4 Analysis of Structure-Specificity-Relationships of PDZ domains

4.1 Problems of PDZ Domain Classification and Ligand Prediction

Several attempts have been made to classify PDZ domains and/or their ligands. The most widely used classification system is based upon the sequences of C-terminal peptide ligands. The last three or four amino acid residues are usually considered and two essential classes are defined. This yields the signatures (S/T)xΦ_{COOH} as class I and ΦxΦ_{COOH} as class II [30] (for further information see Chapter 1.2). This PDZ domain classification of Songyang *et al*. (class I and class II) [30] is based on the interaction between ligand position -2 and the residue located at the N-terminal end of the αB-helix (position αB:1) of the PDZ domain: class I is determined by a polar residue (mostly histidine) whereas class II contains a predominantly hydrophobic amino acid. However, completely different PDZ domain ligands are known which do not fit into the Songyang *et al.* classification, such as the PDZ domain of nNOS with the ligand binding pattern G(E/D)xV$_{COOH}$ [40], the PDZ domain of Mint-1 with the novel recognition sequence (E/D)xW(C/S)$_{COOH}$ [41], the human INAD-like (hINADL) PDZ-3 with xψ(E/D)$_{COOH}$ [155] or the syntenin-2 PDZ domain with xYxC$_{COOH}$ [157].

Recently, a new PDZ classification based on the nature of the amino acids in two critical positions (αB:1 and βB:5) of the PDZ domain fold has been described [154]. Using these positions, all currently known PDZ domains could be arranged into 25 possible groups. This classification provides a method for predicting the specificity of all PDZ domains and relies on a close connection between ligand preference and the amino acid residues at given positions in the PDZ domain. However, within these 25 groups, the first group covers PDZ domains that bind class I peptides and the remaining groups are less clearly determined. Two of them do not corresponded to any known PDZ domains, 14 are not correlated with any ligand sequences, four groups can be positioned into canonical class II domains, and one group includes PDZ domains that are known to have dual specificity [155]. These problems in PDZ domain classification reflect the difficulty to predict the respective ligand.

Other recent publications categorize the PDZ domain containing proteins into three general classes according to their modular organization: 1. multi PDZ proteins, 2. MAGUKs and 3. others (see Chapter 1.2) [27, 156].

The three PDZ domains reported here (the AF6, ERBIN and SNA1 PDZ domains) differ substantially in position (αB:1) from other PDZ domains and show interesting binding specificities. The AF6 PDZ domain has glutamine at position αB:1 and could bind to both, ligand of class I (e.g. BCR: KRQSILFS**T**EV$_{COOH}$) [67] and class II equally (e.g. EPB2: AQMNQIQS**V**EV$_{COOH}$) [64]. The ERBIN PDZ domain contains the typical histidine, resulting in a class I classification. However, interactions of the ERBIN PDZ domain with ligands of class II-type (e.g. ERB2: NPEYLGLD**V**PV$_{COOH}$) are also known. In the contrary, the SNA1 PDZ domain also shows a conserved histidine, which is typical for class I, but now indeed no natural ligands of class II are yet known.

4.2 Facility to Quantify PDZ Domain Specificity

PDZ domains form a β-barrel structures flanked by two α-helices [28] whereby the four C-terminal ligand residues fit in the groove between βB-strand and αB-helix [9, 158]. Although specificity is mainly determined by the four C-terminal residues [31-33], further ligand positions will also contribute specifically, especially to the βB-βC-loop [30, 122]. However, these interactions are restricted to individual PDZ domains or even individual PDZ domain/ligand pairs. In this chapter, a quantitative description of the individual amino acid type-specific affinity contributions for each of the four C-terminal ligand positions relative to a reference peptide was developed. This rigorous treatment of data from binding assays allowed to elucidate PDZ domain specificity profiles. The amino acid type-specific affinity contributions can be related to variations in the binding free energy contributions of the specific ligand side chain/domain interactions according to $\Delta\Delta G_0 = -RT \ln(K_d(\text{peptide}) / K_d(\text{reference})) = -RT \Delta\ln(K_d)$. The specificity profiles allow the calculation of affinities towards the complete peptide sequence space. By comparing the resulting domain-specific ligand sequence spaces of different PDZ domains, the specificity overlap and the selectivity can be determined.

The 6223-Humlib screens of the AF6, ERBIN and SNA1 PDZ domains were used in order to obtain an overview of the recognized sequence diversity. The amount of PDZ domain bound to membrane-attached peptide libraries was quantified by means of a chemiluminescence detection system (measured as Boehringer Light Units - BLU). The signal intensities represent relative affinities. Selected peptide sequences of the 6223-Humlib screens were investigated by using further libraries (substitutional analyses). All these data together provided the basis for designing focused libraries of the type $bbbbB_{-3}B_{-2}B_{-1}B_{0COOH}$ (profile library, B = permutation of a defined set of amino acids, b = mixture of 17 amino acids, without C, M and W), which were used to derive ligand specificity profiles. The amino acid type-specific contributions in each of the four C-terminal ligand positions were then quantified using an Analysis of Variance (ANOVA) of the combined data from the substitutional analyses and the profile libraries. The resulting ANOVA models were finally tested by the prediction of peptides with the highest affinities (super-binders) that were validated experimentally. The latter could be useful as competitive antagonists of natural ligands in cell biology experiments. Finally, using the predicted dissociation constants (K_d) of all potential 130321 ligand sequences for all three PDZ domains, the selectivity and the overlap of the domain-specific ligand sequence spaces in the 1 - 100 μM range are quantitatively compared.

4.3 Results

4.3.1 Overview of AF6, ERBIN and SNA1 PDZ Domain/Peptide Ligand Diversity

A peptide library exposing 6223 C-termini of human proteins (6223-Humlib) was generated by SPOT synthesis of inverted peptides [160] that is considered diverse enough to obtain an overview of the targeted sequence space for three PDZ domains. This library was incubated with the PDZ domains of AF6 (Figure 4.1), ERBIN (Figure 3.7 in Chapter 3) and SNA1 (Figure 3.6 in Chapter 3). About one hundred strong signals due to affinities estimated to be in the micro-molar range, and a multitude of weaker ones were identified in the individual screens (Tables 0.1 to 0.3 in Appendix). To a large extent, these peptide sequences contained the known C-terminal binding motifs. Furthermore, there were non-terminal sequences, potentially forming β-finger-like interactions [16], and a

negligible number of unexpected unconventional motifs. Although the relative intensities (BLU) do not correlate well with dissociation constants, a rough overview of the target sequence space may be obtained by the alignment of the top 100 interacting peptides.

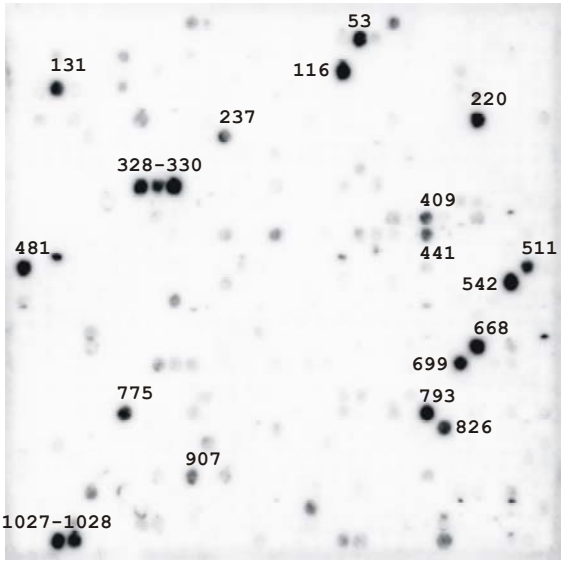


Figure 4.1 ***In vitro*** **Identification of Putative AF6 PDZ Domain Ligands.**
Selected region of the library of 6223 C-termini (11-mers) of human proteins from the Swiss-Prot database incubated with GST-labeled AF6 PDZ domain (cysteine is replaced by serine). The 100 strongest binders are listed in Table 0.1 in the Appendix.

The AF6 PDZ domain revealed a clear preference of 78% for aliphatic amino acids in position 0, especially for valine (62%). Furthermore, phenylalanine occurs in 5% of cases. In position -2, hydrophobic amino acids are tolerated (52%) as well as T/S (39%). Thus, the AF6 PDZ domain targets both, class I and class II ligands. Within the top 100 sequences we often found combinations of valine in position 0, T/S in position -2 and E/S in position -3. A serine in position -2 mainly occurred together with valine in position 0, while a phenylalanine in position 0 seems to correlate with a hydrophobic residue in position -2. Small amino acids (π) such as G/A/P/S are represented with 49% in position -3, in addition to smaller contributions of glutamic acid with 19% and threonine with 11%. In position -1, most of the 19 used amino acids were found except for G/M/R/W, but with the emphasis on Y/V/L/F. The ERBIN PDZ domain also selected aliphatic amino acids in position 0 (78%) with a high preference for leucine (44%). Within the top 100 sequences, 71% of them show T/S at position -2. Only 21% contain a

hydrophobic amino acid, classifying ERBIN more explicitly as a class I PDZ domain. In position -3 acidic amino acids (E/D) occur with a frequency of 52% and serine as the next most frequent amino acid with 12%.

The screen of the SNA1 PDZ domain revealed amino acid preferences for each ligand position similar to those reported in earlier studies [34, 104] (position 0: V/I/L with 67%; position -2: S/T with 67%; position -3: E with 36%; position -4: K/R with 26%).

Table 4.1 Overview of the Amino Acid Frequency within the Top 100 Sequences of the 6223-Humlib.

Domain	Ligand		
Name	Position	Amino Acid	Frequency
AF6	0	Ψ	78%
		F	5%
	-1	x	--
	-2	Φ	52%
		T/S	39
	-3	π	49%
		E	19%
ERBIN	0	Ψ	78%
	-1	x	--
	-2	T/S	71%
		F	21%
	-3	E/D	52%
SNA1	0	Ψ	67%
	-1	x	--
	-2	S/T	67%
	-3	E	36%
	-4	R/K	26%

Footnotes: Position: 0 referring to the C-terminal ligand residue. Amino acid: π representing small amino acids, Φ hydrophobic amino acids and Ψ aliphatic ones. x is denoting any amino acid that occurs without a special frequency.

In summary, the 6223-Humlib screens present the spectrum of interacting peptide sequences for each of the three PDZ domains (Table 4.1). However, they did not show narrow recognition motifs. Although there is a preference for individual amino acids in each of the investigated peptides' C-terminal positions, they occur in various combinations. The preferred combinations of amino acids (corresponding to the

specificity) may thus be better explained by the amino acid type-specific affinity contributions and the cooperativity of the C-terminal ligand residues.

The presented top 100 binding peptides represent in all three cases approximately the top 40 % of the total detected Boehringer Light Unit (BLU) range of the 6223-Humlib screens (Table 4.2). The overlap of the three lists comprising the top 100 peptides recognized by the three PDZ domains is 9 %. Another 6 % of peptides are recognized both the AF6 and ERBIN PDZ domains, 9 % by the AF6 and SNA1 PDZ domains and 20 % by the ERBIN and SNA1 PDZ domains.

Table 4.2 Boehringer Light Unit ranges of top 100 peptides for the 6223-Humlib screens of AF6, ERBIN and SNA1 PDZ domains.

Domain	BLU		Log(BLU)				
	1st	100th	1st	100th	Minimum	Median	*rel. DLog(BLU)
AF6	5 502 000	778 006	6.741	5.891	4.554	4.809	44.0 %
ERBIN	6 763 000	1 281 000	6.830	6.108	4.489	4.868	36.8 %
SNA1	519 140	167 115	5.715	5.223	4.121	4.303	34.8 %

Footnotes: BLU: raw intensities measured in Boehringer Light Units given for the 1st (with the highest intensity) and the 100th peptide sequence (with the 100th highest intensity). Log(BLU): logarithm-transformed BLU values in order to obtain approximately normally distributed values. Minimum: lowest observed Log(BLU) value. Median: median of all 6,223 Log(BLU) values.
*rel. ΔLog(BLU) = {Log(BLU_{1st}) – Log(BLU_{100th})} / {Log(BLU_{1st}) – Median(Log(BLU))}

4.3.2 Substitutional Analyses of PDZ Peptide Ligands

In order to describe the individual PDZ domains' amino acid preferences for each ligand position in more detail, substitutional analyses with up to 20 representative ligands were performed (see Figures 0.1 - 0.3 in the Appendix).

Four C-terminal peptide ligands for each PDZ domain were selected for further investigations (Figure 4.2). For the AF6 PDZ domain, we show the substitutional analyses of the 11 C-terminal residues of the ephrine receptor tyrosine kinase B2 (EPB2, class II), the ephrine receptor tyrosine kinase A7 (EPA7, class II), the forkhead box protein I1 (FXI1, class I) and the muscle sodium channel protein type IV alpha subunit (CIN4, class I) (Figure 4.2 (A)). The targeted sequence can be summarized as **$(\pi/E)(\Psi/T/S)x\Psi_{COOH}$** which is consistent with many hits of the 6223-Humlib screen. Additionally, the

substitutional analyses revealed some restrictions at position -1 (e.g. glycine is not tolerated).

For the ERBIN PDZ domain, we select the substitution analyses of the C-terminal fragments of the receptor tyrosine kinase ERB2 (ERB2, class II), of the activin receptor type II precursor (AVR2, class I), of the muscle sodium channel protein type IV alpha subunit (CIN4, class I) and of the armadillo repeat protein deleted in velo-cardio-facial syndrome (ARVC, class I) (Figure 4.2 (B)). Based on these results, the most preferred ERBIN ligands are described by **(E/D)(T/S/V)x(V/L/I)$_{COOH}$** (see also Chapter 3.2.3). Although this motif is very similar to the one obtained by the sequence analysis of the 6223-Humlib incubation, here, V is revealed as the preferred amino acid in position 0. In the case of the AVR2 ligand, position -5 additionally contributes to ligand specificity, with a preference for aromatic residues F, W and Y and to a minor degree H, I, P and V. Furthermore, the ERBIN PDZ domain revealed a preference for aromatic amino acids in position -1 [17, 128, 164]. This was confirmed by an additional combinatorial library of the type bbbB$_{-4}$ETB$_{-1}$V$_{COOH}$ (Figure 4.3). This library reveals the fully amino acid perturbations of two positions (-1 and -4) within the consensus motif of the ERBIN PDZ domain to display their dependency. If F or W are present at position -1, no preference for one of the other 19 amino acids (cystein omitted) is observed at positions -4. The same phenomenon is observable *vice versa* with lower specificity. Glycine at positions -1 is not tolerated as determined in the substitutional analysis.

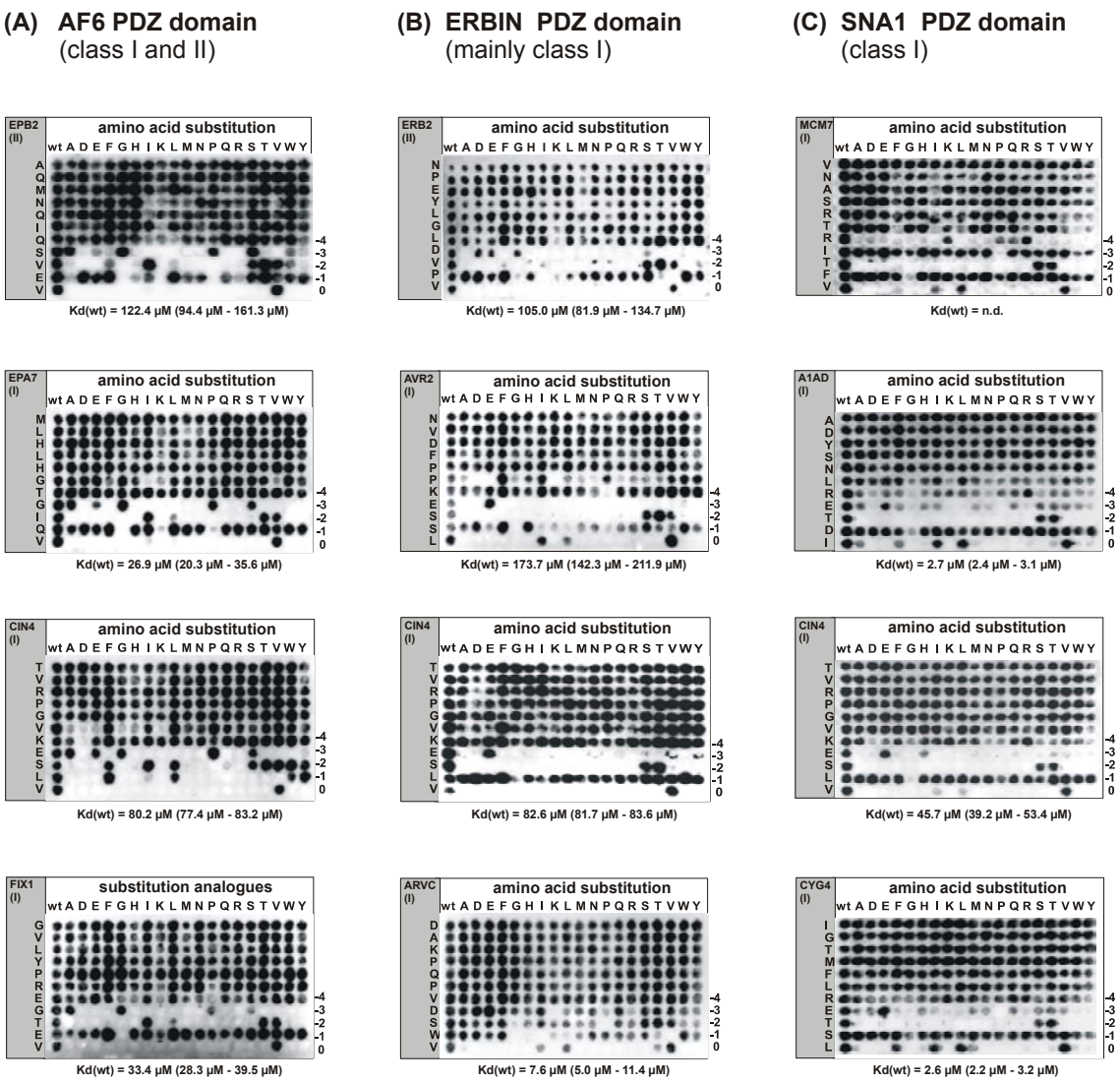


Figure 4.2 Characterization of Ligand/PDZ Domain Interaction of Selected Peptides.
The substitutional analyses of the three investigated PDZ domains revealed a similar pattern of the essential amino acids, which is mainly characterized by the last four C-terminal residues. Furthermore, a class-specific substitution pattern is not observable.
(**A**) Incubations of the AF6 PDZ domain with the C-terminal peptides of the ephrine receptor tyrosine kinase B2 (EPB2), the ephrine receptor tyrosine kinase A7 (EPA7), the muscle sodium channel protein type IV alpha subunit (CIN4) and the forkhead box protein I1 (FXI1).
(**B**) Incubations of the ERBIN PDZ domain with the C-terminal peptides of the receptor tyrosine kinase ERB2 (ERB2), the activin receptor type II precursor (AVR2), the muscle sodium channel protein type IV alpha subunit (CIN4) and the armadillo repeat protein deleted in velo-cardio-facial syndrome (ARVC).
(**C**) Incubations of the SNA1 PDZ domain with the C-terminal peptides of the DNA replication licensing factor MCM7 (MCM7), the α-1D adrenergic receptor (A1AD), the muscle sodium channel protein type IV alpha subunit (CIN4) and the soluble α-2 chain of the guanylate cyclase (CYG4).
All spots in the left hand columns (grey box with class determination given in parentheses) are identical and represent the wild type (wt) peptide. All other spots represent the substitution of one amino acid against the amino acid in the respective column. Hence, each spot bears a single substitution compared to the sequence of the wild type. For example, V in position 0 can only be substituted by itself and in some cases by I and L. The peptide class according to Songyang *et al.* [30] is given in parenthesis next to the PDZ domains and the ligands. The last five C-terminal ligand positions are numbered to the right of each substitutional analysis. The experimentally determined K_d values of the wild type peptides are given below each substitutional analysis, with the respective +/- 1σ interval on log-scale.

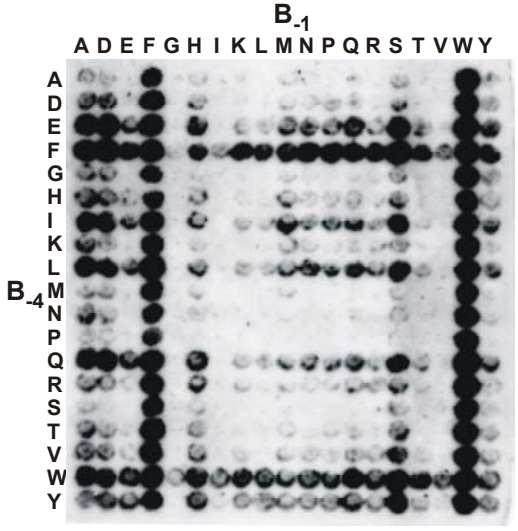


Figure 4.3 Simultaneous Substitutions of two Amino Acids in the Consensus Binding Motif of ERBIN PDZ Domain Ligands.

The GST-ERBIN PDZ incubation of the combinatorial library bbbB_{-4}ET$B_{-1}$$V_{COOH}$ revealed a strong preference for tryptophan and phenylalanine in position -1. These aromatic amino acids compensate the reductionof binding of substitutions in position -4. *Vice versa*, the same is observed for position -4 but in a less specific manner.

Residues denoted with b represent a mixture of all 17 L-amino acids (except C, M and W). B_{-1}/B_{-4} means a substitution of this position through all 20 L-amino acids except C whereas position E, T and V remain constant.

The substitutional analyses showed a specific substitution pattern for every domain rather than a class-specific substitution pattern, irrespective of the investigated ligand sequences. This implies a common binding mode for all ligands. Furthermore, the residues at individual ligand positions mostly contribute independently to the total affinity. For the AF6 and ERBIN PDZ domains it is also apparent that mainly the four C-terminal positions of the peptides are important, while for SNA1 five residues need to be considered. The equal contribution of the last four residues can be attributed to the conserved PDZ domain/ligand binding mechanism, which has already been structurally investigated for the ERBIN and SNA1 PDZ domains. Since no structural data on the AF6 PDZ domain are available, we analyzed the AF6 binding site by NMR and homology modeling.

4.3.3 Analysis of the AF6 PDZ Domain Binding Site

For structural analysis of the AF6 PDZ domain ligand binding mechanism and in particular to determine whether class I and class II ligands bind similarly, several NMR chemical

shift titration experiments (Figure 4.4) with the same four ligands (EPB2, EPA7, CIN4 and FXI1) as shown in Figure 4.2 were performed. The chemical shift changes are summarized in Figure 4.5 (A) and mapped onto the surface representation of the homology modeled AF6 PDZ domain in Figure 4.5 (C-F).

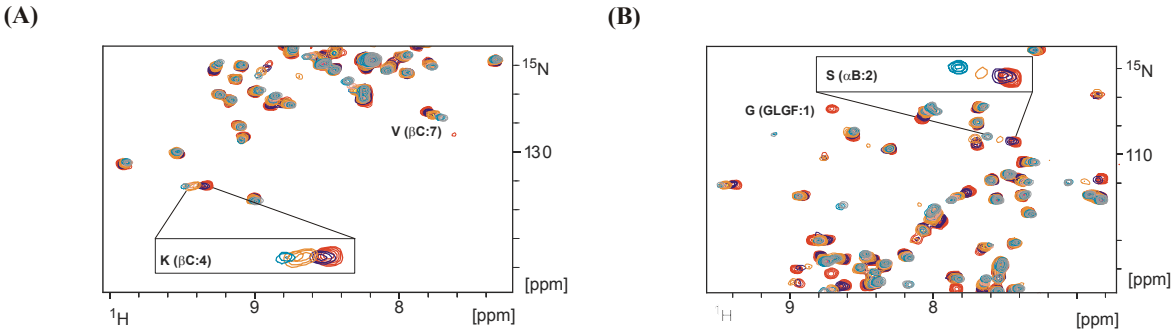


Figure 4.4 Mapping of the AF6 PDZ Domain Binding Site in Presence of the ERB2 Ligand.
Four residues of the AF6 PDZ domain binding site are selected to demonstrate the chemical shift perturbation caused by the ligand binding. These amino acids are found again in Figure 4.5.
^{15}N-AF6 PDZ domain (0.1 mM) titrated with the C-terminal peptide (30 µM - 10 mM) in 20 mM phosphate buffer, 50 mM NaCl, pH 7.0, Bruker DRX600 at 300 K.

The largest chemical shift perturbations are found in the region of the GLGF-loop, βB, βC and αB around the conserved PDZ domain-binding groove. Minor chemical shift perturbations in βC, βD and the C-terminal tail of the domain are probably due to small structural rearrangements. Interestingly, the observed chemical shift perturbation patterns did not differ between class I and class II peptides, which support the conclusions drawn from the substitutional analyses: A hydrophobic amino acid in position -2 could interact with the respective interaction area (IA_{-2}) of the PDZ domain as well as an amino acid containing a hydroxyl group. These contacts are mainly driven by hydrophobic interactions.

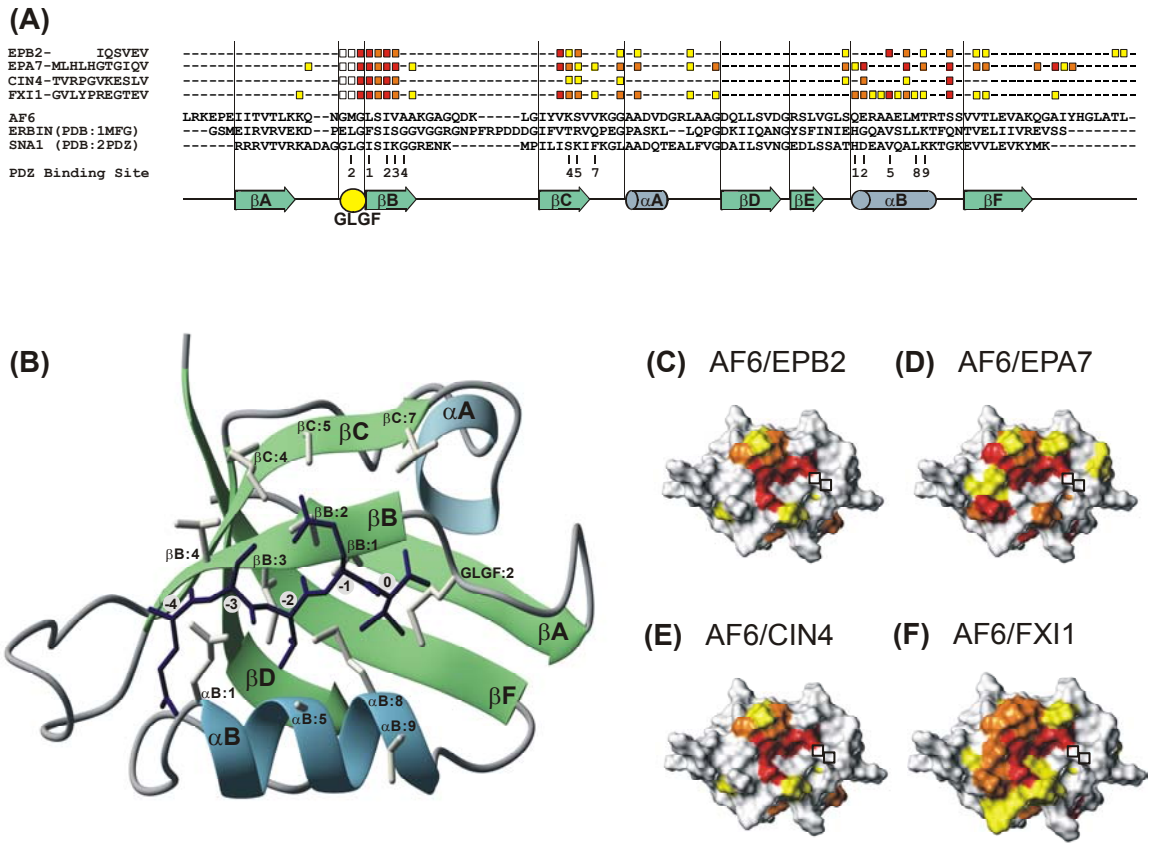


Figure 4.5 Sequence Alignment of AF6, ERBIN and SNA1 PDZ Domains and Ligand-Induced Chemical Shift Perturbations of the AF6 PDZ Domain.

The detected chemical shift changes are mainly situated in the conserved binding region of PDZ domains corresponding to the GLGF-loop, αB , βBand βC, independently of the C-terminal ligand sequences.
(**A**) The chemical shift perturbation patterns observed on the AF6 PDZ domain for four peptides are indicated by rectangles above the alignment of the AF6, ERBIN and SNA1 PDZ domain sequences: vanishing signals are indicated in red, moderate chemical shift changes ($\Delta\delta > 0.140$ ppm) in orange and small chemical shift changes ($\Delta\delta > 0.100$ ppm) in yellow. White rectangles indicate missing chemical shifts.
(**B**) Ribbon representation of the AF6 PDZ domain model. The residues of the ligand binding site are labeled according to the nomenclature outlined in Materials and Methods. The EPB2 ligand ($QSVEV_{COOH}$) is shown in dark blue and the residues are numbered at their Cα-Atoms.
(**C-F**) Mapping of the chemical shift perturbations on the AF6 PDZ domain surface for the C-termini of (**C**) EPB2, (**D**) EPA7, (**E**) CIN4 and (**F**) FXI1. Color-code according to the strengths of the chemical shift perturbations as in (**A**). The surface representations are given in the same orientation as in (**B**). White rectangles indicate missing chemical shifts of the GLGF-loop.

Additionally, the chemical shift perturbation data of the NMR titrations with these four peptides were also used to determine K_d values, which were consistent with the ones determined by surface plasmon resonance (SPR, Table 4.3).

Experiments with the C-terminal carboxamide derivates of the tested peptides yielded only negligible chemical shift perturbations (0.019 ppm ± 0.003 ppm), highlighting the importance of the ligand C-terminal carboxyl group. The scrambled

(VMSVQINEQAQ$_{COOH}$) and the truncated version (AQMNQIQ$_{COOH}$) of the EPB2 C-terminal peptide, as a negative control, showed only insignificant perturbations.

Table 4.3 Dissociation Constants of Representative Ligands.

Domain		Ligand		K_d [µM] determined by		
Name	SP.en.	Acc.no.	Sequence	SPR	NMR	ANOVA prediction
AF6	RGSC	O14924	PKTSAHH**ATFV**$_{COOH}$	**14.8** (12.0 – 18.2)		**10.4** (8.2 – 13.2)
	EPA7	Q15375	MLHLHGT**GIQV**$_{COOH}$	**26.9** (20.3 – 35.6)	**33.1** (29.5 – 37.2)	**30.0** (23.6 – 38.2)
	FXI1	Q12951	GVLYPRE**GTEV**$_{COOH}$	**33.4** (28.3 – 39.5)	**67.8** (59.9 – 76.7)	**67.0** (52.7 – 85.3)
	BCR	P11274	RQSILF**STEV**$_{COOH}$	**64.1** (58.4 – 70.3)		**208.1** (163.6 – 264.8)
	CIN4	P35499	TVRPGVK**ESLV**$_{COOH}$	**80.2** (77.4 – 83.2)	**83.9** (82.5 – 85.4)	**82.1** (64.5 – 104.5)
	CIN4 (short)	P35499	GVK**ESLV**$_{COOH}$	**98.5** (81.6 – 118.9)		**82.1** (64.5 – 104.5)
	PTPZ	P23471	GNIAESL**ESLV**$_{COOH}$	**113.5** (101.4 – 127.1)		**82.1** (64.5 – 104.5)
	REL	Q04864	DSFPYE**FFQV**$_{COOH}$	**118.9** (112.0 – 126.2)		**474.8** (373.1 – 604.1)
	EPB2	P29323	IQ**SVEV**$_{COOH}$	**122.8** (108.9 – 138.3)	**137.4** (119.2 – 158.5)	**194.6** (152.9 – 247.6)
	TAT	P03409	SEKHFR**ETEV**$_{COOH}$	**123.4** (94.4 – 161.3)		**204.0** (160.3 – 259.6)
	ERB2	P04626	PEYLGL**DVPV**$_{COOH}$	**297.0** (261.5 – 337.4)		**625.4** (491.5 – 795.7)
	CYG4 (short)	P33402	FLR**ETSL**$_{COOH}$	**407.9** (396.8 – 419.3)		**289.6** (227.6 – 368.5)
ERBIN	ARVC	O00192	DAKPQPV**DSWV**$_{COOH}$	**7.6** (5.0 – 11.4)		**8.3** (5.4 – 12.7)
	TAT	P03409	SEKHFR**ETEV**$_{COOH}$	**14.8** (12.5 – 17.5)		**50.7** (33.1 – 77.6)
	APC	P25054	HSGSYL**VTSV**$_{COOH}$	**19.2** (18.2 – 20.3)		**74.5** (48.7 – 114.1)
	ATB1	P20020	GSPLHSL**ETSL**$_{COOH}$	**24.2** (19.8 – 29.7)		**62.0** (40.5 – 94.9)
	BCR	P11274	RQSILF**STEV**$_{COOH}$	**41.9** (29.9 – 58.6)		**173.7** (113.4 – 265.9)
	CTNB	P35222	SNQLAWF**DTDL**$_{COOH}$	**53.0** (52.6 – 53.5)		**232.0** (151.5 – 355.3)
	ATB2	Q01814	GSPIHSL**ETSL**$_{COOH}$	**54.1** (49.5 – 59.1)		**62.0** (40.5 – 94.9)
	NME2	Q13244	YKKLSSI**ESDV**$_{COOH}$	**57.1** (55.6 – 58.7)		**23.1** (15.1 – 35.4)
	CIN4 (short)	P35499	GVK**ESLV**$_{COOH}$	**67.4** (57.7 – 78.8)		**119.1** (77.8 – 182.4)
	CYG4 (short)	P33402	FLR**ETSL**$_{COOH}$	**75.4** (27.8 – 205.1)		**62.0** (40.5 – 94.9)
	CIK5	P22459	LCLDTSR**ETDL**$_{COOH}$	**79.3** (76.6 – 82.1)		**50.4** (32.9 – 77.2)
	CIN4	P35499	TVRPGVK**ESLV**$_{COOH}$	**82.6** (81.7 – 83.6)		**119.1** (77.8 – 182.4)
	ERB2	P04626	PEYLGL**DVPV**$_{COOH}$	**105.0** (81.9 – 134.7)		**181.7** (118.7 – 278.2)
	CYG4	P33402	IGTMFLR**ETSL**$_{COOH}$	**115.3** (106.5 – 124.9)		**62.0** (40.5 – 94.9)
	AVR2	P27037	NVDFPPK**ESSL**$_{COOH}$	**173.7** (142.3 – 211.9)		**261.3** (170.6 – 400.1)
SNA1	CYG4	P33402	IGTMFLR**ETSL**$_{COOH}$	**2.6** (2.2 – 3.2)		**11.9** (11.3 – 12.6)
	A1AD	P25100	ADYSNLR**ETDI**$_{COOH}$	**2.7** (2.4 – 3.1)		**15.8** (15.0 – 16.6)
	CIN4	P35499	TVRPGVK**ESLV**$_{COOH}$	**45.7** (39.2 – 53.4)		**20.8** (20.6 – 22.8)
	MCM7	P33993	VNASRTR**ITFV**$_{COOH}$	**n.d.**		**59.6** (56.6 – 62.7)

Footnotes: SP.en.: Swiss-Prot database entry name with the '_HUMAN' suffix omitted. Acc.no.: Swiss-Prot database accession number. Sequence: the four C-terminal residues, which are used for the predictions, are highlighted in bold. SPR: K_d measurements by surface plasmon resonance (n.d. = not determinable). NMR: K_d measurements by nuclear magnetic resonance. ANOVA prediction: K_d prediction by Analysis of Variance regression models. For the AF6 and ERBIN PDZ domain, the predictions were performed using the fixed effects models including pair-wise statistical interaction terms and for the SNA1 PDZ domain with the mixed effects model without interaction terms. The lower and upper boundaries for each K_d value are given in parentheses. For the experimental K_d values these boundaries represent the +/- 1σ interval, for the ANOVA prediction it is the 95% confidence interval, both on log-scale.

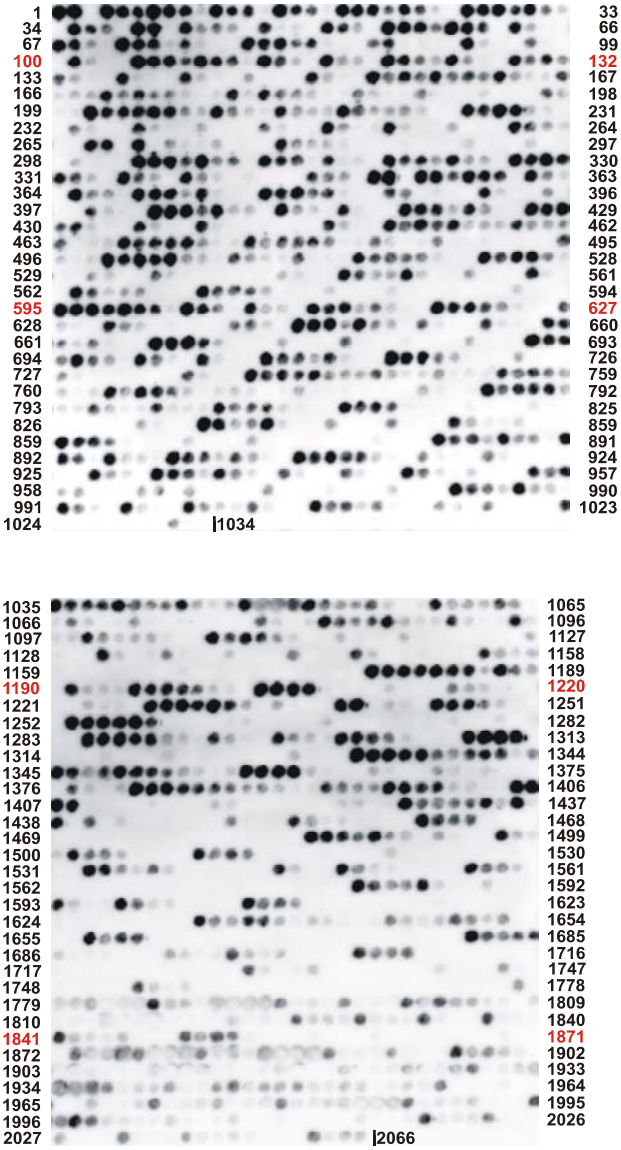
Taken together, the chemical shift perturbation data indicate that the AF6 PDZ domain binds the different peptides via the well-known conserved PDZ domain/ligand binding mechanism, employing identical surface areas.

4.3.4 Prediction of Ligand Specificity by 'Term Schemes'

To quantify the amino acid type-specific affinity contributions of residues in the four C-terminal ligand positions for the AF6 and ERBIN PDZ domains, profile libraries of the type $bbbbB_{-3}B_{-2}B_{-1}B_{0COOH}$ were designed. The identified target sequence sub-space was covered by selecting B_{-3} = A/D/E/F/G/P/S, B_{-2} = F/I/L/S/T/V, B_{-1} = D/E/F/I/L/N/P/Q/S/T/W/Y and B_0 = F/I/L/V (2016 peptides) for the AF6 profile library. For the ERBIN profile library, B_{-3} = D/E/G/S/V, B_{-2} = I/S/T/V, B_{-1} = A/D/E/F/L/P/S/W/Y and B_0 = I/L/V, were permuted resulting in 540 peptides (Figure 4.6).

We calibrated the measured BLU intensities of these libraries using control peptides (each in 5 replicates), which K_d values were determined by SPR before (Table 4.3). Predictive specificity models for the four C-terminal residues were then generated by a fixed effects [126, 127] Analysis of Variance (ANOVA) relating the amino acid sequence of these four positions to the corresponding calibrated BLU intensities obtained from the profile libraries. To analyze the cooperativity of the individual ligand positions, we compared models with and without pair-wise statistical interaction terms (Table 4.4). For both domains, cooperativity between ligand positions is low and accounts only for approximately 10% of the affinity variation. In case of the AF6 PDZ domain, all 4 ligand positions showed weak cooperativity (6 interaction terms), whereas there was no significant cooperativity between ligand position -1 und -2 in case of the ERBIN PDZ domain.

Due to the minimal cooperativity, we included the BLU data from the substitutional analyses to extent the ANOVA models to 19 amino acids (without cysteine) in order to cover the complete ligand sequence space. Since the substitutional analyses were generated independently, we analyzed the dataset consisting of the profile library and selected substitutional analyses using mixed effects [128] ANOVA models (Table 4.4) to cope with the systematic error introduced by using data from different experiments.



Spot-No	BLU [10^4]	Sequence
100	16.9	bbbbGTPF
101	90.6	bbbbGTSV
102	29.1	bbbbGTSI
103	23.6	bbbbGTSL
104	26.3	bbbbGTSF
105	223.0	bbbbGVFV
106	117.5	bbbbGVFI
107	89.8	bbbbGVFL
108	50.6	bbbbGVFF
109	137.9	bbbbGVLV
110	73.8	bbbbGVLI
111	52.6	bbbbGVLL
112	20.9	bbbbGVLF
113	132.5	bbbbGVTV
114	54.9	bbbbGVTI
115	23.1	bbbbGVTL
116	17.9	bbbbGVTF
117	111.0	bbbbGVIV
118	30.1	bbbbGVII
119	24.3	bbbbGVIL
120	23.2	bbbbGVIF
121	107.7	bbbbGVYV
122	66.4	bbbbGVYI
123	38.3	bbbbGVYL
124	27.2	bbbbGVYF
125	81.4	bbbbGVNV
126	28.4	bbbbGVNI
127	16.6	bbbbGVNL
128	16.4	bbbbGVNF
129	84.6	bbbbGVWV
130	37.1	bbbbGVWI
131	56.0	bbbbGVWL
132	30.1	bbbbGVWF
595	126.3	bbbbPIFV
596	1375	bbbbPIFI
597	155.4	bbbbPIFL
598	86.6	bbbbPIFF
599	156.5	bbbbPILV
600	76.2	bbbbPILI
601	57.8	bbbbPILL
602	19.6	bbbbPILF
603	158.3	bbbbPITV
604	68.5	bbbbPITI
605	32.6	bbbbPITL
606	15.2	bbbbPITF
607	71.2	bbbbPIIV
608	23.9	bbbbPIII
609	19.6	bbbbPIIL
610	14.7	bbbbPIIF
611	96.4	bbbbPIYV
612	60.8	bbbbPIYI
613	70.4	bbbbPIYL
614	25.4	bbbbPIYF
615	39.6	bbbbPINV
616	21.2	bbbbPINI
617	14.4	bbbbPINL
618	15.0	bbbbPINF
619	107.1	bbbbPIWV
620	63.1	bbbbPIWI
621	74.3	bbbbPIWL
622	47.1	bbbbPIWF
623	30.5	bbbbPIDV
624	15.9	bbbbPIDI
625	12.0	bbbbPIDL

Spot-No	BLU [10^4]	Sequence
626	10.2	bbbbPIDF
627	37.6	bbbbPIEV
1190	15.8	bbbbEITF
1191	83.3	bbbbEIIV
1192	25.4	bbbbEIII
1193	23.4	bbbbEIIL
1194	22.4	bbbbEIIF
1195	92.8	bbbbEIYV
1196	75.5	bbbbEIYI
1197	76.7	bbbbEIYL
1198	64.8	bbbbEIYF
1199	43.2	bbbbEINV
1200	19.8	bbbbEINI
1201	15.2	bbbbEINL
1202	15.1	bbbbEINF
1203	148.4	bbbbEIWV
1204	120.0	bbbbEIWI
1205	108.8	bbbbEIWL
1206	88.8	bbbbEIWF
1207	15.9	bbbbEIDV
1208	13.3	bbbbEIDI
1209	12.8	bbbbEIDL
1210	13.0	bbbbEIDF
1211	18.7	bbbbEIEV
1212	14.3	bbbbEIEI
1213	12.9	bbbbEIEL
1214	12.6	bbbbEIEF
1215	35.8	bbbbEIQV
1216	24.6	bbbbEIQI
1217	17.9	bbbbEIQL
1218	11.7	bbbbEIQF
1219	9.1	bbbbEIPV
1220	13.0	bbbbEIPI
1841	56.7	bbbbFTYV
1842	19.6	bbbbFTYI
1843	11.9	bbbbFTYL
1844	10.0	bbbbFTYF
1845	18.0	bbbbFTNV
1846	7.1	bbbbFTNI
1847	6.9	bbbbFTNL
1848	7.7	bbbbFTNF
1849	42.4	bbbbFTWV
1850	23.3	bbbbFTWI
1851	45.1	bbbbFTWL
1852	29.0	bbbbFTWF
1853	7.4	bbbbFTDV
1854	6.3	bbbbFTDI
1855	5.8	bbbbFTDL
1856	5.8	bbbbFTDF
1857	6.6	bbbbFTEV
1858	5.0	bbbbFTEI
1859	4.7	bbbbFTEL
1860	4.8	bbbbFTEF
1861	6.6	bbbbFTQV
1862	5.2	bbbbFTQI
1863	5.2	bbbbFTQL
1864	5.3	bbbbFTQF
1865	5.3	bbbbFTPV
1866	5.3	bbbbFTPI
1867	4.7	bbbbFTPL
1868	4.6	bbbbFTPF
1869	6.3	bbbbFTSV
1870	4.1	bbbbFTSI
1871	3.4	bbbbFTSL

Figure 4.6 Permutation of the Last Four C-Terminal Ligand Residues of the AF6 PDZ Domain for K_d Prediction.

We selected the amino acid preferences of the last 4 C-terminal ligand residues based on their frequency deduced from the 6223-Humlib and the substitutional analysis. For the AF6 profile library of the type $bbbbB_{-3}B_{-2}B_{-1}B_{0COOH}$, we choose B_{-3} = A/D/E/F/G/P/S, B_{-2} = F/I/L/S/T/V, B_{-1} = D/E/F/I/L/N/P/Q/S/T/W/Y and B_0 = F/I/L/V (7x6x12x4=2016 peptides) for position -3 through 0, respectively. Position B_0 was permuted first with the above indicated amino acids, then position B_{-1}, B_{-2} and finally position B_{-3}. An additional 30 control peptides (with measured K_d values) were inserted into the profile library next to those peptides having the same four C-terminal amino acids (in total 2046 peptides). Residues denoted with b represent a mixture of all 17 L-amino acids (except C, M and W).
Selected peptide rows of the profile library (highlighted in red) were exemplarily listed in the Table on the right hand. The measured BLU values of the permutated sequences and these of the control peptides were used to calibrate them to the corresponding K_d values.

The mixed effects models agreed well with the corresponding fixed effects ANOVA models, allowing us to base the analysis of the SNA1 PDZ domain solely on the intensity data from substitutional analyses and, furthermore, to include ligand position -4 [34, 104].

The resulting specificity profiles deduced from the final mixed effects models are visualized as 'term schemes' in Figure 4.7 (D-F). The presented relative affinity contributions for all amino acids (without cysteine) in each of the four C-terminal ligand positions allow to estimate the K_d change for each ligand sequence with respect to the reference peptide (K)ESLV$_{COOH}$ (CIN4).

Table 4.4 Statistical Significance and Diagnostics of the Dataset.

Model	Data source	Observations	Fit (R^2)	SDEP	Est. F_{error}	Valid. R^2
AF6						
Fixed effects without IT	Profile Library	2016	0.80	0.50	1.65	0.67
Fixed effects with IT	Profile Library	2016	0.90	0.35	1.42	0.76
Mixed effects without IT	18 SubAna + Profile Library	3280 + 2016	*0.79	0.48	1.62	0.72
ERBIN						
Fixed effects without IT	Profile Library	540	0.82	0.67	1.95	0.63
Fixed effects w IT	Profile Library	540	0.91	0.47	1.60	0.62
Mixed effects without IT	28 SubAna + Profile Library	5020 + 540	*0.78	0.58	1.79	0.55
SNA1						
Mixed effects without IT	12 SubAna	2204	n.a.	0.53	1.70	0.76

Footnotes: ANOVA model diagnostics: statistical pair-wise interaction terms are abbreviated as IT. Data source: the experiments, which contributed raw data to the statistical analysis (SubAna = substitutional analyses). Obs.no.: the number of observed BLU (chemiluminescence intensity) values, each representing a single peptide sequence, which were included into the statistical analysis. Fit R^2: the coefficient of determination (proportion of explained variance) for the fit of the model to the training dataset (calibrated BLU). * For mixed effects models, the Fit R^2 is not well defined, therefore we estimated it by determining the fit of the predictions to the profile library calibrated BLU values. For the SNA1 mixed effects model there was not even a profile library available (n.a. = not applicable). SDEP: standard deviation of error of prediction. Est. F_{error}: multiplicative error of K_d on linear scale estimated from SDEP. Valid. R^2: validated coefficient of determination deduced from the fit of the predicted to the experimentally determined ln(K_d) values.

To validate these models we compared the predicted and experimentally determined K_d values (Table 4.4 and Figure 4.8). For the AF6 and ERBIN PDZ domains the predicted K_d values of the fixed effects ANOVA models deviated in general by no more than a factor of 2 (Table 4.3) from experimentally determined ones, which is in the range of the estimated multiplicative error (Table 4.4). For the AF6 PDZ domain, there are a few exceptions with deviations up to a factor of 4 (BCR, REL, ERB2) and some for the ERBIN PDZ domain deviating by up to a factor of 4.4 (TAT, APC, ATB1, BCR, CTNB, NME2) (Table 4.3). The SNA1 mixed effects model predictions deviated by up to a factor of 5.8. The quality variation of the model can be attributed to the lower number of experiments on which it was based.

Additionally, we built cross-validated models based on reduced data sets. The resulting amino acid type-specific affinity contributions of these models deviated by no more than the 95% confidence interval bandwidths from the 'full' models presented above. We then compared the predictions of cross-validated models with the experimentally determined K_d values. The calculated Q^2 (cross-validated coefficient of determination) did not change significantly with respect to the R^2 (coefficient of determination) determined from the 'full' models (Table 4.4).

4.3.5 Quantification of PDZ Domain/Ligand Specificity

The specificity profiles visualized as 'term schemes' represent the relative amino acid type-dependent affinity contributions ($\Delta\ln(K_d)$) of the ligand residues from position 0 through -3, which correspond to the relative energetic contributions ($\Delta\Delta G^0$). The distribution of the energetic contributions at each position reflects the importance of the respective position for binding specificity (Figure 4.7 (D-F)). Highly selective positions, such as position 0, are characterized by only a few amino acids with low $\Delta\ln(K_d)$, separated far from a tight cluster of all other amino acids.

Less selective positions, such as position -1 in the case of AF6, show a uniformly spread distribution of amino acids. The 'term schemes' include the classical motifs $(\pi/E)(\Psi/T/S)x\Psi_{COOH}$ for the AF6, $(E/D)(T/S/V)x(V/L/I)_{COOH}$ for the ERBIN and $(R/K)E(T/S)x(V/I/L)_{COOH}$ for the SNA1 PDZ domains as combinations of those amino acids which show the lowest $\Delta\ln(K_d)$.

Nevertheless, in these classical sequence motifs, position -1 is considered mostly unspecific, whereas our models show that each of the four C-terminal positions may contribute strongly to the total affinity, depending on the presented amino acid type. At physiological levels of K_d, optimal amino acids in less selective positions, such as W in position -1 for the ERBIN PDZ domain [161], may compensate for non-optimal amino acids in other positions.

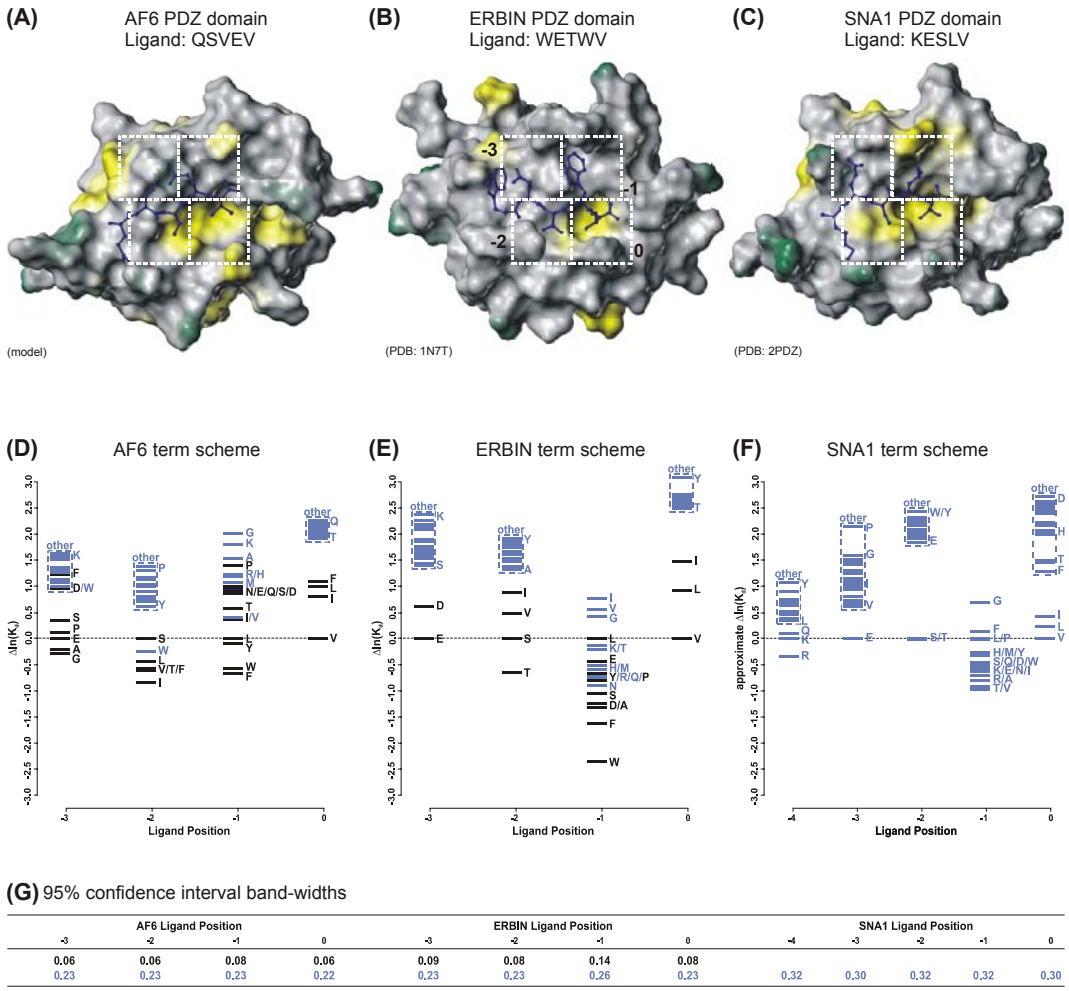


Figure 4.7 Contribution of Conserved PDZ Domain Interaction Areas towards Specific Ligand Binding.

The three complexed PDZ domain models revealed four conserved interaction areas for the four C-terminal ligand residues. Based on this fact, it was possible to predict the affinity contributions of each amino acid (19^4) in these four ligand residues and to determine their energetic contributions, which reflect the importance of the respective positions to for binding specificity.

The interaction areas IA_0 through IA_{-3} are indicated on the PDZ domain surfaces with peptide ligands in blue. (**A**) AF6 PDZ domain complexed with the peptide $QSVEV_{COOH}$ (modeled). (**B**) ERBIN PDZ domain complexed with the peptide $WETWV_{COOH}$ (PDB: 1N7T). (**C**) SNA1 PDZ domain complexed with the peptide $KESLV_{COOH}$ (PDB: 2PDZ). Surface coloring indicates hydrophobic (yellow) and hydrophilic (green) areas.

The specificity profiles for (**D**) the AF6, (**E**) the ERBIN and (**F**) the SNA1 PDZ domains are shown as 'term schemes'. The amino acid type-specific relative affinity contributions ($\Delta\Delta G_0 = \ln(K_d(\text{peptide}) / K_d(\text{reference})) = \Delta\ln(K_d)$) obtained from the mixed effects ANOVA models are plotted versus ligand positions with respect to the C-terminal residue. All contributions are relative to the arbitrary reference $(K)ESLV_{COOH}$ (CIN4) indicated by a horizontal dashed line. Black bars indicate contributions deduced from the profile libraries while those based solely on data from the substitutional analyses are presented in blue. Clusters of amino acids with high $\Delta\ln(K_d)$ are boxed and labeled "other". The K_d for a ligand with a given sequence can be deduced from the 'term scheme' by summing up the corresponding amino acid type-specific affinity contributions ($\Delta\ln(K_d)$) for each position, taking *e* to the power of this sum and multiplying the resulting factor by the K_d of the reference (CIN4: for AF6 80.2 µM, for ERBIN 82.6 µM and for SNA1 45.7 µM): $K_d = K_{d(CIN4)}\, e^{\Sigma\Delta\ln(Kd)}$. The calculations for all ANOVA models can be performed online at http://www.fmp-berlin.de/nmr/pdz.

(**G**) The approximate 95% confidence interval bandwidths (±) are given for the amino acid type-specific contributions in each position of the 'term schemes'. The upper black row gives the confidence intervals for the corresponding black bars, whereas the lower blue rows the ones for blue bars.

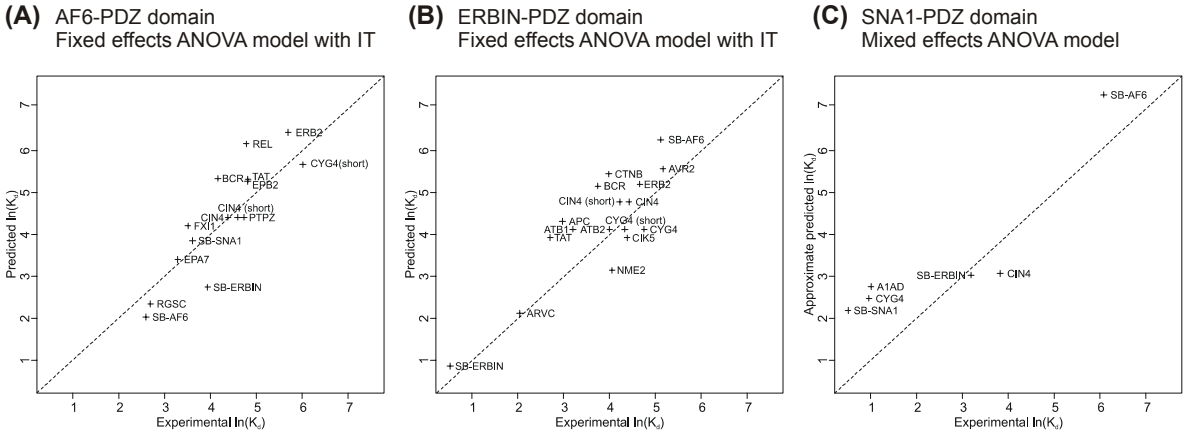


Figure 4.8 Plot of Predicted versus Experimentally Determined K_d Values.
The predictions of the fixed effects ANOVA models including interaction terms are plotted for (**A**) the AF6 and (**B**) the ERBIN PDZ domains. (**C**) The predictions of the mixed effects ANOVA model without interaction terms are plotted for the SNA1 PDZ domain. The peptide sequences are given Table 4.3.

4.3.6 Rational Design of Super-Binding PDZ Domain Ligands

The models derived above were used to find peptides with highest affinity (super-binders) of all 19^4 possible C-terminal peptide sequences, when the last foru residues are considered. The four C-terminal residues for all three super-binders were derived from the ANOVA models by combining those amino acids with the lowest $\Delta\ln(K_d)$ (Figure 4.7 (D-F)). The same super-binding peptides could be obtained from the models with and without pair-wise statistical interaction terms. In addition, we deduced the optimal amino acids in positions upstream of -3 from a semi-quantitative evaluation of the substitutional analyses. The three super-binding peptides (SB-AF6: $LEGIFV_{COOH}$, SB-ERBIN: $WLETWV_{COOH}$, SB-SNA1: $IRETIV_{COOH}$) were indeed found to bind to the respective domains with the highest affinities as compared to all other investigated peptides (Table 4.3 and Table 4.5). The predicted K_d values for the AF6 and ERBIN super-binders deviate by no more than a factor of 2 from the experimental ones. Again, larger errors were observed for the less determinate SNA1 model. The super-binders were then subjected to substitutional analyses, each incubated with all three PDZ domains (Figure 4.9).

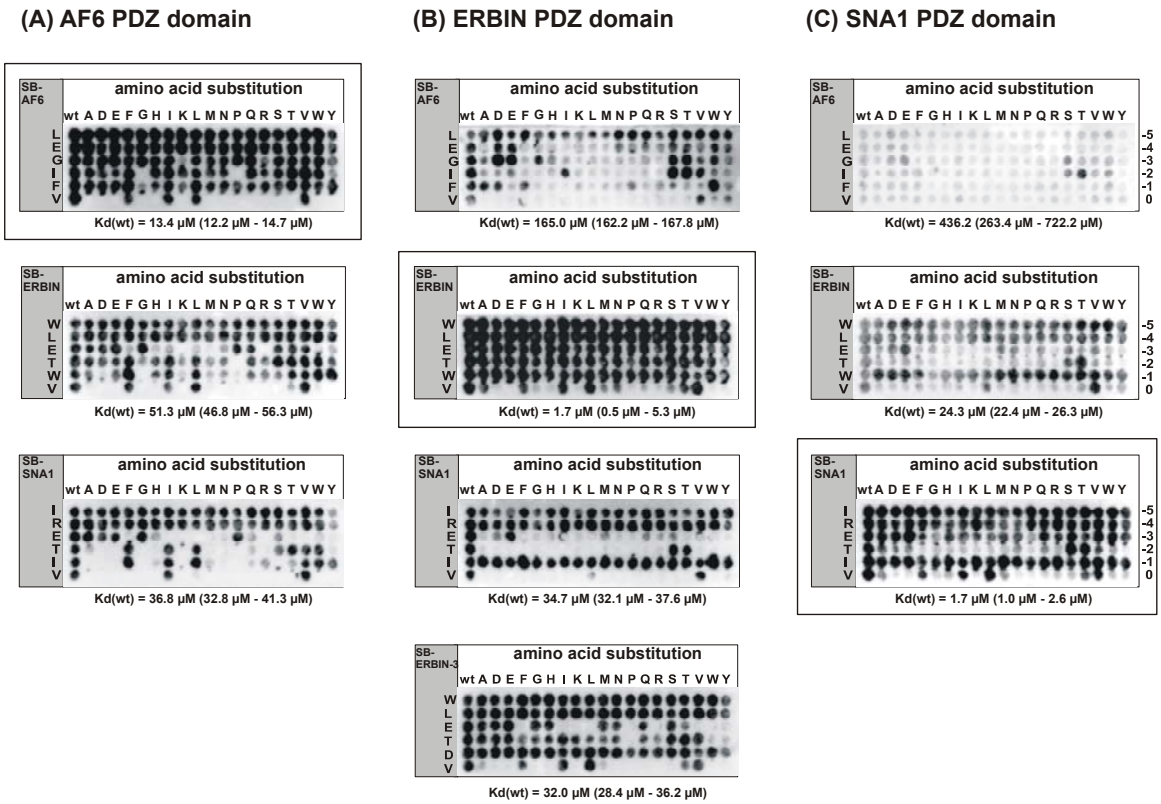


Figure 4.9 Characterization of the Specific PDZ Domain Super-Binders.
All three PDZ domains recognized their super-binders with the highest affinity. Compared to the other super-binders, the SB-AF6 is more specific for the AF6 PDZ domain as for the other PDZ domains which is reflected, in the substitutional pattern and the K_d measurements. On the other hand, the SNA1 PDZ domain seems to be the more selective for its own super binding peptide as the other PDZ domains.
Incubations of (**A**) the AF6 PDZ domain, (**B**) the ERBIN PDZ domain and (**C**) the SNA1 PDZ domain with the super-binding C-terminal peptides SB-AF6 (top), SB-ERBIN (middle) and SB-SNA1 (bottom). All spots in the left hand columns (grey box) are identical and represent the wild type (wt) peptide. All other spots represent the substitution of one amino acid against the amino acid in the respective column. Hence, each spot bears a single substitution compared to the sequence of the wild type. The last six C-terminal ligand positions are numbered to the right of the substitutional analyses. The experimentally determined K_d values of the wild type (wt) peptides are given below each substitutional analysis, with the respective +/- 1σ interval on log-scale. The super-binders of each PDZ domain are boxed.

The experimentally determined K_d values (Table 4.5) and the substitutional analysis in Figure 4.9 show that the ERBIN and SNA1 PDZ domains recognize their own super-binding ligand more specifically than the AF6 PDZ domain. (Figure 4.9, vertical columns). However, the super-binder of AF6 showed the least cross-reactivity (Figure 4.9, horizontal columns). Altogether, these substitutional analyses gave a rough estimate of the overlap between the individual PDZ domain specificity profiles.

Table 4.5 Dissociation Constants of Predicted Super-Binders.

			K_d [µM] determined by					
Domain	**Ligand**	**Sequence**	**SPR**			**ANOVA Prediction**		
AF6	SB-AF6	LE**GIFV**$_{COOH}$	**13.4**	(12.1 –	14.1)	**7.6**	(5.9 –	9.6)
	SB-SNA1	IR**ETIV**$_{COOH}$	**36.8**	(32.8 –	41.3)	**46.8**	(36.8 –	59.6)
	SB-ERBIN	WL**ETWV**$_{COOH}$	**51.3**	(46.8 –	56.3)	**15.5**	(12.1 –	19.7)
ERBIN	SB-ERBIN	WL**ETWV**$_{COOH}$	**1.7**	(0.5 –	5.3)	**2.4**	(1.5 –	3.6)
	SB-SNA1	IR**ETIV**$_{COOH}$	**34.7**	(32.1 –	37.6)	* **153.2**	(121.4 –	193.3)
	SB-AF6	LE**GIFV**$_{COOH}$	**165.0**	(162.2 –	167.8)	**524.9**	(342.8 –	803.7)
	SB-ERBIN-2	WL**ETFV**$_{COOH}$	**22.9**	(20.3 –	25.1)	**5.1**	(3.4 –	7.9)
	SB-ERBIN-3	WL**ETDV**$_{COOH}$	**32.0**	(28.4 –	36.2)	**10.9**	(7.1 –	16.7)
SNA1	SB-SNA1	IR**ETIV**$_{COOH}$	**1.7**	(1.0 –	2.6)	**8.9**	(6.5 –	12.2)
	SB-ERBIN	WL**ETWV**$_{COOH}$	**24.3**	(22.4 –	26.3)	**20.8**	(13.5 –	31.9)
	SB-AF6	LE**GIFV**$_{COOH}$	**436.2**	(263.4 –	722.2)	**1540.8**	(909 –	2610)

Footnotes: Ligand: the super-binders are abbreviated as SB- followed by the cognate PDZ domain name. Sequence: the four C-terminal residues, which are used for the predictions, are highlighted in bold. SPR: K_d measurements by surface plasmon resonance. ANOVA prediction: K_d prediction by Analysis of Variance regression models. For the AF6 and ERBIN PDZ domain, the predictions were performed using the fixed effects models including pair-wise statistical interaction terms. * For the ERBIN/SB-SNA1 prediction we had to resort to the mixed effects model, since I in position -1 was not included in the ERBIN profile library. For the SNA1 PDZ domain we used the mixed effects model without interaction terms. The lower and upper boundaries for each K_d value are given in parentheses. For the experimental K_d values these boundaries represent the +/- 1σ interval, for the ANOVA prediction it is the 95% confidence interval, both on log-scale.

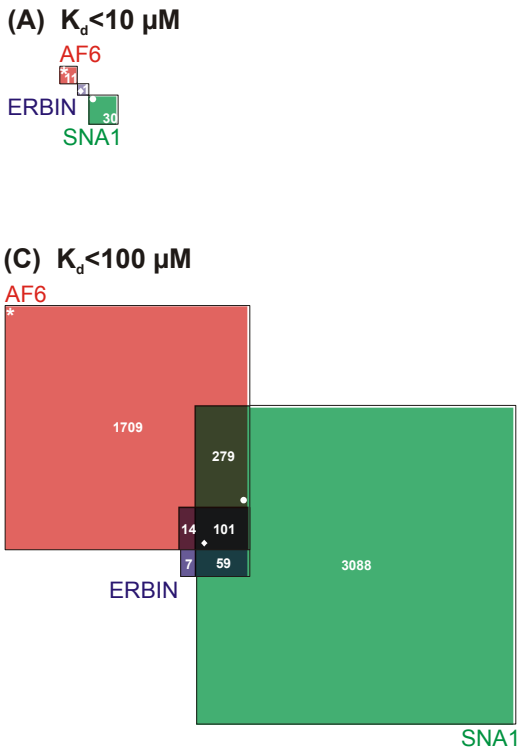
We extended this study by controlling the relevance of ERBIN PDZ domain ligand position -1 with two mutated peptides: WLET**F**V$_{COOH}$ (SB-ERBIN-2) and WLET**D**V$_{COOH}$ (SB-ERBIN-3). Kd measurements (Table 4.5) reflected the importance of W at ligand position -2 demonstrated by the interactions of ERBIN/ARVC with 7.6 µM (5.0 µM - 11.4 µM) and of ERBIN/SB-ERBIN with 1.7 µM (0.5 µM - 5.3 µM). The mutated peptides (ERBIN-2 and ERBIN-3 in Table 4.5) clearly demonstrated the preference of an aromatic instead of a charged amino acid. The substitutional analysis of WLET**D**V$_{COOH}$ (Figure 4.9) shows more selectivity in positions -3 and -2 in comparison to the SB-ERBIN peptide, confirming the high contribution of the W to the total ligand affinity.

4.3.7 Quantification of PDZ Domain/Ligand Selectivity

For the three investigated PDZ domains, the selectivities below a given K_d were determined by calculating the affinities towards all 130321 (19^4) four-residue C-terminal

ligand sequences using the ANOVA models. Figure 4.10 summarizes the number of ligand sequences recognized by each domain below K_d values of 10, 50 and 100 µM. Up to a K_d of 10 µM, the three PDZ domains target separate sequence spaces (Figure 4.10 (A)). In this K_d range, the SNA1 PDZ domain recognizes 30 ligands, the AF6 PDZ domain 11 and the ERBIN PDZ domain only one, namely its super-binder $ETWV_{COOH}$. At higher K_d values, the ligand sequence spaces recognized by the individual PDZ domains overlap (Figure 4.10 (B-C)). The SNA1 and AF6 PDZ domains recognize a large number of ligands whereas the ERBIN PDZ domain binds only few peptides, albeit showing the largest overlap (Numbers are given in the overlapping areas in Figure 4.10 (B-C)). The relative overlap predicted by the ANOVA models is in good agreement with the overlap determined for the top 100 peptides in the 6223-Humlib screens. Furthermore, the ANOVA models predict K_d values below 100 µM for 51 % of the AF6 top 100 peptides, for 29 % of the ERBIN top 100 peptides and for 63 % of the SNA1 top 100 peptides of the 6223-Humlib screens.

The ligand sequence spaces represented by the planes in Figure 4.10 (A-C) contain both class I and class II ligands as shown in Figure 4.10 (D). The AF6 PDZ domain favors class II ligands whereas SNA1 and ERBIN prefer class I motifs. In all cases, the ligands of the less-preferred class appear with less than 20%, even at increasing K_d. In addition, all three PDZ domains recognize sequences, which are not part of these two classes (denoted ‘non-class I/II’ in Figure 4.10 (D)). These ligands are characterized by a higher frequency of the optimal amino acids in positions -1 and -3 (Figure 4.10 (D-F)) and contain always one of the preferred amino acids either in position 0 or -2. The portion of these sequences increases with increasing K_d, while the fraction of peptides from the favored class decreases. The recognized ligand sequence spaces also differ in the location of the affinity maxima (represented by the super-binders) with respect to the overlapping area. The AF6 super-binder is located far away from the overlapping area. On the other hand, the SB-SNA1 is in the overlapping area between SNA1 and AF6 above a predicted K_d of ~50 µM, whereas the SB-ERBIN is recognized by all three PDZ domains down to a predicted K_d threshold of ~20 µM.



(D)

Kd	<10 µM	(a)	<50 µM	(b)	<100 µM	(c)
AF6	**11**		**602**		**2103**	
Class I	2	(18%)	113	(19%)	308	(15%)
Class II	9	(82%)	416	(69%)	1070	(51%)
Non-class I/II	0	(0%)	73	(12%)	725	(34%)
ERBIN	**1**		**48**		**181**	
Class I	1	(100%)	43	(90%)	131	(72%)
Class II	0	(0%)	4	(8%)	25	(14%)
Non-class I/II	0	(0%)	1	(2%)	25	(14%)
SNA1	**30**		**1622**		**3527**	
Class I	30	(100%)	1492	(92%)	2003	(57%)
Class II	0	(0%)	4	(0%)	161	(5%)
Non-class I/II	0	(0%)	126	(8%)	1363	(39%)

Figure 4.10 PDZ Domain Selectivity and Ligand Sequence Space Overlap.
We calculated the K_d values of all 130321 (19^4) C-terminal ligand sequences for the three investigated PDZ domains using the ANOVA model. The recognized sequence space of Kd values below 10 µM (**A**), 50 µM (**B**) and 100 µM (**C**) is represented by planes colored red for the AF6, blue for the ERBIN and green for the SNA1 PDZ domain. The overlap is indicated by the corresponding mixed colors. The numbers of ligand sequences are given in the corresponding areas. The super-binders are indicated by * for AF6, ♦ for ERBIN and • for SNA1.
(**D**) The number of ligand sequences corresponding to the PDZ domain classification according to Songyang *et al.* [30] with class I x(S/T)x(V/I/L)$_{COOH}$, class II xΦxΦ$_{COOH}$ and those corresponding to neither of these two classes ('non-class I/II') are summarized in the Table. The sequences corresponding to the individual areas can be accessed online at http://www.fmp-berlin.de/nmr/pdz.

4.4 Discussion

A notable fraction of PDZ-protein interactions are mediated by a limited number of adaptor domain modules. Classification schemes based on recognized ligand motifs were proposed for SH2, SH3, WW, PDZ and other domains, to partition the ligand sequence space [17]. For PDZ domains, the classification schemes do neither explain all experimental results, nor the respective affinities. The analysis presented here introduces specificity models based on a statistical analysis of *in vitro* binding studies using synthetic peptide libraries. Based on these models, it is now possible to rationalize why PDZ domains bind to a diverse set of peptide sequences.

4.4.1 Super-Binding Peptides

The specificity models for all three investigated PDZ domains allowed us to design super-binding peptides that exhibited the lowest experimental K_d values of all peptides analyzed here. In particular, the ERBIN super-binder showed a 4-fold higher affinity compared to the ARVC peptide (Table 4.3 and Table 4.5). The four C-terminal residues of the SB-ERBIN are identical to the result of a previous phage-display ligand screen [161]. Interestingly, the highest achievable affinity for peptide ligands of the three investigated PDZ domains is in the range of 1 - 10 μM. Furthermore, these super-binding peptide sequences do not match any C-termini of proteins in the current protein databases.

Principally, PDZ domain interactions are involved in the sorting, targeting, and assembly of supramolecular complexes. For this reason, the K_d values are mainly in the micromolar range enabling the switch between their different ligands. Therefore, the super binding peptides could be useful as competitors for specific PDZ-protein interactions in cell biology experiments and also in drug design.

4.4.2 PDZ Domain Selectivity and Overlap

The specificity profiles (‘term schemes’) can also be used to compare the recognized ligand sequence spaces of the three PDZ domains by calculating the affinities towards the complete four residues sequence space (Figure 4.10). The SNA1 PDZ domain, for example, recognizes a much larger set of C-terminal sequences than the PDZ domain from ERBIN, which binds to comparably few sequences. It is intriguing to correlate this finding with the biological context of the corresponding proteins. SNA1, for example, is bound as a scaffolding protein to the dystrophin/utrophin network and its task may be to localize a multitude of proteins to the cytoskeleton via its PDZ domain [34]. On the other hand, ERBIN may fulfill a more specific role as a suppressor of the Ras/Raf pathway [151].

The overlap of the recognized ligand sequence space at K_d values below 50 und 100 μM is surprisingly large for the ERBIN PDZ domain. This suggests that a single C-terminal ligand sequence can bind diverse PDZ domain containing proteins with overlapping specificity profiles and thereby support different cellular functions. However, *in vivo* higher selectivity may be achieved by cooperativity of accompanying domains,

compartmentalization of proteins and therefore reduction of the number of available ligands.

4.4.3 PDZ Domain Structure-Specificity-Relationships

The comparison of the three PDZ domain structures allowed us to rationalize the different amino acid preferences in each ligand position. For this purpose, we extended the interaction area approach introduced by Songyang *et al.* [30] using the combined information from our chemical shift titration experiments and all complex structures in the protein data bank (PDB). Due to integration of the ligand into the β-sheet, the position of the backbone atoms of these four ligand residues is highly conserved, with the backbone RMSD ranging from 0.86 Å for PSD95/PDZ-3 (PDB: 1BE9) [29] to 1.95 Å for INAD/PDZ-1 (PDB: 1IHJ) [159] as compared to SNA1 (PDB: 2PDZ) [34]. Consequently, the individual ligand side chains target always the identical surface patches [30], named hereafter interaction area IA_0 through IA_{-3} (corresponding to the peptide ligand nomenclature). The interaction areas are defined as follows: IA_0 consists of GLGF:2, βB:1, βB:3, αB:5, αB:8 and αB:9; IA_{-1} contains GLGF:3, βB:1 (backbone only), βB:2, βC:4, βC:5 and βC:7; IA_{-2} consists of βB:3, αB:1, αB:2, αB:5 and αB:8; IA_{-3} contains βB:2, βB:4, βC:4 and βC:5 (Figure 4.7 (A-C)).

All three PDZ domains preferred aliphatic amino acids in position 0 due to the fact that pronounced hydrophobic pocket is situated in IA_0. However, obvious structural features do not readily explain the domain-specific discrimination, especially between V/I/L.

On the other hand, the presence of a pocket in IA_{-1} could be correlated with the ability to bind large aromatic residues in ligand position -1. For ERBIN, this pocket has already been demonstrated structurally by Skelton *et al.* [161] to selectively bind aromatic residues, especially W. In the case of AF6, the rim of this pocket is less pronounced whereas the pocket is occupied by F in βC:7 of SNA1 (Figure 4.7 (A-C)).

In contrast to SNA1 and ERBIN, the AF6 PDZ domain shows a manifest dual-specificity, tolerating both hydrophobic amino acids and T/S in ligand position -2. These interactions are mainly due to van-der-Waals contacts to the hydrophobic surface in IA_{-2}, which is indicated by the preference of T over S. This type of recognition via van-der-Waals contacts is in contrast to the hallmark of class I PDZ domains, in which T/S interact

via H-bond formation with αB:1. In this position, class I PDZ domains usually have an H, whereas the AF6 PDZ domain shows a Q. The decreasing preference for T/S in position -2 (SNA1>ERBIN>AF6) suggests a decreasing contribution of the potential H-bond involving the residue in αB:1 which might in turn be compensated by an increasing importance of van-der-Waals contacts (Figure 4.7 (D-F)).

Interaction area IA_{-3} of the AF6 domain is largely occupied by its own V in βB:4, explaining the preference for the small amino acids A, G, P and S (π) in order to avoid repulsive contacts. Glutamate is also tolerated in ligand position -3 probably due to a favorable electrostatic interaction with K in βC:4. In contrast to AF6, the ERBIN and SNA1 PDZ domains show an pronounced pocket in IA_{-3} with R (ERBIN) and K (SNA1) in βC:5, which predominantly selects E and to a minor degree also D.

In addition to the conserved interaction areas IA_0 through IA_{-3} the SNA1 and the ERBIN PDZ domains recognize specific residues in ligand positions -4 and -5, respectively. In the case of SNA1, the specificity for R and K in ligand position -4 can be attributed to interactions with aspartic acid in αB:2 [34]. In case of ERBIN, aromatic residues in positions upstream of -3 can interact with the hydrophobic pocket in the βB-βC-loop [122, 161]. This ability of the SNA1 and ERBIN PDZ domain to make further specific interactions, together with their pronounced pockets in IA_{-1} and IA_{-3} may explain the higher maximally achievable peptide ligand affinity (represented by the super-binders) compared to the AF6 PDZ domain (~10-fold, see Table 4.5).

4.4.4 Conclusions

The specificity models presented here (Figure 4.7) expand our molecular understanding of PDZ domain/ligand recognition by dissecting the ligand specificity into side chain-dependent relative affinity contributions. This facilitates, for example, the rational design of mutational experiments, as demonstrated by our prediction of super-binding peptides. Utilizing the power of predicting K_d values for any ligand sequence prior to biological experiments, a comprehensive picture of all potential PDZ domain interactions can be obtained to guide a rational experimental approach. At the same time, this permits the critical assessment of large-scale protein-protein interaction studies. In addition, the ability to predict K_d values offers a theoretical basis to model protein-protein interaction

networks within the field of systems biology. The ‘term scheme’ can be further converted into a position-specific scoring matrix that can be applied in bioinformatics analyses searching for potential ligands.

Taken together, we propose a general and efficient procedure for profiling protein-protein interaction domains that provides a biophysics-based picture of specificity and selectivity covering the complete ligand sequence space (http://www.fmp-berlin.de/nmr/pdz).

Another interesting area for future studies is to develop small molecular weight compounds that are capable of specifically disrupting certain PDZ domain-mediated interactions. Due to the fact that PDZ domains can specifically recognize small ligands (4 amino acids), the design and development of such PDZ-ligand intervening compounds is conceptually feasible.