

Proteochemometric modeling as a tool to predict clinical response to Anti-Retroviral therapy based on the patient's dominant HIV-genotype

Alwin Hendriks^a, Gerard J.P. van Westen^a, Jörg K. Wegner^b, Andreas Bender^c, Adriaan P. IJzerman^a, Herman W. T. van Vlijmen^{a,b}
^aLeiden / Amsterdam Center for Drug Research, Leiden University, Netherlands; ^bTibotec BVBA, Beerse, Belgium; ^cUniversity of Cambridge, United Kingdom

Introduction

- Single drug treatment in HIV leads to quick occurrence of resistance
- Combination of multiple drugs from different classes (HAART) delays resistance onset
- Predictive *in silico* models can complement assay-based resistance determination on individual targets^{1,2} to determine optimal HAART treatment
- Proteochemometric modeling (PCM)³ enables the modeling of interactions between **several targets (HIV mutants)** against **several small molecules (clinical drugs inhibiting an enzyme)**
- Hereby PCM can possibly determine the optimal HAART regimen per mutant (or per patient)

Aims

- Proof of concept for PCM, based on *in vitro* data of clinical samples
- 3 predictive models (for PI, NRTI and NNRTI drug classes) to determine activity fold change due to viral mutations were generated (see Figure 1)
- Mapping of chemical space, target space and activity space of our dataset (see Figures 2 and 3)

Dataset Composition

- 21 clinical HIV inhibitors used
- Compound descriptors are the Tversky-based similarity vectors to the other compounds in each model
- ECFP₁₂ (PI), ECFP₁₀ (NRTI), ECFP₈ (NNRTI) fingerprints used to construct the similarity matrices
- Physicochemical properties of amino acid side chains (full sequence, Z-scales) serve as the target descriptor

Target	Binding Site	Class	Drugs	Mutants	Datapoints	Amino acids
Reverse Transcriptase	Orthosteric	NRTI	8	10,948	72,727	400
Reverse Transcriptase	Allosteric	NNRTI	4	10,948	35,249	400
Protease	Orthosteric	PI	9	27,472	180,162	99

Model Training

- Models for the learning curves were trained using Support Vector Machines (SVMs)
 - $\text{Gamma: } 1/nx$; $\text{Cost: } 10$; $\text{Epsilon: } 0.25$
- Final models are still in training, with gamma and cost optimization
 - $\text{Gamma: } 2^y/nx$, where y ranges from -6 to 4
 - $\text{Cost: } 10^y$, where y ranges from -6 to 6
 - $\text{Epsilon: } 0.25$
- Model validation performed by external validation
 - Computational validation on unseen sequences
 - Experimental validation on previously untested sequences

References

- Hertogs, K., M.-P. de Bethune, V. Miller, T. Ivens, P. Schel, A. Van Cauwenberge, C. Van den Eynde, V. van Gerwen, H. Azijn, M. van Houtte, F. Peeters, S. Staszewski, M. Conant, S. Bloor, S. Kemp, B. Larder, and R. Pauwels. A Rapid Method for Simultaneous Detection of Phenotypic Resistance to Inhibitors of Protease and Reverse Transcriptase in Recombinant Human Immunodeficiency Virus Type 1 Isolates from Patients Treated with Antiretroviral Drugs. *Antimicrobial Agents and Chemotherapy*, 1998; 42: 269-276
- Vermeiren, H., E. Van Craenenbroeck, P. Alen, L. Bachelier, G. Picchio, and P. Lecocq. Prediction of HIV-1 drug susceptibility phenotype from the viral genotype using linear regression modeling. *Journal of Virological Methods*, 2007; 145: 47-55
- van Westen, G.J.P.; Wegner, J.K.; IJzerman, A.P.; van Vlijmen, H.W.T.; Bender, A.; Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Med. Chem. Commun.*, 2011; 2(1): 16-30

Results

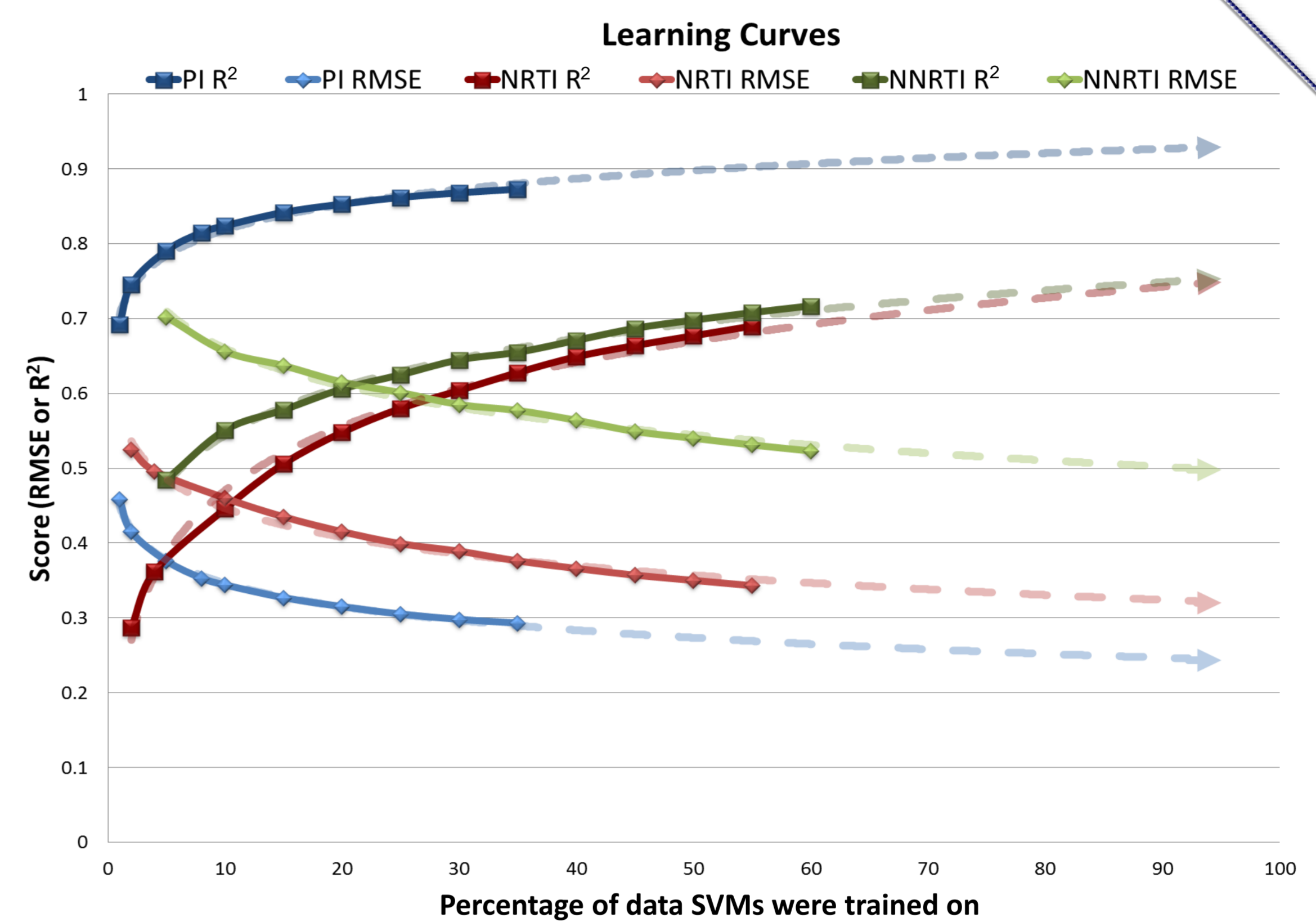


Figure 1: Learning Curves for the three drug classes. The R² represents the prediction capability and the RMSE represents the error of prediction. The dotted lines show the extrapolation of the curves.

Learning curves were Generated to determine maximum model performance to be reached on the different datasets. The Protease inhibitors (PIs) already show good performance on a small data subset. The Nucleoside Reverse Transcriptase Inhibitors (NRTIs) and Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs) need a considerable bigger amounts of data to be trained well.

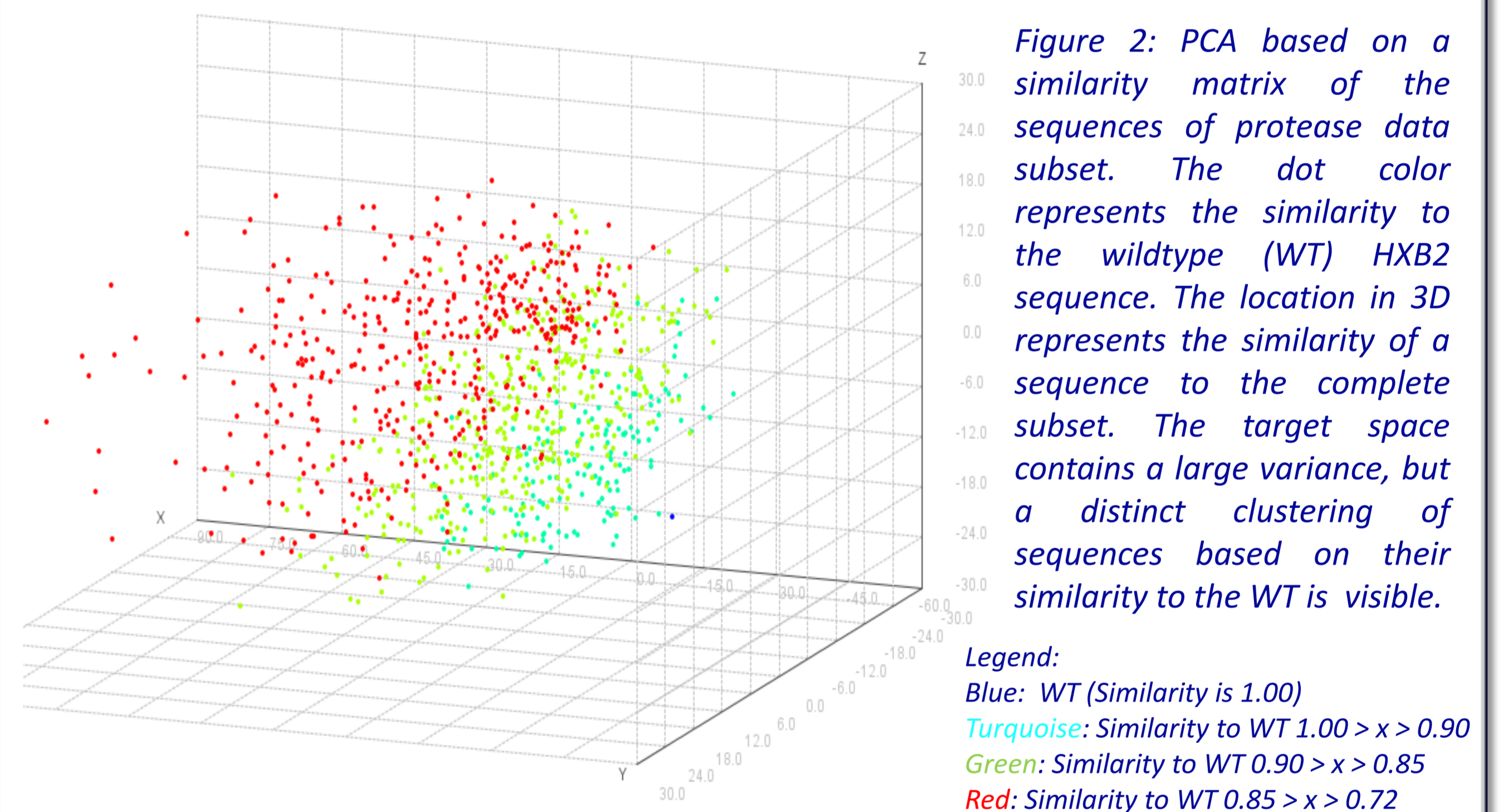


Figure 2: PCA based on a similarity matrix of the sequences of protease data subset. The dot color represents the similarity to the wildtype (WT) HXB2 sequence. The location in 3D represents the similarity of a sequence to the complete subset. The target space contains a large variance, but a distinct clustering of sequences based on their similarity to the WT is visible.

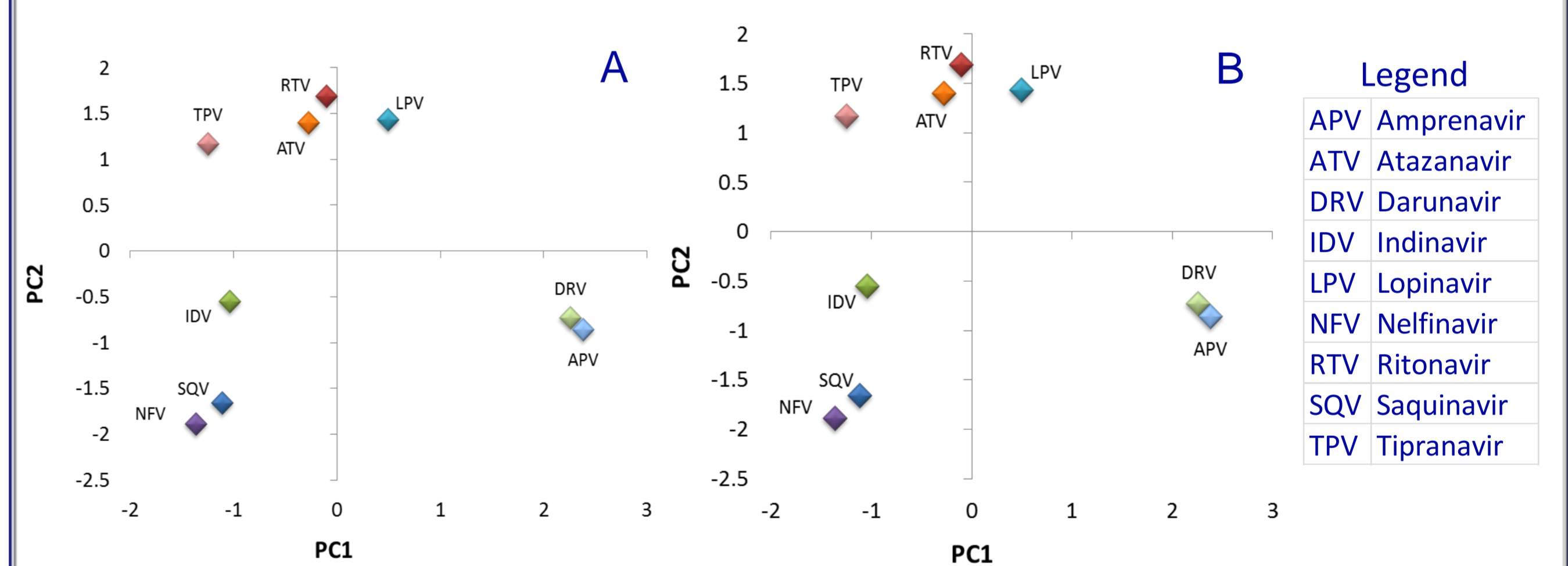


Figure 3: A) PCA loadings of the activity spectrum of the different PI drugs. B) PCA scores of the chemical similarity of PIs.

While there is agreement between chemical and SAR space, they are not identical. Since it is difficult to explain SAR space based solely on chemical similarity, the application of PCM models improves upon the situation, as presented in this work.

Conclusion

- PCM can be used to predict mutant resistance to clinical drugs
- RMSEs of 0.29 (PI), 0.34 (NRTI) and 0.52 (NNRTI) were obtained in the learning curves which are expected to improve further on the complete set
- PCM is able to combine chemical space, target space and activity space in a single model