# Potential utility of docking to identify protein-peptide binding regions

Waqasuddin Khan
waqasuddin.khan@ucdconnect.ie
H.E.J Research Institute of Chemistry,
International Center for Chemical and Biological Sciences,
University of Karachi, Pakistan

Fergal Duffy
fergal.duffy@ucd.ie
Complex and Adaptive Systems Laboratory,
Conway Institute of Biomolecular and Biomedical Sciences and
School of Medicine and Medical Science,
University College Dublin, Ireland

Gianluca Pollastri
gianluca.pollastri@ucd.ie
Complex and Adaptive Systems Laboratory and
School of Computer Science and Informatics,
University College Dublin, Ireland

Denis C. Shields
denis.shields@ucd.ie
Complex and Adaptive Systems Laboratory,
Conway Institute of Biomolecular and Biomedical Sciences and
School of Medicine and Medical Science,
University College Dublin, Ireland

Catherine Mooney
catherine.mooney@ucd.ie
Complex and Adaptive Systems Laboratory,
Conway Institute of Biomolecular and Biomedical Sciences and
School of Medicine and Medical Science,
University College Dublin, Ireland

University College Dublin, Technical Report UCD-CSI-2013-01, May 2013

**Abstract**

Disordered regions of proteins often bind to structured domains, mediating interactions within and between proteins. However, it is difficult to identify a priori the short regions involved in binding. We set out to determine if docking peptides to peptide binding domains would assist in these predictions.

First, we investigated the docking of known short peptides to their native and non-native peptide binding domains. We then investigated the docking of overlapping peptides adjacent to the native peptide. We found only weak discrimination of docking scores between native peptide and adjacent peptides in this context with similar results for both ordered and disordered regions. Finally, we trained a bidirectional recurrent neural network using as input the peptide sequence, predicted secondary structure, Vina docking score and Pepsite score.

We conclude that docking has only modest power to define the location of a peptide within a larger protein region known to contain it. However, this information can be used in training machine learning methods which may allow for the identification of peptide binding regions within a protein sequence.

# Background

Thousands of proteins expressed in cells carry out their specific intracellular and extracellular functions by interacting with each other (protein-protein interactions). These interactions have been acknowledged to play fundamental roles in almost every biological event. Significant biological processes such as protein signalling, trafficking and their synchronised degradation [9, 16], DNA repairing, replication and gene expression [30, 37] require interaction between protein-protein interfaces to perform their tasks. The extent of complexity, co-operativity and diversity for these interactions is enormous, and itself is coordinated by intricate regulatory networks that will ultimately determine the behaviour of biological systems. These interactions are adaptable, that is, many interactions are mediated between the two domains of globular proteins (domain-domain interactions) which tend to be stable (contact surfaces are flat) and require an average size of 1,500 - 3,000 Å$^2$ [10, 21]. Others are intended for fast response to stimuli (domain-motif interactions (DMI)/domain-peptide interactions/peptide-mediated interactions) [35] that occur when a globular domain (50 - 150 residues long) in one protein recognises a short linear peptide (ranging from 3 - 12 amino acid residues) from its corresponding protein partner, creating a comparatively small interface with an average size of 350 Å$^2$ [43]. An estimated 15 - 40% of all interactions in the cell are protein-peptide interactions [29, 37]. These peptide regions are ideal for signalling transduction networks because they are specific, transient and have low-affinity (1 - 150 $\mu$M) [11].

Typically, the peptides involved in DMI or protein-peptide interactions are categorised by a simple sequence pattern, that is, a short linear motif (SLiM).

In general, SLiMs can be expressed as regular expressions (RegEx), a consensus motif with specific conserved residues restricted to particular positions recognised by a binding domain, with a set of similar residues or even arbitrary ones at other locations [40]. Structurally, SLiMs are frequently found in disordered regions at protein termini or between domains [17] with the ability to adopt a variety of conformations [48, 44]. SLiMs may also originate from loops within a structured domain, exposing them to potential binding partners including many of the disordered interaction hubs [14, 19] explaining the many functional roles for these regions. The small binding areas which SLiMs constitute result in weak binding [29] making them suitable for short-lived interactions [36]. But, regardless of their short length, these motifs bind their target protein with sufficient strength to establish a functional relationship [23]. Compared with domain-domain interactions, domain-SLiMs interfaces are attractive drug targets for small molecules and designed inhibitory peptides [20, 30, 34, 53].

The estimated number of protein-peptide interactions in the genome is not reflected in the number of 3D protein-peptide crystal structures available in the Protein Data Bank (PDB) (www.pdb.org) [7]. However, the rapid increase in protein structural data in the PDB does provide an excellent opportunity to investigate how this information could be used to predict novel SLiM mediated protein-peptide interactions using this 3D structural information in conjunction with the protein's sequence.

Computational docking is widely used for the theoretical prediction of small molecule ligand-protein complex. Typically, docking involves two steps: the generation of several alternate conformations of a ligand molecule to sample all possible interaction modes with the receptor binding site, followed by an energy function or a scoring function to rank the poses. In general, docking programs are not optimised for peptide docking and are most successfully used with small molecule compounds. Unconstrained peptides are flexible and tend to adopt several conformations by rotating within the given search space of the receptor site further adding complexity to the docking protocol. Given their poor performance in ranking correct conformations it has generally been regarded to be beyond the scope of current protein docking algorithms to detect interacting partners [1, 18, 42]. Recently however, two high-throughput docking (HTD) experiments have been reported that demonstrated the use of a general docking method to detect interacting partners. Mosca *et al.* [28] used docking to identify pairs of proteins that accurately interact with each other in the *Saccharomyces cerevisiae* interactome by sampling from a large set of alternative possibilities. Wass *et al.* [52] successfully distinguished between interacting (native) and non-interacting (non-native) protein partners.

Here, we investigated if docking can be used to identify protein-peptide interactions with the objective of evaluating if docking could distinguish a peptide binding region from adjacent non-binding regions within a defined stretch of protein sequence. We evaluated the performance of AutoDock Vina [47], to assess whether it is possible to distinguish peptides in this way. First, we generated a non-redundant dataset of high-resolution protein-peptide interacting structures, and further classified the dataset on the basis of peptide length and structure.

We then performed high throughput docking of native peptides to native receptors, native peptides to all other receptors, and finally docking of overlapping peptides, generated by moving a sliding window of peptides from the peptide parent protein sequence, to the native receptor. Finally, we trained a Bidirectional Recurrent Neural Network (BRNN) [4, 5, 51], using as input the peptide sequence, predicted secondary structure, Vina docking score and Pepsite score [46], to predict the peptide binding region within a protein sequence.

Generally, the most important prerequisite for a successful docking program is its ability to reproduce the co-crystallised ligand conformation, that is, a binding pose with a very low RMSD between it and the native ligand poses. But in the case of peptide docking, one has to be pragmatic in terms of the RMSD values which can be achieved due to the number of active rotatable bonds in unconstrained peptides. As we show in Figure 1 there is a linear relationship between the length of the peptide and the number of active rotatable bonds, which increase from a minimum of 9 for a peptide of length 5 residues to a maximum of 132 for a peptide of length 33 residues, in our datasest.

We used AutoDock Vina to dock peptides of length 5 - 35 residues in length extracted from 152 PDB protein-peptide complexes into their respective binding sites. We calculated the backbone RMSD value between the native and the top-ranked Vina pose. The goal here was not to asses if docking could find the exact peptide pose but to investigate what can be learned from docking known peptides to their protein receptor which would help to guide us towards predicting novel protein-peptide interactions, keeping in mind that the longer peptides have a very high number of active rotatable bonds.

Figure 2 shows the distribution of RMSDs between the docked peptide poses and their native co-crystallised peptide poses. The RMSDs of 28% of the peptides is less than 10 Å. Of these, 26% have an RMSD of less than 7 Å. Overall, 16% of the helical and beta, 23% of the helical, 31% of the beta and 37% of the disordered peptides have RSMDs of less than 10 Å. It would seem that the beta and disordered peptides have docked peptide poses closer to their native poses in the receptor site than the helical or helical and beta peptides, however the average length of the helical and helical and beta peptides is longer (16-mer and 15-mer, respectively) than the beta and disordered peptides (8-mer and 7-mer, respectively), and consequently the number of active rotatable bonds is also higher.

### Relationship between RMSD, peptide length, rotatable bonds and Vina score

It has been observed that the greater the number of rotatable bonds, the less likely is that the docked peptide pose will have a low RMSD between it and the native peptide pose [3]. 105 of the peptides in our dataset have more than 32 active rotatable bonds with an average of 55 active rotatable bonds per peptide. We can see from Figure 3(a) that there is a correlation between the number of rotatable bonds and the RMSD ($r = 0.45$), although this is perhaps not as strong as expected, and is somewhat dependent on the structure of the

peptide. Beta and disordered peptides have the strongest correlation ($r = 0.5$), with helical+beta peptides having the weakest ($r = 0.2$). A similar pattern is seen between RMSD and peptide length, with the RMSD between native and docked poses increasing as the peptide length increases (Figure 3(b)). We found that overall there is only a very weak correlation between Vina score and RMSD ($r = 0.18$, Figure 3(c)). The helical peptides have the strongest relationship ($r = 0.27$), but there is no correlation for the other structural classes ($r = 0.1$, $r = -0.06$ and $r = -0.002$ for beta, helical and beta, and disordered respectively). However, if we only look at peptide which are 10 residues or less in length we find a much better correlation ($r = 0.32$) (Figure 3(d)). This would suggest that if we restrict ourselves to shorter peptides ($< 10$ residues) we may be able to use the Vina docking score to guide us in discriminating between binding and non-binding peptides.

## Docking of Overlapping Peptides

We evaluated if docking could be used to identify a peptide binding region in a protein sequence. For example, if we know that there is an interaction between two proteins where one protein has a known 3D structure, could we identify the interacting region of the unstructured protein sequence? For each of our 39 peptides (length 5 to 9 residues), we retrieved the full length UniProtKB [45] protein sequence (ranging in length from 97 to 3969 residues). We selected a window of 97 residues around each of the 39 peptides (see Methods for full details) and prepared sets of overlapping peptides of length 2 to 5 residues by sliding a window along these 39 reduced protein sequence. We performed four sets of docking for each peptide, one for each set of dipeptides, tripeptides, tetrapeptides and pentapeptides, to each of their respective receptors. The best Vina score for each peptide was normalised and the results displayed as Receiver Operating Characteristic (ROC) curves where the true positive rate (TPR) is plotted against the false positive rate (FPR) (Figure 4). The results show only very modest predictive power with AUC of 0.54, 0.52, 0.53 and 0.52 respectively for the pentapeptide, tetrapeptide, tripeptide and dipeptide sliding windows (Figure 4). We compared these results to scores obtained by submitting the same sets of overlapping peptides to the PepSite2 [46] server along with their respective PDB structures. PepSite predicts where on a protein surface a particular peptide is likely to bind. We reasoned that if PepSite could identify the native peptide from other surrounding peptides it would achieve a lower p-value. PepSite is at somewhat of a disadvantage to Vina as the whole protein surface is searched for a likely binding site, as opposed to Vina where a search space is defined in advance.

Our results show that PepSite2, with AUC of 0.55, 0.55, 0.55 and 0.56 respectively for the pentapeptide, tetrapeptide, tripeptide and dipeptide sets, has a similar power to discriminate between binding and non-binding peptides as Vina (Figure 5). We also submitted the full protein sequences to ANCHOR [12] and SLiMPred [26] and compared predictions per residue for the 97 residue sections (Figure 6). ANCHOR, trained specifically to identify peptide binding

regions in disordered sequences struggles with this dataset. SLiMPred performs better as structured binding regions were included during training of this method. However, the performance is still quite modest (AUC 0.63).

Given the poor performance of any individual method we trained a Bidirectional Recurrent Neural Network in 10-fold cross-validation using a similar approach to that used to train SLiMPred [26]. We trained four predictors (one for each of the pentapeptide, tetrapeptide, tripeptide and dipeptide sets) using the peptide sequence, predicted secondary structure (by Porter [39, 6, 38, 25]), Vina score and PepSite2 predicted probability as input. The results were very promising (AUC 0.69 for the tripeptide set) showing that this information can be combined to produce results which are far superior to any single method (Figure 7). At present the dataset available (39 sequences) is too small to allow a full evaluation of this method, however the number of peptide-protein complexes in the PDB continues to grow and as more structures become available we will be able to further develop and evaluate this approach.

# Conclusions

Disordered regions of proteins often bind to structured domains, mediating interactions within and between proteins. We have presented a computational analysis of the performance of peptide docking with AutoDock Vina to asses if Vina could be used to predict protein-peptide interactions. As Vina is designed for small-molecule docking with a restriction on the number of rotatable bonds ($\leq 32$), it is generally assumed that it is not suitable for docking peptides which have many more internal degrees of freedom. Previously, however, we have shown that there is a correlation between the Vina docking scores of dipeptides with ACE and experimentally determined dipeptide ACE inhibition (IC50) [32].

First we investigated docking of known peptides to peptide binding domains. While we did not successfully obtain the native poses of the peptide, Vina scoring can be used to distinguish between interacting and non-interacting peptide ligands and their targets under certain conditions, however, we have shown that normalisation of the Vina score did not offer a significant improvement in the capacity to rank predicted models in this case. Then, we investigated the binding of adjacent overlapping peptides from the peptide's protein sequence. We found only weak discrimination of docking in this context with similar results for PepSite, ANCHOR, and somewhat better results for SLiMPred. We then trained a BRNN using the peptide sequence, predicted secondary structure, Vina score and PepSite probability as inputs. Our analysis shows that although individually Vina and PepSite are unable to identify peptide binding regions within a protein sequence, when used in combination with secondary structure predictions this information may successfully help to predict these binding regions with and AUC of 0.69 if a peptide size of three residues is chosen. In general the ROC curves suggest that this approach is most useful for limiting the search space down by approximately a half, but not particularly good at

pinpointing the exact regions of interest, however it does emphasise the importance of considering information from a number of different sources when trying to predict peptide binding regions.

In this study we have evaluated only one docking method, AutoDock Vina. The Vina docking program has many advantages, it is easy to install and run locally, it is extremely fast and is very suitable for high-throughput docking which is essentially for a study of this size. It is possible, however, that other methods, for example, DynaDock [2], which has been developed more specifically for the docking of peptides into flexible binding sites, may provide better results. This method is slower than Vina and not suitable for high-throughput docking. Another method, the Rosetta FlexPepDock server [24], refines docking conformations given a PDB file and an estimated peptide conformation, however, FlexPepDock is computationally intensive as the protocol samples a significant conformational space and therefore is also not suitable for high-throughput screening. Other target-specific protein-peptide protocols (MHC-peptide interactions [8], PDZ-peptide interactions [31]) are not available as web servers, or for download, at this time.

# Methods

## Dataset Preparation

1,431 protein-peptide complexes were retrieved from the PepX structural database [49]. We removed MHC complexes and immunoglobulins as these proteins make contact only with the antigenic peptides/fragments and not with the motif region embedded within the protein during interaction. Any complex structures that had peptide inhibitors (designed synthetic peptides, peptide analogues/mimetics) or tagged (streptavidin) peptides with modified amino acid residues were also excluded. From the remaining 886 structures, we selected structures whose resolution was 3.0 Å or better, each structure had at least two protein chains, where one of the chains (peptide) was 5-35 amino acid residues long. Clustering of the dataset at varying sequence identity levels was performed to reduce the size of the dataset in such a way that multiple structures whose sequences have at least the specified level of sequence identity were represented by a single structure. Clustering at pair-wise sequence identity below 30% was selected for the final dataset which contains 152 unique peptide sequences along with 152 receptors structures. 22 of the receptor structures have two or more sequence chains that form the fully functional protein-peptide complexes. To limit ourselves to the available structural data, we used the lengths of peptides for which atomic coordinate records are available along with the structural information.

For the reduced dataset of peptides less than 10 residues in length, protein-peptide complexes were retained only if the UniProtKB [45] identifier of the full length protein sequence from which the peptide originated was available, leaving 39 structures. The full length protein sequences were shortened to 97

residues (the length of the shortest protein in the dataset) with the peptide in the centre, unless the peptide was within 45 residues of the start or end of the sequence in which case the shortened sequence is 97 residues from the start or 97 residues from the end of the full length sequence, respectively. Finally, all possible peptides were generated by scanning a sliding windows of either 2, 3, 4 or 5 residues along these subsections of the full length protein sequence from which the peptide is derived.

## Structural Annotation of Dataset

We gathered the structural annotation for peptides from the PDB which uses DSSP [22]. We classify the peptides into four broad structural classes: helical (the peptide has at least one residue labelled as helix by DSSP), beta (the peptide has at least one residue labelled as beta sheet /strand by DSSP), helical and beta (the peptide has at least one residue labelled as helix and at least one labelled as beta sheet /strand by DSSP), disordered (the peptide has no residues labelled as helix or beta sheet /strand, but may have turns or bends, or peptides with no secondary structure assigned by DSSP).

## Docking of Peptides to Proteins

Peptides were converted into the SMILES format using CycloPs [13], and from SMILES to PDB format using Open Babel [33]. AutoDock 4.2 [27] was used to prepare peptide and receptor PDB input files and AutoDock Vina [47] (Version1.1.2) was used to dock the peptides to protein receptors.

## Analysis of Docking Results

The output generated by Vina for each peptide was processed and the pose (binding conformation) having the lowest binding affinity (Vina docking score) was selected for further investigation. RMSD values were calculated against their native co-crystallised peptide structures. To normalised the Vina docking scores and correct for multiple receptors, we calculated the MASC scores in combination with MRE as described by Vigers and Rizzi [50].

To assess the ability of Vina to discriminate between the native peptide and other peptides extracted from the same protein sequence, we measure the TPR and FPR. First, we normalise the Vina docking scores for each peptide so they fall between 0 and 1:

$$Normalised(v_i) = \frac{v_i - V_{min}}{V_{max} - V_{min}} \tag{1}$$

where $V_{max}$ is the absolute value of the minimum Vina binding score (i.e. $-10$) and $V_{min}$ is zero. We used R [41] to plot the linear regressions from which we derived the correlation, $r$ and the "verification" package in R to plot the TPR against the FPR as ROC curves and calculate the area under the curve (AUC). The AUC is a number between 0 and 1 inclusive, where 0.5 indicates

a random model and 1 is perfect, which is equivalent to the probability that a randomly chosen positive instance will rank higher than a randomly chosen negative instance [15].

## Acknowledgements

The authors acknowledge the Research IT Service at University College Dublin and the SFI/HEA Irish Centre for High-End Computing (ICHEC) for providing high performance computing (HPC) resources that have contributed to the research results reported within this paper. The authors thank Kevin Rue for technical assistance.

Funding: This work was supported by Science Foundation Ireland principal investigator grant [grant number 08/IN.1/B1864] to D. C. Shields. W. Khan gratefully acknowledges the award of a European Commission (EC) Erasmus Mundus Europe Asia (EMEA) Split-Doctoral Scholarship Scheme in University College Dublin (UCD), Ireland.

## References

[1] P. Aloy and R.B. Russell. Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Bio*, 7(3):188–197, 2006.

[2] I. Antes. DynaDock: A new molecular dynamics-based algorithm for protein–peptide docking including receptor flexibility. *Proteins*, 78(5):1084–1104, 2010.

[3] J. Audie and J. Swanson. Recent work in the development and application of protein–peptide docking. *Future*, 4(12):1619–1644, 2012.

[4] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15:937–946, 1999.

[5] P. Baldi and G. Pollastri. The principled design of large-scale recursive neural network architectures–dag-rnns and the protein structure prediction problem. *The Journal of Machine Learning Research*, 4:575–602, 2003.

[6] D Baù, AJM Martin, A Mooney, C Vullo, I Walsh, and G Pollastri. Distill: A suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins. *BMC Bioinformatics*, 7:402, 2006.

[7] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, TN Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–242, 2000.

[8] A.J. Bordner and R. Abagyan. Ab initio prediction of peptide-MHC binding geometry for diverse class I MHC allotypes. *Proteins*, 63(3):512–526, 2006.

[9] A. Castro, C. Bernis, S. Vigneron, J.C. Labbé, and T. Lorca. The anaphase-promoting complex: a key factor in the regulation of cell cycle. *Oncogene*, 24(3):314–325, 2005.

[10] L.L. Conte, C. Chothia, and J. Janin. The atomic structure of protein-protein recognition sites. *J Mol Biol*, 285(5):2177–2198, 1999.

[11] F. Diella, N. Haslam, C. Chica, A. Budd, S. Michael, N.P. Brown, G. Travé, and T.J. Gibson. Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front Biosci*, 13:6580–6603, 2008.

[12] Z. Dosztányi, B. Mészáros, and I. Simon. Anchor: web server for predicting protein binding regions in disordered proteins. *Bioinformatics*, 25(20):2745–2746, 2009.

[13] Fergal J Duffy, Mélanie Verniere, Marc Devocelle, Elise Bernard, Denis C Shields, and Anthony J Chubb. Cyclops: generating virtual libraries of cyclized and constrained peptides including nonnatural amino acids. *Journal of chemical information and modeling*, 51(4):829–836, 2011.

[14] A.K. Dunker, M.S. Cortese, P. Romero, L.M. Iakoucheva, and V.N. Uversky. Flexible nets. *FEBS J*, 272(20):5129–5148, 2005.

[15] T. Fawcett. An introduction to ROC analysis. *Pattern Recogn Lett*, 27(8):861–874, 2006.

[16] S.Y. Fuchs, V.S. Spiegelman, and K.G.S. Kumar. The many faces of $\beta$-TrCP E3 ubiquitin ligases: Reflections in the magic mirror of cancer. *Oncogene*, 23(11):2028–2036, 2004.

[17] M. Fuxreiter, P. Tompa, and I. Simon. Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, 23(8):950–956, 2007.

[18] J.J. Gray. High-resolution protein–protein docking. *Curr Opin Struc Biol*, 16(2):183–193, 2006.

[19] C. Haynes, C.J. Oldfield, F. Ji, N. Klitgord, M.E. Cusick, P. Radivojac, V.N. Uversky, M. Vidal, and L.M. Iakoucheva. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol*, 2(8):e100, 2006.

[20] Z. Hayouka, J. Rosenbluh, A. Levin, S. Loya, M. Lebendiker, D. Veprintsev, M. Kotler, A. Hizi, A. Loyter, and A. Friedler. Inhibiting HIV-1 integrase by shifting its oligomerization equilibrium. *P Natl Acad Sci USA*, 104(20):8316–8321, 2007.

[21] S. Jones and J.M. Thornton. Principles of protein-protein interactions. *P Natl Acad Sci USA*, 93(1):13–20, 1996.

[22] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[23] H.Y. Kim, B.Y. Ahn, and Y. Cho. Structural basis for the inactivation of retinoblastoma tumor suppressor by SV40 large T antigen. *EMBO J*, 20(1):295–304, 2001.

[24] N. London, B. Raveh, E. Cohen, G. Fathi, and O. Schueler-Furman. Rosetta FlexPepDock web serverhigh resolution modeling of peptide–protein interactions. *Nucleic Acids Res*, 39(suppl 2):W249–W253, 2011.

[25] C. Mooney and G. Pollastri. Beyond the twilight zone: Automated prediction of structural properties of proteins by recursive neural networks and remote homology information. *Proteins*, 77(1):181–90, 2009.

[26] C. Mooney, G. Pollastri, D.C. Shields, and N.J. Haslam. Prediction of short linear protein binding regions. *J Mol Biol*, 415:193–204, 2011.

[27] G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, and A.J. Olson. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem*, 30(16):2785–2791, 2009.

[28] R. Mosca, C. Pons, J. Fernández-Recio, and P. Aloy. Pushing structural information into the yeast interactome by high-throughput protein docking experiments. *PLoS Comput Biol*, 5(8):e1000490, 2009.

[29] V. Neduva and R.B. Russell. Linear motifs: evolutionary interaction switches. *FEBS lett*, 579(15):3342, 2005.

[30] V. Neduva, R.B. Russell, et al. Peptides mediating interaction networks: new leads at last. *Curr Opin Biotech*, 17(5):465, 2006.

[31] M.Y. Niv and H. Weinstein. A flexible docking procedure for the exploration of peptide binding selectivity to known structures and homology models of PDZ domains. *J Am Chem Soc*, 127(40):14072–14079, 2005.

[32] R. Norris, F. Casey, R.J. FitzGerald, D. Shields, and C. Mooney. Predictive modelling of angiotensin converting enzyme inhibitory dipeptides. *Food Chemistry*, 133(4):1349–1354, 2012.

[33] N.M. OBoyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, and G.R. Hutchison. Open Babel: An open chemical toolbox. *J Cheminform*, 3(1):1–14, 2011.

[34] L. Parthasarathi, F. Casey, A. Stein, P. Aloy, and D.C. Shields. Approved drug mimics of short peptide ligands from protein interaction motifs. *J Chem Inf Model*, 48(10):1943–1948, 2008.

[35] T. Pawson. Dynamic control of signaling by modular adaptor proteins. *Curr Opin Cell Biol*, 19(2):112–116, 2007.

[36] T. Pawson and P. Nash. Assembly of cell regulatory systems through protein interaction domains. *Science Signalling*, 300(5618):445, 2003.

[37] E. Petsalaki and R.B. Russell. Peptide-mediated interactions in biological systems: new discoveries and applications. *Curr Opin Biotech*, 19(4):344–350, 2008.

[38] G. Pollastri, A.J.M. Martin, C. Mooney, and A. Vullo. Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC bioinformatics*, 8(1):201, 2007.

[39] G Pollastri and A McLysaght. Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, 21(8):1719–1720, 2005.

[40] P. Puntervoll, R. Linding, C. Gemünd, S. Chabanis-Davidson, M. Mattingsdal, S. Cameron, D.M.A. Martin, G. Ausiello, B. Brannetti, A. Costantini, et al. ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res*, 31(13):3625–3630, 2003.

[41] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[42] R.B. Russell, F. Alber, P. Aloy, F.P. Davis, D. Korkin, M. Pichaud, M. Topf, and A. Sali. A structural perspective on protein–protein interactions. *Curr Opin Struct Biol*, 14(3):313–324, 2004.

[43] R.B. Russell and T.J. Gibson. A careful disorderliness in the proteome: sites for interaction and targets for future therapies. *FEBS lett*, 582(8):1271–1275, 2008.

[44] A. Stein, R.A. Pache, P. Bernadó, M. Pons, and P. Aloy. Dynamic interactions of proteins in complex networks: a more structured view. *FEBS J*, 276(19):5390–5405, 2009.

[45] The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res*, 40(D1):D71–D75, 2012.

[46] L.G. Trabuco, S. Lise, E. Petsalaki, and R.B. Russell. PepSite: prediction of peptide-binding sites from protein surfaces. *Nucleic Acids Res*, 40(W1):W423–W427, 2012.

[47] O. Trott and A.J. Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*, 31(2):455–461, 2010.

[48] V. Vacic, M.S. Cortese, A.K. Dunker, and V.N. Uversky. Analysis of molecular recognition features (MoRFs). *J Mol Biol*, 362:1043–1059, 2006.

[49] P. Vanhee, J. Reumers, F. Stricher, L. Baeten, L. Serrano, J. Schymkowitz, and F. Rousseau. PepX: a structural database of non-redundant protein–peptide complexes. *Nucleic Acids Res*, 38:D545–D551, 2010.

[50] G.P.A. Vigers and J.P. Rizzi. Multiple active site corrections for docking and virtual screening. *J Med Chem*, 47(1):80–89, 2004.

[51] I. Walsh, A. Vullo, and G. Pollastri. Recursive neural networks for undirected graphs for learning molecular endpoints. *Pattern Recognition in Bioinformatics*, pages 391–403, 2009.

[52] M.N. Wass, G. Fuentes, C. Pons, F. Pazos, and A. Valencia. Towards the prediction of protein interaction partners using physical docking. *Mol Syst Biol*, 7:469, 2011.

[53] L. Zhao and J. Chmielewski. Inhibiting protein–protein interactions using designed molecules. *Curr Opin Struct Biol*, 15(1):31–34, 2005.

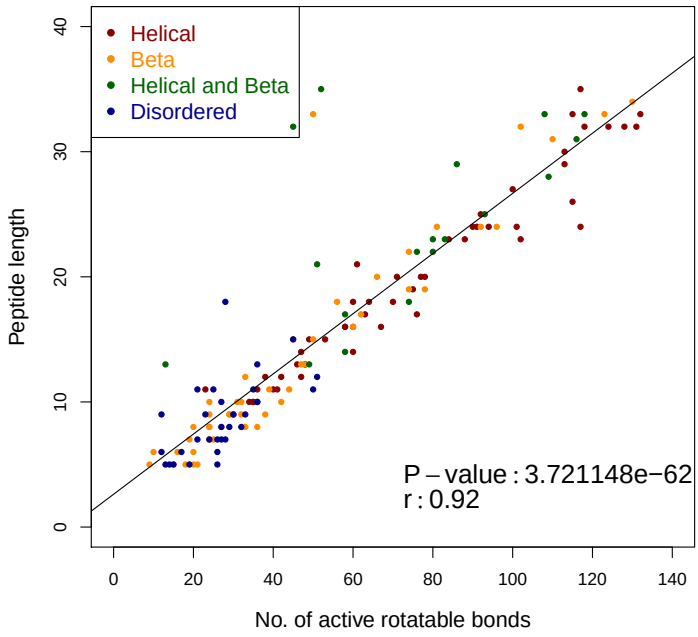



Figure 1: **Active rotatable bonds and peptide length.** Plot showing the relationship between the number of active rotatable bonds and peptide length for the 152 peptide sequences of length 5-35 residues long.

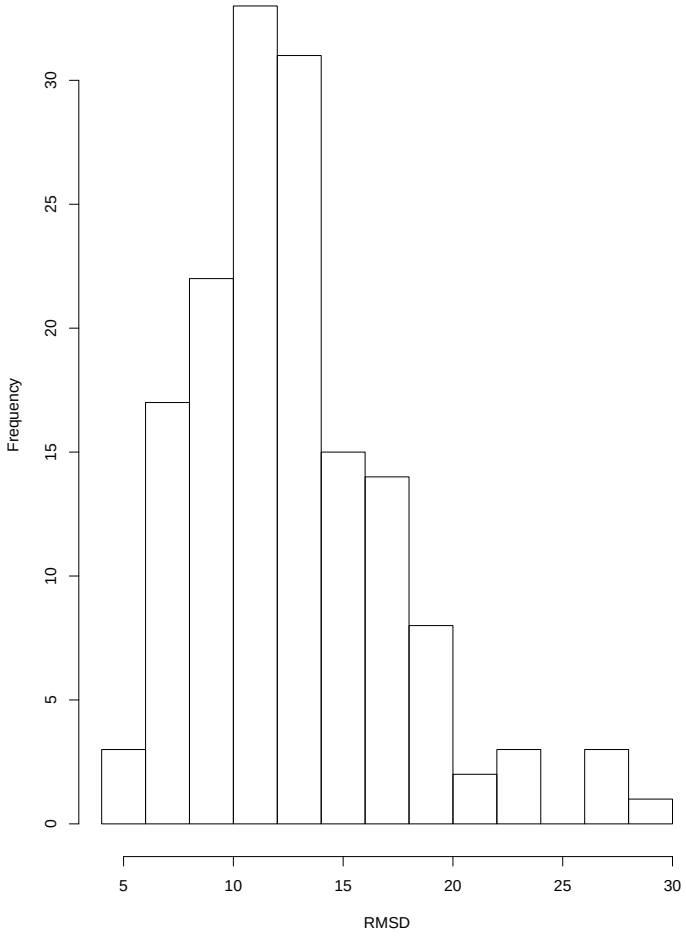



Figure 2: **RMSD distributions.** Histogram showing the distribution of RMSDs between the docked peptide poses and their native co-crystallised peptide poses.

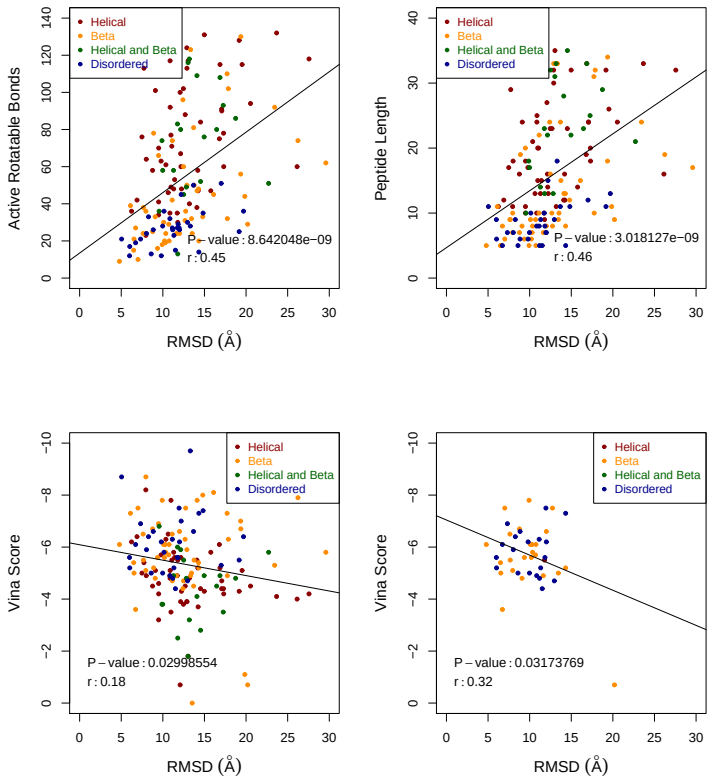
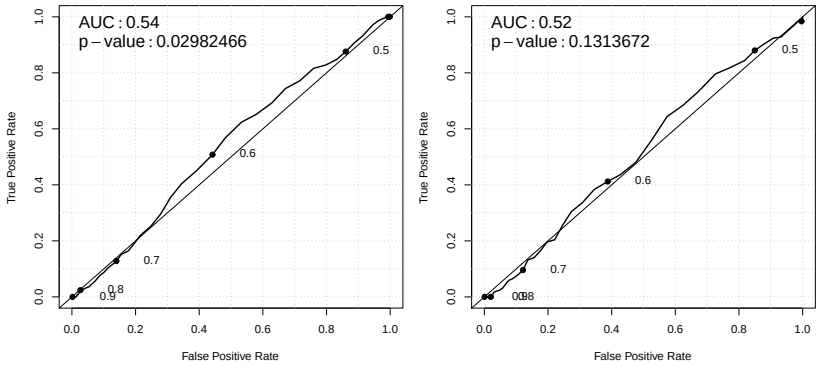


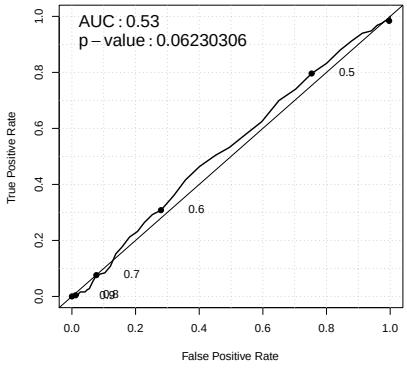

Figure 3: **Plots showing the relationships between the RMSD, peptide length, active rotatable bonds and Vina score.** (a) active rotatable bonds and RMSD (b) peptide length and RMSD (c) Vina score and RMSD (d) Vina score and RMSD (peptides $< 10$ residues in length only).



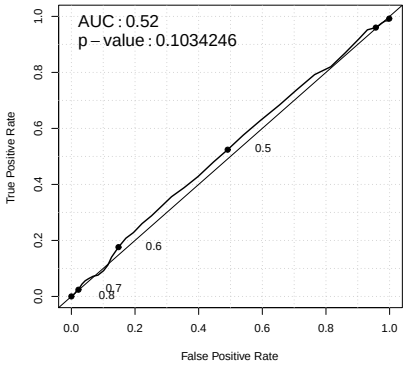

(a) Pentapeptide




(b) Tetrapeptides

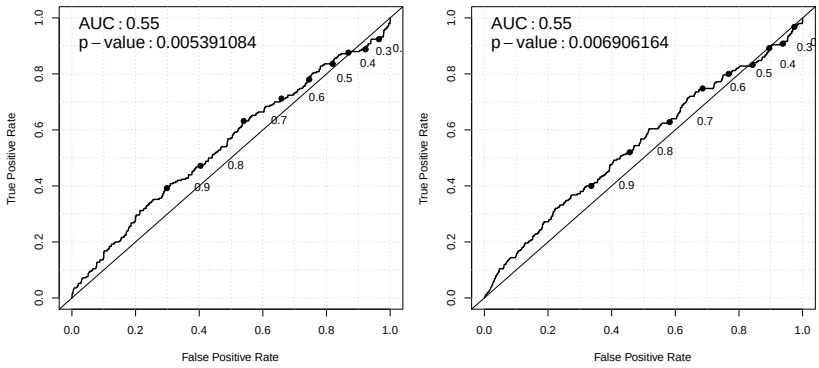



(c) Tripeptides




(d) Dipeptides

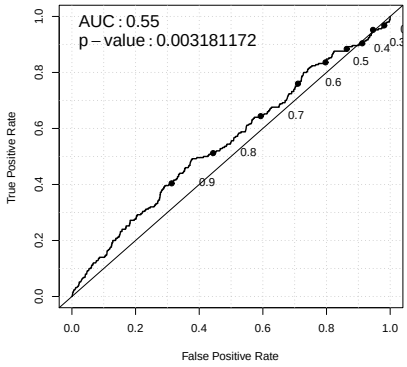
Figure 4: **Vina docking scores – ROC curves for overlapping peptide lengths** (a) Pentapeptide (b) Tetrapeptides (c) Tripeptides (d) Dipeptides.



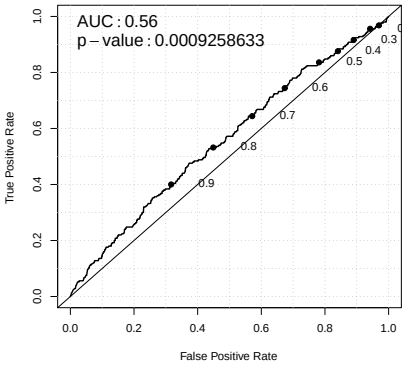

(a) Pentapeptide




(b) Tetrapeptides

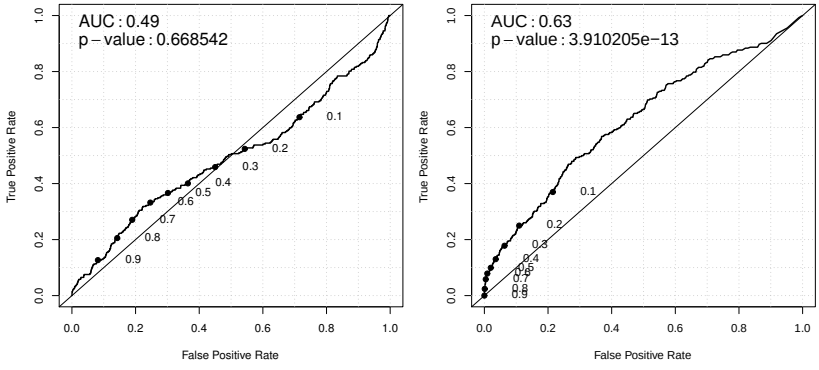



(c) Tripeptides



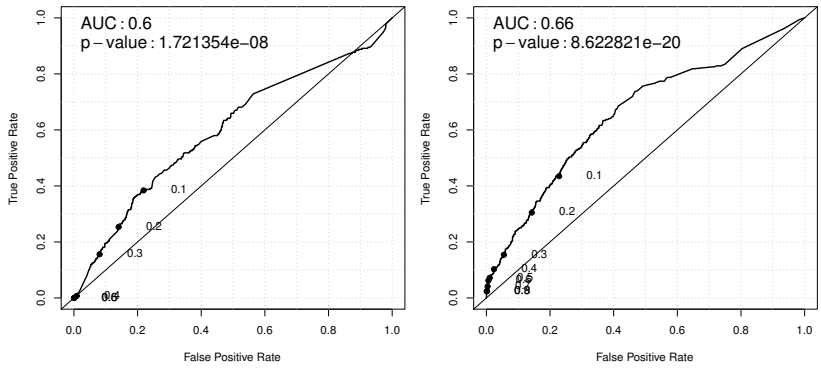

(d) Dipeptides

Figure 5: **PepSite2 – ROC curves for overlapping peptide lengths** The discriminating threshold is 1 - p-value predicted by PepSite2. (a) Pentapeptide (b) Tetrapeptides (c) Tripeptides (d) Dipeptides.




(a) ANCHOR (b) SLiMPred

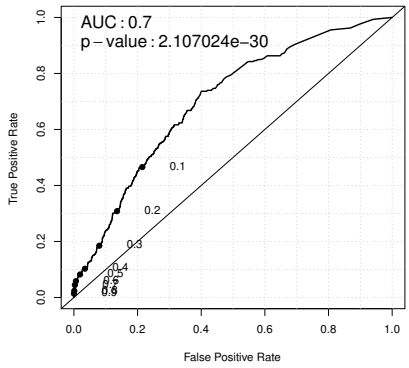
Figure 6: **ROC curves** (a) ANCHOR (b) SLiMPred.



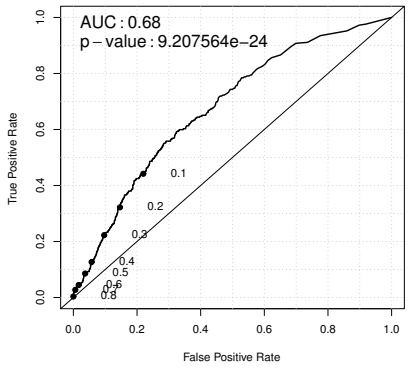

(a) Pentapeptide




(b) Tetrapeptides




(c) Tripeptides




(d) Dipeptides

Figure 7: **PepBindPred – ROC curves BRNN trained on four different overlapping peptide lengths** (a) Pentapeptide (b) Tetrapeptides (c) Tripeptides (d) Dipeptides.