegic Circular Fingerprints
 - Circular, substructure based fingerprints
 - Maximal diameter of 3 bonds from central atom
 - Each substructure is converted to a molecular feature

FCFP_6
0
9
3
-415245925
-587569116
-1272798659
-1272709286
-1343180157
1070061035
136388789
-255848314
1686386090
-1742546106
-2095881698

Target Descriptors

- Select binding site residues from full protein sequence
- Each unique hashed feature represents one amino acid type (comparable with circular fingerprints)

ProtFP
1169372512
-590269326
268201585
268201585
268201585
-58134849
-58134849
-1481898440
558044215
-58134849
4227070002

Machine Learning

- Using R-statistics as integrated with Pipeline Pilot
 - Version 2.11.1 (64-bits)
- Sampled several machine learning techniques
 - SVM
 - Final method of choice
 - PLS
 - Random Forest

Real World Applications of PCM

- Part I: PCM of NNRTIs (analog series) on 14 mutants
 - Output variable: pEC_{50}
 - Data set provided by Tibotec
 - Prospectively validated
- Part II: PCM of small molecules on the Adenosine receptors
 - Output variable pKi
 - ChEMBL_04 / StarLite
 - Both human and rat data combined
 - Prospectively validated

Part I: PCM applied to NNRTIs

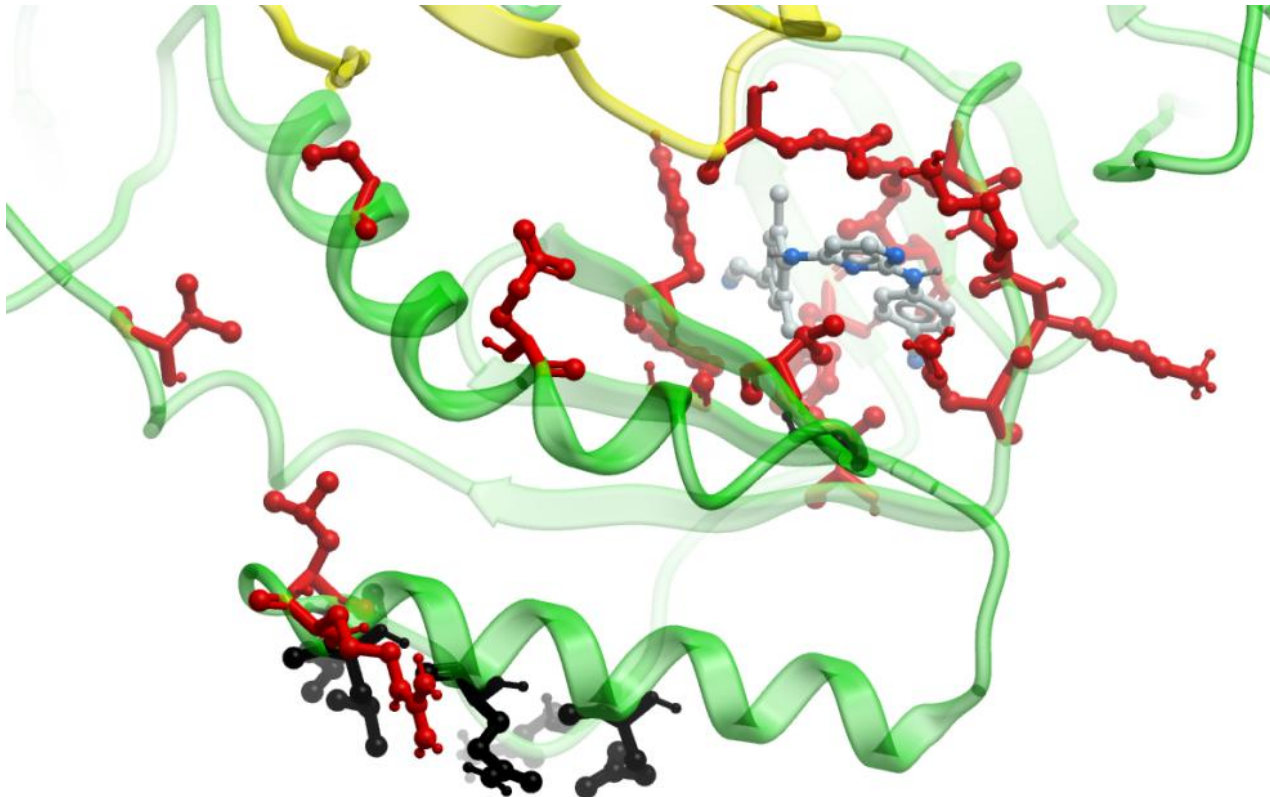
Which inhibitor(s) show(s) the best activity spectrum and can proceed in drug development?

- 451 HIV Reverse Transcriptase (RT) inhibitors
- 14 HIV RT sequences
 - Between zero and 13 point mutations (at NNRTI binding site)
 - Large differences in compound activity on different sequences

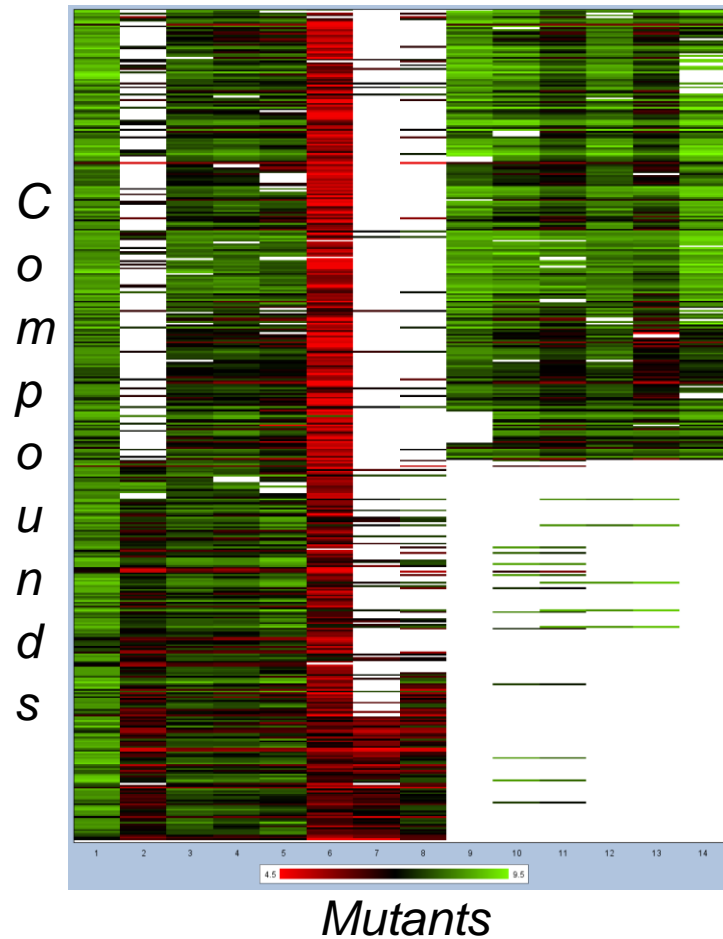
Sequence	Mean pEC ₅₀	StdDev pEC ₅₀	<i>n</i>
1 (wt)	8.3	0.6	451
2	6.9	0.7	259
3	7.6	0.6	444
4	7.5	0.7	443
5	7.4	0.8	429
6	6.0	0.6	316
7	6.5	0.6	99
8	6.9	0.7	147
9	8.3	0.6	222
10	7.9	0.7	252
11	7.5	0.7	257
12	8.0	0.6	242
13	7.4	0.8	244
14	8.2	0.8	220

Binding Site

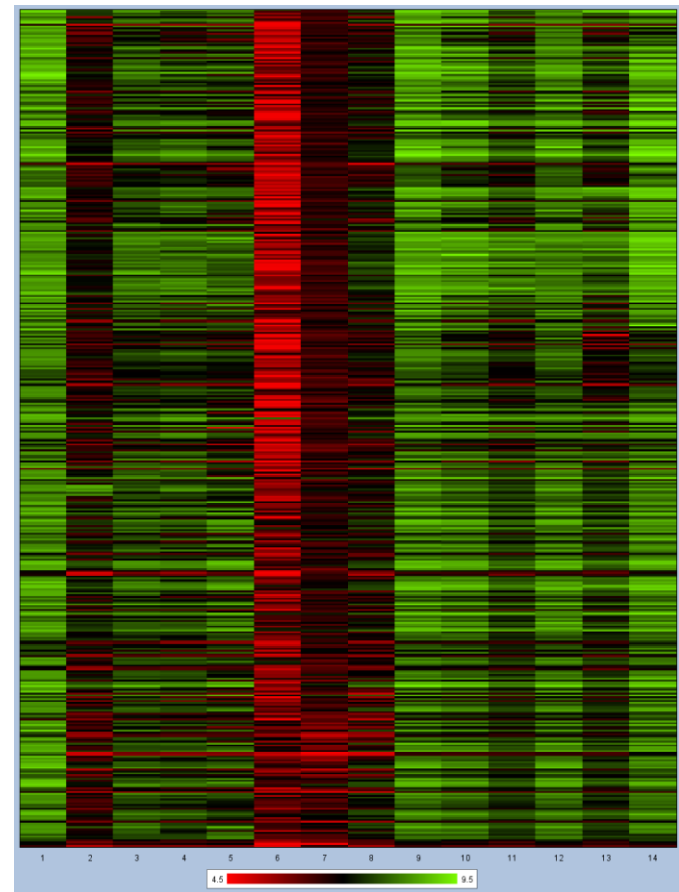
- Selected binding site based on point mutations present in the different strains
- 24 residues were selected



Used model to predict missing values



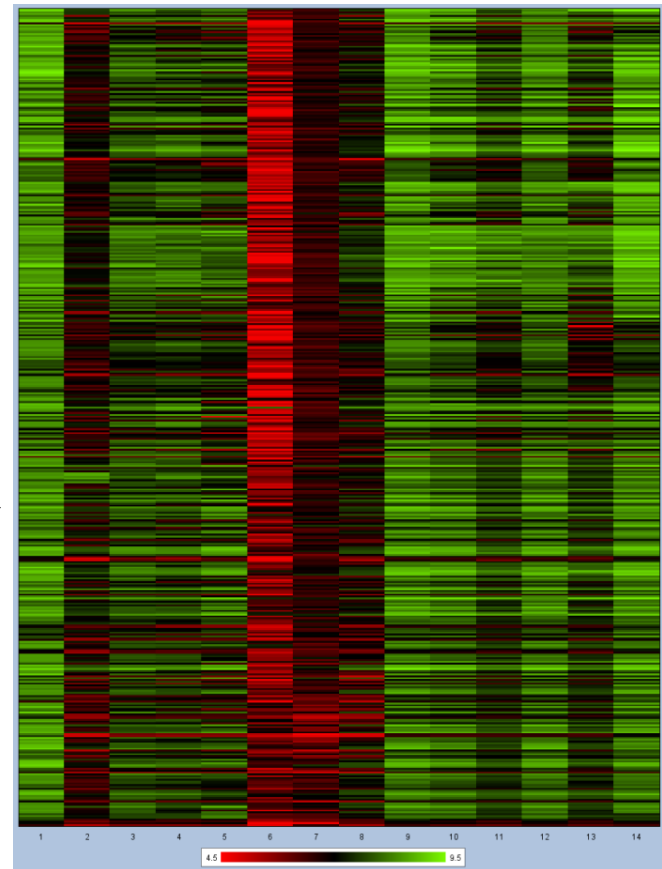
Original Dataset



Completed with model

Prospective Validation

- Compounds have been experimentally validated
 - Predictions where pEC_{50} differs two sd from compound average
 - (69 compound outliers)
 - Predictions where pEC_{50} differs two sd from sequence average
 - (61 sequence outliers)
- Assay validation

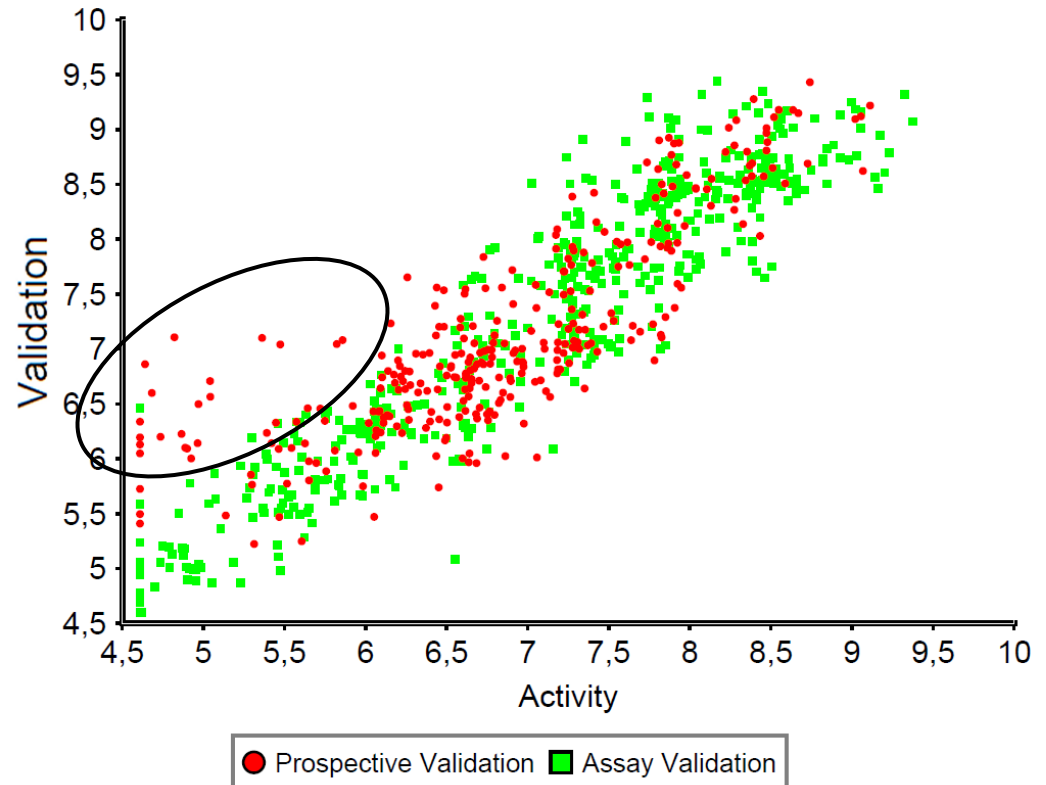


Completed with model



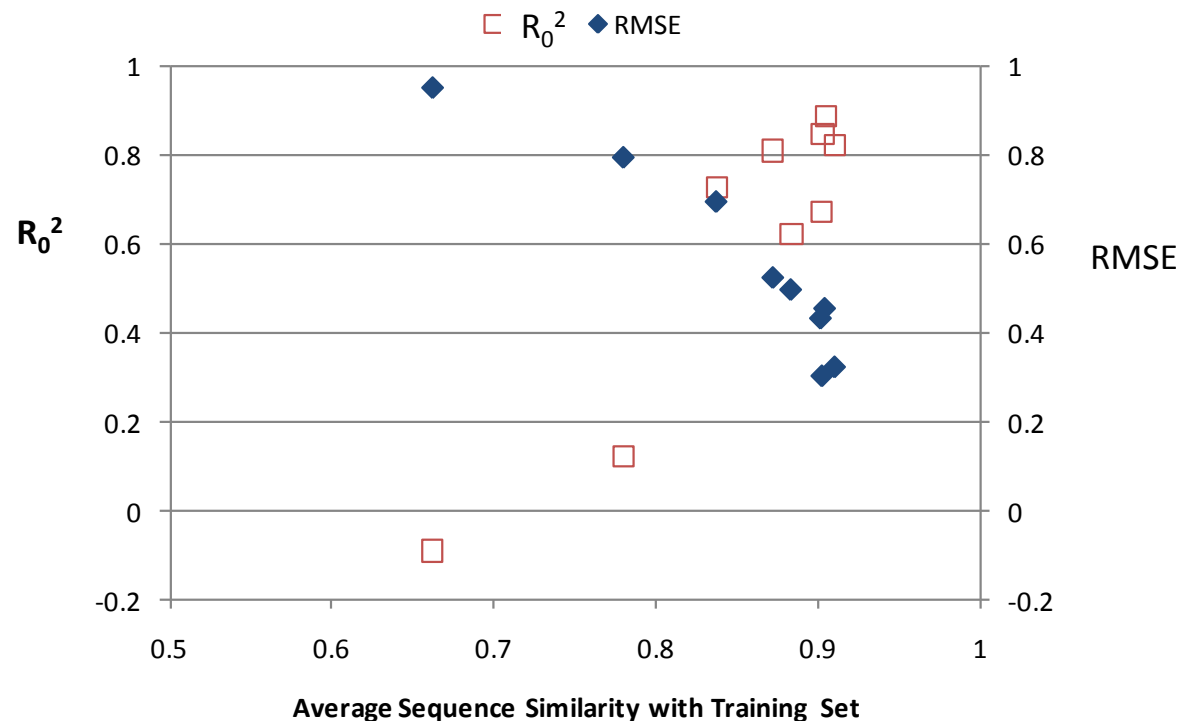
Prospective Validation

- Model:
 - $R_0^2 = 0.69$
 - RMSE = 0.62 log units
- Assay Validation
 - $R_0^2 = 0.88$
 - RMSE = 0.50 log units



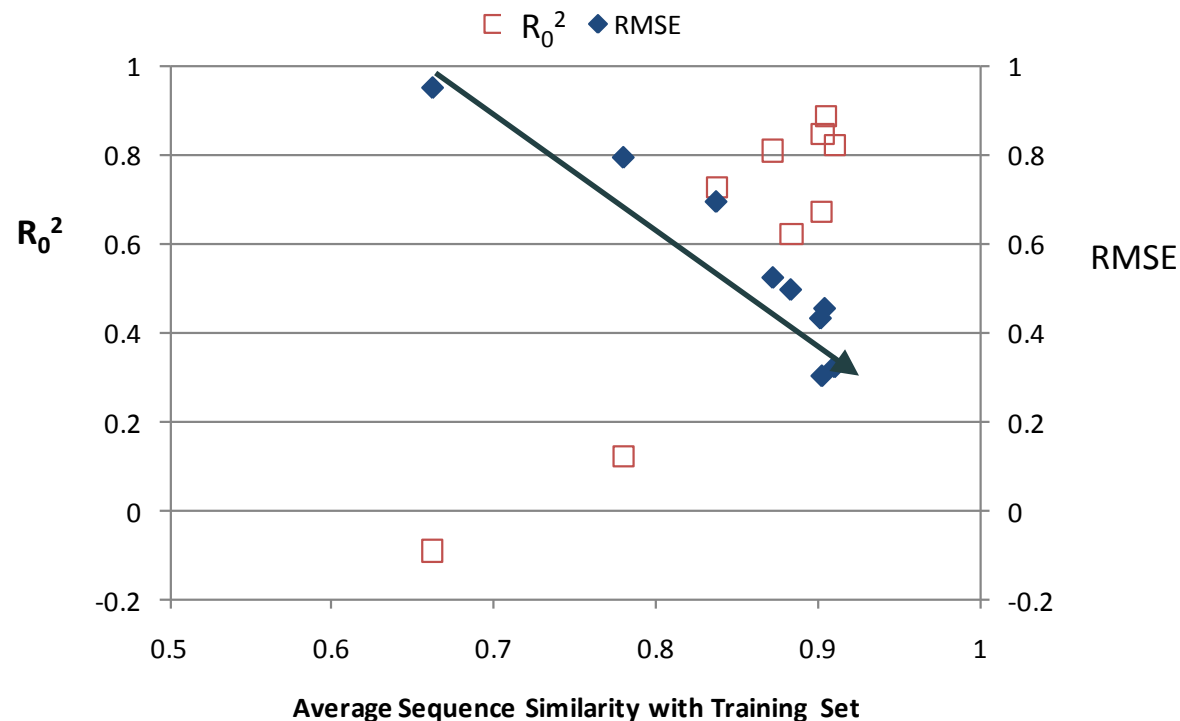
The Applicability Domain Concept Still Holds in Target Space

- Prediction error similarity shows a direct correlation with average sequence similarity to training set



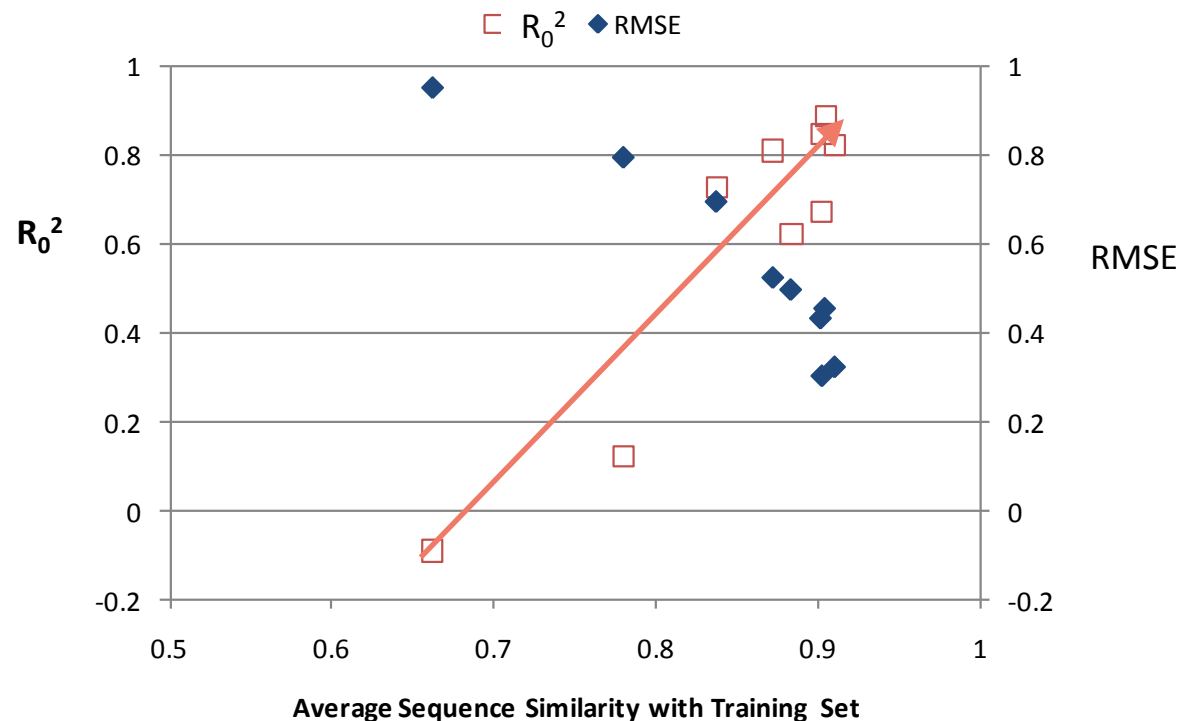
The Applicability Domain Concept Still Holds in Target Space

- Prediction error similarity shows a direct correlation with average sequence similarity to training set



The Applicability Domain Concept Still Holds in Target Space

- Prediction error similarity shows a direct correlation with average sequence similarity to training set



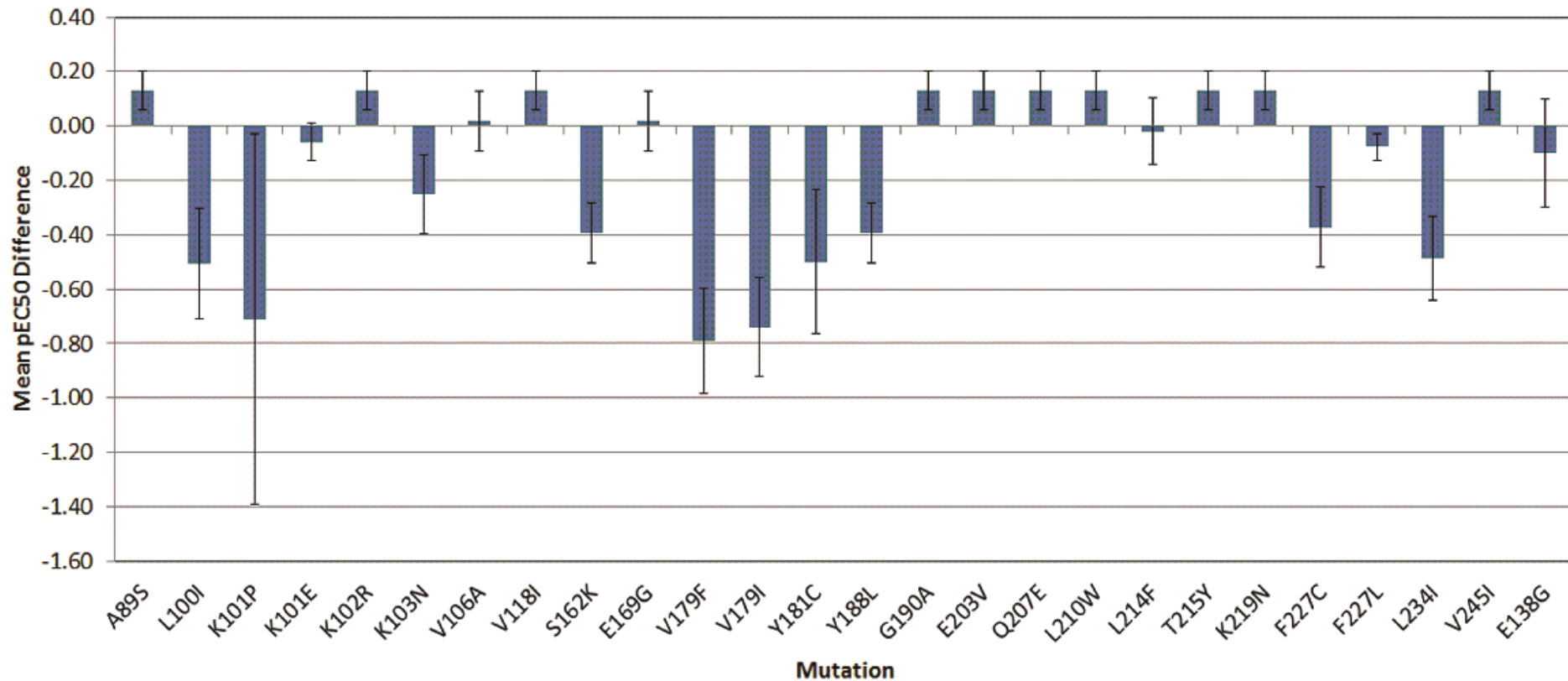
Does PCM outperform scaling and QSAR?

- PCM outperforms QSAR models trained with identical descriptors on the same set
- When considering outliers, PCM outperforms scaling
- PCM can be applied to previously unseen mutants

Validation Experiment	Assay	PCM	pEC ₅₀ scaling	QSAR	10-NN (both)	10-NN (target)	10-NN (cmpd)
R ₀ ² (Full plot)	0.88	0.69	0.69	0.31	0.41	0.21	0.28
R ₀ ² (Outliers)	0.88	0.61	0.59	0.36	0.34	0.32	0.18
RMSE (Full plot)	0.50	0.62	0.57	0.96	0.90	1.29	1.16
RMSE (Outliers)	0.50	0.52	0.58	1.06	0.72	1.39	1.29

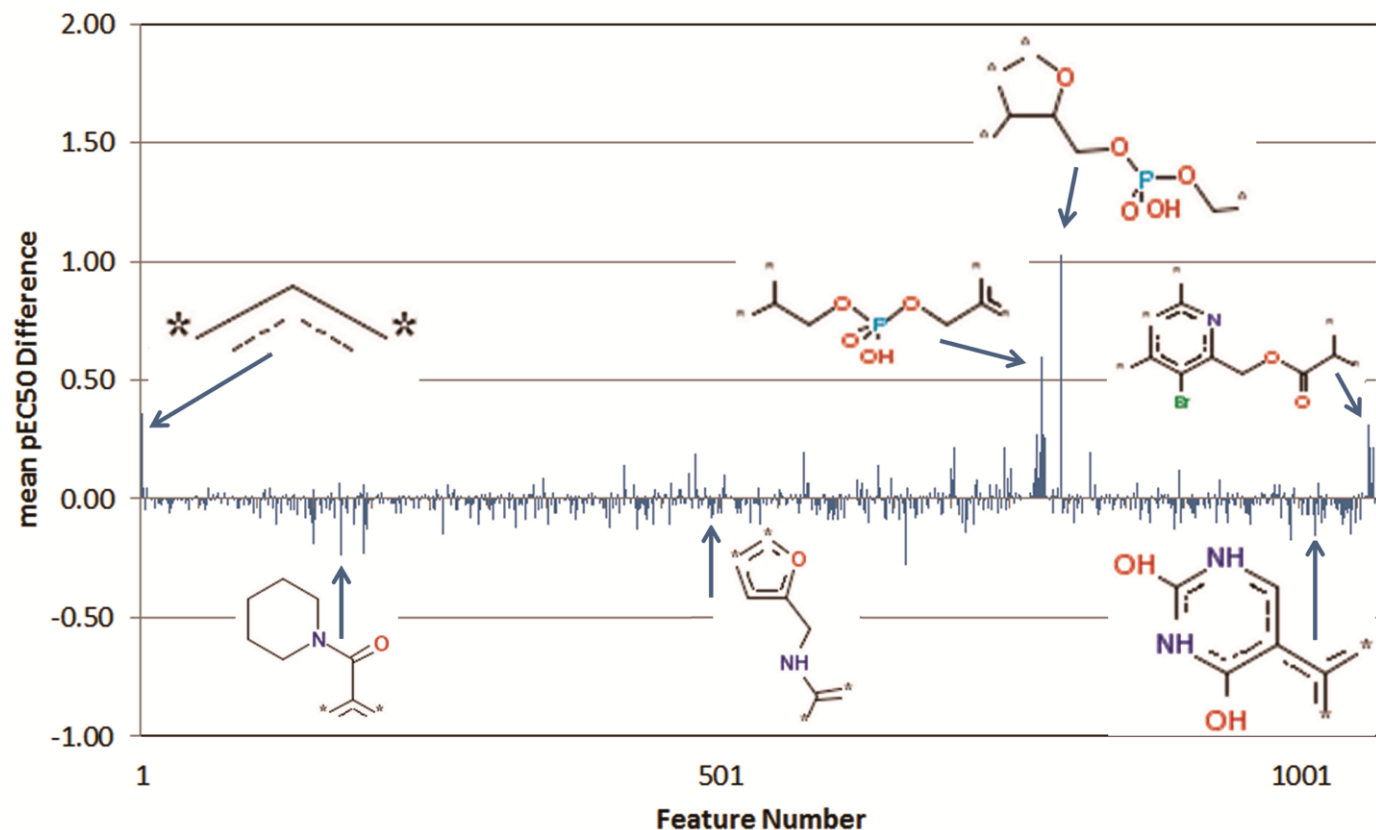
Model Interpretation (Sequences)

- Effect of mutation presence on compound pEC₅₀
- High impact mutations are K101P, V179I and V179F



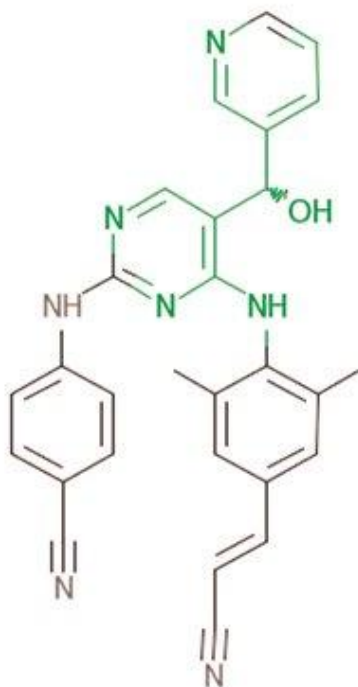
Model Interpretation (Compounds)

- Effect of substructure presence on compound pEC_{50}



Model Interpretation (Compounds)

- Example of positively correlated substructure and negatively correlated substructure



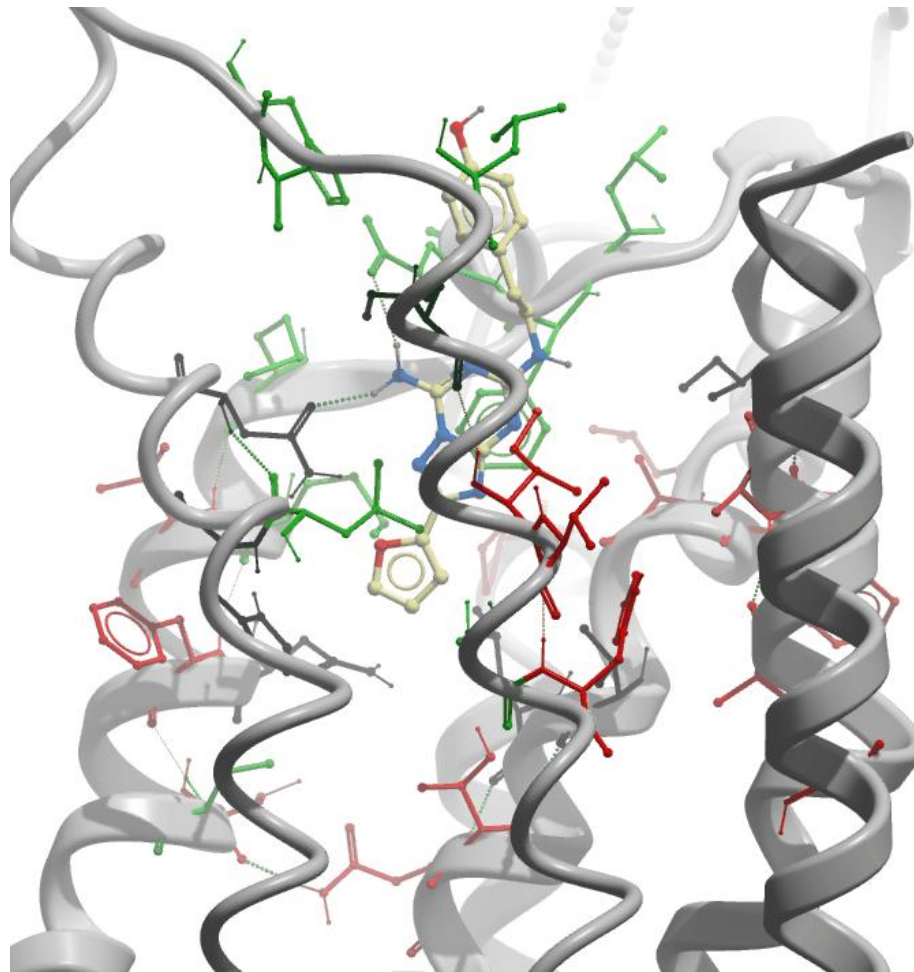
Conclusions

- PCM can guide inhibitor design by predicting bioactivity profiles, as applied here to NNRTIs
- We have shown prospectively that the performance of PCM approaches assay reproducibility (RMSE 0.62 vs 0.50)
- Interpretation allows selection between preferred chemical substructures and substructures to be avoided

Part II: PCM applied to the Adenosine Receptors

- Model based on public data (ChEMBL_04)
- Included:
 - Human receptor data
 - (Historic) Rat receptor data
- Defined a single binding site (including ELs)
 - Based on crystal structure 3EML and translated selected residues through MSA to other receptors
- ***Looking for novel A_{2A} receptor ligands taking SAR information from other adenosine receptor subtypes into account***

Selected Binding Site



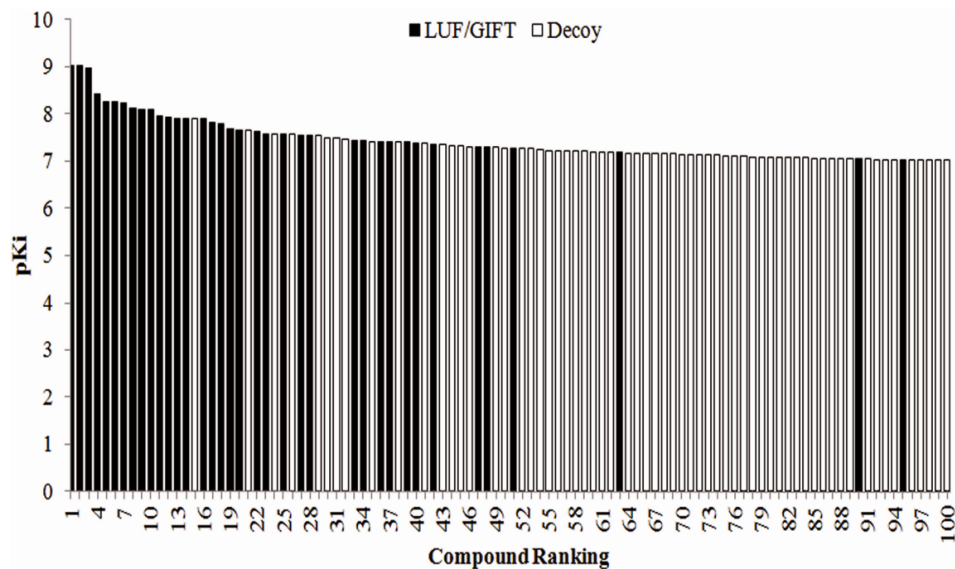
Adenosine Receptor Data Set

- Little overlap between species
- Validation set consists of 4556 decoys and 43 known actives

Receptor	Human	Rat	Overlap	Range (pKi)	External Validation	Decoy
A ₁	1635	2216	147	4.5 - 9.7	130	1139
A _{2A}	1526	2051	215	4.5 - 10.5	57	1139
A _{2B}	780	803	79	4.5 - 9.7	11	1139
A ₃	1661	327	82	4.5 - 10.0	255	1139

In-silico validation

- External validation on in house compound collection
 - Lower quality data set leads to less predictive model
 - Inclusion of Rat data *improves* model (RMSE 0.82 vs 0.87)
- Our final model is able to separate actives from decoys
 - 33 of the 43 known actives were in the top 50



Prospective Validation

- Scanned ChemDiv supplier database (> 790,000 cmpds)
- Selected 55 compounds with focus on diverse chemistry
 - Compounds were tested *in-vitro*



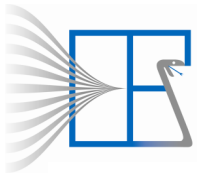
Conclusions

- We have found novel compounds active (in the nanomolar range) on the A_{2A} receptor
 - Hit rate ~11 %
- PCM models benefit from addition of similar targets from other species (RMSE improves from 0.87 to 0.82)
- PCM models can make robust predictions, even when trained on data from different labs

Further discussion

- Poster # 47 A. Hendriks, G.J.P. van Westen *et al.*
 - *Proteochemometric Modeling as a Tool to Predict Clinical Response to Antiretroviral Therapy Based on the Dominant Patient HIV Genotype*
- Poster # 51 E.B. Lenselink, G.J.P. van Westen *et al.*
 - *A Global Class A GPCR Proteochemometric Model: A Prospective Validation*
- Poster # 54 R.F. Swier, G.J.P. van Westen *et al.*
 - *3D-neighbourhood Protein Descriptors for Proteochemometric Modeling*

Acknowledgements



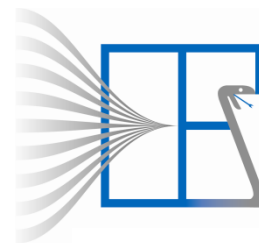
Leiden / Amsterdam
Center for Drug Research



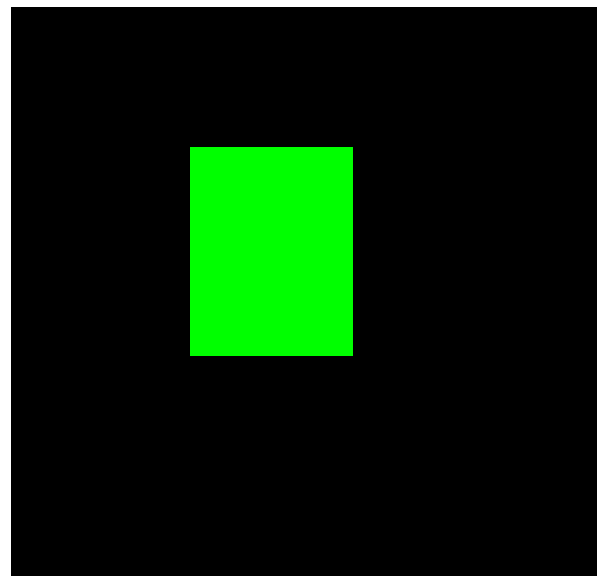
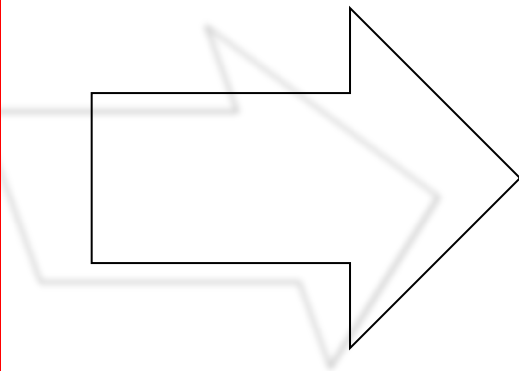
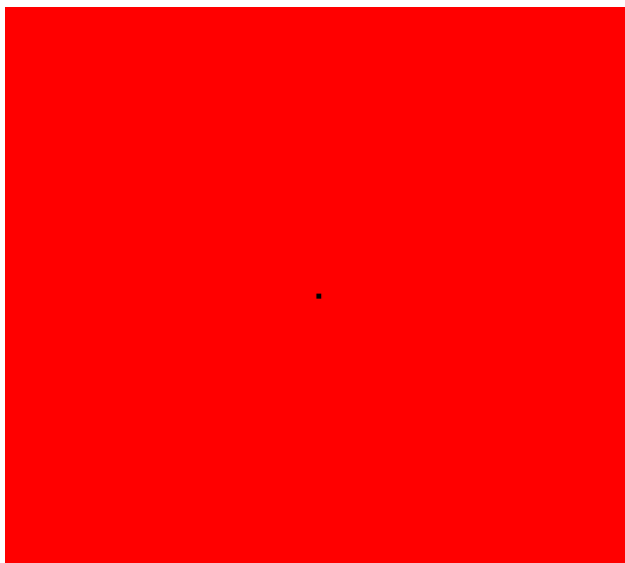
- Prof. Ad IJzerman
- Andreas Bender
- Olaf van den Hoven
- Rianne van der Pijl
- Thea Mulder
- Henk de Vries
- Alwin Hendriks
- Bart Lenselink
- Remco Swier
- Prof. Herman van Vlijmen
- Joerg Wegner
- Anik Peeters
- Peggy Geluykens
- Leen Kwanten
- Inge Vereycken

Real World Applications of Proteochemometric Modeling

*The Design of Enzyme Inhibitors and
Ligands of G-Protein Coupled Receptors*

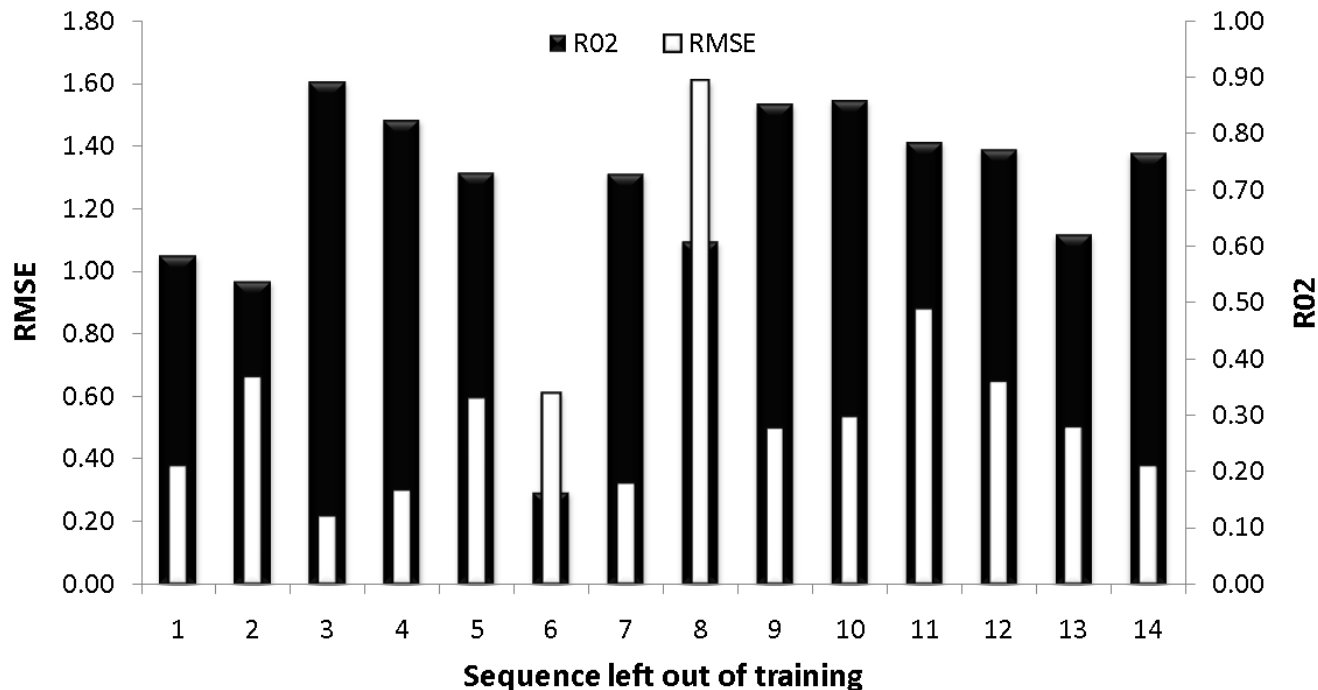


Leiden / Amsterdam
Center for Drug Research



Leave One Sequence Out

- By leaving out one sequence in training and validating a trained model on that sequence, model performance on novel mutants is emulated



Best performing compounds

Sequence	Compound with highest pEC ₅₀	Activity (pEC ₅₀)	Full Model (pEC ₅₀)	Difference (Activity and Model)
All	326	8.39(± 0.61)	8.53(± 0.73)	0.14
1	365	9.16	9.55	0.39
2	221	8.19	8.38	0.19
3	79	8.71	8.81	0.10
4	321	8.83	8.79	0.04
5	321	9.12	8.73	0.39
6	221	8.01	7.93	0.08
7	364	untested	7.50	n/a
8	221	untested	8.42	n/a
9	365	untested	9.43	n/a
10	326	untested	9.23	n/a
11	151	9.05	8.86	0.19
12	321	untested	9.29	n/a
13	100	9.06	8.87	0.19
14	79	9.51	9.62	0.11
			Average	0.18

Worst performing compounds

Sequence	Compound with Lowest pEC ₅₀	Activity (pEC ₅₀)	Full Model (pEC ₅₀)	Difference (Activity and Model)
All	109	5.85(±0.54)	5.82(±0.66)	0.03
1	248	6.09	6.01	0.08
2	109	untested	4.87	n/a
3	422	untested	5.78	n/a
4	84	5.84	5.67	0.17
5	84	5.65	5.54	0.11
6	109	4.60	4.06	0.54
7	439	5.01	5.20	0.19
8	84	4.74	5.20	0.46
9	248	untested	5.96	n/a
10	181	5.82	6.01	0.19
11	181	5.42	5.61	0.19
12	109	5.90	6.09	0.19
13	181	5.11	5.29	0.18
14	181	5.62	5.81	0.19
			Average	0.21