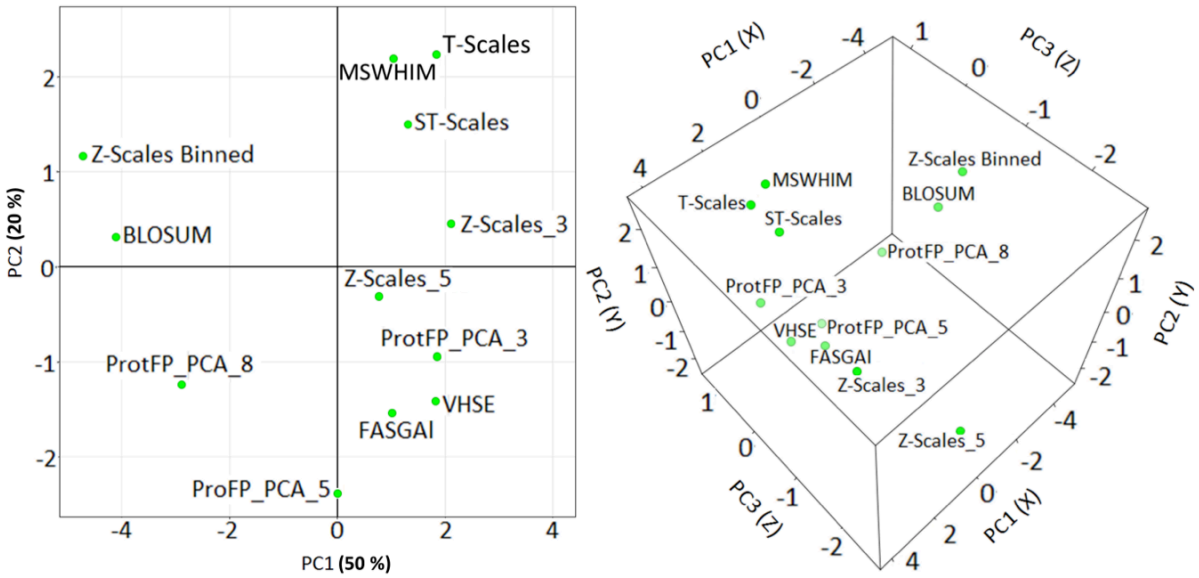
# Déjà Vu - Réjà Vu : on knowledge-based approaches linking ligand and target information to bioactivity

Westen, G.J.P. van

# Chapter 3

# Comparative Study and Benchmarking of 13 Amino Acids Descriptors and Applications to Proteochemometric Modeling




*G.J.P. Van Westen, R.F. Swier, J.K. Wegner, A.P. IJzerman, H.W.T. Van Vlijmen, and A. Bender.*

***(Manuscript submitted)***

# Contents

## 3.1 Abstract

While a large body of work exists on comparing and benchmarking descriptors of molecular structures, a similar comparison on protein descriptors has not yet been performed. Hence, in the current work a total of 13 different protein descriptor sets have been compared with respect to their behavior in perceiving similarities between amino acids, and benchmarked with respect to their ability of establishing bioactivity models. We investigate which descriptors show complementarities in behavior via principal component analysis, and secondly evaluate prediction performance in five structure-activity benchmarks. These comprise one Angiotensin Converting Enzyme inhibitor data (dipeptides), and two proteochemometric data sets (GPCR ligands and multiple GPCRs; enzyme inhibitors and multiple mutants). In describing amino acid similarities, MSWHIM, T-scales and ST-scales show similar behavior, as do VHSE, FASGAI, and ProtFP_PCA (3). The ProtFP_PCA (5), ProtFP_PCA (8), Z-Scales (Binned), and BLOSUM descriptor sets show behavior that is distinct from another and the clusters above. The use of more principal components (>3 per amino acid) leads to a significant difference in the way amino acids are described, despite capturing less variation of the original input data. In bioactivity modeling protein descriptors perform similar (< 0.2 log units RMSE difference), while the performance per protein is still highly variable. T-scales perform the best overall, while one of our ProtFP descriptors performed the worst. Here we provide a comparison of how similar (and different) currently available descriptor sets perceive amino acids to be. We conclude that in a given situation amino acid descriptors from the different clusters should be explored. This is consistent with our observation that while the performance of modeling bioactivity data using different descriptors is overall relatively similar, some descriptors still perform much better than other descriptors on a particular dataset.

## 3.2 Introduction

**3.2.1 Proteochemometric modeling.** Proteochemometric (PCM) modeling uses statistical modeling techniques to model the ligand – target space.[1-4] Related to Quantitative Structure-Activity Relationship (QSAR) modeling, PCM modeling takes both ligand- and target space into account, enabling the models to extrapolate (within limits imposed by the data sets, the descriptors and the modeling method) in both the chemical (ligand) as well as the biological (target) domain. Possible applications include receptor deorphanization, virtual screening for compounds that are selective for a single member of a target family (e.g. the adenosine receptor family), and combined modeling of orthosteric and allosteric compounds (e.g. nucleoside and non-nucleoside HIV reverse transcriptase inhibitors).[3] Hence, the target description is as important as the ligand description. While several publications are available using varying ligand descriptors, on the side of target description there is less literature available, a void we wanted to fill with the current work.[5-7] Previous PCM modeling has been performed using peptide descriptors obtained from the field of Quantitative Sequence-Activity Modeling (QSAM),[2, 8, 9] but later techniques also used different approaches in target description which did not rely on the target sequence (as is the case with QSAM descriptors) but are more structural (e.g. oriented more towards spatial descriptions of the binding site or based on known ligand – target interactions).[10-12]

**3.2.2 Utilization of Quantitative Sequence Activity Modeling (QSAM) derived descriptor sets.** QSAM attempts to quantitatively explain binding affinity of small peptide drugs to protein (or, more generally, macromolecular) targets, similar to QSAR in the field of small molecules and in this context several descriptor sets for amino acids (AAs) have been developed.[13] The majority of these descriptor sets rely on a principal component analysis (PCA) of a large property matrix used to describe the individual AAs. The data is then reduced in dimensionality *via* PCA while still describing typically over 80% of the variation present in the original set.[9] In general this leads to descriptor sets that can correlate peptide make-up with an output variable (as long as this output variable can be described in terms of individual AA properties in the first place). However, Z-scales, the most widely used descriptor set in PCM modeling was intended to be used in research for small peptide drugs and, hence, covers also non-natural AAs. This is also true for the T-scales and ST-scales. Therefore, if the original matrix consists of over 167 AAs (ST-scales) out of which only 20 are natural AAs, then it is not directly clear how large the fraction of the 'AA property space' is formed by the natural amino acids in respect to the total property space.

Hence this leads to potentially less resolution in the space we are particularly interested in modeling accurately, namely the space formed by the natural amino acids.[14] This balance between non-natural and natural AAs leads to principal components after data reduction that are not necessarily the most relevant ones to describe the natural AAs and, hence, previously developed peptide descriptor sets might not be ideal for use in PCM models. In order to capture the current state-of-the-art in describing AA (and peptide) properties, and to potentially improve upon the current situation, in this work we have benchmarked 13 previously published and four novel AA descriptor sets in order to evaluate the performance of QSAM descriptor sets in PCM.

**3.2.3 Amino acid descriptor sets considered in this study.** In the current work we have benchmarked a total of 13 different individual descriptor sets where the AA descriptor sets used belong to different broad classes (**Table 3.1**). Firstly, three descriptor sets, namely Z-scales (all versions), VHSE and ProtFP PCA (all versions), are based on a PCA analysis of physicochemical properties.[9, 15] Secondly, ST-scales and T-scales consist of a principal component analysis of mostly topological properties.[14, 16] FASGAI, part of the third category of descriptor tested is based on a factor analysis of physicochemical properties.[17] Furthermore, we also tested two descriptor sets that are calculated in a very different manner compared to the first six, namely a descriptor set based on three dimensional electrostatic properties calculated per AA (MS-WHIM).[18] Additionally, a descriptor set based on a VARIMAX analysis of physicochemical properties which were subsequently converted to indices based on the BLOSUM62 substitution matrix (BLOSUM).[19] Finally, we tested a descriptor set only describing each AA by a single feature (ProtFP Feature).[20, 21] See **Table 3.1** for an overview.

**Table 3.1.** Descriptor sets included.

| Descriptor Set | Type | Derived by | # of components | Variance explained | AAs Covered |
|---|---|---|---|---|---|
| BLOSUM | Physicochemical and substitution matrix | VARIMAX | 10 | n/a | 20 |
| FASGAI | Physicochemical | Factor Analysis | 6 | 84% | 20 |
| MSWHIM | 3D electrostatic potential | PCA | 3 | 61% | 20 |
| ProtFP (3) | Physicochemical | PCA | 3 | 75% | 20 |
| ProtFP (5) | Physicochemical | PCA | 5 | 83% | 20 |
| ProtFP (8) | Physicochemical | PCA | 8 | 92% | 20 |
| ProtFP (Feature) | Feature based | Hashing | n/a | n/a | 20 |
| ST-scales | Topological | PCA | 5 | 91% | 167 |
| T-scales | Topological | PCA | 8 | 72% | 135 |
| VHSE | Physicochemical | PCA | 8 | 77% | 20 |
| Z-scales (3) | Physicochemical | PCA | 3 | n/a | 87 |
| Z-scales (5) | Physicochemical | PCA | 5 | 87% | 87 |
| Z-scales (Binned) | Physicochemical | PCA followed by binning | n/a | n/a | 20 |

The first column contains the name of the descriptor set as used in the main text. The last column differentiates between descriptor sets only covering the natural amino acids or more. Not available is abbreviated by n/a.

**3.2.4 Summary of the comparative study of AA descriptor sets and benchmarking.** In the current work we characterize the similarity of amino acids, as perceived by each descriptor set considered in this study. Furthermore we benchmark all descriptor sets on three different data sets by constructing structure-bioactivity models and comparing their performance. The datasets are firstly a previously published set of 58 dipeptides that have an inhibitory effect on the angiotensin-converting enzyme (ACE);[22] secondly, a set of 26 GPCRs and approximately 100 active and 100 inactive compounds per receptor obtained from ChEMBL 11;[23] and finally, a set of 451 non-nucleoside reverse transcriptase inhibitors (NNRTIs) and 14 HIV mutants which was also used in a previous publication (but where the protein descriptor used was not varied).[21] It is our hypothesis that the descriptor sets based on solely the natural AAs outperform the descriptor sets created for QSAM work.

## 3.3 Materials and Methods

A more detailed outline of each descriptor set, illustrating the differences and similarities between all of them, is given below. For each descriptor set a short name used in the tables and figures is given in parentheses.

**3.3.1 Z-scales.** Z-scales are based on physicochemical properties of the AAs including NMR data and thin-layer chromatography data. Sandberg *et al.*[9] improved on the original Z-scales published by Hellberg *et al.*[24] by introducing two more Z-scales bringing the total to five scales rather than three. Sandberg *et al.* used 26 properties from 87 AAs. The PCA mainly captures lipophilicity (Z1), bulk (Z2), electrogenicity (Z3). The fourth and fifth scale (Z4 and Z5) are more difficult to interpret relating to properties as electronegativity, heat of formation, electrophilicity and hardness. The total variance explained by these five components is 87 %.

In this study we employed the Z-scales using 5 scales (***Z-scales (5)***) and the Z-scales using 3 scales (***Z-scales (3)***) both of which have been used in previous work.[25] Furthermore, the 5 Z-scales were also binned into several classes per scale (***Z-scales (Binned)***). When an AA fell within one of these bins, the bin property was set '1', otherwise it was set '0'. All natural amino acids were uniquely identifiable based on the classification.

For instance Tryptophan is assigned a '1'for the following classes: Lipophilicity High, Size Large, Electronic Properties High, Electronegativity High and Electrophilicity Low, whereas Glycine is assigned a '1' for the following: Lipophilicity Low, Size Small, Electronic Properties High, Electronegativity Medium Low and Electrophilicity Medium Low. The rationale was that these descriptors would be easier to interpret than descriptors derived from a PCA (see Supporting **Table S1** for the classes).

**3.3.2 Vectors of Hydrophobic, Steric, and Electronic properties (VHSE).** Originally published by Mei *et al.***,** Vectors of Hydrophobic, Steric, and Electronic properties (***VHSE***) are obtained from 18 hydrophobic, 17 steric and 15 electronic properties, giving rise to a total of 50 physicochemical properties of the 20 natural AAs.[15] For each of these three categories a PCA was generated and resulted in Principal Components (PC) of two hydrophobic, two steric and four electronic properties with a total variance of 74.33%, 78.68% and 77.97%, respectively. These eight properties form the VHSE scales.[15]

**3.3.3 T-scales.** Published by Tian *et al.*, the T-scale descriptor (***T-scales***) is derived from several computer programs utilized to generate 67 common topological descriptors of 135 AAs.[16] These topological descriptors are one of the most simplified descriptors since they are derived from an atom-connecting manner in 2D structures of molecules and therefore do not need an optimization of the 3D structures. A PCA calculation of the five most representative descriptors was called the T scales. These five descriptors encompass 91.14% of the total variance of the data.[16]

**3.3.4 ST-scales.** Published by Yang *et al.*, The topological ST-scale (***ST-scales***) descriptor is very similar to the T-scales, extending it by taking 827 properties into account which are mainly constitutional, topological, geometrical, hydrophobic, electronic, and steric properties of a total of 167 AAs.[14] For the ST-scales the molecular structures were first optimized as some of the properties used are conformation-dependent. ST-scale utilizes eight PCs instead of the five PCs of T-scales and describes 71.5% of the total variance of the data.[14]

**3.3.5 MS-WHIM.** Previously published by Zaliani and Gancia, the MS-Whim (***MSWHIM***) descriptor set is derived from 36 electrostatic potential properties derived from the three-dimensional structure of the molecule.[18] These are calculated from 12 statistical parameters starting from x, y, z coordinates of the Connolly surface, which is a solvent-excluded surface (an inverse solvent-accessible surface).[26] On these 36 parameters (3 coordinates by 12 parameters each) of the 20 natural AAs a PCA was performed which gave rise to a set of 3 principal components with a total variance of 61%, as well as a set of 7 principal components with a total of variance of 87%.

However according to the loading plots, the authors concluded that the most representative values were contained in the first three principal components and they hence chose to take only the first three principal components into account in their final descriptor set.[18]

**3.3.6 Factor Analysis Scales of Generalized Amino Acid Information (FASGAI).** Published by Guizhao and Zhiliang**,** Factor Analysis Scales of Generalized AA Information (***FASGAI***) is derived from 335 physicochemical properties of the 20 natural AAs.[17] Contrary to the other descriptor sets a factor analysis is applied rather than a PCA. Factor analysis also simplifies large quantities of data like PCA does, however factor analysis computes a smaller number of factors that describe the *correlated* variables, whereas PCA searches for the parameters with the largest *variance*. After generating these factors, a PCA was applied to get the factors that would describe the data with the most variance. The PCA resulted in the FASGAI protein descriptor of 6 principal components with a total variance of 83.5%.[17]

**3.3.7 BLOSUM.** Published by Georgiev, the BLOSUM matrix-derived amino acid descriptors (***BLOSUM***) is the only AA descriptor set we employed that is not directly based on physical or chemical properties of the AAs, but on both physicochemical properties that have been subjected to a VARIMAX analyses and an alignment matrix of the 20 natural AAs, the BLOSUM62 matrix (for details see the work by Georgiev).[19, 27] This procedure renders scales analogous to the Z-scales.[19] This descriptor was added due to its fundamentally different nature and an anticipated complementarity in capturing AA properties, compared to other descriptor sets.

**3.3.8 Protein Fingerprint (ProtFP).** In addition to the previously published descriptor sets, we also employed a novel AA descriptor set in this work which we termed 'Protein Fingerprint' ('ProtFP'). ProtFP is based on a selection of different AA properties obtained from the AAindex database.[28] However, the difference to descriptor sets mentioned previously is that we started with the full set of indices, while repetitively removing indices with the highest covariance. The final descriptor comes in several flavors. The first ProtFP descriptor (described in more detail below) is based on a PCA of the remaining indices employing 3, 5 or 8 principal components (***ProtFP_PCA (3)***, ***ProtFP_PCA (5)*** or ***ProtFP_PCA (8)***), which allows for quantitative comparison of AAs.

The second variation is based on a hashing approach of all indices values per AA (***ProtFP Feature***), as we published previously.[20, 21] Given the novelty of the ProtFP descriptor sets, their derivation is described in more detail in the following.

**3.3.9 Selection of AAindices (for ProtFP).** The ProtFP descriptor set was constructed from a large initial selection of indices obtained from the AAindex database for all 20 naturally occurring AAs. This is a principal difference to several other AA descriptor sets, where also non-natural AAs were taken into account.[28] Covariance between indices was determined *via* PCA and indices were normalized and scaled to a range between 0 and 1 rather than using the raw indices. The analysis was performed using the Pipeline Pilot implementation, version 6.1.5, of R-statistics and the 'prcomp' package, with the options of 'mean centering' and 'scaling' enabled.[29] Indices showing highest covariance were removed, while at the same time a number of largely independent physicochemical parameters were maintained. The final reduced selection consisted of 58 AAindices, which are hence (a) based on the relevant natural amino acids only, (b) largely independent (since those indices with larges covariance were removed). The final amino acid indices employed in the construction of the ProtFP descriptor set are listed in Supporting **Table S2**.

**3.3.10 PCA of final indices selection (for ProtFP_PCA).** In order to obtain descriptors at lower dimensionality PCA was performed on the final set of 58 amino acid properties. The analysis was performed using default parameters, requiring a minimum explained variance of 75%, but forcing a minimum of 8 principal components (PCs). The first three PCs explained 75% of the variance, 5 PCs explained 83%, and 8 PCs explained 92%. In subsequent experiments three versions were used: the first three PCs (***ProtFP_PCA (3)***), the first 5 PCs (***ProtFP_PCA (5)***) or all eight PCs (***ProtFP_PCA (8)***). See **Table 3.2** for the final principal components.

**Table 3.2.** Principal Components Resulting from the AAindex selection.

| Amino Acid | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | Feature |
|---|---|---|---|---|---|---|---|---|---|
| Variance Explained | 0.43 | 0.24 | 0.08 | 0.06 | 0.04 | 0.03 | 0.03 | 0.02 | n/a |
| Total Variance Explained | 0.43 | 0.67 | 0.75 | 0.81 | 0.85 | 0.88 | 0.90 | 0.92 | n/a |
| G | -5.70 | -8.72 | 4.18 | -1.35 | -0.31 | 2.91 | 0.32 | -0.11 | -176196525 |
| A | -0.10 | -4.94 | -2.13 | 1.70 | -0.39 | 1.06 | -1.39 | 0.97 | 1169372512 |
| C | 4.62 | -3.54 | 1.50 | -1.26 | 3.27 | -0.34 | -0.47 | -0.23 | 892384356 |
| V | 5.04 | -2.90 | -2.29 | 1.38 | 0.06 | 0.08 | 1.79 | -0.38 | -58134849 |
| L | 5.76 | -1.33 | -1.71 | 0.63 | -1.70 | 0.71 | -0.05 | -0.51 | -590269326 |
| I | 6.58 | -1.73 | -2.49 | 1.09 | -0.34 | -0.28 | 1.97 | -0.92 | -1784790725 |
| M | 5.11 | 0.19 | -1.02 | 0.15 | 0.13 | -0.30 | -2.95 | 0.50 | -188476976 |
| F | 6.76 | 0.88 | 0.89 | -1.12 | -0.49 | -0.55 | -0.87 | 1.05 | -1561345091 |
| W | 7.33 | 4.55 | 2.77 | -2.41 | -1.08 | 1.04 | 0.23 | 0.59 | -816166777 |
| Y | 3.14 | 3.59 | 2.45 | -1.27 | -0.06 | -0.29 | 1.99 | 0.30 | 1237879003 |
| H | 0.17 | 2.14 | 1.20 | 0.71 | 1.16 | -0.38 | -1.85 | -2.79 | -1970548995 |
| T | -2.00 | -1.77 | -0.70 | 1.02 | 1.06 | -1.20 | 0.74 | 1.65 | -266397547 |
| P | -3.82 | -2.31 | 3.45 | 1.00 | -3.22 | -3.54 | -0.36 | -0.30 | -576206913 |
| S | -4.57 | -2.55 | -0.67 | 1.11 | 0.99 | -1.02 | 0.11 | 0.65 | -1481898440 |
| D | -6.61 | 0.94 | -3.04 | -4.58 | 0.48 | -1.31 | 0.10 | 0.94 | 1957532765 |
| N | -4.88 | 0.81 | 0.14 | -0.14 | 1.23 | -0.65 | 1.02 | -1.94 | -1593568836 |
| E | -5.10 | 2.20 | -3.59 | -2.26 | -2.14 | 1.35 | -0.45 | -1.31 | 558044215 |
| Q | -3.95 | 2.88 | -0.83 | 0.52 | 0.90 | 0.55 | -0.08 | 0.64 | -1986194934 |
| K | -4.99 | 5.00 | 0.70 | 3.00 | -1.23 | 1.41 | 0.19 | 0.87 | 268201585 |
| R | -2.79 | 6.60 | 1.21 | 2.07 | 1.67 | 0.76 | 0.00 | 0.32 | 1636879004 |

Shown are all eight principal components and the variance explained by these principal components. In addition, the features obtained from the hashing of the AAindex selection are shown. This column represents the feature based ProtFP. Not available is abbreviated by n/a.

**3.3.11 Distance between descriptor sets.** To compare the characteristics of different descriptor sets and their behavior in describing particular AAs as similar and dissimilar, the average 'difference in distances' was calculated for each possible pair of descriptor sets. (See **Figure 3.1** for a scheme of the performed calculations). This value was obtained as follows. Firstly, a full similarity matrix was calculated for each possible AA pair using each descriptor set, thus consisting of 20*20 fields per descriptor set. The distances in this matrix were scaled linearly to a range between 0 (most similar) and 1 (most dissimilar).

Subsequently, for each possible pair of *descriptor sets* the *difference* between the *Euclidian distances of each AA pair* was calculated, giving rise to a total of 400 inter-amino acid distance differences per descriptor set pair. (In other words, we evaluated how differently two descriptors judged the difference between two AAs.

Given that 20 AAs exist, 400 distances exist between all AAs, for a single descriptor set – and the same number of *differences* of those distances for each descriptor set pair.) Of the 400 distances obtained, the average distance and the standard deviation was calculated and subsequently employed as a measure for the distance between amino acid descriptor sets (*i.e.*, if the average distance is high, two amino acid descriptor sets perceive similarities between amino acids in a very different way). The more different those distances are for different descriptor sets, the more different the particular descriptor sets considered behave. We employed a total of 12 descriptor sets for this amino acid descriptor comparison, since the feature based ProtFP descriptor set (***ProtFP (Feature)***) merely uses presence or absence of features and hence could not be included in the distance calculation. In the end, a matrix of 12*12 distances between descriptor sets was obtained which was subject to PCA with the aim to visualize the individual distances between descriptor sets in a graphical way.

(Conceptually, this work is similar to an analysis of chemical descriptors from the ligand side which was performed previously and given the importance of also comparing descriptors from the protein side the current work hence complements this study [30]).

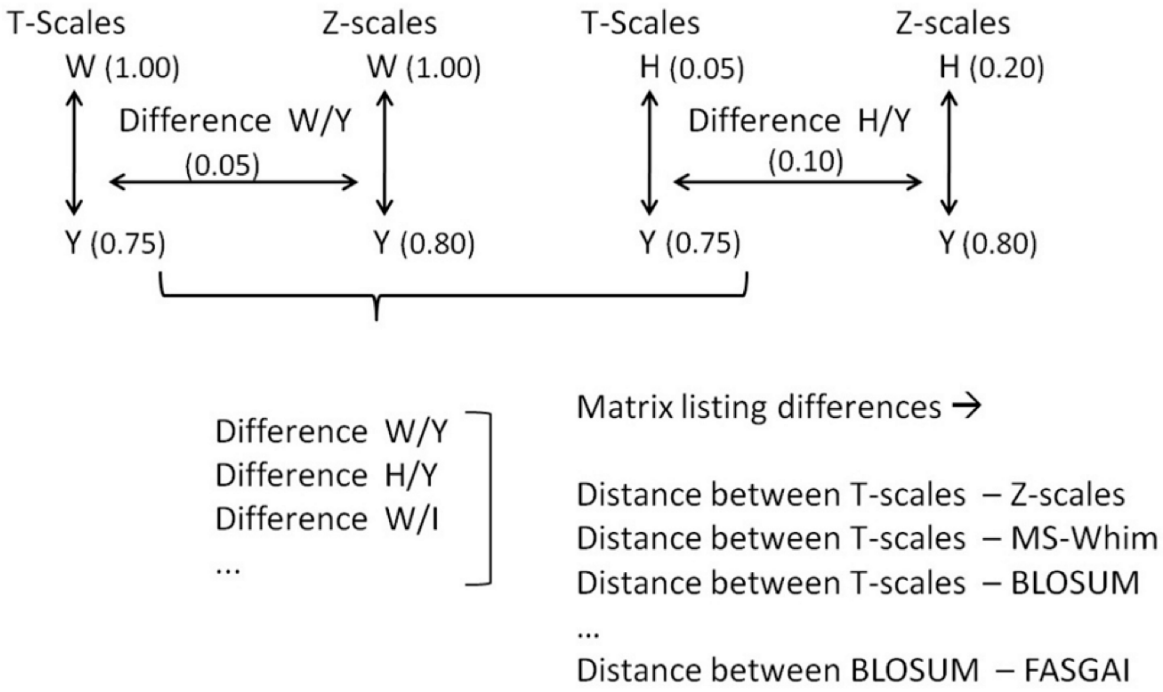



**Figure 3.1: The approach used to characterize descriptor set distances and similarities.** After normalization of all descriptor sets, the difference between a pair of descriptor sets was calculated. This difference was obtained as the difference between the distance separating pair of AAs in descriptor 1 and the same pair in descriptor 2. This was done for all descriptor set pairs. Finally, the average difference was obtained and a full matrix was constructed.

**3.3.12 Benchmark datasets for different descriptors.** While analyzing similar and different behavior of AA descriptor sets is relevant to judge *how similarly* two descriptor sets behave, it does not yet give any information how relevant the information captured by a particular descriptor would be for the generation of bioactivity models. Hence, in order to assess the performance of each descriptor set, three different data sets were used to perform a number of benchmark experiments.

**ACE inhibitor data set.** The first set consisted of 58 dipeptides with a measured ACE inhibiting effect ($pIC_{50}$) and was obtained from literature.[22] The set serves as a benchmark as several of the descriptor sets analyzed here were applied to this set in their original publication. Hence, it can demonstrate that the method we use (Random Forest) performs *on par* or better than the PLS which is conventionally used in QSAM publications (see Supporting **Table S6** for the comparison). See **Table 3.3** for further details about the data set.

**Table 3.3.** The Data Sets Used for the Bioactivity Benchmarks.

| | **ACE Inhibitors** | **GPCRs** | **NNRTIs** |
|---|---|---|---|
| Total Size (Data Points) | 58 | 4,951 | 4,024 |
| Total Compounds | n/a | 3,088 | 451 |
| Average Compound Tanimoto Distance (ECFP_6) | n/a | 0.89 | 0.02 |
| Average Euclidian Distance Compounds (Physicochemical) | n/a | 1.31 | n/a |
| Total Targets (Peptides / Proteins) | 58 | 26 | 14 |
| Average Target Tanimoto Distance (ProtFP (Feature)) | 0.83 | 0.26 | 0.14 |
| Average Euclidian Distance Target (ProtFP_PCA (3)) | 1.35 | 0.89 | 0.47 |
| Completeness (% of total compound - target pairs) | - | 0.06 | 0.64 |

**GPCR data set.** The first bioactivity data set employed for benchmarking different amino acid descriptors in PCM modeling comprised a subset of 26 human monoamine receptors (class A GPCRs listed in Supporting **Table S1**; see also Supporting **Figure S1** regarding the subset of receptors used) obtained from ChEMBL version 11.[23] Receptors were selected only if more than 120 unique ligands with annotated activity were present in ChEMBL. The trans-membrane (TM) binding site was defined according to Gloriam *et al.* and all residues selected were subsequently subject to conversion into numerical values using all protein descriptor sets listed above.[31]

For each of the 26 receptors included in this study all small molecules with an affinity on this receptor available in ChEMBL were selected and further narrowed down to only include Ki annotations with high confidence score (9). Compounds were then classified as 'active' (pKi > 7) or 'inactive' (pKi =< 7). Finally compounds were clustered (using the ECFP_6 fingerprint used to train the models) to obtain a total of 100 chemically diverse 'actives' and 100 chemically diverse 'inactives' per receptor. Compounds were standardized and ionized at pH 7.4 in Pipeline Pilot 8.5.[32]

In total 3,088 distinct compounds were selected to generate a bioactivity model, including 1,863 compounds with measurements on multiple GPCRs, hence leading to a final dataset comprising 4,951 ligand-protein data points (corresponding to 6 % of the total of 80,288 possible compound – receptor combinations in the full matrix of 3,088 compounds and 26 targets; see **Table 3.3** for further details)

**NNRTI data set.** The second bioactivity data set where PCM modeling was applied comprised of 14 mutants of HIV Non-Nucleotide Reverse Transcriptase Inhibitors (NNRTIs) and 451 compounds, hence a total of 6,314 possible compound – receptor combinations out of which for 4,024 a $pEC_{50}$ value was available (66% of the total).[21] The compounds in this case were structural analogues, and hence (as opposed to the GPCR case) the average similarity between the compounds was high, as was the similarity between the protein targets since those were HIV mutants carrying 1 to 13 point mutations. Like in our previous work, the binding site was defined as those AAs that differed between the different mutants (a total of 24 residues).[21] The HXB2 / IIIB reference strain was defined as the wild type (See **Table 3.3** for further details).[33]

**3.3.13 Amino Acid descriptor set benchmarking.** Two different approaches were pursued to benchmark AA descriptor sets with respect to their ability to generate bioactivity models (and hence, to capture protein information relevant to bioactivity and ligand binding); namely 70-30 validation and Leave-One-Sequence-Out (LOSO) which are described in the following.

**70-30 validation.** The primary benchmark was a 70-30 validation experiment. Each descriptor set was used in turn in combination with each of the datasets, and a model was trained on a random 70% of the data available and used to predict the bioactivities of the remaining 30% of the data. This procedure was repeated three times and from the resulting validation parameters the average and standard error of the mean (SEM) was calculated. For the ACE inhibitors this represented a particularly challenging benchmark as this set only includes peptides and no small molecules. For the bioactivity datasets employed for PCM modeling, since these data sets include both proteins and small molecules, this benchmark provides an answer to two different questions.

Firstly, the model was asked to make bioactivity predictions for those compounds that are not present in the training set and hence to extrapolate in the *chemical domain*. This part of the validation was particularly emphasized in case of the GPCR data set due to the low average compound similarity. Hence the model is asked to extrapolate the activity of known compounds and targets to *unknown compounds*.

Secondly, a compound can be present in the training set as annotated on one target, and also be present in the test set as annotated on target 2. This part of the validation was hence emphasized in case of the NNRTI data set due to the high average compound similarity. In this case the model is asked to extrapolate the activity of known compounds and targets to *unknown combinations of the two*, while, individually, each chemical structure and sequence have been seen by the model before (but just not in this particular combination).

**Leave-one-sequence-out validation.** This validation experiment was performed for each *target* in order to assess extrapolation abilities of the PCM models in the biological / target domain. Hence this validation was only applied to the datasets containing targets (GPCR set and NNRTI set). This step is analogous to leaving out ligands from a dataset in more conventional bioactivity modeling – however, since PCM models are also able to extrapolate in the *biological domain* we also need to perform this additional validation here.

In this part of the work, repetitively a single target is left out of the training set and subsequently a model is trained on all bioactivity data points, except for those of the target in the test set, that was left out. Afterwards activity values of all compounds on the target left out of the initial training procedure are predicted and compared to the experimental values.

Again this procedure is repeated three times and the average and SEM were calculated of the validation parameters. These steps are repeated for all targets in the data set in turn. This type of validation is a specialty of PCM modeling since it takes advantage of its ability to extrapolate *also in target space*. It resembles both the real-world situation of deorphanizing receptors, taking only information from related proteins into account and attempting to identify bioactive chemical matter for a receptor for which no ligands have been identified yet.[34, 35]

(Also this concept is applicable to predict which drug to use against a particular receptor mutant in case of *e.g*. personalized medicines, such as in case of the question which drug to use against a particular HIV patient genotype which is addressed also in this work.) Since the ACE inhibitor set consisted of bioactive compounds only, LOSO could not be performed on this set.

**3.3.14 Compound Descriptors.** Ligands were described using ECFP_6 circular fingerprints,[36] which take into account the number of connections to an atom, the element type, the charge, and the atomic mass. These descriptors have previously been shown to perform well in comparative virtual screening studies.[30] This ligand side descriptor was employed for all studies presented in this work containing small molecules. Here an array size of 512 bits (each bit corresponding to a chemical substructure) was used.

In addition, in the GPCR data set compounds were described by their physicochemical properties. These properties were binned into classes, when compounds met one of these classes the property was set as '1', when they did not the property was set as '0' (analog to the Z-scales (Binned) descriptor). The classes that were used are available in Supporting **Table S4** and **S5**.

**3.3.15 PCM Modeling Method.** Both regression and classification models were generated in Pipeline Pilot Version 8.5 using the R-statistics modeling package version 2.12.1.[29, 32] Modeling was performed using the 'forest' package in R Statistics.[37] The size of the forest was experimentally determined to be optimal at 500 trees, the maximum number of descriptors allowed for each tree was set at a fraction 0.5 of the total number. Class size equalization was turned on and a performance estimate during training was obtained using out-of-bag validation. Furthermore data points were fed into the model in a randomized order (differing between repeats of an experiment) to get a more reliable performance estimate.

**3.3.16 Model validation.** To validate our models different parameters were employed depending on the modeling type. In regression models both the Root-Mean-Square Error (RMSE) and the correlation coefficient intersecting the origin ($R_0^2$) were employed.[38] For the classification models the Matthews correlation coefficient (MCC) was used to estimate model performance because of its robustness and the fact that it incorporates both correct and false predictions.[39] However, because of the importance of models to actually retrieve active compounds, we employed model sensitivity as a second performance measure.

**3.3.17 Y-Scrambling.** To make sure that the models created were not based on chance correlations, Y-scrambling or permutation testing was performed. These studies were performed using the same setup as the benchmark experiments (also in triplo) however the modeled variable ($pIC_{50}$, $pEC_{50}$ or activity class) was randomized over the data points. Hence no correlation should exist between the descriptors (ligand and target) and the activity. The results are shown in Supplementary **Figures S36 – S40** and confirm that no predictive models can be trained on this randomized set.

**3.3.18 Descriptor Ranking.** Finally, to obtain a broadly derived performance measure we ranked all 13 amino acid descriptor sets based on their performance per dataset and experiment. This rank-based assessment prevents a single dataset that is modeled very well or very bad (as expressed in RMSE or MCC) unduly influencing the average performance of this descriptor set. Descriptor sets were ranked using the validation parameters ($R_0^2$ and RMSE in the case of regression and MCC and Sensitivity in the case of classification), the final rank per experiment is the sum of both validation ranks. For example in the ACE inhibitor set each descriptor would receive a rank based on the RMSE and one based on the $R_0^{2,}$ the final rank can hence be anywhere between 2 (best score on both validation parameters) and 26 (worst score on both validation parameters). Subsequently the descriptors were re-ranked between 1 and 13 to provide a final rank that could be compared over all three data sets.

## 3.4 Results and Discussion - Section 1 – Similarity between descriptor sets

The first part of our work covers the characterization of descriptor similarity between all benchmarked descriptor sets. Furthermore, we show how ProtFP based descriptor sets were derived.

**3.4.1 PCA of final indices selection (ProtFP_PCA). Figure 3.2A** shows the first two principle components of all 20 natural AAs when employing the ProtFP descriptor set. Overall, the plot shows a general clustering of AAs with similar properties with the first PC corresponding to hydrophobicity (F and I score high whereas D and E score low) and the second PC corresponding to size (W and K score high whereas G and A score low). Noteworthy is the clustering of Leucine and Isoleucine, which is intuitively correct due to their high chemical similarity, however not reproduced by all AA descriptors, like ST-scales (Supporting **Figure S17**). Furthermore, both charged (D, E and R, K) and aromatic residues (F, H, Y, W) cluster together. (The principle components, representing each AA in ProtFP space, can be found in **Table 3.2**.) Hence, overall the ProtFP descriptor set produces a clustering pattern that looks correct from a chemical point of view.

**Figure 3.2B** shows the loadings plot of the first 2 PCs that represent the ProtFP descriptor set. (For a complete list of indices used as input for the PCA please see Supporting **Table S2**.) Here, some interesting observations can be made. For instance, scale 24 and 43 correspond to AAindex FAUJ880112 and MONM990201, respectively. While the former is a measure for negative charge, the latter is a measure for 'averaged turn propensities in a transmembrane helix'. These two properties are close neighbors based on the first two components; however they have a relatively large distance in the third PC. This is interpretable in the following way: it is likely that charged residues, if present in a transmembrane region, initiates a turn and is therefore located at the edges of the TM region. Hence the clustering of these indices together can be rationally explained.

Scales 36 and 39 are another interesting case. The former corresponds to AAindex LEVM760102 (Distance between C-alpha and centroid of side chain) and the latter corresponds to LEVM760105 (Radius of gyration of side chain). It is interesting to see that these two indices end up so close together in the first, second and third principal component. However, this is indeed expected as the maximal range of gyration can only be large if the maximal distance possible between C-alpha and side chain center is large and vice versa.

In conclusion, the division of the AA over the principal component space seems interpretable and in agreement with biochemical intuition; this applies both to the scores and the loadings plot of the PCA we performed. The next step is to compare the new descriptor set ProtFP to existing descriptor sets that have previously been published, both with respect to their ability to capture similarities of AAs and their relative performance in incorporating protein information relevant to bioactivity into SAR models.

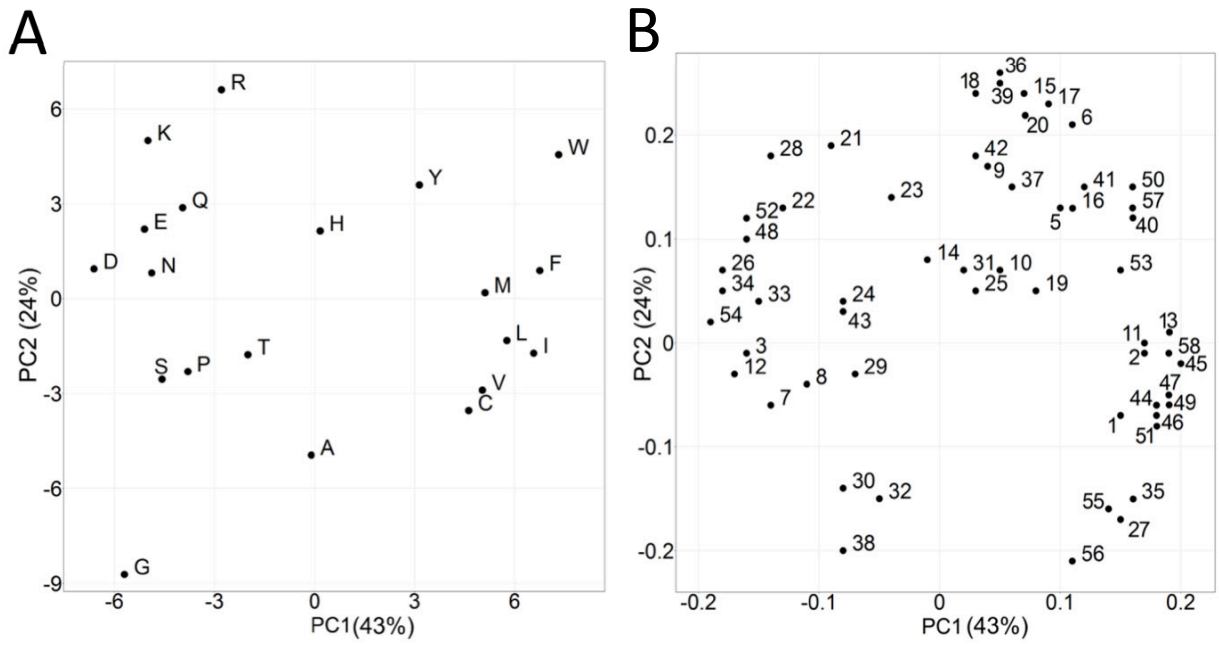



**Figure 3.2: Principal components resulting from the PCA on 58 AAindices.** (A) AAs that share physicochemical properties cluster together. The amount of variance explained by each principal component is shown in brackets. (B) The corresponding loadings plot where the numbers correspond to Supporting **Table S2**.

**3.4.2 Distance between descriptors.** Our first aim of the current study was to compare the behavior of AA descriptor sets, in order to investigate which descriptor sets agree on grouping AAs as similar, and which ones show largely orthogonal behavior. For this purpose, employing each of the AA descriptor sets a Euclidian distance based similarity matrix of all 20 by 20 AAs was calculated and visualized in a heat map. The comparison of ProtFP_PCA (3) with the frequently employed Z-scales (3) is shown in **Figure 3.3**. (The analogous plots, as well as numerical descriptions of the similarity matrices of other AA descriptor sets, are provided in Supporting **Tables S7** to **S18**, as well as Supporting **Figures S2** to **S13** for utilization by the reader in potential future studies).

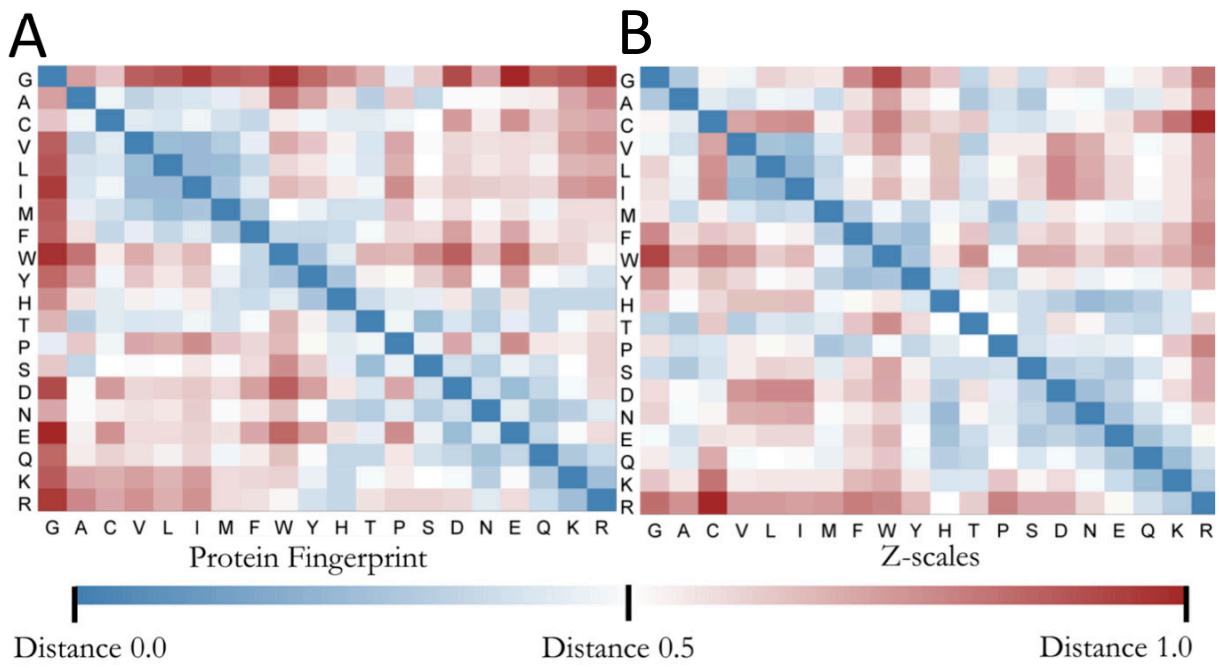



**Figure 3.3: Comparison of the distances between individual AA pairs.** (A) The heat map resulting from the ProtFP_PCA (3) similarity matrix. (B) The heat map resulting from the Z-scales analysis. In particular Histidine and Cysteine show a different distance spectrum when their similarity to the other AAs is compared.

Several clear differences are noteworthy when comparing the two descriptor sets. Firstly, the overall distances in the ProtFP_PCA (3) heat map are larger compared to Z-scales (3) despite the scaling that was applied. Furthermore, Glycine is located further away from the rest of the amino acids. Conversely, Cysteine is located closer to the aliphatic and aromatic AAs, but further away from the charged residues. Finally, Histidine also displays a different profile as it has a central position between the charged residues and aromatic residues in ProtFP PCA (3), whereas it is closely located to the charged AAs in Z-scales. Both descriptor sets therefore interpret the physicochemical space differently, while both views can be rationalized, benchmark experiments are needed to determine which leads to more predictive models.

As a next step it was considered how similar, on average, two descriptor sets perceive any pair of AAs, in order to establish how correlated their similarity perceptions are. **Figure 3.4A** shows the results of the PCA of the average distance between all descriptor sets, hence capturing the similarity in behavior of different AA descriptors. Shown are the 2 first PCs that explain 70 % of the variance. The first thing noteworthy is that MSWHIM, T-scales and ST-scales cluster together (here in the upper right quadrant); similarly, VHSE, FASGAI and ProtFP_PCA (3) form a second cluster (here in the lower right quadrant). The space between these two clusters is occupied by Z-scales (3) (upper right) and Z-scales (5) (lower right). ProtFP_PCA (5) and ProtFP_PCA (8) occupy the lower left quadrant but do not cluster. Finally Z-Scales (Binned) and BLOSUM behave distinctly from all descriptors above, and occupy the upper left quadrant. The distance between Z-scales (5) and Z-scales (Binned) is very large, which was not expected as one is constructed from the other. It could be speculated that the division into bins maximized separation between amino acids that only differ slightly on a continuous scale explaining the very different behavior. **Figure 3.4B** shows the results of the same PCA in three dimensions; now we observe that ProtFP_PCA (3) and Z-scales (5) are in addition to dissimilarities in the first two dimensions also out of the plane of the other descriptors.

The same calculation was repeated using only the absolute distance based on the first two PCs, comparing the descriptors based on the first two dimensions and minimizing the differences generated by a larger set of dimensions (Supporting **Tables S19 – S26** and **Figures S14 to S21**). Since we only use the first PCs the different versions of ProtFP_PCA are identical as are the versions of Z-scales. Again shown are the first two PCs which explain 66 % of the variance. Surprisingly, all descriptor sets based on a PCA of physicochemical properties form a cluster in this case (ProtFP PCA, VHSE and Z-scales), as do the two descriptors based on a topological description (T-scales and ST-scales). Contrarily, the MS-WHIM descriptor behaves most dissimilar to the others, likely due to the fact that this was the only descriptor constructed on an electrostatic potential. Finally, at first it seems surprising that the BLOSUM derived descriptor and the FASGAI descriptor are nearest neighbors in the first two principal components. However, in the 3rd principal component there is a large distance between the two points, rationalizing the difference.

Our results indicate that the different descriptor sets indeed describe the AA space differently, although there are commonalities most often based on the way they are constructed. What can be observed overall is that the use of more principal components (>3 per AA for a particular descriptor set) leads to a significant shift in the way they describe the AA differences.

This is true even while these principal components typically capture less variation of the original underlying matrix on which they were constructed. Therefore it stands to reason that the use of more than 3 principal components per AA might introduce less signal than noise (based on the small amount of variation captured by these components). Since the descriptor sets cluster mainly in the first 2 principal components of the descriptor analysis, these could be used as a guideline to determine complementarity when selecting descriptors to be used in bioactivity modeling (e.g. select one from each quadrant). Another observation is that the descriptor sets here introduced add novelty as they characterize the AA space differently. Assessing similarity in behavior is one aspect of comparing AA descriptor sets, in order to get an idea of the performance of the descriptor sets in this context we have set up several benchmark data sets as described in the following.

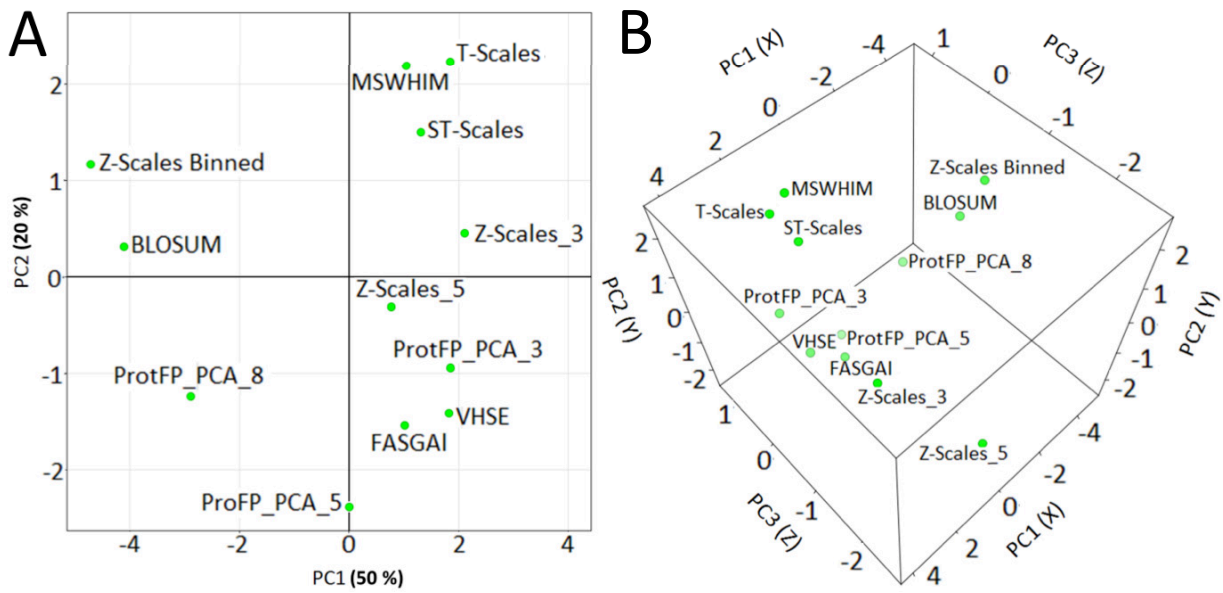



**Figure 3.4: Principal component analysis of the distances between the different descriptor sets.** Shown are the first two components (A). ProtFP_PCA (5) and (8) are seen to cluster away from the others. Furthermore T-scales, ST-scales and MSWHIM cluster together. (B) When the first three PCs are displayed Z-scales (5) and ProtFP_PCA (3) are seen to be distant from their cluster in the first two PCs.

## 3.5 Results and Discussion - Section 2 – Descriptor set performance in bioactivity models.

The second part of our work covers the assessment of descriptor set ability to create bioactivity models.

**3.5.1 ACE inhibitors (70-30).** The first benchmark we performed was a 70-30 validation experiment where ligands were dipeptides inhibiting ACE and where a random 70% of our data set was used for training and 30% for testing. The results of this validation on the test set are shown in **Figure 3.5.** The figure shows that all descriptor sets are capable of capturing the bioactivity space of the peptides as all have a RMSE under 0.8 log units. Interestingly, the best performing descriptor set is the Z-scales (Binned) descriptor (RMSE is 0.40 log units and the $R_0^2$ is 0.84), closely followed by the T-scales (RMSE 0.44 log units and $R_0^2$ 0.86) and the Z-scales (3) (RMSE 0.41 log units and $R_0^2$ 0.78). The worst performing descriptor set is ProtFP (Feature) (RMSE 0.74 log units and $R_0^2$ 0.55), which is not surprising as it does not capture the different degrees of similarity between AAs, only that they are not the same. The ProtFP_PCA descriptor sets are performing better than ProtFP (Feature) but are still lagging compared with the other descriptor sets (RMSE approximately 0.10 log units higher and $R_0^2$ approximately 0.10 lower). The numerical values for the RMSE and $R_0^2$ are included as Supporting **Table S6**, also shown there are the training parameters $Q^2$ and cross validated RMSE (CV_RMSE) which are compared to values from previous studies for the same descriptors on the same set. We have constructed a PCA analysis of the similarity space formed by the dipeptides to explain the differences in behavior we observe. We hope to gain further insight in descriptor set performance by investigating how these descriptor sets characterize the different peptides.

**3.5.2 ACE Inhibitors (Activity Space).** We plotted the first two principal components for each descriptor set and colored the points by their $pIC_{50}$ values (Supporting **Figure S22** – **S24**). We observe a direct correlation in the Z-scales (Binned) descriptor set between location in PCA space and activity. High affinity peptides score negatively on PC2, whereas all marginally active compounds score 0 or higher. Clearly the way the descriptors characterizes the peptides corresponds to their bioactivity. Conversely, the pattern obtained from the ProtFP (Feature) descriptor set does not clearly separate actives and inactives, explaining the poor performance. Well performing descriptor sets T-scales and Z-scales (3) and FASGAI also display a clustering similar to Z-scales (Binned). The PCA shows the highly active peptides to cluster together and the lesser actives are separated from these actives.

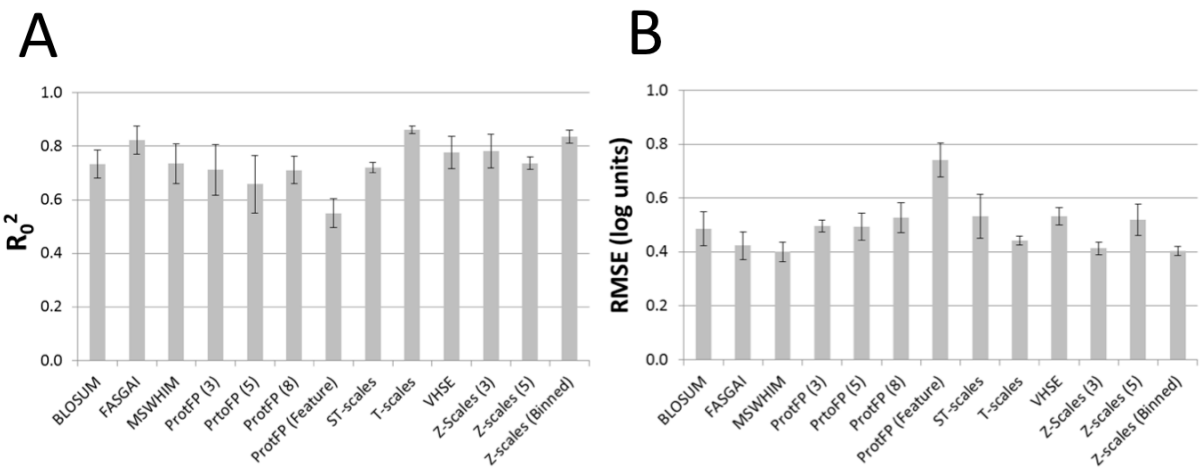



**Figure 3.5: The average performance of the benchmarked descriptor sets in the ACE inhibitors 70-30 validation experiments.** The average is calculated over three different experiments and the error bars represent the SEM. Shown are the $R_0^2$ (A) and the RMSE (B). While all descriptor sets perform similar, Z-scales (Binned) performs the best, followed by the T-scales, and ProtFP (Feature) performs the worst.

**3.5.3 ACE inhibitors (Conclusions).** We conclude that the differences in performance can be explained from the characterization of the peptides by each descriptor set as shown in the PCAs. In addition, we show that we can recreate models based on the individual descriptor sets that are comparable or better than previously published work. Finally, each descriptor set describes the AA space differently (as we have also shown in section 1). Still all were able to capture the bioactivity space and we therefore choose to apply these descriptor sets to PCM sets to see how well they perform.

**3.5.4 GPCR ligands (70-30).** Like we did with the ACE inhibitors, a similar 70-30 validation was performed on the GPCR set, although here a classification model was employed and performance was expressed as average sensitivity and MCC for all descriptors in the study (details are visualized in **Figure 3.6**). Here the descriptor sets perform much closer to each other compared to the ACE inhibitor set (all MCC values lie within the 0.35 – 0.40 range and all sensitivity values between 0.69 - 0.72), which is likely due to the much higher similarity of the targets and hence smaller space. Furthermore, the descriptor sets describe a smaller part of the space that is actually modeled since we now also include the chemical space next to the target space.

The best performance has been obtained in this case by the T-scales (MCC 0.39 and sensitivity 0.72), followed by Z-scales (5) (MCC 0.39 and sensitivity 0.71) and ProtFP_PCA (8) (MCC 0.39 and sensitivity 0.71). ProtFP (Feature) performs the worst (MCC 0.36 and sensitivity 0.69), but the difference is smaller than it was in the ACE inhibitor experiments. Another interesting observation is that all descriptor sets performed the best on the dopamine D5 receptor and the worst the histamine H3 receptor, irrespective of the protein descriptor set selected (Supporting **Figure S28**; although absolute differences in performance could be observed). These two receptors were also modeled the best and the worst respectively in the LOSO experiments as discussed in the following (where also a discussion of the likely underlying reason is given).

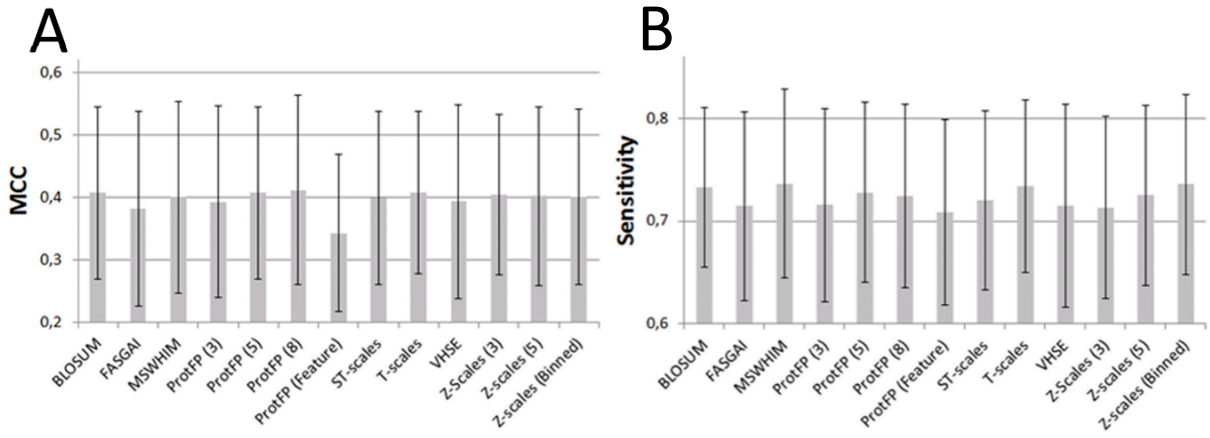



**Figure 3.6: The average performance of the benchmarked descriptor sets in the GPCR 70-30 validation experiments.** The average is calculated over all 26 receptors (performed in triplicate) and the error bar represents the SEM (note that error bars are large due to different performance between models, not between repeats of the individual models. Also see supporting **Figure S28**). Shown are the MCC (A) and the sensitivity (B). The differences between individual descriptor sets are smaller than in the ACE inhibitor experiments, likely due to the fact that models are based on both chemical and protein similarity. For individual receptors larger performance differences occur (main text). Still T-scales (3) performs the best and ProtFP (Feature) again performs the worst.

**3.5.5 GPCR Ligands (LOSO).** In order to benchmark the extrapolation capabilities of the descriptor set we performed a Leave-One-Sequence-Out experiment on the GPCR dataset, the results of which are shown in **Figure 3.7**. The overall performance is similar to the 70-30 validation but slightly worse (MCC between 0.29 – 0.32 and sensitivity between 0.57 – 0.60). However there are some differences, the best performance is by the Z-scales (3) (MCC 0.32 and sensitivity 0.59), followed by the ProtFP_PCA (5) (MCC 0.31 and sensitivity 0.60) and Z-scales (5) (MCC 0.32 and sensitivity 0.58). Surprisingly, the worst performance in this experiment is by ProtFP_PCA (8) (MCC 0.29 and sensitivity 0.57), yet it should be noted that the differences are marginal. Interestingly, the receptor that is modeled the best is again the dopamine D5 receptor and the worst the histamine H3 receptor, irrespective of the protein descriptor set selected (Supporting **Figure S29**). To gain a further understanding of this constant good performance for the D5 receptor and bad performance of the H3 receptor, we performed a PCA analysis analogously to the ACE inhibitors but then applied to the GPCR binding site sequences.

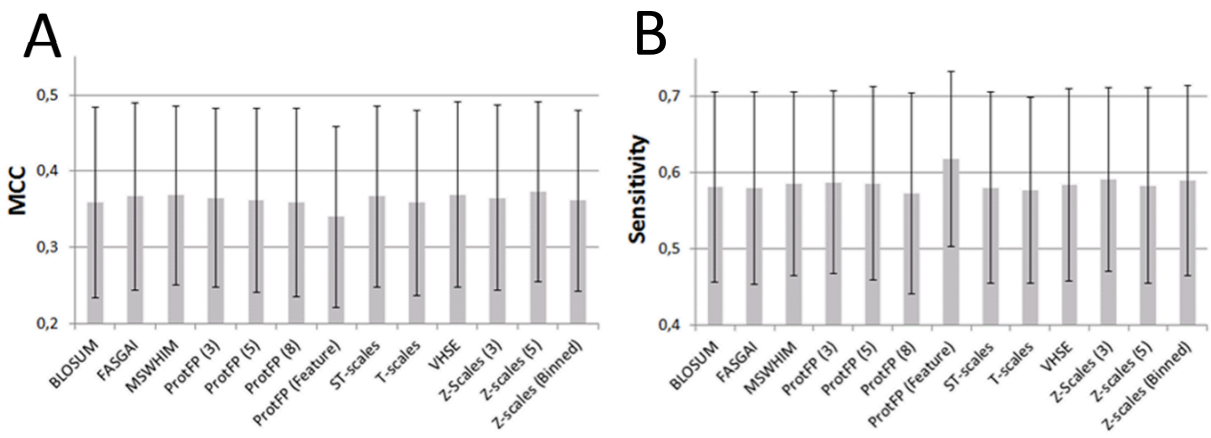



**Figure 3.7: The average performance of the benchmarked descriptor sets in the GPCR LOSO validation experiments.** The average is calculated over all 26 receptors (performed in triplicate) and the error bar represents the SEM (note that error bars are large due to different performance between models trained on different GPCRs, not between repeats of the individual models. Also see supporing **Figure S29**). Shown are the MCC (A) and the sensitivity (B). Here extrapolation takes place on the target side as the test set contains unseen targets. The differences between individual descriptor sets are still small. Again for individual receptors larger performance differences occur (main text). Now, Z-scales (3) performs the best and ProtFP_PCA (8) performs the worst.

**3.5.6 GPCR Ligands (Target Space).** From the PCA analysis of target space we can rationalize the poor performance on the histamine H3 receptor (Supporting **Figures S25** – **S27**). In the PCA of all GPCR targets used in this dataset, and employing the different descriptors, the H3 receptor is located at the edge of the PCA space. Furthermore, the three histamine receptors do not cluster together; in some cases the H3 receptor is located close to the H4 receptor, while in others it shows SAR that is closer to the H1 receptor. It is therefore likely that the models are unable to reliably extrapolate for this receptor based on the other two histamine receptors. Leaving out the H3 receptor removes crucial information from the SAR that cannot be compensated by the other two histamine receptors.

Conversely, the other receptor subtypes (5HT2, beta-adrenergic, and acetylcholine receptors) form clear sub-clusters, which hence allow leaving one receptor out while still retaining much information about the receptor space of that particular protein family. The well-performing dopamine D5 receptor on other hand is located at the center in all cases (always clustered with the other dopamine, 5HT1 and alpha-adrenergic receptors). Leaving this receptor out can therefore be considered straightforward as the target space is well covered

**3.5.7 GPCR Ligands (Conclusions).** We can conclude that all different descriptor sets can be used to create predictive PCM models on this set while still showing an order of (descending) performance as follows: Z-scales (5), ProtFP_PCA (3), T-scales, Z-scales (3), Z-scales (Binned), and BLOSUM (the latter two rank equal). The worst 3 are (descending): FASGAI, ST-scales, and ProtFP (Feature). Furthermore we can conclude that the binding site definition used for the GPCR descriptors is not optimal for all receptors. While the dopamine, 5HT1 and alpha-adrenergic are modeled very well, the histamine receptors clearly suffer, it would therefore be advisable to model these receptors with a different binding site definition. A starting point could be the work by Surgand et al. that also formed the basis for the paper by Gloriam *et al.* however Surgand *et al.* distinguish based on receptor family where Gloriam *et al.* produce a global selection.[40]

**3.5.8 NNRTIs (70-30).** While the above GPCR ligand dataset was based on rather diverse ligands, the NNRTI dataset employed in this study covers a more neatly defined area of both chemical (ligand) space, as well as biological (target) space. The first step is again a 70-30 validation experiment to assess the ability of the different descriptor sets to capture the ligand – target interaction space. The results are shown in **Figure 3.8**. Similar to previous experiments on the GPCR set, the performance of the descriptor sets is very similar (RMSE in the range 0.43 – 0.47 and $R_0^2$ in the range 0.56 – 0.61). However, in this set the ProtFP (Feature) performs the best (RMSE 0.43 and $R_0^2$ 0.61), followed by MSWHIM (RMSE 0.44 and $R_0^2$ 0.61) and Z-scales (3) (RMSE 0.44 and $R_0^2$ 0.60). The worst performance comes from (descending) VHSE (RMSE 0.45 and $R_0^2$ 0.59), BLOSUM (RMSE 0.45 and $R_0^2$ 0.58), and ProtFP_PCA (8) (RMSE 0.46 and $R_0^2$ 0.56).

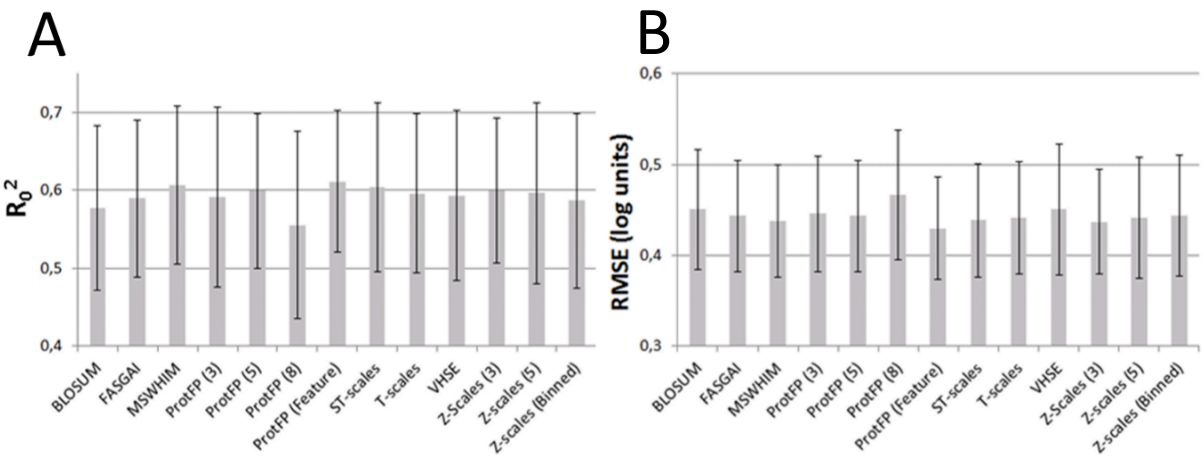



**Figure 3.8: The average performance of the benchmarked descriptor sets in the NNRTIs 70-30 validation experiments.** The average is calculated over all 14 mutants (performed in triplicate) and the error bar represents the SEM (note that error bars are large due to different performance between models trained on different mutants, not between repeats of the individual models. Also see supporting **Figure S30**). Shown are the $R_0^2$ (A) and the RMSE (B). Slightly more variance is seen compared to the GPCR experiments. In this case ProtFP_PCA (8) again performs the worst, while ProtFP (Feature) performs the best.

When focusing on the individual mutants (Supporting **Figure S30**), the best performing mutant is sequence 9 (carrying solely the K103N mutation, which is hence well covered in the remaining training set). All descriptor sets with the exception of BLOSUM are able to model the fraction of the compounds left out with an RMSE of < 0.3 log units on this mutant. The mutant that is modeled the worst is surprisingly not the heavy mutant sequence 7 (which contains a number of 13 total mutations), but rather sequence 2 carrying only two mutations (V179F and Y181C).

When comparing the predicted to the experimentally obtained bioactivities it can be seen that the bad performance is mainly caused by a number of outliers on the extremes. V179F is known to have a high impact on the class of compounds modeled here and the mutation itself (from valine to phenylalanine) is also a large change. Furthermore, this mutation was identified as having the most effect on binding in previous work.[21] The combination of these factors could explain the performance of all descriptors on this sequence. Still it should be noted that there are individual differences between descriptor sets (RMSE ranges between 0.54 – 0.66 and $R_0^2$ between 0.14 – 0.35). The next step we performed was to investigate whether results were transferable to the LOSO experiment, when extrapolation abilities to entirely novel sequences were required.

**3.5.9 NNRTIs (LOSO).** The LOSO validation was performed similar to the GPCR LOSO validation, leaving out one sequence at a time in training and predicting the activity of compounds on the sequence left out. The results are shown in **Figure 3.9**. The best performance is by BLOSUM (RMSE 0.73 and $R_0^2$ 0.66), followed by ProtFP_PCA (3) (RMSE 0.73 and $R_0^2$ 0.66) and ProtFP_PCA (5) (RMSE 0.73 and $R_0^2$ 0.66), while the differences virtually absent. ProtFP (Feature) performs very well based on the RMSE (0.65), but based on the $R_0^2$ ranks 10th (0.64) and hence ranks 5th overall. The worst performance is obtained by (descending) Z-scales (3) (RMSE 0.75 and $R_0^2$ 0.66), MSWHIM (RMSE 0.75 and $R_0^2$ 0.64) and Z-scales (5) (RMSE 0.77 and $R_0^2$ 0.64). Noteworthy is that, while the average RMSE rises to 0.7 log units, the average $R_0^2$ remains over 0.6 for all descriptor sets. This indicates that the descriptors are introducing an absolute error in the predictions, while still in most cases being able to accurately rank the compounds relative to each other.

The mutant modeled the best was sequence 3 (carrying only the Y181C mutation which is present multiple times in the data set, Supporting **Figure S31**). The sequence modeled the worst was sequence 8 (carrying K101P). This sequence was also modeled the worst in previous work.[21] The cause is likely that this particular mutation only occurs in sequence 7 and 8. Since sequence 7 is a heavy mutant, the model is unable to deconvolute the contribution of K101P tot the total effect on lowered binding of inhibitors. It is striking that ProtFP (Feature) performs so much better on this data set than the other two sets. On the NNRTI set, ProtFP (Feature) ranks 1st in the 70-30 validation and 5th in the LOSO validation, in the ACE inhibitor set it ranks 13th and the GPCR set 13th (70-30) and 9th (LOSO). To connect these observations of descriptor set performance to the similarity of the sequences and the way the descriptor sets characterize the space, we again performed a PCA analysis.

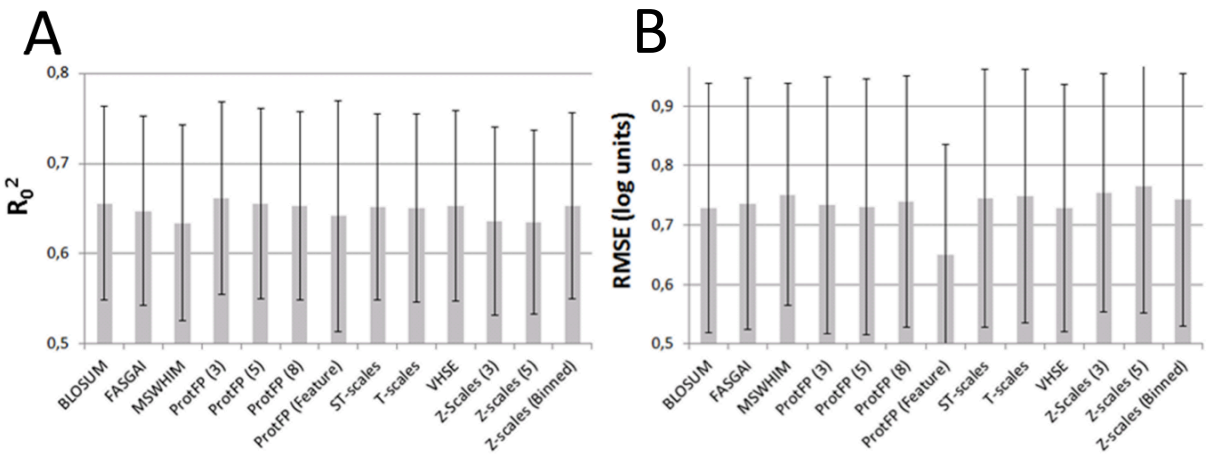



**Figure 3.9: The average performance of the benchmarked descriptor sets in the NNRTIs LOSO validation experiments.** The average is calculated over all 14 mutants (performed in triplicate) and the error bar represents the SEM (note that error bars are large due to different performance between models trained on different mutants, not between repeats of the individual models. Also see supporting **Figure S31**). Shown are the $R_0^2$ (A) and the RMSE (B). Here extrapolation takes place on the target side as the test set contains unseen targets. The differences between individual descriptor sets are still small but the spread of the SEM increases. Again for individual receptors larger performance differences occur (main text). Still ProtFP (Feature) again performs very good, it seems that a simplified representation is favorable for this data set.

**3.5.10 NNRTIs (Target Space).** The PCA analysis can explain the better performance of ProtFP (Feature) (Supporting **Figures S32 – S34**). Due to the fact that the mutants only differ by point mutations and one of the sequences caries 15 mutations (sequence 7), this sequence is set far apart from the other sequences by most descriptor sets. This effect is much less pronounced in ProtFP (Feature) as is does not differentiate between the type of mutations (all AAs are encoded as features so every amino acid difference is equal). The effect is that all the sequences cluster much closer than in the other descriptors, this leads to a better performance on this set.

Another cause for the observed effect could be that, by leaving out the residues that did not mutate in any of the sequences, we have maximized the dissimilarities to an extent that they do not accurately represent the bioactivity space. As the ProtFP (Feature) descriptor set leads to relatively small distances by merely encoding presence or absence of a feature, it partially compensates for this effect. In any case, the completely different way of describing the sequence similarity by ProtFP (Feature) proves to be beneficial for this dataset.

**3.5.11 NNRTIs (Conclusions).** The NNRTI set represented a different data set compared to the GPCR ligand dataset evaluated above as it consists of a number of highly similar sequences and compounds and, hence, resembles a typical data set one might encounter in lead optimization. We conclude that in these cases the feature base descriptor set might perform very well, however its good performance can also be catalyzed by the binding site definition. Therefore this type of descriptor set should be included as a possible candidate when working on a data set consisting of several highly related targets. For example a single GPCR subfamily like the adenosine receptors can also be considered a set of highly similar targets. The findings from this part could therefore also apply to this family. Indeed we found in other work that ProtFP (Feature) also performs well on this set.[20]

**3.5.12 Final Descriptor Set Ranking.** The final ranking of the individual descriptor sets is given in **Table 3.4**. This table included the individual ranks of all descriptor set in each experiment (on a scale of 1 to 13) and a final overall ranking (the sum of the individual rankings). We have also included the average rank and the SEM of this average rank (**Figure 3.10**).

The best performing descriptor sets overall are T-scales (3) (average rank 5.2), ProtFP_PCA (3) (average rank 5.4), Z-scales (3) (average rank 5.6) and Z-scales (Binned) (average rank 6.0). Taking into account that all descriptor sets performed very close and the easy interpretability from the Z-scales (Binned) might make this descriptor set the best choice to use for PCM experiments (it also displays the smallest SEM of the four).

The worst performing descriptor sets are ProtFP (Feature) (average rank 8.4), ST-scales (average rank 9.0), and ProtFP_PCA (8) (average rank 9.4). While their performance was close to the other descriptor sets, they were in the lower performing ranks in 80 % of the experiments. Therefore it might be wise to avoid these descriptor sets on bioactivity modeling in setups such as the PCM modeling employed here; but this again will surely depend on the particular dataset at hand as well.

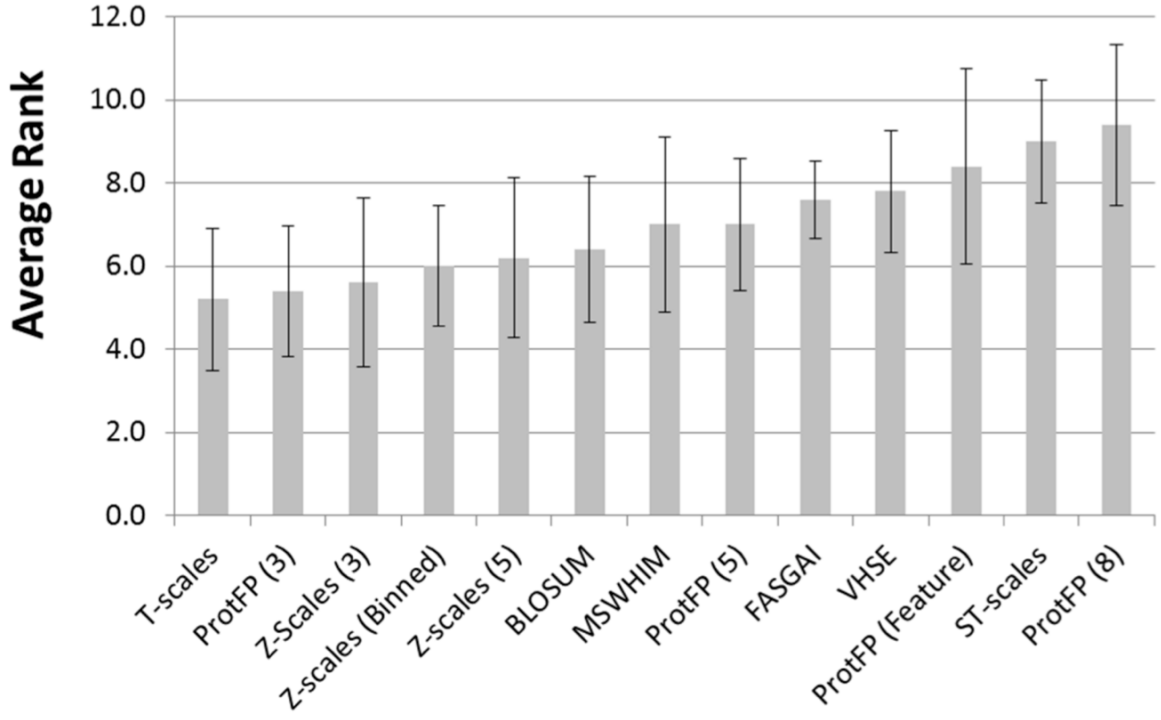



**Figure 3.10: The average rank of the descriptor sets in the bioactivity benchmarks.** The average is calculated over these 5 ranks and the SEM is given by the error bars. The best three descriptor sets perform about equal with an average rank =< 6 (where Z-scales (Binned) shows the smallest spread). The worst performance is by ProtFP (Feature), ST-scales and ProtFP_PCA (8) with an average rank > 8. ProtFP (Feature) and MSWHIM have a large error bar due to their inconsistent performance.

Contrary to our expectations, the best performing descriptor set is based on 135 amino acids (including non-natural amino acids) rather than one built on only the natural amino acids. Another interesting observation is that the top 4 descriptors on average consist on 4 principal components and the worst 4 consist of on average 8 principal components. Furthermore when multiple versions of the same descriptor set are compared, the set employing the least number of amino acids consistently scores better (ProtFP_PCA and Z-scales). This confirms our expectation that including more principal components that describe less variance introduces more noise than information.

**Table 3.4.** Overall Descriptor Set Ranking.

| Descriptor | Final Rank | Rank ACE Inhibitors | Rank GPCR 70-30 validation | Rank GPCR LOSO | Rank NNRTI 70-30 validation | Rank NNRTI LOSO | Mean Rank |
|---|---|---|---|---|---|---|---|
| T-scales | 26 | 2 | 1 | 8 | 5 | 10 | 5.2 (±1.7) |
| ProtFP (3) | 27 | 9 | 5 | 2 | 9 | 2 | 5.4 (±1.6) |
| Z-Scales (3) | 28 | 3 | 10 | 1 | 3 | 11 | 5.6 (±2.0) |
| Z-scales (Binned) | 30 | 1 | 6 | 7 | 10 | 6 | 6.0 (±1.5) |
| Z-scales (5) | 31 | 7 | 2 | 3 | 6 | 13 | 6.2 (±2.0) |
| BLOSUM | 32 | 6 | 7 | 6 | 12 | 1 | 6.4 (±1.8) |
| MSWHIM | 35 | 5 | 12 | 4 | 2 | 12 | 7.0 (±2.1) |
| ProtFP (5) | 35 | 10 | 4 | 11 | 7 | 3 | 7.0 (±1.6) |
| FASGAI | 38 | 4 | 9 | 9 | 8 | 8 | 7.6 (±1.0) |
| VHSE | 39 | 8 | 11 | 5 | 11 | 4 | 7.8 (±1.5) |
| ProtFP (Feature) | 42 | 13 | 13 | 10 | 1 | 5 | 8.4 (±2.4) |
| ST-scales | 45 | 12 | 8 | 12 | 4 | 9 | 9.0 (±1.5) |
| ProtFP (8) | 47 | 11 | 3 | 13 | 13 | 7 | 9.4 (±2.0) |

The descriptor sets are sorted based on their final rank. Also shown are the rank each descriptor set receives in each individual benchmark. The final column shows the average rank for each descriptor set (calculated from the 5 individual ranks) and the SEM of associated with this average. The best performance is achieved by the T-scales (3), closely followed by ProtFP (3), Z-scales (3) and Z-scales (Binned).

**3.5.13 Training Times.** One final property of the descriptor sets has not been highlighted yet. On a workstation with a core i7 860 CPU and 16 GB memory, we found considerable differences in training times for the individual descriptor sets. On the datasets used in this work, as a rule of thumb ProtFP Feature showed the fastest model training while BLOSUM required most time (191% of the training time required for ProtFP Feature). The reason for this large difference is that the feature based descriptor set uses a single variable per amino acid, where the numerical descriptor sets use 3 (ProtFP PCA (3), Z-scales (3) and MS-WHIM) to 10 values (BLOSUM).

## 3.6 Conclusions

Given the large number of AA descriptor sets available we aimed to both characterize those descriptor sets with respect to their perception of similarities between AAs, and to benchmark them in bioactivity models. Descriptor set clustering indicated that they show different behavior from one another when characterizing AA similarities. As might be intuitive, when only considering the first two principal components, descriptor sets cluster the way they are derived, with Z-scales, VHSE and ProtFP PCA falling into one cluster, T-scales and ST-scales forming a second group of descriptor sets, and FASGAI, BLOSUM and MS-WHIM descriptor sets being somewhat distinct to the above groups.

Our results confirm that all QSAM descriptor sets can be used to train predictive bioactivity, including PCM, models. Performance differences between descriptor sets were in the order of magnitude of an RMSE difference of 0.1 log units. Individual targets could cause much larger differences in performance (e.g. the RMSE difference between the HIV mutant modeled best and worst was 1.2 log units). Therefore we conclude that all descriptor sets can be used to create predictive models.

Nevertheless, depending on the problem at hand, it might be wise to do an initial descriptor set selection before training a final model. In particular in data sets where affinity on unknown targets is predicted (like receptor deorphanization exercises, simulated by our LOSO experiments), larger differences in performance can occur. In cases where virtual screening is applied to a data set consisting of known targets, these differences are slightly smaller in magnitude.

## 3.7 Acknowledgements

The financial support of Tibotec BVBA is gratefully acknowledged.

## 3.8 Supporting Information

Additional tables (Supporting **Tables S1 – S26**), figures (**Figures S1 – S39**) are available as pdf. Furthermore, we include a Pipeline Pilot component to convert single letter AA sequences to any of the here tested descriptor sets and a fully functional example protocol, both to be used in Pipeline Pilot 8.5 and up (archive file). These materials are available online at www.gjpvanwesten.nl. The GPCR data set is available upon request but was considered too large to submit with the paper.

## 3.9 References

1. A. Kontijevskis, P. Prusis, et al.; *A look inside HIV resistance through retroviral protease interaction maps.* PLoS Comput. Biol.; *2007*. **3** (3): e48.
2. M. Lapinsh, P. Prusis, et al.; *Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions.* Biochim. Biophys. Acta, Gen. Subj.; *2001*. **1525** (1-2): 180-190.
3. G.J.P. Van Westen, J.K. Wegner, et al.; *Proteochemometric Modeling as a Tool for Designing Selective Compounds and Extrapolating to Novel Targets.* Med. Chem. Commun.; *2011*. **2** (1): 16-30.
4. J.E.S. Wikberg, F. Mutulis, et al.; *Melanocortin receptors: Ligands and proteochemometrics modeling;* in *Melanocortin System*; D. Braaten; Editor *2003*: New York. p. 21-26.
5. J.R. Bock and D.A. Gough; *Virtual screen for ligands of orphan G protein-coupled receptors.* J. Chem. Inf. Model.; *2005*. **45** (5): 1402-1414.
6. M. Lapinsh, P. Prusis, et al.; *Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands.* Mol. Pharmacol.; *2002*. **61** (6): 1465-1475.
7. P. Prusis, S. Uhlén, et al.; *Prediction of indirect interactions in proteins.* BMC Bioinformatics; *2006*. **7**: 167-180.
8. J. Jonsson, T. Norberg, et al.; *Quantitative sequence-activity models (QSAM)--tools for sequence design.* Nucl. Acids Res.; *1993*. **21** (3): 733-739.
9. M. Sandberg, L. Eriksson, et al.; *New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids.* J. Med. Chem.; *1998*. **41** (14): 2481-2491.
10. J. Meslamani, J. Li, et al.; *Protein–Ligand-Based Pharmacophores: Generation and Utility Assessment in Computational Ligand Profiling.* J. Chem. Inf. Model.; *2012*. **52** (4): 943-955.
11. H. Strombergsson, A. Kryshtafovych, et al.; *Generalized modeling of enzyme-ligand interactions using proteochemometrics and local protein substructures.* Proteins: Struct., Funct., Bioinf.; *2006*. **65** (3): 568-579.
12. N. Weill and D. Rognan; *Development and Validation of a Novel Protein−Ligand Fingerprint To Mine Chemogenomic Space: Application to G Protein-Coupled Receptors and Their Ligands.* J. Chem. Inf. Model.; *2009*. **49** (4): 1049-1062.
13. P. Zhou, F. Tian, et al.; *Quantitative Sequence-Activity Model (QSAM): Applying QSAR Strategy to Model and Predict Bioactivity and Function of Peptides, Proteins and Nucleic Acids.* Current Computer - Aided Drug Design; *2008*. **4** (4): 311-321.

14. L. Yang, M. Shu, et al.; *ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues.* Amino Acids; *2010*. **38** (3): 805-816.

15. H. Mei, Z.H. Liao, et al.; *A new set of amino acid descriptors and its application in peptide QSARs.* Biopolymers; *2005*. **80** (6): 775-786.

16. F. Tian, P. Zhou, and Z. Li; *T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides.* J. Mol. Struct.; *2007*. **830** (1-3): 106-115.

17. L. Guizhao and L. Zhiliang; *Factor Analysis Scale of Generalized Amino Acid Information as the Source of a New Set of Descriptors for Elucidating the Structure and Activity Relationships of Cationic Antimicrobial Peptides.* QSAR Comb. Sci.; *2007*. **26** (6): 754-763.

18. A. Zaliani and E. Gancia; *MS-WHIM Scores for Amino Acids: A New 3D-Description for Peptide QSAR and QSPR Studies.* J. Chem. Inf. Comput. Sci.; *1999*. **39** (3): 525-533.

19. A.G. Georgiev; *Interpretable numerical descriptors of amino acid space.* J. Comput. Biol.; *2009*. **16** (5): 703-723.

20. G.J.P. Van Westen, O.O. van den Hoven, et al.; *Identifying Novel Adenosine Receptor Ligands by Simultaneous Proteochemometric Modeling of Rat and Human Bioactivity Data.* J. Med. Chem.; *2012*: **55** (16): 7010-7020.

21. G.J.P. Van Westen, J.K. Wegner, et al.; *Which Compound to Select in Lead Optimization? Prospectively Validated Proteochemometric Models Guide Preclinical Development.* PLoS One; *2011*. **6** (11): e27518.

22. S. Hellberg, L. Eriksson, et al.; *Minimum analogue peptide sets (MAPS) for quantitative structure-activity relationships.* Int. J. Pept. Protein Res.; *1991*. **37** (5): 414-424.

23. A. Gaulton, L.J. Bellis, et al.; *ChEMBL: a large-scale bioactivity database for drug discovery.* Nucleic Acids Res.; *2011*. **40**: D1100 - D1107.

24. S. Hellberg, M. Sjoestroem, et al.; *Peptide quantitative structure-activity relationships, a multivariate approach.* J. Med. Chem.; *1987*. **30** (7): 1126-1135.

25. M. Lapins, M. Eklund, et al.; *Proteochemometric modeling of HIV protease susceptibility.* BMC Bioinformatics; *2008*. **9** (1): 181-192.

26. M. Connolly; *Analytical molecular surface calculation.* J. Appl. Crystallogr.; *1983*. **16** (5): 548-558.

27. S. Henikoff and J.G. Henikoff; *Amino acid substitution matrices from protein blocks.* Proc. Natl. Acad. Sci. U. S. A.; *1992*. **89** (22): 10915-10919.

28. S. Kawashima, H. Ogata, and M. Kanehisa; *AAindex: Amino Acid Index Database.* Nucleic Acids Res.; *1999*. **27** (1): 368-369.

29. R Development Core Team; *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing *2006*.
30. A. Bender, J.L. Jenkins, et al.; *How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space.* J. Chem. Inf. Model.; *2009*. **49** (1): 108-119.
31. D.E. Gloriam, S.M. Foord, et al.; *Definition of the G Protein-Coupled Receptor Transmembrane Bundle Binding Pocket and Calculation of Receptor Similarities for Drug Design.* J. Med. Chem.; *2009*. **52** (14): 4429-4442.
32. Accelrys Software Inc *Pipeline Pilot Professional Edition* Scitegic Version 8.5
33. B.T. Korber, B.T. Foley, et al. *Numbering Positions in HIV Relative to HXB2CG*. *1998*.
34. E. van der Horst, J. Peironcely, et al.; *A novel chemogenomics analysis of G protein-coupled receptors (GPCRs) and their ligands: a potential strategy for receptor de-orphanization.* Bmc Bioinformatics; *2010*. **11** (1): 316.
35. E. Van der Horst, J.E. Peironcely, et al.; *Chemogenomics Approaches for Receptor Deorphanization and Extensions of the Chemogenomics Concept to Phenotypic Space.* Curr. Top. Med. Chem.; *2011*. **11** (15): 1964-1977.
36. D. Rogers and M. Hahn; *Extended-Connectivity Fingerprints.* J. Chem. Inf. Model.; *2010*. **50** (5): 742-754.
37. A. Liaw and M. Wiener; *Classification and Regression by randomForest.* R News; *2002*. **2** (3): 18-22.
38. A. Tropsha; *Predictive Quantitative Structure-Activity Relationships Modeling;* in *Handbook of Chemoinformatics Algorithms*; J. Faulon and A. Bender; Editors. *2010*.
39. P. Baldi, S. Brunak, et al.; *Assessing the accuracy of prediction algorithms for classification: an overview.* Bioinformatics; *2000*. **16** (5): 412-424.
40. J.-S. Surgand, J. Rodrigo, et al.; *A chemogenomic analysis of the transmembrane binding cavity of human G-protein-coupled receptors.* Proteins: Struct., Funct., Bioinf.; *2006*. **62** (2): 509-538.