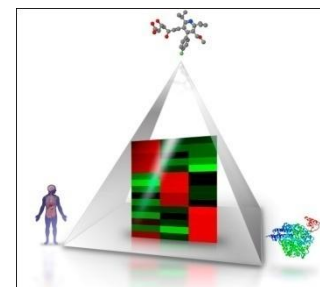


# From Single-Target Models to Multiple-Target Models

## Extrapolating in Target Space Using Proteochemometrics Approaches

Andreas Bender, Leiden / Amsterdam Center for Drug Research & Unilever Centre for Molecular Informatics, University of Cambridge



# How can we anticipate bioactivity spectra of compounds early on?

- Where single-target models are useful
- The importance of more complex approaches: bioactivity profiles and phenotypic screening
- Single-target bioactivity models vs. predicting bioactivity profiles
- Prospective studies on non-nucleoside HIV reverse transcriptase inhibitors



# Single-Target drugs vs. Selective Promiscuity

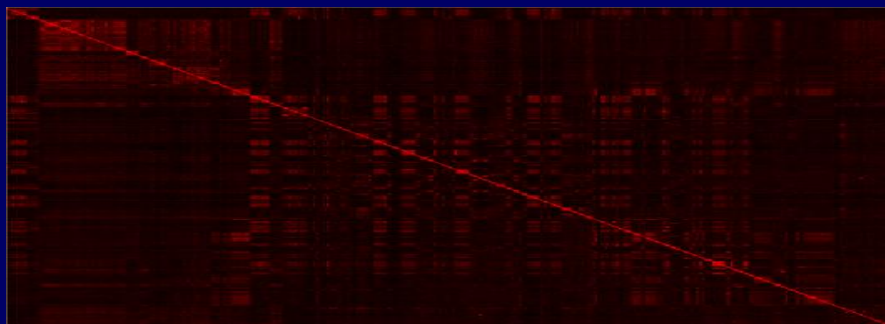
- Drug discovery thousands of years ago:  
Try this herb!  
(If it works, you are healed. If it doesn't, you are dead.  
But you did it for your children.)
- Drug discovery in the 1980's: Isolated receptors (HTS), millions of compounds (CombiChem). High throughput, but less relevance.
- Today: Realization that activity against *sets of targets* is relevant for both efficacy and safety
- More relevant for modulating the phenotype



# Conventional Approach: Single-target models ('QSAR')

- Bioactivity models can (ideally) predict activity of compounds on a single target, based on chemical structure
- Models *cannot*:
  - Predict activity on *related* targets based on chemical structure
  - Extrapolate *reliably* beyond the original training-set

QSAR : Looking at ligand properties alone



- However, in some cases they are also useful...

# Example: *In silico* Target Fishing with orphan ligands to rationalize phenotypes

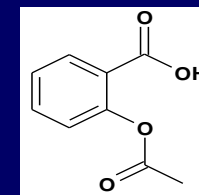
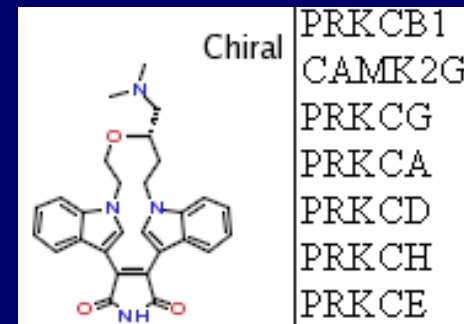
## Motivation

Often we know the *compound*, and the *phenotype* (adverse reaction, high-content screening output, ...) but not the *mode of action*



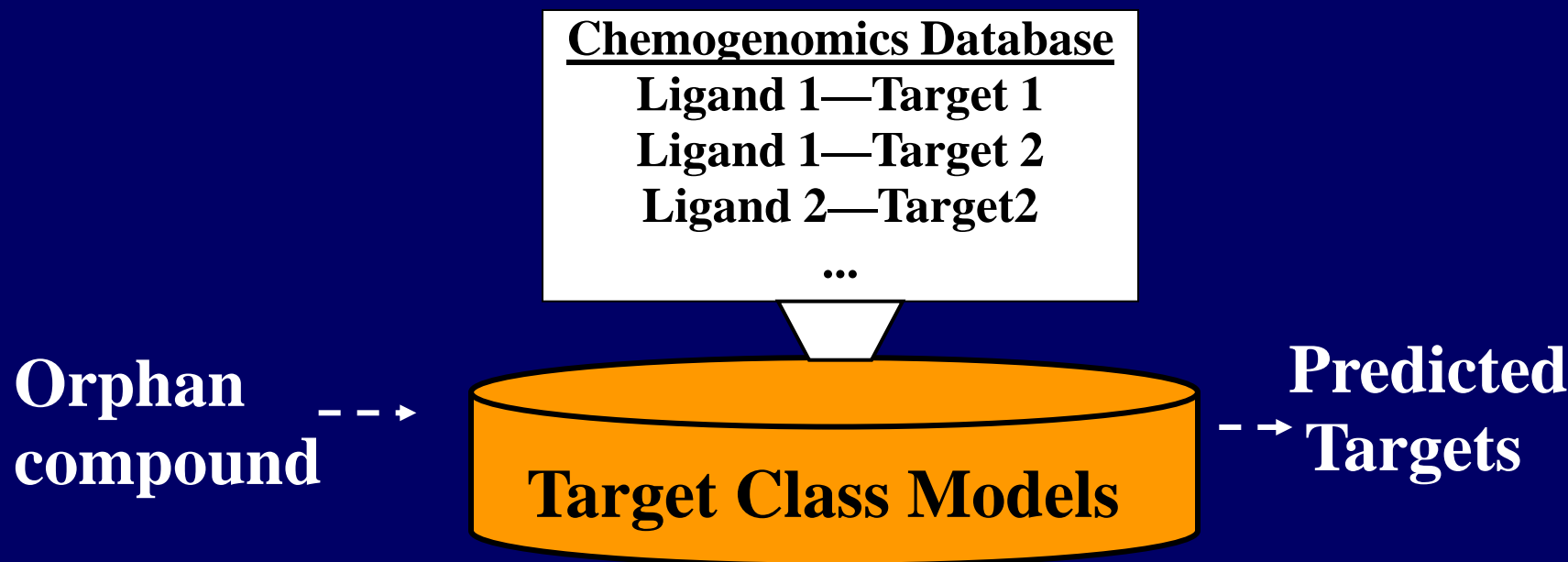
At the same time, decades of public and proprietary data on protein-ligand interactions provide a wealth of minable SAR information

Review: J.L. Jenkins, A.Bender, J.W. Davies.  
*Drug Discov. Today: Technol.* **2007** (3), 413- 421.




# Ligand-Target Prediction

- We have millions of datapoints which molecular structures are binding to which protein in the body
- Thus, we can build statistical ligand-target relationships to *predict* the targets of molecules, based on their molecular structure
- We used WOMBAT DB, but ChEMBLDB ,GVK, Jubilant, in-house data, ... are other viable options

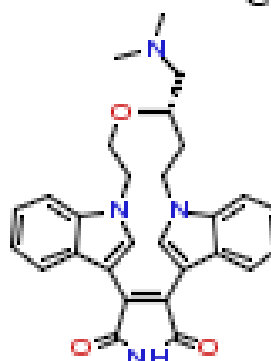


# Prediction Examples: Gleevec, Ruboxistaurin

- Gleevec (Novartis),
  - Launched
  - Targets Bcr-Abl, c-kit, PDGFRb

Molecule	Targets	Scores
	ABL1	46.50
	PDGFRB	28.99
	KIT	22.02
	CDK9	21.30
	BRAF	16.13
	FLT1	13.09
	PLK1	8.05
	BTK	5.44

- Ruboxistaurin (Lilly/Takeda), Phase III
  - PKCb

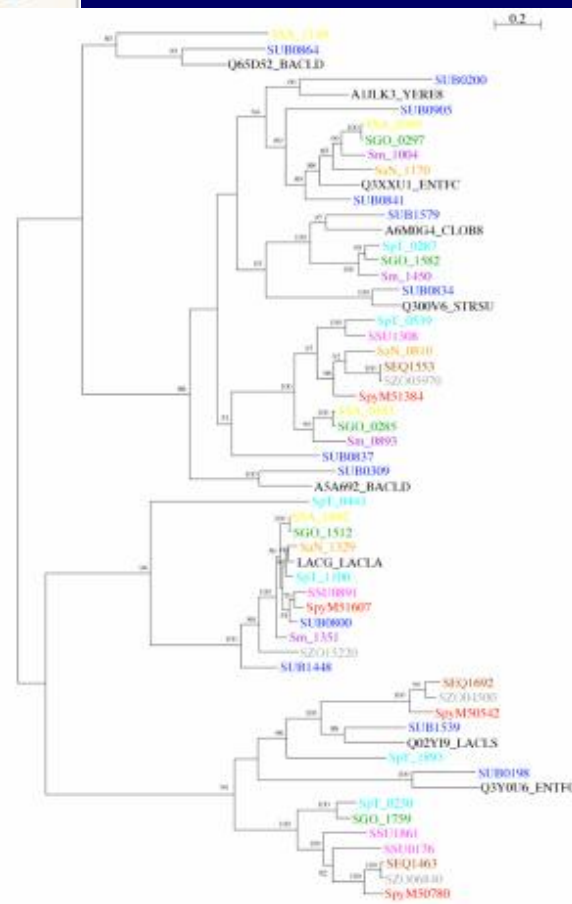
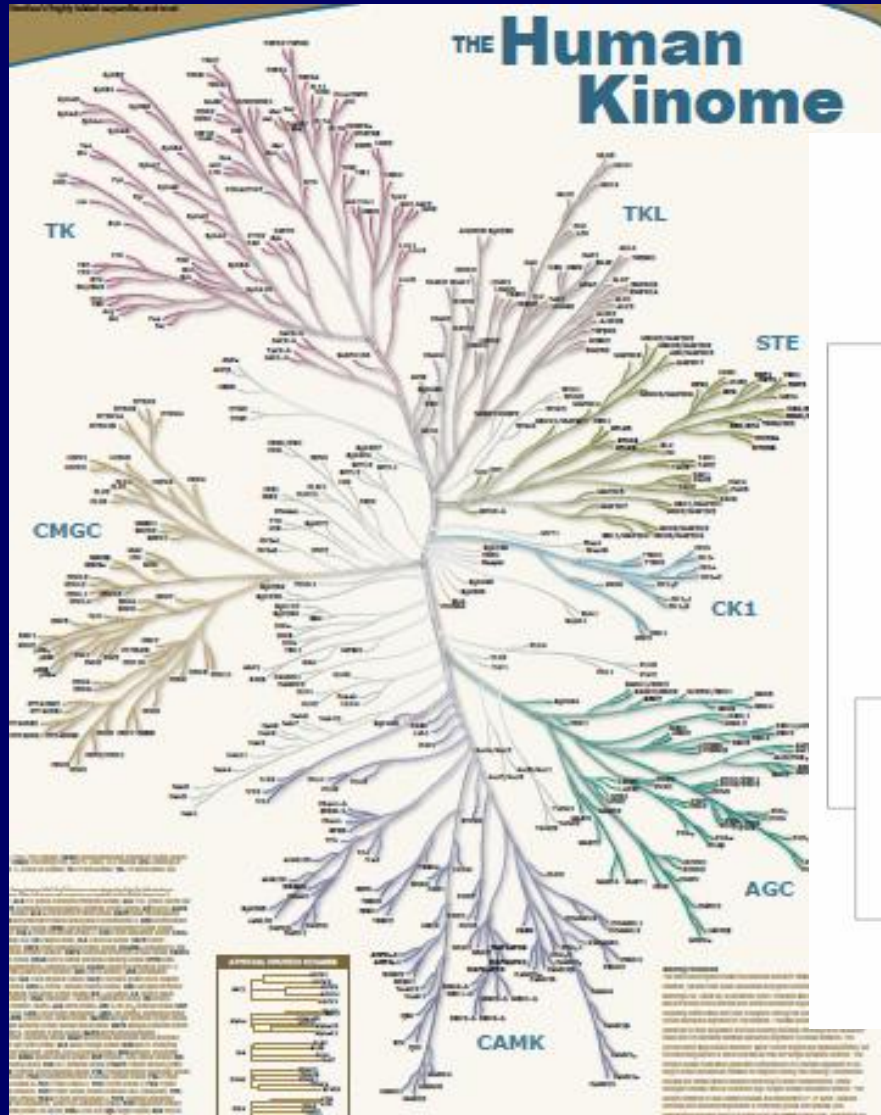
Molecule	Targets	Scores
 <p>Chiral</p>	PRKCB1	95.81
	CAMK2G	87.48
	PRKCG	66.35
	PRKCA	56.99
	PRKCD	52.44
	PRKCH	51.41
	PRKCE	50.42
	PRKCZ	42.48

# Summary Target Prediction

- We can predict molecular targets, based on existing ligand-target bioactivity databases
- This works better the more chemistry is 'known'
- Examples shown: Prediction of kinase selectivity/promiscuity
- However, until this stage we do *not take target similarities into account* when predicting targets!
- All target prediction models are basically *independent*



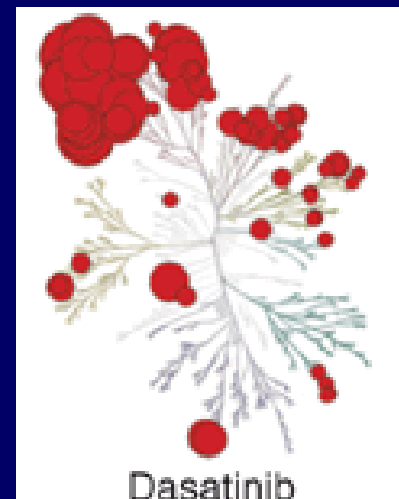
So what we need to consider is the inter-target relationships to generalize interaction space: Can be GPCRs, kinases, enzyme mutants, ...



- GP57
- GP58
- HH2R
- 5H2A
- 5H2C
- 5H2B
- D4DR
- D3DR
- D2DR
- 5H1A
- 5H7
- 5H5A
- 5H1F

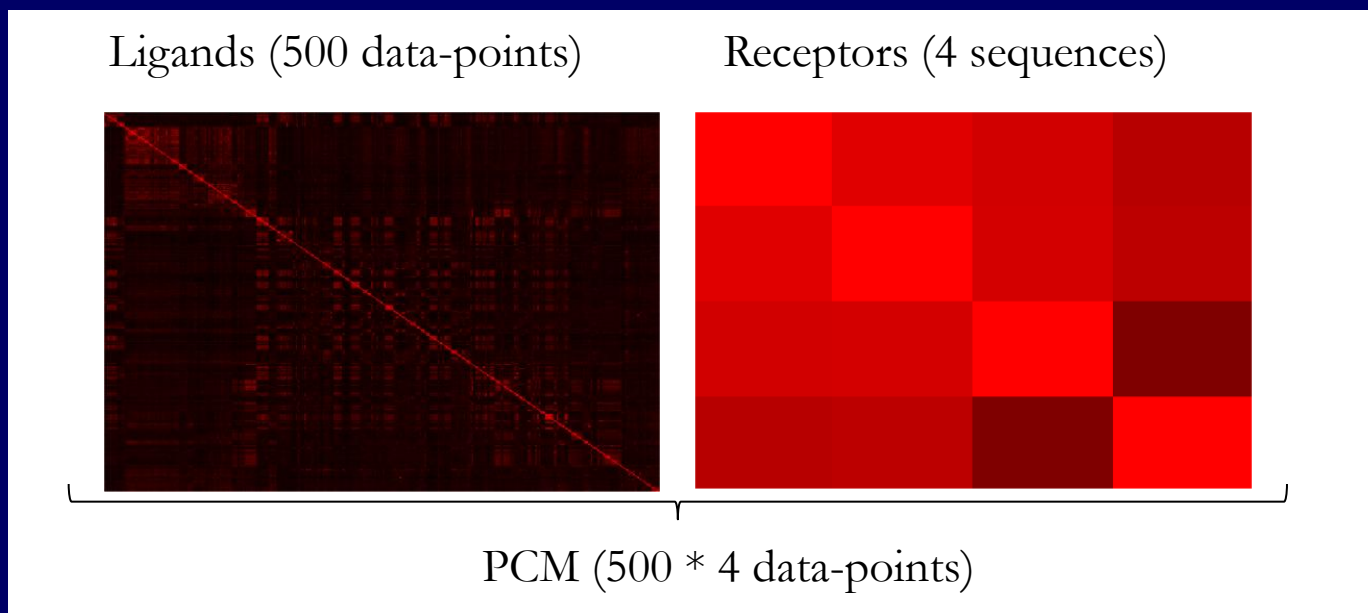
# Prospective Prediction of Bioactivity Profiles

- Hit 'relevant' targets – either single ones, or combinations thereof
- Question: How to anticipate activity spectra *early on*?
- Together with Leiden University and Tibotec, we are developing 'proteochemometrics' approaches to predict selectivity profiles of compounds
- van Westen *et al.*, Proteochemometric Modeling as a Tool for Designing Selective Compounds and Extrapolating to Novel Targets. *MedChemComm in press* (2010).



# Proteochemometric Modeling takes target properties into account

- Models the ligand – target interaction space
- Predict activity on related targets



- van Westen *et al.*, Proteochemometric Modeling as a Tool for Designing Selective Compounds and Extrapolating to Novel Targets. *MedChemComm in press* (2010).

# Case Study: HIV Reverse Transcriptase Inhibitors

- Overview:
  - HIV mutates
  - So does HIV RT
  - Q: *Which drug should we use on a patient?*
- A:
  - We model HIV RT inhibitors
  - We take also information about mutations into account
  - We show (a) that we can *predict activities for a sequence left out* (retrospectively) – ‘known drugs, new patient’
  - We show (b) that we can *predict novel ligand-target pairs* (prospectively) – ‘which drug for which patient’



# The Problem With HIV (I)

- HIV is a fatal disease...  
... but can be controlled today
- However, HIV still is a worldwide epidemic

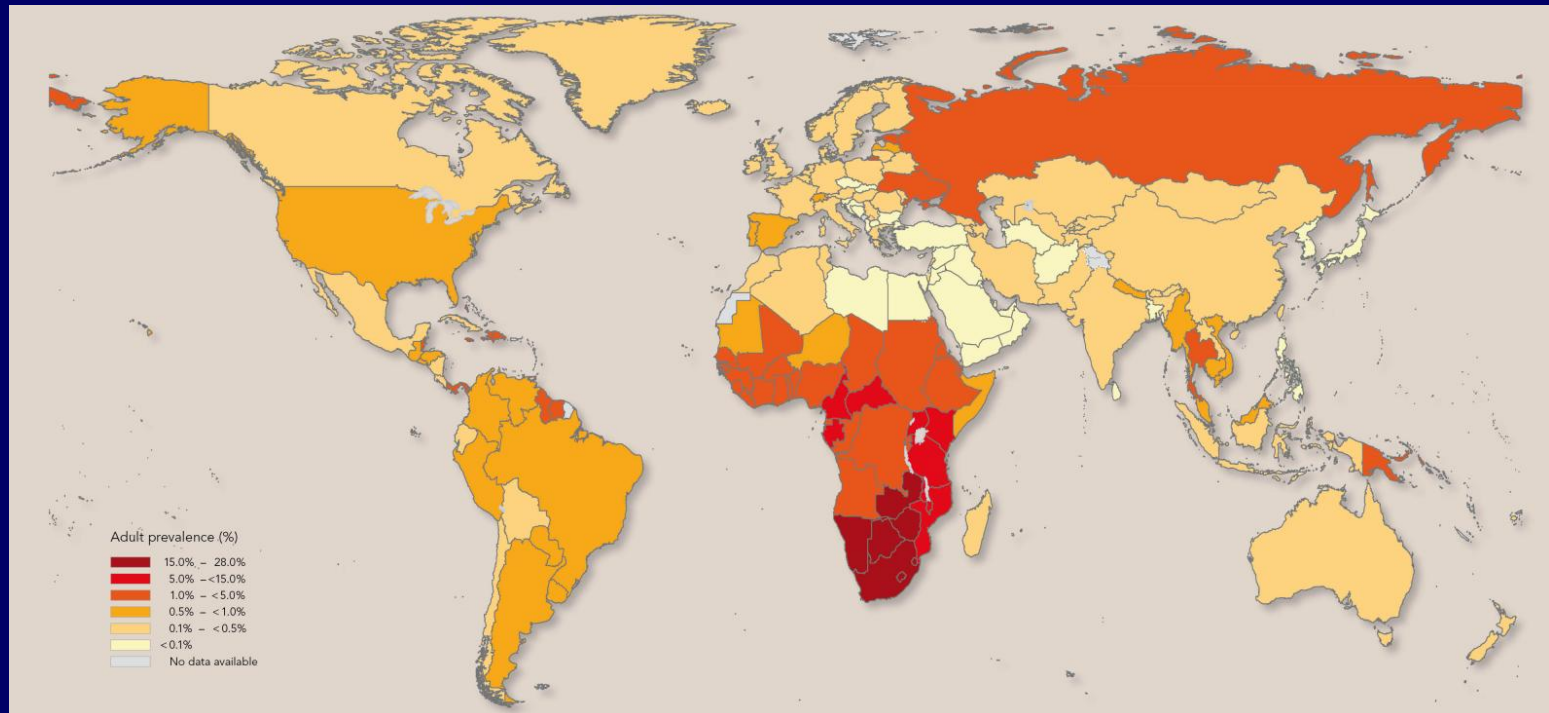
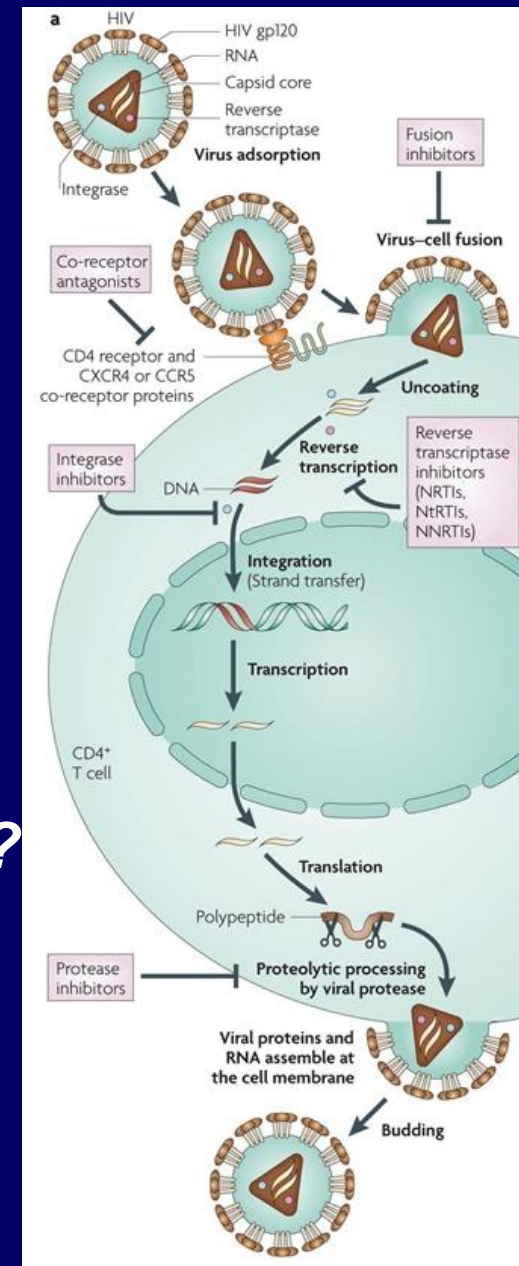


Figure: 2008 Report on the global AIDS epidemic (UNAIDS)

# The Problem With HIV (II)

- Inhibitors exist acting on different targets
  - Reverse Transcriptase (RT), Protease, Integrase, etc
- However, the problem is mutation
  - A large number of drugs, quickly became ineffective due to resistance
- **The main question then is: *Which drug should I use for which patient?***



1: Frankel, A.D. & Young, J.A. Annu Rev Biochem (1998) 67, 1-25

Figure: Nature Reviews Drug Discovery, Dec2007, Vol. 6 Issue 12, p1001-1018, 18p; found on p1004

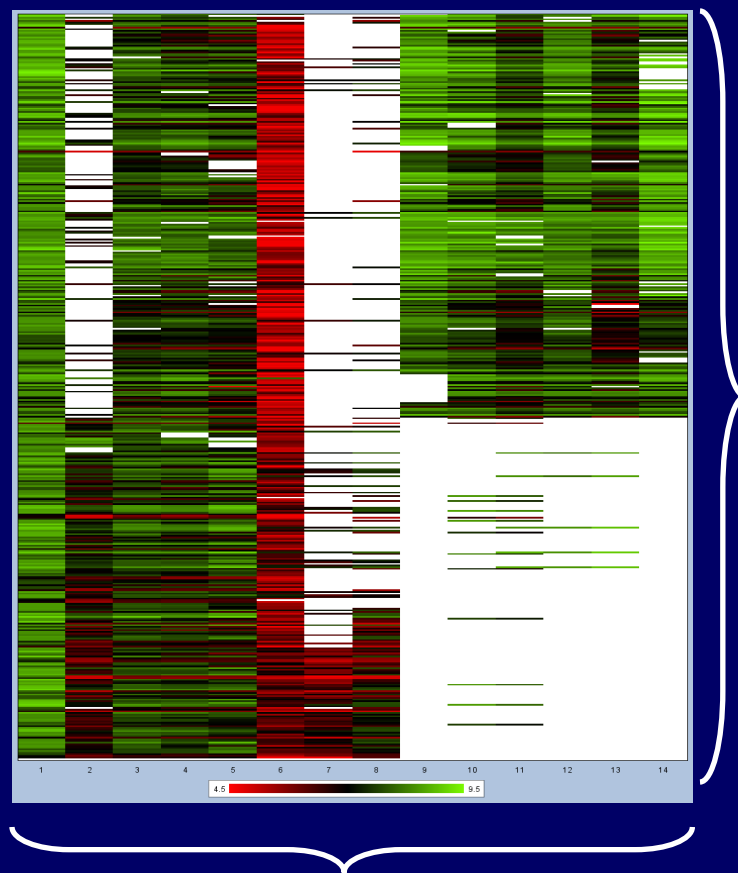
# HIV Reverse Transcriptase Inhibitor Dataset

- 451 HIV Reverse Transcriptase (RT) inhibitors
  - Allosteric (Non-Nucleoside Reverse Transcriptase Inhibitors)
  - Analog series (2 scaffolds)
- 14 HIV RT sequences
  - Between zero and 13 point mutations (at NNRTI binding site)
  - Large differences in compound activity on different sequences

Sequence	Mean pEC50	Stdev pEC50	n
1	8.3	0.6	450
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

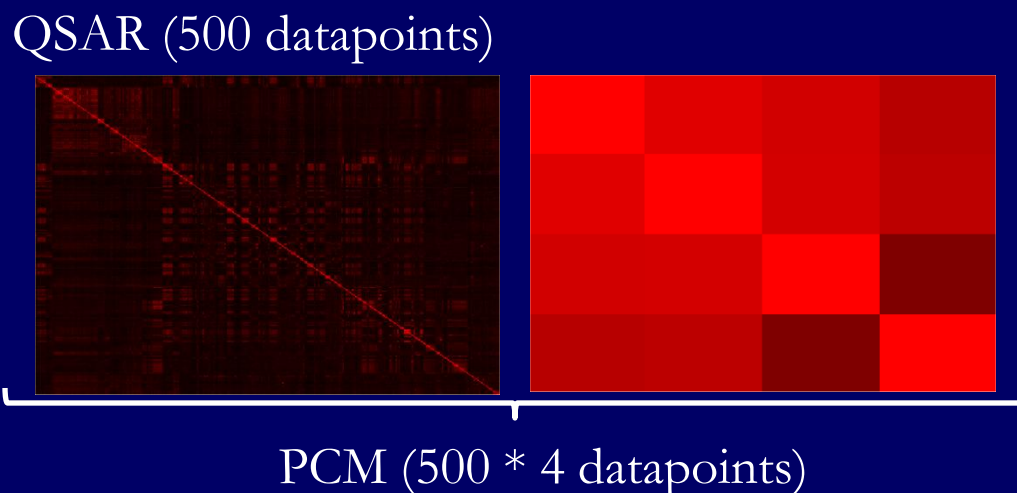
# The inhibitor-mutant bioactivity matrix is incomplete

- Each sequence can be paired with each compound
- Activity definition:  $pEC_{50}$ 
  - Matrix was 64 % complete
    - 4024 out of 6314 compound – sequence pairs had  $pEC_{50}$  value
  - Individual compounds can perform very different on different sequences
- **How can we extrapolate to unknown ligand-target pairs?**



# Proteochemometric Modeling (I)

- PCM Introduced by Wikberg *et al.* in 2001 <sup>2</sup>
  - Analogue to QSAR modeling however, includes target descriptors in addition to compound descriptors
  - Models the ligand – target interaction space
  - Improves conceptually the extrapolation capabilities of models



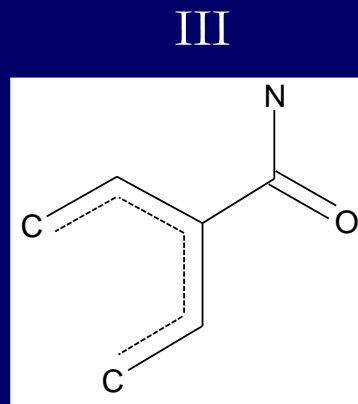
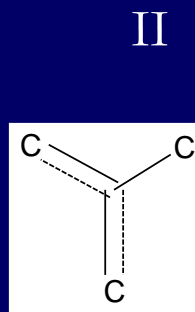
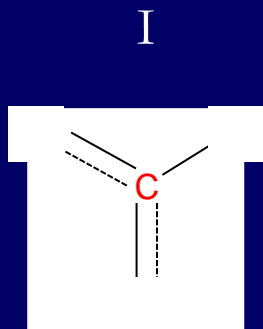
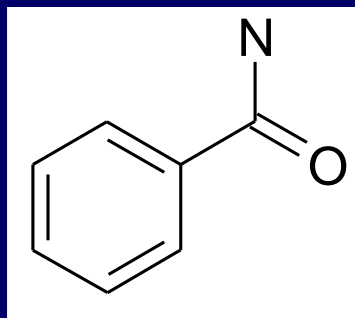
# Proteochemometric Modeling (II)

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



# Compound Descriptors

- Compounds (451) were described using Scitegic FCFP\_6 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
-2005884698

# Target Descriptors

- Use feature based target descriptor for compatibility
  - Constructed from Aaindex derived parameters
    - Hydrophobicity, Flexibility, etc...
  - Use a PCA to *select* a broad number of indices to generate a unique hashed feature for each amino acid
- Z-scales: *PCA* on 80+ amino acids
- Here: Natural AAs, *PCA* only for *selection!*
- Each unique hashed feature represents one amino acid type (comparable with FCFP fingerprints)
  - Currently being extended to take *distance* between amino acids into account



# Compound – target descriptors

- Compound descriptors are combined with target descriptors and represent the input vector
- This input vector is linked to an activity value (pEC50) which represent the variable to be modeled

*Compound Descriptors*

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

*Target Descriptors*

ProtFP
1169372512
-590269326
268201585
268201585
268201585
-58134849
-58134849
-1481898440
558044215
-58134849
4327070000

*Combined  
with*

*Activity*

pEC50
7.8



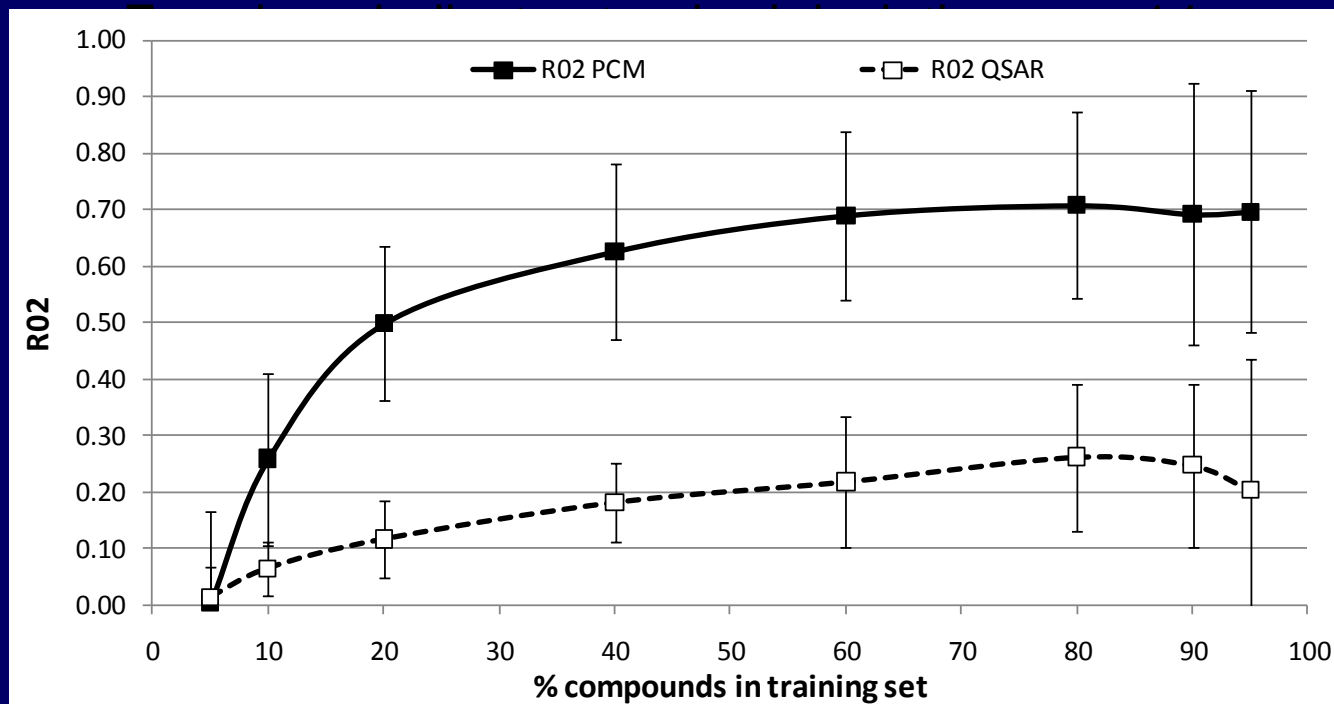
# Modeling Software used

- Models generated using Pipeline Pilot 6.1.5
- Machine learning in R-Statistics 2.3.1
  - Support Vector Machines
    - Sampled Gamma and Cost over an exponential range using 10-fold CV
    - Epsilon was set at 0.2 (equal to experimental error)
    - Fingerprints folded to fixed length array of counts (256 with first columns carrying most data)
- Validation using :
  - $R_0^2$  (ranking of compound – target pairs)
  - RMSE (prediction error)



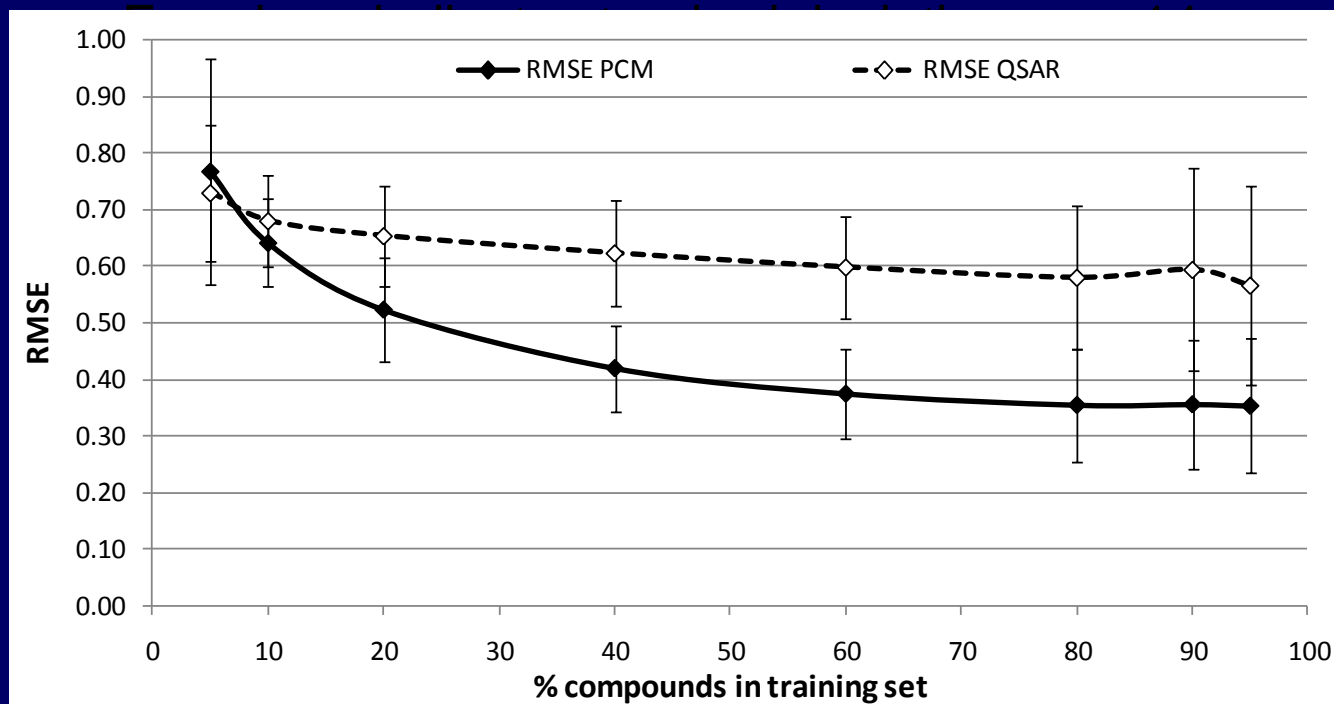
# *In-silico* validation: $R_0^2$

- Identical data set
  - Average  $R_0^2$  was measured over all predictions *per* sequence
  - PCM included targets, QSAR only compounds



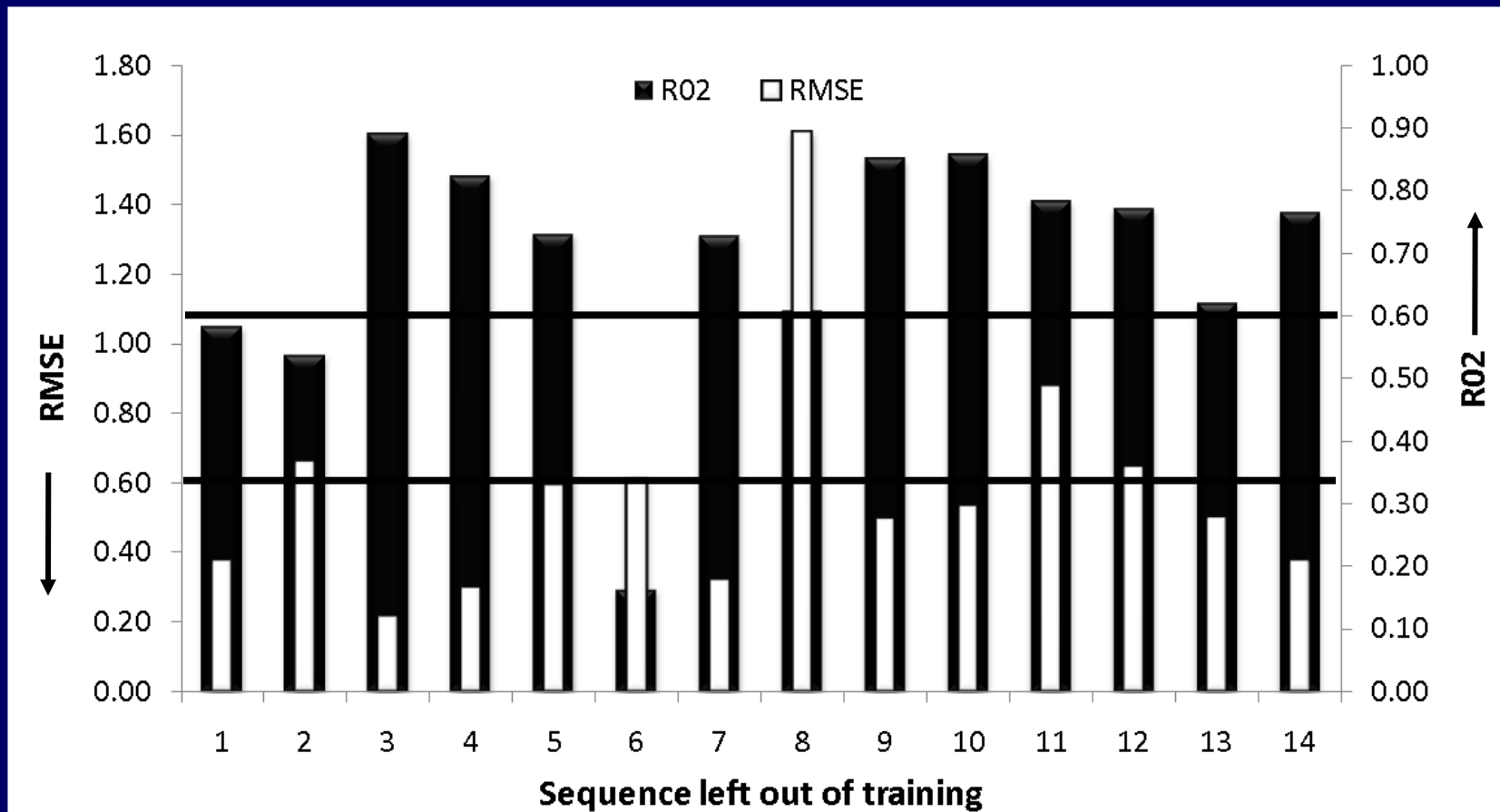
# *In-silico* validation: RMSE

- Identical data set
  - Average RMSE was measured over all predictions *per* sequence
  - PCM included targets, QSAR only compounds



# Which drug to use on a novel patient? Leave-One-Sequence-Out Validation

- Inter-/Extrapolation in target space



# Why did 6 and 8 underperform?

- Sequences 6 and 8 represent a singleton
  - Sequence 6 is the only sequence containing high impact mutation E138G
- Sequence 8 contains K101P, which is only present in heavy (13) mutant (sequence 7)
  - Model has difficulty scaling the contribution of K101P in the background 12 other mutations
  - Reversed experiment does work (leaving out 7)

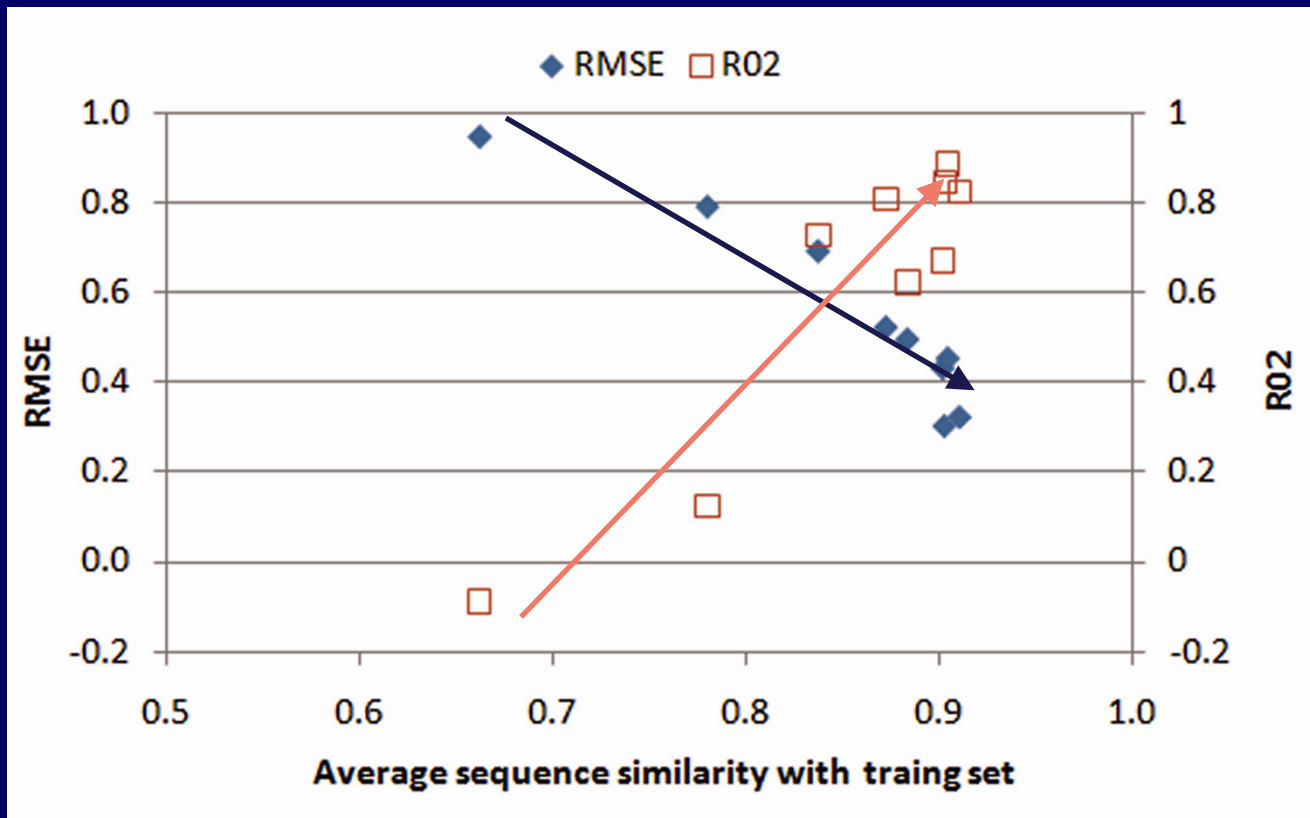
	89	100	101	102	103	106	118	162	169	179	181	188	190	203	207	210	211	214	215	219	227	234	245	138
1	A	L	K	K	K	V	V	S	E	V	Y	Y	G	E	Q	L	R	L	T	K	F	L	V	E
6	A	I	K	K	N	V	V	S	E	I	C	Y	G	E	Q	L	R	F	T	K	F	L	V	G
7	S	L	P	R	K	V	I	S	E	V	C	Y	A	V	E	W	R	F	Y	N	F	L	I	E
8	A	L	P	K	K	V	V	S	E	V	Y	Y	G	E	Q	L	R	L	T	K	F	L	V	E

# Extending the Applicability Domain Concept From Ligand to Target Space

- Applicability Domain mentions in which cases the model is applicable for the *ligand side*
  - Based on similarity between training set and unknown compound
- Here we extend it to the target side
- Similar to chemogenomics approaches <sup>3</sup>, PCM could e.g. be applied for GPCR deorphanization

# The Applicability Domain Concept Still Holds in Target Space

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

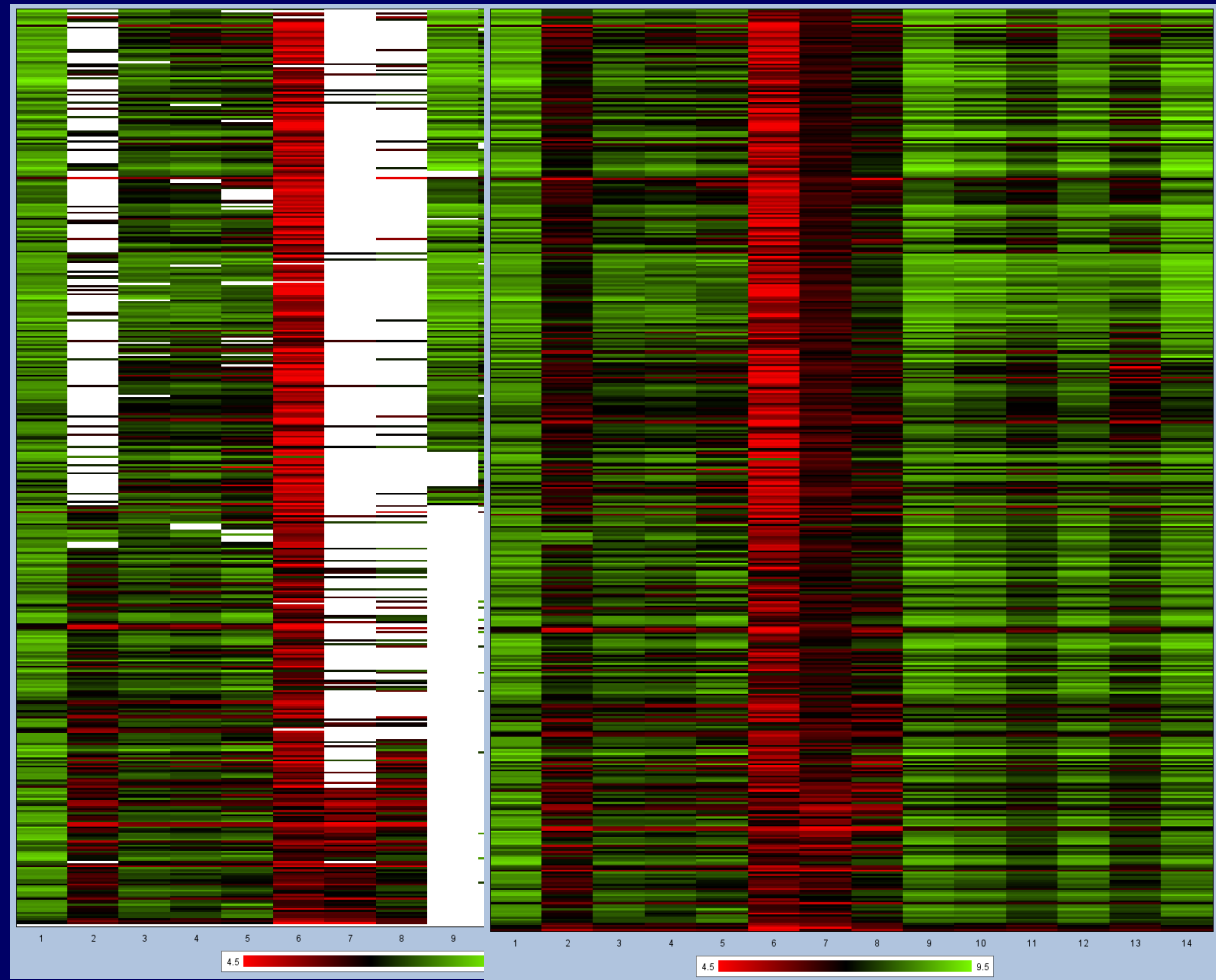


# Applicability Domain Concept

- Prediction error similarity shows a direct correlation with sequence similarity
- We successfully extended the applicability domain concept to the target side
- A full 2D plot is currently being investigated, analyzing ligand and sequence similarity in parallel and relating these properties to prediction error



# Prospective Validation of HIV RT inhibitors: Picking the right 'personalized medicine'

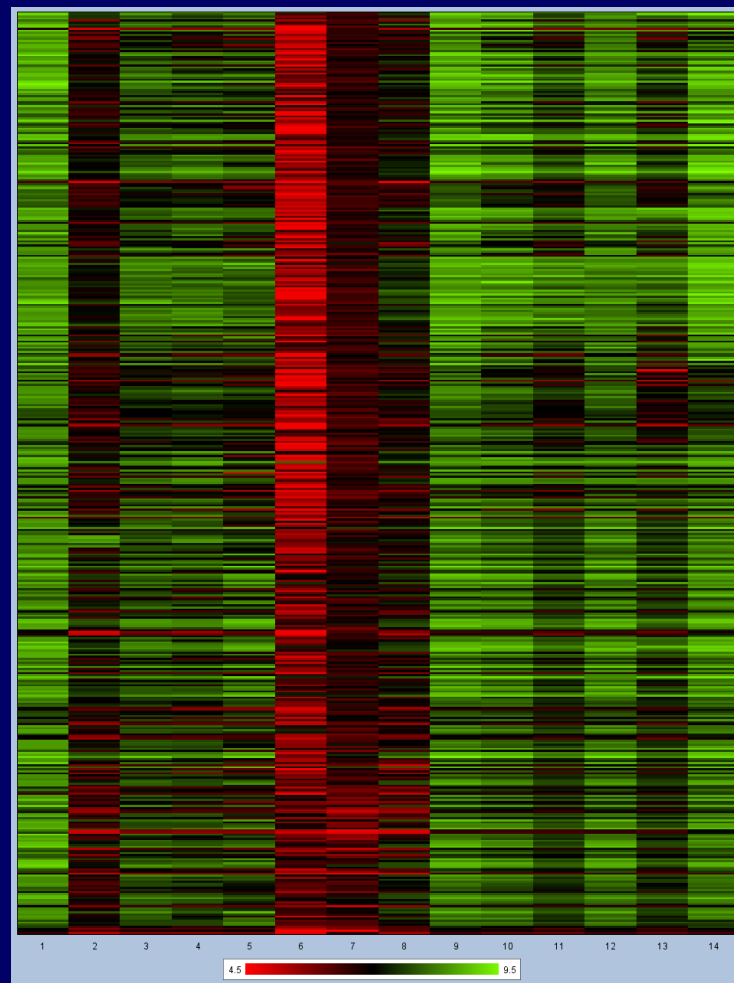


Original  
Dataset

Completed with  
model

# Prospective Validation: Compound Selection

- Compounds have been experimentally validated
  - Predictions where pEC50 differs two standard deviations from compound average (69)
  - Prediction where pEC50 differs two standard deviations from sequence average (61)
- Assay validation

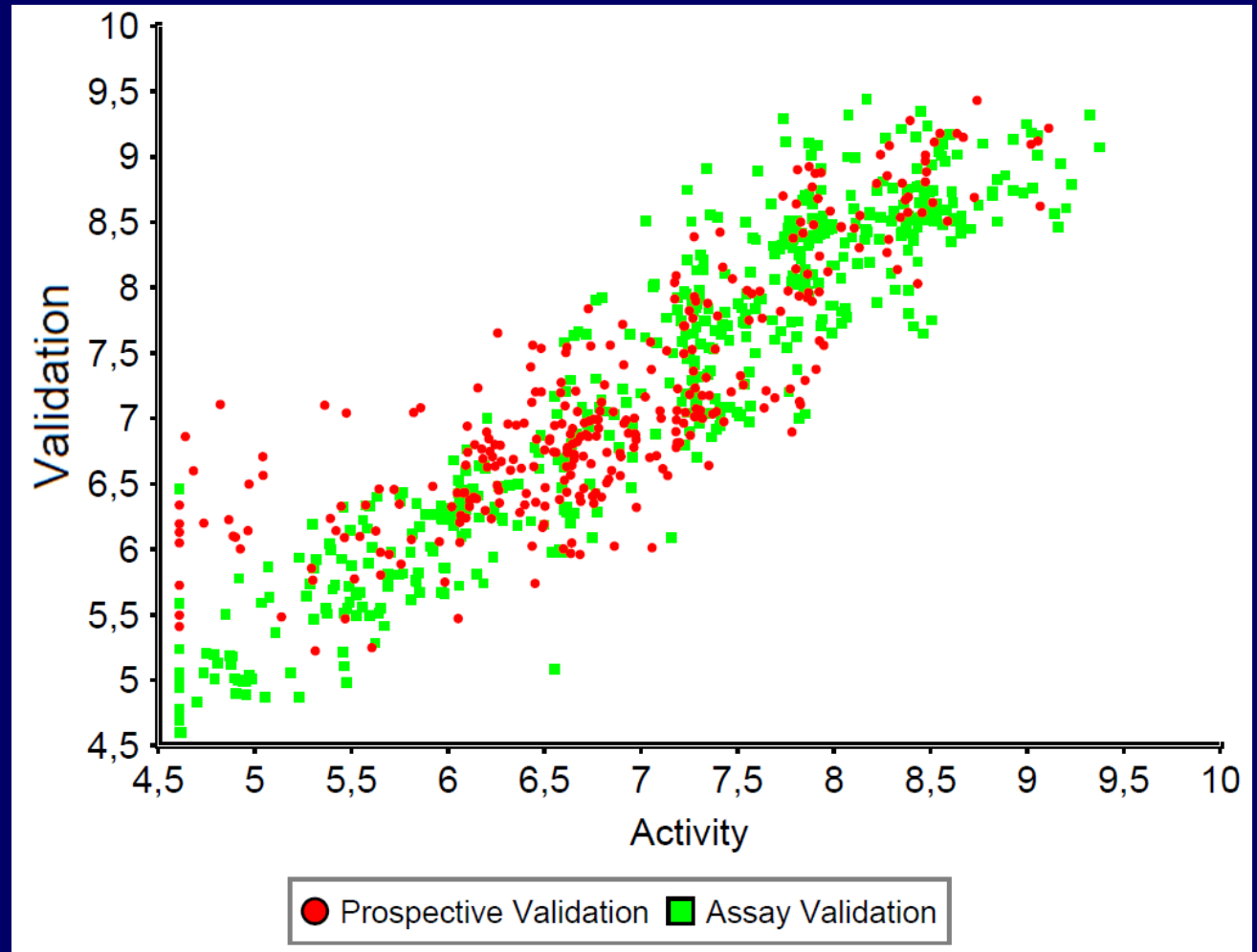


Completed with  
model ↑

# Prospective Predictions on Novel Compounds

## Approach Assay Performance

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



# Summary of Bioactivity Profile Prediction

- We can (and should!) exploit knowledge of chemical and biological space to predict *bioactivity profiles*
- Including knowledge about target relationships improves our ability to do this
- This has been shown prospectively experimentally for HIV RT mutants, we now work on a subset of the GPCR family
- This is relevant for both *efficacy (for on-targets)* and *safety (for off-targets)*
- For review see: van Westen *et al.*, Proteochemometric Modeling as a Tool for Designing Selective Compounds and Extrapolating to Novel Targets. *MedChemComm in press* (2010).



# Acknowledgements

Leiden University

- Gerard van Westen, Ad IJzerman

Tibotec

- Herman van Vlijmen, Joerg Wegner

