**Technische Universität München**

**Max-Planck-Institut für Biochemie**

# Structural analysis of 20S Proteasome and Development of Structure-Based Virtual Screening Methods

Marcelino Arciniega Castro

Vollständiger Abdruck der von der Fakultät für Chemie der Technischen Universität München zur Erlangung des akademischen Grades eines

**Doktors der Naturwissenschaften (Dr. rer. nat.)**

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. M. Groll

Prüfer der Dissertation: 1. apl. Prof. Dr. Dr. h.c. R. Huber (i. R.)

2. Univ.-Prof. Dr. I. Antes

Die Dissertation wurde am 16.06.2014 bei der Technischen Universität München eingereicht und durch die Fakultät für Chemie am 23.07.2014 angenommen.

# Acknowledgments

I would like to express my gratitude to Prof. Robert Huber for giving me the great opportunity of working under his supervision.

I also would like to thank all my other bosses, Dr. Michael Groll, Dr. Gerhard Müller, Dr. Tad A. Holak, Dr. Grzegorz Popowicz and especially to Dr. Oliver Lange. Thanks for providing me with working conditions, for shearing expertise and knowledge, for cheering me up and telling me off.

I am deeply grateful to the institutions that provided financial support during my PhD studies: CONACYT, DAAD and "*Peter und Traudl Engelhorn Stiftung*".

My thanks to Monika Schneider for helping in many ways.

I want to thank to Dra. Alicia Ortega for kicking me into the plane to Munich.

I also want to say "thank you very much guys!!!!" to all the people that I referred to as "my friends" for making my PhD student time bearable. Thank you very much guys for the all those: coffees, beers, dinners, advices, jokes, trips, cheers, and beautiful experiences that brought joy to my being.

Special thanks to my life companion Anna Rojowska.

I would like to thank to my parents, brothers, aunts, uncles and cousins, for the support, love and good wishes that they gave me over all these years. You are always in my mind (specially my aunt Rosa Maria).

Finally and most importantly, I want to express my most profound gratitude to my mother, without her none of all the wonderful things that I have experienced in my life would not have happened.

# List of publications

M. Arciniega, P. Beck, O.F. Lange, M. Groll and R. Huber. Differential Global Structural Changes in the Core Particle of Yeast and Mouse Proteasome Induced by Ligand Binding. *PNAS* (**2014**) 111:9479-9484. [http://www.pnas.org/content/111/26/9479].

M. Arciniega and O. F. Lange. Improvement of Virtual Screening Results by Docking Data Feature Analysis. *J Chem Inf Model.* (**2014**) 54:1401-11. [http://dx.doi.org/10.1021/ci500028u].

M. Bista, S. Wolf, K. Khoury, K. Kowalska, Y. Huang, E. Wrona, M. Arciniega, G. M. Popowicz, T. A. Holak and A. Dömling. Transient Protein States in Designing Inhibitors of the MDM2-p53 Interaction. *Structure* (**2013**) 21:2143-51.

S. Baek, N.J. Kang, G.M. Popowicz, M. Arciniega, S.K. Jung, S. Byun, N.R. Song, Y.S. Heo, B.Y. Kim, H.J. Lee, T.A. Holak, M. Augustin, A. M. Bode, R. Huber, Z. Dong and K.W. Lee. Structural and Functional Analysis of the Natural JNK1 Inhibitor Quercetagetin. *J Mol Biol.* (**2013**) 425:411-23.

N. Gallastegui, P. Beck, M. Arciniega, R. Huber, S. Hillebrand and M. Groll. Hydroxyureas as Noncovalent Proteasome Inhibitors. *Angew Chem Int Ed Engl* (**2012**) 51:247-9.

# Table of Contents

1. Abstract ..... 1
2. Introduction ..... 3
2.1 Virtual Screening ..... 3
2.1.1 Docking ..... 3
2.1.2 Virtual Screening workflow ..... 5
2.2 20S proteasome ..... 7
2.2.1 Biological Role and Structure ..... 7
2.2.2 Cleavage mechanism and its inhibition ..... 8
2.3 Initial virtual screening on the 20S ..... 13
2.3.1 Reliability score and docking map ..... 13
2.3.2 Initial test ..... 14
2.3.3 Conclusions from the initial test ..... 17
3. Methods ..... 18
3.1 AutoDock4.2 ..... 18
3.2 AutoDock Vina ..... 18
3.3 Rosetta Ligand ..... 19
3.4 Molecular Dynamic Simulations (Gromacs) ..... 20
3.5 Principal Component Analysis ..... 21
3.6 Artificial Neural Networks ..... 21
4. Description of the first publication ..... 25
5. Description of the second publication ..... 26
6. References ..... 27

# 1. Abstract

This work presents results that constitute the required basis for establishing a successful structure-based Virtual Screening (VS) towards the identification of 20S proteasome inhibitors. The construction of this basis is divided into two aspects: i) structural and functional understanding of 20S proteasome and ii) a reliable method for analyzing VS results.

This document consists of an introductory part where first is presented a brief review of the most important aspects of VS methodologies followed by a description of the biological relevance of 20S proteasome as well as the current state proteasome inhibitors design. Subsequently, the preliminary VS results obtained which led to deeper understanding of the problem that forms the cardinal part of this dissertation are outlined. Thereafter the methodology is described depicting the analytical tools used in the development of the research. Finally, included in this work are two peer-reviewed publications that demonstrate the contribution to the development of protocols for the successful design of 20S proteasome inhibitors.

In the first publication a Principal Component Analysis (PCA) on X-ray crystal structures of 20S proteasome's β5 subunit unveils a domain movement induced upon peptidic inhibitor binding, thereby making possible to classify structures into pep- and apo-clusters. Similar displacements are observed in the mouse constitutive β5 subunit, but not in the mouse immuno β5 subunit, where the apo and liganded states are similar to the yeast structures in the pep-cluster. Further structural analysis, constituted by Molecular Dynamic simulations and a new X-ray crystal structure, confirmed the relevance that the peptide binding has on the induction of the observed domain motion. By assessing the implication that this structural transition has on the rest of the 20S proteasome two possible allosteric pathways, which had been predicted previously in the literature, were found. These findings not only identify a completely new landscape of possible drugable pockets for an allosteric inhibition, but also provided a possible explanation for the enhanced peptidolytic activity observed in the immunoproteasome with respect to its constitutive counterpart.

In the second publication an automated method for analyzing structure-based VS results was developed. The need for the development of this algorithm is based on the two aspects: i) current docking methodologies are able to identify the right binding mode for a single ligand

based on the docking score, but the same docking score is a rather inaccurate measure when comparing binding modes of different ligands, and ii) an important part of the VS protocol is still the human expertise in the evaluation of the docking binding modes aiming to identify false positives. To address these aspects, an Artificial Neural Network (ANN) is trained, validated and tested on VS benchmark. Each ligand has assigned features derived from the analysis of the docking data, which are the input on which the ANN evaluates the ligand chances of being an active inhibitor. The features are important aspects based on which an expert user would take the decision whether to trust or not the suggested docked molecule. The developed algorithm not only provides results comparable with the best of the methodologies found in the literature, but also proves to be robust and efficient method in VS.

# 2. Introduction

## 2.1 Virtual Screening

In the early stages of drug discovery process, a huge experimental effort is required to identify molecules capable to inhibit the activity of the target protein (Khanna 2012). Complications arise mainly from the vast amount of molecules that can compose the screening library (here also referred as chemical library). To reduce the experimental cost and to accelerate the process of identifying molecules with inhibitory capabilities (active molecules), a set of computational techniques, known as Virtual Screening (VS), have been devised (Tanrikulu, Krüger et al. 2013).

VS techniques can be divided in ligand-based and structure-based (Scior, Bender et al. 2012). In the ligand-based approach, the search on the chemical library is performed to identify molecules that share similar characteristics to that of previously known active molecules, that is, actual binders or ligands. In the structure-based approach atomic models of both, the protein´s active site and screening molecule, are used to evaluate their binding affinity. In this case, estimating the binding affinity of each molecule in the chemical library conforms the screening. This structure-based approach is preferred when novel lead compounds are a prerequisite (Drwal and Griffith 2013). In following section is provided a brief description on the main aspects of the structured-based approach (from now on referred to as docking).

### 2.1.1 Docking

The objective of the docking procedure is to accurately predict the binding affinity of a molecule (from now on also referred as ligand) in the protein's active site by evaluating the atomic interactions that conforms the binding complex(Xuan-Yu Meng 2011). This evaluation is dependent on the relative orientation of the ligand's atoms with respect to those forming the protein active site. In a docking experiment every single conformation of the binding complex is referred to as a *pose*, and each evaluated pose has an associated scoring value. For an accurate evaluation of binding energy a synergistic cooperation between two elements is needed: i) a scoring function assessing the quality of the interactions, and ii) a pose search algorithm that optimize those interactions. Thus, the scoring function is responsible of discern

favorable from unfavorable conformations, while the pose search algorithm samples the conformational space to optimize the evaluation of the scoring function.

There are three types of scoring functions: i) force-field-based, ii) empirical and iii) knowledge-based (Cheng, Li et al. 2012). Force-field-based scoring functions evaluated the binding by computing all the non-bonded interactions among the atoms of the binding complex. The van der Waals interactions are represented by Leonard-Jones potentials, while the electrostatic contributions are computed from a Coulombic perspective, using a distance dependent dielectric function. The empirical scoring functions assess the binding affinity by a weighted sum of energy terms such as hydrogen bond, ionic interaction, hydrophobic effect and binding entropy. The weights scale the contribution that each term provides to the final score and are calibrated to reproduce the experimental binding affinities from a set of known complexes. The knowledge-based scoring functions use distance dependent potentials between each pair of atoms of the protein-ligand complex. These potentials are calibrated based on the statistical analysis of X-ray crystal structures complexes. The rationale behind this approach is that the more frequent an interacting pair is found in crystal complexes, the more favored is its interaction.

The pose search algorithms are very diverse, thus here are briefly described three of the most commonly used (Xuan-Yu Meng 2011): i) incremental construction methods, ii) Monte Carlo methods and iii) Genetic methods. In the incremental construction methods, the ligand is firstly divided into fragments. Secondly, each of these fragments is positioned in to the active site in a sequential fashion. Fragments are usually defined by rotatable bonds and the largest fragment referred to as anchor fragment is docked first. After the anchor fragment is positioned, it becomes a fixed part of the active site and the next fragment is docked. This process is continued until all the fragments are being docked and the original molecule can be reconstructed. This methodology is used in DOCK 4.0 (Moustakas, Lang et al. 2006), FlexX (Gastreich, Lilienthal et al. 2006), and eHiTS (Zsoldos, Reid et al. 2007). In the Monte Carlo methods, the protein-ligand poses are generated by sampling the configuration space of the rigid-body translations and rotations, in addition to the sampling of rotatable bonds. The pose selection is made obeying an energy based criterion. If the current pose satisfies the criteria, the pose is first stored and then modified to generate the next conformation. Monte Carlo sampling methods are used ICM (Neves, Totrov et al. 2012) and RosettaLigand (Combs, DeLuca et al. 2013). In Genetic methods each pose is described by a set of genes, each of them encoding a particular state the ligand´s degrees of freedom. Several poses, conforming a “generation” or

"set of individuals", are evaluated and compared among each other. The individuals with better evaluations are then chosen to perform on them "gene mutations" that render the next generation. Genetic algorithms are used in AutoDock (Morris, Huey et al. 2009) and GOLD (Verdonk, Cole et al. 2003).

Regardless of the scoring function and pose search algorithm, a typical output from docking a single molecule in the active site is composed by a small set of the best scored poses. Although in principle these suggested poses are ranked by its scoring values, inaccuracies of the scoring functions make necessary the performance of a visual inspection of the suggested poses in order to identify false positives. This visual analysis is crucial in the success of a docking experiment (Klebe 2006; Cosconati, Forli et al. 2010; Broccatelli and Brown 2014).

### 2.1.2 Virtual Screening workflow

Several considerations are needed in a traditional structure-based VS work flow (Figure 1). Firstly, a decision needs to be made in regard of the active site model to be used: single structure or a set of structures, experimental data or modeling data, etc. The influence that the active site conformation has on the identification of active molecules is well documented (Sinko, Lindert et al. 2013). Secondly, the preparation of the screening library usually requires a pre-filtering of molecules that form the library (Lipinski, Lombardo et al. 2001). Removing molecules that have a low probability to result as actives, for example by looking at a specific physicochemical property, reduces considerably the computation time, eases the analysis and increases the chances of finding an active molecule. After the preparation of the active site and the chemical library, the VS consist in docking each molecule of the library into the active site. As described in the previous section, form each molecule $(M)$ a set of docking poses is generated

$$M(Pose_k, Score_k) \quad where \quad M \in Screening\ Library; \quad k \in number\ of\ docking\ poses$$

Moreover, if docking is performed on a set of protein structures, then of each molecule $(M)$

$$M(Pose_k, Score_k, Receptor_j) \quad where\ Receptor_j \in Set\ of\ active\ sites$$

From this perspective, the visual inspection of all possess is no longer feasible, even for a few hundreds of screened molecules, thereby the use of the docking scoring value as a main

ranking method imperative. Identifying the actual binding mode based on the docking score from a set of poses of the same ligand is now a days considered as a partially solved problem (Damm-Ganamet, Smith et al. 2013). However, far more challenging is the comparison of poses between different ligands. The difficulties emerge, for example, in that conformational entropies canceled out when comparing poses of the same ligand, but not when comparing poses of different ligands.

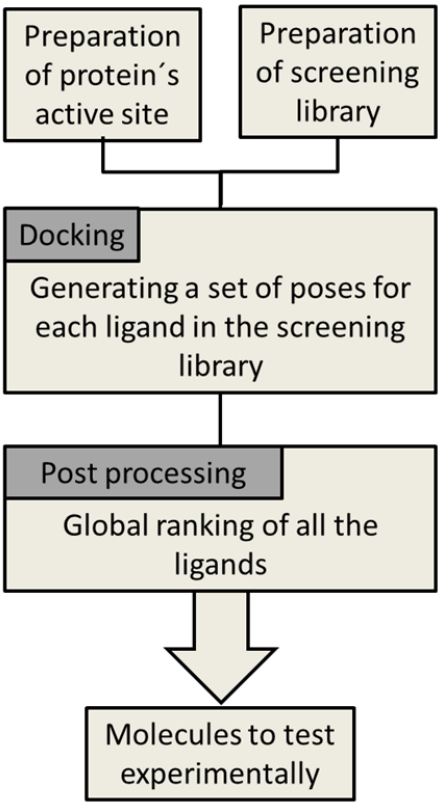



Figure 1. Virtual Screening Workflow in a structure-based approach.

Several methods and tools have been proposed to produce an automatic and reliable ranking of the docked library (Fukunishi 2010), nevertheless the problems is still far from being solved (Damm-Ganamet, Smith et al. 2013). The appropriate treatment of all these issues in the VS workflow is preponderant in identifying interesting hits that boost the drug discovery process.

## 2.2 20S proteasome

### 2.2.1 Biological Role and Structure

90% of all the non-lysosomal protein cleavage that takes place in the cell is performed through a protein degradation pathway, known as Ubiquitin Proteasome System (UPS) (Hershko and Ciechanover 1998). In order to maintain this process strictly regulated, target proteins for degradation are tagged with a polyubiquitin chain that is subsequently recognized by the heart of this degradation pathway, the 26S proteasome(Pickart 2004).

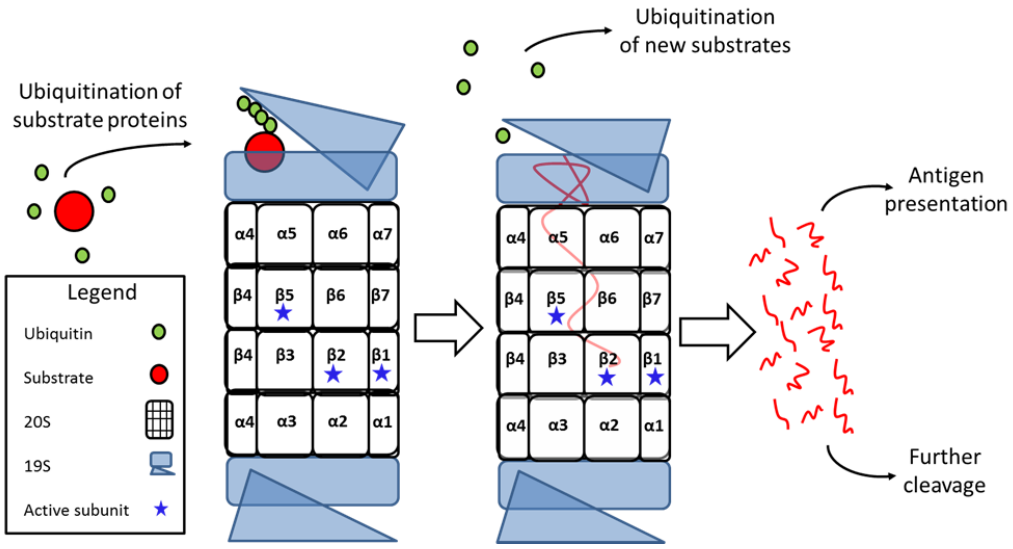



Figure 2. Protein degradation through the UPS. Proteins that need to be degraded are marked with a poli-ubiquitin chain. This attachment is an ATP dependent process that involves the action the enzymes E1-E3 (not shown in the scheme for clarity reasons). This ubiquitination is the signal recognized by the regulatory particle 19S to allow the protein to reach the inner part of the 20S, where the active subunits will cleave the protein into peptides. These peptides can be cleaved further by other peptidases or used as antigens in an immune response.

The 26S proteasome is composed by two major substructures: the 19S Regulatory Particle (RP) and 20S proteasome Core Particle (CP). The RP, has been shown to be situated in one or two ends of the barrel-shaped structure of the 20S CP, and has the function of recognizing and unfolding poly-ubiquitinated proteins and subsequently translocating them into the catalytic lumen this macromolecule, known as the 20S (Figure 2). The 20S proteasome has a barrel-like-shaped structure given by four stacked heptameric rings in α-β-β-α stoichiometry. In eukaryotes seven different α- and β-subunits form the ring structure, whereas in simpler organisms, such as

archeabacteria, just one type of α- and β- subunits conform the heptameric arrangement. In the case of eukaryotic proteasome the catalytic centers of the CP are located inside the β-subunit rings, specifically at subunits β1, β2 and β5. After the peptidolityc activity has occurred in these sites the peptides leave the 20S and are further cleavage by other peptidases. However, alternatively, the peptides may reach the major histocompatibility complex (MHC) class I. In this scenario the peptides form part of an adaptive immune response (occurring in vertebrates) (Figure 2). To facilitate the recognition by the MHC I, antigenic peptides need to have a hydrophobic residues at their C-terminal (Strehl, Textoris-Taube et al. 2008). To effectively produce peptides with the proper characteristics, the constitutive actives subunits, β1, β2 and β5, are substituted by immuno counterparts, β1i, β2i and β5i. Despite high degree of sequence similarity, biochemical and cellular assay have shown that the 20S form with the immuno subunits (immunoproteasome) has an increased peptidolytic activity with respect to constitutive proteasome (Sijts, Ruppert et al. 2000).

### 2.2.2 Cleavage mechanism and its inhibition

From the seven different β-subunits found in CP of eukaryotes, just three are proteolytically active; β1, β2 and β5. These active subunits differ from the rest in that they have eliminated a pro-peptide from their N-terminal structure, leaving exposed a Threonine, Thr1, which itself defines the core of the catalytic triad. Surrounding the Threonine are different specificity pockets (S) numbered in relation with their position from the hydrolyzable peptide bond, the corresponding side of the peptide substrate (P) are numbered consequently (Figure 3A). In 20S proteasome of eukaryotic, the sequence diversity of the subunits confers structural variation to the three active sites given rise to different substrate affinities. These specificities have been characterized as caspase-, trypsin- and chymotripsyn-like, for the subunits β1, β2 and β5, respectively (Nussbaum, Dick et al. 1998). Structural and mutational analyses have shown that residues forming the S1 pocket, and in particular the residue at the sequence position 45 of the β active subunit, has a preponderant role in the definition of these specificities (Groll, Bochtler et al. 2005). In the subunit β1 the Arg45 provides to the pocket an acidic nature, while the Gly45 in the β2 and Met45 in the β5 render basic and hydrophobic milieus, respectively. Nevertheless, all three active subunits share a common cleavage mechanism (Groll and Huber 2004). First, the Thr1O$\gamma$ is attached to the carbonyl carbon at the peptide bond of the substrate forming an acyl-ester intermediate (Figure 3B). In a second reaction step, Thr1N atom accomplish a proton acceptor role, thus promoting the cleavage of the intermediate (Figure 3C). During the catalysis, the presence of a water molecule in the proximity of the Thr1 accomplishing a nucleophilic

function result crucial, since it not only acts as proton shuttle between Thr1Oγ and ThrN, but also participates in the acyl-ester bond disruption of the intermediate.

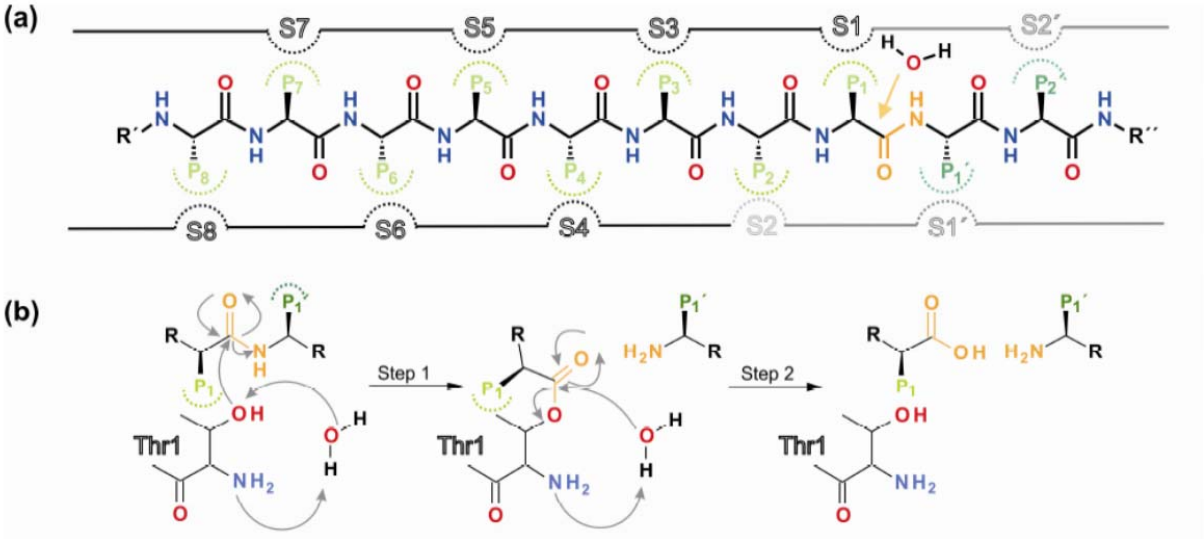



Figure 3. Cleavage mechanism of the CP. (a) Schematic representation of non-primed substrate binding pockets S1, S2, S3…Sn (colored black except for S2 (grey) due to its lack of presence in constitutive CPs) and primed S1', S2', S3'… Sn' sites, depending on their location to the peptide bond. Ligand side chains with the proteasomal specificity pockets, are referred to as P1, P2, P3…Pn and P1', P2', P3'…Pn', accordingly. (b) Cleavage mechanism by Thr1 in the active β subunits of the 20S proteasome ( Figure taken form (Gallastegui de la Rosa 2012) ).

Participation of the CP 20S proteasome in crucial processes of the cellular cycle is well acknowledged (Glickman and Ciechanover 2002). Similarly well recognized is its participation in the development of main human diseases such as Alzheimer´s disease, autoimmune diseases, muscle atrophy and different cancer, e.g. multiple myeloma (Kisselev and Goldberg 2001). Consequently copious efforts have been made to develop inhibitors of the 20S proteasome activity. Dozens of proteasome inhibitors are currently found in the literature (Beck, Dubiella et al. 2012; Kisselev, van der Linden et al. 2012), and can be classified under two schemes: i) peptidic or non-peptidic and ii) covalent or non-covalent inhibitors. In the first classification, the number of proteasome inhibitors having a peptide-like structure is clearly superior to the number of non-peptidic, whereas in the second classification the covalent proteasome inhibitors are more numerous than the non-covalent. This testifies that landscape of proteasome inhibitors in dominated by covalent peptidic molecules.

The 20S proteasome inhibition by covalent peptidic molecules has three components: i) the allocation of the peptide side chains into the specificity pockets, ii) the formation of an antiparallel β-sheet at the actives site, and iii) the covalent bond formation. The side chains of the peptidic inhibitor module the binding at the different catalytic subunits, with P1-site being the main contributor. Nonetheless, it is well documented that the residue at P3 has also strong influence modulating the peptide binding (Rydzewski, Burrill et al. 2006). The formation of the antiparallel β-sheet is a process occurring in all three active sites, since the peptidic inhibitor mimics the backbone hydrogens bonds that the peptide substrates originally undertake. Finally, the covalent bond formation is highly dependent on the inhibitor head group (the milieu of the different active sites also play a role). The head group defines properties as kinetics and specificities. Some examples of head groups inhibiting the 20S proteasome are Aldehydes, boronic acids and epoxyketones.

Peptidic aldehydes were the first discovered CP 20S proteasome synthetic inhibitors (Borissenko and Groll 2007). The X-ray crystal structures of yeast CP with complexed structures of Calpain inhibitor I and II (CAL I and CAL II) aided to unveil the cleavage as well as the inhibitory mechanism by the proteasome reacts, demonstrating the key role played by residue Thr1. The inhibition of CAL I is accomplished by the formation of a hemiacetal bond between the aldehyde group and the Thr1 (Kisselev and Goldberg 2001). The main disadvantage of the aldehyde inhibitors is their lack specificity proteasome as they also react with serine and cysteine proteases.

Peptidic boronic acids are probably the most successful proteasome inhibitors (Pellom ST Jr. 2012). They offer a set of advantages over their aldehyde counterparts: i) low specificity towards other proteases, ii) low dissociation rate from 20S proteasome and iii) higher inhibitory activity. A peptidic boronic acid molecule (Bortezomib) is since 2003 approved by the Food and Drug Administration (FDA) as drug against multiple myeloma. Nonetheless, peptide boronates still present unwanted side effects as a consequence of their slow reversibility (Ruschak, Slassi et al. 2011).

Peptidic epoxyketones, such as epoxomycin derived from an actinomycete strain, possess a high specificity towards the chymotrypsin-like active site, although inhibition effects on the other two active subunits are also observed (Meng, Mohan et al. 1999). Epoxyketons binds irreversibly to the Thr1 by forming a morpholino ring with the Thr1Oγ and ThrN. This peculiar moiety, form upon binding, renders epoxyketones highly specific towards the 20S proteasome,

since Thr1 is not present at the aminoterminal position in other proteases. Carfilzomib, an epoxomycin analogue, is since July 2012 an approved drug by the FDA to treat multiple myeloma (Steele 2013). However, Carfilzomib, as Bortezomib, is unable to penetrate solid tumors (Stein, Cui et al. 2014).

Despite of the success of the covalent peptidic inhibitors in the clinical field there is place for improvement. Form a medicinal chemistry perspective, covalent inhibitors are usually avoided due to the risk their over reactivity and therefore possible secondary effects (Zhou, Chan et al. 2005). Peptidic inhibitors, on the other hand, present a series of drawback such as poor metabolic stability, poor membrane permeability and poor oral bioavailability (Craik, Fairlie et al. 2013). In the specific case of the covalent peptidic inhibitor of the 20S proteasome already in the drug market, they have shown inability to target solid tumors, and undesired secondary effects (Stein, Cui et al. 2014). This situation calls for development of non-covalent and non-peptidic molecules. In this regard, there are only few X-ray crystal structures of the 20S proteasome in complex with non-covalent inhibitors. These inhibitors are to TMC-95A and its derivatives, hydroxyurea derivatives, and K-7174 a homopiperazine derivative.

TMC-95A and its derivatives are potent non-covalent proteasome inhibitors isolated from Apiospora montagnei (Koguchi 2000). These natural products, although still peptidic, have a cyclic structure that provides them a low entropic cost upon binding. TMC-95A is highly specific and potent towards the 20S proteasome, it inhibitory activity reaches the nanomolar regime at the chymotrysin-like active site and the micromolar range at the other two catalytic centers (Koguchi 2000). Despite of these advantages, the cyclic structure of TMC-95A and their derivatives represents a complex synthetic challenge that prevents their use for commercial purposes. Linear analogs have been made, in order to simplify the chemical synthesis process, but a considerably loss in potency has been encounter (Groll, Gallastegui et al. 2010).

The yeast 20S proteasome in complex with a hydroxyurea compound was the first reported structure showing a non-covalent and non-peptidic ligand binding mode (Gallastegui, Beck et al. 2012). These hydroxyurea inhibitors show specificity for the chymotrypsin-like active site of yeast 20S proteasome. Its non-covalent interaction proved to be unique in comparison with other inhibitors uncovering two side pockets to S1 and S3 termed the S1’ and the S3’ pockets. These inhibitors were optimized leading to a highly potent inhibitor with a Ki value in the nanomolar range termed H10. H10 showed to make two strong hydrogen bonds between the hydroxyurea moiety and two highly important residues. This inhibitor had two side chains; one

located in the S1' pocket (methyl group) and the second one in the S3' (the adamantly group). Unfortunately, this inhibitor showed poor solubility and cell permeability properties (personal communication with the authors).

K-7174 a homopiperazine derivate is another interesting non-covalent and non-peptidic proteasome inhibitor (Kikuchi, Shibayama et al. 2013). This molecule inhibits the all active subunits by occupying the primed binding pockets. Due these characteristics, it is able to work concomitantly with Bortezomib in the induction of apoptosis in myeloma cells (Kikuchi, Yamada et al. 2013) . Promising results are expected from Homopiperazine derivatives that compensate for the weak points of the covalent peptidic inhibitors described before.

It is important to highlight that most of the 20S inhibitors design towards hindering the β5 subunit activity, which has been identify the essential for the 20S proteasome operation (Parlati, Lee et al. 2009). Nevertheless, despite of the high degree of homology, the structural variation among subunits of different species, including the immuno variations, render different active sites, thus opening the possibilities of achieving specificities among species. For instance, the development and design of inhibitors specific for human immunoprotesome is highly desirable for treatment of autoimmune diseases (Huber and Groll 2012).

## 2.3 Initial virtual screening on the 20S

The use of Structure-based Virtual Screening offers the possibility of exploring the molecular landscape in search of new inhibitors against the 20S proteasome. Here is presented the initial steps taken towards the challenging goal that represents the identification of novel proteasome inhibitors. It is important to stress that this preliminary results are intentionally presented as part of the introduction, since it provided the platform for the development of the research that constitutes the main part of this dissertation.

### 2.3.1 Reliability score and docking map

As previously mentioned, two aspects of the structured-based Virtual Screening (VS) have great influence on the results: i) Human visual inspection of the docking poses and ii) the model of the active site. The visual evaluation of the docking results is crucial in the reduction of false positives (Klebe 2006). However, to perform this assessment for a large compound library is utterly unreliable, thus calling the development a post-analysis that provides a trustable ranking of screened molecules. On the other hand, the dependency of the protein model can be mild by using either several active site models or by considering a flexible active site (Sinko, Lindert et al. 2013). Given these scenarios, an automated selection a process was devised attempting to reduce de number of molecules on which perform the visual evaluation. Due to the vast structural information on the β5 subunit, the protocol was performed to work on several crystal structures.

The core idea of the protocol was that within a single screening experiment, i.e., docking of a chemical library into a single active site, molecular candidates were selected automatically based on a "reliability score", which tries to capture important aspects that would render a docked molecule worth to evaluate visually. The reliability score was computed as the weighted sum of values derived from the virtual experiment: *i*) average of high scored poses, *ii*) rmsd among the poses, and *iii*) number of poses in the top half of the rank comprised by whole screened library,

$$reliability\ score = W_{Dscore}D_{Score} + W_{rmsd}Rmsd + W_{Freq}Freq$$

Only molecules satisfying these cut-off values were selected as candidates and their reliability score was computed. Since the docking performance of the molecule depends on the protein crystal structure, each model of the active site had its own set of candidates. The performance of all these molecules over the whole set of protein structures was depicted in a "Docking Map"

(Figure 4) with idea to provide a fast visual way to identify if a molecule has a preference for a certain set of structures.

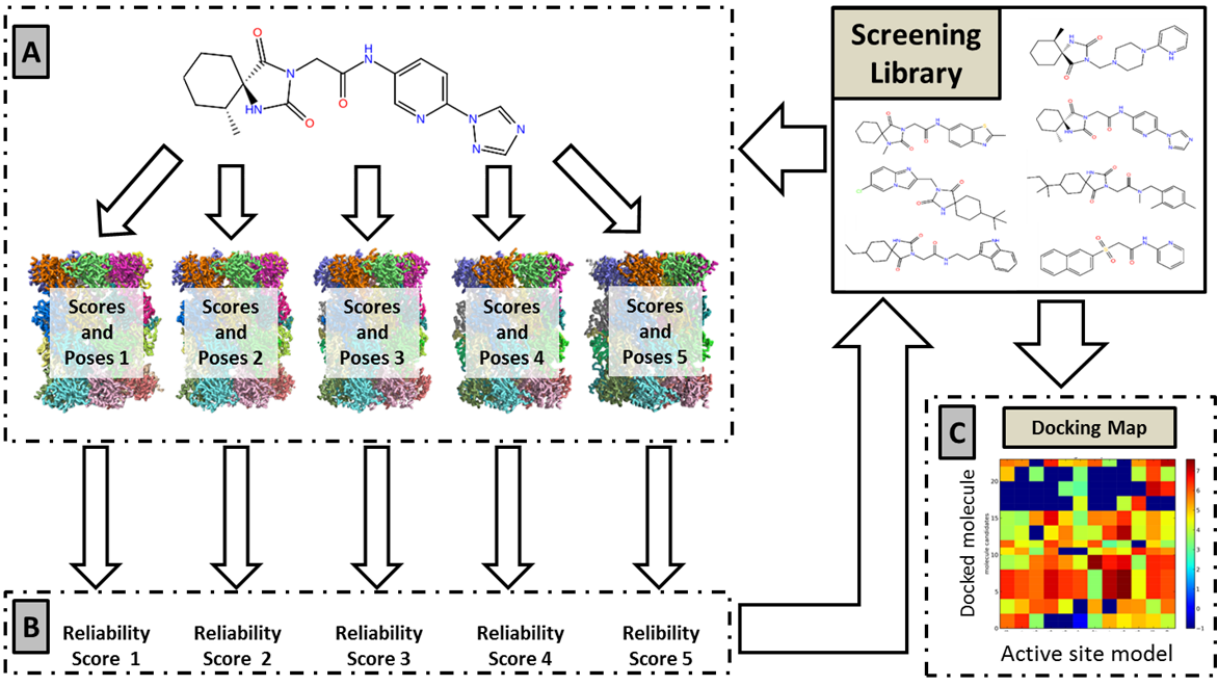



Figure 4. Docking map algorithm for virtual screening. A) Each molecule in the chemical library is docked in each structure of the ensemble. B) For each structure, the molecules are evaluated by the "reliability score". C) The a set of candidate molecules are reported in docking map scheme, where the performance each candidate molecule on ensemble of protein structures is shown.

### 2.3.2 Initial test

The "Docking Map" was tested on a library of 3000 purchasable compounds containing a hydroxyurea-like scaffold. This library was docked into an ensemble constituted by yeast and mouse constitutive as well as the mouse immuno proteasome structures. The objective of the test was to evaluate the capacity of the algorithm to provide a fast visual evaluation on whether a molecule has preference for certain set of structures and to reduce the amount of molecules on which to perform human evaluation. From the screening of the 3000 compounds (first library), a very noisy map was obtained, although some interesting features were observed (Figure 5A). For example, molecules from 5 to 8 showed a very promiscuous binding by scoring relatively good in all the active site structures. In contrast, candidate molecules 9 and 10 showed preference for a sub-set of the yeast structures, thus indicating presence of conserved

structural difference among yeast active sites. The candidate molecule 1 generated particular interest, since it showed slightly more preference for the iCP structures than to the cCP. To test whether additional molecules resembling molecule 1 could enhance this signal, 200 new molecules were included in the analysis. These molecules were selected from online databases satisfying the condition to be at least 0.7 similar to candidate molecule 1 in the tanimoto scale. With this enriched library (second library) the specificity signal increased in intensity for just one molecule (Figure 5A). Interestingly, this new molecule (candidate molecule number 1 in the 2nd screening) showed a very similar scaffold to some molecules of the first screening (Figure 5B). Based on the visual inspection of these molecules docking poses a new set of 50 molecules was manually created and included in the analysis (third library). The docking of this new set of molecules produced an outstanding result in their preference towards the mouse immune proteasome (Figure 5A), especially candidate molecule number 7.

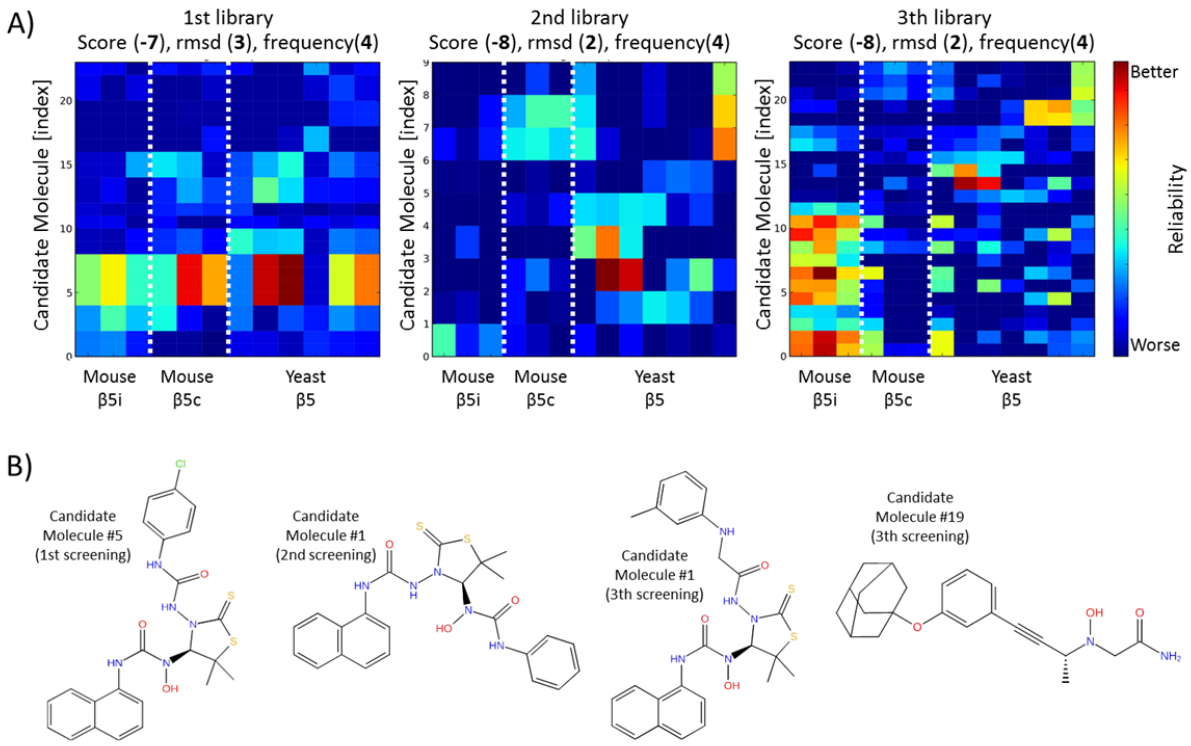



Figure 5. Testing of the Docking Map algorithm. A) Docking maps computed from virtual screening data from three chemical libraries referred in the text. B) Examples of the molecules showing an interesting docking performance in accordance with the docking map.

Despite of the interesting results indicating the possible identification of a set of molecules with a clear preference for the immunoproteasome, it was decided to test experimentally the

candidate molecules for the yeast system. This decision was taken due to the complicated procedure that is required to purify and crystalize the mouse immunoproteasome, and the more economical screening system established for yeast proteasome in our collaborators laboratory, Prof. Groll. From results based only on yeast proteasome, 9 of the suggested molecules were acquired and tested in a flourometric assay (Philipp Beck, Prof. Groll lab.). Interestingly within this small set of compounds 2 molecules were able to provide a weak inhibition at β5 active site (Figure 5). However, their crystallization did not show any electron density.

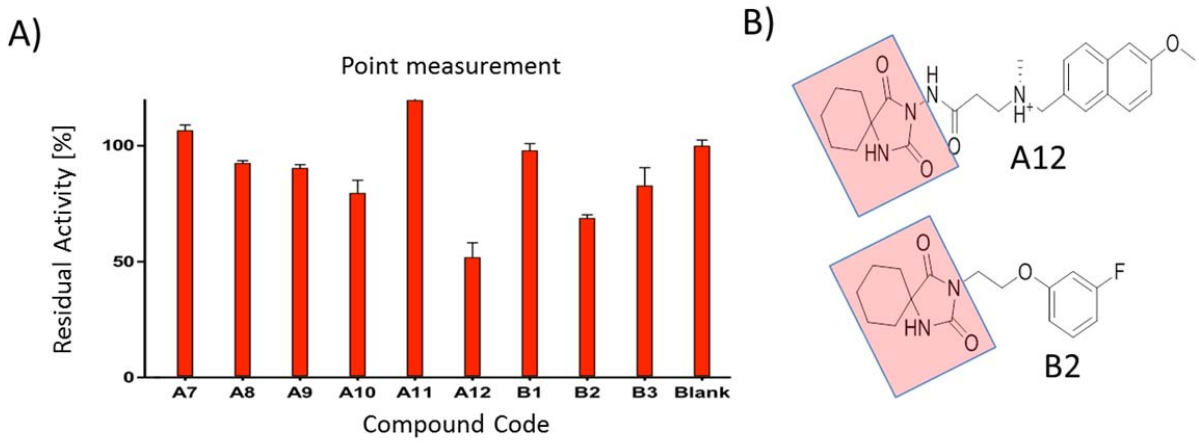



Figure 6. Testing of commercial molecules on yeast 20S proteasome. A) Compounds bought directly from chemical providers were tested on fluorometric assay on yeas 20S proteasome. B) The two molecules showing some inhibitory activity (approx. $IC_{50}$ 150μM) show a similar scaffold (pink squares).

### 2.3.3 Conclusions from the initial test

Several conclusions are drawn from this first attempt to perform Virtual High Throughput Screening on 20S proteasome.

- Despite of the lack of electron density in the crystallization of the weak active molecules, their identification by structured-based virtual screening resulted quite promising, considering that two out of nine tested molecules show weak activity. In addition these molecules shared a common chemical feature, thereby offering inhibitor design possibilities.

- The “Docking Map” algorithm provided evidence on the existence of structural features in the ensemble of β5 subunits that made possible to distinguish between proteasome types, constitutive, immune, yeast and ,even, among different yeast structures.

- The “reliability score” proved to be an interesting concept that was worth to develop further, for example by providing an adequate parameterization of the weighting factors.

From the perspective provided by these conclusions, it was clear that to increase the certainty of the Virtual Screening results on 20S proteasome two main aspects needed to be addressed:

- A deeper understanding of the structural differences among the different proteasome species in imperative, since despite of the high similarity the subtle differences among them are utterly crucial for the structured-based Virtual Screening purposes.

- A refinement of the “reliability score” idea as post-analysis method for Virtual Screening experiments needed to be formulated. This refined methodology should be rigorously benchmarked against broad spectrum of protein targets.

These two aspects motivated the work developed on two peer-review publications presented in the sections following the methods section and that constitute the rest of this document.

# 3. Methods

## 3.1 AutoDock4.2

AutoDock4.2 uses semi-empirical scoring function to estimate the energy difference ($\Delta G$) between the unbound and bound states of the protein-ligand complex (Huey, Morris et al. 2007).

$$\Delta G = \left(V_{bound}^{P\ intra} - V_{unbound}^{P\ intra}\right) + \left(V_{bound}^{P-L} - V_{unbound}^{P-L} + \Delta S_{conf}\right)$$

Where $\Delta S_{conf}$ is an entropy approximation penalizing the restriction of rotatable bonds upon binding and the potentials ($V$) are computed from the pair wise interaction of atoms within the protein ($P\ intra$) and between the protein-ligand complex ($P - L$). The potentials take the form:

$$V = W_{vdw} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + W_{Hbond} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right)$$

$$+ W_{elec} \sum_{i,j} \left(\frac{q_i q_j}{e(r_{ij}) r_{ij}}\right) + W_{sol} \sum_{i,j} \left(S_i v_j + S_j v_i\right) e^{\left(-r_{ij}^2 / 2\sigma^2\right)}$$

The weights ($W$) are calibrated to fit experimental binding constants. The first term accounts for attraction/repulsion interactions. The second term evaluates the hydrogen bond interactions based on distance ($r_{ij}$) and directionality ($E(t)$). The third term accounts the electrostatic interaction using a columbic potential. The fourth term is the desolvation potential based on the volume of atoms ($v$) that surround a given atom and shelter it from the solvent. This interaction is weighted ($S$) and decays exponentially ($\sigma = 3.5$ Å). The pose search is performed by a genetic algorithm under a Lamarckian scheme.

## 3.2 AutoDock Vina

In AutoDock Vina the scoring function is combination of knowledge-based potentials and empirical scoring functions (Trott and Olson 2010). The conformation is evaluated ($C$) by accounting interactions *intra* and *inter* the molecules

$$C = C_{inter} + C_{intra}$$

The summations run over all pair of atoms (*i,j*), excluding atoms separated by three consecutive covalent bonds (1-4 interactions). The interactions are defined in terms of the surface distance $d_{ij} = (r_{ij} - Atomic\ radius_i - Atomic\ radius_j) * (1\text{Å})^{-1}$ and have associated a weight ($W$)

$$C = W_{gauss1} \sum_{i<j} e^{\left(-d_{ij}/0.5\right)^2} + W_{gauss2} \sum_{i<j} e^{\left(-(d_{ij}-3)/2\right)^2} + W_{repul} \sum_{i<j} \begin{cases} d_{ij}^2, & if\ d_{ij} < 0 \\ 0, & if\ d_{ij} \geq 0 \end{cases}$$

$$+ W_{hydrophobic} \sum_{\substack{i<j \\ if\ applies}} \begin{cases} 1, & if \quad d_{ij} < 0.5 \\ 0, & if \quad d_{ij} > 1.5 \\ -d_{ij} + 1.5, & other\ wise \end{cases}$$

$$+ W_{Hbond} \sum_{\substack{i<j \\ if\ applies}} \begin{cases} 1, & if \quad d_{ij} < -0.7 \\ 0, & if \quad d_{ij} > 0 \\ \frac{d_{ij}}{-0.7}, & other\ wise \end{cases}$$

Note the last two terms are applicable only for the appropriate pair of atoms, whereas the steric terms (the first three terms) consider all atom pairs. The pose search algorithm used by Autodock Vina is an Iterated Local Search global optimizer adapting a Broyden-Fletcher-Goldfarb-Shanno (BFGS) method for the local optimization.

## 3.3 Rosetta Ligand

The all-atom scoring function in Rosetta comprises weighted individual terms that are summed to create an estimation of the total energy (Combs, DeLuca et al. 2013). Most of these terms are knowledge-based potentials. The van der Waal interactions are represented by a 6–12 Lennard-Jones potential with attractive and repulsive terms. The solvation energy is modeled by an implicit water scheme where the burial of polar atoms is penalized. The electrostatic interactions are modeled by the scoring function contains Newtonian physics–based terms, including a 6–12 Lennard-Jones and a solvation potentials. The 6–12 Lennard-Jones potential is split into attractive and repulsive terms, and represents the van der Waals interactions. The solvation potential is an implicit water model that penalizes the burial of polar atoms. The electrostatic interactions are captured through a pair potential, and a hydrogen bond potential that accounts for long-range and short-range, as well as, directionality hydrogen bonding. Additionally, the

scoring function has rotametic terms that dictate side chain conformations according to the Dunbrack rotamer library.

## 3.4 Molecular Dynamic Simulations (Gromacs)

In a system of *N* interacting atoms, for example a protein solvated in water, molecular dynamics (MD) simulation allow to evaluate dynamical properties of that system by solving Newton´s equation of motion (D. van der Spoel, E. Lindahl et al. 2010). With each atom having a coordinate ($\boldsymbol{r}_i$), usually provided by X-ray structures, an atomic mass ($m_i$), provided by the atom type, and a velocity, usually randomly assigned according to the Boltzmann distribution, the system of equations to solve is:

$$\frac{\partial^2 \boldsymbol{r}_i}{\partial t^2} = \boldsymbol{F_i}, \qquad i = 1, \dots, N$$

With the forces $\boldsymbol{F_i}$ being the negative derivate of a potential function $V$ that defines the interactions among the particles

$$\boldsymbol{F_i} = -\frac{\partial V}{\partial \boldsymbol{r_i}}$$

Any other properties specific to each atom in the system are considered to be within the potential function $V$. Additionally note that $\boldsymbol{F_i}$ and $\boldsymbol{r_i}$ are vectors.

This system of equations is solved in small time steps, thus providing a time evolution of the system of $\boldsymbol{N}$ interaction atoms that is recorded in a set of individual frames, the *trajectory* of the system. There are several thermostats and barostats that, in principle, allow the simulation to generate conformations according to the canonical (NVT) and isothermal-isobaric (NPT) ensembles. The terms in the potential function can be classified in three categories: i) non-bonded, ii) bonded and iii) restrains. The non-bonded interactions are pair-wise additive and centro-symmetric. They contain repulsion and dispersion terms in the form of Leonard-Jones potentials (6-12 interaction) together with a columbic interaction, which is computed from the assignment of partial charges to the atoms. The bonded interactions contains terms that consider bond lengths, bond angles and dihedral angles, thus involving interactions between two, three and four atoms, respectively. While the non-bonded and bonded interactions attempt to capture the effects of the electron clouds, thus representing real interactions, the restrains

terms in the potential (which are optional to include) represent fictitious interactions that impose motion restrictions on the system, either to avoid disastrous deviations, or to include knowledge from experimental data.

## 3.5 Principal Component Analysis

The principal component analysis (PCA) is a multivariate statistical tool that allows the reduction of high-dimensional data sets onto a group of collective variables. PCA has been successfully applied to extract collective modes of motion in proteins from molecular dynamics simulation trajectories (Kitao, Hirata et al. 1991; García 1992). To study the protein motion of a protein, an ensemble of structures, taken from a molecular dynamic simulation trajectory or from as set of experimental structures, is averaged to generate a reference structure. Every single frame is then compared against the average structure to generate a co-variance matrix, which is also averaged over the whole ensemble.

$$C_{ij} = \langle (\, x_i - \langle x_i \rangle \,)\big(\, x_j - \langle x_j \rangle \,\big) \rangle$$

Thus, the matrix elements $C_{ij}$ contain averaged information of the correlated atomic displacements, $x_i$ and$x_j$, with respect to their average positions, $\langle x_i \rangle$ and $\langle x_j \rangle$. By diagonalizing $C$ this correlated motions can be represented by a set of orthogonal vectors that dissect the global displacements. This diagonalization is solved as an eigenvalue problem

$$A^t C A = \lambda$$

The eigenvectors are represented by the matrix $A$, and the column vector $\lambda$ contains the corresponding eigenvalues. The larger the eigenvalue the larger the movement described by the associated eigenvector. Usually only first few eingenvectors are needed to capture the most relevant domain displacements.

## 3.6 Artificial Neural Networks

An Artificial Neural Network (ANN) is a machine learning model inspired by the operation process observed in biological neural networks (Bishop 2006; Jalali-Heravi 2008). This

computational model is formed by artificial neurons, also referred to as nodes, organized in a set of hierarchical layers, each layer constituted by a certain number of nodes. The nodes of different layers are connected between them for signal transmission. The outputs of the nodes in a given layer become the inputs of the nodes in the subsequent layer. The connections are weighting factors that scale the input received by each node, thus affecting the output signal that is going to be propagated to the nodes of the next layer (Figure 7A).

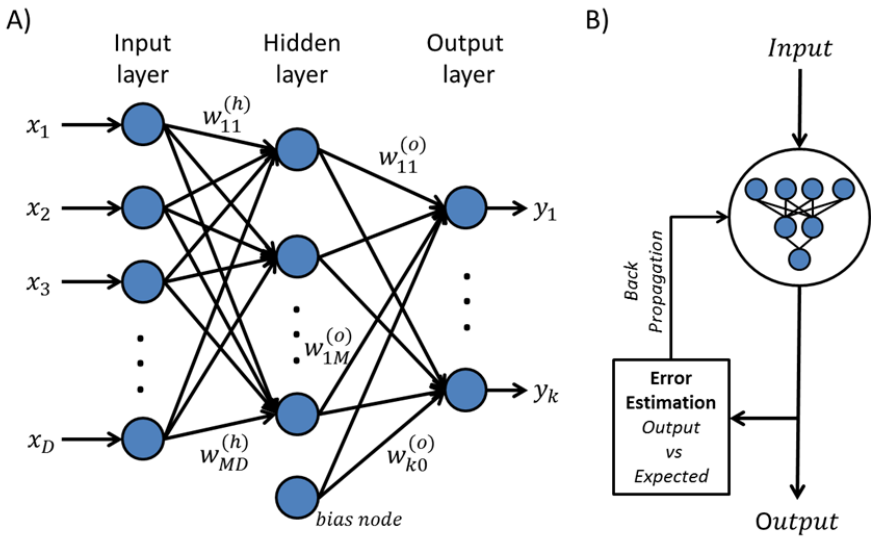



Figure 7. Typical organization of an Artificial Neural Network (ANN) trained under a backpropagation protocol. A) ANN connectivity. The scheme shows a three layer network: input layer consisting of *D* number of nodes, hidden layer with *M* nodes and output layer constituted by *k* nodes. The nodes and the weighted connections are represented by circles and lines, respectively. In sake of clearance, only some connection weights are labeled. The superscripts (h) and (o) are used to differentiate the weights from the hidden and output layers, respectively. In the hidden layer there is an extra bias node that contributes in the transition of signal to the output layer. B) Training with backpropagation. After the set of inputs are analyzed by the ANN, the difference between the ANN´s output and the expected value is computed and used to modify the weights of the network. This error estimation is an iterative process that is used to adjust the performance of the network.

As a machine learning algorithm, the ANN needs to be trained before it can be used for prediction purposes. In one of the most popular trainings, the backpropagation algorithm, the ANN is trained by providing examples of the result expected given certain inputs were provided. A typical architecture of a ANN with a backpropagation training consists of three layers: i) input layer dealing with the input data directly, ii) a hidden layer taking information from the input layer

and sending its own output to next layer, ii) and output layer which returns values that can be evaluated with respect to of the expected results. The signal processing within each node is defined as the sum of the weighted input signals

$$a_j = \sum_{i=1}^{D} w_{ij} x_i + w_{j0}$$

The quantities $a_j$ are linear combinations of the input values (D input values in this case) reaching the node *j* plus an additional bias term $w_{j0}$ (optional). The $a_j$ values, also known as activation values, are then transformed into a single output using a differentiable and nonlinear activation function

$$z_j = h(a_j)$$

This output value ($z_j$) becomes the input that is transmitted to the next layer of nodes (feed-forward architecture). Common functional forms of $h(\cdot)$ are either a logistic sigmoid function

$$z_j = \frac{1}{1 + e^{-a_j}}$$

or a softmax activation function

$$z_j = \frac{e^{a_j}}{\sum_i^D e^{a_i}}$$

The $k$ output values of the ANN ($y$) are compared against a set of $k$ target values in terms of a squared difference over $p$ training patterns

$$E = \frac{1}{2} \sum_p \sum_k (y_{pk} - t_{pk})^2$$

The target values are an essential part of the training process. In back-propagation learning, the error in prediction is used to adjust the weights in order to minimize the error of the next iteration (Fig 7B)

$$\Delta w_{ij}(n) = \eta \delta_i O_j + \alpha \Delta w_{ij}(n-1)$$

Where $\Delta w_{ij}$ represents the change in the weight factor for each network node, $\delta_i$ is the error associated with the node *i* and $O_j$ is the output value of the node *j*. The *learning rate* ($\eta$) and

*momentum factor* ($\alpha$) are parameters that need to be optimized before the training, since they control the velocity and the efficiency of the learning process. Similarly, the number of nodes forming the hidden layer also affects the training of the network. Too many hidden nodes can cause overtraining, thus making the ANN useless for predictions besides the training set, and too few can render the network unable to learn. After training the ANN to a satisfactory level, the weights linking the nodes are used to predict results for new input datasets.

# 4. Description of the first publication

*Differential global structural changes in the core particle of the yeast and mouse proteasome induced by ligand* binding by Marcelino Arciniega, Philip Beck, Oliver F. Lange, Michael Groll and Robert Huber. [http://www.pnas.org/content/111/26/9479] (Marcelino Arciniega contributed in: research design, performing the research, data analysis, writing the paper)

The vast amount of X-ray crystal structures of the 20S proteasome in complex with small molecule inhibitors, mainly on the β5 subunit of yeast proteasome, proved the need of a detailed analysis of these structures to select appropriate models for developing structured-based virtual screening. The influence that the active model has on the virtual screening results was experienced by first hand on the "Docking map" analysis described in the introduction of this dissertation. The initial idea was to generate a classification of the known crystal structures of the β5 subunit of yeast 20S proteasome, and based on this classification identify the set of complexes that best represented the structural plasticity of this active site. Thus this clustering problem was approached using Principal Component Analysis (PCA).

Interesting results were obtained when PCA was applied, on the backbone atoms, of the β5 ensemble of the yeast structures. The PCA showed two well defined clusters that correlated with the presence or absence of peptidic inhibitor at the active site, thus allowing the classification of the structures in pep- and apo-clusters (please notice the apo cluster also contained structures in complex with non-peptidic inhibitors). These results were surprising, since in 15 years of structural work on inhibitor design for the 20S proteasome, such movement has never been described. The only reported backbone movement, induced by a peptidic inhibitor, was that of β5 subunit in the mouse constitutive. Interestingly, the analysis of mouse β5 constitutive and immunoproteasome structures from the perspective of the yeast ensemble revealed that the backbone movement observed in the constitutive mouse proteasome is essentially the same than that of the yeast proteasome. The mouse immunoproteasome structure appeared, on the other hand, to maintain a structure similar yeast pep-cluster even in its apo state. The relevance of the peptidic binding in inducing the backbone movement was supported by Molecular Dynamic simulations results and the X-ray crystal structure of a Boc-$(Ala)_3$-COH inhibitor in complex with the yeast β5 subunit. The implications that this movement has on the rest of the subunits of the 20S proteasome were also investigated. From this analysis, possible communications pathways between i) the *cis* and *tras* β5 subunits and ii) the β5 active site and the α-subunits were revealed.

# 5. Description of the second publication



This work corresponds to the necessity to develop an automated algorithm that allows a better selection of compounds in the context of structured-based virtual screening and corresponds to the development of the concept of “reliability score” presented in the introduction of this dissertation. The idea behind it is to evaluate the docked compounds based on an evaluation of a set of features that would render a docked molecule interesting for experimental testing. These features, derived from the analysis of their docking data, represent concepts that would be considered by a human user as indicators of the molecule´s inhibitory activity. Since the relative importance or weight that each of these features would receive from a human user depend on his expertise in working with a given docking program, the problem of adjusting such weights called the use of a machine learning approach using an Artificial Neural Network (ANN).

To test this idea, three different docking programs were used on a virtual screening benchmark consisting of datasets of ligands and decoys for 40 different proteins. Additionally, a consensus approach was established combing information for the three independent docking programs. From the data generated by each individual program five docking features were observed on each screened molecule: i) best docking score, ii) ligand efficiency, iii) scores from similar molecules, iv) the position of the ligand’s poses within the general rank, and v) structural consistency of the ligand’s poses. The information associated to each screened molecule was used to feed the input layer of an ANN that returned a single value assessing the activity chances of a given molecule. The ANN was trained, validated and applied on each the 40 datasets.

The virtual screening performance of this approach, called Docking Data Feature Analysis (DDFA), was compared to that of the traditional method, which consists in rank the screened libraries based only on the docking score. The DDFA methodology outperformed the traditional ranking method, achieving substantial improvement. The robustness of the methodology and the significance of the improvement were also assessed. When comparing the DDFA performance with other methodologies found in the literature, the DDFA results were as good as the best of the reported algorithms.

# Differential global structural changes in the core particle of yeast and mouse proteasome induced by ligand binding

Marcelino Arciniega[a,b,1], Philipp Beck[b], Oliver F. Lange[c], Michael Groll[b], and Robert Huber[a,b,d,e,1]

[a]Emeritus Group Structure Research, Max Planck Institut für Biochemie, 82152 Martinsried, Germany; [b]Center for Integrated Protein Science at the Department Chemie, Lehrstuhl für Biochemie, Technische Unversität München, 85748 Garching, Germany; [c]Biomolecular NMR and Munich Center for Integrated Protein Science, Lehrstuhl für Chemie, Technische Universität München, 85747 Garching, Germany; [d]Zentrum für Medizinische Biotechnologie, Universität Duisburg-Essen, 45117 Essen, Germany; and [e]School of Biosciences, Cardiff University, Cardiff CF10 3US, Wales, United Kingdom

Contributed by Robert Huber, May 8, 2014 (sent for review April 14, 2014)

**Two clusters of configurations of the main proteolytic subunit β5 were identified by principal component analysis of crystal structures of the yeast proteasome core particle (yCP). The apo-cluster encompasses unliganded species and complexes with nonpeptidic ligands, and the pep-cluster comprises complexes with peptidic ligands. The murine constitutive CP structures conform to the yeast system, with the apo-form settled in the apo-cluster and the PR-957 (a peptidic ligand) complex in the pep-cluster. In striking contrast, the murine immune CP classifies into the pep-cluster in both the apo and the PR-957–liganded species. The two clusters differ essentially by multiple small structural changes and a domain motion enabling enclosure of the peptidic ligand and formation of specific hydrogen bonds in the pep-cluster. The immune CP species is in optimal peptide binding configuration also in its apo form. This favors productive ligand binding and may help to explain the generally increased functional activity of the immunoproteasome. Molecular dynamics simulations of the representative murine species are consistent with the experimentally observed configurations. A comparison of all 28 subunits of the unliganded species with the peptidic liganded forms demonstrates a greatly enhanced plasticity of β5 and suggests specific signaling pathways to other subunits.**

20S proteasome | PCA analysis | allosteric regulation

Among the many factors involved in protein degradation through the ubiquitin-proteasome pathway, the core particle (CP) 20S proteasome plays the key role of the protease component. With the regulatory particle (RP), it forms a complex that selectively degrades ubiquitin-protein conjugates (1, 2). The CP in eukaryotes is a multisubunit complex composed of four stacked heptameric rings: two identical outer rings formed by seven different α subunits and two identical inner rings formed by seven different β subunits. The $\alpha_{1-7}\beta_{1-7}\beta_{1-7}\alpha_{1-7}$ organization defines a cylindrical structure (3). The α-rings control substrate entry into the lumen of the particle, where it is processed at the peptidolytic active centers, which are located at the inner walls of the β rings, specifically at subunits β1, β2, and β5. These active subunits are characterized by an N-terminal Thr residue. The other four β subunits have unprocessed N-terminal propeptides and are enzymatically inactive.

All three active subunits share a common peptide hydrolyzing mechanism with two main steps (4): (*i*) the positioning of the substrate peptide in the active site by antiparallel alignment in between segments 47–49 and 21 of the active β subunits and (*ii*) peptide bond cleavage initiated by a nucleophilic attack of the hydroxyl group of the N-terminal Thr1 on the carbonyl carbon atom of the scissile peptide. Sequence diversity among β subunits endows them with distinctive structural features and different specificity pockets (S1, S2, S3, etc.) where the substrate side chains (P1, P2, P3, etc.) are bound (5). Consequently, the correlation of structural features of the S1 pockets with the distinctive cleavage products has led to the association of β1, β2, and β5 with caspase-like, trypsin-like, and chymotrypsin-like activities, respectively (6).

The catalytically active subunits are substituted in immune cells of vertebrate organisms by the immune β-subunits β1i, β2i, and β5i as part of an adaptive immune response. These substitutions cause substantial functional differences between the constitutive (cCP) and immuno (iCP) species, reflected in higher yield of peptides that are recognized by the major histocompatibility complex (MHC) class I generated by iCP (7). Additionally, it has been observed that iCP achieves higher degradation rates than cCP, in both in vitro and cellular assays (8–13).

Some sequence variations between the constitutive and immune subunits provide explanations to the observed catalytic differences. Most conspicuously, and first seen in the eukaryotic proteasome crystal structure from yeast (yCP) (3) and confirmed by the murine constitutive and immune CP structures (mcCP and miCP) (14), Arg45 of the β1 subunit, located at the base of the S1 pocket, is replaced by leucine in β1i, thereby causing a specific change of the electrostatic milieu, in line with the observed low postacidic activity of the iCP (15).

Despite the high sequence similarity between β5 subunits of mcCP and miCP including identical active sites, a peptidic α-β-epoxyketone inhibitor, PR-957, showed higher affinity to iCP by one order of magnitude. The structural comparison of cCP

## Significance

We analyzed 46 molecular structures of the yeast proteasome core particle (CP) by principal component analysis (PCA) and discovered two distinct configurations of the principal proteolytic subunit β5: the apo-cluster encompassing complexes with nonpeptidic ligands and the pep-cluster of complexes with peptidic ligands. Both configurations differ by a small domain motion and numerous slight global changes, thus enabling intersubunit communication. PCA was expanded to the mouse CP and revealed a striking difference between the constitutive CP and the immune CP. The former conforms to the yeast system and executes the structural change seen in yeast, although both immune apo and liganded CP classify into the pep configuration, a possible explanation for the generally higher activity of the immune proteasome.

Author contributions: M.A., M.G., and R.H. designed research; M.A. performed research; P.B. contributed new reagents/analytic tools; M.A., O.F.L., and R.H. analyzed data; and M.A. and R.H. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The structure has been deposited in the Protein Data Bank, www.pdb.org (PDB ID code 4QBY).

[1]To whom correspondence may be addressed. E-mail: castro@biochem.mpg.de or huber@biochem.mpg.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1408018111/-/DCSupplemental.

and iCP in their apo and PR-957 liganded states suggested an explanation. On binding of PR-957, the cCP β5 backbone displays significant deformations, whereas the iCP β5 backbone remains unchanged. This observation, together with our experience in constructing β5 models for virtual screening purposes, prompted us to reinvestigate the vast amount of structural data for yCP by a procedure that facilitates discovery of global changes: principal component analysis (PCA).

We focus our study on the β5 subunit, because β5 inactivation in yeast renders a lethal phenotype (16) and therefore β5 harbors an essential enzymatic activity, and because almost all crystallographically defined complexes are liganded at their β5 active site.

Here we present a detailed investigation of the wealth of yeast and mouse proteasome ligand complex structures that led us to embark on structural comparisons beyond the immediate vicinity of the ligands to obtain a view of the global response of the core particle of yeast and mouse proteasome to complex formation. This study (*i*) is evidence of the structural plasticity of the β, specifically β5, subunits; (*ii*) offers perspectives for the analysis of the structure-function relationship of the CP; and (*iii*) provides an aid for the design and development of ligands as drugs for this intensively studied target for cancer and autoimmune diseases.

## Results

**Structural Transition of Subunit β5 on Peptidic Ligand Binding.** We analyzed 46 β5 subunits from yCP crystal structures (yβ5) reported in the Protein Data Bank (PDB) database (Tables S1 and S2) aiming to identify backbone transitions induced on inhibitor binding by PCA because it eases structure classification (17–19).

The projection of each structure onto the first two PCA eigenvectors revealed two clusters clearly distinguished by the first principal component (PC1) (Fig. 1). Although PC1 captures 78% of the structural variances, the second principal component (PC2) accounts for less than 5% (Fig. 1, *Inset*). The clusters are highly correlated with inhibitor binding. On the positive *x* axis, structures in either the apo-state or in complex with nonpeptidic inhibitors are found, whereas the projections on the negative side correspond to structures with peptidic inhibitors. In the following, we refer to these clusters as apo- and pep-clusters, respectively. Striking results are obtained by projecting the β5 structures of the murine (mβ5) and bovine (20) (bβ5) species on the eigenvectors of the PCA analysis of the yeast structures. The constitutive subunits, mβ5c and bβ5c, are classified into the apo-cluster in their ligand free forms, whereas the uncomplexed structure of the immune subunit, mβ5i, colocalizes in the pep-cluster together with the mammalian liganded structures.

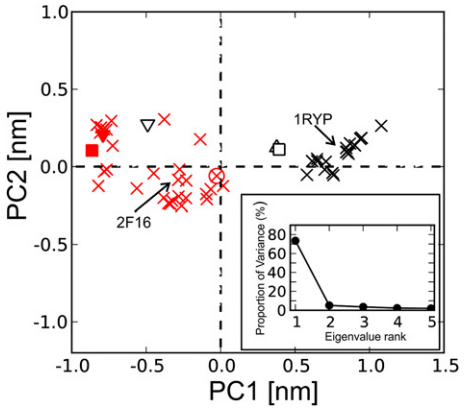



**Fig. 1.** Projection of the β5 structures on the first two eigenvectors of PCA (PC1, PC2). Structures of yβ5 (×), mβ5c (□), mβ5i (▽), and bβ5c (△) are classified into two groups; structures in complexes: (*i*) peptidic inhibitors (red) and (*ii*) nonpeptidic inhibitors or apo structures (black). Filled symbols indicate structures in complex with PR-957. The Boc-$(Ala)_3$-al yCP complex is represented by ⊗. The arrows indicate the position of apo (1RYP) and Bortezomib complexed (2F16) structures. (*Inset*) Eigenvalue spectrum representing the percentage of the total variance captured by the corresponding eigenvector.

These observations led us to investigate the contribution of the amino acid side chains of the peptidic inhibitors to the observed structural perturbation. Thus, we synthesized the peptidic Boc-$(Ala)_3$-al inhibitor (*SI Methods*) and determined the yCP cocrystal structure with the aim to discern contributions from the main chain and side chains. The small and neutral methyl groups do not fill the subsites and are expected to exert minimal influence. Interestingly, the complex structure clusters with the pep-series (Fig. 1), thus excluding a decisive role of the side chains and underpinning the importance of the main chain on the classifications.

To highlight the domain motion induced by binding of peptidic ligands, structures from each of the two clusters were overlaid considering residues 1–39 and 125–190 (Fig. 2). The protein segments containing residues T21, G47, and A49 accept and donate, respectively, four hydrogen bonds to peptidic ligands, configuring a short three-stranded antiparallel β-sheet (Fig. 2*A*). This binding appears to trigger the closure of the pocket relative to the apo-series by shifting the α-helix (H1), comprising residues 49–70, of yβ5 (Fig. 2*B*). A similar reorganization is observed for the mβ5c on binding of a peptidic inhibitor (Fig. 2*C*) but is absent in mβ5i, whereas both apo and liganded structures present a closed conformation to establish the antiparallel β-structure (Fig. 2*D*). Similar comparison of apo and peptidic liganded bacterial proteasomes, using the mycobacterial CP structures (21), did not indicate molecular rearrangements as seen in yeast.

To support the previous observations on the role of peptidic binding, we performed a series of molecular dynamics (MD) simulations of 20 ns length on a truncated model of the CP (*SI Methods* and Table S3). We simulate mβ5c and mβ5i under three different starting conditions: (*i*) apo structures, (*ii*) structures in complex with a peptidic ligand, and (*iii*) complexed sctructures but with the ligand removed. The stability of the open and closed conformers is backed up by the simulations (Fig. 3 and Fig. S1). Notably, projecting the MD trajectories of the apo and the liganded forms onto PC1, computed from yeast coordinates, shows that during the simulated time, the structures remain within their corresponding apo- and pep-cluster regions (Figs. 1 and 3 and Fig. S1). In contrast, deletion of the ligand from mβ5c and mβ5i impacted the simulation very differently. After 400 ps (Fig. 3) and 2 ns (Fig. S1) in two independent runs, the mβ5c shifts from the pep- to the apo-cluster region, remaining there for the rest of the simulation, whereas mβ5i stays in the pep-cluster region. Both runs of the MD trajectories sampled regions that deviate quantitatively from those defined by the PCA analysis of the yeast structures (Fig. 1), but the histograms reveal a qualitative agreement and clearly two populations, the apo- and pep-clusters, respectively. We attribute this quantitative deviation to the MD equilibration process of the crystal structures and to the truncated approximation of the CP. Taken together, the PCA analysis (Fig. 1), the structural overlays (Fig. 2), and the MD simulations (Fig. 3 and Fig. S1) document that mβ5c follows the pattern seen in yeast, showing a domain closure on peptide binding and hydrogen bond formation. Notably this transition does not occur in mβ5i, where both apo and pep structures remain in the closed conformation.

**Implications of Peptidic Binding on the CP Structure.** An identical PCA analysis for all 14 subunits indicated similar, albeit less pronounced, clustering only for β4 and β6, which are adjacent to β5 in the heptameric ring. Nonetheless, to further explore whether the state of the β5 active site has an influence on the rest

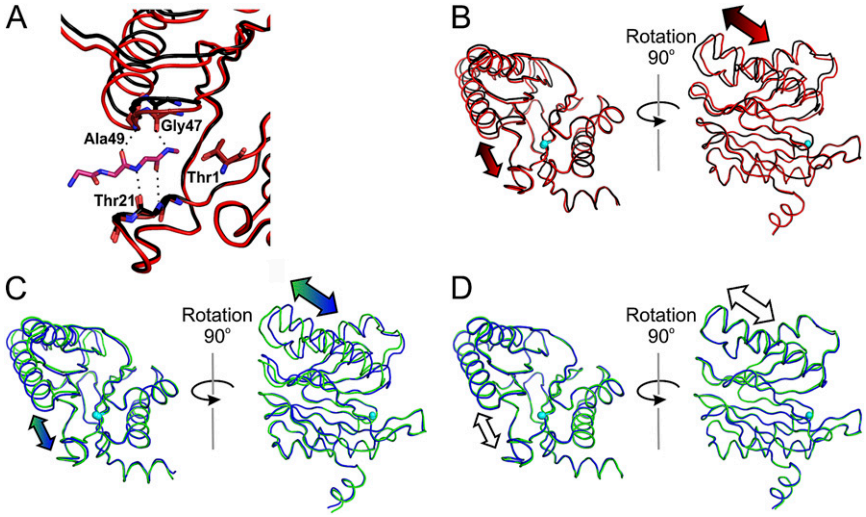



**Fig. 2.** Domain motion in subunit β5 induced by binding of peptidic ligands. (*A*) Peptidic binding at the active site of yeast β5 (apo in black, liganded in red). (*B*) Structural superposition of yβ5 structures: apo (1RYP, black) and complex with a peptidic inhibitor (2F16, Bortezomib, red). (*C*) Superposition of mβ5c: apo (3UNE, green) and complex with a peptidic inhibitor (3UNB, PR-957, blue). (*D*) Superposition of mβ5i: apo (3UNH, green) and complex with a peptidic inhibitor (3UNF, PR-957, blue). Arrows highlight the domain movement of yβ5 and mβ5c. Ligands are not shown in *B*–*D* for clarity.



of the CP structure, each of the 13 remaining subunits was analyzed from the β5 perspective. For this purpose, the structures were classified into the apo-cluster, the elements of the β5 pep-cluster with β5 specific ligands, and the elements of the β5 pep-cluster with ligands also bound at β1 and/or β2. Average structures from these three sets were computed, and the differences for each subunit in their Cα coordinates were mapped along the polypeptide chain (Fig. 4). In accordance with the PCA data, the binding of a peptidic inhibitor at β5 induces significant backbone shifts over the whole subunit. Notably, these are propagated to the neighboring subunits: β4 and β6, and, to a smaller extent, to β2. Interestingly, only minor displacements are induced by peptidic binding at β1 and β2. In contrast to the clear perturbations observed in β subunits, only small and localized signals are observed in the α subunits. The backbone shift in α3 around residue 220 can be associated with the displacements in the β rings. This structural change is further supported by a hydrogen bond between $\alpha 3N_{221}N\gamma$ and $\beta 2D_{220}O$ and by the observed displacement of the last 20 residues of β2 that are in close contact with β3 and the shifted region of β6. Perturbation pathways explaining the other displacements of the α subunits are not evident. Intriguingly, residues known to be involved in the assembly of the 26S particle (residue 66 in the α subunits) and located in proximity of the gate channel (residue 129) are found within the shifted regions. To assess the effect of peptidic binding from a different perspective, the crystallographic "temperature" B (disorder) factors of the main chain atoms were averaged over the same set of structures as in the Cα analysis (Fig. S2). Interestingly, the analysis shows that the variations of the B-factors of the β-subunits are substantially higher than in the α-subunits and follow the trend seen for the structural alterations on ligation (Fig. 4). Taken together, the observed variations of Cα positions and B-factors testify to enhanced structural plasticity of the β subunits, specifically β5, and mark possible pathways for intersubunit communication.

To gain further insights into possible shift conferment pathways, we analyzed Cα differences in β5 (Fig. 5*A*). The shift induced in β5 propagates to β6 and β2 through the segments 30–41 and 204–212, whereas the transfer to β4 occurs through the segment 115–144. It is worth mentioning that a single β5 subunit interacts with β4, in both *cis* and *trans*, i.e., within its heptameric ring and across with the adjacent ring, thus enabling communication to the *trans* β5. Helix H1 of β5 neighbors α4 and α5 offers a possible signal pathway to the α rings (Fig. 5*B*). Interestingly, only minor influences are observed in β3, despite its proximity to the ligand binding site of β5. Aside from these observations, at least five distal segments that are not in direct contact with the active site depict considerable backbone displacements (Fig. 5*A*). The role of these five segments was examined for the definition of an apo- or pep-structures by a PCA analysis (Fig. 5*C*). Residues 42–53 produce just enough signal to discriminate among the clusters, which is expected because most of these amino acids are directly involved in peptidic binding. However, when residues 30–41 are considered, a clearer differentiation is observed. This finding is surprising, because most of them are located at the outer surface of the CP, thus far from the ligand binding site. Progressive inclusion of the other distal segments in the analysis increases the distance between the clusters, thereby demonstrating that the signal of peptidic binding propagates to the back of the subunit and thus describing a possible communication pathway between the inner and outer surfaces of the CP. The importance of these segments for peptidic binding at the active site is supported by backbone shifts observed in mβ5c (Fig. 5*D*).

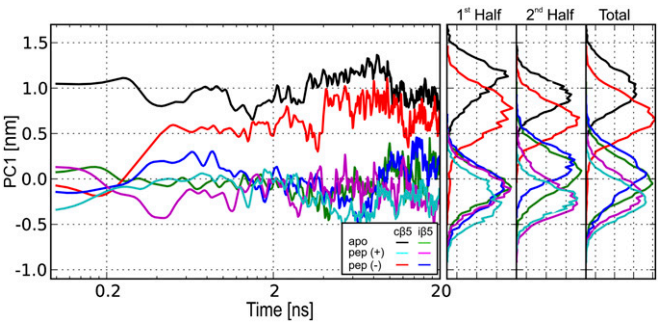



**Fig. 3.** Molecular dynamics simulation. Projections of the trajectories of six mouse β5 structures on PC1 of PCA of the yβ5 structures. The simulated system corresponds to structures starting from the apo (*apo*), liganded [*pep(+)*] and liganded state after removing the inhibitor [*pep(–)*]. The solid lines represent the running averages of 50 frames windows. The histograms on the right are computed with the raw trajectory data.

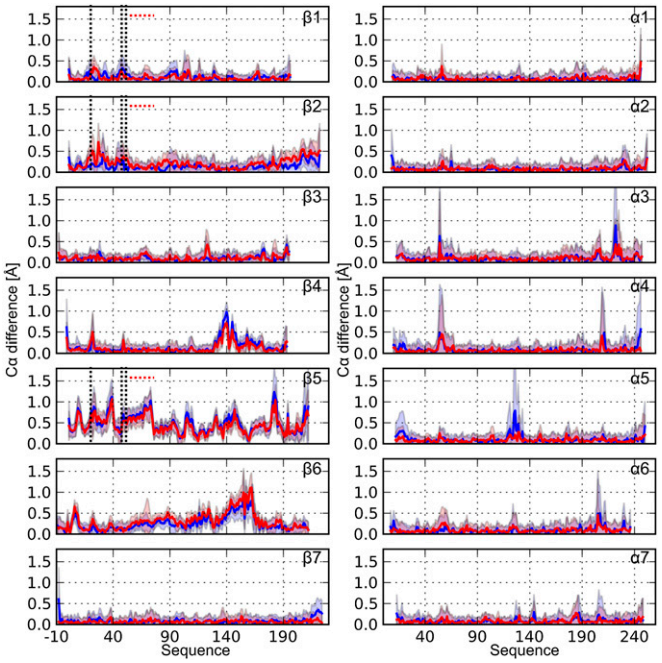



**Fig. 4.** Cα differences of each of the 14 different subunits of yCP. Cα differences between average apo-structure and average complex structures with (*i*) peptidic ligand bound to β5 (blue) or (*ii*) peptidic ligands also in other active subunits (red). The associated error is computed from the SD of the averaged coordinates (shaded regions). Dashed lines highlight residues involved in the peptidic binding (black) and forming the H1 α-helix (red), respectively.

Although similar backbone displacements have been identified in yβ5 and mβ5c, the mβ5i main chain atoms retain the configuration of the pep-structure.

## Discussion

The regulation of an enzyme activity by ligand binding at an allosteric, i.e., distant from the active, center is a frequent phenomenon manifested structurally in changes of quarternary structures as in the founding case of hemoglobin, in large-scale domain rearrangements, disorder-order transitions, and global small scale changes. The underlying mechanisms may be a selection from a preexisting population of configurations (conformational selection) or ligand induced structural shifts (induced fit). An unambiguous distinction between these limiting cases requires measurement of the kinetics of ligand binding. A structural definition may be given by the plethora of tools of structural biology including theoretical and molecular dynamics calculations, with X-ray crystallography being a main source of experimental data. However, the detection of small scale global rearrangements from crystallographic data is challenging when using the common simple visual inspection of superimposed atomic models, which are of limited accuracy. PCA has been introduced and applied to the analysis of conformational ensembles from molecular dynamics simulations and offers also a tool for comparing large sets of experimental structures with the aim of detecting and defining common modes of deformation (22, 23).

We were inspired to apply this method on the proteasome crystallographic data for several reasons. In a large set of structures of yeast CP ligand complexes, no conformational changes were described, whereas functional activity measurements indicated an allosteric interaction between the different active sites (24, 25), which should be reflected in structural changes. However, the interpretations were controversial (26, 27). Additionally, atomic force microscopy (28) and biochemical assay (29) data suggested a correlation between the status of the active site at the β-subunits and features of the α-subunits in yeast CP. In the same line, NMR measurements of the archaeal 20S proteasome indicated an allosteric communication of the active sites with the α-subunits (30, 31). These observations, together with the small domain movement seen in the crystal structures of murine cCP on ligand binding but not in iCP (14), called for a more detailed analysis.

The reinspection of the multitude of differently liganded yCP crystal structures provides further insights into the mechanism of the 20S proteasome. The presented PCA analysis of the yCP was restricted to main chain atoms and focused essentially on the β5 active subunits. Surprisingly, it revealed two conformers, apo- and pep-clusters, unveiling differential specific structural changes on ligand binding and deformations by peptidic ligands which form, in contrast to other types of ligands, a characteristic antiparallel β-sheet with the backbone of residues T21, G47, and A49. We ascribe the wide space drawn out by peptidic ligands in the PCA map compared with other ligands (Fig. 1) to the large contact area in their binding pockets, provoking additional differential distortions (Fig. S3). Peptidic ligand binding at β5 is the principal trigger of all observed conformational changes in yCP and cCP. Additional binding at β1 and β2 has much less influence on those subunits and on β5 (Fig. 4 and Fig. S4). Covalent linkage with Thr1 by various reactants of either nonpeptidic or peptidic ligands does not have a noticeable bearing on the main chain configurations as defined by PCA analysis. Notably, mammalian CPs blended into the clusters showing that there is a consistent structural transformation on peptidic ligand binding among eukaryotic CPs. The dominant influence of the main chain ligand protein interaction is revealed by the fact that yβ5 in complex with Boc-$(Ala)_3$-al clusters with the pep-series. In line with this finding are the results provided by the MD simulations, which give support for the existence of two β5 conformations and highlight the role that peptidic binding has in differentiating these conformations. Although the apo and liganded structures of the mβ5c and mβ5i are stable during the simulation time, the removal of the peptidic ligand from the mβ5c causes a switch from the pep-conformation to the apo-structure within 400 ps and 2 ns, respectively, in two runs. In strong contrast and in agreement with the structural data, removing the peptidic ligand from the mβ5i structure has no effect.

The tight packing of the subunits in the CP together with the observed conformational changes opens the possibility of signal propagation from β5 to other subunits. Most of the conformational alterations in β5 induced by the peptidic binding are distant from Thr1 and located at the surface of the CP (Fig. 5*A*) and hence are not in contact with other subunits and therefore are unlikely involved in direct signal transduction. However, they are evidence for the transit among conformations. Their relevance for the definition of the apo- and pep-clusters is supported by PCA and the murine β5 crystal structures (Fig. 5*C*). Among the backbone displacements that contact other subunits and may generate a direct signal, two are preeminent: the backbone shift induced in both β4 subunits of the segment of residues 115–150 (Fig. 4) and the possible communication with α5 and α4 via the displacement of α-helix H1.

The conformational effects on the β4 subunits by peptidic binding at β5 opens a possible allosteric pathway between the β5 subunits. Interestingly, positive cooperativity of the chymotryptic (β5′s) activity is consistent with this proposal (24, 26, 27, 32).

In regard to the structural rearrangements of α-helix H1, our observation is also in line with the reported importance of this helix in the global motions and allosteric communication observed in the archaeal 20S (30) and in the HslV protease (33), the prokaryotic homolog of the eukariotic proteasome (34, 35).

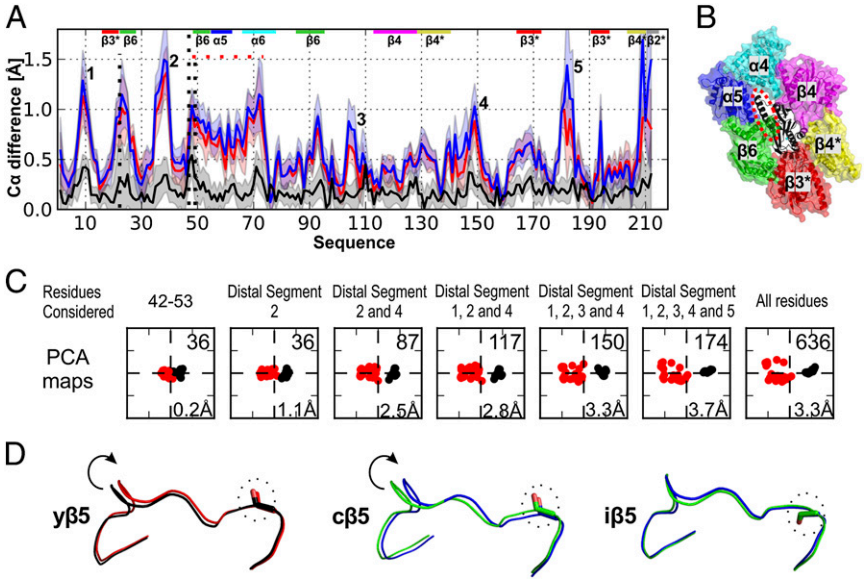



**Fig. 5.** Intersubunit propagation of ligand-induced β5 structural changes. (*A*) Cα difference of β5 between the apo structure (1RYP) and the average nonpeptidic complexes (black); average peptidic complexes with β5 specific ligands (blue), and average peptidic complexes with ligands also in other active subunits (red). Dashed lines and shaded regions presented as in Fig. 4. The colored bars on top highlight contact regions with neighboring subunits. The bold numbers indicate five segments, distal from Thr1, where high variations are observed. (*B*) Subunits surrounding β5 (black structure) with H1 α-helix highlighted (red dashed ellipsoid). (*C*) PCA maps considering the contribution of the main chain atoms of the indicated residues. The five distal segments consist of residues 5–14, 30–41, 101–111, 138–154, and 179–186, respectively. In each map, numbers on the right upper and lower corners indicate the number of atoms involved in the analysis and the distance in the PC1 axis between the closest elements of each cluster, respectively. Colored coding is according to Fig. 1. (*D*) The backbone displacement of segment 4 is shown for yβ5, mβ5c, and mβ5i. The structures are colored as in Fig. 2. The arrows highlight the displacement occurring in yβ5 and mβ5c on peptide binding. The dashed circle highlights the backbone of residues S141, R142, and S142 of yβ5, mβ5c, and mβ5i, respectively.

The results from the HslV protease (33, 36, 37), thermoplasma acidophilum 20S (30, 31), murine 20S (14), and the here reported murine and yeast 20S analysis suggest that the displacement of the α-helix H1 is a common allosteric trigger linking the status of the active site with the entry gate into the proteasome with diverse functional consequences (28, 29).

This question led us to analyze available structural data on the yCP-open gate mutant where the N-terminal segment of the α3 subunit, which is central in the entangled structure of the entry port, had been deleted (38). This mutation causes structural disorder of the entry pore, which seals the particle lumen in the WT species, thus opening an axial channel into the proteolytically active inner chamber and displaying strongly enhanced peptidase activity. For that same reason, a comparison with the proteasome component of the complexes with the 11S activator and Blm10, respectively, which also display an open entry port, was added (39, 40). Both molecular structures are unliganded and lie in the apo-cluster region of the PCA, failing to provide structural evidence for a reciprocal signal from the gate to the active site in the sense of the described structural alterations in β5. Other signals that escaped our analysis may exist, and we are aware that possible conformational shifts can be suppressed by constraints of the crystal lattice.

The discovery of two clusters of CP conformers and their specific structural differences related to peptidic ligand binding in the yeast system cannot, to our knowledge, offer a functional correlation in the sense of enhanced affinity or binding rates of nonpeptidic ligands that do not induce/require structural changes in β5, because of the lack of strictly comparable pairs of ligands. However, recent work (41) showed substantial discrimination between cCP and iCP and specificity for cCP by a nonpeptidic inhibitor in accordance with the structural features described here, but the contribution of different side chain interactions cannot be singled out.

The mammalian system, however, offers a conspicuous correlation between structural and functional data. The binding of a peptidic ligand causes domain closure and movements of about 1 Å in yβ5 and similarly in mβ5c (Fig. 2). In contrast, mβ5i is in a preformed configuration optimal for peptidic ligand (and presumably substrate) binding in the apo state. The domains are closed and geared up for peptide binding without requiring domain motion. These observations suggest that the formation of the antiparallel β-sheet when the ligand peptide aligns with protein segments containing residues T21, G47, and A49 is the principal driving force for the overall change of the β5 backbone, which is not restricted to neighboring protein segments, but affects the entire subunit in various ways. The energies associated with these structural transitions of β5 and their propagation to the adjacent subunits β4 and β6 are difficult to evaluate. It is obvious, however, that they can contribute to the activation energy of ligand binding and may be major factors for the observed enhanced activity of the immune proteasome mentioned earlier.

The presented study is relevant for pharmacology and specifically for advanced design of ligands that discriminate between cCP and iCP. There is experimental evidence for reduced toxicity of specific immune proteasome ligands and added benefits in the therapy of autoimmune disorders. We suggest that inhibitors displaying the characteristic main chain bonding scheme of peptidic ligands have a genuine preference for the iCP. Certainly, side chain and subsite interactions also have a major impact in binding affinity and selectivity. Moreover, our findings suggest that assisting or impeding the backbone shift of the distal segments may have significant effects on affinity and kinetics of peptidic inhibitors. The design of external binding inhibitors distant from the active site therefore might emerge as an option. In either case, molecular dynamics simulations, as shown here, can be used to guide design and experiments.

## Methods

**Structural Classification of Reported PDB Structures.** Currently the RCSB protein databank (42) lists 46 X-ray crystal structures of yCP together with four

murine and one bovine CPs. This ensemble, including the here reported structure in complex with the Boc-$(Ala)_3$-al inhibitor, constitutes the currently available high-resolution structural information (Tables S1 and S4). The structures belong to two disjoint classes: structures with the active site occupied by peptidic inhibitors and structures in the apo-state or in complex with nonpeptidic inhibitors. We define a CP ligand as peptidic if it aligns antiparallel in between protein segments and establishes the corresponding hydrogen bonds with Thr21N, Thr21O, Gly47O, and Ala49N (Fig. 2*A*), regardless of a covalent bond with –Thr1.

**PCA.** PCA is a statistical tool that has been successfully applied to identify protein domain motion (17–19). Given an ensemble of structures, the algorithm consists of generating an average structure and using it as a common reference for calculating a covariance matrix

$$c_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle. \quad [1]$$

The matrix elements $c_{ij}$ contain averaged information on the correlated deviations of the atomic coordinates, $x_i$ and $x_j$, from the corresponding ensemble average, $\langle x_i \rangle$ and $\langle x_j \rangle$. Diagonalization of $C$ provides a space transformation to represent these correlations in a set of orthogonal vectors, thereby dissecting the global displacements into independent components, and is performed by solving the eigenvalue problem

$$A^t C A = \lambda. \quad [2]$$

The matrix $A$ represents the eigenvectors and $\lambda$ represents the associated eigenvalues. The magnitude of the eigenvalues reflects the magnitude of the displacements described by its associated eigenvectors. We performed a PCA analysis on the ensemble of β5 yeast crystal structures (Table S1 and *SI Methods*) using the GROMACS (43) subroutines g_covar and g_anaeig.

**MD Simulations.** The MD simulations were performed using the GROMACS (43) (v. 4.6.2) molecular simulation package (*SI Methods*).

**ACKNOWLEDGMENTS.** We thank R. Feicht for large-scale purification of yeast 20S proteasomes. M.A. is a recipient of a fellowship of the Peter und Traudl Engelhorn-Stiftung. This work was supported by Deutsche Forschungsgemeinschaft Grant LA 1817/3-1 (to O.F.L.).

1. Hershko A, Ciechanover A (1998) The ubiquitin system. *Annu Rev Biochem* 67(1): 425–479.
2. Gallastegui N, Groll M (2010) The 26S proteasome: Assembly and function of a destructive machine. *Trends Biochem Sci* 35(11):634–642.
3. Groll M, et al. (1997) Structure of 20S proteasome from yeast at 2.4 A resolution. *Nature* 386(6624):463–471.
4. Borissenko L, Groll M (2007) 20S proteasome and its inhibitors: Crystallographic knowledge for drug development. *Chem Rev* 107(3):687–717.
5. Groll M, Huber R, Moroder L (2009) The persisting challenge of selective and specific proteasome inhibition. *J Pept Sci* 15(2):58–66.
6. Orlowski M, Cardozo C, Michaud C (1993) Evidence for the presence of five distinct proteolytic components in the pituitary multicatalytic proteinase complex. Properties of two components cleaving bonds on the carboxyl side of branched chain and small neutral amino acids. *Biochemistry* 32(6):1563–1572.
7. Kloetzel P-M, Ossendorp F (2004) Proteasome and peptidase function in MHC-class-I-mediated antigen presentation. *Curr Opin Immunol* 16(1):76–81.
8. Deol P, Zaiss DMW, Monaco JJ, Sijts AJAM (2007) Rates of processing determine the immunogenicity of immunoproteasome-generated epitopes. *J Immunol* 178(12): 7557–7562.
9. Krüger E, Kloetzel P-M (2012) Immunoproteasomes at the interface of innate and adaptive immune responses: Two faces of one enzyme. *Curr Opin Immunol* 24(1): 77–83.
10. Seifert U, et al. (2010) Immunoproteasomes preserve protein homeostasis upon interferon-induced oxidative stress. *Cell* 142(4):613–624.
11. Sijts AJAM, et al. (2000) Efficient generation of a hepatitis B virus cytotoxic T lymphocyte epitope requires the structural features of immunoproteasomes. *J Exp Med* 191(3):503–514.
12. Strehl B, et al. (2008) Antitopes define preferential proteasomal cleavage site usage. *J Biol Chem* 283(26):17891–17897.
13. Voigt A, et al. (2010) Generation of in silico predicted coxsackievirus B3-derived MHC class I epitopes by proteasomes. *Amino Acids* 39(1):243–255.
14. Huber EM, et al. (2012) Immuno- and constitutive proteasome crystal structures reveal differences in substrate and inhibitor specificity. *Cell* 148(4):727–738.
15. Gaczynska M, Rock KL, Goldberg AL (1993) Gamma-interferon and expression of MHC genes regulate peptide hydrolysis by proteasomes. *Nature* 365(6443):264–267.
16. Chen P, Hochstrasser M (1996) Autocatalytic subunit processing couples active site formation in the 20S proteasome to completion of assembly. *Cell* 86(6):961–972.
17. Kitao A, Hirata F, Gō N (1991) The effects of solvent on the conformation and the collective motions of protein: Normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. *Chem Phys* 158(2-3):447–472.
18. García AE (1992) Large-amplitude nonlinear motions in proteins. *Phys Rev Lett* 68(17): 2696–2699.
19. Amadei A, Linssen ABM, Berendsen HJC (1993) Essential dynamics of proteins. *Proteins* 17(4):412–425.
20. Unno M, et al. (2002) The structure of the mammalian 20S proteasome at 2.75 A resolution. *Structure* 10(5):609–618.
21. Hu G, et al. (2006) Structure of the Mycobacterium tuberculosis proteasome and mechanism of inhibition by a peptidyl boronate. *Mol Microbiol* 59(5):1417–1428.
22. Skjaerven L, Martinez A, Reuter N (2011) Principal component and normal mode analysis of proteins; a quantitative comparison using the GroEL subunit. *Proteins* 79(1):232–243.
23. David C, Jacobs D (2014) *Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins. Protein Dynamics, Methods in Molecular Biology*, ed Livesay DR (Humana Press, New York), Vol 1084, pp 193–226.
24. Kisselev AF, Akopian TN, Castillo V, Goldberg AL (1999) Proteasome active sites allosterically regulate each other, suggesting a cyclical bite-chew mechanism for protein breakdown. *Mol Cell* 4(3):395–402.
25. Stein RL, Melandri F, Dick L (1996) Kinetic characterization of the chymotryptic activity of the 20S proteasome. *Biochemistry* 35(13):3899–3908.
26. Myung J, Kim KB, Lindsten K, Dantuma NP, Crews CM (2001) Lack of proteasome active site allostery as revealed by subunit-specific inhibitors. *Mol Cell* 7(2):411–420.
27. Schmidtke G, Emch S, Groettrup M, Holzhütter H-G (2000) Evidence for the existence of a non-catalytic modifier site of peptide hydrolysis by the 20 S proteasome. *J Biol Chem* 275(29):22056–22063.
28. Osmulski PA, Hochstrasser M, Gaczynska M (2009) A tetrahedral transition state at the active sites of the 20S proteasome is coupled to opening of the α-ring channel. *Structure* 17(8):1137–1147.
29. Kleijnen MF, et al. (2007) Stability of the proteasome can be regulated allosterically through engagement of its proteolytic active sites. *Nat Struct Mol Biol* 14(12): 1180–1188.
30. Ruschak AM, Kay LE (2012) Proteasome allostery as a population shift between interchanging conformers. *Proc Natl Acad Sci USA* 109(50):E3454–E3462.
31. Sprangers R, Kay LE (2007) Quantitative dynamics and binding studies of the 20S proteasome by NMR. *Nature* 445(7128):618–622.
32. Kuckelkorn U, et al. (1995) Incorporation of major histocompatibility complex–encoded subunits LMP2 and LMP7 changes the quality of the 20S proteasome polypeptide processing products independent of interferon-γ. *Eur J Immunol* 25(9): 2605–2611.
33. Shi L, Kay LE (2014) Tracing an allosteric pathway regulating the activity of the HslV protease. *Proc Natl Acad Sci USA* 111(6):2140–2145.
34. Bochtler M, Ditzel L, Groll M, Huber R (1997) Crystal structure of heat shock locus V (HslV) from Escherichia coli. *Proc Natl Acad Sci USA* 94(12):6070–6074.
35. Bochtler M, et al. (2000) The structures of HsIU and the ATP-dependent protease HsIU-HsIV. *Nature* 403(6771):800–805.
36. Sousa MC, et al. (2000) Crystal and solution structures of an HslUV protease-chaperone complex. *Cell* 103(4):633–643.
37. Kwon A-R, Kessler BM, Overkleeft HS, McKay DB (2003) Structure and reactivity of an asymmetric complex between HslV and I-domain deleted HslU, a prokaryotic homolog of the eukaryotic proteasome. *J Mol Biol* 330(2):185–195.
38. Groll M, et al. (2000) A gated channel into the proteasome core particle. *Nat Struct Biol* 7(11):1062–1067.
39. Whitby FG, et al. (2000) Structural basis for the activation of 20S proteasomes by 11S regulators. *Nature* 408(6808):115–120.
40. Sadre-Bazzaz K, Whitby FG, Robinson H, Formosa T, Hill CP (2010) Structure of a Blm10 complex reveals common mechanisms for proteasome binding and gate opening. *Mol Cell* 37(5):728–735.
41. Kazi A, et al. (2014) Discovery of PI-1840, a novel non-covalent and rapidly reversible proteasome inhibitor with anti-tumor activity. *J Biol Chem* 289(17):11906–15.
42. Berman HM, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28(1):235–242.
43. Pronk S, et al. (2013) GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29(7):845–854.

# Supporting Information

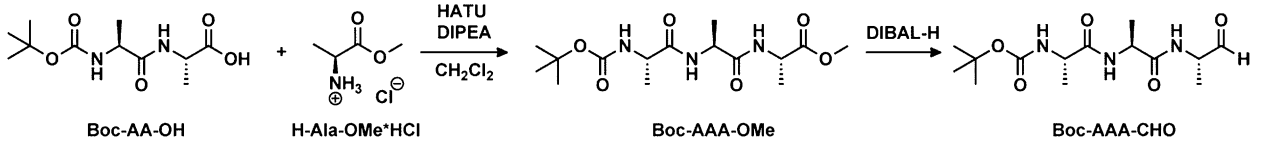
Arciniega et al. 10.1073/pnas.1408018111

## SI Methods

**Principal Component Analysis.** In the yeast system, we computed the ensemble average of the main chain atoms of β5 after positional and orientational superimposition of all used crystal structures, which are mostly derived from isomorphous crystal forms (Table S2).

The principal component analysis (PCA) covariance was computed considering the main chain coordinates (Cα, C, N) from the 46 yeast proteasome core particle (yCP) structures available in the Protein Data Bank (PDB) database (Table S1). As the last 10 residues in mammalian β5/β5i differ considerably from yeast, the covariance matrix was calculated considering residues 1–106 and 108–192. Residue Lys107 is an insertion in yeast and was also omitted. The PCA analysis was computed using the GROMACS (1) subroutines g_covar and g_anaeig.

**Molecular Dynamics Simulations.** The starting coordinates of the modeled systems, mcCP and miCP in their apo and complexed forms, were taken from the respective refined crystallographic structures (Table S1). The model consisted of two parts: (*i*) the active part, which comprises the complete β5 and β6 subunits; and (*ii*) the segments surrounding the active part (Table S3), held in place by positional restraints applied to their Cα atoms.




All segments, with the exception of β7-T1, β3-S1, and β3-M204, were capped with acetyl and methylamine groups at the N terminus and C terminus, respectively. This truncated model of the CP allows a reduction of the computational time while maintaining the molecular surrounding of each active site as defined by the crystal structures. The peptidic ligand used in the simulations was a noncovalent Ala-Ala-Ala-$NCH_3$ moiety. Its starting backbone coordinates were taken from the backbone of the crystallographic PR-957 complex, thus maintaining the hydrogen bonds. During the simulations, positional restraints were applied to the Ala Cαs and to the methyl carbon of the capping group. The same setup and equilibration protocol were applied to all of the modeled systems. Production runs consisted of 20 ns performed in the isothermal-isobaric (NTP) ensemble and were carried out twice (Fig. 3 and Fig. S1).

All systems were simulated using the AMBER99SB all-atom force field (2), an explicit water solvent scheme; and the transferable intermolecular potential with 3 sites water molecule model (3), with 0.15 NaCl concentration and a small surplus of ions for electric charge neutralization and periodic boundary conditions. The long-range electrostatic interactions were computed using fast particle-mesh ewald (PME) (4), using a grid spacing of 1.2 Å. The van der Waals and short-range electrostatic interactions were computed using cutoffs of 14 and 11 Å, respectively. The neighbor list was updated every five steps. All bonds were constrained using LINCS (5), enabling a simulation time step of 0.002 ps.

The following startup protocol was applied to all modeled systems. First, two consecutive minimization processes with (*i*) 1,000 steps under the steepest descendant algorithm and (*ii*) 200 steps using conjugate gradient algorithm. Subsequently, to equilibrate the system at 300 K, we performed 100 ps in the canonical ensemble at 300 K using a V-rescale thermostat ($\tau = 0.1$ ps) (6), followed by 100 ps in the NTP ensemble using Parrinello-Rahman (7) pressure coupling at 1 bar ($\tau = 0.1$ ps). Finally, 20 ns of molecular dynamics (MD) simulation was performed in the NTP ensemble, maintaining the conditions from the last equilibration process.

**Computation of Average Structures.** For each CP subunit, the coordinates were taken from the structures listed in Table S1 and aligned with the yeast apo CP structure (1RYP) as reference. From these aligned coordinates, the averages were computed for apo- and pep-clusters of structures. To compare B-factors among different structures, the B-factors of the main chain atoms for each polypeptide chain were rescaled by Z-score (Eq. **S1**), a statistical standardization method that allows a comparison between B-factors from different structures (8, 9)

$$\text{Z-score}(\text{residue of residue } i) = (\text{B-factor}_i - \mu)/\sigma, \qquad \textbf{[S1]}$$

where B-factor *i* is the B-factor of the *i*th residue. The average μ and SD σ are computed from the B-factors of all N, Cα, and C main chain atoms of the polypeptide chain.

**Synthesis of Boc(Ala)$_3$-CHO.**

***Boc-AAA-OMe.*** Boc-AA-OH (513 mg, 1.970 mmol), H-Ala-OMe*HCl (250 mg, 1.791 mmol), and *O*-(7-azabenzotriazole-1-yl)-*N,N,N'N'*-tetramethyluronium hexafluorophosphate (HATU) (817 mg, 2.149 mmol) were dissolved in $CH_2Cl_2$ (20 mL) and cooled to 0 °C before addition of *N,N*-diisopropylethylamine (DIPEA) (1.251 mL, 7.16 mmol). The cooling bath was removed, and the reaction was allowed to warm to room temperature for 2 h. The reaction mixture was washed with water (2 × 20 mL) and brine (20 mL). The organic fractions were collected, dried over $MgSO_4$, filtered, and concentrated to give a colorless solid. Flash column chromatography (100–0% petrol ether in ethyl acetate) gave Boc-AAA-OMe (600 mg, 1.737 mmol, 97% yield) as a colorless powder. $^1$H NMR (360 MHz, DMSO) δ = 4.36–4.17 (m, 2H), 3.14 (m, 1H), 2.69 (s, 3H), 1.37 (s, 9H), 1.30–1.23 (m, 9H). Electrospray ionization with tandem mass spectrometry (ESI-MS) calculated for $C_{15}H_{28}N_3O_6$ [M+H], 346.20; observed, 345.75.

***Boc-AAA-CHO.*** To a solution of Boc-AAA-OMe (100 mg, 0.290 mmol) in $CH_2Cl_2$ (dry) (3.5 mL) at −78 °C was added diisobutylaluminium hydride (DIBAL-H, 1 M in hexane, 1.158 mL, 1.158 mmol). The reaction mixture was stirred at −78 °C for 90 min before it was quenched by addition of saturated $NH_4Cl$ (0.35 mL) and saturated potassium sodium tartrate (4 mL). The mixture was stirred at room temperature for another 1 h and extracted with $CH_2Cl_2$ (3 × 10 mL). The combined organic phases were washed with $H_2O$ (15 mL) and brine (15 mL), dried over $Na_2SO_4$, filtered, and concentrated in vacuo. The obtained colorless powder of **Boc-AAA-CHO** was used without further

purification. ESI-MS calculated for $C_{14}H_{26}N_3O_5$ [M+H], 316.19; observed, 315.83.

**Crystallization and Structure Determination.** Crystals of the 20S proteasome from *Saccharomyces cerevisiae* were grown using hanging drop method at 24 °C as previously described (10, 11). Crystals were incubated for at least 72 h with Boc-Ala-Ala-Ala-al (concentration: 50 mM in DMSO). The protein concentration used for crystallization was 40 mg/mL in Tris·HCl (10 mM, pH 7.5) and EDTA (1 mM). The drops were composed of 3 μL of protein and 2 μL of the reservoir solution, containing 30 mM of magnesium acetate ($MgAc_2$), 100 mM of morpholino-ethane-sulphonic acid (Mes) (pH 6.8), and 10% (vol/vol) of 2-methyl-2,4-pentandiol (MPD). Crystals were soaked in a cryoprotecting buffer (30% MPD, 20 mM $MgAc_2$, 100 mM Mes, pH 6.8) and flash-cooled in a stream of liquid nitrogen gas at 100 K (Oxford Cryo Systems) for data collection. The space group of the complex belongs to $P2_1$ with cell dimensions of about $a$ = 135 Å, $b$ = 300 Å, $c$ = 145 Å and β =113° (Table S4). Datasets were collected using synchrotron radiation with λ = 1.0 Å at the X06SA-beamline in SLS (Villingen, Switzerland). X-ray intensities and data reduction were evaluated using the XDS program package (12). Conventional crystallographic rigid body, positional, and temperature factor refinements were carried out with REFMAC5 (13) using coordinates of the yeast 20S proteasome structure as a starting model (PDB ID code: 1RYP) (14). For model building, the programs SYBYL and MAIN (15) were used. The completed complex structure yielded excellent R factors, as well as root mean square deviation bond and angle values. Coordinates were confirmed to fulfill the Ramachandran plot (Table S1).

**ACKNOWLEDGMENTS.** M.G. and P.B. thank the staff of the beamline X06SA at the Paul Scherrer Institute, Swiss Light Source (Villigen, Switzerland), for assistance during data collection and the Deutsche Forschungsgemeinschaft SFB1035 for financial support.

1. Pronk S, et al. (2013) GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29(7):845–854.
2. Hornak V, et al. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* 65(3):712–725.
3. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79(2):926–935.
4. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J Chem Phys* 98(12):10089–10092.
5. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) LINCS: A linear constraint solver for molecular simulations. *J Comput Chem* 18(12):1463–1472.
6. Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. *J Chem Phys* 126(1):014101.
7. Parrinello M, Rahman A (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys* 52(12):7182–7190.
8. Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G (2003) Improved amino acid flexibility parameters. *Protein Sci* 12(5):1060–1072.
9. Pontius J, Richelle J, Wodak SJ (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J Mol Biol* 264(1):121–136.
10. Gallastegui N, Groll M (2012) Analysing properties of proteasome inhibitors using kinetic and X-ray crystallographic studies. *Methods Mol Biol* 832:373–390.
11. Groll M, Huber R (2005) Purification, crystallization, and X-ray analysis of the yeast 20S proteasome. *Methods Enzymol* 398:329–336.
12. Kabsch W (1993) Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J Appl Cryst* 26:795–800.
13. Vagin AA, et al. (2004) REFMAC5 dictionary: Organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr D Biol Crystallogr* 60(Pt 12 Pt 1): 2184–2195.
14. Groll M, et al. (1997) Structure of 20S proteasome from yeast at 2.4 A resolution. *Nature* 386(6624):463–471.
15. Turk D (1992) Improvement of a programme for molecular graphics and manipulation of electron densities and its application for protein structure determination. PhD thesis (Technische Universität München, Munich).

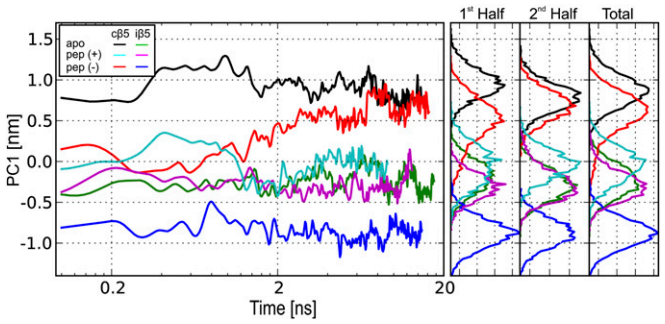



**Fig. S1.** Molecular dynamics simulation. Second independent simulation ran under identical starting conditions as in Fig. 3, but with different initial velocities.

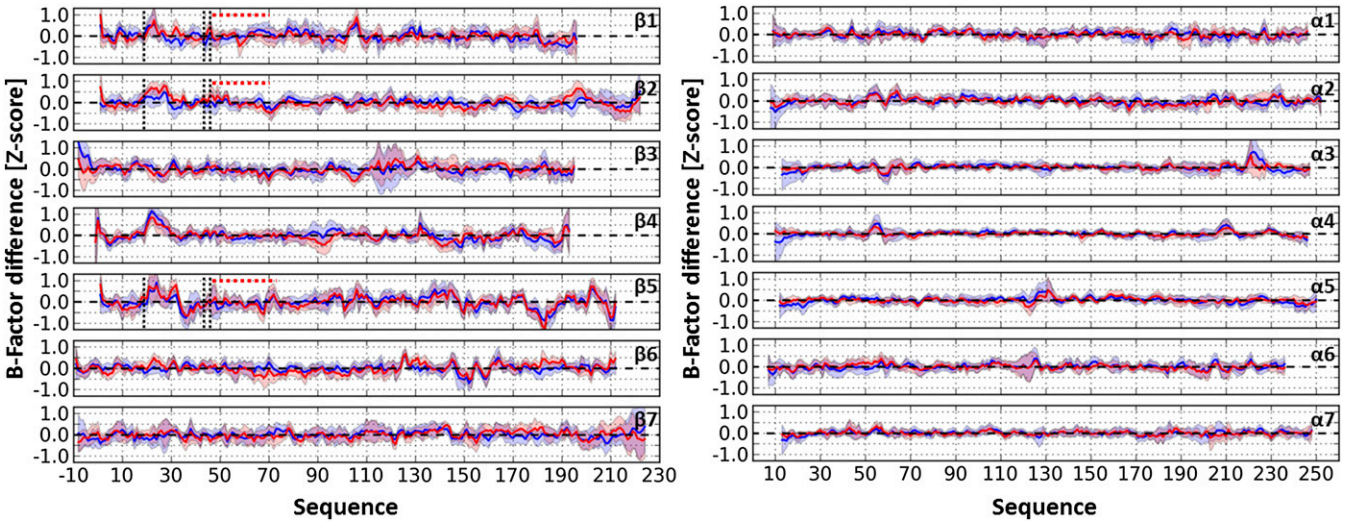



**Fig. S2.** B-factor differences all 14 subunits of yCP. B-factor differences between average apo-structure and average complex structures with a peptidic ligand in β5 (blue) or with peptidic ligands also in other active subunits (red). The associated error is computed from the SD of the averaged Z-scores (shaded regions). Dashed lines highlight residues involved in the peptidic binding (black) and forming the H1 α-helix (red), respectively.

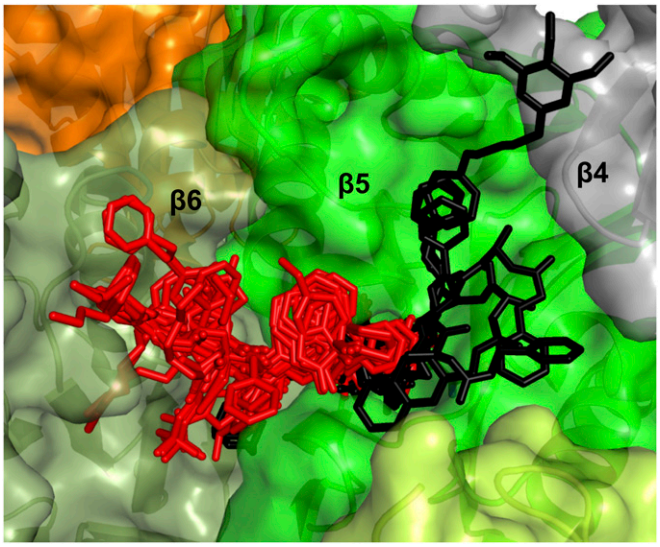



**Fig. S3.** Superposition of ligands of the β5 subunit. Peptidic and nonpeptidic inhibitors are shown in red and black sticks, respectively.

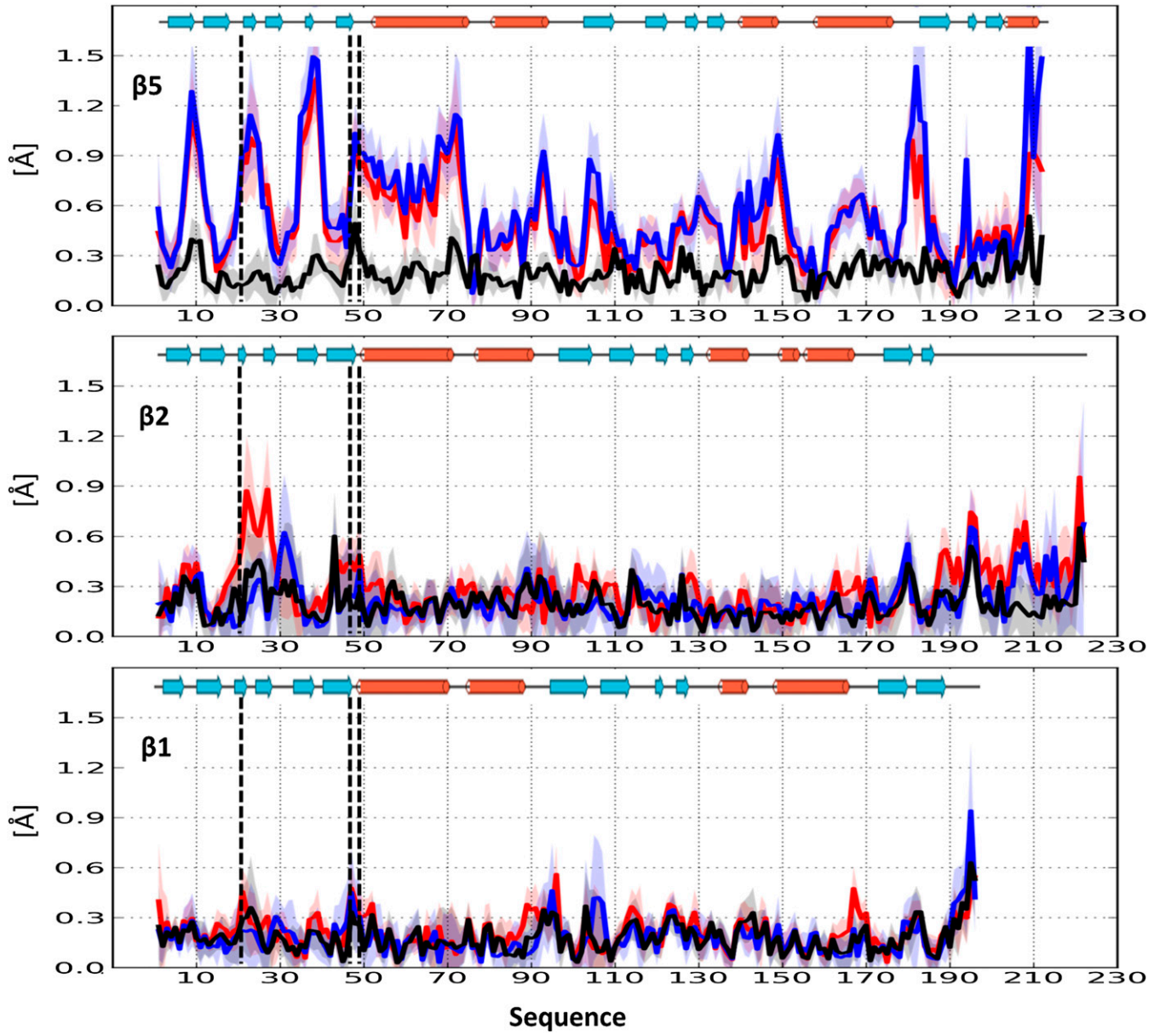



**Fig. S4.** Cα difference against apo structures of the active subunits. Cα difference between the apo structure (1G0U) and the average nonpeptidic complexes (black); average peptidic complexes with β5 specific ligands (blue), and average peptidic complexes with ligands in other active subunits (red). Dashed lines highlight residues involved in the peptidic binding (black). On top of each plot, the secondary structure scheme is marked. The associated error is computed from the SD of the averaged coordinates (shaded regions).

**Table S1. Summary of used structures**

| PDB ID | Molecule | β5 | β2 | β1 | Covalent | PC1 | PC2 |
|---|---|---|---|---|---|---|---|
| Peptidic | | | | | | | |
| 1G65 (1) | Epoxomicin | + | – | – | Yes | –0.73 | 0.14 |
| 3MG4 (2) | LXT[a] | + | – | – | No | –0.14 | 0.18 |
| 3MG6 (2) | LYT[a] | + | – | – | No | –0.80 | 0.25 |
| 3MG7 (2) | L2T[a] | + | – | – | No | –0.73 | 0.30 |
| 3MG8 (2) | L3T[a] | + | – | – | No | –0.83 | 0.27 |
| 3NZJ (3) | TMC-95A mimic 2a | + | – | – | No | –0.23 | –0.09 |
| 3NZW (3) | TMC-95A mimic 2b | + | – | – | No | –0.76 | –0.02 |
| 3NZX (3) | TMC-95A mimic 2c | + | – | – | No | –0.56 | –0.14 |
| 3OEU (4) | PRD_000944 | + | – | – | No | –0.77 | 0.24 |
| 3OEV (4) | PRD_000959 | + | – | – | No | –0.38 | 0.31 |
| 3OKJ (5) | PRD_001051 | + | – | – | Yes | –0.78 | –0.03 |
| 3SDI (4) | PRD_001071 | + | – | – | No | –0.79 | 0.24 |
| 3SDK (4) | PRD_001075 | + | – | – | No | –0.81 | 0.22 |
| 4JSQ (6) | TMC-95A mimic dimer 4e | + | – | – | No | –0.34 | –0.24 |
| 4JT0 (6) | TMC-95A mimic dimer 4a | + | – | – | No | –0.07 | –0.15 |
| 4QBY (This study) | Boc-$(Ala)_3$-al | + | – | + | Yes | –0.03 | –0.06 |
| 3BDM (7) | Glidobactin A | + | + | – | Yes | –0.09 | –0.21 |
| 3GPJ (8) | Syringolin B | + | + | – | Yes | –0.23 | –0.20 |
| 4INR (9) | Vinyl Sulfone Lu102 | + | + | – | Yes | –0.28 | –0.19 |
| 4INT (9) | Vinyl Sulfone LU122 | + | + | – | Yes | –0.38 | –0.20 |
| 4INU (9) | Vinyl Sulfone LU112 | + | + | – | Yes | –0.26 | –0.25 |
| 4JSU (6) | TMC-95A mimic dimer 3a | + | + | – | No | –0.35 | –0.24 |
| 1JD2 (10) | TMC-95A | + | + | + | No | 0.02 | –0.12 |
| 2ZCY (7) | Syringolin A | + | + | + | Yes | –0.45 | –0.04 |
| 2F16 (11) | Bortezomib | + | + | + | Yes | –0.29 | –0.09 |
| 3D29 (12) | Fellutamide B | + | + | + | Yes | –0.82 | –0.12 |
| 3MG0 (2) | Bortezomib | + | + | + | Yes | –0.28 | –0.02 |
| 3UN4 (13) | PR-957 Morpholine | + | + | + | Yes | –0.33 | –0.22 |
| 4GK7 (14) | Syringolin-Glidobactin | + | + | + | Yes | –0.10 | –0.18 |
| Apo/Nonpeptidic | | | | | | | |
| 2FAK (15) | Salinosporamide A | + | + | + | Yes | 0.65 | 0.03 |
| 3GPT (16) | GPT[a] | + | + | + | Yes | 0.65 | 0.05 |
| GPW (16) | Salinosporamide | + | + | + | Yes | 0.61 | 0.04 |
| 4EU2 (17) | k-7174 | + | + | + | No | 0.94 | 0.18 |
| 1G0U (18) | Open gate apo mutant | – | – | – | No | 0.87 | 0.15 |
| 1RYP (19) | Apo structrure | – | – | – | No | 0.85 | 0.10 |
| 1VSY (20) | Open gate Blm10 complex | – | – | – | No | 0.95 | 0.19 |
| 1Z7Q (21) | Open gate 11S complex | – | – | – | No | 1.08 | 0.26 |
| 3L5Q (20) | Open gate Blm10 complex | – | – | – | No | 0.95 | 0.19 |
| 2GPL (22) | BIQ* | – | + | – | No | 0.71 | 0.03 |
| 3DY3 (23) | Spirolactacystin D | + | – | – | Yes | 0.75 | –0.04 |
| 3DY4 (23) | Spirolactacystin S | + | + | – | Yes | 0.76 | –0.05 |
| 3E47 (24) | Homobelactosin C | + | – | – | Yes | 0.84 | 0.12 |
| 3SHJ (25) | Hydroxyurea (HU10) | + | – | – | No | 0.70 | –0.02 |
| 3TDD (26) | Belactosin C | + | – | – | Yes | 0.86 | 0.08 |
| 3UN8 (13) | PR-957 Epoxy | + | – | – | Yes | 0.58 | –0.05 |
| 4LQI (27) | Vibralactone | + | – | + | Yes | 0.84 | 0.09 |
| 4J70 (28) | Belactosin derivate 3e | – | + | – | Yes | 0.90 | 0.14 |
| Mammalian | | | | | | | |
| 1IRU (29) | Apo Bovine | – | – | – | No | 0.38 | 0.12 |
| 3UNB (13) | PR-957 mouse | + | + | + | Yes | –0.86 | 0.10 |
| 3UNE (13) | Apo Mouse | – | – | – | No | 0.39 | 0.11 |
| 3UNF (4) | PR-957 mouse immuno | + | + | + | Yes | –0.79 | 0.21 |
| 3UNH (13) | Apo Mouse immuno | – | – | – | No | –0.49 | 0.27 |

Column 1 corresponds to the classification of the 20S structures as peptidic, apo/nonpeptidic, and mammalian. Columns 2 and 3 correspond to PDB ID of the 20S structure and ligand's name, respectively. Columns 4–6 indicate the presence (+) or absence (–) of the ligand at the active site of subunits β5, β2, and β1, respectively. Column 7 indicate the whether the ligand binds covalently to the 20S. Columns 8 and 9 show the values plotted in Fig. 1, where PC1 and PC2 correspond to the projections of the each structure on the first two principal components of the ensemble of yeast β5 structures, respectively.

*Small molecule ID in the PDB.

1. Groll M, Kim KB, Kairies N, Huber R, Crews CM (2000) Crystal structure of epoxomicin:20S proteasome reveals a molecular basis for selectivity of α′,β′-epoxyketone proteasome inhibitors. *J Am Chem Soc* 122(6):1237–1238.
2. Blackburn C, et al. (2010) Characterization of a new series of non-covalent proteasome inhibitors with exquisite potency and selectivity for the 20S β5-subunit. *Biochem J* 430(3): 461–476.
3. Groll M, et al. (2010) 20S proteasome inhibition: Designing noncovalent linear peptide mimics of the natural product TMC-95A. *ChemMedChem* 5(10):1701–1705.
4. Blackburn C, et al. (2010) Optimization of a series of dipeptides with a P3 threonine residue as non-covalent inhibitors of the chymotrypsin-like activity of the human 20S proteasome. *Bioorg Med Chem Lett* 20(22):6581–6586.
5. Gräwert MA, et al. (2011) Elucidation of the α-keto-aldehyde binding mechanism: A lead structure motif for proteasome inhibition. *Angew Chem Int Ed Engl* 50(2):542–544.
6. Desvergne A, et al. (2013) Dimerized linear mimics of a natural cyclopeptide (TMC-95A) are potent noncovalent inhibitors of the eukaryotic 20S proteasome. *J Med Chem* 56(8): 3367–3378.
7. Groll M, et al. (2008) A plant pathogen virulence factor inhibits the eukaryotic proteasome by a novel mechanism. *Nature* 452(7188):755–758.
8. Clerc J, et al. (2009) Synthetic and structural studies on syringolin A and B reveal critical determinants of selectivity and potency of proteasome inhibition. *Proc Natl Acad Sci USA* 106(16):6507–6512.
9. Geurink PP, et al. (2013) Incorporation of non-natural amino acids improves cell permeability and potency of specific inhibitors of proteasome trypsin-like sites. *J Med Chem* 56(3): 1262–1275.
10. Groll M, Koguchi Y, Huber R, Kohno J (2001) Crystal structure of the 20 S proteasome:TMC-95A complex: A non-covalent proteasome inhibitor. *J Mol Biol* 311(3):543–548.
11. Groll M, Berkers CR, Ploegh HL, Ovaa H (2006) Crystal structure of the boronic acid-based proteasome inhibitor bortezomib in complex with the yeast 20S proteasome. *Structure* 14(3): 451–456.
12. Hines J, Groll M, Fahnestock M, Crews CM (2008) Proteasome inhibition by fellutamide B induces nerve growth factor synthesis. *Chem Biol* 15(5):501–512.
13. Huber EM, et al. (2012) Immuno- and constitutive proteasome crystal structures reveal differences in substrate and inhibitor specificity. *Cell* 148(4):727–738.
14. Archer CR, et al. (2012) Activity enhancement of the synthetic syrbactin proteasome inhibitor hybrid and biological evaluation in tumor cells. *Biochemistry* 51(34):6880–6888.
15. Groll M, Huber R, Potts BCM (2006) Crystal structures of Salinosporamide A (NPI-0052) and B (NPI-0047) in complex with the 20S proteasome reveal important consequences of β-lactone ring opening and a mechanism for irreversible binding. *J Am Chem Soc* 128(15):5136–5141.
16. Groll M, McArthur KA, Macherla VR, Manam RR, Potts BC (2009) Snapshots of the fluorosalinosporamide/20S complex offer mechanistic insights for fine tuning proteasome inhibition. *J Med Chem* 52(17):5420–5428.
17. Kikuchi J, et al. (2013) Homopiperazine derivatives as a novel class of proteasome inhibitors with a unique mode of proteasome binding. *PLoS ONE* 8(4):e60649.
18. Groll M, et al. (2000) A gated channel into the proteasome core particle. *Nat Struct Biol* 7(11):1062–1067.
19. Groll M, et al. (1997) Structure of 20S proteasome from yeast at 2.4 A resolution. *Nature* 386(6624):463–471.
20. Sadre-Bazzaz K, Whitby FG, Robinson H, Formosa T, Hill CP (2010) Structure of a Blm10 complex reveals common mechanisms for proteasome binding and gate opening. *Mol Cell* 37(5):728–735.
21. Förster A, Masters EI, Whitby FG, Robinson H, Hill CP (2005) The 1.9 A structure of a proteasome-11S activator complex and implications for proteasome-PAN/PA700 interactions. *Mol Cell* 18(5):589–599.
22. Groll M, Götz M, Kaiser M, Weyher E, Moroder L (2006) TMC-95-based inhibitor design provides evidence for the catalytic versatility of the proteasome. *Chem Biol* 13(6):607–614.
23. Groll M, Balskus EP, Jacobsen EN (2008) Structural analysis of spiro β-lactone proteasome inhibitors. *J Am Chem Soc* 130(45):14981–14983.
24. Groll M, Larionov OV, Huber R, de Meijere A (2006) Inhibitor-binding mode of homobelactosin C to proteasomes: New insights into class I MHC ligand generation. *Proc Natl Acad Sci USA* 103(12):4576–4579.
25. Gallastegui N, et al. (2012) Hydroxyureas as noncovalent proteasome inhibitors. *Angew Chem Int Ed Engl* 51(1):247–249.
26. Korotkov VS, et al. (2011) Synthesis and biological activity of optimized belactosin C congeners. *Org Biomol Chem* 9(22):7791–7798.
27. List A, et al. (2014) Omuralide and vibralactone: Differences in the proteasome- β-lactone-γ-lactam binding scaffold alter target preferences. *Angew Chem Int Ed Engl* 53(2):571–574.
28. Kawamura S, et al. (2013) Potent proteasome inhibitors derived from the unnatural cis-cyclopropane isomer of Belactosin A: Synthesis, biological activity, and mode of action. *J Med Chem* 56(9):3689–3700.
29. Unno M, et al. (2002) The structure of the mammalian 20S proteasome at 2.75 A resolution. *Structure* 10(5):609–618.

**Table S2. Unit cell parameters of the 20S structures shown in Table S1**

| Structure | Space group | Angle β (°) | Length, α (Å) | Length, β (Å) | Length, γ (Å) |
|---|---|---|---|---|---|
| cCP bovine apo | P 21 21 21 | 90.0 | 316.7 | 205.9 | 116.0 |
| cCP mouse PR-957 | P 1 21 1 | 106.6 | 171.7 | 198.6 | 226.8 |
| cCP mouse apo | P 1 21 1 | 108.07 | 171.0 | 201.3 | 226.0 |
| iCP mouse PR-957 | P 1 21 1 | 107.1 | 117.3 | 194.6 | 157.7 |
| iCP mouse apo | P 1 21 1 | 105.7 | 118.3 | 205.2 | 161.9 |
| 11S-yCP complex | P 21 21 21 | 90.0 | 193.0 | 232.1 | 296.8 |
| Blm10-yCP complex | P 21 21 21 | 102.9 | 236.1 | 127.7 | 532.7 |
| Yeast (average ± SD) | P 1 21 1 | 112.9 ± 0.3 | 135.4 ± 1.0 | 300.7 ± 1.0 | 144.6 ± 1.0 |

**Table S3. Residues constituting the model used in the MD simulations**

| Subunit | Active part |
|---|---|
| β5 | Complete subunit |
| β6 | Complete subunit |
| **Subunit** | **Surrounding segments** |
| β4 | R19-F36, G48-Q61 and R95-V100 |
| β7 | 1T-N3, H89-W107 and L122-V159 |
| α4 | S92-R108 |
| α5 | S54-I66, I80-V104 and D139-K141 |
| α6 | K58-I67, K79-D97 |
| β2* | D17-S30, S129-A133, T160-N172 and P190-T213 |
| β3* | S1-G7, K25- Q39, V131-W153, T165-V184 and T198-M204 |
| β4* | K126-T150 and K162-F171 |

*Subunits from the *trans* β ring.

**Table S4. Data collection and refinement statistics**

| Crystallographic data | yCP:Boc-Ala-Ala-Ala-al |
|---|---|
| Crystal parameter | |
| Space group | $P2_1$ |
| Cell dimensions | $a$ = 135 Å, $b$ = 300 Å, $c$ = 145 Å, β = 113° |
| Molecules per AU* | 1 |
| Data collection | |
| Beam line | SLS, PX06SA |
| Wavelength (Å) | 1.0 |
| Resolution range (Å)† | 20–3.0 (3.1–3.0) |
| No. observations‡ | 586,676 |
| No. unique reflections‡ | 194,983 |
| Completeness (%)† | 92.4 (89.7) |
| $R_{merge}$ (%)†,§ | 8.6 (51.4) |
| $I/\sigma$ ($I$)† | 10.4 (2.3) |
| Refinement (CNS) | |
| Resolution range (Å) | 15–3.0 |
| No. refl. working set | 185,233 |
| No. refl. test set | 9,750 |
| No. nonhydrogen | 49,655 |
| Solvent (water, $Mg^{2+}$) | 191 |
| Inhibitor (none-hydrogen) | 88 |
| $R_{work}/R_{free}$ (%)¶ | 17.8/20.7 |
| RMSD bond (Å/°)‖ | 0.005/1.10 |
| Average B-factor (Å$^2$) | 68.7 |
| Ramachandran plot (%)** | 95.7/3.9/0.4 |
| PDB accession code | 4QBY |

*Asymmetric unit.

†The values in parentheses of resolution range, completeness, $R_{merge}$ and $I/\sigma$ ($I$) correspond to the last resolution shell.

‡Friedel pairs were treated as identical reflections.

§$R_{merge}(I) = \Sigma_{hkl}\Sigma_j |[I(hkl)j - I(hkl)]|/\Sigma_{hkl} I$hkl, where $I$(hkl)j is the measurement of the intensity of reflection hkl and $<I(hkl)>$ is the average intensity.

¶$R = \Sigma_{hkl} \mid |F_{obs}| - |F_{calc}| \mid /\Sigma_{hkl} |$Fobs$|$, where $R_{free}$ is calculated without a σ cutoff for a randomly chosen 5% of reflections, which were not used for structure refinement, and $R_{work}$ is calculated for the remaining reflections.

‖Deviations from ideal bond lengths/angles.

**Number of residues in favored region/allowed region/outlier region.

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

# Improvement of Virtual Screening Results by Docking Data Feature Analysis

Marcelino Arciniega*[,†,‡] and Oliver F. Lange[‡,§]

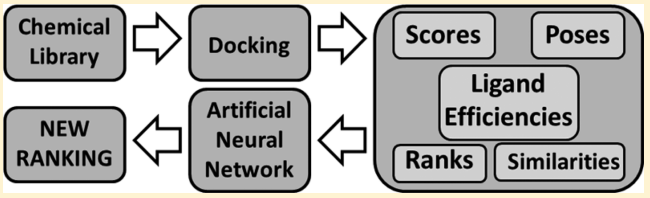
[†]Max Planck Institute Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

[‡]Biomolecular NMR and Munich Center for Integrated Protein Science, Department Chemie, Technische Universität München, 85747 Garching, Germany

[§]Institute of Structural Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany

Supporting Information

**ABSTRACT:** In this study, we propose a novel approach to evaluate virtual screening (VS) experiments based on the analysis of docking output data. This approach, which we refer to as docking data feature analysis (DDFA), consists of two steps. First, a set of features derived from the docking output data is computed and assigned to each molecule in the virtually screened library. Second, an artificial neural network (ANN) analyzes the molecule's docking features and estimates its activity. Given the simple architecture of the ANN, DDFA can be easily adapted to deal with information from several docking programs simultaneously. We tested our approach on the Directory of Useful Decoys (DUD), a well-established and highly accepted VS benchmark. Outstanding results were obtained by DDFA not only in comparison with the conventional rankings of the docking programs used in this work but also with respect to other methods found in the literature. Our approach performs with similar good results as the best available methods, which, however, also require substantially more computing time, economic resources, and/or expert intervention. Taken together, DDFA represents an automatic and highly attractive methodology for VS.




## 1. INTRODUCTION

In the early stages of the drug discovery process, large chemical libraries are screened to identify lead molecules that allow the development of new drugs. The large amount of experimental resources that are required renders this search highly expensive and time-consuming.[1] In this context, in sillico virtual screening (VS) has appeared as a fast and economic approach that increases the efficiency of the lead discovery process.[2−4]

There are two main approaches to perform VS: ligand-based and structure-based.[5] In the former, previously known active ligands are used to identify other molecules with similar characteristics; in the latter, protein and ligand structural models at atomic resolution are used to evaluate their binding affinity. Structure-based methods perform generally better than ligand-based methods in identifying new lead compounds.[6] In structured-based methods, the screening is performed by docking each of the library's molecules into the receptor's active site while optimizing the atomic interactions between the binding partners.

Docking methods are foremost developed to identify the ligand's actual binding mode from the large set of sampled conformations tested.[7] In this regard, docking software like Autodock, AutodockVina, and RosettaLigand have achieved high performances.[8−10] Compared to the discrimination of correct and incorrect conformations of the same ligand, it is far more challenging to discriminate active from inactive ligands, as it is required for structured-based screening. This complication arises from the difficulty to evaluate free energy terms of the unbound ligand state, such as solvent and conformational entropies. Whereas these terms cancel out when comparing conformations of the same ligand, they do not cancel out when treating different molecules. Nevertheless, due to the lack of better ranking methods, it is common practice to rank ligands based only on the docking score of their best docked conformations.

Many attempts have been made to improve the ranking of ligands beyond the accuracy obtained by using the plain docking score.[11−16] For instance, García-Sosa et al.[14] reported that a better correlation between docking scores and experimental binding energies can be achieved by dividing the docking score by the ligand's size; resulting in a descriptor often referred to as ligand efficiency (LE). Others have analyzed several high ranking conformations per ligand, rather than considering only the best scored conformations. For example, Seok et al.[15] augmented the binding energy evaluation by adding an entropy term that was estimated from the populations of clusters of high ranked binding modes. To establish a new ranking, Wallach et al.[16] compared scores of query molecules with that of physically similar (molecular weight, number of rotational bonds, number of hydrogen acceptor/donor, etc.) but chemically dissimilar (different

Received: January 14, 2014
Published: May 5, 2014

   dx.doi.org/10.1021/ci500028u | J. Chem. Inf. Model. 2014, 54, 1401−1411

topology and functional groups) decoys. For each molecule in the screening library, a set of decoys is generated. The approach is based on the assumption that an active ligand should obtain a significantly higher score than the decoys' score distribution. Although the method turned out as a success, its main disadvantage is that the amount of molecules to be docked increases by 2 orders of magnitude.

The common theme of the methods mentioned above is to rank the screening library by a modified form of a single scoring function. An alternative class of approaches attempts to overcome the deficiencies of individual scoring functions by employing two or more docking programs in a consensus scheme.[17] One of the most popular ways to combine multiple ranked lists is the "rank-by-rank" approach.[18] For each molecule the average of its ranks is computed and used to establish the final ranking list. A different method would be to average scores from different scoring functions. It has been shown that both of these methods rely crucially on the diversity and high quality of the scoring functions.[19] A more sophisticated analysis was developed by Jacobsson et al.[20] In their work, data mining techniques were used to create "if−then" rules that yielded upper and lower bounds to seven scoring functions.

Here we describe a novel framework to predict the ligand activity based on a diverse set of docking features rather than focusing on a single kind; such as the docking score. This framework, which we named docking data feature analysis (DDFA), converts this set of docking features into a feature score. The signal conversion is performed by an artificial neural network (ANN) that can be trained to work with data from either single or several docking programs. In our particular case we performed the analysis using three programs and five docking features: (i) best docking score, (ii) ligand efficiency, (iii) scores from similar molecules, (iv) the position of the ligand's poses within the general rank, and (v) structural consistency of the ligand's poses. These features were selected to capture different aspects that are typically employed in a human expert analysis to identify active binding molecules from the VS ranking. Bearing this in mind, the docking score feature represents the traditional approach, which assumes correlation between score and activity. The ligand efficiency feature contributes to a size independent comparison among ligands. Monitoring the performance of chemically similar molecules is inspired by the structural activity relationship (SAR) central idea, which is that similar molecules have similar binding energies. The feature that monitors the ranks of the ligand poses assumes that poses from an active molecule are not distributed randomly through the entire rank. The pose variability feature exploits that active ligands often show better converged poses. It is important to mention that the DDFA can be easily extended and/or adapted to include other features.

To test the DDFA approach, we docked the broadly used Directory of Useful Decoys[21] (DUD) using three different docking programs—Autodock4,[22] Autodockvina,[9] and RosettaLigand[23]—and predicted ligand activity. DUD is widely accepted for benchmarking VS protocols. It consists of 40 receptors of pharmaceutical relevance and a screening library of over 100 000 molecules. To predict the ligand activities for a receptor of DUD benchmark, the DDFA ANN was trained using 22 receptors from the remaining 39 data sets. The 22 receptors of the training set were randomly selected after removing receptors with similar biological activity or with reported positive cross-enrichment, with respect to the receptor to be evaluated. We repeated this process with a different receptor left out of the training set each time to obtain ligand activity predictions for each receptor in DUD. As a control, DDFA's performance was compared to that of the individual docking scores of the used programs and a consensus ranking; with the latter generated by the ligand's best rank in any program. The performance evaluation was carried out using well established and broadly accepted metrics, such as enrichment factor (ef) and the area under the curve (auc) from the receiver operator characteristic (ROC) curve.

## 2. METHODS

**2.1. Docking Programs.** In this study three docking programs Autodock4.2[22] (AD4), Autodockvina1.2[9] (ADV), and RosettaLigand3.4[23] (RL) were used. Although AD4 and ADV were developed by the same lab, they differ in the sampling methods and weights of individual score terms. RL is part of the Rosetta's software suite for modeling macromolecular structures. We used AD4 and ADV with a rigid receptor model and RL with flexible side-chains for the receptor.

*2.1.1. Docking Using AD4 and ADV.* The receptors and ligands were prepared following the standard setup protocols using Gasteiger partial charges.[22] The grid sizes were set up to 27 Å × 27 Å × 27 Å in both programs, using as grid center the center of mass of the ligand provided by the DUD to localize the binding pocket. For AD4, the receptor grid was generated using autogrid4 with 0.375 Å of grid spacing. The docking parameter file was generated with the prepare_dpf42.py script in AutoDockTools.[22] The Lamarckian genetic algorithm with default parameters was selected as pose search method.[8] Ten output poses were requested. For ADV, a maximum of ten output poses was kept using a restriction of 3 kcal/mol in the score difference between the best and worse poses. The global search exhaustiveness parameter was set to 16 (default value 8).

*2.1.2. Docking Using Rosetta Ligand.* For Rosetta Ligand (RL) the receptor side chain conformations were first optimized with the *fixbb* application of Rosetta.[24] The ligands were adapted to the RL format using scripts provided in the Rosetta distribution (molfile_to_params.py).[24] RL searches for docking poses by cycling through a predetermined library of intraligand conformations simultaneously to optimizing the ligand's rigid body degrees of freedom and receptor sidechain dihedral angles. Usually the ligand conformational library is generated with the external program OpenEye's Omega.[25]

In the context of this work, ligand conformations were already available through the AD4 and ADV docking output, and thus, all output poses from AD4 and ADV were used for the ligand conformation library of RL. For every run, the ligand initial placement was provided by the center of mass of a randomly selected member of the conformation library. Docking was performed using the RossetaScripts[26] application with the parameters reported by Davis et al.[27] The number of runs per ligand was set to 50. The top ten structures in interface score were selected for analysis and comparison with the other docking software.

**2.2. RAW Rankings.** The screening library of each DUD receptor was docked using the docking programs AD4, ADV, and RL a ranking based exclusively on the docking score was generated. A fourth ranking (ALL) was created by assigning to each ligand the best achieved position within any of the individual rankings. Tied cases were resolved by comparing the ligand's standardized docking scores of the individual programs. Docking scores were standardized by subtracting the average

and dividing by the standard deviation of the score-distribution. This standardization procedure is commonly known as Z-Score. We refer to this set of scores as RAW (RAW-AD4, RAW-ADV, RAW-RL, and RAW-ALL) since they represent the most straightforward approach to establish a ranking of a docked library.

**2.3. DDFA Rankings.** In the DDFA approach, a feature vector is assigned to each ligand and used as input layer of a feed-forward ANN. The term "docking feature" refers to characteristic information computed from the docking data of the screened library. Details of the docking features used in this work are given in section 2.4. The analysis was performed considering docking data from either a single docking program (DDFA-AD4, DDFA-ADV, and DDFA-RL) or from analyzing all three sources simultaneously (DDFA-ALL). ANN's architecture and training procedure is described in section 2.5.

**2.4. Docking Features.** Docking data is analyzed to derive features that help to discriminate between active and inactive ligands. In this work five features are used in the analysis (DockScore, DockLE, DockSimi, DockPoses, and DockRmsd) and are described in the following sections.

*2.4.1. DockScore.* This feature is given by the best docking score of the ligand poses. It represents the traditional approach, in which the docking score helps to provide enough information to discriminate an active molecule from an inactive one. Prior to analysis the docking scores were standardized as Z-scores.

*2.4.2. DockLE.* The ligand efficiency (LE) was computed as the quotient between the best ligand's score and the number of heavy atoms of the ligand.

*2.4.3. DockSimi.* The DockSimi feature of a ligand is the weighted average of the best docking scores of the five most similar ligands in the docked library. The Tanimoto coefficients (Tc) were used as both similarity measures and weighting factors in the computation of the average. The FP2 molecular fingerprints as implemented in OpenBabel[28] version 2.3.1 were used to compute the Tc. Only ligands with Tc > 0.70 were considered as similar. Whenever no similar ligands existed in the docked library, DockSimi was set to zero.

*2.4.4. DockPoses.* This feature is a five-dimensional vector composed of the number of ligand poses that are within the top 5%, 10%, 15%, 20%, and 25%, respectively, of all pose-scores in the docked library.

*2.4.5. DockRmsd.* This feature is a five-dimensional vector given by the RMSD of the second−sixth ranked poses of a ligand when superimposed to the first ranked pose.

**2.5. Evaluation of Docking Features Using Artificial Neural Networks.** *2.5.1. Architecture of the ANN.* Artificial neural networks (ANNs) are known to perform well on pattern recognition and classification problems.[29] Here we train an ANN to identify active molecules based on the information provided by docking features.

Figure 1 shows a schematic representation of the ANN topology. It consists of 13, 8, and 1 nodes for the input, hidden, and output layers, respectively (Supporting Information Figure S1). The network has full-connectivity among the layers, with linear, sigmoidal, and softmax activation functions for the input, hidden, and output layers, respectively. The ANN was constructed using the PyBrain[30] package. Given the ligand's docking features at the input layer, the returned value at the ANN's output node can be interpreted as a confidence assessment on the ligand's activity chances. Consequently the ligands of the screened library are ranked based on the ANN's output. If DDFA is applied to a single docking program, the docking features give rise to 13 input nodes as follows: one node for features DockScore, DockLE, and DockSimi and five nodes each for the features DockPose and DockRmsd. In the DDFA-ALL, where DDFA is applied to three docking programs simultaneously, the number of input nodes is tripled.

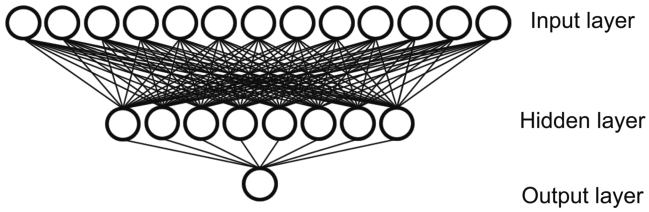



**Figure 1.** Schematic representation of the ANN used with single docking program. The ANN has a feed-forward architecture consisting of three layers with 13 and 8 nodes in the input and hidden layers, respectively, and a single output node. There is full connectivity among the layers using linear, sigmoid, and softmax as activation functions for the input, hidden, and output nodes, respectively. A detailed description of the input values taken by each of the input nodes can be found in the Methods section. In the case where the information from three docking program is used, the input layer nodes are triplicated.

*2.5.2. Training and Application of the ANN.* In order to apply the DDFA approach to any of the DUD receptor's, the docking data was divided into three nonoverlapping sets: training, validating, and testing (Figure 2). The training set is used during parameter estimation; the validation set is used to control hyperparameters and to monitor training progress; the test set is used to measure the performance of the methodology as in the reported results. To test the method we use a leave-one-out approach. Thus, one receptor and it's DUD ligands are used as the testing set and remaining receptors and associated DUD ligands are used for training and validation. However, because similarities between the receptor used for testing and the receptors used for training or validation might cause overestimation of the performance for truly new and unseen cases, we further remove any receptors similar to the test receptor from training and validation sets. As similar we consider receptors in the same biological class (Table 1) and also receptors for which positive cross-enrichment has been reported[21] (Table 1, column 2). Because this would cause varying numbers of receptors in training and validation sets, we further reduce their number to always get a total of 22 receptors, which reflects the smallest number of nonsimilar receptors which would ever occur. This final selection is done randomly. Thus, the analysis on the testing set represents a realistic evaluation of the DDFA performance and similar performance would be expected for unknown receptors and screening ligands.

To generate balanced training and validation sets, all active molecules are taken together with the same amount of randomly selected decoys. From this pool with a balanced active ligand to decoy ratio, 70% was used for training and 30% for validation (Supporting Information Figure S2). The training process was conducted under a back-propagation protocol with a value of 0.001 for all three training parameters: weight decay, learning rate, and momentum (Supporting Information Figure S3). The testing set was used to monitor the ANN performance over the training epochs. Training was terminated when a plateau for the test-set performance had been reached. This plateau occurred after 800 epochs for the AD4, ADV, and the

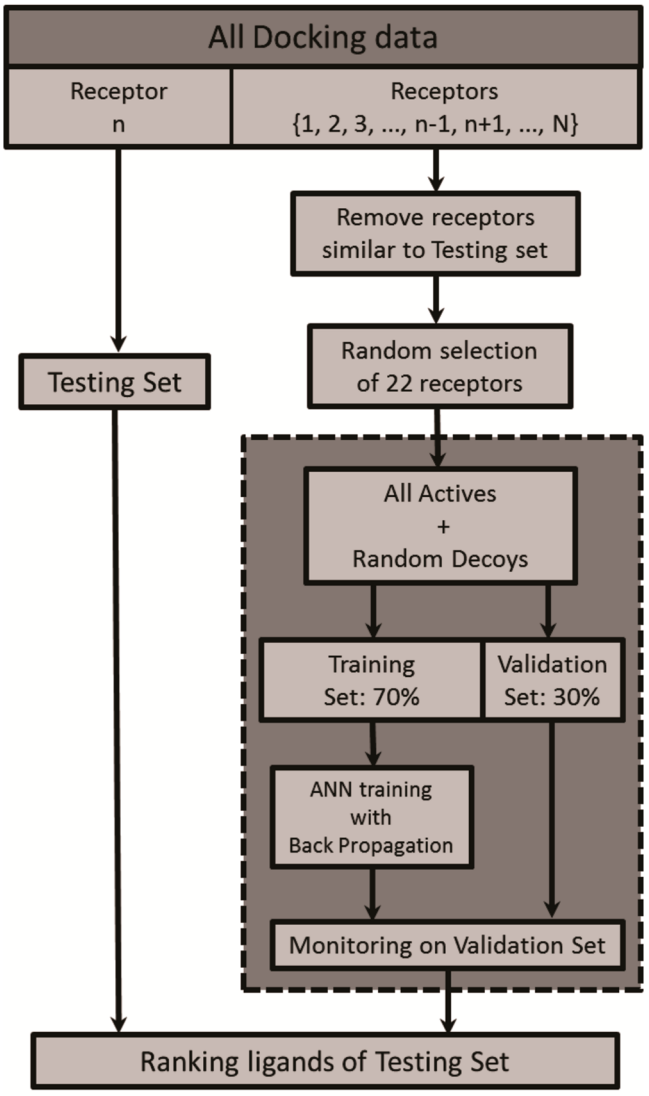



**Figure 2.** Flowchart depicting the cross-validation and application procedures of the DDFA approach. VS data is separated into three nonoverlapping sets: training, validation, and testing. The receptor to be analyzed constitutes the testing set; any of its data is considered during the training and validation processes (enclosed dashed region). The training and validation sets are formed using 22 receptors. None of these receptors can be similar to the receptor used for testing (Table 1). Docking data belonging to all active molecules, together with same amount of information from random selected decoys, is partitioned into training and validation sets in a 70:30 ratio. The ANN is trained under a back-propagation protocol using the training set, whereas the validation set is used to monitor the ANN performance. As final step, the trained ANN is applied to the test receptor.

ALL-scheme and after 1200 epochs for DDFA-RL (Supporting Information Figure S4). After the training process, ligands from the testing receptor were ranked based on the ANN's output. This ANN training procedure was repeated 40 times, each time with a different set of the 40 receptors of DUD selected as the testing receptor and thus excluded from the training and validation sets.

**2.6. VS Performance Evaluation.** In order to evaluate the performance of VS experiments, several metrics were computed on the benchmark receptors based on the generated rankings. These metrics are the area under the curve (auc) of the receiver operator characteristic (ROC) curve for sensitivity versus

**Table 1. Similarity Relationships between Receptors as Considered to Build Training and Validation Sets[a]**

| class: nuclear hormone receptors | |
|---|---|
| receptor | receptors with positive cross-enrichment |
| AR | TK, ADA, ALR2, PARP, PNP, SAHH |
| ERagonist | PNP |
| ERantagonist | *none* |
| GR | *none* |
| MR | PARP |
| PPARg | *none* |
| PR | *none* |
| RXRa | COX-1 |
| **class: kinases** | |
| receptor | receptors with positive cross-enrichment |
| CDK2 | *none* |
| EGFr | *none* |
| FGFr1 | *none* |
| HSP90 | *none* |
| P38MAP | *none* |
| PDGFrb | *none* |
| SRC | PDE5 |
| TK | ADA, COMT, ALR2, COX-1, GPB, PARP, PNP, SAHH |
| VEGFr2 | *none* |
| **class: serine proteases** | |
| receptor | receptors with positive cross-enrichment |
| FXa | DHFR, GART |
| thrombin | DHFR, ERantagonist |
| trypsin | PPARg, ADA, DHFR |
| **class: metallo enzymes** | |
| receptor | receptors with positive cross-enrichment |
| ACE | ALR2 |
| ADA | *none* |
| COMT | RXRa, ALR2, AmpC, PNP |
| PDE5 | P38MAP |
| **class: folate enzymes** | |
| receptor | receptors with positive cross-enrichment |
| GART | PPARg |
| DHFR | PPARg |
| **class: other enzymes** | |
| receptor | receptors with positive cross-enrichment |
| AChe | FXa |
| ALR2 | GART, ACE, RXRa, PPARg, AmpC, COX-1, COX-2 |
| AmpC | GART, ACE, RXRa, PPARg, ALR2, COX-1, COX-2 |
| COX-1 | ALR2, COX-2 |
| COX-2 | HSP90, ALR2, PARP |
| GPB | COMT |
| HIVPR | *none* |
| HIVTR | PNP |
| HMGR | RXRa, ACE, GART, ARL2,AmpC,COX-1 |
| InhA | *none* |
| NA | PPARg, thrombin, trypsin, ADA |
| PARP | COX-1, PNP |
| PNP | TK, ADA, COMT, COX-1, GPB, PARP, SAHH |
| SAHH | TK, ADA, COMT, COX-1, PARP, GPB, PNP |

[a]The 40 DUD receptors sorted in 6 biological classes. For a given receptor used as test set, all receptors within the same classification and in the list of reported cross-enrichment[21] are excluded from training and validation sets.

specificity [eqs 1 and 2] and enrichment factor (ef) [eq 3]. These metrics were chosen due to their popularity and acceptance in the field.[31]

$$\text{sensitivty} = \frac{\text{true positives}}{\text{total actives}} \quad (1)$$

$$\text{specificiy} = \frac{\text{true negatives}}{\text{total decoys}} \quad (2)$$

$$\text{ef}_{X\%} = \frac{\text{actives found at } X\%}{\text{molecules at } X\%} \frac{\text{total molecules}}{\text{total actives}} \quad (3)$$

To estimate the significance of the difference $\Delta X$ in a metric $X$ between a pair of methods, with $X$ being either the auc or ef, we compute the $p$-value on the average difference[31]

$$p = \frac{1}{2}\left\{1 - \text{erf}\left[\langle \Delta X \rangle \sqrt{\frac{N}{2\text{Var}(\Delta X)}}\right]\right\} \quad (4)$$

Where erf is the error function, $N$ is the number of receptors in the DUD benchmark, and $\text{Var}(\Delta X)$ denotes the variance of $\Delta X$.

## 3. RESULTS AND DISCUSSION

We have developed a novel method for improving virtual screening (VS) results called docking data feature analysis (DDFA). In this approach, all ligands are docked several times with different docking programs. Features derived from the full library of docked poses and scores are assessed by an artificial neural network (ANN) to identify potential active molecules. To test our approach, we used the Directory of Useful Decoys[21] (DUD), which is an established VS benchmark consisting of 40 receptors, each of them having its own screening library with a 1:36 active to decoy ratio. The DUD is a challenging test for receptor-based VS algorithms since the decoys were selected specifically to be similar to the active molecules of each receptor.[21] Each of the 40 DUD libraries were docked using three different programs: Autodock4.2[22] (AD4), Autodockvina[9] (ADV), and RosettaLigand3.4[23] (RL). Docking was conducted with a rigid receptor molecule in AD4 and ADV and with flexible receptor sidechains in RL. Two rankings were generated from each of the three data sets: (i) based on the docking score (RAW) and (ii) based on the novel feature score (DDFA). Additionally, RAW and DDFA rankings were generated by combining all three docking data sets (denoted as RAW-ALL and DDFA-ALL). In the following we compare the VS performance between the two ranking approaches (RAW and DDFA) applied to the four docking data sets (AD4, ADV, RL, ALL). To evaluate docking performance, we computed the area under the curve (auc) of the receiver operator characteristic (ROC) curve given by the various rankings. An AUC of 0.5 reflects a random selection, whereas a value of 1.0 reflects the perfect identification of active compounds. As a second performance measure, we computed the enrichment factor (ef), which compares the active-to-decoy ratio computed at a given cutoff rank.

Compared to all three individual docking programs, DDFA-ALL significantly improves the auc (Figure 3A, C, and E). Notably, DDFA-ALL yields performances above the random level (auc > 0.5) for all the receptors, with 30 of them registering auc values above 0.7 (Table 2). In contrast, RAW-AD4, RAW-ADV, and RAW-RL, yield good performance (auc > 0.7) only in 11, 15, and 15 cases, respectively (Table 3). Even more striking differences are observed in the number of receptors performing poorly (auc < 0.5); with 13, 7, 7, and 0 for RAW-AD4, RAW-ADV, RAW-RL, and DDFA-ALL, respectively. In line with these results, DDFA-ALL obtains an average auc of 0.77, which exceeds the corresponding values of RAW-AD4, RAW-ADV, and RAW-RL by 28%, 20%, and 18%. DDFA-ALL not only clearly outperformed the individual scoring programs in the auc metric but also in the enrichment factor (ef) (Figure 3B, D, and F). Within the first 20% of the ranking, DDFA-ALL's ef is around 50% larger than the efs of the conventionally obtained rankings. Taken together, these findings indicate that the DDFA-ALL is a robust method for evaluating VS experiments, not only because it effectively yields higher average performance in terms of auc and ef but also due to its strong reduction of poor performing receptors.

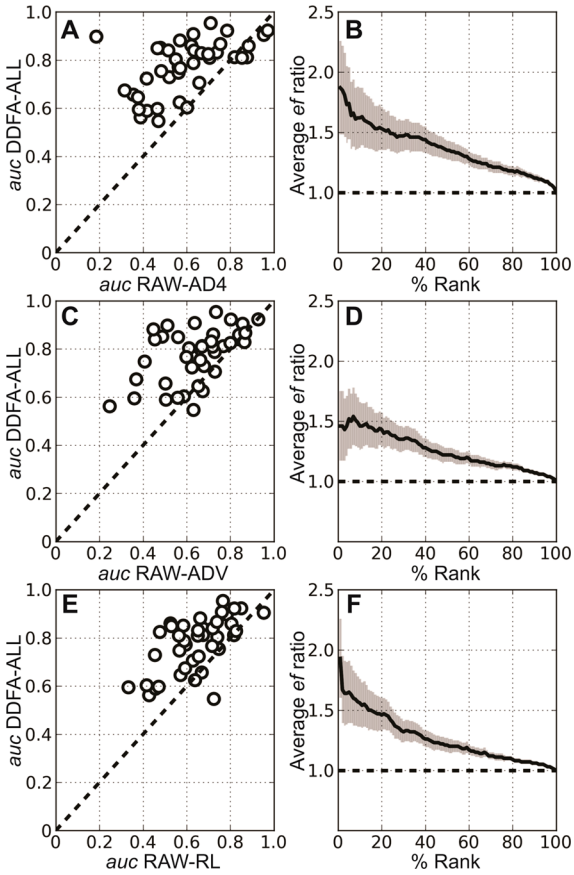



**Figure 3.** DDFA-ALL vs individual RAW rankings. DDFA-ALL is compared against RAW-AD4 (A, B), RAW-ADV (C, D), and RAW-RL (E, F). In plots comparing the auc (A, C, and E) the circles represent each of the 40 DUD receptors. Plots comparing ef (B, D, and F) show the DDFA-ALL to individual RAW average ratio. In all the plots, the dashed line indicates the limit where both methods perform equally.

Next we asked whether the remarkable gain in performance of DDFA-ALL stems from the feature-based analysis of the docking data or from the combination of complementary docking programs. With this objective, we applied the DDFA approach to the data from single docking programs. Interestingly, these individual versions of DDFA still outperform the RAW approach (Figure 4) yielding 28, 27, and 28 receptors with auc > 0.7 for DDFA-AD4, DDFA-ADV, and DDFA-RL, respectively, which has to be compared to the 11, 15, and 15 cases of good performance for RAW approaches


dx.doi.org/10.1021/ci500028u | J. Chem. Inf. Model. 2014, 54, 1401−1411

Table 2. DDFA Rankings Metrics[a]

| | AD4 | | | | ADV | | | | RL | | | | ALL | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $ef_{max}$ | $ef_{2\%}$ | $ef_{20\%}$ | auc | $ef_{max}$ | $ef_{2\%}$ | $ef_{20\%}$ | auc | $ef_{max}$ | $ef_{2\%}$ | $ef_{20\%}$ | auc | $ef_{max}$ | $ef_{2\%}$ | $ef_{20\%}$ | auc |
| average | 12.9 | 9.6 | 2.8 | 0.74 | 13.7 | 10.3 | 2.8 | 0.75 | 13.1 | 9.7 | 3.0 | 0.76 | 13.5 | 10.3 | 3.0 | 0.77 |
| conf 95% | 3.0 | 2.2 | 0.3 | 0.04 | 3.0 | 2.2 | 0.3 | 0.03 | 2.4 | 2.0 | 0.3 | 0.03 | 2.6 | 2.0 | 0.3 | 0.03 |
| ACE | 1.8 | 0.0 | 1.7 | 0.57 | 4.3 | 2.1 | 2.0 | 0.63 | 2.9 | 0.0 | 2.7 | **0.70** | 2.3 | 0.0 | 1.9 | 0.57 |
| AChE | 19.3 | 11.9 | 2.6 | 0.71 | 21.3 | 18.1 | 3.8 | **0.88** | 20.1 | 12.8 | 2.2 | 0.67 | 4.7 | 3.3 | 2.2 | 0.73 |
| ADA | 1.0 | 0.0 | 0.4 | 0.40 | 8.3 | 6.5 | 1.7 | 0.60 | 13.7 | 9.1 | 2.1 | **0.70** | 5.5 | 2.6 | 1.7 | 0.64 |
| ALR2 | 20.4 | 16.3 | 3.4 | 0.81 | 31.4 | 27.5 | 4.0 | **0.87** | 23.5 | 17.7 | 3.3 | 0.74 | 19.6 | 15.7 | 3.9 | 0.85 |
| AmpC | 4.8 | 4.8 | 1.2 | 0.56 | 4.8 | 2.4 | 0.7 | 0.55 | 4.8 | 2.4 | 1.2 | **0.57** | 4.8 | 2.4 | 0.7 | 0.51 |
| AR | 7.3 | 5.8 | 2.7 | 0.78 | 23.3 | 17.5 | 4.0 | **0.88** | 5.1 | 3.2 | 1.4 | 0.54 | 9.7 | 9.7 | 3.9 | 0.82 |
| CDK2 | 14.4 | 10.8 | 3.4 | 0.83 | 21.9 | 16.8 | 3.1 | **0.81** | 11.7 | 8.0 | 2.9 | 0.72 | 13.1 | 10.2 | 3.2 | 0.80 |
| COMT | 7.6 | 0.0 | 3.2 | 0.80 | 43.5 | 29.0 | 3.7 | **0.88** | 4.8 | 4.8 | 2.3 | 0.66 | 23.9 | 10.6 | 4.0 | 0.87 |
| COX-1 | 8.7 | 8.7 | 2.9 | 0.75 | 29.1 | 16.6 | 3.8 | **0.83** | 12.5 | 10.4 | 3.2 | 0.71 | 24.9 | 14.5 | 3.6 | 0.82 |
| COX-2 | 10.6 | 9.8 | 3.0 | 0.78 | 24.7 | 20.2 | 3.7 | 0.84 | 4.8 | 4.2 | 3.3 | 0.80 | 13.5 | 13.0 | 3.9 | **0.86** |
| DHFR | 17.2 | 12.8 | 4.0 | 0.89 | 12.1 | 9.4 | 3.0 | 0.79 | 19.5 | 18.6 | 4.6 | **0.95** | 19.5 | 15.9 | 4.2 | 0.91 |
| EGFr | 2.0 | 1.7 | 1.8 | 0.67 | 7.3 | 7.3 | 2.4 | 0.75 | 13.7 | 11.7 | 3.5 | **0.84** | 11.8 | 8.9 | 3.1 | 0.80 |
| ERagonist | 15.1 | 14.4 | 3.4 | 0.71 | 12.5 | 12.1 | 3.4 | 0.84 | 15.1 | 9.1 | 3.7 | 0.84 | 18.2 | 18.2 | 4.3 | **0.92** |
| ERantagonist | 27.2 | 21.0 | 3.3 | **0.84** | 6.9 | 6.6 | 3.2 | 0.77 | 19.1 | 15.8 | 3.2 | 0.84 | 8.2 | 6.6 | 3.3 | 0.79 |
| FGFrl | 7.5 | 7.5 | 2.0 | 0.62 | 5.9 | 5.0 | 2.5 | **0.73** | 2.1 | 1.3 | 1.6 | 0.61 | 4.2 | 2.9 | 1.6 | 0.57 |
| FXa | 9.2 | 5.6 | 1.6 | 0.58 | 5.6 | 4.5 | 1.9 | 0.62 | 11.2 | 6.9 | 2.3 | **0.65** | 7.1 | 5.0 | 1.6 | 0.61 |
| GART | 2.5 | 0.0 | 2.4 | 0.74 | 2.6 | 2.6 | 1.8 | 0.68 | 12.8 | 11.5 | 3.5 | **0.84** | 15.3 | 7.7 | 3.4 | 0.76 |
| GPB | 8.2 | 5.0 | 1.4 | 0.60 | 8.0 | 5.9 | 3.2 | 0.81 | 18.1 | 14.7 | 3.9 | 0.83 | 22.5 | 19.0 | 4.3 | **0.89** |
| GR | 14.2 | 11.6 | 2.8 | **0.76** | 10.5 | 7.2 | 1.3 | 0.65 | 2.6 | 1.3 | 1.2 | 0.55 | 4.5 | 3.9 | 1.5 | 0.61 |
| HIVPR | 20.3 | 14.0 | 2.6 | 0.72 | 18.6 | 14.9 | 2.9 | 0.74 | 16.9 | 13.2 | 3.4 | **0.80** | 15.5 | 10.9 | 2.6 | 0.76 |
| HIVRT | 8.4 | 8.4 | 2.1 | 0.57 | 16.9 | 10.5 | 2.1 | 0.66 | 12.1 | 5.9 | 1.8 | 0.61 | 9.7 | 9.4 | 2.3 | **0.66** |
| HMGR | 7.2 | 5.8 | 3.6 | 0.80 | 8.7 | 7.2 | 1.7 | 0.64 | 20.2 | 11.5 | 3.2 | **0.86** | 11.5 | 5.8 | 3.6 | 0.82 |
| HSP90 | 1.1 | 0.0 | 0.8 | 0.47 | 3.3 | 1.4 | 2.7 | 0.68 | 5.5 | 2.7 | 1.8 | 0.67 | 1.4 | 0.0 | 1.4 | **0.62** |
| InhA | 25.1 | 18.2 | 3.6 | 0.80 | 34.6 | 21.2 | 3.5 | 0.81 | 29.5 | 17.3 | 3.2 | 0.76 | 23.4 | 17.5 | 3.8 | **0.84** |
| MR | 38.7 | 21.4 | 3.9 | **0.88** | 16.0 | 10.0 | 4.3 | 0.87 | 21.7 | 20.0 | 4.0 | 0.82 | 38.7 | 21.4 | 3.6 | 0.84 |
| NA | 8.0 | 8.0 | 2.4 | 0.71 | 8.3 | 7.2 | 2.6 | 0.68 | 5.2 | 5.2 | 3.7 | **0.80** | 16.0 | 11.5 | 2.5 | 0.74 |
| P38MAP | 2.8 | 2.8 | 1.7 | 0.62 | 8.0 | 7.2 | 3.1 | **0.77** | 3.1 | 2.4 | 2.0 | 0.67 | 4.5 | 3.7 | 2.5 | 0.71 |
| PARP | 40.7 | 28.7 | 4.7 | 0.95 | 9.4 | 7.5 | 4.4 | 0.89 | 33.5 | 26.4 | 4.9 | **0.96** | 31.3 | 25.7 | 4.6 | 0.95 |
| PDE5 | 15.3 | 14.3 | 3.7 | 0.83 | 22.3 | 16.0 | 3.6 | 0.83 | 20.0 | 16.0 | 4.3 | 0.89 | 21.4 | 18.5 | 4.1 | **0.90** |
| PDGFrb | 11.9 | 9.8 | 3.4 | **0.81** | 5.3 | 4.2 | 1.6 | 0.66 | 12.5 | 9.5 | 3.3 | 0.80 | 5.4 | 5.1 | 2.4 | 0.68 |
| PNP | 8.8 | 7.2 | 3.2 | 0.75 | 13.0 | 9.3 | 3.4 | 0.75 | 10.9 | 9.3 | 2.8 | **0.81** | 10.9 | 9.3 | 3.1 | 0.77 |
| PPARg | 7.8 | 7.0 | 2.2 | 0.72 | 7.1 | 7.1 | 2.6 | 0.70 | 9.4 | 8.9 | 3.5 | **0.82** | 5.2 | 3.8 | 2.2 | 0.76 |
| PR | 13.3 | 6.3 | 4.6 | **0.90** | 2.6 | 1.9 | 2.0 | 0.67 | 11.3 | 11.3 | 3.0 | 0.77 | 19.0 | 19.0 | 4.0 | 0.85 |
| RXRa | 33.0 | 25.6 | 5.0 | **0.96** | 17.9 | 17.9 | 4.8 | 0.94 | 16.5 | 10.3 | 4.8 | 0.93 | 20.5 | 20.5 | 4.3 | 0.91 |
| SAHH | 2.0 | 0.0 | 1.5 | 0.53 | 3.8 | 1.6 | 2.3 | 0.72 | 25.7 | 24.7 | 4.3 | **0.90** | 16.1 | 15.5 | 3.8 | 0.80 |
| SRC | 20.4 | 18.0 | 4.0 | **0.87** | 12.1 | 11.1 | 3.4 | 0.82 | 7.0 | 5.1 | 2.9 | 0.77 | 9.6 | 8.8 | 3.0 | 0.83 |
| thrombin | 12.6 | 12.6 | 3.4 | 0.82 | 23.9 | 14.7 | 3.3 | **0.85** | 15.4 | 10.5 | 3.3 | 0.82 | 14.0 | 11.2 | 3.4 | 0.84 |
| TK | 2.9 | 0.0 | 2.5 | **0.66** | 1.8 | 0.0 | 0.7 | 0.62 | 3.0 | 0.0 | 1.6 | 0.54 | 4.6 | 2.3 | 2.3 | 0.61 |
| trypsin | 17.8 | 13.4 | 4.0 | **0.88** | 10.3 | 8.2 | 2.5 | 0.72 | 4.1 | 2.1 | 2.7 | 0.74 | 13.4 | 11.1 | 3.3 | 0.76 |
| VEGFr2 | 19.2 | 13.6 | 2.9 | 0.77 | 22.5 | 14.0 | 2.9 | 0.76 | 21.4 | 14.0 | 3.2 | **0.79** | 16.8 | 11.8 | 2.9 | 0.77 |

[a]Enrichment factor (ef) at 2%, 20%, and maximal reached, in addition to the ROC auc. The bold values indicate the highest auc value achieved in the given receptor.

reported above. Also the number of receptors with auc < 0.5 remains low; the only two observed cases are angiotensin converting enzyme (ACE) and heat shock protein 90 (HSP90), for which auc of 0.40 and 0.47, respectively, are obtained with DDFA-AD4 (Table 2). The average auc values, for DDFA-AD4, ADDF-ADV, and ADDF-RL, are 0.74, 0.75, and 0.76, respectively. These results still correspond to improvements of 23%, 17%, and 17% with respect to their RAW counterparts. Also the ef improves with the DDFA individual versions (Figure 4B, D, and F). For DDFA-AD4 and DDFA-RL, the ef is around 50% larger than that of the corresponding RAW version over the first 10% of the ranking, whereas for DDFA-ADV this degree of improvement is just observed at the starting point of the ranking. This analysis confirms the robustness of the DDFA approach, since a significant enhancement in performance is already obtained even when information from a single docking program only is used.

The above-mentioned observation suggests that some part of the performance gain in DDFA-ALL stems from the combination of different docking programs. To assess this influence, the RAW rankings of the docking programs were combined in to a single list, RAW-ALL. Indeed, the RAW-ALL ranking also outperforms individual RAW rankings (Figure 5A, C, and E), although to a lesser extent than the DDFA-ALL (Figure 5G). In the RAW-ALL approach, 21 proteins reported auc values above 0.70; which exceeds the 11, 15, and 15 of these cases for RAW-AD4, RAW-ADV, and RAW-RL, respectively; but, it is still inferior to the 30 cases for DDFA-


dx.doi.org/10.1021/ci500028u | J. Chem. Inf. Model. 2014, 54, 1401−1411

Table 3. RAW rankings metrics[a]

| | AD4 | | | | ADV | | | | RL | | | | ALL | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $ef_{max}$ | $ef_{2\%}$ | $ef_{20\%}$ | auc | $ef_{max}$ | $ef_{2\%}$ | $ef_{20\%}$ | auc | $ef_{max}$ | $ef_{2\%}$ | $ef_{20\%}$ | auc | $ef_{max}$ | $ef_{2\%}$ | $ef_{20\%}$ | auc |
| average | 7.8 | 5.6 | 2.0 | 0.60 | 9.7 | 7.1 | 2.1 | 0.64 | 7.7 | 6.2 | 2.1 | 0.65 | 9.3 | 7.4 | 2.6 | 0.70 |
| conf 95% | 2.2 | 1.9 | 0.4 | 0.06 | 2.7 | 2.3 | 0.3 | 0.05 | 1.7 | 1.6 | 0.3 | 0.04 | 2.1 | 2.0 | 0.4 | 0.05 |
| ACE | 2.1 | 0.0 | 1.0 | **0.38** | 1.3 | 0.0 | 0.8 | 0.36 | 1.0 | 0.0 | 0.5 | 0.33 | 1.2 | 0.0 | 1.2 | 0.35 |
| AChE | 1.9 | 1.9 | 1.3 | 0.52 | 5.4 | 4.8 | 2.8 | **0.68** | 1.0 | 0.0 | 0.7 | 0.45 | 2.9 | 2.4 | 2.3 | 0.61 |
| ADA | 1.0 | 0.0 | 0.0 | 0.36 | 1.2 | 0.0 | 0.9 | 0.50 | 2.3 | 0.0 | 1.3 | **0.67** | 1.5 | 0.0 | 0.6 | 0.57 |
| ALR2 | 2.9 | 0.0 | 2.6 | 0.62 | 9.8 | 3.9 | 2.7 | 0.72 | 5.9 | 5.9 | 1.5 | 0.53 | 6.2 | 2.0 | 2.7 | **0.73** |
| AmpC | 1.1 | 0.0 | 0.2 | 0.39 | 1.0 | 0.0 | 0.2 | 0.25 | 1.0 | 0.0 | 0.5 | **0.43** | 1.0 | 0.0 | 0.2 | 0.30 |
| AR | 9.0 | 7.0 | 2.7 | 0.70 | 19.4 | 18.8 | 3.7 | 0.80 | 5.1 | 3.2 | 1.5 | 0.48 | 15.4 | 14.7 | 3.9 | **0.84** |
| CDK2 | 8.6 | 4.3 | 2.1 | 0.56 | 13.1 | 10.9 | 2.1 | 0.62 | 4.4 | 2.9 | 1.6 | 0.60 | 8.6 | 6.5 | 2.5 | **0.63** |
| COMT | 1.4 | 0.0 | 1.4 | 0.48 | 32.6 | 14.5 | 1.4 | 0.49 | 4.8 | 4.8 | 0.9 | **0.58** | 10.9 | 9.7 | 1.8 | 0.55 |
| COX-1 | 2.2 | 2.2 | 0.8 | 0.52 | 16.6 | 10.4 | 3.8 | **0.84** | 6.2 | 6.2 | 1.6 | 0.67 | 12.5 | 12.5 | 3.6 | 0.83 |
| COX-2 | 7.8 | 7.3 | 2.8 | 0.75 | 25.6 | 23.0 | 4.0 | 0.87 | 4.0 | 3.3 | 2.6 | 0.74 | 18.8 | 17.3 | 4.3 | **0.90** |
| DHFR | 17.7 | 14.7 | 4.8 | **0.95** | 11.1 | 8.9 | 3.5 | 0.86 | 17.5 | 17.4 | 4.7 | 0.95 | 14.5 | 12.5 | 4.7 | 0.94 |
| EGFr | 1.7 | 1.5 | 1.2 | 0.55 | 2.5 | 1.1 | 1.7 | 0.61 | 7.0 | 6.0 | 2.6 | **0.74** | 4.9 | 3.6 | 2.2 | 0.72 |
| ERagonist | 21.2 | 18.2 | 3.1 | 0.79 | 18.2 | 18.2 | 3.3 | 0.80 | 16.7 | 12.1 | 3.6 | 0.82 | 18.2 | 15.9 | 4.6 | **0.93** |
| ERantagonist | 21.8 | 19.7 | 3.5 | **0.82** | 13.6 | 9.2 | 2.2 | 0.67 | 13.6 | 11.8 | 2.3 | 0.67 | 13.2 | 13.2 | 3.3 | 0.77 |
| FGFrl | 1.0 | 0.8 | 0.7 | 0.42 | 1.2 | 0.4 | 0.9 | **0.50** | 1.0 | 0.0 | 0.8 | 0.46 | 1.0 | 0.0 | 0.7 | 0.46 |
| FXa | 2.1 | 1.8 | 1.2 | 0.57 | 2.4 | 2.4 | 1.8 | **0.67** | 9.1 | 6.6 | 1.9 | 0.64 | 5.5 | 5.5 | 1.8 | 0.66 |
| GART | 4.4 | 1.3 | 4.0 | **0.88** | 2.9 | 0.0 | 2.6 | 0.77 | 5.6 | 5.1 | 3.5 | 0.82 | 3.9 | 1.3 | 3.9 | 0.85 |
| GPB | 1.0 | 0.0 | 0.1 | 0.19 | 2.9 | 2.9 | 1.2 | 0.51 | 10.4 | 9.8 | 3.5 | 0.78 | 6.3 | 2.9 | 3.2 | **0.82** |
| GR | 6.5 | 5.8 | 2.1 | **0.60** | 7.9 | 5.2 | 1.4 | 0.59 | 1.3 | 0.7 | 0.6 | 0.42 | 9.1 | 6.5 | 1.6 | 0.57 |
| HIVPR | 5.1 | 4.1 | 2.5 | 0.66 | 5.8 | 5.8 | 2.7 | 0.73 | 6.8 | 5.8 | 2.3 | 0.63 | 6.5 | 4.8 | 2.6 | **0.73** |
| HIVRT | 2.5 | 2.4 | 0.7 | 0.38 | 12.1 | 7.0 | 1.9 | **0.65** | 7.3 | 5.9 | 1.9 | 0.57 | 9.7 | 7.0 | 2.0 | 0.63 |
| HMGR | 3.9 | 2.9 | 1.7 | 0.63 | 1.2 | 0.0 | 0.7 | 0.45 | 2.9 | 1.4 | 0.9 | **0.72** | 1.9 | 1.4 | 1.4 | 0.62 |
| HSP90 | 1.2 | 0.0 | 0.4 | 0.47 | 1.7 | 0.0 | 1.2 | 0.63 | 2.4 | 0.0 | 2.4 | **0.72** | 1.6 | 0.0 | 1.0 | 0.64 |
| InhA | 22.7 | 13.5 | 1.9 | 0.47 | 19.1 | 12.4 | 1.8 | **0.56** | 10.3 | 5.1 | 1.6 | 0.53 | 15.4 | 11.1 | 2.2 | 0.53 |
| MR | 15.5 | 10.7 | 4.6 | **0.88** | 23.3 | 23.3 | 4.0 | 0.84 | 16.7 | 16.7 | 3.3 | 0.81 | 21.7 | 16.7 | 4.0 | 0.84 |
| NA | 2.3 | 1.2 | 0.9 | 0.56 | 1.1 | 0.0 | 0.4 | 0.41 | 4.1 | 3.1 | 1.6 | **0.57** | 2.1 | 2.1 | 0.8 | 0.49 |
| P38MAP | 1.6 | 1.2 | 0.7 | 0.42 | 3.1 | 2.0 | 2.3 | 0.62 | 3.6 | 2.0 | 1.9 | **0.65** | 2.4 | 2.3 | 2.0 | 0.63 |
| PARP | 25.1 | 21.1 | 2.7 | 0.71 | 9.4 | 4.5 | 2.8 | 0.73 | 15.2 | 13.2 | 3.1 | 0.77 | 15.2 | 14.7 | 4.3 | **0.91** |
| PDE5 | 11.7 | 6.9 | 2.3 | 0.63 | 11.7 | 8.0 | 2.0 | 0.64 | 10.6 | 9.2 | 2.6 | **0.76** | 15.3 | 9.7 | 2.9 | 0.75 |
| PDGFrb | 7.7 | 4.7 | 0.9 | 0.32 | 5.3 | 3.0 | 0.6 | 0.37 | 4.2 | 3.6 | 1.7 | **0.59** | 6.5 | 3.5 | 1.3 | 0.50 |
| PNP | 6.1 | 3.1 | 2.4 | 0.63 | 4.8 | 4.1 | 2.6 | 0.73 | 5.6 | 3.1 | 1.9 | 0.59 | 4.2 | 2.1 | 3.1 | **0.79** |
| PPARg | 1.2 | 0.0 | 1.1 | 0.48 | 4.7 | 3.5 | 1.8 | 0.66 | 7.1 | 5.9 | 2.7 | **0.75** | 3.5 | 1.8 | 1.9 | 0.63 |
| PR | 13.3 | 8.4 | 1.7 | 0.57 | 1.9 | 0.0 | 1.1 | 0.45 | 19.7 | 15.0 | 1.9 | 0.66 | 15.8 | 9.4 | 3.2 | **0.81** |
| RXRa | 16.5 | 15.4 | 5.0 | 0.97 | 33.0 | 28.2 | 4.3 | 0.93 | 17.9 | 17.9 | 4.0 | 0.85 | 28.2 | 28.2 | 5.0 | **0.98** |
| SAHH | 2.6 | 0.0 | 2.1 | 0.67 | 22.5 | 20.1 | 3.5 | **0.86** | 17.0 | 17.0 | 3.5 | 0.83 | 13.3 | 12.4 | 3.3 | 0.81 |
| SRC | 14.0 | 10.7 | 2.8 | 0.70 | 5.0 | 4.1 | 2.4 | 0.72 | 7.6 | 4.4 | 2.0 | 0.65 | 12.1 | 10.1 | 3.9 | **0.85** |
| thrombin | 8.4 | 7.0 | 2.8 | 0.74 | 11.2 | 8.4 | 3.0 | 0.71 | 7.0 | 7.0 | 2.4 | 0.65 | 11.2 | 9.1 | 3.1 | **0.79** |
| TK | 1.5 | 0.0 | 0.9 | 0.47 | 1.6 | 0.0 | 0.7 | 0.56 | 4.6 | 2.3 | 1.1 | 0.47 | 2.3 | 2.3 | 1.1 | **0.57** |
| trypsin | 17.8 | 14.5 | 3.7 | **0.85** | 8.2 | 4.1 | 2.5 | 0.67 | 3.1 | 3.1 | 1.7 | 0.56 | 8.2 | 6.2 | 3.5 | 0.83 |
| VEGFr2 | 15.6 | 8.9 | 1.9 | 0.57 | 14.2 | 8.7 | 2.1 | 0.60 | 16.6 | 9.9 | 2.7 | **0.72** | 20.2 | 13.4 | 2.8 | 0.70 |

[a]Enrichment factor (ef) at 2%, 20%, and maximal reached, in addition to the ROC auc. The bold values indicate the highest auc value achieved in the given receptor.

ALL. On the side of poor-performing receptors (auc < 0.5), their number is reduced to four, which certainly improves with respect to the individual RAW rankings, but not in comparison with DDFA-ALL, with its zero cases with auc < 0.5. An equivalent picture is observed with the ef metrics. RAW-ALL outperforms the individual rankings (Figure 5B, D, and F), but not DDAF-ALL where RAW-ALL is at least 20% smaller over the initial 10% of the ranking (Figure 5H). These results provide evidence on the beneficial effect that is obtained from the combination of three docking data sources.

To evaluate the significance of the observed differences between methods in the performance metrics auc and $ef_{2\%}$, we computed their *p*-value[31] (Methods). The comparison of the four different versions of RAW and DDFA yields remarkably low *p*-values ($<1 \times 10^{-3}$; Table4A). The further improvement in auc achieved by DDFA-ALL with respect to DDFA-AD4, DDFA-ADV, and DDFA-RL is confirmed by the low *p*-values, 0.02, 0.04, and 0.07, respectively (Table 4B). In contrast, DDFA-ALL does not yield significantly better $ef_{2\%}$ than the individual versions of DDFA (Table 4B). Taken together, DDFA is significantly better than RAW in both metrics, whereas DDFA-ALL outperforms the individual versions of DDFA only in the auc metric.

After confirming the significance of the results yielded by DDFA, we wanted to assess their stability with respect to the number of receptors used during the training and validation process. The systematic reductions of receptors used for training causes a gradual decay in performance for all four


dx.doi.org/10.1021/ci500028u | J. Chem. Inf. Model. 2014, 54, 1401−1411

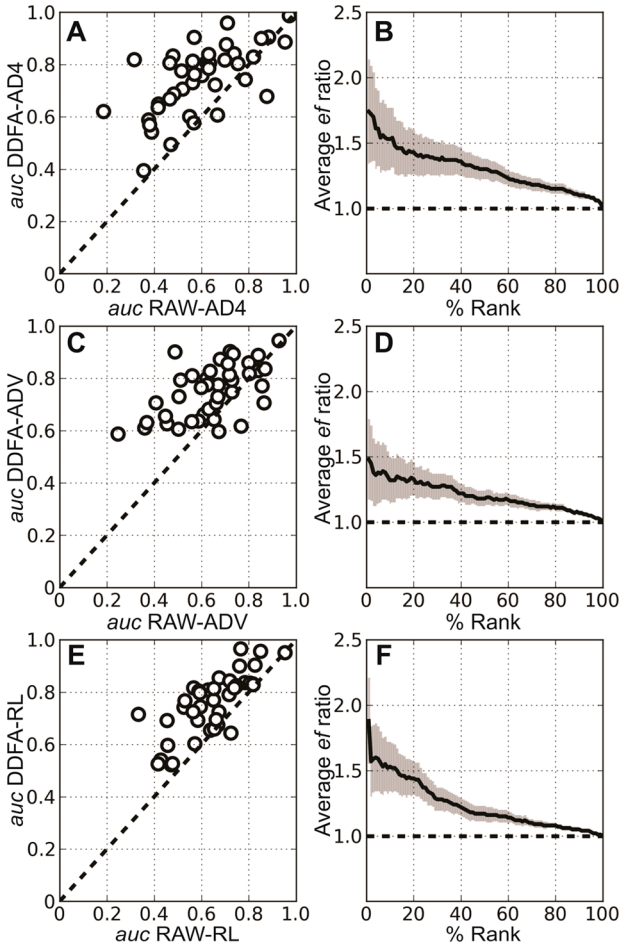



**Figure 4.** Individual DDFA vs individual RAW rankings. Individual versions of both, DDFA and RAW, are compared for AD4 (A, B), ADV (C, D), and RL (E, F). In plots comparing the auc (A, C, and E) the circles represent each of the 40 DUD receptors. Plots comparing ef (B, D, and F) show the individual DDFA to individual RAW average ratio. In all the plots, the dashed line indicates the limit where both methods perform equally.

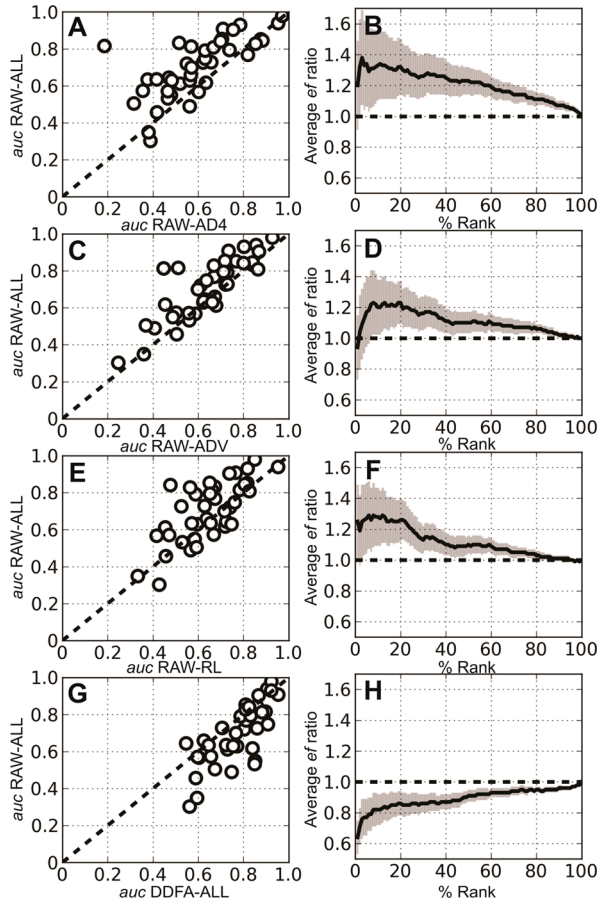



**Figure 5.** RAW-ALL vs individual RAW and DDFA-ALL rankings. RAW-ALL is compared against RAW-AD4 (A, B), RAW-ADV (C, D), RAW-RL (E, F), and DDFA-ALL (G, H), respectively. In plots comparing the auc (A, C, E, and G) the circles represent each of the 40 DUD receptors. Plots comparing ef show the RAW-ALL to individual RAW average ratio (B, D, and F), and the RAW-ALL to DDFA-ALL average ratio. In all the plots, the dashed line indicates the limit where both methods perform equally.

DDFA cases (Figure 6), as expected when using less training data. Nevertheless, DDFA performance is always at least as good as the RAW performance (Figure 6; Table 3), such that one can say with confidence that DDFA is robust in the sense, that it never hurts average performance to use it. Moreover, we should note that in this test we did not reoptimize the hyperparameters that control training for each number of receptors such that one might be able to improve performance by hyperparameter tuning (Supporting Information Figure S5). The result in Figure 6 also shows that the method has potential to achieve an even better performance than demonstrated here, if more receptors than 22 were available for training.

In addition to benchmarking the DDFA approach on DUD, our study also provides valuable insight into the VS performance of the individual docking programs (Table 3). On average, RAW-ADV yielded better results than RAW-AD4 and RAW-RL. The superior performance of RAW-ADV over RAW-AD4 in not surprising, since it matches with previously reported observations.[9,32,33] This is the first time, however, that results for RL obtained on the DUD benchmark were published. RL obtained averages values of 0.65 and 6.19 in auc and $ef_{2\%}$, respectively, thereby yielding similar results to ADV and better than AD4 (Table 3). This result is in line with the outstanding performances that RAW-RL obtained in pose recovery benchmarks.[10,27,34] Table 3 shows that for three receptors none of the docking programs reached auc values above the random level: (i) angiotensin converting enzyme (ACE), (ii) Amp-C beta lactamase (AmpC), and (iii) fibroblast growth factor receptor 1 (FGFr1). These are the three receptors for which DDFA-ALL yielded also the poorest results with auc of 0.57, 0.51, and 0.57 for ACE, AmpC, and FGFr1, respectively. This observation suggest that the improvement produced by DDFA-ALL is somewhat limited by the quality of the individual docking results. Another interesting example is the platelet derived growth factor receptor (PDGFrb), a receptor for which another seven different scoring functions report auc values under 0.5.[35] In our hands, the aucs yielded by RAW-AD4 and RAW-ADV for PDGFrb are also below 0.5, whereas RAW-RL obtains an auc of 0.59. In contrast, the performances of the individual version of DDFA are undoubtedly better; 0.81, 0.66, and 0.80 for DFFA-AD4, DDFA-ADV, and DDFA-RL, respectively.

As shown above DDFA represents a highly attractive alternative to traditional ranking approaches for analyzing VS experiments. This finding is also supported by comparing DDFA performances with those found in the literature (Table

**Table 4. *p*-Values of the Difference in Metrics (auc, $ef_{2\%}$) between Each Pair of Methods[a]**

| | AD4 | | ADV | | RL | | ALL | |
|---|---|---|---|---|---|---|---|---|
| (A) Significance of the Difference in Performance between RAW and DDFA | | | | | | | | |
| | RAW | DDFA | RAW | DDFA | RAW | DDFA | RAW | DDFA |
| | *(0.60, 5.6)* | *(0.74, 9.5)* | *(0.64, 7.1)* | *(0.75, 10.3)* | *(0.65, 6.2)* | *(0.76, 9.7)* | *(0.70, 7.4)* | *(0.77, 10.3)* |
| *p*-value in auc | $<1 \times 10^{-5}$ | | $<1 \times 10^{-5}$ | | $<1 \times 10^{-5}$ | | $<1 \times 10^{-5}$ | |
| *p*-value in $ef_{2\%}$ | $<1 \times 10^{-5}$ | | $3 \times 10^{-4}$ | | $<1 \times 10^{-5}$ | | $3 \times 10^{-4}$ | |

| | DDFA | | DDFA | | DDFA | |
|---|---|---|---|---|---|---|
| (B) Significance of the Difference in Performance between DDFA-ALL and the Individual Versions of DDFA | | | | | | |
| | AD4 | ALL | ADV | ALL | RL | ALL |
| | *(0.74, 9.5)* | *(0.77, 10.3)* | *(0.75, 10.3)* | *(0.77, 10.3)* | *(0.76, 9.7)* | *(0.77, 10.3)* |
| *p*-value in auc | 0.02 | | 0.04 | | 0.07 | |
| *p*-value in $ef_{2\%}$ | 0.27 | | 0.59 | | 0.42 | |

[a]The lower the *p*-value, the more significant the differences in performance.

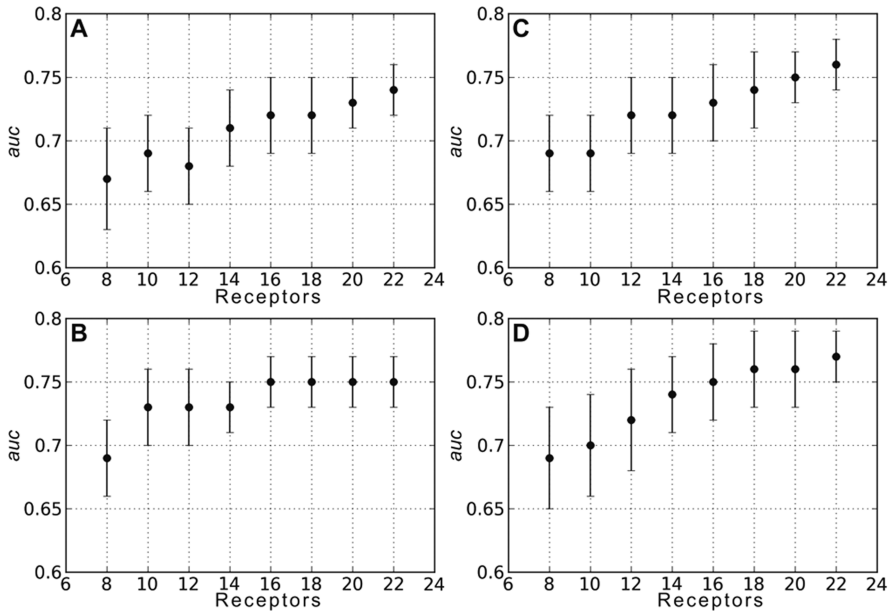



**Figure 6.** Average auc on the DUD benchmark yielded by (A) DDFA-AD4, (B) DDFA-ADV, (C) DDFA-RL, and (D) DDFA-ALL, with a different number of training receptors. The plotted values correspond to the average over five independent runs using a different subset of receptors. Error bars correspond to the associated standard deviations.

5). Considering structured-based methodologies tested on the DUD benchmark, the different versions of the DDFA approach (ALL, AD4, ADV, and RL) obtained performances that situate them among the best methods available. Certainly, the commercial docking software, ICM and Glide SP, achieve the top performances in the auc and $ef_{2\%}$ metrics, respectively. Nonetheless, their corresponding performances fall within the 95% confidence limits of DDFA-ALL; auc 0.77 ± 0.03 and ef 10.3 ± 2.0 (Table 2). One of the best methods we also found is the methodology developed by Durrant et al.[37] in which NNScore[38] is used. This methodology resembles ours in the sense that it combines academic docking software with an artificial neutral network. However, while NNScore is trained on the characteristic interactions of protein−ligand complexes, thus proposing an interaction rescoring scheme, our DDFA is trained on the characteristic features of the docking data associated with active molecules, thereby representing a reranking scheme. Additionally our DDFA approach also yields high ef values at 2%, which, together with the averaged ef curves presented previously (for example Figure 3), provide confidence on the performance stability that our approach has on this metric. These findings, together with the inherent flexibility of DDFA (easily extended to combine several docking programs and docking features), render our novel approach as highly attractive for analyzing VS experiments.

## 3. CONCLUSION

The DDFA approach introduced in this work was able to improve considerably the selection of active compounds from the output of popular docking programs. This was achieved by extending the analysis of the docking data beyond the traditional docking score. Although the usefulness of rescoring, consensus rankings, and machine learning methods has already been noted,[39−41] what distinguishes our study is that we could convincingly show a possible way to combine all these elements together synergistically. It must be emphasized, however, that the success on combining several docking features and/or scoring programs resides in their diversity.[42,43] Each element should account for different characteristics that contribute to the active-decoy discrimination. Although establishing the

Table 5. Reported Performances on DUD[a]

| methodologies | auc | $ef_{2\%}$ |
|---|---|---|
| ICM [ref 36][b] | 0.79 | |
| AutodockVina-NN1 [ref 37] | 0.78 | |
| Glide SP [ref 35][b] | 0.77 | 12.2 |
| **DDFA-ALL** | **0.77** | **10.3** |
| **DDFA-RL** | **0.76** | **9.7** |
| AutodockVina-NN2 [ref 37] | 0.76 | |
| **DDFA-ADV** | **0.75** | **10.3** |
| normalization score [ref 16][c,d,e] | 0.75 | |
| **DDFA-AD4** | **0.74** | **9.5** |
| Glide HTVS [ref 37] | 0.73 | |
| Surflex [ref 35] | 0.72 | 12.0 |
| Glide HTVS [ref 35] | 0.72 | 10.7 |
| ICM [ref 36] | 0.71 | |
| **RAW-ALL** | **0.70** | **7.4** |
| Autodock Vina [ref 37] | 0.70 | |
| eHiTS [ref 16][d,e] | 0.70 | |
| Glide SP [ref 16][d,e] | 0.70 | |
| Surflex [ref 35][b] | 0.66 | 7.9 |
| **RosettaLigand** | **0.65** | **6.2** |
| **AutodockVina** | **0.64** | **7.1** |
| NNScore 1.0 [ref 38][d] | 0.64 | |
| ICM [ref 35] | 0.63 | 8.0 |
| FlexX [ref 35] | 0.61 | 7.2 |
| **Autodock4.2** | **0.60** | **5.6** |
| PhDock [ref 35] | 0.59 | 7.7 |
| NNScore 2.0 [ref 32][d] | 0.59 | |
| AutodockVina [ref 32][d] | 0.58 | |
| Dock [ref 35] | 0.55 | 8.2 |
| $Autodock_{fast}$ [ref 32][d] | 0.51 | |
| $Autodock_{rigorous}$ [ref 32][d] | 0.50 | |

[a]Methodologies reported in this work are highlighted in bold letters. [b]Tuned by expert knowledge. [c]Computational expensive. [d]Subset of the DUD receptors. [e]Subset of decoys used.

optimal docking feature selection for a given set of scoring programs is a challenging task, it certainly opens a pathway to possible further improvements.

In terms of the well-established virtual screening metrics, auc and ef, DDFA performance is statistically similar to that reported by commercial software under expert intervention[35,36] or by methods that increase the computational cost by 2 orders of magnitude.[16] Additionally, DDFA shows an excellent stability in its results and, in strong contrast to simple ranking schemes, performs better than random selection for every single receptor in the DUD benchmark. Overall, DDFA represents a new, simple, and automatic reranking treatment that not only is easy to implement and extend to other docking software or docking data features but also provides high VS performance with minimal extra computing time.

## ■ ASSOCIATED CONTENT

### Supporting Information

Benchmark of the parameters used to setup the ANN of DDFA. This material is available free of charge via the Internet at http://pubs.acs.org. The scripts used to evaluate the ligands with DDFA are available upon request.

## ■ AUTHOR INFORMATION

### Corresponding Author

*E-mail: castro@biochem.mpg.de.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work has been supported by the CONACYT-DAAD scholarship 209523 (M.A.) and the DFG grant LA 1817/3-1 (O.F.L.)

## ■ ABBREVIATIONS

AD4, Autodock4; ADV, AutodockVina; RL, Rosetta Ligand; ANN, artificial neural network; DDFA, docking data feature analysis

## ■ REFERENCES

(1) Khanna, I. Drug Discovery in Pharmaceutical Industry: Productivity Challenges and Trends. *Drug Discovery Today* **2012**, *17*, 1088−1102.

(2) Cavasotto, C. N.; Orry, A. J. W. Ligand Docking and Structure-based Virtual Screening in Drug Discovery. *Curr. Top. Med. Chem.* **2007**, *7*, 1006−1014.

(3) Tanrikulu, Y.; Krüger, B.; Proschak, E. The Holistic Integration of Virtual Screening in Drug Discovery. *Drug Discovery Today* **2013**, *18*, 358−364.

(4) Kar, S.; Roy, K. How Far Can Virtual Screening Take us in Drug Discovery? *Expert Opin. Drug Discovery* **2013**, *8*, 245−261.

(5) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martínez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *J. Chem. Inf. Model.* **2012**, *52*, 867−881.

(6) Drwal, M. N.; Griffith, R. Combination of Ligand- and Structure-Based Methods in Virtual Screening. *Drug Discovery Today: Technologies* **2013**, *10*, e395−e401.

(7) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.

(8) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.

(9) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.

(10) Davis, I. W.; Raha, K.; Head, M. S.; Baker, D. Blind Docking of Pharmaceutically Relevant Compounds Using RosettaLigand. *Protein Sci.* **2009**, *18*, 1998−2002.

(11) Zhong, S.; Zhang, Y.; Xiu, Z. Rescoring Ligand Docking Poses. *Curr. Opin. Drug Discovery Dev.* **2010**, *13*, 326−34.

(12) Rajamani, R.; Good, A. C. Ranking Poses in Structure-Based Lead Discovery and Optimization: Current Trends in Scoring Function Development. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 308−15.

(13) Brewerton, S. C. The Use of Protein-Ligand Interaction Fingerprints in Docking. *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 356−64.

(14) García-Sosa, A. T.; Hetényi, C.; Maran, U. Drug Efficiency Indices for Improvement of Molecular Docking Scoring Functions. *J. Comput. Chem.* **2010**, *31*, 174−184.

(15) Lee, J.; Seok, C. A Statistical Rescoring Scheme for Protein−Ligand Docking: Consideration of Entropic Effect. *Proteins* **2008**, *70*, 1074−1083.

(16) Wallach, I.; Jaitly, N.; Nguyen, K.; Schapira, M.; Lilien, R. Normalizing Molecular Docking Rankings Using Virtually Generated Decoys. *J. Chem. Inf. Model.* **2011**, *51*, 1817−1830.

(17) Feher, M. Consensus Scoring for Protein-Ligand Interactions. *Drug Discovery Today* **2006**, *11*, 421−428.
(18) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus Scoring for Ligand/Protein Interactions. *J. Mol. Graph. Model.* **2002**, *20*, 281−295.
(19) Yang, J. M.; Chen, Y. F.; Shen, T. W.; Kristal, B. S.; Hsu, D. F. Consensus Scoring Criteria for Improving Enrichment in Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 1134−1146.
(20) Jacobsson, M.; Lidén, P.; Stjernschantz, E.; Boström, H.; Norinder, U. Improving Structure-Based Virtual Screening by Multivariate Analysis of Scoring Data. *J. Med. Chem.* **2003**, *46*, 5781−5789.
(21) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.
(22) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. Autodock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexiblity. *J. Comput. Chem.* **2009**, *16*, 2785−91.
(23) Meiler, J.; Baker, D. ROSETTALIGAND: Protein-Small Molecule Docking with Full Side-Chain Flexibility. *Proteins* **2006**, *65*, 538−548.
(24) Lemmon, G.; Meiler, J. Rosetta Ligand Docking with Flexible XML Protocols. In *Computational Drug Discovery and Design*; Baron, R., Ed.; Springer: New York, 2012; Vol. *819*, pp 143−155.
(25) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.
(26) Fleishman, S. J.; Leaver-Fay, A.; Corn, J. E.; Strauch, E. M.; Khare, S. D.; Koga, N.; Ashworth, J.; Murphy, P.; Richter, F.; Lemmon, G.; Meiler, J.; Baker, D. RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. *PLoS One* **2011**, *6*, e20161.
(27) Davis, I. W.; Baker, D. ROSETTALIGAND Docking with Full Ligand and Receptor Flexibility. *J. Mol. Biol.* **2009**, *385*, 381−392.
(28) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3*, 33.
(29) Dreiseitl, S.; Ohno-Machado, L. Logistic Regression and Artificial Neural Network Classification Models: a Methodology Review. *J. Biomed. Inform.* **2002**, *35*, 352−359.
(30) Schaul, T.; Bayer, J.; Wierstra, D.; Sun, Y.; Felder, M.; Sehnke, F.; Rückstieß, T.; Schmidhuber, J. PyBrain. *J. Mach. Learn. Res.* **2010**, *11*, 743−746.
(31) Nicholls, A. What Do We Know and When Do We Know it? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239−255.
(32) Durrant, J. D.; McCammon, J. A. NNScore 2.0: A Neural-Network Receptor−Ligand Scoring Function. *J. Chem. Inf. Model.* **2011**, *51*, 2897−2903.
(33) Chang, M. W.; Ayeni, C.; Breuer, S.; Torbett, B. E. Virtual Screening for HIV Protease Inhibitors: A Comparison of AutoDock 4 and Vina. *PLoS One* **2010**, *5*, e11955.
(34) Kaufmann, K. W.; Meiler, J. Using RosettaLigand for Small Molecule Docking into Comparative Models. *PLoS One* **2012**, *7*, e50769.
(35) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455−1474.
(36) Neves, M. A.; Totrov, M.; Abagyan, R. Docking and Scoring with ICM: The Benchmarking Results and Strategies for Improvement. *J. Comput. Aided Mol. Des.* **2012**, *26*, 675−686.
(37) Durrant, J. D.; Friedman, A. J.; Rogers, K. E.; McCammon, J. A. Comparing Neural-Network Scoring Functions and the State of the Art: Applications to Common Library Screening. *J. Chem. Inf. Model.* **2013**, *53*, 1726−1735.
(38) Durrant, J. D.; McCammon, J. A. NNScore: A Neural-Network-Based Scoring Function for the Characterization of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2010**, *50*, 1865−1871.
(39) Houston, D. R.; Walkinshaw, M. D. Consensus Docking: Improving the Reliability of Docking in a Virtual Screening Context. *J. Chem. Inf. Model.* **2013**, *53*, 384−390.
(40) Planesas, J. M.; Claramunt, R. M.; Teixidó, J.; Borrell, J. I.; Pérez-Nueno, V. I. Improving VEGFR-2 Docking-Based Screening by Pharmacophore Postfiltering and Similarity Search Postprocessing. *J. Chem. Inf. Model.* **2011**, *51*, 777−787.
(41) Teramoto, R.; Fukunishi, H. Supervised Consensus Scoring for Docking and Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47*, 526−534.
(42) Hsu, D. F.; Chung, Y.-S.; Kristal, B. S., Combinatorial Fusion Analysis: Methods and Practice of Combining Multiple Scoring Systems. In *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*; IGI Global, 2008; pp 1157−1181.
(43) Hsu, D. F.; Kristal, B.; Schweikert, C., Rank-Score Characteristics (RSC) Function and Cognitive Diversity. In *Brain Informatics*; Yao, Y., Sun, R., Poggio, T., Liu, J., Zhong, N., Huang, J., Eds.; Springer: Berlin and Heidelberg, 2010; Vol. *6334*, pp 42−54.

# Improvement of Virtual Screening results by Docking Data Feature Analysis

AUTHORS

*Marcelino Arciniega*[1,2]

*Oliver F. Lange*[2,3]

AUTHOR ADDRESS

1) Max Planck Institute Biochemistry. Am Klopferspitz 18, 82152 Martinsried Germany

2) Biomolecular NMR and Munich Center for Integrated Protein Science, Department Chemie, Technische Universität München, 85747 Garching, Germany

3) Institute of Structural Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany

**Corresponding Author**

Marcelino Arciniega*

* castro@biochem.mpg.de

ASSOCIATED CONTENT

**Figure S1**. Number of nodes in the hidden layer.

**Figure S2.** Partition of the Train and Test Sets.

**Figure S3.** Back propagation train parameters.

**Figure S4.** Number training Epochs.

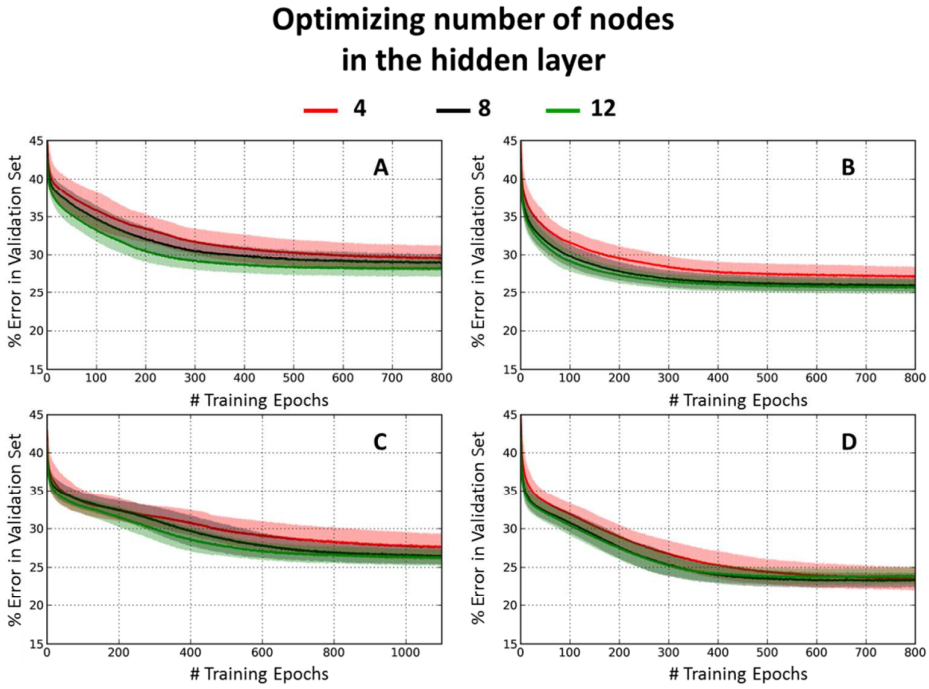



**Figure S1**. Number of nodes in the hidden layer. The error on evaluating the Validation Set averaged over forty receptors for Autodock4 (A), AutodockVina (B), RosettaLigand (C), and ALL scheme (D). As the number of nodes increases, the standard deviation (shaded regions) is reduced and a faster convergence is obtained.

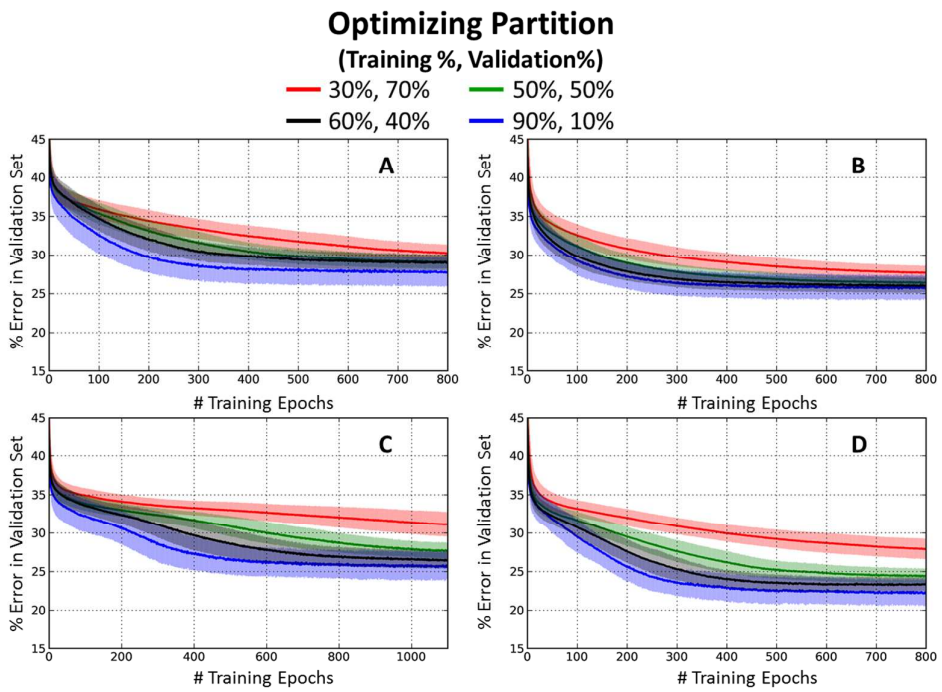



**Figure S2.** Partition of the Train and Validation Sets. The error on evaluating the Validation Set averaged over forty receptors for Autodock4 (A), AutodockVina (B), RosettaLigand (C), and ALL scheme (D).

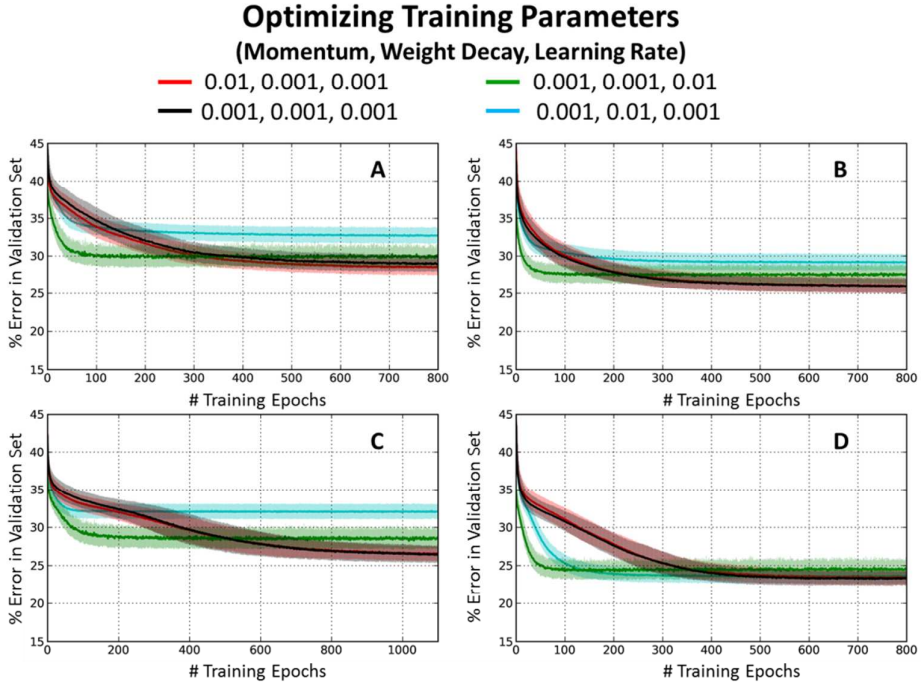



**Figure S3**. Back propagation train parameters. The error on evaluating the Validation Set averaged over forty receptors for Autodock4 (A), AutodockVina (B), RosettaLigand (C), and ALL scheme (D). The optimal performance curve (black) is obtained with by setting the parameters of back propagation trainer of pybrain to: momentum = 0.001, weight decay = 0.001, and learning rate = 0.001.

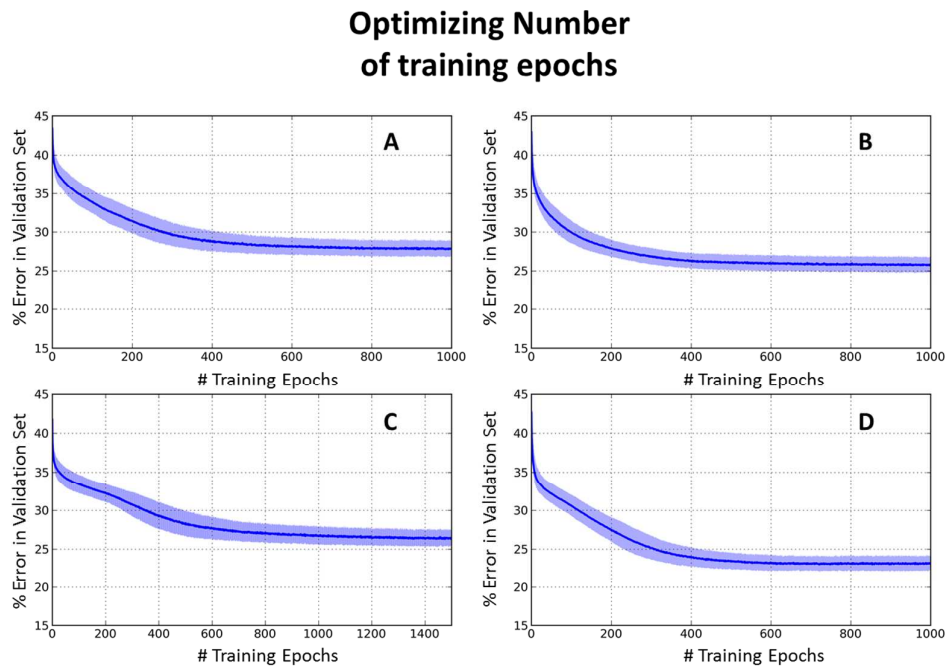



**Figure S4**. Number training Epochs. The error on evaluating the Validation Set averaged over forty receptors for Autodock4 (A), AutodockVina (B), RosettaLigand (C), and ALL scheme (D). Training during 800 epochs appears to be sufficient for Autodock4, AutodockVina and the ALL scheme, whereas 1300 epochs are required for RosettaLigand.

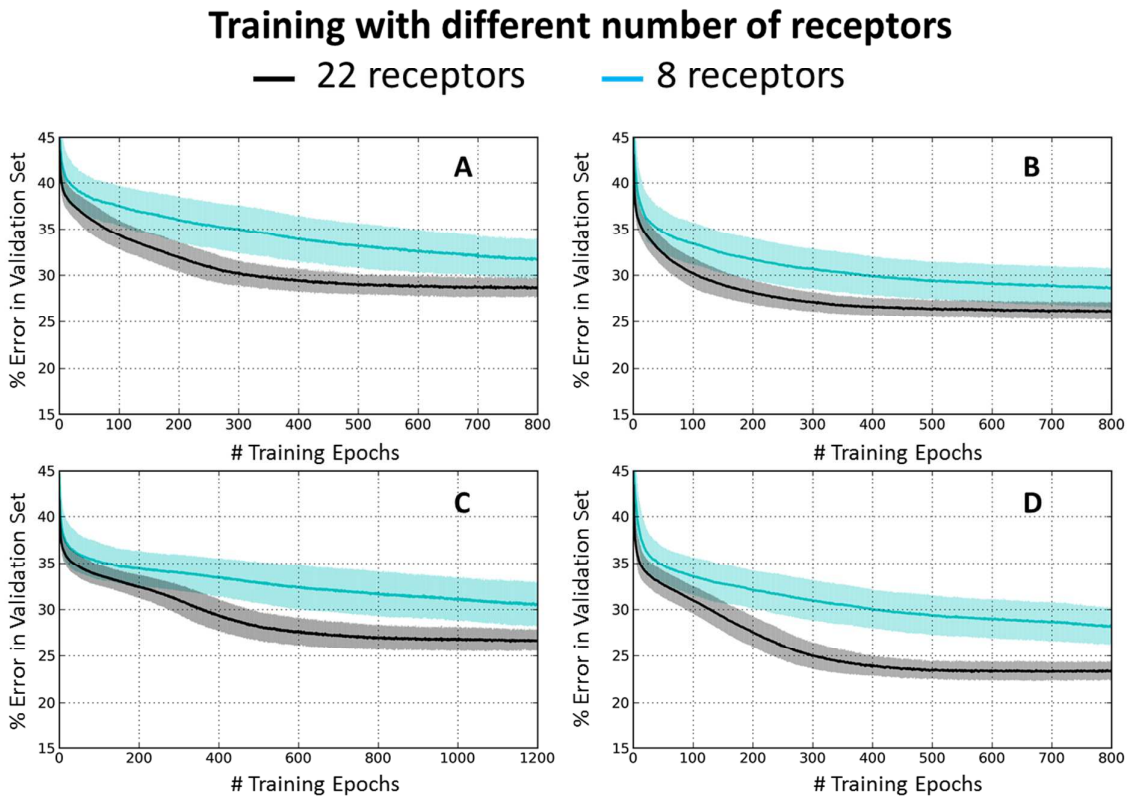



**Figure S5**. Number training receptors. The error on evaluating the Validation Set averaged over forty receptors for Autodock4 (A), AutodockVina (B), RosettaLigand (C), and ALL scheme (D). The reduction of training information yields highly noisy data, although longer training may produce slightly better results.

# 6. References


Beck, P., C. Dubiella, et al. (2012). "Covalent and non-covalent reversible proteasome inhibition." Biological Chemistry **393**(10): 1101-1120.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Singapore, Springer.

Borissenko, L. and M. Groll (2007). "20S Proteasome and Its Inhibitors: Crystallographic Knowledge for Drug Development." Chemical Reviews **107**(3): 687-717.

Broccatelli, F. and N. Brown (2014). "Best of Both Worlds: On the Complementarity of Ligand-Based and Structure-Based Virtual Screening." Journal of Chemical Information and Modeling.

Cheng, T., Q. Li, et al. (2012). "Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review." The AAPS Journal **14**(1): 133-141.

Combs, S. A., S. L. DeLuca, et al. (2013). "Small-molecule ligand docking into comparative models with Rosetta." Nat. Protocols **8**(7): 1277-1298.

Cosconati, S., S. Forli, et al. (2010). "Virtual screening with AutoDock: theory and practice." Expert Opinion on Drug Discovery **5**(6): 597-607.

Craik, D. J., D. P. Fairlie, et al. (2013). "The Future of Peptide-based Drugs." Chemical Biology & Drug Design **81**(1): 136-147.

D. van der Spoel, E. Lindahl, et al. (2010). "Gromacs User Manual version 4.6-beta1."

Damm-Ganamet, K. L., R. D. Smith, et al. (2013). "CSAR Benchmark Exercise 2011–2012: Evaluation of Results from Docking and Relative Ranking of Blinded Congeneric Series." Journal of Chemical Information and Modeling **53**(8): 1853-1870.

Drwal, M. N. and R. Griffith (2013). "Combination of ligand- and structure-based methods in virtual screening." Drug Discovery Today: Technologies **10**(3): e395-e401.

Fukunishi, Y. (2010). "Post Processing of Protein-Compound Docking for Fragment-Based Drug Discovery (FBDD): In-Silico Structure-Based Drug Screening and Ligand-Binding Pose Prediction." Current Topics in Medicinal Chemistry **10**(6): 680-694.

Gallastegui de la Rosa, N. (2012). Characterisation and optimisation of 20S proteasome inhibitors Doktors der Naturwissenschaften (Dr. rer. nat.) Dissertation Technische Universität München.

Gallastegui, N., P. Beck, et al. (2012). "Hydroxyureas as Noncovalent Proteasome Inhibitors." Angewandte Chemie International Edition **51**(1): 247-249.

García, A. E. (1992). "Large-amplitude nonlinear motions in proteins." Physical Review Letters **68**(17): 2696-2699.

Gastreich, M., M. Lilienthal, et al. (2006). "Ultrafast de novo docking combining pharmacophores and combinatorics." Journal of Computer-Aided Molecular Design **20**(12): 717-734.

Glickman, M. H. and A. Ciechanover (2002). "The Ubiquitin-Proteasome Proteolytic Pathway: Destruction for the Sake of Construction." Physiological Reviews **82**(2): 373-428.

Groll, M., M. Bochtler, et al. (2005). "Molecular Machines for Protein Degradation." ChemBioChem **6**(2): 222-256.

Groll, M., N. Gallastegui, et al. (2010). "20S Proteasome Inhibition: Designing Noncovalent Linear Peptide Mimics of the Natural Product TMC-95A." ChemMedChem **5**(10): 1701-1705.

Groll, M. and R. Huber (2004). "Inhibitors of the eukaryotic 20S proteasome core particle: a structural approach." Biochimica et Biophysica Acta (BBA) - Molecular Cell Research **1695**(1–3): 33-44.

Hershko, A. and A. Ciechanover (1998). "THE UBIQUITIN SYSTEM." Annual Review of Biochemistry **67**(1): 425-479.

Huber, E. M. and M. Groll (2012). "Inhibitors for the Immuno- and Constitutive Proteasome: Current and Future Trends in Drug Development." Angewandte Chemie International Edition **51**(35): 8708-8720.

Huey, R., G. M. Morris, et al. (2007). "A semiempirical free energy force field with charge-based desolvation." Journal of Computational Chemistry **28**(6): 1145-1152.

Jalali-Heravi, M. (2008). Neural networks in analytical chemistry. Artificial Neural Networks methods and Applications. D. J. Livingstone. Sandown, UK, Humana Press. **458:** 81-121.

Khanna, I. (2012). "Drug discovery in pharmaceutical industry: productivity challenges and trends." Drug Discovery Today **17**(19–20): 1088-1102.

Kikuchi, J., N. Shibayama, et al. (2013). "Homopiperazine Derivatives as a Novel Class of Proteasome Inhibitors with a Unique Mode of Proteasome Binding." PLoS ONE **8**(4): e60649.

Kikuchi, J., S. Yamada, et al. (2013). "The Novel Orally Active Proteasome Inhibitor K-7174 Exerts Anti-myeloma Activity in Vitro and in Vivo by Down-regulating the Expression of Class I Histone Deacetylases." Journal of Biological Chemistry **288**(35): 25593-25602.

Kisselev, A. F. and A. L. Goldberg (2001). "Proteasome inhibitors: from research tools to drug candidates." Chemistry & Biology **8**(8): 739-758.

Kisselev, Alexei F., W. A. van der Linden, et al. (2012). "Proteasome Inhibitors: An Expanding Army Attacking a Unique Target." Chemistry & Biology **19**(1): 99-115.

Kitao, A., F. Hirata, et al. (1991). "The effects of solvent on the conformation and the collective motions of protein: Normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum." Chemical Physics **158**(2–3): 447-472.

Klebe, G. (2006). "Virtual ligand screening: strategies, perspectives and limitations." Drug Discovery Today **11**(13–14): 580-594.

Koguchi, K. J., Nishio M, Takahashi K, Okuda T, Ohnuki T, Komatsubara S. (2000). "TMC-95A, B, C, and D, Novel Proteasome Inhibitors Produced by Apiospora montagnei Sacc. TC 1093 Taxonomy, Production, Isolation, and Biological Activities." The Journal of Antibiotics **53**(2): 105-109.

Lipinski, C. A., F. Lombardo, et al. (2001). "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings." Advanced Drug Delivery Reviews **46**(1–3): 3-26.

Meng, L., R. Mohan, et al. (1999). "Epoxomicin, a potent and selective proteasome inhibitor, exhibits in vivo antiinflammatory activity." Proceedings of the National Academy of Sciences **96**(18): 10403-10408.

Morris, G. M., R. Huey, et al. (2009). "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility." Journal of Computational Chemistry **30**(16): 2785-2791.

Moustakas, D., P. T. Lang, et al. (2006). "Development and validation of a modular, extensible docking program: DOCK 5." Journal of Computer-Aided Molecular Design **20**(10-11): 601-619.

Neves, M. C., M. Totrov, et al. (2012). "Docking and scoring with ICM: the benchmarking results and strategies for improvement." Journal of Computer-Aided Molecular Design **26**(6): 675-686.

Nussbaum, A. K., T. P. Dick, et al. (1998). "Cleavage motifs of the yeast 20S proteasome β subunits deduced from digests of enolase 1." Proceedings of the National Academy of Sciences **95**(21): 12504-12509.

Parlati, F., S. J. Lee, et al. (2009). "Carfilzomib can induce tumor cell death through selective inhibition of the chymotrypsin-like activity of the proteasome." Blood **114**(16): 3439-3447.

Pellom ST Jr., A. S. (2012). "Development of Proteasome Inhibitors as Therapeutic Drug." J Clin Cell Immunol **S5:5**.

Pickart, C. M. (2004). "Back to the Future with Ubiquitin." Cell **116**(2): 181-190.

Ruschak, A. M., M. Slassi, et al. (2011). "Novel Proteasome Inhibitors to Overcome Bortezomib Resistance." Journal of the National Cancer Institute **103**(13): 1007-1017.

Rydzewski, R. M., L. Burrill, et al. (2006). "Optimization of Subsite Binding to the β5 Subunit of the Human 20S Proteasome Using Vinyl Sulfones and 2-Keto-1,3,4-oxadiazoles: Syntheses and Cellular Properties of Potent, Selective Proteasome Inhibitors." Journal of Medicinal Chemistry **49**(10): 2953-2968.

Scior, T., A. Bender, et al. (2012). "Recognizing Pitfalls in Virtual Screening: A Critical Review." Journal of Chemical Information and Modeling **52**(4): 867-881.

Sijts, A. J. A. M., T. Ruppert, et al. (2000). "Efficient Generation of a Hepatitis B Virus Cytotoxic T Lymphocyte Epitope Requires the Structural Features of Immunoproteasomes." The Journal of Experimental Medicine **191**(3): 503-514.

Sinko, W., S. Lindert, et al. (2013). "Accounting for Receptor Flexibility and Enhanced Sampling Methods in Computer-Aided Drug Design." Chemical Biology & Drug Design **81**(1): 41-49.

Steele, J. M. (2013). "Carfilzomib: A new proteasome inhibitor for relapsed or refractory multiple myeloma." Journal of Oncology Pharmacy Practice **19**(4): 348-354.

Stein, M. L., H. Cui, et al. (2014). "Systematic Comparison of Peptidic Proteasome Inhibitors Highlights the α-Ketoamide Electrophile as an Auspicious Reversible Lead Motif." Angewandte Chemie International Edition **53**(6): 1679-1683.

Strehl, B., K. Textoris-Taube, et al. (2008). "Antitopes Define Preferential Proteasomal Cleavage Site Usage." Journal of Biological Chemistry **283**(26): 17891-17897.

Tanrikulu, Y., B. Krüger, et al. (2013). "The holistic integration of virtual screening in drug discovery." Drug Discovery Today **18**(7–8): 358-364.

Trott, O. and A. J. Olson (2010). "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading." Journal of Computational Chemistry **31**(2): 455-461.

Verdonk, M. L., J. C. Cole, et al. (2003). "Improved protein–ligand docking using GOLD." Proteins: Structure, Function, and Bioinformatics **52**(4): 609-623.

Xuan-Yu Meng, H.-X. Z., Mihaly Mezei and Meng Cui (2011). "Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery." Current Computer Aided-Drug Design **7**(2): 146-157.

Zhou, S., E. Chan, et al. (2005). "Drug Bioactivation Covalent Binding to Target Proteins and Toxicity Relevance." Drug Metabolism Reviews **37**(1): 41-213.

Zsoldos, Z., D. Reid, et al. (2007). "eHiTS: A new fast, exhaustive flexible ligand docking system." Journal of Molecular Graphics and Modelling **26**(1): 198-212.