Approximate and efficient prediction of entropy changes upon complex formation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf
vorgelegt von

Ido Yitshak Ben-Shalom

aus Jerusalem
Düsseldorf, July 2016

Aus dem Institut für Medizinische Chemie und Pharmazie

der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der

Mathematisch-Naturwissenschaftlichen Fakultät der

Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. Holger Gohlke

Korreferent: Prof. Dr. Gunnar Schröder

Tag der mündlichen Prüfung:

Eidesstattliche Erklärung

lch versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist.

Diese Dissertation wurde in der vorgelegten oder einer ähnlichen Form noch bei keiner anderen Institution eingereicht und es wurden bisher keine erfolglosen Promotionsversuche von mir unternommen.

Düsseldorf, im July 2016:

Acknowledgements

First and foremost I thank my supervisor Prof. Dr. Holger Gohlke for setting the expectation bar as high as he did, with his support, wisdom, and enthusiasm. Mostly I want to thank for enhancing my understanding and appreciation for sciences and a very exciting doctorate culminating in this thesis.

To Dr. Stefania Pfeiffer-Marek, our collaboration partner from Sanofi-Aventis, I thank her for support and fruitful discussions. It was a pleasure to work with her both on a personal and professional level.

I thank Christian Hanke, Dr. Alexander Metz, Dr. Christopher Pfleger, and Michele Bonus who always provided assistance and a helping hand. To Hannes Kopitz I thank for the preliminary work on the vibrational entropy project. To my other lab members I thank for being not just co-workers but also very good friends. They always lent an ear, provided fruitful ideas, often offered a different perspective, and made coming to work a joy.

Finally I would like to thank my family, for their support on every step of the way.

To my Family

Table of contents

Abbreviations

Abl	Abelson murine leukemia
AIDS	Acquired immunodeficiency syndrome
ADME	Absorption, distribution, metabolism, and excretion
ATP	Adenosine triphosphate
Bcr	Breakpoint cluster region
BEERT	Binding Entropy Estimation for Rotation and Translation
DDD	Distance-dependent dielectric
DNA	Deoxyribonucleic acid
EGF	Epidermal growth factor
EM	Electron microscopy
ER	Endoplasmatic reticulum
FBDD	Fragment-based drug design
FIRST	Floppy Inclusion and Rigid Substructure Topography
FIXa	Factor IXa
FVa	Factor Va
FVIIa	Factor VIIa
FX	Factor X
FXa	Factor Xa
GA	Geldanamycine
GB	Generalized Born
GDS	Guanine nucleotide dissociation stimulator
GHKL	Gyrase, Hsp90, Histidine kinase, and MutL
Gla	γ-Carboxyglutamic acid
Grp94	Glucose-regulated protein 94
HAC	Heavy atom count
HIV	Human immunodeficiency virus
HSP	Heat shock protein
Hsp90	Heat shock protein 90

HTS	High-throughput screening
LE	Ligand efficiency
LogP	Logarithm of octanol/water partition coefficient
MD	Molecular dynamics
MM	Molecular mechanics
MM-GBSA	Molecular mechanics generalized Born surface area
MM-PBSA	Molecular mechanics Poisson-Boltzmann surface area
NMA	Normal mode analysis
NMR	Nuclear magnetic resonance
PARSE	Parameters for solvation (free) energy
PB	Poisson-Boltzmann
PDB	Protein Data Bank
PSA	Polar surface area
QSAR	Quantitative structure-activity relationship
QHA	Quasi-harmonic analysis
RD	Radicicol
RMSD	Root-mean-square deviation
RMSG	Root-mean-square gradient
RNA	Ribonucleic acid
RRHO	Rigid rotor harmonic oscillator
SAR	Structure-activity relationship
SASA	Solvent-accessible surface area
TNF	Tumor necrosis factor
TRAP-1	TNF receptor-associated protein 1
TPR	Tetratricopeptide repeat

Zusammenfassung

Alle biologischen Prozesse, einschließlich Metabolismus, Signaltransduktion, Pathogenese und Arzneimittelwirkung, sind abhängig von der spezifischen molekularen Erkennung und der Bindungsaffinität zwischen Liganden und deren makromolekularen Zielstrukturen. Daher erfordert das rationale Design potenter Liganden eine genaue Kenntnis der Bindungaffinität. Ein Maß für die Bindungaffinität ist die freie Bindungsenergie (ΔG), die aus enthalpischen (ΔH) und entropischen Beiträgen (ΔS) zusammengesetzt ist. Die Vorhersage von ΔS ist jedoch schwierig, weil dazu die Kenntniss aller thermodynamischen Zustände benötigt wird, weshalb dieser Term bei Vorhersagen von ΔG oft vernachlässigt wird. Diese Vereinfachung führt aber oftmals zu schlechten Vorhersagen der Bindungsaffinität. Ein Verständnis von ΔS ist somit von essenzieller Bedeutung für die Untersuchungen von Bindungsaffinität zwischen Ligand und Zielstruktur im Zuge der Leitstrukturidentifizierung und -optimierung.

Daher habe ich im ersten Projekt die Methode BEERT (Binding Entropy Estimation for Rotation and Translation) entwickelt, durch die Rotations- und Translationsentropieänderungen ($\Delta S_{\text{config.}}$) bei der Ligandbindung effizient berechnet werden. Diese Entropieänderungen resultieren aus der unterschiedlich starken Einschränkung von Rotations- und Translationsfreiheitsgraden bei der Bindung und werden durch BEERT in drei Schritten approximiert. Zuerst wird ein Ensemble gedockter Bindeposen mit AutoDock generiert, um die Bindungsenergielandschaft repräsentativ durchzumustern. Durch nachfolgende Clusteranalyse bezüglich der intermolekularen Interaktionsmuster dieser Bindeposen wird die Population zugänglicher Mikrozustände ermittelt. Zusammen mit der Kenntnis der Tiefe und Breite der zugehörigen Energieminima erlaubt dies die Vorhersage von $\Delta S_{\text{config.}}$. Dazu wird $\Delta S_{\text{config.}}$ aus den Einschränkungen der Rotations- und Translationsvolumina bei der Bindung abgeleitet. Insgesamt schätzt BEERT $\Delta S_{\text{config.}}$ durch Modellierung der natürlich vorkommenden Reduktion von Mikrozuständen ab, die zu einer Änderung der Rotations- und Translationsentropie führt.

Zur Validierung korrelierte ich experimentelle Bindungsenergien ($\Delta G_{\text{exp.}}$) mit berechneten Bindungsenergien ($\Delta G_{\text{calc.}}$) für HIV-1 Protease-, Faktor Xa- und Hsp90-Inhibitoren. $\Delta G_{\text{calc.}}$ berechnete ich durch effektive Bindungsenergien ($\Delta G_{\text{eff.}}$) aus MM-PBSA Rechnungen kombiniert mit $T\,\Delta S_{\text{config.}}$ Werten, bestimmt mit BEERT. Die Korrelationen von $\Delta G_{\text{calc.}}$ und $\Delta G_{\text{exp.}}$ aller Datensätze sind signifikant ($R^2 = 0{,}54 - 0{,}72$; $p < 0{,}001$) und deutlich besser als

$\Delta G_{eff.}$ allein (R^2 = 0,01 - 0,38) oder in Kombination mit Abschätzungen von $T\Delta S$ anhand der Zahl rotierbarer Bindungen (R^2 = 0,01 - 0,42). Diese Korrelationen sind auch robust hinsichtlich einer Leave-One-Out-Kreuzvalidierung (q^2 = 0,34 - 0,66; $p < 0,05$). Insgesamt ermöglicht BEERT die effiziente und genaue Berechnung von $\Delta S_{config.}$ und Bindungsenergien zur Leitstrukturoptimierung.

In einem zweiten Projekt verglich ich zwei Ansätze zur Berechnung der Änderung der Schwingungsentropie ($\Delta S_{vib.}$). $\Delta S_{vib.}$ entsteht aus der Variation der Schwingungsfreiheitsgrade des Protein-Ligand-Komplexes und seiner ungebundenen Komponenten. Normalmodenanalyse (normal mode analysis, NMA) stellt die am weitesten verbreitete Methode zur Berechnung von $\Delta S_{vib.}$ dar. Die Berechnung mittels NMA ist allerdings ein sehr zeitintensives Verfahren und benötigt mehrere Stunden für einen einzelnen Protein-Ligand-Komplex. Daher wurde das Programm „floppy inclusion and rigid substructure topography" (FIRST) erweitert, um $\Delta S_{vib.}$ für einen einzelnen Protein-Ligand-Komplex in nur wenigen Minuten berechnen zu können. Zur Validierung von FIRST Entropie verglich ich die mittels FIRST Entropie berechnete $\Delta S_{vib.}$ ($\Delta S_{vib., FIRST}$) mit mit NMA berechneter $\Delta S_{vib.}$ ($\Delta S_{vib., NMA}$) für HIV-1 Protease-, Faktor Xa- und Hsp90-Protein-Inhibitor-Komplexe. Für Faktor Xa-Komplexe korreliert $\Delta S_{vib., FIRST}$ und $\Delta S_{vib., NMA}$ signifikant (r^2 = 0.46; $p < 0.001$). Im Falle von HIV-1 Protease- und Hsp90-Inhibitoren korreliert $\Delta S_{vib., FIRST}$ hingegen nicht mit $\Delta S_{vib., NMA}$. Die Inhibitoren der HIV-1 Protease und des Hsp90 sind strukturell sehr ähnlich. Berechnete $\Delta S_{vib., NMA}$ liegen in einem Bereich von 20 cal mol^{-1} K^{-1} und befinden sich damit fast im Bereich der Berechnungsgenauigkeit von ~17 cal mol^{-1} K^{-1}. Faktor Xa-Inhibitoren, hingegen, zeigen eine breitere strukturelle Diversität, so dass auch $\Delta S_{vib., NMA}$ Werte einen breiteren Bereich von 30 cal mol^{-1} K^{-1} einnehmen, außerhalb der Berechnungsgenauigkeit. Insgesamt ist FIRST Entropie ein neues und effektives Programm zur Berechnung von $\Delta S_{vib.}$ im Zuge der Leitstrukturoptimierung.

Abstract

All biological processes, including metabolic pathways and signal transduction pathways, depend on the specific molecular recognition and binding affinity between ligands and their macromolecular targets. A measure for the binding affinity is the binding energy (ΔG), which is composed of enthalpic (ΔH) and entropic (ΔS) contributions. The rational design of potent ligands requires an accurate knowledge of the binding free energy. ΔS is difficult to predict as it requires a full understanding of all possible states of the system and is thus often neglected. This neglection often leads to inadequate ΔG predictions. Understanding ΔS is therefore crucial for understanding the ΔG between a ligand and its target macromolecule in the lead discovery process.

In the first project, I developed BEERT (Binding Entropy Estimation for Rotation and Translation), a method for fast predictions of translational and rotational entropy contributions of the ligand ($\Delta S_{\text{config.}}$) to the binding free energy. Differences in $\Delta S_{\text{config.}}$ originate from the varying extent of restricting translational and rotational motion upon binding, which is approximated by BEERT in three steps. First, an ensemble of docked ligand binding poses is generated, which accounts for their bound energy landscape. Subsequently, these poses are clustered by intermolecular interaction pattern to distinguish between accessible microstates, whose populations reflect the depths and widths of the underlying energy minima. Finally, $\Delta S_{\text{config.}}$ is derived from the comparison between the translational and rotational volumes in these microstates with the unbound ligand state. Altogether, BEERT estimates $\Delta S_{\text{config.}}$ by modeling the naturally occurring reduction of accessible microstates that leads to changes in translational and rotational entropy.

To validate BEERT, I fitted predicted $\Delta S_{\text{config.}}$ and MM-PBSA effective energies ($\Delta G_{\text{eff.}}$) to experimental binding affinities of HIV-1 protease, Factor Xa (FXa), and Heat shock protein 90 (Hsp90) inhibitors using multiple linear regression. The multiple linear regression combines the estimate of the translational and rotational entropy from BEERT with the enthalpy and solvation free energy from MM-PBSA. For all datasets, the obtained correlations are highly significant ($R^2 = 0.54 – 0.72$, $p < 0.001$), markedly improved compared to MM-PBSA results alone ($R^2 = 0.01 – 0.38$) or combined with $-T\Delta S_{\text{config.}}$ estimates based on the number of rotatable bonds ($R^2 = 0.01 – 0.42$), and robust in a leave-one-out cross-validation ($q^2 = 0.34 – 0.66$, $p < 0.05$). In summary, I could show that BEERT allows an

efficient calculation of the change in translational and rotational entropy and the binding free energy which can be used for the ligand optimization process.

In the second project, I calculated changes in the vibrational entropy ($\Delta S_{vib.}$) upon ligand binding to proteins. $\Delta S_{vib.}$ originate from the varying vibrational degrees of freedom in a protein-ligand complex with the unbound partners. $\Delta S_{vib.}$ accounts for about 50 cal mol^{-1} K^{-1} (equals to ~15 kcal mol^{-1} at 298 K) of the binding free energy. [1-4] $\Delta S_{vib.}$ is most commonly estimated using normal mode analysis (NMA). [5-7] However, NMA is a very time consuming method, takes up to several hours for a single structure, and for a single analysis usually about 500 structures are needed to achieve sufficient precision. Hence, our working group introduced a computationally highly efficient approximation of changes in the vibrational entropy ($\Delta S_{vib.}$) upon binding to biomolecules based on rigidity theory. In constraint network representations of the binding partners, $\Delta S_{vib.}$ is estimated from changes in the variation of the number of low (i.e., zero) frequency modes with respect to variations in the networks' coordination number. Compared to $\Delta S_{vib.}$ computed by NMA as a gold standard, our approach yields significant and good to fair correlations for datasets of protein-protein complexes ($r^2 = 0.80$; $p < 0.001$) as well as in alanine scanning ($r^2 = 0.51$; $p < 0.001$). On my aforementioned datasets, HIV-1 protease, FXa, and Hsp90 this resulted in significant correlation for the FXa dataset ($r^2 = 0.46$; $p < 0.001$) and poor correlations for the HIV-1 protease and Hsp90 datasets ($r^2 < 0.11$). The reason for the poor correlation is the width of the distribution of $\Delta S_{vib.}$ calculated using NMA, which is very similar in magnitude to the average standard deviation of the computed $\Delta S_{vib.}$. Therefore, according to Kramer *et al.*, [8] the maximum possible squared Pearson correlation coefficient (r^2_{max}) on these datasets vanishes. On an additional protein-small molecule dataset, the trypsin dataset, this resulted in a fair and significant correlation ($r^2 = 0.40$; $p < 0.001$).

1 Introduction

Molecular recognition is essential for all biological processes, e.g., in metabolic pathways or signal transduction pathways, and, therefore, has a great biological importance. [9-13] In the drug development process, small molecules are designed to alter biological processes. The search for drug candidates can be assisted via virtual screening. Virtual screening is a computational method in which large compound libraries are screened against a protein target, predicting the binding free energy searching for potent inhibitors. Virtual screening is the computational equivalent of experimental high-throughput screening (HTS). [14-16]

Scoring functions used in the field of drug design for binding free energy ($\Delta G_{bind.}$) predictions show a pronounced dependence of the predicted $\Delta G_{bind.}$ on the size of the ligands. This results in larger ligands having the ability to make more interactions, receiving better scores. [17-20] A possible correction is to normalize the score according to the number of heavy atoms. [20] The preference towards high molecular weight ligands contradicts the observation of enthalpy-entropy compensation. [9-13] Enthalpy-entropy compensation postulates that larger ligands, with more degrees of freedom, make more interactions, and are therefore more restricted, and lose more entropy upon binding. [13, 21, 22]

Upon ligand binding to a protein, the protein and the ligand lose entropy as a result of the decreased rotational, translational, conformational, and vibrational motion. [23-25] In contrast entropy is gained from the release of restricted solvent molecules from the surface of the protein and the ligand to the solution. This is accompanied by an enthalpic component, resulting from the loss of interactions between the water molecules and the solute molecules. Therefore, this term is called solvation free energy ($\Delta G_{sol.}$).

I focused on the configurational entropy change, i.e., the restriction of the overall translational, rotational, and vibrational motions of the protein and the ligand upon binding. Solvation free energy is successfully incorporated in many scoring functions and is therefore not addressed here. [26-33] The configurational entropy change is usually separated into two terms. The first accounts for the translational and rotational entropy change. The second accounts for the internal entropy change. [1] These terms correspond to both projects I worked on.

In the first project, I investigated the translational and rotational entropy change that accompanies the ligand binding to a protein. Usually, it is assumed that in its bound state the ligand loses its translational and rotational motion completely. However it was shown that some degrees of translational and rotational motions persist even for the bound ligand. [34-39] In order to estimate the change in the degrees of freedom of the ligand between its bound and unbound state, the residual translational and rotational motion of the ligand in its bound state is determined and compared to the unbound state by considering the relative motion between the protein and the ligand as rigid bodies following the work of Steinberg and Scheraga. [40] This decouples the internal vibrational motions of the body from the rotational and translational motion of the rigid body and allows to estimate the translational, rotational, and vibrational motion separately. [41, 421, 43, 44] Whereas the unbound translational and rotational motions are straightforward to calculate, i.e., considering free rotation and translation, many methods were developed for estimating the translational and rotational motions of the bound space.

Carlsson and Aqvist used molecular dynamics simulations in order to generate predicted binding geometries of the ligand (docking poses) in its bound state, [45] which were then used to predict rotational and translational entropy changes ($\Delta S_{R/T}$). They calculated $\Delta S_{R/T}$ only for molecular fragments (e.g., methane, ethane, benzene) and not drug-like ligands. [46] The group of Gilson also used molecular dynamics simulations for predicting the bound ligand space. Their method was tested in two cases. The first test was a supramolecular host-guest systems. [47, 48] Supermolecular host-guest systems are small sized structures compared to protein-ligand systems. They are often used as models for biomolecules, but have limited conformational flexibility. Therefore, for host-guest systems, determining the conformational space is more feasible. [2, 47] The second test was a single protein-ligand complex, where the entropy change was numerically estimated but the value was not used for $\Delta G_{bind.}$ prediction. [43] Ruvinsky and Kozintsev used docking to generate bound ligand poses. The docked poses were clustered by structural similarity, whereat each cluster is assumed to represent an energy minimum. These clusters were then used successfully to identify near-native ligand binding poses. [49-51] However, there was no improvement in the $\Delta G_{bind.}$ predictions. [52] Heretofore, no method used $\Delta S_{R/T}$ predictions for $\Delta G_{bind.}$ predictions.

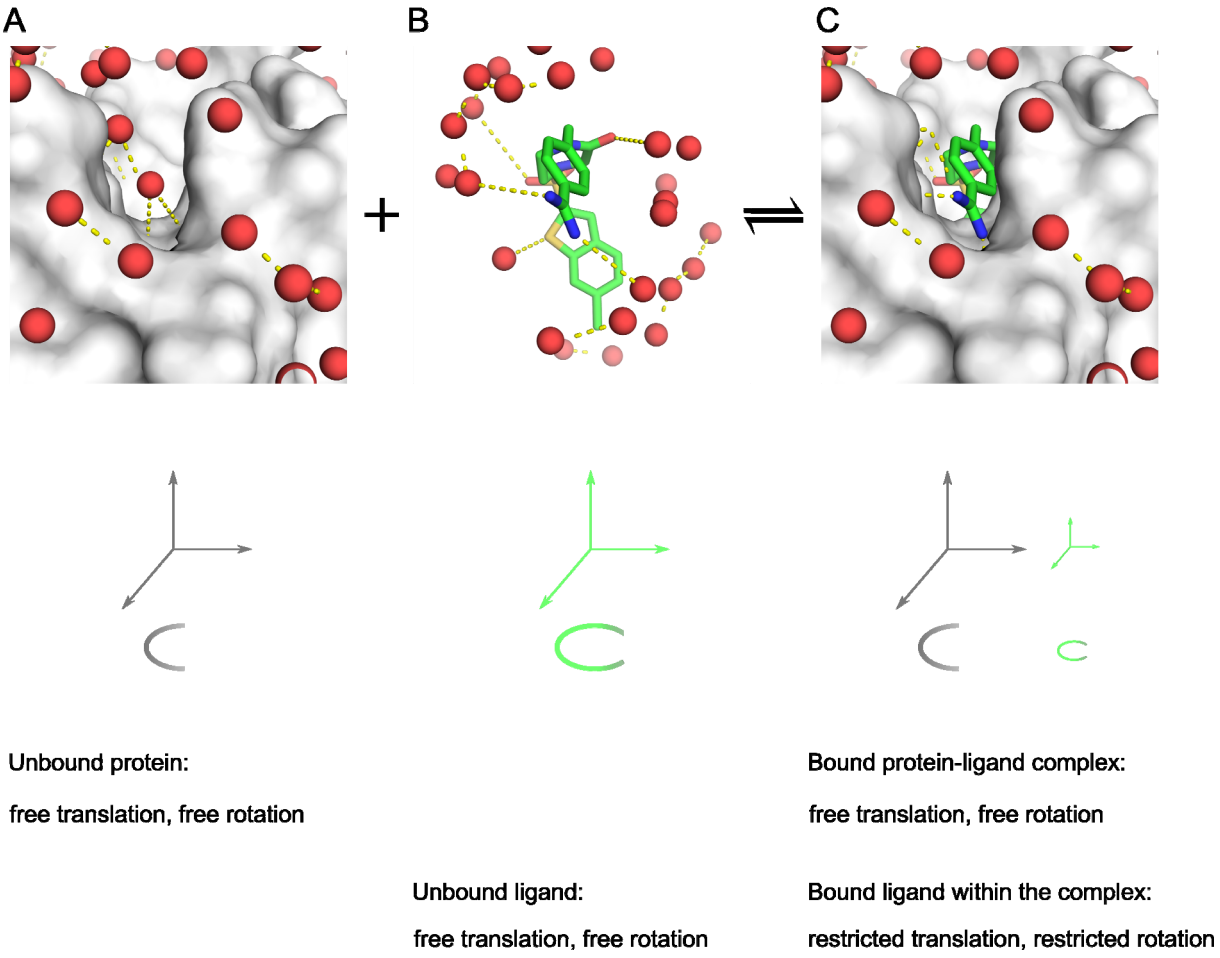

Figure 1: (A) The solvated structure of a protein, which maintains free rotation and translation. (B) The solvated structure of the ligand, which also maintains free rotation and translation. (C) The solvated structure of the protein-ligand complex. The complex as a whole maintains free rotation and translation, whereas the ligand has a restricted translational and rotational freedom. The water molecules from the buried surface area of the protein and the ligand are released to the bulk upon binding, resulting in the solvation free energy.

In the second project I investigated the change in the flexibility of the protein-ligand complex that results from the ligand binding. This is governed by two main terms, the conformational and vibrational term. Conformational entropy is associated with the large scale conformational change, i.e., the different energy minima. Vibrational entropy is associated with local fluctuations around a defined structure, i.e., the width of each energy minimum (see Figure 2). [53-61]

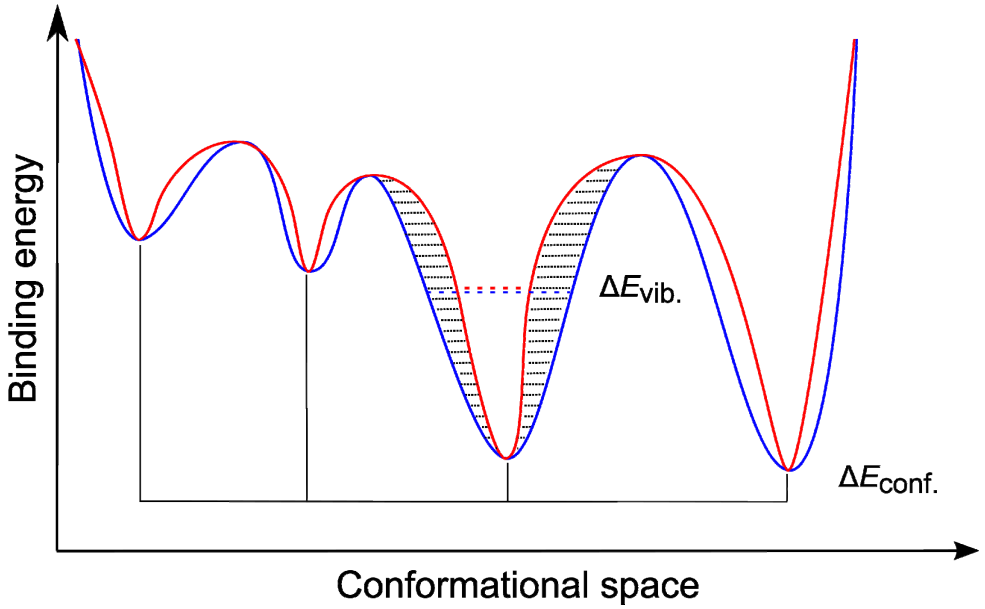

Figure 2: The energy landscape of the bound and unbound states of the ligands is presented in red and blue respectively. The conformational entropy is associated with the change in the number of accessible conformations, which is shown to be neglectable. [43] The vibrational entropy is associated with the change of the width of the energy wells (an example is shown as the area marked in dots, between the bound (red) and unbound (blue) energy wells). [43]

2 Background

In the following chapter I will review the available computational methods for studying the entropic effect in molecular recognition and highlight its significance in modern drug discovery. I will further point out the importance of an accurate calculation of the entropic contribution to the total binding free energy upon ligand binding, which is most often neglected in the binding free energy predictions. Finally, I will describe the datasets used for validation: HIV-1 protease, Factor Xa, and Heat shock protein 90.

2.1 Drug discovery process

Drug discovery aims at the development of novel active molecules for the treatment of diseases or pathophysiological conditions. The cost of the development from the initial idea to the marketed drug is estimated at over 1 billion US dollars and can last up to 15 years. [62] This process faces many challenges. The costs continue to grow, whereas the success rate (approval of a drug) stagnates. This indicates the need of new approaches and technologies in the drug discovery process. Implementation of new technologies in the preclinical phase can reduce the overall costs. [63] In my thesis I focus on protein structures as they are known to be important as drug targets. [64-69] In the following chapter I will outline classical experimental methods, as well as computer based approaches commonly used in the drug discovery process.

2.2 Enthalpy and entropy in drug design

In drug discovery it is desirable to improve the binding affinity of a ligand to a protein. This is often attempted by adding potential substituents to the ligand to maximize the interaction surface and increase the enthalpy gain. However, with increasing ligand size, the number of degrees of freedom of the ligand also increases. Therefore, a larger ligand suffers from a larger number of restrictions upon binding, which imposes a larger entropic penalty on the system. This phenomenon is referred to as enthalpy-entropy compensation. [9-13, 21, 22, 70] Though highly correlated, considering enthalpy and entropy alone were shown to be insufficient for binding energy prediction. [11] As a consequence, both the enthalpy and the entropy terms must be taken into account in accurate binding energy predictions. The prediction of the enthalpy and entropy change requires knowledge of the accurate three-dimensional structure of the protein and the ligand.

2.3 Structure-based drug design

Structure-based drug design uses the three-dimensional structure of a biological target as a starting point for the rational design of drugs. [71, 72] Structure-based drug design examines three-dimensional aspects of compatibility between a ligand and its target protein. [73] The structure of the protein-ligand complex is usually determined by X-ray crystallography, NMR, Cryo-EM, or homology modeling. As this structure is the basis for all further studies, the quality of the structure is highly important. [16]

2.4 Computational methods in structure-based drug design

Computational methods for structure-based drug design reduce the costs and time required in the drug development process. [74] The basis for this is the analysis of geometrical and energetical complementarity of the ligand inside the binding pocket of the protein for understanding their complex formation. [75] In the cases where the binding pocket is not known there is a need to identify and characterize it. Therefore, many algorithms for pocket identification were developed. [76-78] An important property of binding pockets is, though mostly buried, that they contain polar atoms. [79] These polar atoms can act as anchoring sites for a ligand, and can contribute significantly to the total binding energy. [80, 81] Predicting the biophysical properties of a binding site allows quantifying the maximal affinity achievable by a ligand, and can help estimate the ligand's probability to become a drug. [82] Correctly identified binding pockets can be further used in molecular docking.

2.4.1 Molecular docking

Molecular docking is a method used in computer-aided drug design in order to match a candidate ligand to a target protein binding site by detecting favorable and unfavorable interactions between a ligand and a protein. Docking seeks to identify the most favorable position of the ligand inside the binding pocket. [83] It offers a computational, relatively fast alternative for experimental techniques, allowing to filter large libraries of compounds into focused libraries that can be further experimentally tested. [84]

Docking explores different ligand orientations and conformations inside the binding pocket of the protein. [73, 85] The important factors for binding are: geometric compatibility, electrostatic compatibility, low-energy conformation of the ligand and the protein, hydrophobic interactions, and stacking forces. [86] The binding affinity is then estimated using the

interactions between the ligand, the target protein, and the solvent. The docking process accounts for the many degrees of freedom of the ligand. Six degrees of freedom result from the translational and rotational mobility of the ligand relative to the protein, in addition to $3n$-6 (n is the number of atoms) internal degrees of freedom. Most docking programs consider the protein as rigid, and account only for the flexibility of the ligand. [87] In order to deal with the large number of degrees of freedom, every docking program must include an efficient conformational search algorithm. [88] Various search algorithms are applied in different docking programs, including Monte Carlo, Genetic Algorithm, Lamarckian Genetic Algorithm, Neural Networks, Iterative Stochastic Elimination, and Simulated Annealing.

Molecular docking has two main aims. The first one is to identify the three-dimensional configuration of the ligand inside the binding pocket. The second is to score the quality of the binding mode. [83] The determination of the ligand configuration can be validated by comparing docked ligand poses to experimentally determined ligand structures. [89] When considering a 2 Å root-mean-square deviation (RMSD) of the atomic distances from the crystal structure as acceptable, docking programs are able to predict up to 70% of the experimental binding poses properly. Some docking programs show better results on specific protein types: e.g., GOLD performs better with a more hydrophilic binding site. [90]

2.4.2 Scoring functions

Each docking program has an implemented scoring function. In addition, a scoring function can also be external, and score the quality of a given protein-ligand complex structure. In both cases it aims to quantify the binding affinity of a given protein-ligand binding mode. [91] A scoring function should accurately represent the thermodynamic properties of a protein-ligand interaction. [83] In principle a scoring function usually calculates the binding free energy of a ligand under standard conditions. However, in practice the success of absolute binding free energy predictions is quite low, and at best a ranking of the binding energies of different ligands can be achieved. [87] Scoring functions can be divided into four main classes: knowledge-based, force field-based, empirical-based, and descriptor-based.

2.4.2.1 Knowledge-based scoring functions

Knowledge-based scoring functions use statistically evaluated data from structural databases in order to derive rules on preferred interaction geometries. This approach assumes that atom

types interacting more frequently in certain distances, reflecting favorable interactions. The distance distribution of an interaction (also referred to as a "distance-dependent pseudopotential") represents the energetic character of this interaction, where the more common distances represent an energetically favorable interaction. The mathematical formulation is based on an inverse Boltzmann law, where the "potential of mean force" is calculated from the statistical probability distribution. This can be further used to score the energetics of this interaction in other structures. [91, 92] An advantage of this method is that it does not require a full understanding of the physical character of the interaction, it relies directly on the statistical appearance of the interaction. A disadvantage might be that this not necessarily represents the binding free energy surface, but rather only the crystal structure orientation. [92] Examples for knowledge-based scoring functions are BLEEP, [93, 94] PMF, [18, 95] and SMoG. [96]. Another example is DrugScore, [17, 80] which I used for binding energy predictions.

DrugScore is a knowledge-based scoring function used for predicting binding free energies and identifying binding poses, and was tested on protein-ligand complexes. [17, 80, 97-101] DrugScore uses structural information from experimentally determined protein-ligand complexes. This information was retrieved from the PDB, [102] converted into distance-dependent pair preferences as well as into singlet potentials scaled by the solvent-accessible surface area. Therefore, DrugScore-predicted binding free energies correspond to the enthalpic contributions to the binding energy, and implicitly, the solvation free energy. The translational, rotational, vibrational, and conformational entropy components are neglected in DrugScore.

2.4.2.2 Force field-based scoring functions

Force field-based scoring functions (also referred to as physics-based scoring functions) describe the physical interactions between the protein and the ligand. [103-105] They account for all atom-atom interactions. They sum the bond, angle, dihedral, electrostatic, and van der Waals terms. [106] Force fields depend on empirical parameters, which are derived from physical measurements and quantum mechanical calculations. As the force field is empirically parameterized, it can usually only be applied on molecular systems similar to the ones it was parameterized on. Considering that force field-derived binding energies only account for physical interactions and consequently neglect entropic contributions results in a large size

dependence, as larger ligands make more interactions, and gain better scores. Therefore, addition of an entropic term is necessary. [91] Examples for force field-based scoring functions are AMBER, [105, 107] CHARMM, [108, 109] and OPLS. [110-112] Another example is MM-PB/GBSA, which incorporates a force field-based scoring function (in addition it incorporates solvation free energy). I used MM-PBSA for binding energy predictions. [26, 113, 114]

MM-PBSA (molecular mechanics Poisson-Boltzmann surface area) [23, 26] is a post-processing end-point free energy calculation method for calculating free energies of molecules in solution. [115] Snapshots for the generation of conformational ensembles can either be obtained from a single trajectory of the complex ("single-trajectory approach") or from separate trajectories of the complex, receptor, and ligand ("separate trajectory approach/three-trajectory approach").

The MM-PBSA approach separates the binding free energy ($\Delta G_{\text{bind.}}$) to the different energy components as described in eq. 1. [26, 116]

$$\Delta G_{bind.} = \Delta E_{MM} + \Delta G_{solv.} - T\Delta S \qquad 1$$

In this equation, ΔE_{MM} is the change in the gas-phase molecular mechanics energy, $\Delta G_{\text{solv.}}$ is the change in the solvation free energy, and $T\Delta S$ is the change in the entropy of the solute molecules.

ΔE_{MM} uses a force field-based scoring function as described in eq. 2.

$$\Delta E_{MM} = \Delta E_{Internal} + \Delta E_{Electrostatic} + \Delta E_{vdW} \qquad 2$$

In this equation, $\Delta E_{\text{Internal}}$ is the internal energy change resulting from the bond, angle, and dihedral energies. $\Delta E_{\text{Electrostatic}}$ is the electrostatic energy change, and ΔE_{vdW} is the van der Waals energy change.

$\Delta G_{\text{solv.}}$ is modeled separately according to eq. 3.

$$\Delta G_{solv.} \approx \Delta G_{PB} + \Delta G_{SA} \qquad 3$$

In this equation, ΔG_{PB} is the contribution describing the polar part of the solvation free energy calculated using a Poisson-Boltzmann (PB) model. ΔG_{SA} is the contribution describing the non-polar part of the solvation free energy calculated according to the buried solvent-accessible surface area (SASA). The solute entropy term of eq. 1 ($T\Delta S$) can be calculated using quasi-harmonic analysis or normal mode analysis (NMA). [23, 26] However, this term is difficult to calculate and is often neglected. [26, 117]

2.4.2.3 Empirical scoring functions

Empirical scoring functions estimate the free binding energy as a sum of individual contributions, each representing an important energetic term, e.g., ionic interactions, hydrogen bonds, hydrophobic contacts, or entropic factors. The contribution of each individual term is obtained using a weighting coefficient, determined by regression or statistical analysis of data retrieved from test sets of protein-ligand complexes with solved three-dimensional structure and known experimental binding affinities. [91, 118-120] The scoring function then tries to generalize the information on external protein-ligand complexes. [121] Examples for empirical scoring functions are SCORE1, [119] SCORE2, [122] ChemScore, [123] GlideScore, [124] and X-Score. [125]

Another example is Surflex, which I used for binding free energy predictions. Surflex models the non-covalent binding of small organic molecules to proteins. It was trained on a broad range of structurally and functionally different proteins. [126] It contains parameters which consider hydrophobic complementarity, polar complementarity, and entropy terms. The parameters of the hydrophobic and polar terms are determined using pair-wise distances between atoms, considering factors such as the atom type, the formal charge, and possible hydrogen bonds. The directionality of interactions is also taken into account. Each of the parameters is scaled and optimized according to the training dataset. [121, 127] The entropic term is also taken into account. First it penalizes the score linearly to the number of rotational

bonds, intended to model their restriction. Second it penalizes the score linearly to the logarithm of the molecular weight, intended to model the rotational and translational restriction. [128]

2.4.2.4 Descriptor-based scoring functions

Descriptor-based scoring functions are relatively new, they incorporate quantitative structure-activity relationships (QSAR) into the protein-ligand complex interaction evaluation. Structure-activity relationships (SAR) determine the correlation between structure and biological activity, i.e. how changes in the structure of a ligand influence the biological activity. [129] QSAR are usually performed using regression models, trying to predict the biological activity given different chemical structures. [130] QSAR correlates physicochemical properties (molecular descriptors) with biological activity. [131, 132] Experimental data regarding activity and selectivity of the ligands is used for the generation of a model correlating the biological activity with the molecular descriptors. [133] Usually the scoring function starts from a large set of molecular descriptors and uses machine learning algorithms in order to determine the descriptors and their values that influences the biological activity. A statistically valid QSAR model can be used to predict the biological activity of a compound without performing any biological evaluations. [118] Examples for descriptor-based scoring functions are NNScore, [134, 135] RF-Score, [136, 137] SFCscore^{RF}, [138] and ID-Score. [139]

2.5 Determining similarity between ligand poses

Determining the similarity between different ligand poses is of high importance for structure-based drug design. The first use of it is to measure the ability to reproduce known crystallographic structures (re-docking) that way allowing to examine the success of re-docking. Furthermore, it allows to examine the variability within a given set of molecules. The most commonly used method for similarity comparison between different conformers (in my case poses) of the same molecule is root-mean-square deviation (RMSD). It gives a quantitative single-number measurement of the structural similarity, as described in equation 4. [140-142]

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(v_i - w_i)^2} \qquad 4$$

Where *v* and *w* are the coordination vector of two poses and *N* is the number of atoms. The lower the value, the more similar the two conformers are to each other.

2.6 Normal mode analysis

Normal mode analysis (NMA), also called harmonic analysis, identifies the natural movement of atoms relative to each other. NMA characterizes each energy well of the conformational energy landscape as a parabolic approximation of the relative movements caused by thermally induced fluctuations. [60, 61] These relative movements are the internal vibrations of a molecule and describe the dynamics of the system. [143]

2.6.1 Energetic representation of the conformational states

Harmonic approximation of the energy potential well around a given conformation is calculated using eq. 5.

$$U(r) = \frac{1}{2}(r - R) \times K(R) \times (r - R) \qquad 5$$

In this equation *R* is the 3n-dimensional vector (*n* is the number of atoms) describing the center of the energy well. *r* is the 3n-dimensional vector examined. *K* is the Hessian matrix defining the shape of the potential well. [56, 144] This allows identifying the natural concerted motion of the macromolecule, by investigating the vibrational normal modes that are energetically accessible for each energy minimum.

2.6.2 Uses and limitations of NMA

Identification of the underlying motions of a macromolecule can correspond to large atomic displacements, revealing functional importance. [6, 145, 146] Examples for this were shown for the hinge motion of lysozyme [147, 148] and citrate synthase. [149] Large conformational changes were shown in the GroEL chaperonin [150] and aspartate transcarbamylase. [151-153] As NMA investigates the energy minimum of a structure it can be used to refine structures of macromolecules solved by X-ray crystallography [154, 155] and NMR. [156] The dynamic structure of tRNA was investigated using NMA revealing the three different segments and the interactions between them. [157] DNA molecules were also examined displaying the supercoiling mechanism on the molecular level. [158, 159] Furthermore, the vibrational frequencies of the molecules can be used to calculate the vibrational entropy, as was

developed by Hagler *et al.* on oligopeptide structures [160] and further used by Levitt *et al.* on trypsin inhibitor, crambin, ribonuclease and lysozyme. [7] Tidor and Karplus quantified the entropy change of insulin upon dimerization. [36] NMA is considered nowadays as the gold standard method for vibrational entropy analysis. [5, 6, 60]

An important limitation of NMA is that it assumes that the energy landscape is harmonicly shaped. However, many evidence indicates that the energy landcape is an-harmonic. [161-163] Another limitation is that if the structure is displaced from equilibrium, the harmonic assumption is no longer valid. [60]

2.7 Floppy inclusion and rigid substructure topography

Floppy inclusion and rigid substructure topography (FIRST) is a method which uses techniques taken from graph theory and applies them to protein structures in order to describe the flexibility of a protein. [164]

2.7.1 Rigidity theory

A bar and joint framework is a system of joints connected by fixed-length bars. The joints are allowed full free motion, which is limited by the bars. The rigidity of the system is determined by the constraint network of bars between the joints. [165] A rigid framework has no internal degrees of freedom. An example can be given using a framework in a two-dimensional space. If a framework has a total of f degrees of freedom, three degrees of freedom are attributed to the framework as a rigid body, i.e., vertical, horizontal, and rotational degrees of freedom. [165] The number of internal degrees of freedom of the system is f-3. Internal degrees of freedom are also called floppy modes. [166] Each joint has two degrees of freedom. For an N-atomic system, there are $2N$ degrees of freedom of which $2N - 3$ are internal. [167] The bars hold the connected joints fixed, reducing the degrees of freedom. $2N - 3$ bars are the minimal number of bars required to cancel all the internal degrees of freedom of the framework (Figure 3B). [167] A system with less than $2N - 3$ bars maintains some degrees of internal flexibility, i.e., it is underconstrained (Figure 3A). In comparison, a system with more than $2N - 3$ bars contains more bars than needed to rigidify the system and is considered overconstrained (Figure 3C). [164, 168]

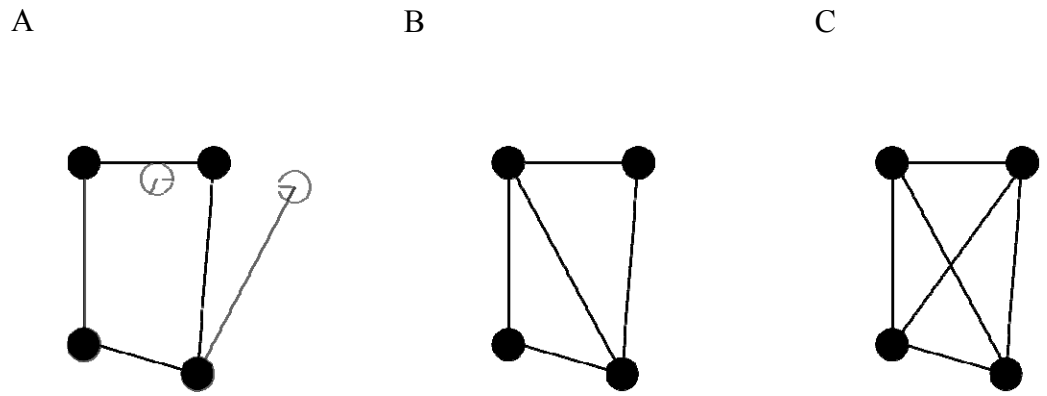

Figure 3: A two-dimensional bar and joint framework where the four joints are represented as dots, and the bars are represented as lines. (A) An underconstrained network, maintaining internal degrees of freedom. (B) A minimally constrained network, with five constraints which are bars between the four joints preventing any internal degrees of freedom. This equals $2N-3$, which is the minimal number of constraints required to remove all internal degrees of freedom. (C) An overconstrained network, with more than $2N-3$ constraints, removing all internal degrees of freedom (figure adapted from [168]). [165]

An important factor to consider when counting the number of joints and bars is the distribution of the bars, i.e., when considering one system, parts of the system can be overconstrained, whereas other parts might be underconstrained (Figure 4). Hence, the total number of bars is not sufficient to determine whether the entire system has internal degrees of freedom or not. [169]

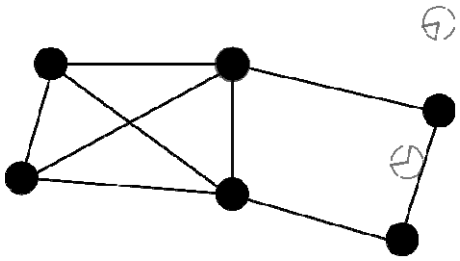

Figure 4: A two-dimensional bar and joint framework where the six joints are represented as dots, and the bars are represented as lines. The total number of constraints equals $2N-3$, which is sufficient to give a minimally constrained network on average. However, the right half of the framework is underconstrained, whereas the left part is overconstrained (figure adapted from [168]).

The constraint network can be also represented as a three-dimensional framework. However, then an additional tern needs to be added in order to account for the dihedral rotation. [170] This is called the body-bar-hinge framework, where each joint is defined as a body with six degrees of freedom. Dihedral rotation about the bar axis makes the bar a hinge. [169]

2.7.2 Three-dimensional pebble game

The pebble game is a recursive algorithm used to describe the rigidity of a three-dimensional joint framework. [167] Each body is assigned three pebbles, representing the three degrees of freedom. Each bar between two bodies is assigned a pebble from one of the bodies attached to it. Pebbles that are not assigned to any bar represent the degrees of freedom of that body. Rearrangement of the framework is allowed as long as the bars that were covered by pebbles remain covered by pebbles of one of the two bodies they are connected to. The result of the pebble game is a framework where the free pebbles represent the number of degrees of freedom of each body to indicate which constraints are independent and which are redundant. [171]

2.7.3 The application of rigidity theory in protein structure

FIRST describes proteins as a system of body-bar-hinge, where the atoms are the bodies, and the bonds are the bars and hinges. Two types of bars are defined, the first allows dihedral rotation and are therefore hinges and the second does not allow rotation and is used to describe non-rotatable bonds. [171] The internal degrees of freedom of each atom of the protein can indicate and show the rigid and flexible substructures of the structure. [164, 172]

2.8 Datasets used for validation

Datasets of pharmacologically relevant targets were used for validation, HIV-1 protease, FXa, Hsp90, trypsin, and a protein-protein dataset.

2.8.1 HIV-1 protease

The Human immunodeficiency virus (HIV) is the etiologic agent for the acquired immunodeficiency syndrome (AIDS). The viral proteins and enzymes that comprise the viral core are processed by the HIV-1 protease, which is an aspartyl protease. [173-175] The essential role of the HIV-1 protease in the HIV life cycle, and its unique specificity, makes it a major therapeutic target. [176]

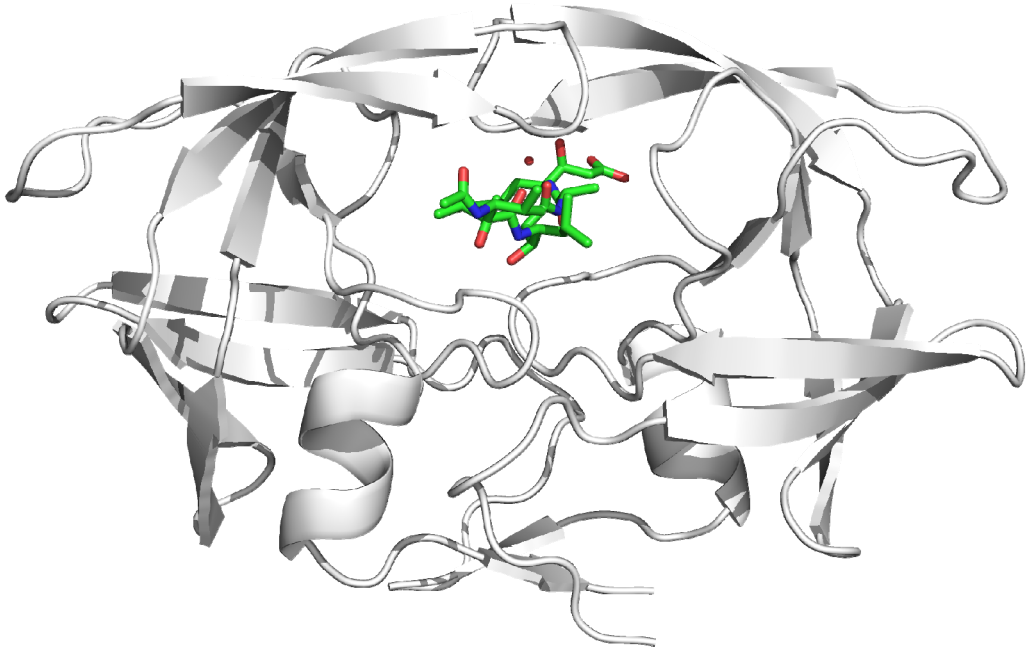

Figure 5: Crystal structure of the homodimeric HIV-1 protease bound to Acetyl-pepsatin (PDB ID 5HVP), shown as cartoon and sticks representation, respectively.

The HIV-1 protease enzyme is a homodimer, where each monomers is composed of 99 amino acids. The active site is located at the interface between the two monomers, where the flaps are also a part of the binding pocket. The triad Asp25, Thr26, and Gly27 from each monomer is located at the bottom of the binding pocket. The two aspartic acid residues which are involved in the catalysis are Asp25, and Asp25´. They share one hydrogen, and therefore have only one negative charge. The protein has a two-fold rotational symmetry (C_2 symmetry), the C_2 axis lies between the two monomers, perpendicular to the catalytic aspartic acids (Asp25, and Asp25`). [177, 178]

The S1 and S1´ subsites are very hydrophobic. S1 is the subsite of the protein, which binds the first residue preceding the cleavage point at the substrate. S2 is the subsite binding the P2 position, which is the second position preceding the cleavage point; on the other side of the cleavage point the substrate residues are P1´, P2´ etc. and of the protein S1´, S2´ etc. [179, 180] The S2 subsites are also hydrophobic, except the Asp29 residues, and the Asp30 residues, the S3 subsites are adjacent to the S1 subsite and are also mostly hydrophobic, except for Arg8. [181, 182]

The two flexible flaps of the HIV-1 protease cover the active site and thereby restrict access to it. It is assumed that in the unbound state the flaps are semi-open, and close upon binding of the substrate. This domain movement accounts for the high flexibility of the binding pocket. [183]

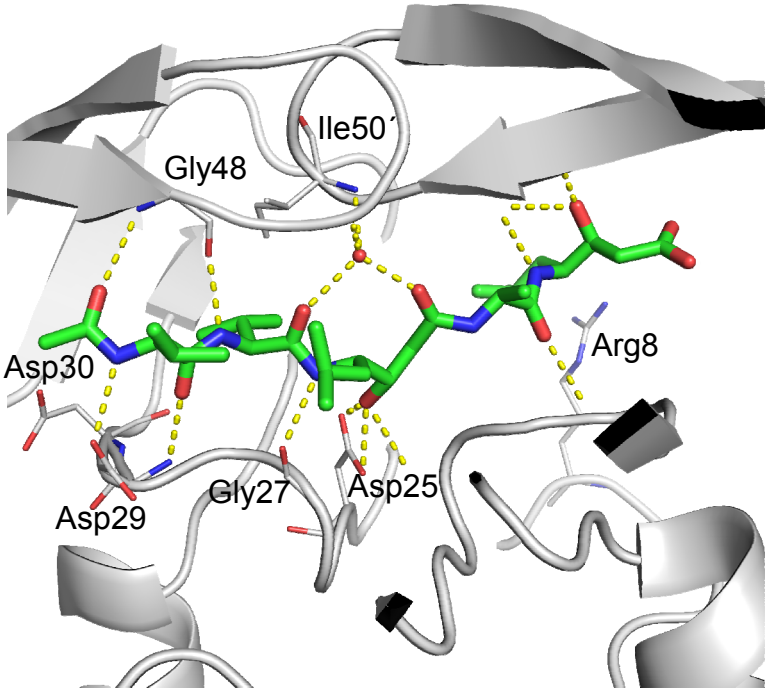

Figure 6: Binding mode of Acetyl-pepsatin, shown in sticks with CPK coloring with carbons colored green in the HIV-1 protease shown in cartoons and colored in white. The binding pocket is shown in sticks and the side chains are colored in CPK with white carbons (PDB ID 5HVP). Dashed lines indicate potential polar interactions.

2.8.2 Factor Xa

Factor Xa (FXa) is a trypsin-like serine protease, which is a pivotal component in the blood coagulation process. [184]

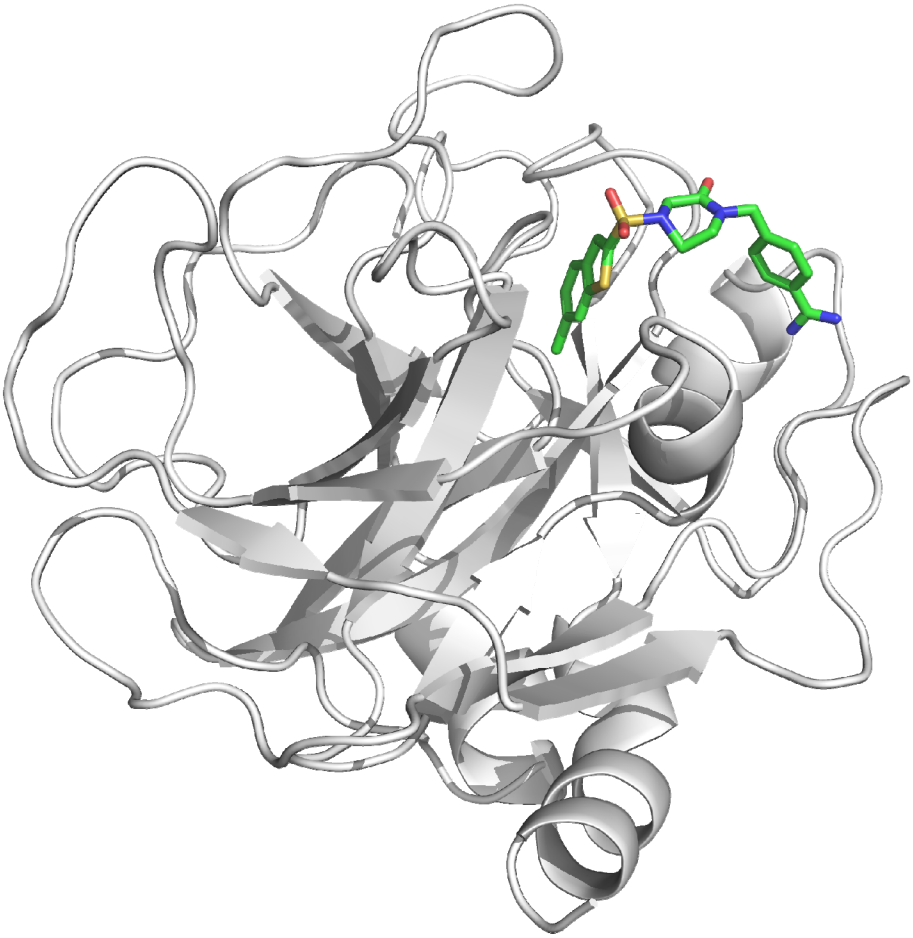

Figure 7: Crystal structure of the FXa bound to RTR (PDB ID 1NFY), shown as cartoon and sticks representation, respectively.

The trypsin family of serine proteases is formed by a structurally conserved globular catalytic domain consisting of two β-barrel subdomains. A cleft between these two subdomains creates the binding pocket. The substrate is a peptide that is cleaved by the catalytic triad Asp102, His57, and Ser195 (the numbering used is from chymotrypsin, [185] which was the first trypsin-like serine protease to be discovered. The use of this numbering is consistent with literature on the trypsin-like serine protease family). [186]

FXa cleaves Phe-Phe-Asn-Pro-Arg-Thr-Phe and Tyr-Ile-Asp-Gly-Arg-Ile-Val in prothrombin. There is a strong preference for arginine as the P1 residue. The positively charged arginine at

the P1 residue forms a salt bridge with the negatively charged aspartic acid (Asp189) at the bottom of the S1 pocket. [187] The shallow character of the S2 pocket, lined by the side chain of tyrosine (Tyr99) allows only a small amino acid in this position, preferably a glycine. The S3 pocket is flat and exposed to the solvent, and the S4 pocket is hydrophobic. [186]

Inhibition of FXa has been shown to be useful against thrombosis, by reducing the prothrombinase activity towards prothrombin, and thereby decreasing and delaying the formation of thrombin upon activation of the clotting cascade. [188]

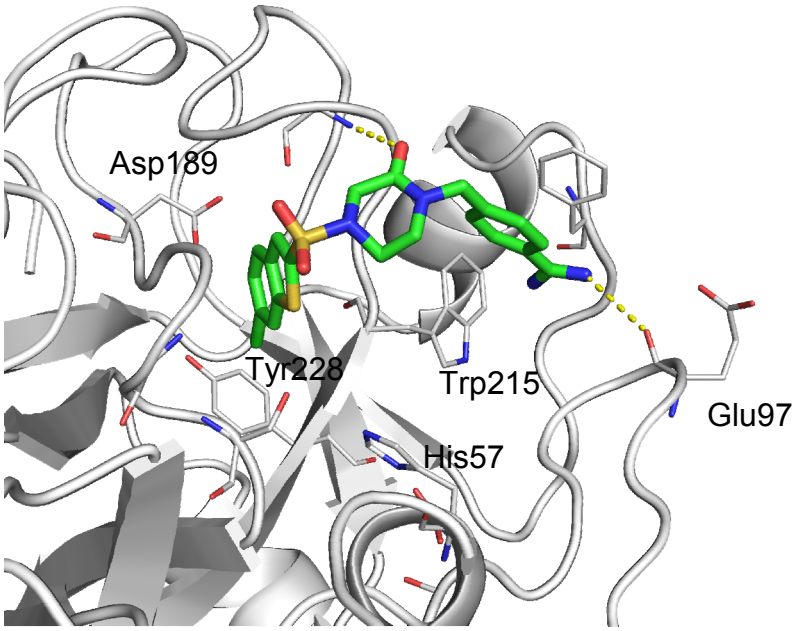

Figure 8: Binding mode of ligand RTR shown in sticks with CPK coloring with carbons colored green in the FXa shown in cartoons and colored in white. The binding pocket is shown in sticks and the side chains are colored in CPK with white carbons (PDB ID 1NFY). Dashed lines indicate potential polar interactions.

The S2 pocket is shallow, lined by the side chain of Tyr99. Therefore, FXa inhibitors aiming to occupy the binding pocket must have an elongated conformation in order to bridge this subsite. Most inhibitors of the FXa contain two basic moieties to occupy the S1 and S4 sites, separated by functional groups operating as extension. The first basic group replaces the P1

arginine of the substrate in the S1 pocket to generate a salt bridge to the Asp189. [188] The second basic group binds in the S4 hydrophobic pocket by generating π–cation interactions to the three aromatic amino acids: Trp215, Phe174, and Tyr99 composing this sub-pocket, this replaces the Ile P4 residue of the substrate. [187]

2.8.3 Heat shock protein 90

Heat shock protein (Hsp) 90 kilodalton (kDa) is an ATP-dependent molecular chaperone. It belongs to a family of highly abundant chaperones. [189, 190]

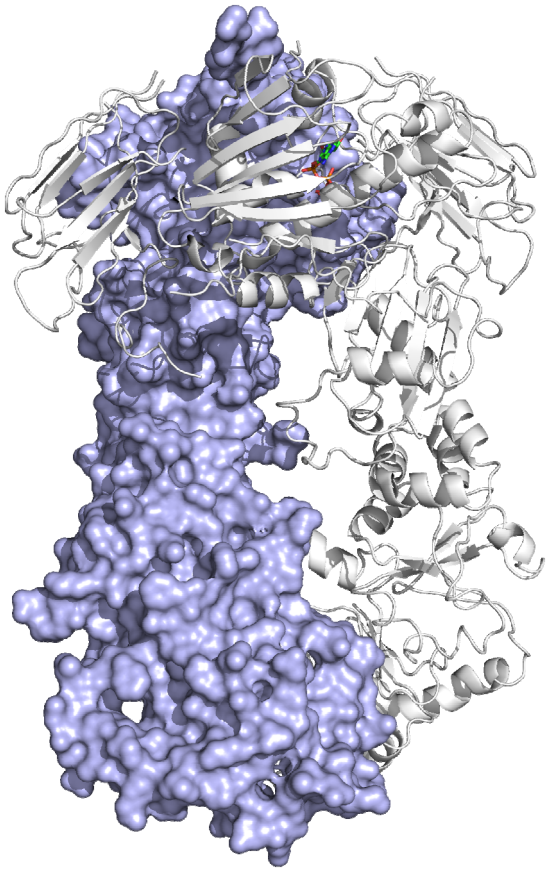

Figure 9: Crystal structure of the homodimeric Hsp90 bound to ATP (PDB ID 2CG9). One monomer is shown as surface representation, one as cartoon representation, and the ligand as sticks representation.

Hsp90 is a homodimer, where each monomer is composed of three domains, the C-terminal domain, the middle domain, and the N-terminal ATP-binding domain. [189] The C-terminal domain is a dimerization site. The middle domain is involved in the ATP hydrolysis, as it is

part of the ATP-binding site, and contains typical catalytic residues. In addition, the middle domain is involved in the binding of client proteins, and co-chaperones. The N-terminal domain contains a unique ATP and ADP binding site. Upon binding and hydrolysis of the ATP to ADP conformational changes occur, which regulate the binding of target proteins, and co-chaperones. [190]

The Hsp90 machinery is still not completely understood. It is assumed that the middle and N-domain of each monomer act as molecular clamp (connected at the C-terminal domain), trapping most of the client proteins, thereby leading to their conformational alteration to the active form. Nucleotide binding facilitates the transition between the open (apo) and closed conformations. [191-194] More than 200 proteins to date are identified as client proteins of Hsp90, among them are kinases, e.g., Bcr-Abl [195], transcription factors, e.g., p53 [196], and other chaperones, e.g., Hsp70. [196]

Hsp90 is a key component in the ability of the cell to handle stress conditions as many of its target proteins regulate cell survival, proliferation and apoptosis. Cancer cells experience extreme stress conditions (i.e., lack of nutrients, hypoxia, proteotoxic stress, genetic instability, etc.). Overexpression of Hsp90 "buffers" these stress conditions and allows the cell to survive and maintain its cancerous character. [197] As a result of the increasing stress levels, the expression of Hsp90 also increases from about 1-2% of the total cellular protein under normal conditions to 4-6% under stress conditions. [198, 199] Therefore, several researches showed that Hsp90 is significant in the progress of cancer, and is considered a major target for cancer therapy. [200-203]

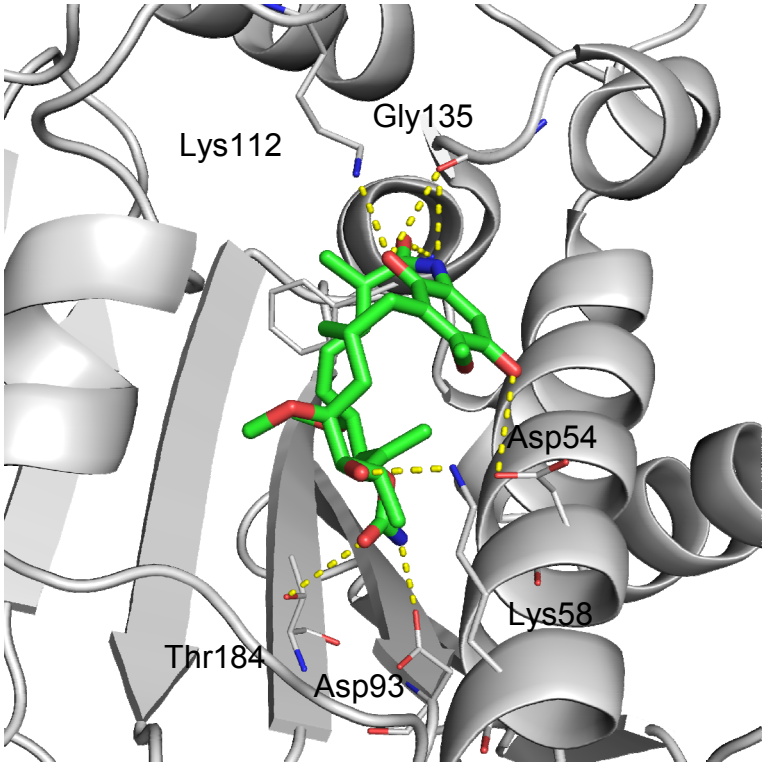

Figure 10: Binding mode of Geldanamycin shown in sticks with CPK coloring with carbons colored green and the N-terminal domain of Hsp90 shown in cartoons and colored in white. The binding pocket is shown in sticks and the side chains are colored in CPK with white carbons (PDB ID 1YET). Dashed lines indicate potential polar interactions.

The first identified inhibitors of Hsp90 were the natural products Geldanamycine (GA), and Radicicol (RD). [204] They were shown to adopt an ATP-like bound conformation inside the N-terminal ATPase binding site, thereby inhibiting the ATP binding and subsequent hydrolysis. This prevents the chaperone activity of the Hsp90 on its oncogenic target protein, leads to the accumulation of these oncogenic proteins, and in turn leads to cell apoptosis. [205-207]

2.8.4 Trypsin

Trypsin is a digestive enzyme, which like FXa belongs to the trypsin family of serine proteases. It contains the same catalytic triad, Asp102, His57, and Ser195. [208] Cleavage of the N-terminal domain of trypsinogen results in the active form of trypsin. [209] Trypsin cleaves C-terminal to positively charged side chains. [210]

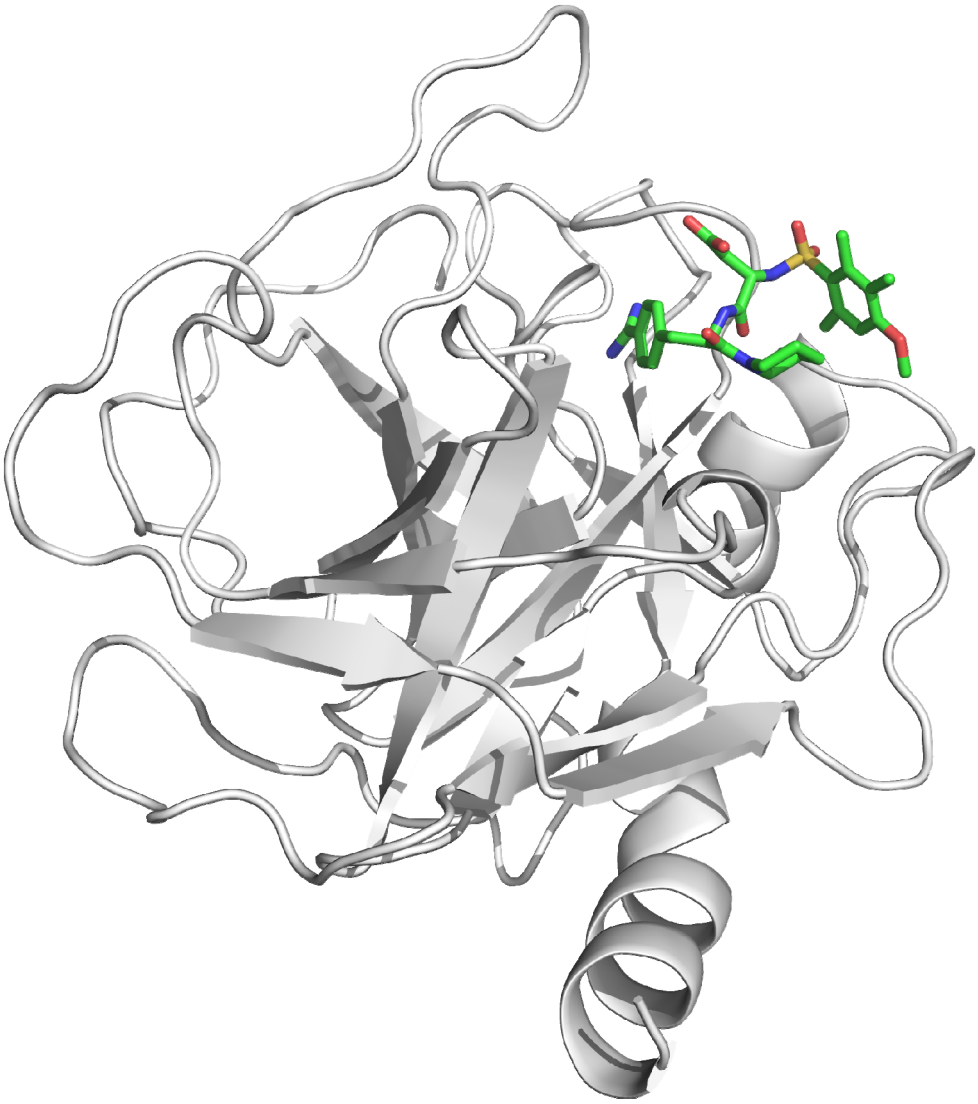

Figure 11: Crystal structure of trypsin bound to CRC200 (PDB ID 1K1N), shown as cartoon and sticks representation, respectively.

The positively charged amino acid of the substrate forms a salt bridge with Asp189 at the bottom of the S1 pocket. [211] Gly 193 and Ser 195 form the oxianionic binding hole. [209]

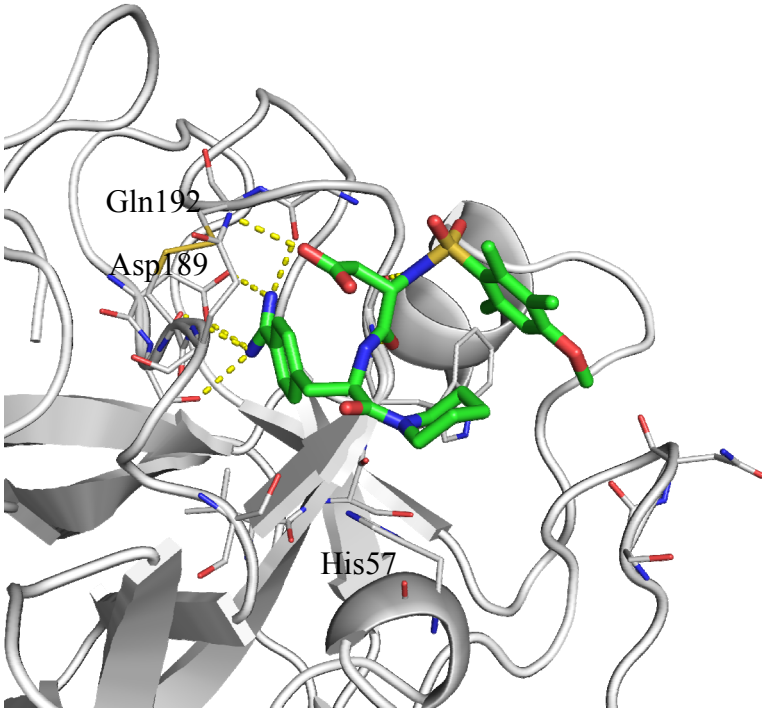

Figure 12: Binding mode of CCR shown in sticks with CPK coloring with carbons colored green and trypsin is shown in cartoons and colored in white. The binding pocket is shown in sticks and the side chains are colored in CPK with white carbons (PDB ID 1K1N). Dashed lines indicate potential polar interactions.

2.8.5 Protein-protein dataset

In this dataset structures of four protease-protease inhibitors, four antibody-antigen and two signal transduction complexes were used. The dataset is comprised of a chymotrypsin with the third domain of the Kazal-type ovomucoid inhibitor from Turkey complex, Ras-Raf complex, Ras-RalGDS (guanine nucleotide dissociation stimulator) complex, human leukocyte elastase complexed with the third domain of the Kazal-type ovomucoid inhibitor from Turkey, β-trypsin complexed with a peptidic inhibitor, and of subtilisin complexed with a Streptomyces subtilisin inhibitor. The complexes in this dataset exhibit diverse folds, protein sizes between 775 and 8398 atoms, and binding affinities from the μM to pM range.

3 Theory

3.1 Molecular recognition and Gibbs free energy

My aim is the calculation of the free energy of binding upon non-covalent association of a protein and a ligand, forming a complex. This calculation is valid for the non-covalent binding of any two molecules, but for my case I consider a protein and a ligand. It is the basis for many enzymatic reactions, such as catalysis, hydrolysis, and phosphorylation. [212, 213]

The process of molecular recognition and binding between a protein and a ligand to form a complex can be described as the equilibrium of the associated and dissociated states

$$P + L \underset{K_d}{\overset{K_a}{\rightleftharpoons}} PL \qquad 6$$

In this equilibrium P is the protein, L the ligand, and PL the protein-ligand complex.

The total differential of the Gibbs free energy (dG) of the binding process between a protein and a ligand is described in eq. 7. [42]

$$dG \approx -SdT + Vdp + \sum_{\alpha} \mu_{\alpha} dN_{\alpha} \qquad 7$$

Where S is the entropy, p is the pressure, V is the volume, and μ_α is the thermodynamic potential of molecule α (protein, ligand, and protein-ligand complex).

Upon protein-ligand binding under constant pressure and temperature conditions ($dp = dT = 0$, $dN_{PL} = -dN_P = -dN_L = 1$), the Gibbs free energy (G) equals the thermodynamic potential (μ), and the change in the binding free energy (ΔG) is [42]

$$\Delta G_{bind} = \mu_{PL} - (\mu_P + \mu_L) \quad (8)$$

μ_α is expressed in eq. 9. [42]

$$\mu_\alpha = -k_B T \ln \frac{Q_\alpha / V}{C_\alpha} \quad (9)$$

Where C_α is the concentration of molecule α, k_B is the Boltzmann constant, T is the temperature and Q is the molecular partition function or the statistical weight.

The molecular partition function describes a system in thermodynamic equilibrium. It sums over all possible microstates. [214] There are two main approaches for describing these partition functions, the first is the rigid rotor harmonic oscillator approximation and the second is the flexible molecule approach.

3.1.1 Formulation in terms of the rigid rotor harmonic oscillator approximation

The rigid rotor harmonic oscillator (RRHO) formulation of the partition function approximates the protein-ligand complex, protein and ligand as rigid rotors, i.e., the only internal motions of the molecules are vibrational. [36, 42, 215] This allows to approximate the translational, rotational, and vibrational components of the partition function (from eq. 9) separately. [42]

The translational part of the partition function is

$$Q_T = V\left(\frac{2\pi m}{\beta h^2}\right)^{3/2} \quad (10)$$

In this equation h is Planck's constant, m is the molecular mass, V is volume, and β is defined as expressed in eq. 11 [216]

$$\beta \equiv \left(k_B T\right)^{-1} \qquad 11$$

In the RRHO approximation, as the molecules are rigid, the three moments of inertia are treated as constants and therefore, the rotational part of the partition function is [42]

$$Q_R = \frac{8\pi^2}{\sigma_{ext}}\left(\frac{2\pi m}{\beta h^2}\right)^{3/2}\left(I_1 I_2 I_3\right)^{1/2} \qquad 12$$

In this equation I_1, I_2, and I_3 are the three principal moments of inertia of the molecule and σ_{ext} is the symmetry number for external symmetry operations that leave internal molecular coordinates unchanged.

The vibrational part of the partition function is calculated from the internal vibrations, as described by eq. 13. [42]

$$Q_{Vib} = e^{-\beta E_0}\prod_i \frac{e^{-\beta h\omega_i/4\pi}}{1-e^{-\beta h\omega_i/2\pi}} \qquad 13$$

In this equation i is an internal vibration, ω_i is its angular frequency, and E_0 is the energy minimum. [59].

The full partition function is the combination of the translational, rotational, and vibrational partition functions (eqs. 11, 12, and 13) as expressed by eq. 14. [217]

$$Q = Q_T Q_R Q_{Vib} \qquad 14$$

3.1.2 Formulation in terms of the flexible molecule approach

The following is taken from the manuscript "Efficient approximation of ligand rotational and translational entropy changes upon binding for use in MM-PBSA calculations" by Ben-Shalom I. Y., Pfeiffer-Marek S., Baringhaus K. H., and Gohlke H. (submitted).

Our approach to estimate translational and rotational entropy change upon ligand binding to a protein starts from the formulation of the molecular partition function in what has been termed the flexible molecule (FM) approach, [42] which distinguishes itself from the RRHO approximation[36, 215] in that it does not depend on a rigid rotor approximation and, hence, is better suited for flexible molecules. The FM approach relies on classical statistical mechanics, which allows to omit kinetic energy contributions to the partition function. Furthermore, as applies to the RRHO approximation, in the FM approach in the classical limit, the free energy or entropy of binding become independent of mass. [42]

In the FM approach, the molecular partition function is given by the configurational integral over the n spatial coordinates, which can be separated into three coordinates of translation, three coordinates of rotation, and $(3n - 6)$ internal coordinates

$$Q = V\left(\frac{8\pi^2}{\sigma_{ext}}\right)Q_{in} \qquad 15$$

Overall translation contributes a factor of V, overall rotation a factor of $8\pi^2 / \sigma_{ext}$, and Q_{in} is the internal contribution. [42, 218]

Considering $S \sim \ln Q$ [218] and eq. 15, the complete configurational entropy can be decomposed into

$$S = S_T + S_R + S_{in} \qquad 16$$

without the requirement to consider mutual information terms that account for the degree of correlation between two coordinates [219-221] as such terms are zero in the FM approach when they involve overall translation and rotation. [42]

For our purposes, it is convenient for the protein-ligand complex C to further separate the six relative translational (T') and rotational (R') coordinates of the ligand L with respect to the protein P from the remaining $3n - 12$ internal coordinates (in')[42]

$$S^C = S_T^C + S_R^C + S_{T'}^C + S_{R'}^C + S_{in'}^C + \frac{1}{2}\sum_{i,j\in\{T',R',in'\};i\neq j} I_{i,j} + I_{T';R';in'} \qquad 17$$

which now incurs likely finite second- and third-order mutual information terms I. As a first approximation, we neglect these terms as they are computationally expensive to evaluate. [220, 221] However, it has been suggested that these terms can be similar in magnitude to the individual entropy terms. [43, 222]

The change in configurational entropy upon formation of the protein-ligand complex from the two independently moving binding partners then is

$$\Delta S_{config.} = S_T^C + S_R^C + S_{T'}^C + S_{R'}^C + S_{in'}^C - \left(S_T^P + S_R^P + S_{in}^P + S_T^L + S_R^L + S_{in}^L\right) \qquad 18$$

Consider that S^{C}_{T} and S^{P}_{T} are identical because the standard volume V (eq. 15) applies to both species and that S^{C}_{R} and S^{P}_{R} will likewise cancel if σ_{ext} of both species are identical (which holds for asymmetric proteins, and which we apply as a second approximation if the protein has rotational symmetry but the binding ligand is asymmetric). The contributions resulting from S_{in} and $S_{in'}$ are described in the chapter " Estimating changes in the vibrational entropy". Here, we make use of a third approximation, we neglect contributions by S_{in} and $S_{in'}$ for all three species. We do so assuming that considering relative translational and rotational motions (T', R') between protein and ligand in the complex ("librational motions") captures a major contribution to $\Delta S_{config.}$. Along these lines, contributions due to restrictions of the conformational space of the binding partners, related to drops in the *number* of energy wells

accessed before and after complex formation, have previously been shown not to be the primary source of $\Delta S_{\text{config.}}$;[43, 47] for the ligand, this may be due to conformational preorganization already in the unbound state [223, 224] and for the protein due to a restricted rotamer space of side chains located in the concave surface of the binding pocket. [225, 226] Not considering changes in the *width* of energy wells upon complex formation appears more severe, however. [43, 47] Yet, for reasons of computational efficiency, we are going to treat the protein as rigid for evaluating $\Delta S_{\text{config.}}$ (see below), excluding *per se* the possibility to compute changes in the width of its energy wells. Finally, as to the ligand, [2, 43] this results in

$$\Delta S_{config.} = \left(S_{T'}^{C} - S_{T}^{L}\right) + \left(S_{R'}^{C} - S_{R}^{L}\right) \qquad 19$$

3.1.2.1 Approximation of the change in translational entropy

The term within the first brackets in eq. 19 can be evaluated as [31, 213]

$$S_{T'}^{C} - S_{T}^{L} = \Delta S_{T} = k_{B} \ln\left(\frac{V_{bound}}{V_{unbound}}\right) \qquad 20$$

with V_{unbound} = 1660 Å^3 as found by integrating over the translational volume of a ligand in solution at a standard concentration of 1 M and k_B being the Boltzmann constant. [40, 227] V_{bound} is the effective translational volume accessible to the ligand after binding. [42] We determine the configurational space of the bound ligand by docking (see Figure 13A). Similar poses are clustered together assuming that a cluster represents a minimum in the energy landscape (see Figure 13B). It is important to note that a change in translational entropy defined that way in general depends on how the relative translational coordinates are defined for the complex. [227] Treating the protein as rigid as done here (see below) eliminates one source of ambiguity arising from motions of some of the protein atoms used to define the protein's reference point relative to the reference point of the ligand. [42] Here, we compute V_{bound} separately for each cluster from

$$V_{bound} = [\max(X) - \min(X)] \times [\max(Y) - \min(Y)] \times [\max(Z) - \min(Z)] \quad 21$$

with $\min\{X, Y, Z\}$ and $\max\{X, Y, Z\}$ being the respective minimal and maximal positions of the ligand's center of mass along the Cartesian axes (see Figure 13C). This procedure has been used previously [39, 43, 49, 52, 228] and makes use of the fourth approximation that the ligand in the bound state resides in a square well potential of mean force.

3.1.2.2 Approximation of the change in rotational entropy

By analogy, we compute the term within the second brackets in eq. 19, again for each cluster separately, as [31, 213]

$$S_{R'}^{C} - S_{R}^{L} = \Delta S_R = k_B \ln\left(\frac{\Omega_{bound}}{\Omega_{unbound}}\right) \quad 22$$

with $\Omega_{\text{unbound}} = 8\pi^2 / \sigma_{\text{ext}}$ [40, 49, 52] and Ω_{bound} being the effective rotational volume accessible to the ligand after binding. [49, 52, 229-231] Following Ruvinsky, [52] we approximate Ω_{bound} as

$$\Omega_{bound} = [-\cos(\max(\theta)) + \cos(\min(\theta))] \times [\max(\varphi) - \min(\varphi)] \times [\max(\psi) - \min(\psi)] \quad 23$$

For determining θ, ϕ, and ψ, we treat the ligand as a rigid body. We center the ligand, determine its principal axes and the corresponding eigenvalues, order the eigenvalues according to their magnitude, and calculate the rotation matrix. [232] The quaternion is then computed from the rotation matrix according to ref. [233], and from it θ, ϕ, and ψ. [234] The cosine function for the angle θ results from the integration over $\sin(\theta)$ in the rotational partition function as described in eq. 20 in ref. [51] and ref. [235].

3.1.2.3 Multiple energy wells in the bound state

In the bound state, multiple binding modes of a ligand can sometimes be observed, particularly for weakly binding ligands, [101, 236-241] reflecting an energy surface with energetically similar energy wells. Previous work showed [43, 53] that the entropy across these wells is the weighted average of the entropies S_i associated with an individual well plus an entropy associated with the distribution of the system across the energy wells $\{i\}$ ("mixing entropy", eq. 24). [42]

$$S = \sum_i p_i S_i - k_B \sum_i p_i \ln p_i \qquad 24$$

p_i is the probability of finding the system in energy well i and is computed from all bound ligand configurations $\{j\}$ in a well according to (eq. 25) [42, 242]

$$p_i = \frac{e^{-E_i / k_B T}}{\sum_j e^{-E_j / k_B T}} \qquad 25$$

where E_i is the energy of the lowest (best) scored pose in well i and T = 298 K.

As p_i computed according to eq. 25 depends on the accuracy of the docking energy in our approach, which may be limited, we tested two alternatives. In the first, p_i is computed from the number of poses N_i in energy well i (eq. 26)

$$p_i = \frac{N_i}{\sum_j N_j} \qquad 26,$$

and in the second, all *n* energy wells get the same weight (eq. 27)

$$p_i = \frac{1}{n} \qquad 27.$$

As we do not consider multiple energy wells, i.e. conformations, for a ligand in the free state, we approximate eq. 24 by omitting the "mixing entropy" for the bound state as well. See above with respect to the influence of the drop in the number of energy wells on $\Delta S_{\text{config.}}$.
Considering eqs. 19-22, finally, this results in the expression for approximating $\Delta S_{\text{config.}}$ used in this work

$$\Delta S_{config.} = k_B \sum_i p_i \times \left(\Delta S_{R,i} + \Delta S_{T,i} \right) \qquad 28$$

Note that this way of averaging entropies S_i associated with an individual well distinguishes our work from that of Ruvisnky *et al*. [50, 51].

3.1.2.4 Sampling of energy wells in the bound state

According to the predominant states approximation introduced by Gilson, [243] the largest contributions to the configurational integral are found in or near energy minima. [42] We thus approximate eq. 24 for the bound state by considering a finite, and usually small, number of well-defined energy wells (see below for how such energy wells are identified). Following a suggestion of Ruvinsky *et al.*, [49, 52] we use a global optimization technique combined with a local energy minimization as implemented in the Lamarckian genetic algorithm of AutoDock 3.0 [244] for generating bound ligand configurations located in energy wells (see chapter 4.3 "Molecular docking" section in Materials and Methods for details).
Note that for smooth energy landscapes or systems with many degrees of freedom, sampling the energy wells will be slow to converge such that non-negligible contributions to the overall $\Delta S_{\text{config.}}$ may be missed. [243, 245-247] A correction has been devised for that case by Gilson and coworkers. [243, 245] We feel it safe to assume, however, that the energy landscape of the bound

state is dominated by a small number of low-energy states only [248-250] such that we do not need to consider such a correction.

Proper identification of which of the generated bound ligand configurations belong to one energy well is important for appropriately estimating residual translational and rotational mobility according to eqs. 27 and 29. An obvious and widely used criterion is to cluster the ligand configurations based on the root-mean-square deviation (RMSD) of their coordinates. [49, 52, 140-142, 244] We tested this criterion, too (see chapter 4.4 "Clustering of ligand binding poses" in Materials and Methods for details). However, as this criterion can lead to essentially identical binding modes being sorted into differential clusters due to different conformations of ligand parts that remain solvated, in addition, we devised and tested an interaction-based clustering inspired by interaction fingerprints introduced in previous studies. [251, 252] See chapter 4.4 "Clustering of ligand binding poses" in Materials and Methods for details. This way of clustering bound ligand configurations for approximating $\Delta S_{\text{config.}}$ distinguishes our work from that of Ruvisnky *et al.* in that only an RMSD-based clustering was used there. [49, 52]

The overall workflow yielding the $\Delta S_{\text{config.}}$ approximation (eq. 28) (Figure 13) has been termed BEERT (*B*inding *E*ntropy *E*stimation for (changes in) *R*otation and *T*ranslation).

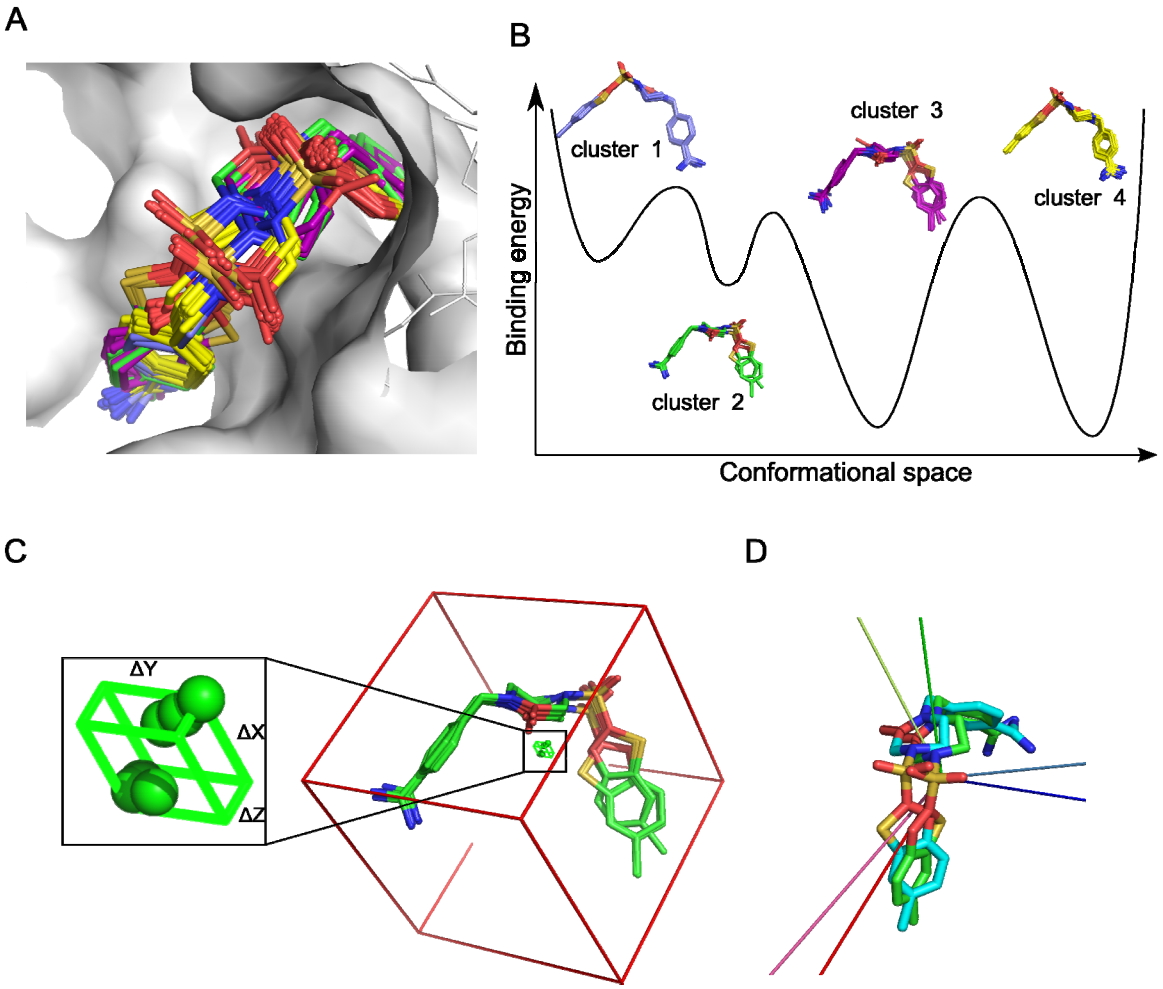

Figure 13: Procedure to approximate $\Delta S_{\text{config.}}$ by BEERT (eq. 28). (A) Sampling ligand binding poses by docking. (B) The different clusters represent energy wells on the binding (free) energy landscape. (C) For each ligand pose, the center of mass is computed (marked as a green point). For each cluster, the effective translational volume is calculated from the encapsulated volume of the centers of mass of all ligand poses in the cluster (eq. 21). (D) For each cluster, the effective rotational volume is calculated from the principal axes of the ligand poses (eq. 23).

3.1.3 Estimating changes in the vibrational entropy

The following is taken from the manuscript "Rigidity theory-based approximation of vibrational entropy changes upon binding to biomolecules" by Gohlke H., Ben-Shalom I. Y., Kopitz H., Pfeiffer-Marek S., and Baringhaus K. H..

We introduce a computationally highly efficient approximation of vibrational entropy change ($\Delta S_{vib.}$) upon binding to biomolecules based on rigidity theory and compare its results for datasets of protein-protein and protein-ligand complexes to those obtained with NMA-based $\Delta S_{vib.}$. The principle idea underlying our approach is that, rather than estimating $\Delta S_{vib.}$ from changes in the vibrational frequencies of normal modes and, hence, the *width* of energy wells upon binding, we estimate $\Delta S_{vib.}$ from changes in the *variation of the number* of low (i.e., zero) frequency modes. This will be described in detail in the following.

In normal mode analysis, a potential energy function $V(x)$ is expanded in a Taylor series expansion about some point x_0. [61] If x_0 denotes the location of a minimum of $V(x)$, the gradient of $V(x)$ vanishes. If also third and higher-order derivatives of $V(x)$ are ignored, the dynamics of the system can be described in terms of linearly independent normal modes obtained from diagonalizing the Hessian matrix, each one associated with a frequency ν_i. From the ν_i, the vibrational contributions to thermodynamic properties can be determined. [253, 254] For $S_{vib.}$, one obtains

$$S_{vib.} = T^{-1} \sum_{i=1}^{3N-6} \left[\frac{h\nu_i}{e^{\frac{h\nu_i}{k_B T}} - 1} - k_B T \ln\left(1 - e^{\frac{-h\nu_i}{k_B T}}\right) \right] \qquad 29$$

As is obvious from eq. 29, $S_{vib.}$ is particularly sensitive to the frequencies of the lowest modes of vibration. [36, 253] The low-frequency modes reflect the presence of weak forces in the biomolecular system, encoded, e.g., as torsion angle and van der Waals potentials in current state-of-the-art biomolecular force fields. [255]

To make now the connection to approximating $S_{vib.}$ based on rigidity theory, we first neglect weak forces in $V(x)$, resulting in a Kirkwood [256] or Keating [257] potential V_K, schematically written as [165]

$$V_K = \frac{\alpha}{2}(\Delta l)^2 + \frac{\beta}{2}(\Delta\theta)^2 \qquad 30$$

V_K describes small displacements from an equilibrium structure in a bond-bending network in terms of changes in bond length (Δl) and bond angle ($\Delta\theta$), with α and β being the force constants for bond stretching and bending, respectively. Diagonalizing the Hessian from eq. 30 ascertains a number F of vibrational modes with zero frequency. [165] These so-called floppy modes correspond to the ways in which the network can be continuously deformed at no cost in energy by rotations around bonds; F decreases with an increasing mean coordination $\langle r \rangle$ in the network (Figure 14A). [165] Note that the floppy modes will become "spongy", i.e. will have a small finite frequency, if weak forces are present in the network.

Four points are important for estimating $S_{vib.}$ from F. First, if α and β become (infinitely) large in eq. 30, the bond stretching and bending forces become bond and angle constraints, leading to a constraint network. A representation of biomolecules in terms of constraint networks has been successfully used in the analysis of bimolecular rigidity and flexibility previously, where, in addition to covalent interactions, non-covalent interactions (hydrogen bonds, salt bridges, and hydrophobic tethers) are modelled via bond and angle constraints. [164, 258, 259] Second, as to computational efficiency, rather than by diagonalizing the Hessian from eq. 30, F can also be determined by an advanced constraint (Maxwell) counting [165, 260] on the constraint network as implemented in the combinatorial "pebble game" algorithm. [167, 261] This algorithm performs with a time complexity of, on average, $O(N)$, providing for a dramatic speed up for large (biomolecular) systems compared to the time complexities of $O(N^3)$ for matrix diagonalization. Third, for the description of a system's dynamics by NMA, the system must reside at a local minimum on the potential energy hypersurface. In MM/PB(GB)SA-type applications, typically, structures have been minimized to a root mean-square gradient (RMSG) of the potential energy of $10^{-5} - 10^{-3}$ kcal mol^{-1} Å $^{-1}$ before applying NMA. [23] Performing minimizations of biomolecules to such low RMSG is computationally demanding. In contrast, no minimization is required prior to applying the "pebble game" algorithm to constraint networks. Fourth, in rigidity and flexibility analysis, the negative of the number of floppy modes, $-F$, has been shown to act as a free energy: F is a convex function of $\langle r \rangle$ (i.e.,

$F^{(2)} = d^2F / d\langle r\rangle^2 \geq 0$) (Figure 14A) such that if there is an ambiguity, the system will always be in the lowest free energy, i.e. maximum floppy modes, state. [169, 262] Considering $\langle r\rangle$ as a temperature-like quantity, [263] $F^{(2)}$ has been regarded as a specific heat and used to characterize the order of transitions of constraint networks switching between rigid and flexible states. [262] Continuing the thermodynamic interpretation of F and its derivatives, here we introduce the negative of $F^{(1)} = dF / d\langle r\rangle$ as an entropy-like quantity, based on the defining equation for entropy [254] (eq. 31)

$$S = -(\partial G / \partial T)_{N,p} \sim -(d(-F) / d(-\langle r\rangle)) = -F^{(1)} \qquad 31$$

In eq. 31, we consider that $\langle r\rangle$ decreases with increasing temperature, as already successfully applied in thermal unfolding simulations of constraint network representations of biomolecules. [264, 265] To the best of our knowledge, no such thermodynamic interpretation of $F^{(1)}$ has yet been presented.

For $F^{(1)}$, a relation with respect to F and the total number of bonds N_B in constraint networks with a fixed number of nearest neighbors has been derived (eq. 32) [169, 262]

$$-F^{(1)} \sim (3N - F) / N_B \geq 0 \qquad 32$$

first, yielding that $-F^{(1)} \geq 0$, as required of an entropy. Second, $-F^{(1)}$ depends on the actual network state (Figure 14B): if N_B is low, related to a very flexible network, $-F^{(1)}$ approaches a positive limit, indicating the maximum entropy of the system; if N_B is high, related to a very rigid network, $-F^{(1)}$ approaches zero, as expected for a system for which only one state of realization exists. Note that adding constraints, e.g. due to binding of a ligand, to a constraint network representation of a biomolecule with either low or high N_B will not lead to marked changes in $-F^{(1)}$ (Figure 14B). In contrast, marked changes are to be expected when constraints are added to a network with intermediate N_B (Figure 14B), reminiscent of a biomolecule with marginal stability. [266]

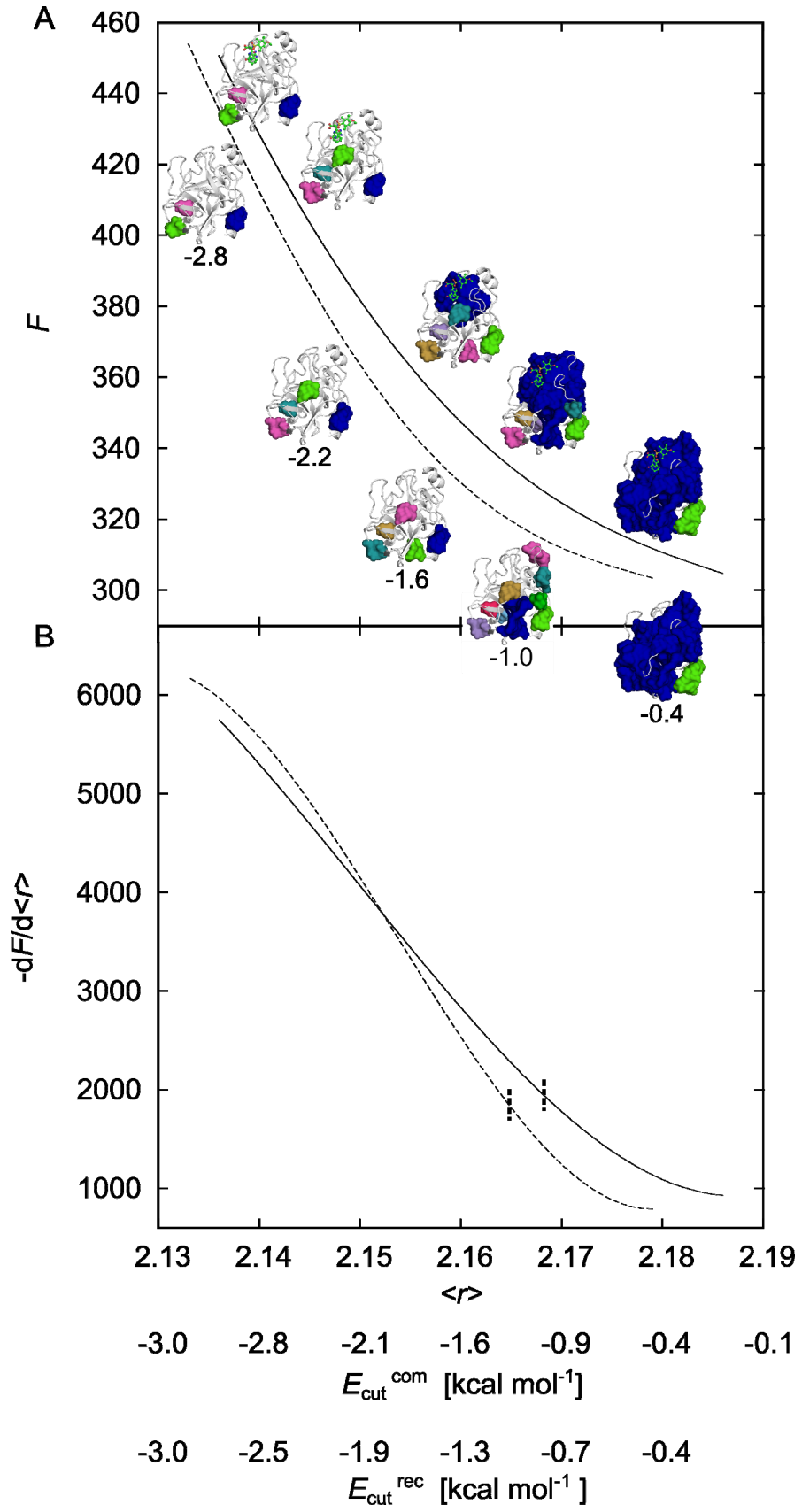
A
460
440
420
400
380
360
340
320
300
F
-2.8
-2.2
-1.6
-1.0
-0.4
B
6000
5000
4000
3000
2000
1000
-dF/d<r>
2.13
2.14
2.15
2.16
2.17
2.18
2.19
<r>
-3.0
-2.8
-2.1
-1.6
-0.9
-0.4
-0.1
E_cut com [kcal mol-1]
-3.0
-2.5
-1.9
-1.3
-0.7
-0.4
E_cut rec [kcal mol-1]

Figure 14: (A) The number of floppy modes (F) as a function of the mean coordination $\langle r \rangle$ is shown exemplarily for one MD simulations-generated conformation of the trypsin-ligand complex (PDB ID 1K1N, solid line) and the protein only (dashed line). Along the lines rigid cluster decompositions of both structures computed at identical E_{cut} values, respectively, are depicted and marked under the structures, with each colored blob indicating one rigid cluster; the largest rigid cluster is colored in blue. (B) $-F^{(1)} = -dF / d\langle r \rangle$, introduced here as an entropy-like quantity, is shown as a function of the mean coordination $\langle r \rangle$, corresponding to the curves in panel A. E_{cut} values corresponding to $\langle r \rangle$ are depicted along the abscissa for the complex and the protein. To compare $-F^{(1)}$ for different biomolecular systems, the respective $\langle r \rangle$ at a fixed E_{cut} value was determined. The vertical dashed and dotted lines depict $-F^{(1)}$ for complex and protein at $E_{cut} = -1.0$ kcal mol^{-1}. Note that in this E_{cut} range, the rigid cluster decompositions (panel A) differ the most.

The change in vibrational entropy upon binding to a biomolecule is then approximated as (eq. 33)

$$\Delta -F^{(1)} = (-F^{(1)}_{com}) - (-F^{(1)}_{rec}) - (-F^{(1)}_{lig}) \qquad 33$$

where com, rec, and lig refer to the complex, receptor, and ligand, respectively.

Computing $-F^{(1)}$ proceeds in three steps. First, a structural ensemble of the complex is generated by all-atom molecular dynamics (MD) simulations. Performing the subsequent analyses on an ensemble rather than a single structure overcomes the problem that results from constraint counting are sensitive to the input structural information. [267, 268] Second, a constraint network is generated for each complex conformation, as done in previous studies of biomolecular rigidity and flexibility. [164, 258, 259] In addition, a constraint network is generated for the receptor conformation extracted from the respective complex, as is for the extracted "ligand" in the case of protein-protein complexes. In contrast, small molecule ligands lack the typical network character and, thus, are not suitable for evaluation by constraint counting. In all, a so-called single-trajectory approach is pursued, as often applied in end-point (free) energy methods, which neglects possible conformational changes of the unbound structures but usually gives less noisy results than the three-trajectory alternative. [269] Third, a "constraint

dilution trajectory" of network states $\{\sigma\}$ is generated from each initial constraint network by successively removing non-covalent constraints. [264, 265, 270-272] Here, hydrogen bond constraints (including salt bridges) are removed in the order of increasing strength [264, 270, 273] such that for network state σ only those hydrogen bonds are retained that have an energy $E_{HB} \leq E_{cut}$. The hydrogen bond energy E_{HB} is determined from an empirical energy function [274] successfully used by us [275-278] and others [265, 270, 271, 279, 280] in this context. For each σ, $F(\langle r \rangle)$ is computed by constraint counting with the program *first*, [164, 167] and from that $-F^{(1)}$ at a given $\langle r \rangle$ by numerical differentiation. Using eq. 32 instead is not possible because atoms in constraint networks generated from biomolecules have a variable number of nearest neighbors. To compare $-F^{(1)}$ for different biomolecular systems, the respective $\langle r \rangle$ at a fixed E_{cut} value was determined. Here, E_{cut} = -1.0 kcal mol^{-1} was used unless otherwise noted, motivated from previous studies. [267]

4 Materials and Methods

4.1 Datasets used for validation

The BEERT workflow was evaluated on three datasets of protein-ligand complexes of pharmacologically relevant targets, HIV-1 protease, Factor Xa (FXa), and Heat shock protein 90 (Hsp90). The structures were retrieved from the Protein Data Bank (PDB), [281] using only complexes that contain a wild-type protein and an inhibitor for which experimental binding affinity information is available. We chose datasets with at least 15 crystal structures of the protein with different ligands. For the HIV-1 protease and FXa datasets, we obtained information about the experimental binding free energy from

$$\Delta G_{bind} = RT \ln K_i \qquad 34$$

For the Hsp90 dataset, we used pIC_{50} (logarithm of the half maximal inhibitory concentration) values instead. As these experimental data were retrieved for all Hsp90 ligands using the same experimental settings, [282-287] they, too, can be used for computing relative binding free energies. For competitive inhibitors the relation between IC_{50} and K_i can be expressed in eq. 35. [212]

$$K_i = \frac{IC_{50}}{1 + \frac{[S]}{K_M}} \qquad 35$$

In this equation [S] is the concentration of the substrate and K_M is the Michaelis constant. [288]

pK_i and pIC_{50} values were taken from the databases PDBbind, [289-291] binding MOAD, [292, 293] and Binding DB. [294-297]

The vibrational entropy project was evaluated in addition on the trypsin dataset and on the protein-protein dataset.

4.2 General preparation of protein and ligand structures

For the BEERT project, for each protein-ligand complex the structural coordinates were retrieved from the PDB. [281] The ligands were removed from the complexes, as were all water molecules, and the ligands were assigned Sybyl atom types and saved separately. For docking, hydrogens and charges are not considered and were therefore not added. [17, 99] For MD simulations, hydrogens were added to the receptors and ligands using PrepWizard. [298] Default protonation states were assigned to all protonatable amino acids except: I) The protonation and rotamer states of histidines, asparagines, and glutamines were assigned manually to optimize local interactions. II) HIV-1 protease contains two aspartic acid residues in the catalytic site, Asp25 and Asp25'. According to several studies, the protonation states of these residues were assigned such that one of the two aspartic acids is monoprotonated with a proton placed on the oxygen in position OD2 of the side chain. [181, 299-301] The protonation tates of the ligands were determined using the Epik program. [302, 303]

For the trypsin dataset that was used in the vibrational entropy project, the protonation state of histidines, asparagines, and glutamines was determined by the REDUCE software [304] and the protonation states of the ligands were determined using the PRODRG server. [305]

Chapters 4.3 until 4.6 were performed for the BEERT project and the following text is taken from the manuscript "Efficient approximation of ligand rotational and translational entropy changes upon binding for use in MM-PBSA calculations" by Ben-Shalom I. Y., Pfeiffer-Marek S., Baringhaus K.H., and Gohlke H. (submitted).

4.3 Molecular docking

All docking runs were performed following established procedures. [99] We used AutoDock 3.05 [244] with DrugScore pair potentials [98] as an objective function, which has been used successfully previously for generating good docking solutions. [80, 306, 307] The docking protocol for flexible ligand docking considered 100 independent runs per ligand using an initial population size of 100 individuals, 5.0×10^3 generations, a maximum number of 10.0×10^6

energy evaluations, a mutation rate of 0.02, a crossover rate of 0.8, and an elitism value of 1. The spacing of the precomputed potential grids was set to the default value of 0.375 Å. In order to probe for the convergence of the $\Delta S_{\text{config.}}$ approximation, in addition, for each complex ensembles with 500 and 1,000 docking poses were generated.

4.4 Clustering of ligand binding poses

Subsets of the docked ligand configurations belonging to a well of the energy landscape of the bound state were, first, identified by RMSD-based clustering as implemented in AutoDock 3.05, [244] using RMSD thresholds of 1 Å. The RMSD-based clustering considers internal symmetries (e.g., in the case of phenyl substituents). [244]

Second, we developed a clustering of bound ligand configurations based on the interaction pattern between the ligand and the protein, which was inspired by previous studies. [251, 252] For this, we identify all pairs of ligand and protein heavy atoms that are closer than a cutoff distance (d_{cut}) for each docking pose (Figure 20) and generate a union set of all protein atoms. We assign "1" only for the pairs of ligand and protein heavy atoms with $d < d_{\text{cut}}$ and arrange them in a matrix such that the value of a matrix element becomes 1 if the actual distance is closer than d_{cut} and 0 otherwise. All such matrices contain the same protein and ligand atoms for one protein-ligand complex in the same order in rows i and columns j, respectively. The number of matrix elements (i, j) that are 1 in both matrices then defines the similarity between two docking poses. The poses are clustered according to their similarity by hierarchical clustering [308-310] as implemented in R. [311] Identifying pairs of ligand and protein atoms in the first step of this approach is efficiently performed making use of a cell data structure [312] ubiquitously applied in MD simulations. [313, 314] For this, the space of the protein is partitioned into cubes $\{c\}$ of edge length d_{cut}. For a given ligand atom in a cube c, only protein atoms within this cube or within the neighboring 26 cubes are considered for distance calculations.

4.5 Estimating binding affinities by DrugScore scoring

For comparison, relative binding affinities are estimated using the DrugScore pair potentials for the docked protein-ligand complex. DrugScore is a knowledge-based scoring function and has been previously used for estimating binding affinities on protein-ligand complexes. [17, 80, 97, 98, 100, 101] DrugScore potentials encode distance-dependent interaction energies between

ligand and protein atoms derived from statistical preferences and implicitly include solvation contributions. [98] Changes in $\Delta S_{config.}$ are not considered in DrugScore.

4.6 Estimating binding affinities by Surflex scoring

For comparison, we also used Surflex as an external scoring function developed by Jain and coworkers to estimate relative binding affinities of docked protein-ligand complexes, [126, 128, 315, 316] which performed very well in an external evaluation. [317] We used Surflex to score our existing docked poses, because it incorporates the number of rotatable bonds of the ligand as a measure for the configurational entropy. [128]

4.7 Molecular dynamics simulations

MD simulations to generate conformational ensembles of the protein-ligand complexes for post-processing with MM-PBSA (see below) were performed with the AMBER11 suite of molecular simulation programs, [318] except for the trypsin dataset, for which AMBER10 was used, [319] following established procedures. [115] The Cornell *et al.* force field [320] with modifications introduced by Hornak *et al.* (ff99SB) [321] and the general amber force field (GAFF) [322] were used for proteins and ligands, respectively. Partial charges of the small molecules were generated according to the RESP procedure. [322, 323] The structures were solvated in a rectangular box of TIP3P water molecules where the distance between the edges of the box and the closest solute atom was at least 11 Å. Periodic boundary conditions were applied using the particle mesh Ewald (PME) method [313] to treat long-range electrostatic interactions. Bond lengths of bonds involving hydrogen atoms were constrained using the SHAKE algorithm. [324, 325] The time step for all MD simulations was 2 fs, and a direct-space non-bonded cutoff of 8 Å was applied.

Initially, each complex crystal structure was minimized by 50 steps of steepest descent minimization, followed by 500 steps of conjugate gradient minimization. After minimization, the system was heated from 100 K to 300 K using canonical ensemble (NVT) MD simulations for 50 ps. Then, the solvent density was adjusted using isothermal-isobaric ensemble (NPT) MD simulations for 250 ps. Positional restraints with a force constant of 5 kcal mol^{-1} $Å^{-2}$ applied during thermalization were reduced in a stepwise manner over 50 ps followed by 50 ps of unrestrained NVT-MD simulations at 300 K with a time constant of 2 ps for heat bath coupling. Temperature control was done using the Berendsen thermostat. [326] The HIV-1

protease, FXa, and Hsp90 complexes were then subjected to 250 ns of NVT-MD simulations for production, extracting snapshots in time intervals of 20 ps. The trypsin complexes were then subjected to 20 ns of NVT-MD simulations for production, extracting snapshots in time intervals of 20 ps.

In addition, for the vibrational entropy project, the dataset of protein-protein complexes was used. The four antibody-antigen and four protein-protein complexes have been used to investigate the energetics of protein-protein complex formation by Brooijmans *et al.*. [327] Conformations extracted from molecular dynamics (MD) simulations between 300 and 600 ps were provided by N. Brooijmans; the MD simulations had been performed with the AMBER suite of molecular simulation programs [319] at 298 K in a box of TIP3P water [328] using the ff99 force field. [320] The signal transduction complexes have been used in a study on protein-protein binding by Gohlke *et al.*. [329] Conformations extracted from molecular dynamics (MD) simulations between 2 ns and 12 ns were used; the MD simulations had been performed with the AMBER suite of molecular simulation programs [319] at 300 K in a box of TIP3P water [328] using the Cornell *et al.* force field.[330]

4.8 Effective energies from MM-PBSA computations

The following was performed for the BEERT project and the following text is taken from the manuscript "Efficient approximation of ligand rotational and translational entropy changes upon binding for use in MM-PBSA calculations" by Ben-Shalom I. Y., Pfeiffer-Marek S., Baringhaus K.H., and Gohlke H. (submitted).

MM-PBSA [23, 26] (molecular mechanics Poisson-Boltzmann surface area) is a post-processing end-point free energy calculation method. [115] MM-PBSA was performed as described previously. [115] Snapshots for the generation of conformational ensembles can either be obtained from a single trajectory of the complex ("single-trajectory approach") or from separate trajectories of the complex, receptor, and ligand ("separate trajectory approach/three-trajectory approach"). Previous studies showed larger noise when pursuing the latter approach. [116, 331] Hence, we followed the "single-trajectory approach" here. Snapshots of the binding partners were extracted every 20 ps from MD trajectories of the complexes of 250 ns length. A sampling interval of 20 ps is well above the correlation time of the effective energy and results in statistically independent snapshots in that respect. [23, 329, 332] All counter ions and

water molecules were stripped from the snapshots. The gas-phase energy was calculated based on the ff99SB force field [321] without applying any non-bonded cutoff. The polar part of the solvation free energy was determined by solving the linearized Poisson-Boltzmann (PB) equation as implemented in AMBER11 [318, 333] and applying PARSE radii. [334] A dielectric constant of 1 and 80 for the interior and exterior of the solute was applied, respectively. The polar contributions were computed at 100 mM ionic strength, with a solvent probe radius of 1.4 Å. The non-polar part of the solvation free energy was calculated by a solvent-accessible surface area-dependent term, using $\gamma = 0.0072$ kcal mol^{-1} Å^{-2} for the surface tension. The sum of the gas-phase energy and the polar and non-polar parts of the solvation free energy will be termed effective energy ($\Delta G_{\text{eff.}}$) below.

Chapters 4.9 and 4.10 were performed for the vibrational entropy project and the following text is taken from the manuscript "Rigidity theory-based approximation of vibrational entropy changes upon binding to biomolecules" by Gohlke H., Ben-Shalom I. Y., Kopitz H., Pfeiffer-Marek S., and Baringhaus K. H..

4.9 Calculation of $S_{\text{vib.}}$ by normal mode analysis

$S_{\text{vib.}}$ was calculated by normal mode analysis (NMA). [5-7] For NMA, the system must reside at a local minimum on the potential energy hypersurface. [61, 335] Therefore, conformations of each *protein-protein complex*, and the respective "receptor" and "ligand" extracted from it, had been energy minimized in the gas phase using a distance-dependent dielectric constant (DDD) of $\varepsilon(r) = 4r$ when calculating Coulombic interactions, with r being the distance between two solute atoms, until the root-mean-square of the elements of the gradient vector (RMSG) was $< 10^{-4}$ kcal mol^{-1} Å^{-1}. [23] From the frequencies of the vibrational modes, $S_{\text{vib.}}$ had been computed according to eq. 29 using $T = 300$ K. [23] For minimization and NMA, the programs sander and nmode of the AMBER7 suite of molecular simulation programs [319] had been used. [23] Changes in the vibrational entropy upon binding were then calculated as

$$\Delta S_{\text{vib.}} = S_{\text{vib.,com}} - S_{\text{vib.,rec}} - S_{\text{vib.,lig}} \qquad 36$$

where com, rec, and lig refer to the complex, receptor, and ligand, respectively.

In the case of *protein-ligand complexes*, in general, each complex conformation from the MD simulations was minimized to RMSG < 10^{-4} kcal mol^{-1} $Å^{-1}$ using the force fields as for the MD simulations and a DDD of $\varepsilon(r) = 4r$ when calculating Coulombic interactions, with r being the distance between two solute atoms. [335] For the minimizations, the program SANDER of the AMBER11 suite of molecular simulation programs was used. [319] Starting structures for the separate minimizations of receptor and ligand were taken from these minimized complex structures. This follows a recommendation by Page and Bates, [336] according to which, for single-trajectory calculations, smaller fluctuations in the computed vibrational entropies are obtained compared to the common approach of extracting starting structures for receptor and ligand from a non-minimized complex structure. NMA was performed as previously established in our group [335] using the program NAB of the AMBER11 suite of molecular simulation programs. [319] In general, all atoms of the respective complex, receptor, or ligand conformations were considered using a DDD of $\varepsilon(r) = 4r$. Changes in vibrational entropy upon binding were then calculated according to eq. 29 and eq. 36, using $T = 300$ K. To further reduce the influence of imprecisions in the $S_{vib.}$ calculations on $\Delta S_{vib.}$, the average over vibrational entropies computed for all receptors of one dataset was used for $S_{vib.,rec}$ in eq. 36.

In addition $S_{vib.}$ values for the trypsin datatset from ref. [335] were used. These values were generated using the generalized Born (GB^{HCT}) model as proposed by Still *et al.* [337] together with the Hawkins *et al.* [338] pair-wise descreening approximation for computing effective Born radii and *mbondi* intrinsic GB radii. [339] Note that $\Delta S_{vib.}$ values calculated with either the DDD or GB^{HCT} models have been found to be highly correlated (see Figure 2 in ref. [335]).

4.10 Constraint network generation and constraint counting

For all conformations containing a biomolecule, constraint networks in the bar and joint representation [167] were generated using the program *ambpdb* with the option "-first" from the AMBER11 suite of molecular simulation programs. [319] To remain consistent with the $S_{vib.}$ calculations, the minimized conformations were used for this, unless otherwise noted. Here, atoms are represented as nodes, and covalent and non-covalent interactions (hydrogen bonds, salt bridges, and hydrophobic tethers) as distance and angle constraints. [164] Hydrogen bonds

and salt bridges were taken into account subject to geometric criteria detailed here [164] and if their energy $E_{HB} \leq E_{cut}$; E_{HB} was computed from an empirical energy function [274] successfully used by us [275-278] and others [265, 270, 271, 279, 280] in this context. Hydrophobic interactions between carbon or sulfur atoms were taken into account if the distance between these atoms was less than the sum of their van der Waals radii (C: 1.7 Å, S: 1.8 Å) plus D_{cut} = 0.15 Å.

For each conformation, a "constraint dilution trajectory" of network states $\{\sigma\}$ is generated from the initial constraint network by successively removing hydrogen bond and salt bridge constraints in the order of increasing strength [264, 270, 273] such that for network state σ only those hydrogen bonds are retained that have an energy $E_{HB} \leq E_{cut}$; for this, E_{cut} was varied in the range of -0.01 to -3.0 kcal mol^{-1} in 100 steps. For a given σ, F and $\langle r \rangle$ are computed by the program *first*, [164, 167]. This results in a smooth function $F(\langle r \rangle)$ (Figure 14). From that $-F^{(1)}$ is computed by numerical differentiation at a given $\langle r \rangle$. To compare $-F^{(1)}$ for different biomolecular systems, the respective $\langle r \rangle$ at a fixed E_{cut} value was determined. Here, E_{cut} = -1.0 kcal mol^{-1} was used unless otherwise noted, motivated from previous studies. [267] Finally, $-F^{(1)}$ results are averaged over all conformations of a complex, receptor, or ligand species, respectively, and Δ-$F^{(1)}$ is calculated according to eq. 33.

In the case of the protein-ligand complexes, to reduce the influence of imprecisions in the $-F^{(1)}$ calculations on Δ-$F^{(1)}$ and remain consistent with the $S_{vib.}$ calculations, the average over $-F^{(1)}$ computed for all receptors of one dataset was used for $-F^{(1)}_{rec}$ in eq. 33. Furthermore, small molecule ligands lack the typical network character and thus are not suitable for evaluation by constraint counting. For such ligands, $-F^{(1)}_{lig}$ in eq. 33 is replaced by a scaled $c * S_{vib.,lig}$ value. The scaling coefficient c was determined as an average over the ratios $-F^{(1)}_{com} / S_{vib.,com}$ of the complexes of the trypsin dataset and is 0.17 cal^{-1} mol K; c is not a fitting coefficient but rather scales the magnitude of $S_{vib.}$ towards that of $-F^{(1)}$. The same c value was then applied to all other protein-ligand complex datasets. Note that computing $S_{vib.,lig}$ for small molecules is computationally cheap and does not impair much the overall computational efficiency of computing Δ-$F^{(1)}$.

4.11 Multiple linear regression

The following was performed for the BEERT project and the following text is taken from the manuscript "Efficient approximation of ligand rotational and translational entropy changes upon binding for use in MM-PBSA calculations" by Ben-Shalom I. Y., Pfeiffer-Marek S., Baringhaus K.H., and Gohlke H. (submitted).

A linear combination of $\Delta G_{\text{eff.}}$ from MM-PBSA and $T\Delta S_{\text{config.}}$ from BEERT was used as an approximation to the binding free energy (eq. 37). The coefficients in eq. 34 were determined by multiple linear regression against experimental ΔG_{bind}.

$$\Delta G_{predicted} = a\Delta G_{eff.} + b\left(-T\Delta S_{config.}\right) + c \qquad 37$$

As previously shown, for different proteins, the ratio between enthalpic and entropic contributions to the binding free energy is different.[340-342] Therefore, multiple linear regression was performed separately for each dataset.

4.12 Quality measures and error estimates

The following was performed for both projects and the following text is taken from the manuscripts "Efficient approximation of ligand rotational and translational entropy changes upon binding for use in MM-PBSA calculations" by Ben-Shalom I. Y., Pfeiffer-Marek S., Baringhaus K.H., and Gohlke H. (submitted) and "Rigidity theory-based approximation of vibrational entropy changes upon binding to biomolecules" by Gohlke H., Ben-Shalom I. Y., Kopitz H., Pfeiffer-Marek S., and Baringhaus K. H..

The results of the multiple linear regression were evaluated in terms of the coefficient of determination (R^2) between experimental and predicted binding energies, Fisher's F value, and the root mean squared error $S = \sqrt{(\text{RSS}/[\text{n-h-1}])}$, where RSS is the sum of squared differences between fitted and experimentally determined binding affinities, n is the sample size, and h is the number of regressors in eq. 34.[101] For statistical validation, we calculated coefficients of determination for a leave-one-out cross-validation (q^2). To do so, for each ligand in a dataset, coefficients in eq. 34 are determined by multiple linear regression for all

but this ligand, and the ligand is then used for testing. q^2 is then computed as 1-PRESS/SSD, where PRESS equals the sum of squared differences between and experimentally determined binding affinities and SSD is the sum of the squared differences between experimentally determined binding affinities and the mean of the training set binding affinities. The root mean squared error is $S_{PRESS} = \sqrt{}(PRESS/[n-h-1])$ in this case. [101] Coefficients of determination were also calculated for Y-randomization (Y-R^2), where the values of the experimental ΔG_{bind} were randomly shuffled prior to performing the multiple linear regression (eq. 34). This randomization process tells one how well random values could be fitted by our model, and is a negative control. [343] *p* values were computed using the program R, using a specific form of the *F*-test that compares the predicted model to the experimental data. [311]

Coefficients of correlation (r^2) were determined by comparing DrugScore scores, MM-PBSA effective energies, BEERT entropies, and the MW each separately against experimental $\Delta G_{bind.}$. In addition coefficients of correlation were determined by comparing $S_{vib.}$ (eq. 29) or $\Delta S_{vib.}$ (eq. 36) calculated using NMA to $-F^{(1)}$ (eq. 32) or Δ-$F^{(1)}$ (eq. 33) calculated by constraint counting, respectively. r^2_{max} is the maximal achievable correlation between ΔG_{bind} and a computational prediction, given the experimental uncertainty and the standard deviation of experimental $\Delta G_{bind.}$ (see eq. 17 in ref. [8]). We calculated 95% confidence intervals for r^2 by performing bootstrapping as previously done by us [331] using the boot package [344] of the program R [311] and 10,000 bootstrap replicas, employing bias-corrected, accelerated percentile intervals. *p* values were computed using the program R, using the *F*-test on the residual sum of squares. [311]

Reported uncertainties of the docking results and the different energy terms as well as $S_{vib.}$, $\Delta S_{vib.}$, $-F^{(1)}$, and Δ-$F^{(1)}$ are the standard error of the mean (SEM), i.e., the standard deviation divided by the square root of the number of samples. [345] Error propagation for eq. 34 was computed according to refs. [346, 347] (eq. 38)

$$SEM_{Total} = \sqrt{a^2\left(SEM_{\Delta G_{eff.}}\right)^2 + b^2\left(SEM_{\Delta S_{R/T}}\right)^2} \qquad 38,$$

regarding the uncertainty in the coefficients *a* and *b*.

Alternatively, error propagation for $S_{vib.}$ (eq. 29), $\Delta S_{vib.}$ (eq. 36), $-F^{(1)}$ (eq. 32), and Δ-$F^{(1)}$ (eq. 33) was considered according to refs. [346, 347] (eq. 39)

$$SEM_{Total} = \sqrt{(SEM_{com.})^2 + (SEM_{rec.})^2 + c^2(SEM_{lig.})^2} \quad 39$$

where c is the scaling coefficient, *i.e.*, 0.17 cal^{-1} mol K when $S_{vib.,lig}$ values are used for computing Δ-$F^{(1)}$ for protein-small molecule complexes, and 1.00 otherwise.

5 Results and Discussion

5.1 Datasets used for validation

The BEERT pipeline was evaluated on the HIV-1 protease, FXa, and Hsp90 datasets and the following is taken from the manuscript "Efficient approximation of ligand rotational and translational entropy changes upon binding for use in MM-PBSA calculations" by Ben-Shalom I. Y., Pfeiffer-Marek S., Baringhaus K.H., and Gohlke H. (submitted).

HIV-1 protease is a homodimeric aspartic protease with two extended *β*-hairpin loops, which are flexible and open and close upon substrate binding. [348] The HIV-1 protease dataset is composed of 20 complex structures with inhibitors with molecular weights of 500-750 Dalton and experimental pK_i values in the range of 7 to 12. PDB IDs and ligand properties for all complexes are provided in Table S1 in the Supporting Information (SI). All HIV-1 protease inhibitors in our dataset are asymmetric, and all contain a secondary hydroxyl group as the transition state-mimicking unit (Figure S1 in the SI). They decompose into subsets with different molecular scaffolds: (hydroxyethylamino)sulfonamides, [12, 349-352] *N*-phenyloxazolidinone-5-carboxamides, [353] 2-methyl-3-hydroxybenzamides, [12] and azaphenylalanines. [12]

FXa is a trypsin-like serine protease with a globular catalytic domain. The binding site is created by a relatively shallow cleft between two *β*-barrel subdomains. The active site is mostly hydrophobic with an aspartic acid important for the recognition. [186] The FXa dataset is composed of 20 complex structures with inhibitors with molecular weights of 400-600 Dalton and experimental pK_i values in the range of 6-10. PDB IDs and ligand properties for all complexes are provided in Table S2 in the SI. The FXa inhibitors are all asymmetric and are composed of different subsets (Figure S2 in the SI): One group is formed by β-amino ester derivatives, with ligands containing benzamidine, aminomethyl-biphenyl, [354] or pyridine *N*-oxide moieties. [355] A second group contains a sulfonamide moiety linked to thienopyridine [354]or a 3-amino-2-pyrrolidinone scaffold, [356-359] or as a part of a ring in the form of sulfonylpiperazinone. [360] A third group contains a 3-amidinobenzyl-1*H*-indole-2-carboxamide scaffold [361] or are indole-2-carboxylic acid-based. [362]

Hsp90 is an ATP-dependent chaperone. It is a homodimer where each monomer is composed of an N-terminal ATPase domain, a middle domain, and a C-terminal dimerization

domain. [189] The N-terminal binding site in the ATPase domain is formed by four helices shaping a compact cavity. [190, 197, 287] The Hsp90 ligand dataset contains 17 complex structures with inhibitors binding to the N-terminal ATPase domain. The ligands are ATP mimetics, with molecular weights in the range of 150-500 Dalton and experimental pIC_{50} values in the range of 3-8. PDB IDs and ligand properties for all complexes are provided in Table S3 in the SI. The Hsp90 inhibitors are all asymmetric and are composed of different subsets (Figure S3 in the SI): purine-based inhibitors, [287] molecules in which the purine scaffold is reduced to a pyrazole ring, [286] thienopyrimidines and analogs including triazin-based and phenyldiazenyl-pyrimidin-based ligands, [282] as well as diaryl isoxazole-, 1-(2-hydroxyphenyl)-2-naphthol-, [363] and resorcinylic isoxazole amide-based ligands. [364]

Datasets used in previous studies aiming at predicting binding free energies often showed a strong correlation between molecular weight and experimental binding free energies. [17, 70, 365] Hence, for such datasets, binding free energy predictions that show a ligand size-dependency can yield fair but trivial results. [17] For the three datasets used here, the coefficients of correlation between molecular weight and experimental binding free energies, derived from Tables S1, S2, and S3 in the SI, are 0.06 (HIV-1 protease), 0.06 (FXa), and 0.44 (Hsp90; bootstrapped 95% confidence interval: $0.03 < r^2 < 0.79$) (Table 2), excluding outliers as discussed in the next chapter. The related p values indicate a non-significant correlation for the HIV-1 protease and FXa datasets; in the case of the Hsp90 dataset, $p = 0.02$ indicating a significant correlation. The same result holds if the logarithm [366, 367] or square root [368] of the molecular weight is correlated with the experimental binding free energies (data not shown).

The vibrational entropy project was evaluated in addition on the trypsin dataset and on the protein-protein dataset and the following text is taken from the manuscript “Rigidity theory-based approximation of vibrational entropy changes upon binding to biomolecules” by Gohlke H., Ben-Shalom I. Y., Kopitz H., Pfeiffer-Marek S., and Baringhaus K. H..

Trypsin belongs to the trypsin family of serine proteases, like FXa. It contains the same catalytic triad, Asp102, His57, and Ser195, and the an the same Asp189, which is important for the recognition. [186, 208]. The trypsin dataset comprises 23 trypsin – small molecule complex structures used in ref. [335] (PDB IDs: 1C5S, 1F0T, 1G36, 1K1N, 1K1O, 1K1P, 1MTW, 1O2K, 1O36, 1QB6, 1QBO, 1QCP, 1RXP, 1S0R, 1TX7, 1V2N, 2AYW, 2FX4, 2OTV, 2ZDK, 2ZDL, 2ZDN, 2ZFS). The structures cover a broad spectrum of ligands with varying size,

ranging from benzamidine, which only fills the S1 pocket, to Crc200 (Chiron-Behring), which forms multiple polar and hydrophobic interactions within the trypsin active site. Binding affinities span a range from ~25 μM to ~10 nM, i.e., about 3.4 log units (see Figure S4 in the SI). The dataset of protein-protein complexes comprises four antibody-antigen, four protease-protease inhibitor, and two signal transduction complexes (PDB IDs 1CHO, 1DVF, 1GUA, 1LFD, 1MLC, 1PPF, 1VFB, 2JEL, 2PTC, 2SIC).

Chapters 5.2 until 5.8 were performed for the BEERT project and the following text is taken from the manuscript "Efficient approximation of ligand rotational and translational entropy changes upon binding for use in MM-PBSA calculations" by Ben-Shalom I. Y., Pfeiffer-Marek S., Baringhaus K.H., and Gohlke H. (submitted).

5.2 Sampling the configurational space of bound ligands by molecular docking

For sampling energy wells in the bound state according to the predominant states approximation, [243] we applied molecular docking using AutoDock 3.05 [244] as a search engine and DrugScore pair potentials [98] as an objective function. [80] As a means to probe the validity of this procedure, we evaluated to what extent the molecular docking identifies the global minimum on the energy surface of the bound state, assuming that this is represented by the crystal structure of the protein-ligand complex. To do so, for each protein-ligand complex, the lowest energy docking pose in the largest cluster was compared to the ligand pose in the crystal in terms of the RMSD. Results for the single complexes are shown in Tables S4, S5, and S6 in the SI for all three datasets. To assess the statistical significance of this result, the comparison was done for five independent docking runs, resulting in a SEM of the RMSD of < 2 Å in all cases.

Following the convention that a docking result can be considered "good" if its RMSD over all non-hydrogen atoms is < 2.0 Å to the crystal structure, [369, 370] we obtain "good" results in 100% of the cases for the HIV-1 protease and FXa datasets, and 90% for the Hsp90 dataset (Tables S4, S5, and S6 in the SI). This result is superior to the docking success reported for the combination of AutoDock/DrugScore on a mixed dataset of protein-ligand complexes [80] but resembles very well docking results obtained when using DrugScore as a scoring function on another dataset of HIV-1 protease (100% of the 52 docked ligands resulted in a predicted

pose < 2 Å RMSD from the crystallographic structure). [371] DrugScore also showed a success rate of 80% in recognizing the native pose of HIV-1 protease and FXa ligands. [365] On an Hsp90 dataset, 72% of the docked poses were within 1 Å RMSD to the crystal structure, which is a more stringent criterion than ours. [372] Overall, these results demonstrate that the representation of the energy surface of the bound state by DrugScore and the use of AutoDock for sampling it is highly suitable for identifying the experimental global energy minimum of the bound state in these cases of rigid protein/flexible ligand re-docking experiments.

In order to examine to what extent we sample bound configurations by docking, for each protein-ligand complex we ordered the clusters of docking poses according to their size and calculated the RMSD of the lowest energy pose from the crystal structure. For each dataset, we averaged the size of the cluster as well as the calculated RMSD value of respective clusters over all complexes. Figure 15 reveals that poses of the largest cluster are always most similar to the crystal structure, whereas the smaller clusters tend to have higher RMSD values (up to 3.5 Å RMSD). A dense sampling of bound ligand configurations in the vicinity of the global minimum is indicated by the fact that different clusters showing similar RMSD values from the crystal structure are present. Furthermore, the smaller clusters usually show high standard deviations (up to ~3 Å RMSD) (Figure 15), indicating that across a dataset ligands occupy different energy wells of the binding (free) energy landscape to a different extent.

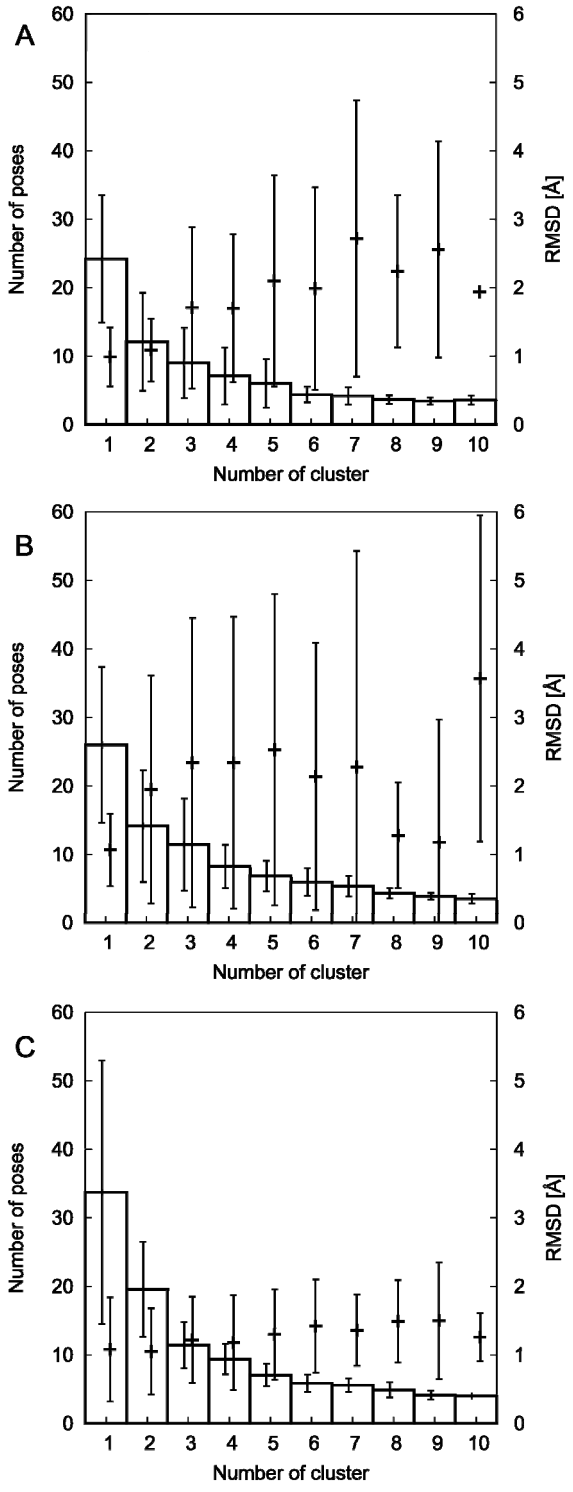

Figure 15: Assessment of the sampling of bound ligand poses by docking. Depicted is the size of the clusters of docked ligand poses (bars, sorted with respect to the size along the abscissa) and the RMSD of the lowest energy pose with respect to the crystal structure, averaged over respective clusters obtained for all ligands of the datasets of (A) HIV-1 protease, (B) FXa, and (C) Hsp90. Error bars depict the standard deviation.

For some of the protein-ligand complexes the docking resulted in ligand poses that do not seem suitable for identifying residual ligand motions in the binding pocket because the docked poses were either too similar, i.e., they were located in a single energy minimum, or too diverse, i.e., pronounced energy minima could not be recognized. Hence, first, we required for a docking result for the FXa and Hsp90 datasets to be taken into account for BEERT calculations that the largest cluster contains at least 20% of the docked ligand poses. Such a criterion has been applied previously by us as an indicator for convergence of the docking. [99] From the FXa dataset, 15% of the ligands did not fulfill this criterion (PDB IDs 1LPK, 1LPG, 1LPZ, and 2J95); from the Hsp90 dataset, 6% of the ligands did not fulfill this criterion (PDB ID 1UYH) (Figure 16). For the HIV-1 protease dataset, no threshold was considered because the binding pocket and the ligands are large, resulting in many degrees of freedom and therefore more diverse sets of ligand poses. [373] Second, in one case (6%; Hsp90 complex with PDB ID 2WI2), the docking resulted in a single cluster with essentially indistinguishable ligand configurations representing a single, very narrow energy minimum (Figure 16). As this ligand has a rather low binding affinity (IC_{50} = 350 μM), such a tight binding mode seemed unrealistic to us. We thus discarded this complex from further analyses. Finally, in one case (5%; HIV-1 protease complex with PDB ID 3EKV), the docking resulted in single pose clusters for > 50% of the generated poses, although one large cluster was present. As we are unable to compute $\Delta S_{\text{config.}}$ according to eq. 24 for single poses in energy wells, we discarded this complex from further analyses.

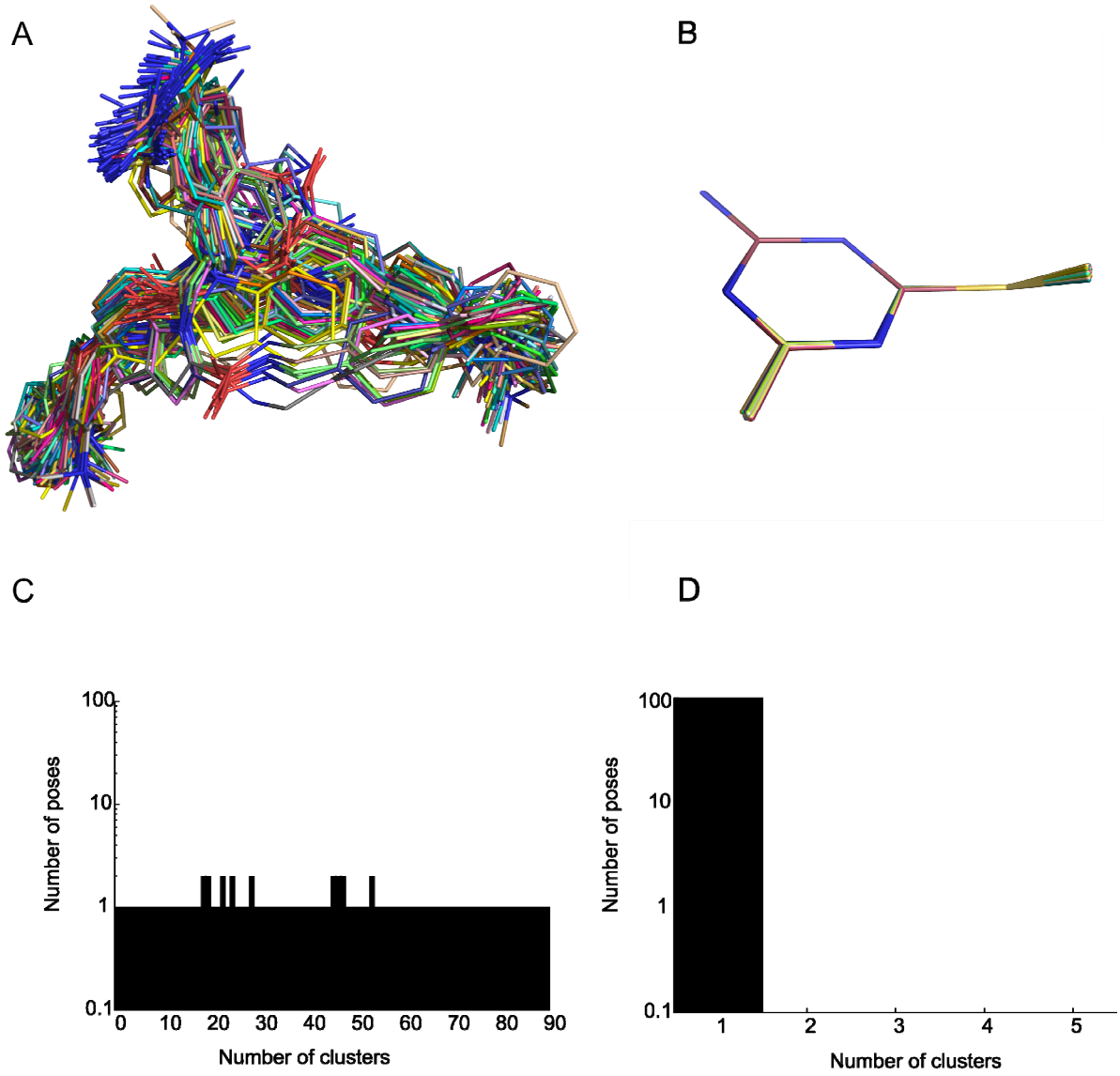

Figure 16: One hundred ligand poses generated by docking with AutoDock using DrugScore as the objective function. [80] (A) The docked poses of the ligand taken from PDB code 2WI2 (from the Hsp90 dataset). (B) The docked poses of the ligand taken from PDB code 1LPG (from the FXa dataset). The carbon atoms of each ligand pose are colored differently. (C) and (D) Clustering profiles according to panels (A) and (B), respectively.

5.3 Structural analyses of MD simulations

At the structural level, three RMSD values were computed for 12,500 snapshots extracted from trajectories of 250 ns length each:[374] For C_α atoms of the protein with respect to the

starting structure used for the MD simulation; this reveals conformational changes of the protein. For ligand heavy atoms with respect to the starting structure; this reveals conformational changes of the ligand. For ligand heavy atoms with respect to the starting structure after superimpositioning only the protein of the respective complex; this reveals potential movements of the ligand (plus conformational changes) with respect to the protein structure.

In the HIV-1 protease dataset, all protein structures that were further used for calculations show RMSD values < 2 Å, which compares favorably with other studies of MD simulations of HIV-1 protease complexes. [375] RMSD values of the ligand when the protein is fitted are < 5 Å in all cases, with 15 out of 19 structures showing an RMSD < 3 Å. For the remaining four structures, conformational changes of the ligand contributed to the RMSD, as shown by the ligand-internal RMSD of ~2 Å. For further details, see Table S7 and Figure S5 in the SI.

The structure of 2Q54 was shown to be unstable during the MD simulation. The protein RMSD was higher compared to the other proteins in the dataset, and the ligand leaves the binding pocket as revealed by the ligand RMSD > 10 Å after fitting the protein. The value also kept increasing in the course of the MD simulations (Figure S5 in the SI). This structure was therefore not used for further analysis.

In the FXa dataset, all protein structures show RMSD values < 2.5 Å, which compares favorably with other studies of MD simulations of FXa complexes. [376] RMSD values of the ligand when the protein is fitted are < 8 Å in all cases. Although some of the RMSD values are high, they are stable during the MD simulations and resulted to a large extent from conformational changes of the ligands, as shown by the ligand-internal RMSD of ~4 Å. For further details, see Table S8 and Figure S6 in the SI.

In the Hsp90 dataset, all protein structures show RMSD values < 4 Å, with 16 out of 18 structures showing RMSD values < 3 Å. The two remaining structures show structural modifications during the first 10-20 ns but then remain stable for the remainder of the MD simulations. This is comparable with other studies of MD simulations of Hsp90. [377] RMSD values of the ligand when the protein is fitted are < 3.5 Å in all cases. For further details, see Table S9 and Figure S7 in the SI.

5.4 Comparison of predicted and experimentally determined binding affinities

5.4.1 Predicting binding affinities using MM-PBSA effective energies

12,500 equally distributed conformations of the protein-ligand complexes were extracted from trajectories of 250 ns length each. The structural stability during the MD simulations was examined and is reported in the section 5.4 "Structural analyses of MD simulations", as well as in Tables S7, S8, and S9 and Figures S5, S6, and S7 in the SI. We calculated MM-PBSA effective energies following the single-trajectory approach for all three datasets. We first investigated the robustness and precision of the MM-PBSA computations on our datasets. The drift of the MM-PBSA effective energies over time was computed from the slope of the linear regression line and is very low, with an absolute value < 0.07 kcal mol^{-1} ns^{-1} for all structures and 41 out of 57 complexes (72%) having a value ≤ 0.02 kcal mol^{-1} ns^{-1} (Tables S10, S11, and S12 in the SI), demonstrating robust MM-PBSA computations. The SEM of each MM-PBSA effective energy calculation is < 0.001 kcal mol^{-1} for all complexes, demonstrating the high precision of the calculations. We next compared the relative effective energies computed by MM-PBSA for the three datasets to the experimental data (Tables S13, S14, and S15 in the SI). This resulted in weak and insignificant correlations for the HIV-1 protease and Hsp90 datasets ($r^2 = 0.02$ and $r^2 = 0.01$, respectively) and a moderate and significant correlation for the FXa dataset ($r^2 = 0.38$, bootstrapped 95% confidence interval: $0.04 < r^2 < 0.68$, $p < 0.05$) (Table 2). These results are in line with previous work by Yang *et al.*, who used MM-PBSA calculations in order to determine the binding energies of 156 ligands of six groups of protein families; correlations of $r^2 = 0.5$ between predicted and experimental binding energies were found for four out of six groups. [378] MM-PBSA effective binding energies were calculated for nine of the ligands in our HIV-1 protease dataset using MM-PBSA and compared to the experimental values; this resulted in a very low to no correlation, even after modifying the parameters of MM-PBSA. [350] Another study was performed on the FXa inhibitors containing the 3-amidinobenzyl-1*H*-indole-2-carboxamide scaffold. The ligands were examined using MM-PBSA. Calculations were performed on an initial scaffold, then the different ligands were built from this scaffold. This yielded a poor correlation with $r^2 = 0.22$, which was not statistically significant. [379, 380] Various computational methods (MM-GBSA, FEP, Docking) were used for the binding free energy prediction of Hsp90 inhibitors reaching a high

predictive ability of $r^2 \approx 0.7$ for the different methods. However, the r^2 values between the experimental binding free energies and the molecular weights were almost equally high. [381] This indicates that the prediction results may be related to the size of the molecules, too.

5.4.2 Predicting binding affinities by a linear combination of MM-PBSA effective energies and BEERT configurational entropies

Calculating the translational and rotational entropies for the three datasets resulted in an disfavorable contribution to the binding free energy ($-T\Delta S_{\text{config.}}$ at $T = 300$ K) ranging between 8-16 kcal mol^{-1} (see Tables S16, S17, and S18 in the SI). These values are in line with literature values for contributions resulting from restricting translation and rotation of ligands upon binding to a protein. [43, 340, 382, 383] Erikson estimated $-T\Delta S_{\text{config.}}$ resulting from the immobilization of an actin subunit when it is bound to the actin polymer as 7 - 11 kcal mol^{-1}. [382] Verkhivker *et al.* estimated a value of 11 kcal mol^{-1} associated with the translational and rotational entropy change for different protein-ligand complexes of HIV-1 protease. [340] On the upper side of the scale, Chang *et al.* estimated $-T\Delta S_{\text{config.}}$ that results from the association of amprenavir to HIV-1 protease as 15.7 kcal mol^{-1}. [47] Lower values were presented by Horton and Lewis who calculated $-T\Delta S_{\text{config.}}$ for 15 different protein-ligand complexes resulting in an average of 6.2 kcal mol^{-1}. [383]

Despite their high structural similarity, the ligands in each dataset display a broad range in $T\Delta S_{\text{config.}}$, thereby indicating that changes in the translational and rotational entropy cannot be neglected in binding energy predictions. Remarkably, in the case of ligand 895 (PDB ID 2UWL) and ligand 894 (PDB ID 2UWP) from the FXa dataset, which differ only in one bond order (Figure 17, Table 1), the binding affinity of the former is 4 nM, whereas that of the latter is 154 nM. However, $\Delta G_{\text{eff.}}$ values calculated by MM-PBSA for both complexes were very similar compared to the spread across all other FXa complexes (Table S14). $\Delta S_{\text{config.}}$ computed by BEERT resulted in a significant entropic penalty for ligand 894, which apparently can better adapt to the binding pocket due to the higher degree of flexibility, resulting in a stronger restriction of the translational and rotational degrees of freedom (Figure 17, Table 1). This effect is likely amplified by the larger restriction of conformational degrees of freedom for ligand 894 and can explain the difference in the experimental binding affinities between the two ligands (see Figure 17 and Table 1).

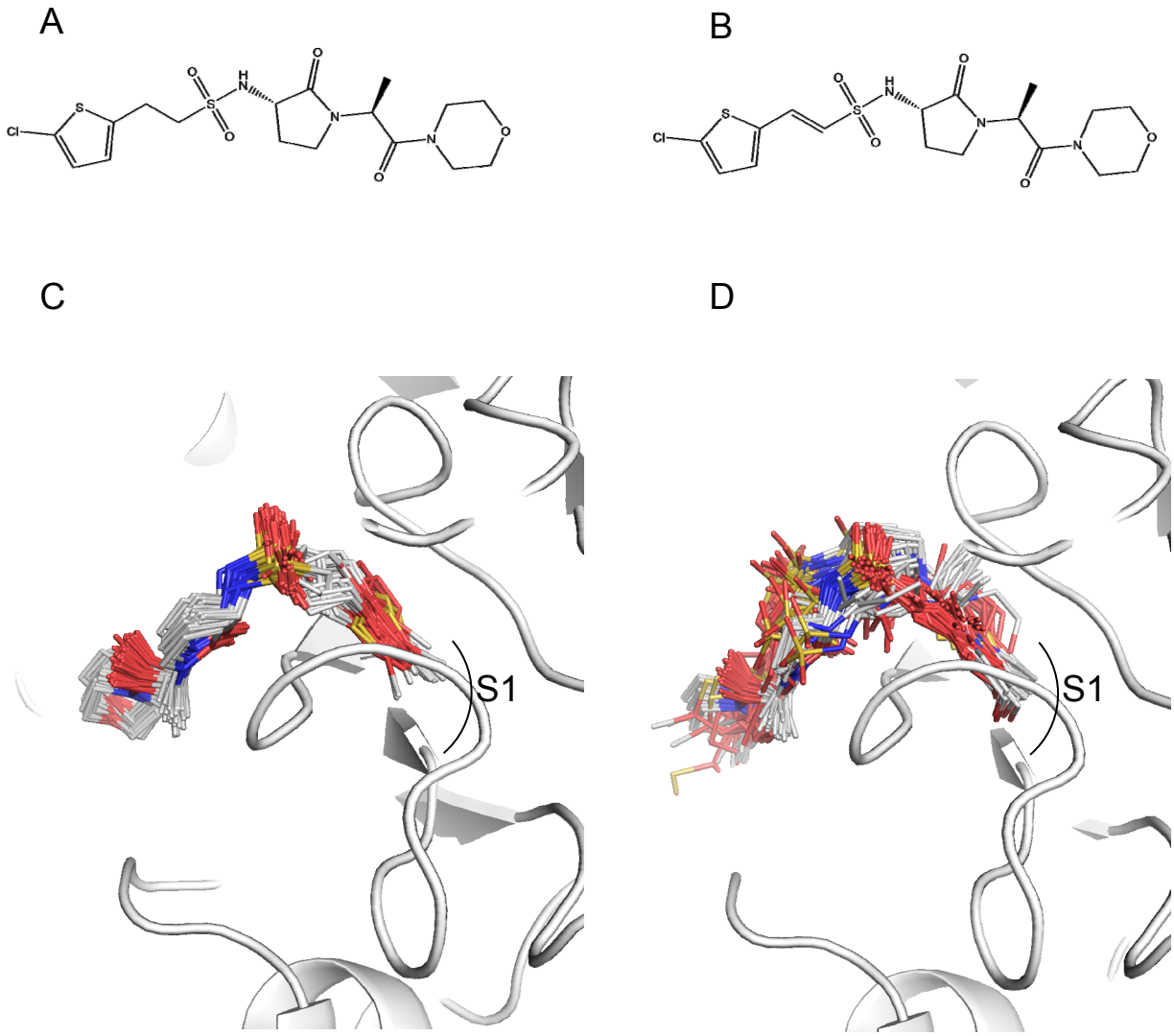

Figure 17: Similar ligands showing different residual mobility in the binding pocket. Structures of the ligands 894 (A) and 895 (B) taken from PDB codes 2UWP and 2UWL, respectively. One hundred poses of the ligands from panels (A) and (B) generated by redocking to the structures of 2UWP (C) and 2UWL (D).

Table 1: Comparison of the rotational and translational entropy changes computed by eq. 28 of the two similar ligands 894 (PDB ID 2UWP) and 895 (PDB ID 2UWL).

PDB code	894	895
Ω	0.252	0.573
V [Å^3]	0.004	0.051
$T\Delta S_{rot.}$ [kcal mol^{-1}]	-3.43	-2.94
$T\Delta S_{trans.}$ [kcal mol^{-1}]	-7.74	-6.19
$T\Delta S_{config.}$ [kcal mol^{-1}]	-11.17	-9.13
$\Delta G_{eff.}$ [kcal mol^{-1}]	-22.0	-23.7
$\Delta G_{bind.}$ [kcal mol^{-1}]	-12.9	-13.9
Experimental pK_I [nM]	154	4

To predict binding affinities for each dataset of protein-ligand complexes, we approximated $\Delta G_{bind.}$ by a linear combination of $\Delta G_{eff.}$ and $\Delta S_{config.}$ according to eq. 34. Multiple linear regression was performed separately for each dataset to incorporate that the ratio between enthalpic and entropic contributions to the binding free energy may be different for different proteins. [340, 341] The resulting coefficients are provided in Table 3. For the HIV-1 protease and FXa datasets, the coefficients for the MM-PBSA effective energies are ~10- and ~4-fold lower than the coefficients for $-T\Delta S_{config.}$, respectively, that way compensating that the MM-PBSA effective energies are larger in magnitude than $-T\Delta S_{config.}$ values (Tables S13, S14, S16, 17 in the SI). In the case of the Hsp90 dataset, the coefficient for the MM-PBSA effective energies is close to zero, corroborating that the MM-PBSA effective energies convey almost no information with respect to explaining $\Delta G_{bind.}$ (Table 3). The coefficients of determination R^2 between the fitted (eq. 34, Table 2) and experimental binding energies are 0.72 for the HIV-1 dataset ($F = 20.79$), 0.54 for the FXa dataset ($F = 7.65$), and 0.63 for the Hsp90 dataset ($F = 10.01$) (Table 2 and Figure 18), respectively.

To further check the statistical significance of the derived models, cross-validation runs were performed by means of the "leave-one-out" (LOO) procedure. For all three data sets, q^2 values > 0.3 were obtained (HIV-1 protease: 0.67; FXa: 0.34; Hsp90: 0.46) (Table 2), qualifying the models as "good". [384] The correlations of the predicted *versus* experimental ΔG are statistically significant ($p < 0.05$) in all three cases (Figure 19, Table 2), and the root mean squared error s_{PRESS} is ≤ 1.36 kcal mol^{-1} (Table 2) and, hence, only ~2-fold larger than the experimental uncertainties (Table 2) and close to the limit of chemical accuracy. It has been a matter of debate if internal validation, as performed by cross-validation here, suffices to judge the robustness of a model. [385, 386] For small datasets as used in our case, internal validation was suggested to be more appropriate, however, as information for deriving the model would be lost if the dataset were split to obtain an external test set. [387] Further indication for the significance of the obtained models is provided by the results obtained for the data sets with randomly scrambled experimental $\Delta G_{bind.}$. Here, the coefficients of determination R^2 are close to zero in all three cases (Table 2). Overall, these validations strongly suggest that the developed models are reliable and predictive.

Table 2: Results of statistical analyses related to scoring protein-ligand complexes after excluding outliers.

	Method												
	MW	MM-PBSA	MM-PBSA & BEERT						DrugScore	Surflex	MM-PBSA & RB	DrugScore & RB	
Dataset	r^2[a]	r^2[b]	R^2[c]	F[d]	S[e]	q^2[f]	s_{PRESS}[g]	Y-R^2[h]	r^2[i]	r^2[j]	R^2[k]	R^2[l]	r^2_{max}[m]
HIV-1 protease	0.06	0.02	0.72****	20.79	1.19	0.67****	1.36	0.04	0.03	0.02	0.12	0.08	0.79
FXa	0.06	0.38**	0.54***	7.65	0.78	0.34**	1.07	0.01	0.11	0.00	0.42*	0.18	0.99
Hsp90	0.44	0.01	0.63****	10.01	1.44	0.46**	1.25	0.08	0.14	0.37*	0.01	0.33	0.98

[a] Correlation between experimental ΔG and the MW of the ligands (Tables S1, S2, and S3 in the SI).
[b] Correlation of $\Delta G_{eff.}$ calculated by MM-PBSA based on MD simulations of 250 ns length with experimental ΔG (Tables S10, S11, and S12 in the SI).
[c] Coefficient of determination for a multiple linear regression according to eq. 34; sample size HIV-1 protease: $n = 18$; FXa: $n = 16$; Hsp90: $n = 15$.
[d] Fisher's F value.
[e] Root mean squared error, see text for definition. [101] In kcal mol^{-1}.
[f] Leave-one-out cross-validated coefficient of determination for the multiple linear regression according to eq. 34.
[g] Root mean squared error, see text for definition. [9] In kcal mol^{-1}.
[h] Coefficients of determination for the multiple linear regression according to eq. 34 after Y-scrambling of experimental ΔG values.
[i] Correlation of the DrugScore score obtained from the pair-wise distance-dependent potentials (eq. 5 in ref. [98]) with experimental ΔG (Tables S4, S5, and S6 in the SI).
[j] Correlation of the scoring function Surflex with experimental ΔG (Tables S4, S5, and S6 in the SI).
[k] Coefficient of determination for a multiple linear regression using $\Delta G_{eff.}$ and the number of rotatable bonds as $\Delta S_{config.}$ according to eq. 34.
[l] Coefficient of determination for a multiple linear regression using DrugScore as ΔH and the number of rotatable bonds as independent variables against experimental ΔG.
[m] The maximal achievable correlation in any computational method, considering the experimental uncertainty and the range of experimental ΔG values (eq. 17 in ref. [8]).
Statistical significance: *: $p < 0.01$, **: $p < 0.005$, ***: $p < 0.001$, ****: $p < 0.0005$.

Table 3: Coefficients of the multiple linear regression (eq. 34) for default parameters of the BEERT approach.

Dataset	Coefficients[a]		
	a	*b*	*c*
HIV-1 protease	0.12 (0.10, 0.11)	1.14 (1.08, 1.20)	-21.60 (-22.19,-21.24)
FXa	0.10 (0.08, 0.11)	0.45 (0.30, 0.54)	-12.58 (-13.20, -11.45)
Hsp90	-0.01 (-0.10, -0.01)	-0.58 (-0.36, -0.73)	-2.22 (-2.02, -2.45)

[a] Values in parentheses denote the 95% confidence interval for the regression coefficient.

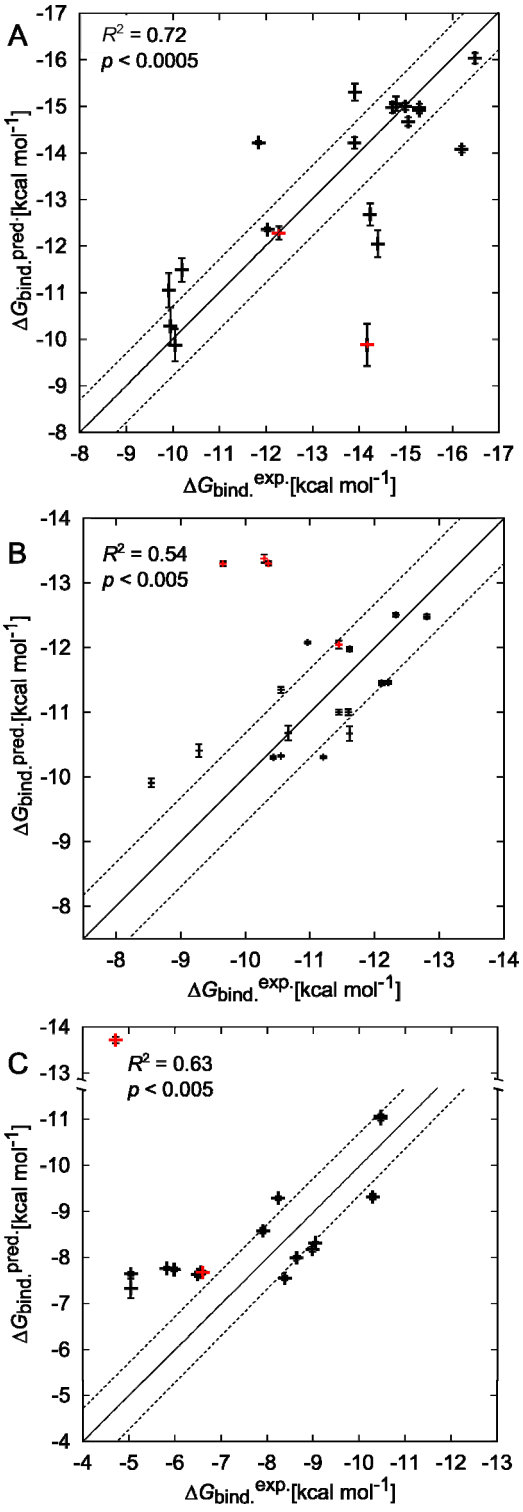

Figure 18: Scatter plots of fitted versus experimental ΔG values for the (A) HIV-1 protease, (B) FXa, and (C) Hsp90 datasets. Fitted values were computed according to eq. 34 using coefficients given in Table 3. Dashed lines depict uncertainty in the experimental ΔG_{bind} values. Datapoints excluded prior to performing the multiple linear regression are shown in red. The vertical error bars depict the SEM according to eq. 38.

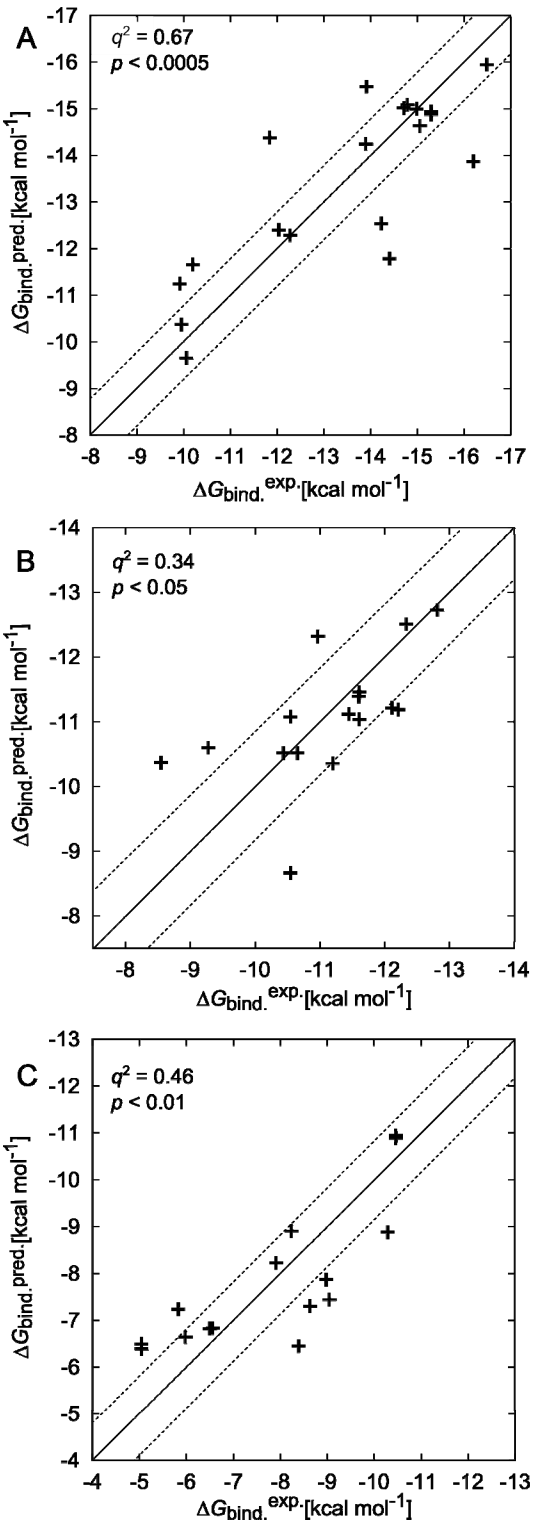

Figure 19: Scatter plots of predicted versus experimental ΔG values using a leave-one-out cross-validation for the (A) HIV-1 protease, (B) FXa, and (C) Hsp90 datasets. Dashed lines depict the experimental error range of the correlation.

5.5 Influence of the identification of energy wells on the regression results

In order to determine the influence of the identification of energy wells on the estimate of ligand translational and rotational entropy changes (eq. 24), we varied the parameters of the interaction-based clustering, the sizes of ensembles of binding poses to be clustered, and the weighting of the energy wells and repeated the multiple linear regression analyses according to eq. 34.

The distance threshold for defining interactions between protein and ligand atoms was varied between 3 Å and 7 Å in intervals of 1 Å. The lower bound is close to or below the sum of van der Waals radii of two atoms, and the upper distance allows to consider protein-ligand interactions mediated by one water molecule. Across all three datasets, the largest R^2 values are found for a distance threshold of 4 Å, with minor (FXa) or more pronounced (HIV-1 protease, Hsp90) changes of the coefficient of determination for values below or above (Table 4).

Table 4: Coefficients of determination (R^2) between experimental and fitted ΔG using different distance thresholds for interactions.[a]

	Distance[b]				
Dataset	3	4	5	6	7
HIV-1 protease	0.49	0.72	0.53	0.55	0.56
FXa	0.52	0.54	0.49	0.46	0.44
Hsp90	0.23	0.61	0.54	0.54	0.54

[a] Multiple linear regression according to eq. 34.

[b] Threshold that defines an interaction between a protein and a ligand atom; in Å.

The similarity between two ligand poses is determined as the fraction of common interactions that they share with the protein. The higher the similarity threshold is, the more similar are ligand poses within each cluster, and therefore more clusters are generated. The similarity threshold was varied between 0.2 and 0.4 in steps of 0.1 (Table 5). A threshold of 0.2 yielded the best regression results across all three datasets, which decreased at higher thresholds. Thresholds > 0.4 resulted in a small number of clusters (2-3) and therefore were not

considered. A threshold of 0.1 led to clusters being very small, often containing only a single pose, for which we are unable to compute $\Delta S_{\text{config.}}$ according to eq. 24 (data not shown).

Table 5: Coefficients of determination (R^2) between experimental and fitted ΔG using different similarity thresholds for interaction-based clustering.[a]

	Threshold[b]		
Dataset	0.2	0.3	0.4
HIV-1 protease	0.72	0.53	0.50
FXa	0.54	0.47	0.44
Hsp90	0.61	0.52	0.50

[a] Multiple linear regression according to eq. 36.

[b] Threshold used for similarity determination as the fraction of common interactions that two poses share.

The minimum number of poses in a cluster required to consider it an energy well was modified between 2 and 5. Single pose clusters were not tested (see above); thresholds > 4 resulted in the exclusion of more than 50% of the docked poses of the complexes in all datasets. The best regression results across all datasets were obtained when the minimum number of poses was 2 (Table 6). Thus, excluding clusters with more than one pose resulted in lower correlations.

Table 6: Coefficients of determination (R^2) between experimental and fitted ΔG using different minimal sizes of cluster.[a]

	Minimal cluster size[b]		
Dataset	2	3	4
HIV-1 protease	0.72	0.51	0.36
FXa	0.54	0.51	0.52
Hsp90	0.61	0.46	0.35[c]

[a] Multiple linear regression according to eq. 34.

[b] Minimal number of poses required to define a cluster as an energy well.

[c] R^2 was calculated based on 16 out of 18 complexes because for two of the complexes more than 50% of the docked poses were excluded.

Ensembles of binding poses with population sizes of 100, 500, and 1,000 were generated by docking in order to test how many binding poses are required to map the energy landscape of the ligand bound to protein. Although for all three ensemble sizes the regressions resulted in $R^2 \geq 0.41$, the best results were achieved for 100 ligand poses across all three data sets (Table 7). As a possible explanation, larger ensembles lead to more equally populated clusters across all ligands, that way blurring entropy differences between binding poses: The standard error of the mean for the cluster population is 0.53 and 0.52 for the ensembles of 500 and 1,000 poses, respectively, compared to 0.66 for the ensemble of 100 poses.

Table 7: Coefficients of determination (R^2) between experimental and fitted ΔG using different number of ligand poses generated by AutoDock.[a]

	Number of poses[b]		
Dataset	100	500	1.000
HIV-1 protease	0.72	0.42	0.41
FXa	0.54	0.59	0.54
Hsp90	0.63	0.56	0.56

[a] Multiple linear regression according to eq. 36.

[b] Number of poses generated by AutoDock and used for the calculation.

We tested three weighting functions (eqs. 20-22) for computing the entropy across energy wells as the weighted average of the entropies associated with an individual well (eq. 24). Across all three datasets, the largest R^2 values were obtained when the weights were computed from the occupancy of the clusters (eq. 21) (Table 8). This result likely reflects that, although the Boltzmann averaging used in eq. 20 is rigorous, it suffers from inaccuracies in the docking energies used to compute it. Considering the weight of each energy well equal (eq. 22) resulted in likewise worse results.

Table 8: Coefficients of determination (R^2) between experimental and fitted ΔG using different weighting methods for $\Delta S_{\text{config.}}$ calculation.[a]

	Method		
Dataset	Occupancy weighting[b]	Equal weighting[c]	Boltzmann weighting[d]
HIV-1 protease	0.72	0.58	0.46
FXa	0.54	0.45	0.53
Hsp90	0.61	0.54	0.49

[a] Multiple linear regression according to eq. 34.

[b] The cluster occupancy was used as weighting factor (eq. 21).

[c] All energy wells were equally weighted (eq. 22).

[d] Boltzmann-weighted averaging with respect to the docking energy computed by DrugScore was applied (eq. 20).

Finally, the interaction-based clustering was compared to RMSD-based clustering; the latter has been used by Ruvisnky *et al.* [70, 83] in the context of estimating changes in the translational and rotational entropy upon ligand binding. When inspecting docking poses clustered by RMSD as implemented in AutoDock, we observed that poses with similar binding modes often belonged to different clusters. This resulted from them showing different conformations of the solvent-accessible parts, while their bound parts were similar (Figure 20). The interaction-based clustering developed here was inspired by interaction fingerprints, which have been used previously to represent 3D pharmacophore interactions. [251, 252] As a major difference between these uses and ours, we compare multiple binding poses of one ligand in a protein rather than aiming at comparing types of interactions formed by different ligands. Using RMSD-based clustering with an RMSD threshold of 1 Å yielded R^2 values in the regressions (eq. 34) of 0.08 and 0.09 for the HIV-1 protease and Hsp90 datasets, respectively, and 0.48 for the FXa dataset (Table 9). Using the interaction-based clustering yielded $R^2 \geq 0.54$ across all datasets. Thus, clustering ligand poses based on their bound parts only is superior, likely because then both the restricted mobility of the bound parts and the residual mobility of the solvent-accessible parts can be correctly identified, which results in better $\Delta S_{\text{config.}}$ estimates.

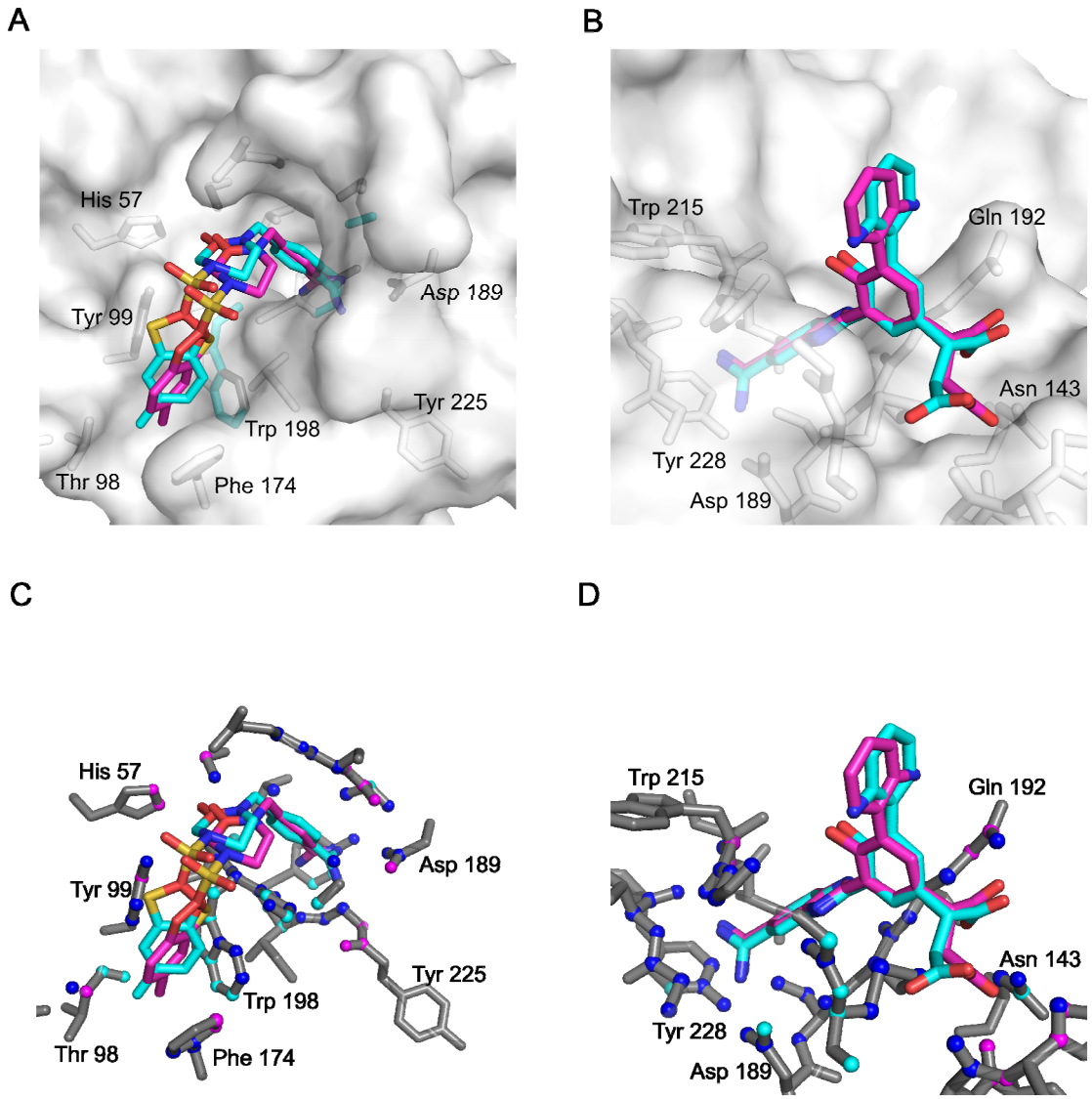

Figure 20: Advantages of interaction-based clustering. (A) Structure of FXa (PDB code: 1NFY) in surface representation with two docked ligand poses (ligand **RTR**). The RMSD between the two ligand poses is ~2.5 Å, resulting mainly from the rotation of the solvent-exposed chlorobenzothiophene substituent. (B) Structure of trypsin (PDB code: 1O36) in surface representation with two docked ligand poses (ligand **607**). The RMSD between the two ligand poses is ~2.0 Å, resulting mainly from the rotation of the solvent-exposed aniline substituent. Both ligand poses in panels A and B introduce relatively high RMSD values despite overall similar binding modes. (C) and (D) Identical binding poses of the ligand **RTR** in the structure of FXa (PDB code: 1NFY) (C) and of ligand **607** in the structure of trypsin (PDB code: 1O36) (D) as in panels A and B, respectively. Protein atoms interacting only with one ligand pose are colored in magenta (24% and 16% of the total

interacting atoms for A and B, respectively), protein atoms interacting with the other ligand pose are colored in light blue (20% and 13% of the total interacting atoms for A and B, respectively), and protein atoms interacting with either one of the ligand poses are colored in blue (56% and 71% of the total interacting atoms for A and B, respectively).

Table 9: Coefficients of determination (R^2) between experimental and fitted ΔG using different clustering methods.[a]

	Clustering method	
Dataset	Interaction clustering[b]	RMSD clustering[c]
HIV-1 protease	0.72	0.08
FXa	0.54	0.48
Hsp90	0.61	0.09

[a] Multiple linear regression according to eq. 34.

[b] Interaction-based clustering was used, applying the optimal parameters identified in Tables 4 - 9.

[c] RMSD-based clustering was used as implemented in AutoDock, applying an RMSD threshold of 1 Å.

In summary, the best results in the regressions (eq. 36) were obtained using interaction-based clustering with a distance threshold of 4 Å, a similarity threshold of 0.2, a minimum number of poses in a cluster of 2, an ensemble size of 100, and a weighting based on the occupancy of a cluster (eq. 21). Results obtained with these settings are reported in Table 2.

5.6 Binding affinities predicted by DrugScore

For comparison, we predicted relative binding affinities for the docked poses obtained by AutoDock with DrugScore. This resulted in weak and insignificant correlations with experimental binding energies (HIV-1 protease: R^2 = 0.03; FXa: R^2 = 0.11; Hsp90: 0.14) (Table 2 and Tables S4, S5, and S6 in the SI). These results are at variance with a previous study where DrugScore was used for relative binding affinity predictions on the "Wang dataset", a dataset of 100 protein-ligand complexes from different protein and ligand groups, [388] which resulted in a fair correlation (Spearman's rank correlation coefficient R_S = 0.624). [17] DrugScore also showed a high ability in binding affinity prediction compared

to other scoring functions on five groups of protein-ligand complexes, each with 15-61 ligands, resulting in $0.35 < R^2 < 0.56$. [97] Another study showed the ability of DrugScore in predicting binding affinities on complexes of serine proteases, metalloproteases, and lyases, each with 10-25 ligands, resulting in $0.67 < R^2 < 0.87$. [365] The difference in the predictive power of DrugScore observed here and in the previous studies can likely be accounted to the difference in the used datasets: In the previous studies, the experimental binding energies covered a range of up to 16 kcal mol^{-1}. In our datasets, the experimental binding energies cover a range of ≤ 7 kcal mol^{-1}, a more realistic scenario in the context of lead optimization.

5.7 Binding affinities predicted by Surflex

For an additional comparison, we also predicted relative binding affinities by Surflex as an external scoring function for the docked poses obtained by AutoDock with DrugScore. Surflex was previously rated one of the best scoring functions in its ability to rank known inhibitors compared to other scoring functions on a dataset of 100 different protein-ligand complexes. [317] We used Surflex to score our existing docked poses, as it incorporates the number of rotatable bonds of the ligand as a measure for the change in configurational entropy. [128] Correlating experimental binding affinities to predicted Surflex scores resulted in weak and insignificant correlations for the HIV-1 protease and FXa datasets (r^2 = 0.02 and 0.00, respectively), and fair results for the Hsp90 dataset (r^2 = 0.37; bootstrapped 95% confidence interval: $0.05 < r^2 < 0.70$; $p < 0.05$) (Table 1; Tables S19, S20, and S21 in the SI). As in the case of scoring with DrugScore, these results are at variance with previous studies (see above) but can likely be explained with the different ranges of binding energies of the datasets used in these studies and here. Similar difficulties of scoring functions in correlating scores to binding affinities for congeneric series of ligands and/or ligand datasets with small ranges of binding affinities have been reported before. [389]

5.8 Number of rotatable bonds as a measure for the change in configurational entropy

The number of rotatable bonds has been used frequently as an approximate measure of changes in configurational entropy upon ligand binding in scoring functions.[22, 50-53] Here, we replaced $\Delta S_{config.}$ in eq. 36 by the number of rotatable bonds determined for each ligand (Tables S1, S2, and S3 in the SI) and performed separate multiple linear regression analyses

against experimental binding affinities of the three datasets. This resulted in R^2 values of 0.12 and 0.01 for the HIV-1 protease and Hsp90 datasets, respectively (Table 1, Table S29 in the SI). For the FXa dataset, $R^2 = 0.42$ ($p < 0.05$). Thus, for all three datasets, the use of the number of rotatable bonds yielded inferior results than when using $\Delta S_{\text{config.}}$ (eq. 21) in eq. 36. We also combined the number of rotatable bonds with DrugScore scores instead of $\Delta G_{\text{eff.}}$ in eq. 29 and performed multiple linear regression analyses. This resulted in R^2 values of 0.08, 0.18, and 0.33 for the HIV-1 protease, FXa, and Hsp90 datasets, respectively, again showing inferior models compared to using eq. 36 (Table 2 and Table 10). Thus, together with the results obtained for Surflex described above, using the number of rotatable bonds as a measure for the change in configurational entropy yielded regression models that are either not statistically significantly different from an intercept-only model or showed R^2 values that are markedly smaller than those obtained when applying $\Delta S_{\text{config.}}$ (eq. 28) in the context of eq. 36.

Table 10: Coefficients of determination (R^2) between experimental and fitted ΔG estimating the entropy term from the number of rotatable bonds.

Dataset	ΔH[a]	$\Delta G_{\text{eff.}}$[b]
HIV-1 protease	0.08	0.10
FXa	0.18	0.42
Hsp90	0.34	0.01

[a] Multiple linear regression according to eq. 36, where $\Delta S_{\text{config.}}$ is the number of rotatable bonds, and ΔH is DrugScore score.

[b] Multiple linear regression according to eq. 36, where $\Delta S_{\text{config.}}$ is the number of rotatable bonds, and $\Delta G_{\text{eff.}}$ is MM-PBSA effective energy.

5.9 Comparisson of vibrational entropy calculation using NMA and FIRST

The following is taken from the manuscript “Rigidity theory-based approximation of vibrational entropy changes upon binding to biomolecules” by Gohlke H., Ben-Shalom I. Y., Kopitz H., Pfeiffer-Marek S., and Baringhaus K. H..

The validity of eq. 31 and 33 was assessed on one dataset of protein-protein complexes and four datasets of protein-small molecule complexes by comparison to vibrational entropies computed by NMA. The dataset of protein-protein complexes comprises four antibody-antigen, four protease-protease inhibitor, and two signaling complexes with diverse folds, protein sizes between 775 and 8398 atoms, and binding affinities from the μM to pM range. Initially, we probed if $-F^{(1)}$ behaves as an extensive property, as required of an entropy-like quantity and confirmed in the case of insulin dimerization for $S_{vib.}$ computed by NMA. [36] We computed $-F^{(1)}$ for each complex, receptor, and ligand of the dataset members, resulting in 3 x 10 values. When plotted against the mass of the proteins, a significant and very good correlation results (r^2 = 0.92; bootstrapped 95% confidence interval: 0.84 < r^2 < 0.96; $p = 2.2 * 10^{-15}$; Figure 21), demonstrating a strong dependence of $-F^{(1)}$ on the system size, indicative of an extensive property.

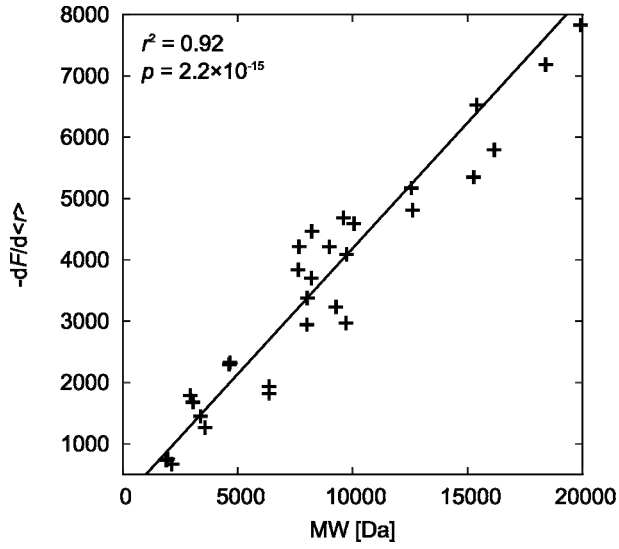

Figure 21: Correlation of $-F^{(1)}$ *versus* the protein mass for complexes, receptors, and ligands of the protein-protein complex dataset. In addition, the correlation line is shown.

Not surprisingly, $-F^{(1)}$ and $S_{vib.}$ of the 3 x 10 complexes, receptors, and ligands, respectively, also yield a very good correlation (r^2 = 0.95; data not shown). More importantly, we next correlated Δ-$F^{(1)}$ and $\Delta S_{vib.}$ (eq. 36), *i.e.* estimates of *changes* in the vibrational entropy upon binding, for the ten protein-protein complexes, yielding a significant and good correlation (r^2 = 0.80; bootstrapped 95% confidence interval: 0.19 < r^2 < 0.96; p = 0.0005; Figure 22A).

In contrast, $\Delta S_{vib.}$ correlated against the area of the epitope buried upon complex formation yields a weak correlation ($r^2 = 0.36$; data not shown). Together, this demonstrates that the good correlation of Δ-$F^{(1)}$ *versus* $\Delta S_{vib.}$ does not have a trivial, *i.e.* size-dependent, origin; rather, Δ-$F^{(1)}$ describes alterations in the density of vibrational states of the complexes relative to the binding partners apparently with good accuracy.

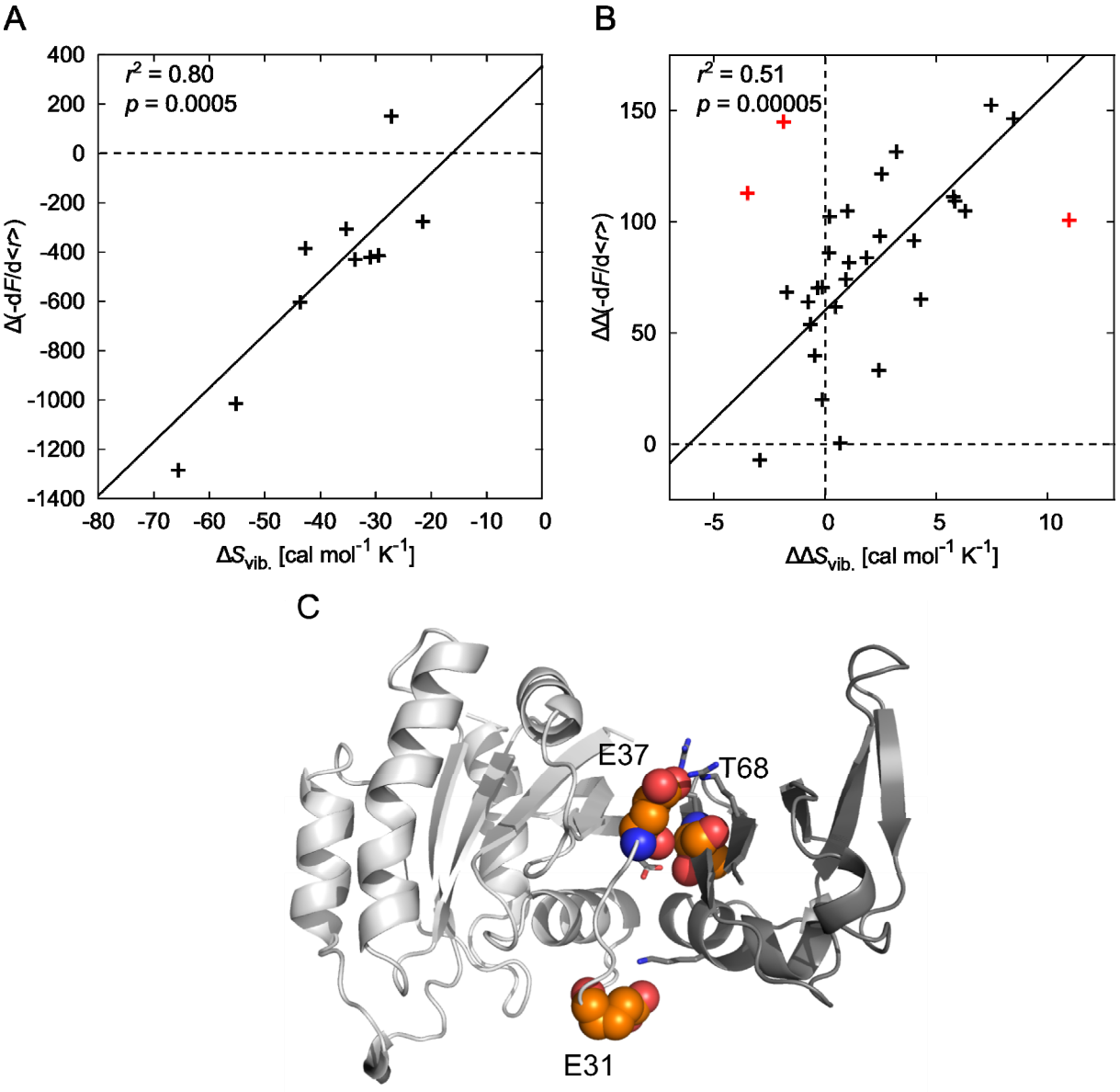

Figure 22: (A) Correlation of Δ-$F^{(1)}$ *versus* $\Delta S_{vib.}$ computed for ten protein-protein complexes. The average SEM of Δ-$F^{(1)}$ and $\Delta S_{vib.}$ are ~210 and ~10.0 cal mol^{-1} K^{-1}, respectively. (B) Correlation of $\Delta\Delta$-$F^{(1)}$ *versus* $\Delta\Delta S_{vib.}$ computed for 30 alanine mutations in the interface of Ras-Raf (PDB ID 1GUA) [329] using E_{cut} = -0.2 kcal mol^{-1}. The red symbols denote mutations E31$_{Ras}$, E37$_{Ras}$, and T68$_{Raf}$

considered outliers. The average SEM of Δ-$F^{(1)}$ and $\Delta S_{vib.}$ are ~70 and ~1.3 cal mol^{-1} K^{-1}, respectively. Dashed lines indicate Δ-$F^{(1)}$, $\Delta S_{vib.}$ = 0; the correlation line is represented by a straight line. (C) The structure of the of Ras-Raf complex (PDB ID 1GUA), where Ras is colored in white and Raf in gray. The three outliers corresponding to Figure 22B are depicted in spheres; residues interacting with them across the interface are depicted in sticks.

Note that significant and good correlations were also obtained if E_{cut} was set to -0.6 or -1.4 kcal mol^{-1} (r^2 = 0.63, 0.83; Figure 23), demonstrating that our approach is robust with respect to the choice of E_{cut}.

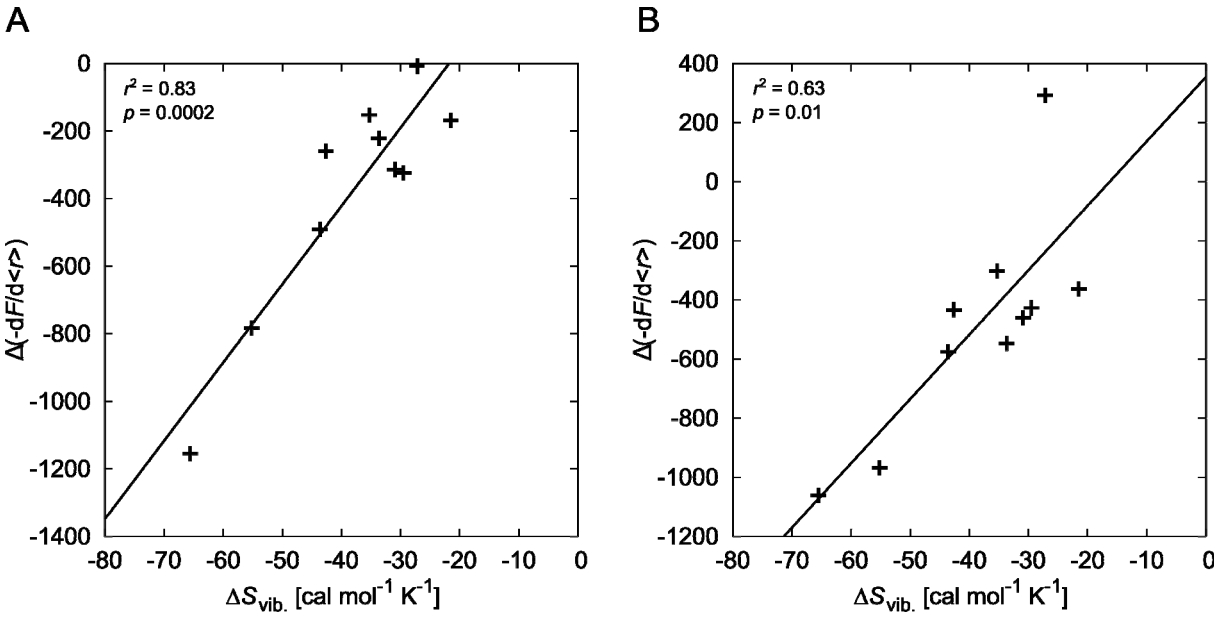

Figure 23: Correlation of Δ-$F^{(1)}$ *versus* $\Delta S_{vib.}$ computed for the protein-protein complex dataset using (A) E_{cut} = -1.4 and (B) -0.6 kcal mol^{-1}. In addition, the correlation line is shown.

Finally, we used structures directly extracted from the MD trajectories for computing Δ-$F^{(1)}$, rather than the minimized ones used as input for NMA. Not considering PDB ID 2JEL, Δ-$F^{(1)}$ of which deviates most between non-minimized and minimized structures, resulted in a correlation with r^2 = 0.54 (bootstrapped 95% confidence interval: 0.01 < r^2 < 0.96; p = 0.02;

Figure 24). Thus, despite structural deviations between respective conformations used for constraint counting and NMA, still a good correlation is obtained.

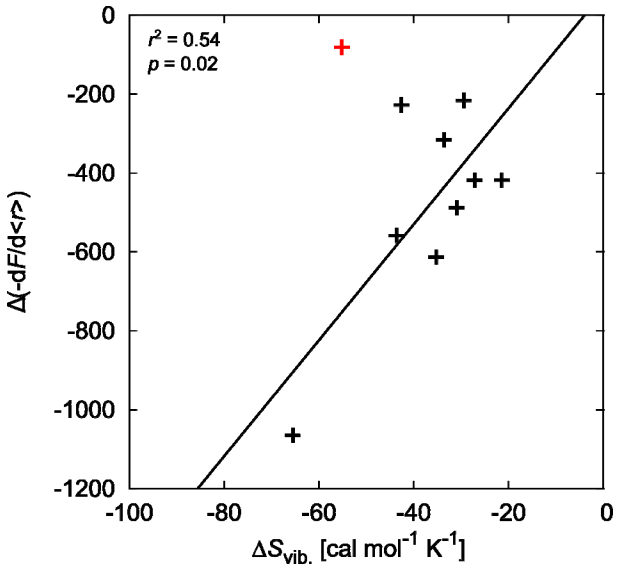

Figure 24: Correlation of Δ-$F^{(1)}$ *versus* $\Delta S_{vib.}$ computed for the protein-protein complex dataset using structures directly extracted from the MD trajectories for computing Δ-$F^{(1)}$, rather than the minimized ones used as input for NMA. In addition, the correlation line is shown. The red symbol denotes PDB ID 2JEL considered an outlier.

Computational alanine scanning allows from a single MD simulation an estimate of the individual contribution of each residue of a protein-protein complex to the binding and has proven valuable for identifying "hot spot" residues in protein-protein epitopes. [115, 329, 390-392] For the vibrational entropy contribution, the difference in the change of $S_{vib.}$ upon binding ($\Delta\Delta S_{vib.}$) is computed by NMA from the wild type and an alanine mutant; the mutant is generated from the wild-type conformations by removing respective atoms. $\Delta\Delta$-$F^{(1)}$ is computed analogously. The correlation of $\Delta\Delta$-$F^{(1)}$ *versus* $\Delta\Delta S_{vib.}$ for 30 alanine mutations in the interface of Ras-Raf (PDB ID 1GUA)[329] is significant and weak (R^2 = 0.24; bootstrapped 95% confidence interval: 0.01 < r^2 < 0.58; p = 0.01; Figure 25) if E_{cut} = -1.0 kcal mol^{-1} is used and three outliers are disregarded. The correlation can be markedly improved (R^2 = 0.51; bootstrapped 95% confidence interval: 0.22 < r^2 < 0.72; $p = 2.7*10^{-5}$; Figure 22B) if E_{cut} = -0.2 kcal mol^{-1} is used, again disregarding three outliers (residues E31$_{Ras}$, E37$_{Ras}$, and

$T68_{Raf}$; Figure 22C). Apparently, analyzing stiffer constraint networks is favorable here, likely because the $\Delta\Delta$-$F^{(1)}$ values for side chains on the protein surface become less noisy when contributions from the protein core become less pronounced. Side chains of residues $E37_{Ras}$ and $T68_{Raf}$ are involved in salt bridges or a polar hydrogen bond across the center of the epitope, and their influence on the vibrational entropy change is underestimated by constraint counting (Figure 22B). Residue $E31_{Ras}$ engages in a salt bridge interaction at the edge of the epitope, and its influence on the vibrational entropy change is overestimated by constraint counting (Figure 22B). Neglecting solvent influences or cooperative effects on the strength of the polar interactions in the energy function E_{HB} might cause these deviations. Note that the vibrational entropy contributions of most of the side chains in the Ras-Raf epitope disfavor binding, as indicated by $\Delta\Delta S_{vib.} > 0$. The $\Delta\Delta$-$F^{(1)}$ values mirror this finding for all but seven of the side chains (disregarding the three outliers).

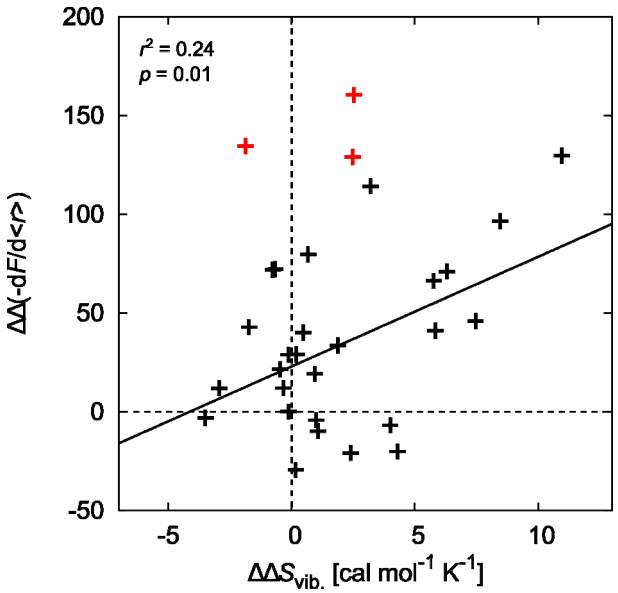

Figure 25: Correlation of $\Delta\Delta$-$F^{(1)}$ *versus* $\Delta\Delta S_{vib.}$ computed for 30 alanine mutations in the interface of Ras-Raf (PDB ID 1GUA) using E_{cut} = -1.0 kcal mol^{-1}; the $\Delta\Delta S_{vib.}$ values were taken from ref. [329]. In addition, the correlation line is shown. The red symbols denote three outliers.

As to the protein-small molecule complexes, the trypsin dataset encompasses 23 complexes with ligands ranging in size from filling only the S1 pocket to those capturing the entire active site, and binding affinities covering a range of ~3.4 log units. The correlation of Δ-$F^{(1)}$ *versus*

$\Delta S_{vib.}$ is significant and fair (r^2 = 0.40; bootstrapped 95% confidence interval: $0.09 < r^2 < 0.66$; $p = 0.001$; Figure 26A). Some ligands lead to $\Delta S_{vib.} > 0$, whereas others show $\Delta S_{vib.} < 0$. Ligands of the former group are usually small and make few interactions with the protein (Figure 26D), allowing for librational motions of the ligand;[393] in contrast, those of the latter group usually make many interactions with different parts of the protein (Figure 26C), stiffening the protein.[394] Notably, this distinction between ligands is almost perfectly reflected in the Δ-$F^{(1)}$ values (Figure 26A), revealing that constraint counting can distinguish between ligand binding that leads to favorable *versus* disfavorable vibrational entropy contributions to the binding affinity. This property is of high importance when ranking potential ligands.[36] The Factor Xa dataset contains 20 complex structures with small molecule ligands that are more similar in size (400 – 600 Da) and show a narrower distribution of binding affinities (range: ~2.7 log units). As an additional challenge, the dataset contains both ligands that form the well-known salt bridge with Asp189 in the S1 pocket and those that place non-polar moieties there. The correlation of Δ-$F^{(1)}$ *versus* $\Delta S_{vib.}$ is significant and fair ($r^2 = 0.46$; bootstrapped 95% confidence interval: $0.06 < r^2 < 0.74$; $p = 0.001$; Figure 26B and Table S23 in the SI). Again, both Δ-$F^{(1)}$ and $\Delta S_{vib.}$ distinguish between ligand binding that leads to favorable *versus* disfavorable vibrational entropy contributions to the binding affinity (Figure 26B). A ligand of the former group is IIA (PDB ID 2BOH), which places a chlorothiophen moiety into the S1 pocket (Figure 26F), one of the latter group is IMA (PDB ID 1LPG), which places a benzamidine moiety there (Figure 26E). As the ligands are otherwise similar in size and interaction pattern with the protein, one can speculate that it is the locking-in of protein and ligand by a salt bridge that leads to disfavorable vibrational entropy contributions in contrast to the less restrictive interactions of the chlorothiophen moiety.

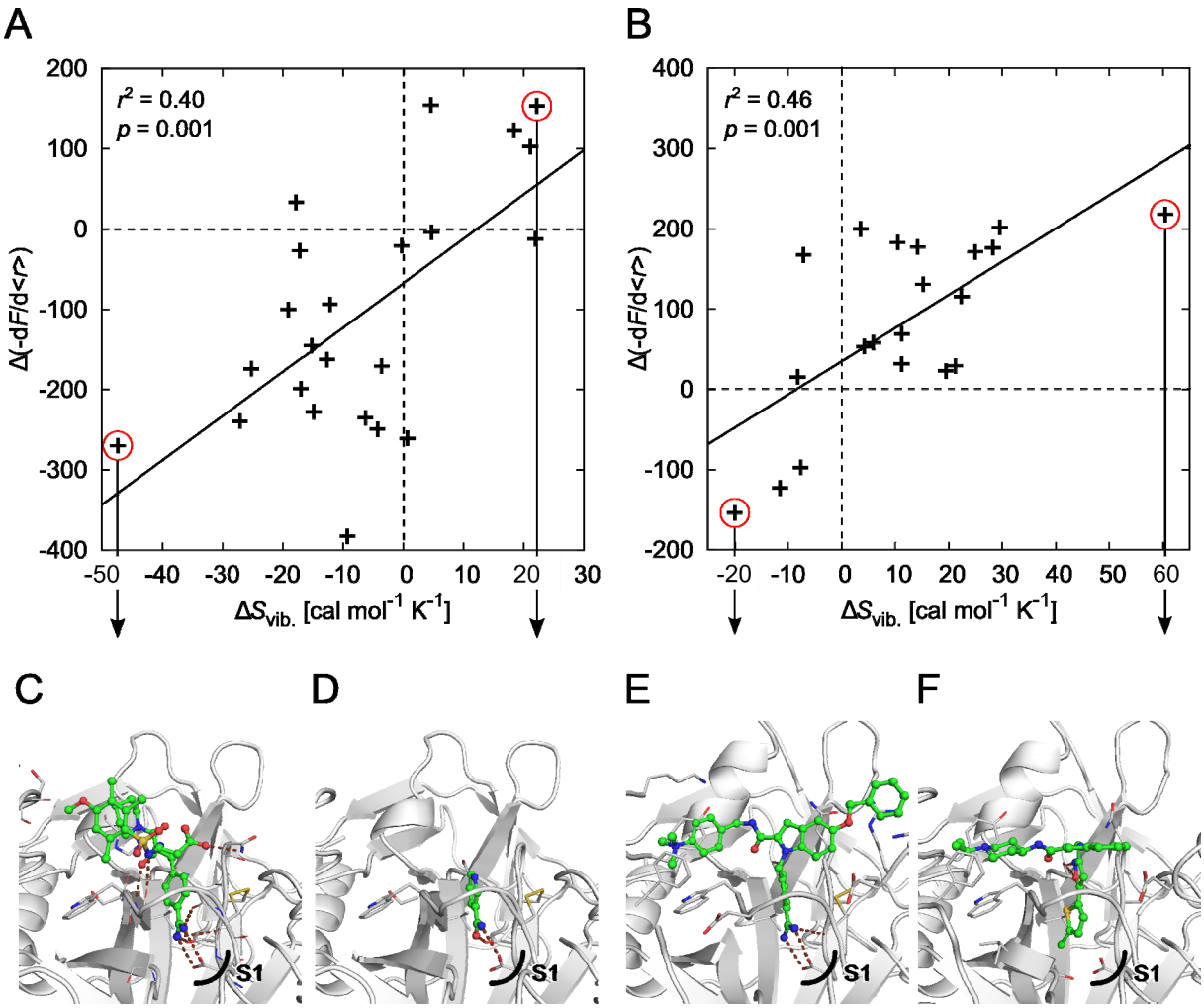

Figure 26: Correlation of Δ-F$^{(1)}$ versus Δ$S_{vib.}$ computed for protein-small molecule complexes of (A) the trypsin dataset and (B) the FXa dataset. Dashed lines indicate Δ-F$^{(1)}$, Δ $S_{vib.}$ = 0; the correlation line is represented by a straight line. Data points of complexes that result in maximal favorable or disfavorable vibrational entropy changes are circled, and the respective crystal structures of the complex are depicted: (C) ligands CRC200 (taken from PDB ID 1K1N) and (D) nicotinamide (taken from PDB ID 2OTV); (E) ligands IMA (taken from PDB ID 1LPG) and (F) IIA (taken from PDB ID 2BOH). The location of the S1 pocket in trypsin and FXa is indicated by a black arc; polar interactions between protein and ligand are depicted by red dashed lines. The average SEM of Δ-F$^{(1)}$ and Δ $S_{vib.}$ are 0.1 and 1.0 cal mol^{-1} K^{-1}, respectively.

Finally, we investigated two additional datasets of Hsp90- and HIV-1 protease-small molecule complexes. These datasets were rather similar to the trypsin and Factor Xa datasets with respect to the number of data points and the range of ligand sizes and binding affinities

(Hsp90 dataset: 16 complex structures with small molecule ligands ranging from 150 – 500 Da and binding affinities spanning 4.7 log units; HIV-1 protease dataset: 20 / 500 – 750 Da / 4.2 log units). As a major difference, however, the width of the distribution of $\Delta S_{vib.}$ computed by NMA across each dataset is only ~1/3 of that of the trypsin and Factor Xa datasets (~22-25 cal mol^{-1} K^{-1}), that way being very similar in magnitude to the average standard deviation of the computed $\Delta S_{vib.}$ (~23-29 cal $mol^{-1}K^{-1}$). Therefore, according to Kramer *et al.*, [8] the maximum possible squared Pearson coefficient of determination (r^2_{max}) on these datasets vanishes; in agreement, Δ-$F^{(1)}$ *versus* $\Delta S_{vib.}$ did not yield significant correlations (Tables S3 and S4 in the SI). Note that these last results restate the fundamental challenge of computing precise changes in vibrational entropy in general, [395] rather than showing a limitation of approximating them by Δ-$F^{(1)}$.

Using the "pebble game" algorithm provides a dramatic speed up. For a normal-sized protein (~250 residues), the computing time for determining F is ~8 s, compared to ~2.5 hours in NMA for the energy minimization and diagonalization of the Hessian, when both computations are performed on a single CPU core with 2.5 GHz.

6 Conclusions

The aim of my thesis was the development of a method for calculating the entropy change upon protein-ligand complex formation, for improving binding free energy predictions. To this end, my thesis contains two projects. The first project included the calculation of the translational and rotational entropy change of the ligand upon complex formation and the second project includes comparing the vibrational entropy change.

For the first project I developed the python-based BEERT (Binding Entropy Estimation for Rotation and Translation) software, for predicting the translational and rotational entropy change of a ligand upon binding. The varying extent of translational and rotational restriction upon binding results in differences in the translational and rotational entropy. In order to estimate the restriction, I first generated ensembles of ligand poses using the AutoDock software. The different poses then were clustered according to their intermolecular interaction pattern which represents the different accessible microstates. The ligand poses in each microstate reflect the depths and widths of the underlying energy wells. The translational and rotational entropy is then calculated from the restriction of the rotation and translation volume between the bound and unbound state of the ligand.

For validation of the BEERT software I used three datasets of well-established drug targets, namely HIV-1 protease, FXa, and Hsp90. The datasets were composed of highly similar inhibitors with a small range of molecular weights and inhibitory activity which represents a real-life scenario in lead optimization. Binding affinity predictions were performed using multiple linear regression of the translational and rotational entropy calculated by BEERT and effective binding energy calculated by MM-PBSA. The predicted binding affinities were compared with those from experiments. For all three datasets, using only MM-PBSA resulted in weak correlations, while incorporating the entropy calculations from BEERT improved the binding energy prediction dramatically ($R^2 = 0.56 – 0.72$). As an alternative, the number of rotatable bonds is often used for an estimation of the entropy change. However, using the number of rotatable bonds in a multiple linear regression together with MM-PBSA does not improve the results, which indicates the importance of incorporating the BEERT. BEERT was further tested for robustness using a leave-one-out cross-validation resulting in $q^2 = 0.34 – 0.66$, which was statistically significant for all three datasets. As a negative control y-randomization resulted in no correlation.

In the second project, I compared changes in the vibrational entropy upon binding to biomolecules based on rigidity theory, to $\Delta S_{vib.}$ computed by NMA as a gold standard. This was also validated on the same three datasets as used in the first project. In addition it had been previously validated in our group on a dataset of trypsin inhibitors, a protein-protein dataset as well as in alanine scanning, yielding significant and good to fair correlations for datasets of protein-protein and protein-small molecule complexes. As our approach is computationally highly efficient, it is a valuable alternative to NMA-based vibrational entropy computations in end-point (free) energy methods.

The significance of the presented approach BEERT is that it allows calculating the often neglected translational and rotational entropy contribution to binding free energy. In addition, we provided a highly efficient method for calculating vibrational entropy. Taken together, both methods allow calculating the configurational entropy change upon ligand binding in a highly efficient manner, thus complementing existing scoring functions or free energy calculation. To the best of our knowledge, this is the first time that a prediction of configurational entropy was successfully implemented in a way that allows large scale virtual screening, making our approach a valuable tool for identifying, understanding, and optimizing drug molecules and molecular interactions in general.

7 Supporting Information

7.1 Supplemental tables

Table S1: Physicochemical and structural properties of ligands from the HIV-1 protease dataset.

PDB ID	Resolution[a]	MW[b]	Rotatable bonds[c]	K_I
2I0D	1.95 [349]	637.754	14	0.8 pM [289-297]
2Q54	1.85 [353]	656.779	15	0.98 nM [289-297]
2Q55	1.90 [353]	655.794	14	2.04 nM [289-297]
2Q5K	1.90 [353]	628.812	16	5 pM [289-293]
2QHY	1.85 [350]	583.653	13	33 nM [289-293]
2QHZ	1.85 [350]	606.610	13	53 nM [289-293]
2QI0	2.10 [350]	582.717	13	42 nM [289-293]
2QI1	2.00 [350]	582.698	15	50 nM [289-293]
2QI3	1.95 [350]	547.740	15	63 pM [289-293]
2QI4	1.80 [350]	567.730	14	36 pM [289-293]
2QI5	1.85 [350]	588.793	16	14 pM [289-293]
2QI6	1.85 [350]	553.703	13	27 pM [289-293]
2QI7	1.85 [350]	532.701	17	62 pM [289-293]
3EKV	1.75 [12]	505.636	14	40 pM [294-297]
3EKX	1.97 [12]	567.793	12	10 pM [294-297]
3EKY	1.80 [12]	704.867	18	9 pM [294-297]
3GI4	1.85 [351]	677.699	13	16 pM [289-297]
3GI5	1.80 [351]	651.738	13	6 pM [289-297]
3GI6	1.84 [351]	663.715	14	6 pM [289-297]
3MXD	1.95 [352]	625.700	13	1.47 nM [289-293]

[a] Resolution of the crystallographic structure in Å.

[b] Molecular weight of the ligand in g mol^{-1}.

[c] Number of rotatable bonds of the ligand.

Table S2: Physicochemical and structural properties of ligands from the FXa dataset.

PDB ID	Resolution[a]	MW[b]	Rotatable bonds[c]	K_I
1EZQ	2.20 [354]	482.756	12	0.9 nM [289-297]
1F0R	2.10 [354]	475.720	5	22 nM [289-297]
1F0S	2.10 [354]	447.666	5	18 nM [289-297]
1KSN	2.10 [355]	470.701	10	0.4 nM [289-293]
1LPG	2.00 [396]	576.937	11	82 nM [289-297]
1LPK	2.20 [396]	451.701	11	28 nM [289-293]
1LPZ	2.40 [396]	489.577	7	25 nM [289-297]
1LQD	2.70 [396]	448.741	7	9 nM [289, 290, 293-297]
1NFU	2.05 [397]	483.139	6	18 nM [289-297]
1NFW	2.10 [397]	455.085	5	1.1 nM [289-297]
1NFX	2.15 [397]	525.177	7	3 nM [289-297]
1NFY	2.10 [397]	483.139	6	1.3 nM [289-297]
2BOH	2.20 [362]	501.180	6	3 nM [289-297]
2CJI	2.10 [398]	482.087	5	6 nM [289-297]
2J34	2.01 [356]	486.096	5	15 nM [289-297]
2J94	2.10 [357]	468.641	6	534 nM [289-297]
2J95	2.01 [357]	518.160	6	4 nM [289-297]
2UWL	1.90 [358]	460.058	6	4 nM [289-297]
2UWP	1.75 [358]	460.058	7	154 nM [289-297]
2VH0	1.70 [359]	544.199	8	3.1 nM [289, 290]

[a] Resolution of the crystallographic structure in Å.

[b] Molecular weight of the ligand in g mol^{-1}.

[c] Number of rotatable bonds of the ligand.

Table S3: Physicochemical and structural properties of ligands from the Hsp90 dataset.

PDB ID	Resolution[a]	MW[b]	Rotatable bonds[c]	IC_{50}
1UY7	1.90 [287]	311.387	6	200 µM [289-293]
1UY9	2.00 [287]	325.37	6	15.3 µM [289-293]
1UYC	2.00 [287]	341.413	8	41 µM [289-293]
1UYD	2.20 [287]	405.884	9	200 µM [289-293]
1UYG	2.00 [287]	303.296	5	53.5 µM [289-293]
1UYH	2.20 [287]	359.404	8	14.3 µM [289-293]
1UYK	2.20 [287]	343.361	6	17.1 µM [289-293]
2BYH	1.90 [286]	371.780	6	259 nM [289-293]
2BYI	1.60 [286]	422.849	7	461 nM [289-293]
2BZ5	1.90 [363]	460.337	5	700 nM [289-291]
2UWD	1.90 [364]	388.807	7	28 nM [294-297]
2VCI	2.00 [283]	465.549	9	21 nM [289-297]
2VCJ	2.50 [283]	457.913	8	21 nM [289-297]
2WI2	2.09 [282]	156.211	1	350 µM [289-297]
2WI4	2.40 [282]	344.202	3	1.56 µM [289-297]
2WI5	2.10 [282]	353.405	5	900 nM [289-297]
2WI6	2.18 [282]	367.258	3	230 nM [289-297]

[a] Resolution of the crystallographic structure in Å.

[b] Molecular weight of the ligand in g mol^{-1}.

[c] Number of rotatable bonds of the ligand.

Table S4: Average RMSD values and DrugScore scores of ligands of the HIV-1 protease dataset derived from re-docking experiments.

PDB ID	RMSD[a]	SEM[b]	DrugScore[c]	SEM[d]
2I0D	0.93	0.04	-25.14	0.03
2Q54	0.95	0.03	-26.94	0.06
2Q55	0.48	0.02	-26.90	0.01
2Q5K	0.99	0.05	-26.99	0.02
2QHY	0.61	0.02	-24.96	0.01
2QHZ	1.20	0.05	-24.53	0.01
2QI0	0.57	0.01	-25.69	0.01
2QI1	0.94	0.03	-25.80	0.01
2QI3	0.95	0.04	-23.05	0.01
2QI4	0.86	0.08	-25.06	0.01
2QI5	0.58	0.01	-24.52	0.01
2QI6	0.73	0.05	-24.95	0.01
2QI7	1.06	0.04	-23.31	0.08
3EKV	1.13	0.10	-23.77	0.01
3EKX	0.35	0.01	-24.73	0.01
3EKY	1.47	0.02	-26.30	0.08
3GI4	0.45	0.01	-25.20	0.14
3GI5	1.01	0.01	-25.53	0.03
3GI6	0.74	0.04	-24.62	0.13
3MXD	0.43	0.01	-25.14	0.06

[a] RMSD for the pose with the lowest predicted binding energy of the largest cluster averaged over 5 independent docking runs in Å.

[b] Standard error of the mean.

[c] DrugScore relative energy for the pose with the lowest predicted binding energy of the largest cluster averaged over 5 independent docking runs in kcal mol^{-1}, converted from DrugScore arbitrary units using the regression calculated from the experimental binding energy.

[d] Standard error of the mean.

Table S5: Average RMSD values and DrugScore scores of ligands of the FXa dataset derived from re-docking experiments.

PDB ID	RMSD[a]	SEM[b]	DrugScore[c]	SEM[d]
1EZQ	1.93	0.02	-23.73	0.01
1F0R	0.57	0.03	-22.01	0.00
1F0S	1.15	0.00	-21.28	0.00
1KSN	0.50	0.04	-22.92	0.01
1LPG	1.25	0.02	-23.72	0.08
1LPK	1.03	0.01	-24.02	0.01
1LPZ	1.13	0.04	-22.45	0.02
1LQD	1.07	0.04	-22.72	0.03
1NFU	1.55	0.11	-21.09	0.06
1NFW	0.27	0.01	-19.50	0.02
1NFX	0.55	0.01	-21.40	0.07
1NFY	0.87	0.05	-20.65	0.05
2BOH	0.40	0.00	-21.46	0.02
2CJI	0.36	0.01	-21.51	0.00
2J34	0.18	0.01	-20.16	0.00
2J94	0.47	0.01	-20.54	0.02
2J95	0.64	0.04	-20.11	0.03
2UWL	0.32	0.01	-19.26	0.05
2UWP	0.58	0.06	-20.10	0.00
2VH0	1.18	0.00	-21.72	0.01

[a] RMSD for the pose with the lowest predicted binding energy of the largest cluster averaged over 5 independent docking runs in Å.

[b] Standard error of the mean.

[c] DrugScore relative energy for the pose with the lowest predicted binding energy of the largest cluster averaged over 5 independent docking runs in kcal mol^{-1}, converted from DrugScore arbitrary units using the regression calculated from the experimental binding energy.

[d] Standard error of the mean.

Table S6: Average RMSD values and DrugScore scores of ligands of the Hsp90 dataset derived from re-docking experiments.

PDB ID	RMSD[a]	SEM[b]	DrugScore[c]	SEM[d]
1UY7	1.72	0.03	-19.35	0.01
1UY9	1.86	0.01	-19.35	0.01
1UYC	1.55	0.04	-19.40	0.00
1UYD	1.94	0.01	-20.11	0.00
1UYG	1.50	0.16	-18.53	0.01
1UYH	1.87	0.07	-19.47	0.01
1UYK	1.85	0.01	-19.89	0.00
2BYH	1.07	0.00	-19.47	0.04
2BYI	1.06	0.01	-19.89	0.00
2BZ5	1.46	0.00	-20.65	0.01
2UWD	1.20	0.04	-20.07	0.00
2VCI	1.54	0.01	-22.53	0.00
2VCJ	1.59	0.00	-21.07	0.00
2WI2	0.21	0.00	-14.14	0.00
2WI4	0.19	0.01	-17.36	0.00
2WI5	0.54	0.03	-19.10	0.00
2WI6	0.91	0.08	-17.85	0.00
1UYE	3.45	0.12	-20.38	0.01
1UYF	3.03	0.01	-20.65	0.04

[a] RMSD for the pose with the lowest predicted binding energy of the largest cluster averaged over 5 independent docking runs in Å.

[b] Standard error of the mean.

[c] DrugScore relative energy for the pose with the lowest predicted binding energy of the largest cluster averaged over 5 independent docking runs in kcal mol^{-1}, converted from DrugScore arbitrary units using the regression calculated from the experimental binding energy.

[d] Standard error of the mean.

Table S7: Average RMSD values of the protein-ligand complexes in the HIV-1 protease dataset derived from MD simulations.

PDB ID	Protein		Ligand fitted		Ligand not fitted	
	RMSD[a]	Std[b]	RMSD[c]	Std[d]	RMSD[e]	Std[f]
2I0D	1.68	0.30	2.59	1.16	3.67	1.63
2Q54	2.52	0.54	3.94	0.51	10.84	4.45
2Q55	1.25	0.17	1.41	0.31	2.15	0.40
2Q5K	1.50	0.22	1.37	0.25	1.67	0.32
2QHY	1.37	0.22	1.51	0.42	2.15	0.66
2QHZ	1.38	0.23	2.72	0.74	4.75	0.99
2QI0	1.65	0.30	1.43	0.27	1.97	0.31
2QI1	1.61	0.27	1.43	0.42	1.76	0.49
2QI3	1.61	0.29	2.17	0.27	2.48	0.34
2QI4	1.34	0.17	1.68	0.29	2.05	0.31
2QI5	1.49	0.26	1.62	0.23	2.27	0.48
2QI6	1.26	0.23	1.03	0.33	1.64	0.74
2QI7	1.26	0.17	1.71	0.25	1.94	0.27
3EKV	1.50	0.20	1.61	0.28	2.43	0.42
3EKX	1.31	0.17	1.98	0.29	3.88	0.94
3EKY	1.79	0.19	2.42	0.24	3.95	0.45
3GI4	1.61	0.35	2.06	0.76	2.93	1.01
3GI5	1.28	0.16	1.76	0.26	2.05	0.36
3GI6	1.54	0.34	1.36	0.30	2.03	0.43
3MXD	1.35	0.34	1.33	0.92	1.94	0.98

[a] Average RMSD in Å calculated over 12,500 snapshots taken from 250 ns of MD simulation. RMSD for each snapshot was calculated over all C_α atoms of the protein, relative to the initial structure, after superimposing the protein.

[b] Standard deviation calculated over the average from all snapshots, in Å.

[c] Average RMSD in Å calculated over 12,500 snapshots taken from 250 ns of MD simulation. RMSD for each snapshot was calculated over all ligand atoms, after superimposing the ligand.

[d] Standard deviation calculated over the average from all snapshots, in Å.

[e] Average RMSD in Å calculated over 12,500 snapshots taken from 250 ns of MD simulation. RMSD for each snapshot was calculated over all ligand atoms, after superimposing the protein.

[f] Standard deviation calculated over the average from all snapshots, in Å.

Table S8: Average RMSD values of the protein ligand complexes in the FXa dataset derived from MD simulations.

PDB ID	Protein		Ligand fitted		Ligand not fitted	
	RMSD[a]	Std[b]	RMSD[c]	Std[d]	RMSD[e]	Std[f]
1EZQ	1.74	0.25	1.07	0.40	1.57	0.61
1F0R	2.04	0.27	2.43	0.43	4.82	1.59
1F0S	1.67	0.27	2.47	0.70	5.88	1.70
1KSN	1.47	0.13	1.33	0.24	1.67	0.27
1LPG	1.86	0.23	1.53	0.31	3.46	0.50
1LPK	2.03	0.42	1.02	0.28	3.81	0.60
1LPZ	1.82	0.29	2.38	0.52	4.91	1.22
1LQD	1.69	0.21	2.02	0.50	3.59	1.05
1NFU	1.97	0.38	2.61	0.29	5.57	1.32
1NFW	1.68	0.25	2.00	0.19	5.65	1.00
1NFX	1.84	0.32	2.24	0.35	4.85	0.77
1NFY	2.35	0.40	2.17	0.52	7.86	1.43
2BOH	2.11	0.26	1.90	0.30	2.64	0.45
2CJI	2.10	0.43	2.69	0.64	6.62	1.75
2J34	2.23	0.26	1.71	0.41	2.79	0.95
2J94	1.94	0.30	1.28	0.60	3.27	1.31
2J95	1.95	0.19	1.62	0.31	4.39	0.82
2UWL	2.03	0.24	2.71	0.54	5.80	1.43
2UWP	1.95	0.24	2.73	0.41	7.94	1.56
2VH0	1.76	0.22	2.46	0.43	5.38	1.19

[a] Average RMSD in Å calculated over 12,500 snapshots taken from 250 ns of MD simulation. RMSD for each snapshot was calculated over all C_α atoms of the protein, relative to the initial structure, after superimposing the protein.

[b] Standard deviation calculated over the average from all snapshots, in Å.

[c] Average RMSD in Å calculated over 12,500 snapshots taken from 250 ns of MD simulation. RMSD for each snapshot was calculated over all ligand atoms, after superimposing the ligand.

[d] Standard deviation calculated over the average from all snapshots, in Å.

[e] Average RMSD in Å calculated over 12,500 snapshots taken from 250 ns of MD simulation. RMSD for each snapshot was calculated over all ligand atoms, after superimposing the protein.

[f] Standard deviation calculated over the average from all snapshots, in Å.

Table S9: Average RMSD values of the protein ligand complexes in the Hsp90 dataset derived from MD simulations.

PDB ID	Protein		Ligand fitted		Ligand not fitted	
	RMSD[a]	Std[b]	RMSD[c]	Std[d]	RMSD[e]	Std[f]
1UY7	2.14	0.33	1.70	0.25	2.89	0.56
1UY9	2.80	0.76	1.83	0.32	2.94	0.65
1UYC	3.60	0.94	1.83	0.26	3.36	0.62
1UYD	2.21	0.53	1.37	0.41	1.89	0.54
1UYG	2.05	0.47	1.39	0.67	2.64	1.61
1UYH	2.04	0.39	1.79	0.44	3.27	1.12
1UYK	2.34	0.39	1.31	0.31	1.77	0.35
2BYH	2.07	0.48	0.91	0.26	1.88	0.72
2BYI	2.75	0.76	1.39	0.20	2.11	0.37
2BZ5	2.02	0.26	1.58	0.46	2.77	0.71
2UWD	2.80	0.76	0.83	0.17	1.70	0.47
2VCI	2.03	0.46	2.09	0.49	2.79	0.70
2VCJ	2.35	0.31	1.75	0.57	2.34	0.67
2WI2	2.09	0.30	0.98	0.22	2.13	0.42
2WI4	2.03	0.34	1.33	0.41	2.74	0.50
2WI5	2.16	0.32	0.77	0.20	1.17	0.33
2WI6	2.50	0.41	1.04	0.17	1.44	0.24

[a] Average RMSD in Å calculated over 12,500 snapshots taken from 250 ns of MD simulation. RMSD for each snapshot was calculated over all C_α atoms of the protein, relative to the initial structure, after superimposing the protein.

[b] Standard deviation calculated over the average from all snapshots, in Å.

[c] Average RMSD in Å calculated over 12,500 snapshots taken from 250 ns of MD simulation. RMSD for each snapshot was calculated over all ligand atoms, after superimposing the ligand.

[d] Standard deviation calculated over the average from all snapshots, in Å.

[e] Average RMSD in Å calculated over 12,500 snapshots taken from 250 ns of MD simulation. RMSD for each snapshot was calculated over all ligand atoms, after superimposing the protein.

[f] Standard deviation calculated over the average from all snapshots, in Å.

Table S10: The drifts in the effective binding energies, determined from the slopes of the linear regression lines against the time for each complex from the HIV-1 protease dataset.

PDB ID	Drift[a]
2I0D	-0.02
2Q54	-0.01
2Q55	-0.02
2Q5K	0.01
2QHY	0.01
2QHZ	-0.04
2QI0	-0.01
2QI1	-0.05
2QI3	-0.01
2QI4	0.00
2QI5	-0.02
2QI6	0.00
2QI7	-0.02
3EKV	-0.02
3EKX	-0.07
3EKY	-0.02
3GI4	0.00
3GI5	0.01
3GI6	-0.04
3MXD	-0.01

[a] Effective binding energies were calculated over 250 ns, and the drift is given in kcal mol^{-1} ns^{-1}.

Table S11: The drifts in the effective binding energies, determined from the slopes of the linear regression lines against the time for each complex from the FXa dataset.

PDB ID	Drift[a]
1EZQ	0.04
1F0R	-0.03
1F0S	0.03
1KSN	-0.01
1LPG	0.04
1LPK	0.00
1LPZ	0.01
1LQD	0.01
1NFU	-0.03
1NFW	0.00
1NFX	-0.03
1NFY	0.03
2BOH	0.00
2CJI	0.01
2J34	-0.02
2J94	0.05
2J95	0.00
2UWL	-0.02
2UWP	0.01
2VH0	-0.04

[a] Effective binding energies were calculated over 250 ns, and the drift is given in kcal mol^{-1} ns^{-1}.

Table S12: The drifts in the effective binding energies, determined from the slopes of the linear regression lines against the time for each complex from the Hsp90 dataset.

PDB ID	Drift[a]
1UY7	0.00
1UY9	0.02
1UYC	-0.03
1UYD	-0.02
1UYG	0.05
1UYH	0.00
1UYK	0.01
2BYH	0.01
2BYI	0.00
2BZ5	0.01
2UWD	-0.01
2VCI	-0.01
2VCJ	-0.02
2WI2	0.00
2WI4	-0.01
2WI5	0.03
2WI6	0.00

[a] Effective binding energies were calculated over 250 ns, and the drift is given in kcal mol^{-1} ns^{-1}.

Table S13: MM-PBSA effective energies of the HIV-1 protease dataset using the one-trajectory approach.

PDB ID	MM-PBSA[a]	SEM[b]
2I0D	-39.60	$4.80 \cdot 10^{-04}$
2Q54	-21.90	$5.84 \cdot 10^{-04}$
2Q55	-36.10	$5.28 \cdot 10^{-04}$
2Q5K	-40.20	$4.32 \cdot 10^{-04}$
2QHY	-34.20	$4.00 \cdot 10^{-04}$
2QHZ	-33.80	$4.80 \cdot 10^{-04}$
2QI0	-46.00	$3.84 \cdot 10^{-04}$
2QI1	-37.20	$6.24 \cdot 10^{-04}$
2QI3	-39.00	$4.16 \cdot 10^{-04}$
2QI4	-32.70	$4.48 \cdot 10^{-04}$
2QI5	-35.70	$6.40 \cdot 10^{-04}$
2QI6	-32.80	$4.24 \cdot 10^{-04}$
2QI7	-37.70	$3.76 \cdot 10^{-04}$
3EKV	-41.80	$3.92 \cdot 10^{-04}$
3EKX	-42.20	$6.16 \cdot 10^{-04}$
3EKY	-34.30	$5.04 \cdot 10^{-04}$
3GI4	-41.20	$4.48 \cdot 10^{-04}$
3GI5	-36.40	$4.32 \cdot 10^{-04}$
3GI6	-35.20	$9.68 \cdot 10^{-04}$
3MXD	-30.70	$9.52 \cdot 10^{-04}$

[a] Average MM-PBSA effective binding energy in kcal mol^{-1} calculated over 12,500 snapshots from 250 ns of MD simulations.

[b] Standard error of the mean, in kcal mol^{-1}.

Table S14: MM-PBSA effective energies of the FXa dataset using the one-trajectory approach.

PDB ID	MM-PBSA[a]	SEM[b]
1EZQ	-49.5	$4.64 \cdot 10^{-04}$
1F0R	-25.2	$5.04 \cdot 10^{-04}$
1F0S	-28.7	$4.56 \cdot 10^{-04}$
1KSN	-50.9	$3.84 \cdot 10^{-04}$
1LPG	-48.6	$5.52 \cdot 10^{-04}$
1LPK	-48.8	$6.32 \cdot 10^{-04}$
1LPZ	-41.2	$5.36 \cdot 10^{-04}$
1LQD	-37.1	$5.20 \cdot 10^{-04}$
1NFU	-27.8	$5.36 \cdot 10^{-04}$
1NFW	-25.6	$4.32 \cdot 10^{-04}$
1NFX	-35.1	$5.04 \cdot 10^{-04}$
1NFY	-27.0	$5.20 \cdot 10^{-04}$
2BOH	-28.4	$3.04 \cdot 10^{-04}$
2CJI	-24.4	$3.52 \cdot 10^{-04}$
2J34	-28.1	$4.56 \cdot 10^{-04}$
2J94	-22.3	$4.96 \cdot 10^{-04}$
2J95	-30.9	$3.12 \cdot 10^{-04}$
2UWL	-23.7	$4.08 \cdot 10^{-04}$
2UWP	-22.0	$3.12 \cdot 10^{-04}$
2VH0	-31.3	$5.68 \cdot 10^{-04}$

[a] Average MM-PBSA effective binding energy in kcal mol^{-1} calculated over 12,500 snapshots from 250 ns of MD simulations.

[b] Standard error of the mean, in kcal mol^{-1}.

Table S15: MM-PBSA effective energies of the Hsp90 dataset using the one-trajectory approach.

PDB ID	MM-PBSA[a]	SEM[b]
1UY7	-28.1	$2.88 \cdot 10^{-04}$
1UY9	-28.3	$3.44 \cdot 10^{-04}$
1UYC	-34.3	$3.68 \cdot 10^{-04}$
1UYD	-34.8	$3.52 \cdot 10^{-04}$
1UYG	-21.6	$4.56 \cdot 10^{-04}$
1UYH	-28.0	$3.28 \cdot 10^{-04}$
1UYK	-26.9	$3.04 \cdot 10^{-04}$
2BYH	-19.2	$4.08 \cdot 10^{-04}$
2BYI	-26.7	$5.20 \cdot 10^{-04}$
2BZ5	-33.1	$3.68 \cdot 10^{-04}$
2UWD	-29.4	$3.12 \cdot 10^{-04}$
2VCI	-33.5	$3.76 \cdot 10^{-04}$
2VCJ	-33.3	$3.76 \cdot 10^{-04}$
2WI2	-15.0	$3.20 \cdot 10^{-04}$
2WI4	-22.9	$3.12 \cdot 10^{-04}$
2WI5	-28.1	$3.92 \cdot 10^{-04}$
2WI6	-32.9	$3.12 \cdot 10^{-04}$

[a] Average MM-PBSA effective binding energy in kcal mol^{-1} calculated over 12,500 snapshots from 250 ns of MD simulation.

[b] Standard error of the mean, in kcal mol^{-1}.

Table S16: -$T\Delta S_{\text{config.}}$ as calculated by BEERT for the HIV-1 protease dataset.

PDB ID	BEERT[a]	SEM[b]
2I0D	9.07	0.08
2Q54	10.52	0.10
2Q55	10.30	0.02
2Q5K	10.85	0.05
2QHY	12.51	0.19
2QHZ	12.86	0.27
2QI0	15.16	0.25
2QI1	13.90	0.29
2QI3	10.61	0.09
2QI4	11.30	0.17
2QI5	9.53	0.11
2QI6	11.87	0.21
2QI7	9.51	0.14
3EKV	14.73	0.33
3EKX	10.26	0.06
3EKY	9.71	0.07
3GI4	10.17	0.09
3GI5	9.73	0.05
3GI6	9.55	0.06
3MXD	11.38	0.03

[a] Average entropies at T = 300 K in kcal mol^{-1} computed according to eq. 28 over 5 independent docking runs, each with 100 poses.

[b] Standard error of the mean, in kcal mol^{-1}.

Table S17: $-T\Delta S_{config.}$ as calculated by BEERT for the FXa dataset.

PDB ID	BEERT[a]	SEM[b]
1EZQ	10.78	0.05
1F0R	10.40	0.04
1F0S	11.10	0.03
1KSN	11.14	0.07
1LPG	8.86	0.08
1LPK	8.72	0.14
1LPZ	7.24	0.07
1LQD	9.05	0.04
1NFU	8.68	0.10
1NFW	7.94	0.05
1NFX	8.85	0.07
1NFY	8.26	0.07
2BOH	10.29	0.25
2CJI	10.23	0.03
2J34	10.19	0.25
2J94	10.66	0.14
2J95	7.81	0.14
2UWL	8.54	0.08
2UWP	9.50	0.22
2VH0	10.17	0.10

[a] Average entropies at T = 300 K in kcal mol^{-1} computed according to eq. 28 over 5 independent docking runs, each with 100 poses.

[b] Standard error of the mean, in kcal mol^{-1}.

Table S18: $-T\Delta S_{config.}$ as calculated by BEERT for the Hsp90 dataset.

PDB ID	BEERT[a]	SEM[b]
1UY7	9.28	0.36
1UY9	9.88	0.06
1UYC	10.09	0.07
1UYD	9.93	0.07
1UYG	9.91	0.05
1UYH	9.86	0.06
1UYK	9.77	0.08
2BYH	10.58	0.09
2BYI	10.39	0.06
2BZ5	9.74	0.08
2UWD	12.71	0.11
2VCI	15.79	0.08
2VCJ	15.71	0.08
2WI2	20.04	0.12
2WI4	11.34	0.10
2WI5	12.64	0.06
2WI6	11.05	0.03

[a] Average entropies at T = 300 K in kcal mol^{-1} computed according to eq. 28 over 5 independent docking runs, each with 100 poses.

[b] Standard error of the mean, in kcal mol^{-1}.

Table S19: Surflex as an external scoring function on the re-docking poses retrieved by AutoDock and DrugScore for the HIV-1 protease dataset.

PDB ID	pK_i[a]	Surflex[b]
2I0D	12.10	8.51
2Q54	9.01	10.25
2Q55	8.69	11.01
2Q5K	11.89	5.13
2QHY	7.48	9.62
2QHZ	7.28	7.37
2QI0	7.38	7.73
2QI1	7.30	10.53
2QI3	10.20	9.40
2QI4	10.44	11.17
2QI5	10.85	8.82
2QI6	10.57	10.29
2QI7	10.21	8.82
3EKV	10.40	8.73
3EKX	11.00	10.16
3EKY	11.05	5.82
3GI4	10.80	9.96
3GI5	11.22	10.41
3GI6	11.22	11.00
3MXD	8.83	10.83

[a] Experimental pK_i.

[b] Surflex energy for the pose with the lowest predicted binding energy of the largest cluster in kcal mol^{-1}.

Table S20: Surflex as an external scoring function on the re-docking poses retrieved by AutoDock and DrugScore for the FXa dataset.

PDB ID	pK_i[a]	Surflex[b]
1EZQ	9.05	3.73
1F0R	7.66	3.52
1F0S	7.74	5.05
1KSN	9.40	4.51
1LPG	7.09	2.68
1LPK	7.55	7.62
1LPZ	7.60	3.67
1LQD	8.05	4.27
1NFU	7.74	6.77
1NFW	8.96	6.46
1NFX	8.52	3.74
1NFY	8.89	7.37
2BOH	8.52	5.97
2CJI	8.22	6.58
2J34	7.82	5.33
2J94	6.27	6.88
2J95	8.40	4.57
2UWL	8.40	4.74
2UWP	6.81	3.02
2VH0	8.51	3.24

[a] Experimental pK_i.

[b] Surflex energy for the pose with the lowest predicted binding energy of the largest cluster in kcal mol^{-1}.

Table S21: Surflex as an external scoring function on the re-docking poses retrieved by AutoDock and DrugScore for the Hsp90 dataset.

PDB ID	pIC_{50}[a]	Surflex[b]
1UY7	3.70	5.72
1UY9	4.82	5.83
1UYC	4.39	5.19
1UYD	3.70	3.24
1UYG	4.27	5.39
1UYH	4.84	5.61
1UYK	4.77	5.12
2BYH	6.59	4.88
2BYI	6.34	6.52
2BZ5	6.15	6.82
2UWD	7.55	5.99
2VCI	7.68	7.51
2VCJ	7.68	6.58
2WI2	3.46	2.68
2WI4	5.81	6.26
2WI5	6.05	4.80
2WI6	6.64	5.43

[a] Experimental pIC_{50}.

[b] Surflex energy for the pose with the lowest predicted binding energy of the largest cluster in kcal mol^{-1}.

Table S22: ΔS_{vib} computed by NMA and Δ-$F^{(1)}$ computed by constraint counting for the HIV-1 protease dataset.

PDB ID	ΔS_{vib} [a]	SEM_{Total}[b]	Δ-$F^{(1)}$ [c]	SEM_{Total}[d]
2I0D	0.23	0.29	45.01	0.08
2Q54	-18.76	0.31	118.97	0.08
2Q55	-5.10	0.31	50.90	0.08
2Q5K	-6.36	0.32	-58.40	0.08
2QHY	4.19	0.29	71.39	0.07
2QHZ	6.99	0.31	133.92	0.08
2QI0	7.82	0.32	63.01	0.09
2QI1	-4.27	0.33	50.88	0.08
2QI3	-9.08	0.22	85.00	0.09
2QI4	2.87	0.31	129.85	0.10
2QI5	0.00	0.29	25.85	0.10
2QI6	3.82	0.37	78.74	0.09
2QI7	-8.28	0.32	82.53	0.09
3EKV	-3.85	0.17	121.03	0.09
3EKX	-6.48	0.28	83.18	0.09
3EKY	4.24	0.34	93.58	0.10
3GI4	-6.28	0.31	-15.64	0.09
3GI5	-8.62	0.30	51.83	0.09
3GI6	-9.74	0.43	174.23	0.09
3MXD	-5.72	0.35	175.37	0.09

[a] ΔS_{vib} (eq. 36 in the main text) was averaged over 500 snapshots taken from the last 10 ns of a 20 ns MD simulation, in cal mol^{-1} K^{-1}.

[b] Standard error of the mean of ΔS_{vib} (eq. 39 in the main text), in cal mol^{-1} K^{-1}.

[c] Δ-$F^{(1)}$ (eq. 33 in the main text) was averaged over 500 snapshots taken from the last 10 ns of a 20 ns MD simulation, in cal mol^{-1} K^{-1}.

[d] Standard error of the mean of Δ-$F^{(1)}$ (eq. 39 in the main text).

Table S23: ΔS_{vib} computed by NMA and Δ-$F^{(1)}$ computed by constraint counting for the FXA dataset.

PDB ID	ΔS_{vib} [a]	SEM_{Total}[b]	Δ-$F^{(1)}$ [c]	SEM_{Total}[d]
1EZQ	-7.67	0.34	-97.59	0.08
1F0R	29.53	0.46	201.78	0.09
1F0S	19.44	0.31	22.95	0.09
1KSN	-11.52	0.38	-122.82	0.09
1LPG	-19.95	0.33	-153.90	0.09
1LPK	21.25	0.66	29.37	0.10
1LPZ	-8.18	0.30	15.30	0.09
1LQD	11.24	0.36	69.20	0.08
1NFU	15.18	0.33	130.84	0.09
1NFW	5.90	0.43	58.14	0.10
1NFX	4.19	0.34	53.18	0.08
1NFY	22.37	0.38	115.76	0.08
2BOH	60.39	0.41	217.85	0.08
2CJI	24.94	0.39	171.33	0.10
2J34	28.23	0.36	176.34	0.08
2J94	-7.10	0.35	167.34	0.09
2J95	14.16	0.34	177.25	0.09
2UWL	10.48	0.41	182.96	0.09
2UWP	3.55	0.38	200.01	0.10
2VH0	11.21	0.42	31.65	0.10

[a] ΔS_{vib} (eq. 36 in the main text) was averaged over 500 snapshots taken from the last 10 ns of a 20 ns MD simulation, in cal mol^{-1} K^{-1}.

[b] Standard error of the mean of ΔS_{vib} (eq. 39 in the main text), in cal mol^{-1} K^{-1}.

[c] Δ-$F^{(1)}$ (eq. 33 in the main text) was averaged over 500 snapshots taken from the last 10 ns of a 20 ns MD simulation, in cal mol^{-1} K^{-1}.

[d] Standard error of the mean of Δ-$F^{(1)}$ (eq. 39 in the main text).

Table S24: ΔS_{vib} computed by NMA and Δ-$F^{(1)}$ computed by constraint counting for the Hsp90 dataset.

PDB ID	ΔS_{vib} [a]	SEM_{Total}[b]	Δ-$F^{(1)}$ [c]	SEM_{Total}[d]
1UY7	13.82	0.30	-96.75	0.06
1UY9	13.99	0.28	-108.81	0.07
1UYC	16.26	0.38	-75.89	0.06
1UYD	10.65	0.29	-94.67	0.06
1UYG	15.48	0.31	-42.29	0.05
1UYH	19.85	0.32	-61.22	0.06
1UYK	16.81	0.28	-90.08	0.06
2BYH	9.43	0.32	-68.58	0.05
2BYI	1.59	0.32	-148.32	0.06
2BZ5	5.58	0.27	-82.93	0.06
2UWD	6.49	0.31	-97.28	0.05
2VCI	5.33	0.37	-53.75	0.05
2VCJ	3.55	0.34	-97.27	0.05
2WI2	12.32	0.32	37.02	0.06
2WI4	9.51	0.31	-29.88	0.05
2WI6	10.77	0.40	-153.93	0.06

[a] ΔS_{vib} (eq. 36 in the main text) was averaged over 500 snapshots taken from the last 10 ns of a 20 ns MD simulation, in cal mol^{-1} K^{-1}.

[b] Standard error of the mean of ΔS_{vib} (eq. 39 in the main text), in cal mol^{-1} K^{-1}.

[c] Δ-$F^{(1)}$ (eq. 33 in the main text) was averaged over 500 snapshots taken from the last 10 ns of a 20 ns MD simulation, in cal mol^{-1} K^{-1}.

[d] Standard error of the mean of Δ-$F^{(1)}$ (eq. 39 in the main text).

Table S25: ΔS_{vib} computed by NMA and Δ-$F^{(1)}$ computed by constraint counting for the Trypsin dataset.

PDB ID	ΔS_{vib} [a]	SEM_{Total}[b]	Δ-$F^{(1)}$ [c]	SEM_{Total}[d]
1C5S	-0.25	0.63	-20.96	7.52
1F0T	-12.68	0.64	-162.47	6.24
1G36	-17.77	0.70	33.52	6.73
1K1N	0.67	0.54	-260.65	6.30
1K1O	-47.41	0.83	-270.03	7.09
1K1P	-9.24	0.78	-382.79	6.87
1MTW	21.96	0.80	-12.37	5.69
1O2K	-17.15	0.73	-26.97	7.43
1O36	-12.15	0.74	-93.91	7.51
1QB6	-16.99	0.63	-198.87	7.22
1QBO	-25.18	0.79	-174.11	6.08
1QCP	-3.56	0.56	-170.75	5.99
1RXP	-19.08	0.62	-100.17	6.59
1S0R	18.42	0.78	123.13	7.43
1TX7	4.73	0.65	-3.99	6.58
1V2N	4.59	0.63	154.22	7.71
2AYW	-15.2	0.63	-145.27	7.28
2FX4	21.09	0.78	102.76	7.50
2OTV	22.22	0.75	153.23	8.07
2ZDK	-6.25	0.78	-235.15	8.10
2ZDL	-14.83	0.68	-227.76	7.77
2ZDN	-4.24	0.99	-249.22	6.62
2ZFS	-27.09	0.68	-239.42	7.22

[a] ΔS_{vib} (eq. 36 in the main text) was averaged over 30 snapshots taken from the last 6 ns of a 10 ns MD simulation, in cal mol^{-1} K^{-1}.

[b] Standard error of the mean of ΔS_{vib} (eq. 39 in the main text), in cal mol^{-1} K^{-1}.

[c] Δ-$F^{(1)}$ (eq. 33 in the main text) was averaged over 30 snapshots taken from the last 6 ns of a 10 ns MD simulation, in cal mol^{-1} K^{-1}.

[d] Standard error of the mean of Δ-$F^{(1)}$ (eq. 39 in the main text).

7.2 Supplemental figures

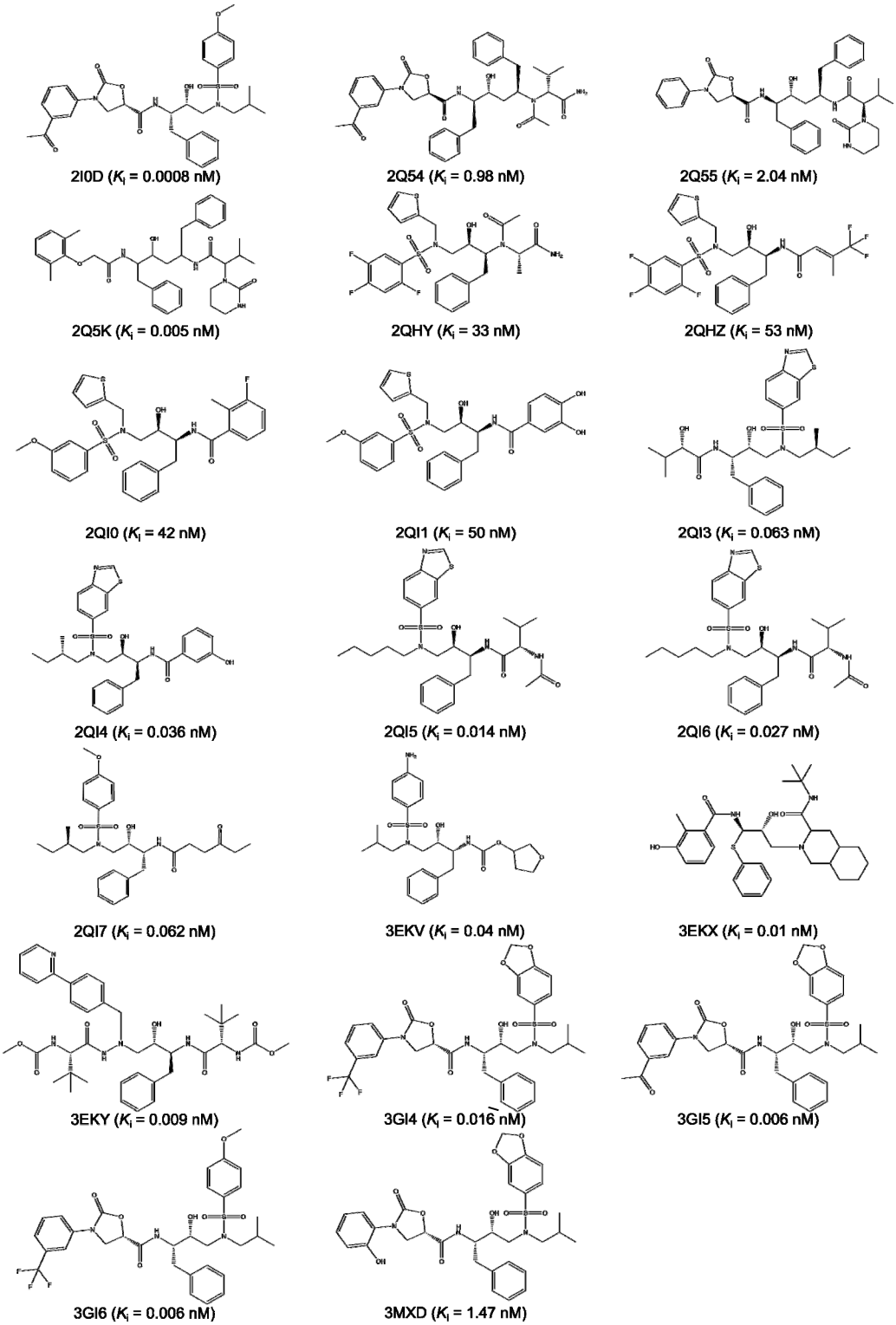

Figure S1: Structures of the ligands of the HIV1 protease dataset, the PDB ID of the complex, and the experimental binding affinity (K_i).

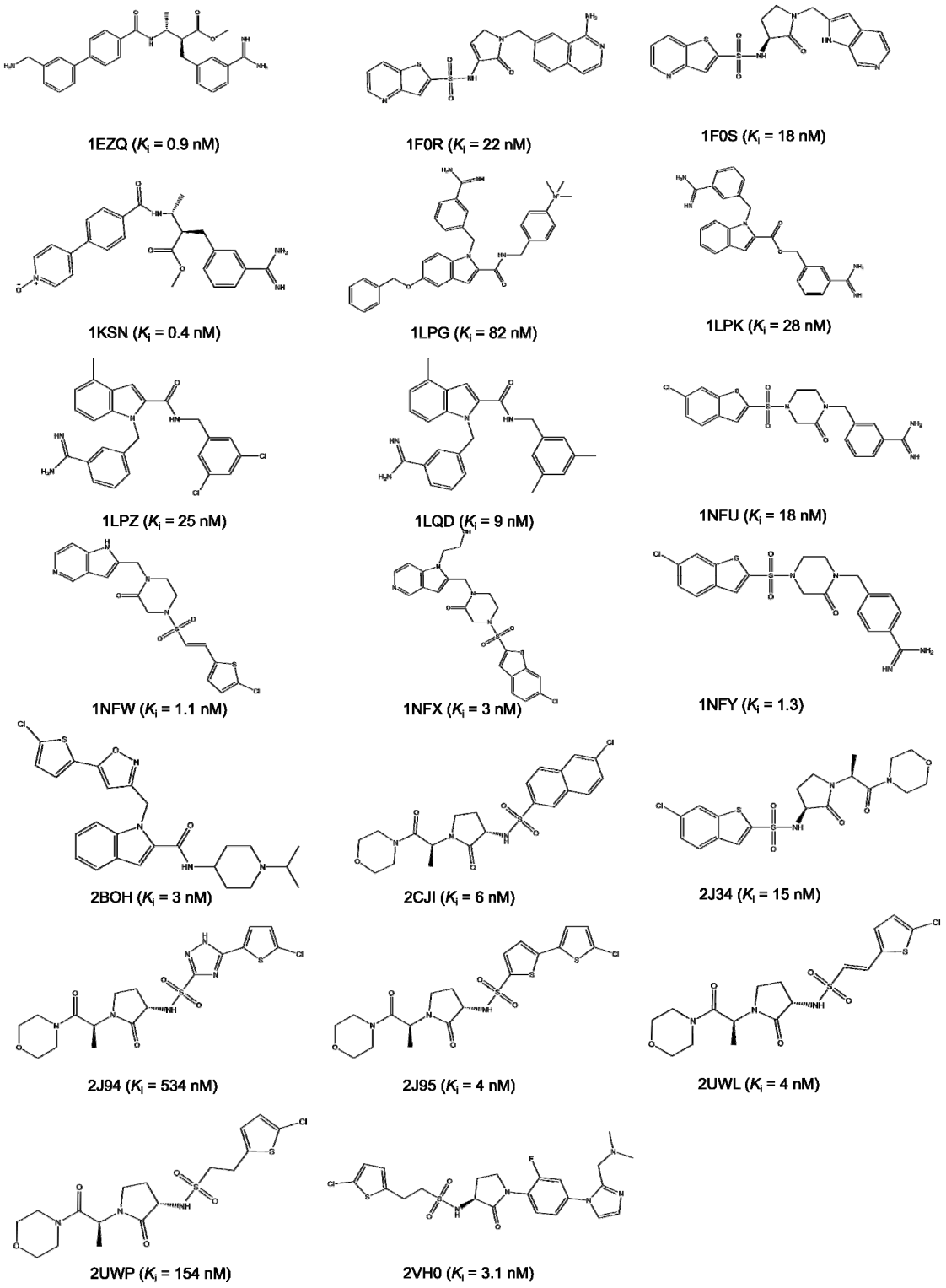

Figure S2: Structure of the ligands of the FXa dataset, the PDB ID of the complex, and the experimental binding affinity (K_i).

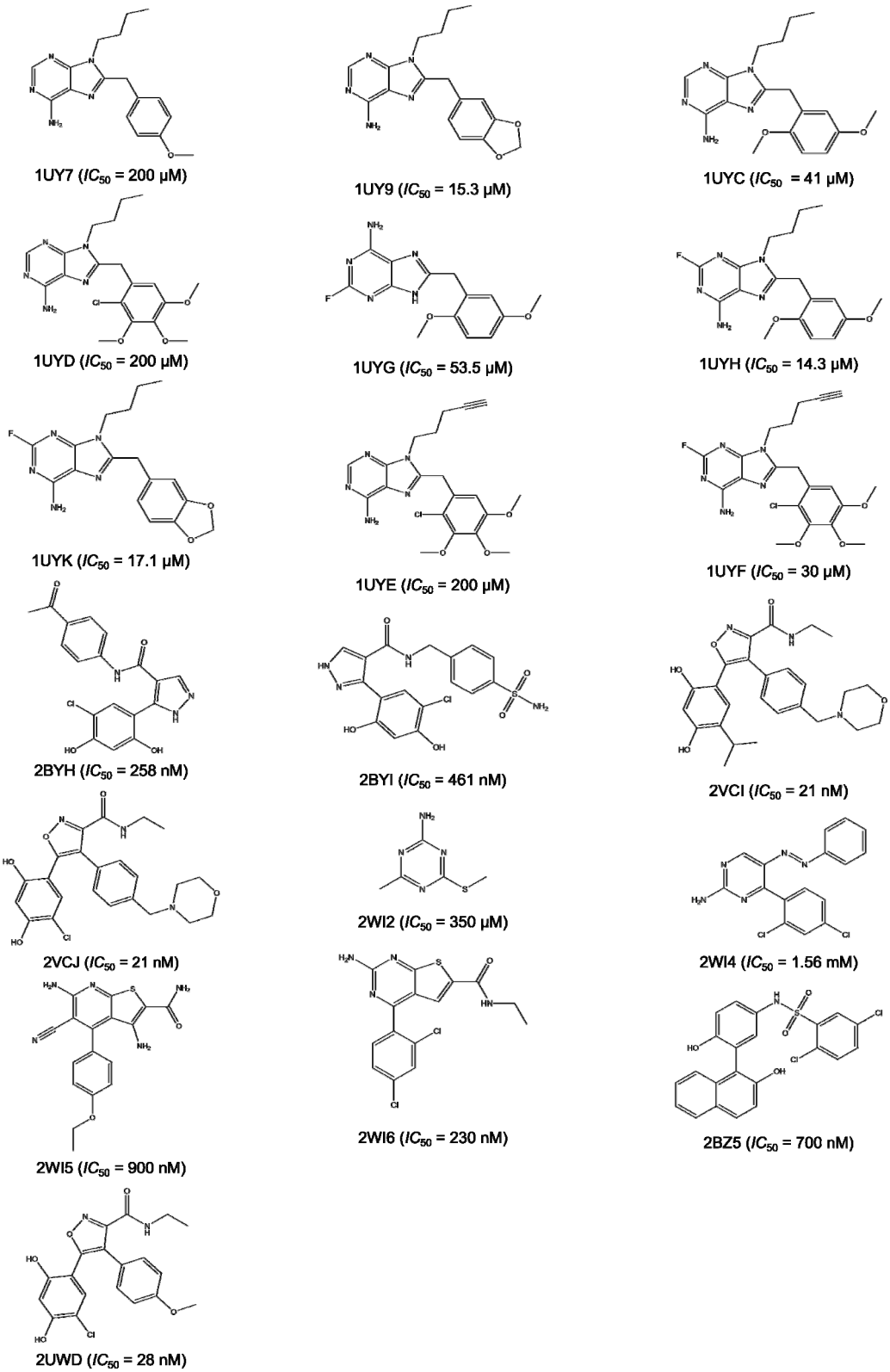

Figure S3: Structures of the ligands of the Hsp90 dataset, the PDB ID of the complex, and the half maximal inhibitory concentration (IC_{50}).

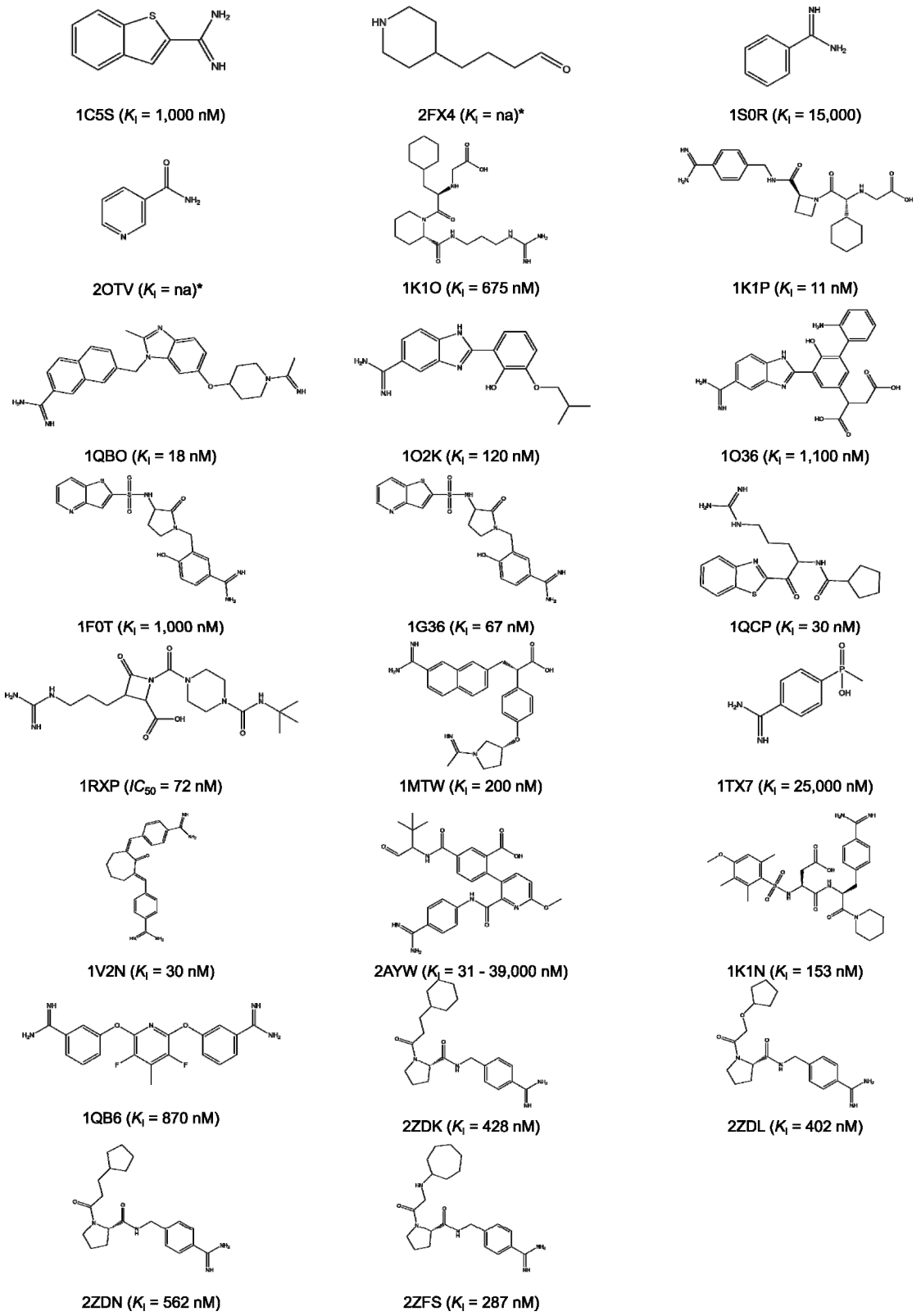

Figure S4: Structure of the ligands of the trypsin dataset, the PDB ID of the complex, and the experimental binding affinity (K_i) or the half maximal inhibitory concentration (IC_{50}). *No K_i/IC_{50} value available.

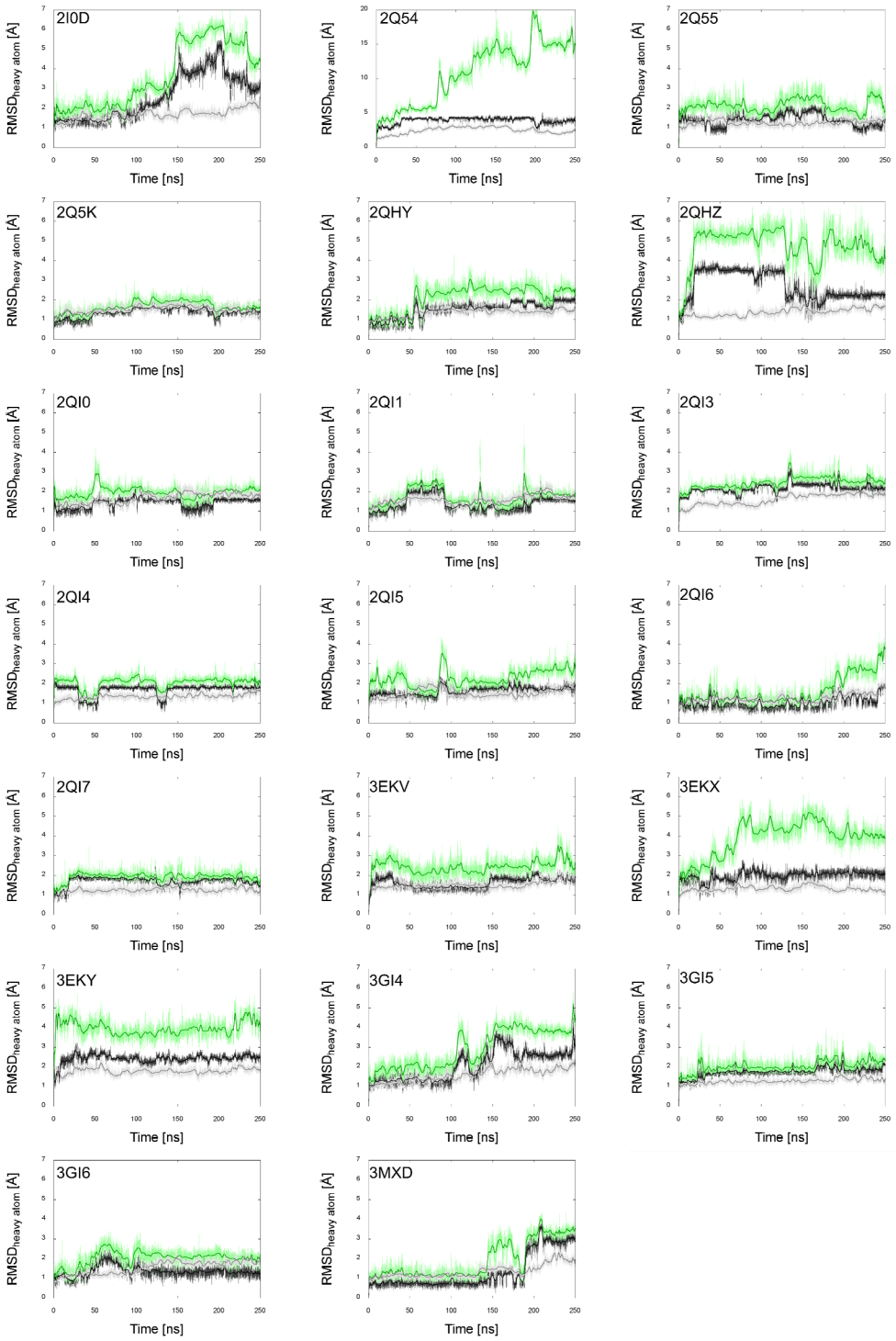

Figure S5: RMSD profile over 250 ns MD simulations for each of the protein-ligand complexes in the HIV-1 protease dataset. RMSD is calculated with respect to the starting structure. Results over the C_α atoms of the protein are shown in grey, those over the ligand heavy atoms in black, those over the ligand heavy atoms after superimpositioning only the protein in green.

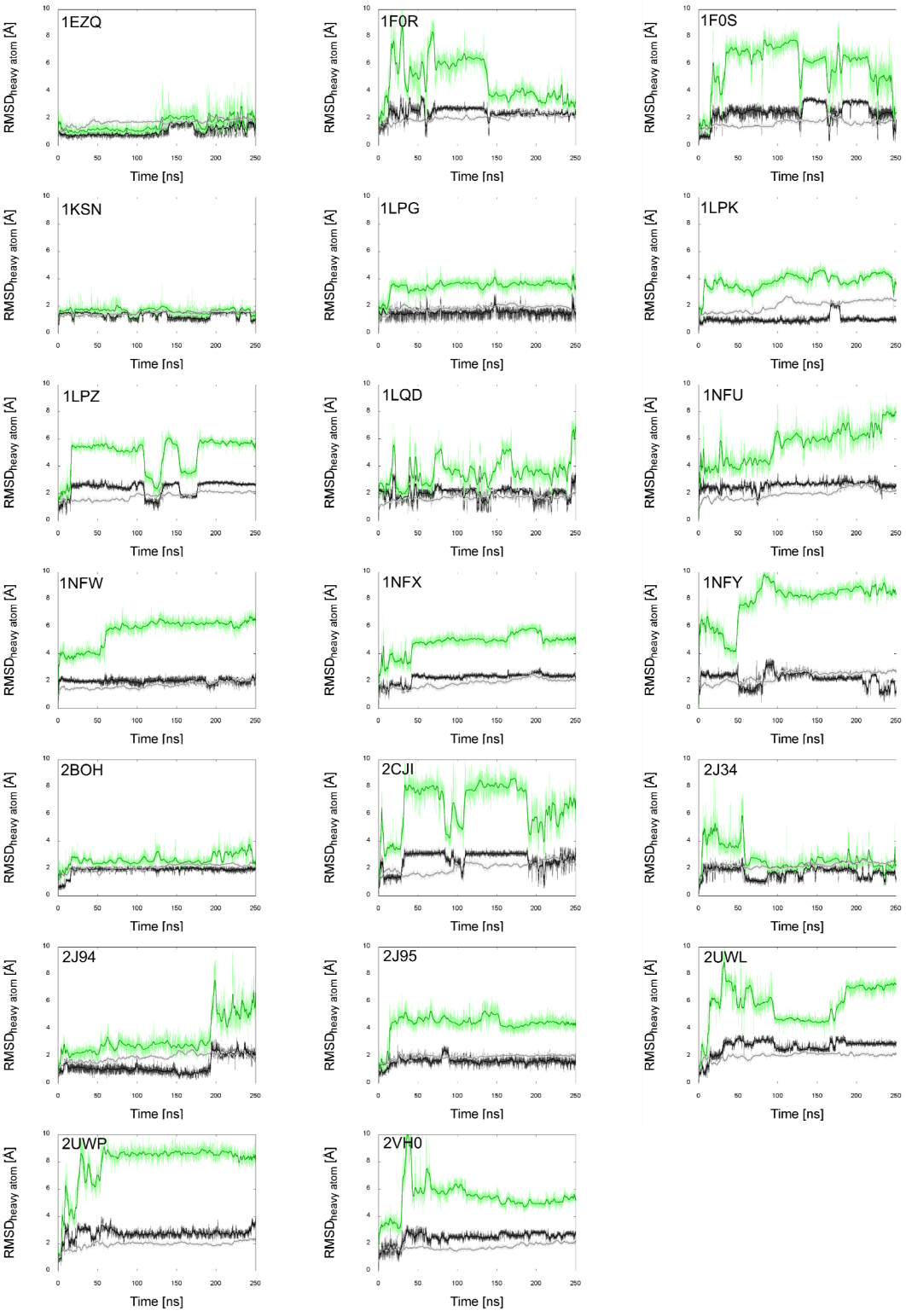

Figure S6: RMSD profile over 250 ns MD simulations for each of the protein-ligand complexes in the FXa dataset. RMSD is calculated with respect to the starting structure. Results over the C_α atoms of the protein are shown in grey, those over the ligand heavy atoms in black, those over the ligand heavy atoms after superimpositioning only the protein in green.

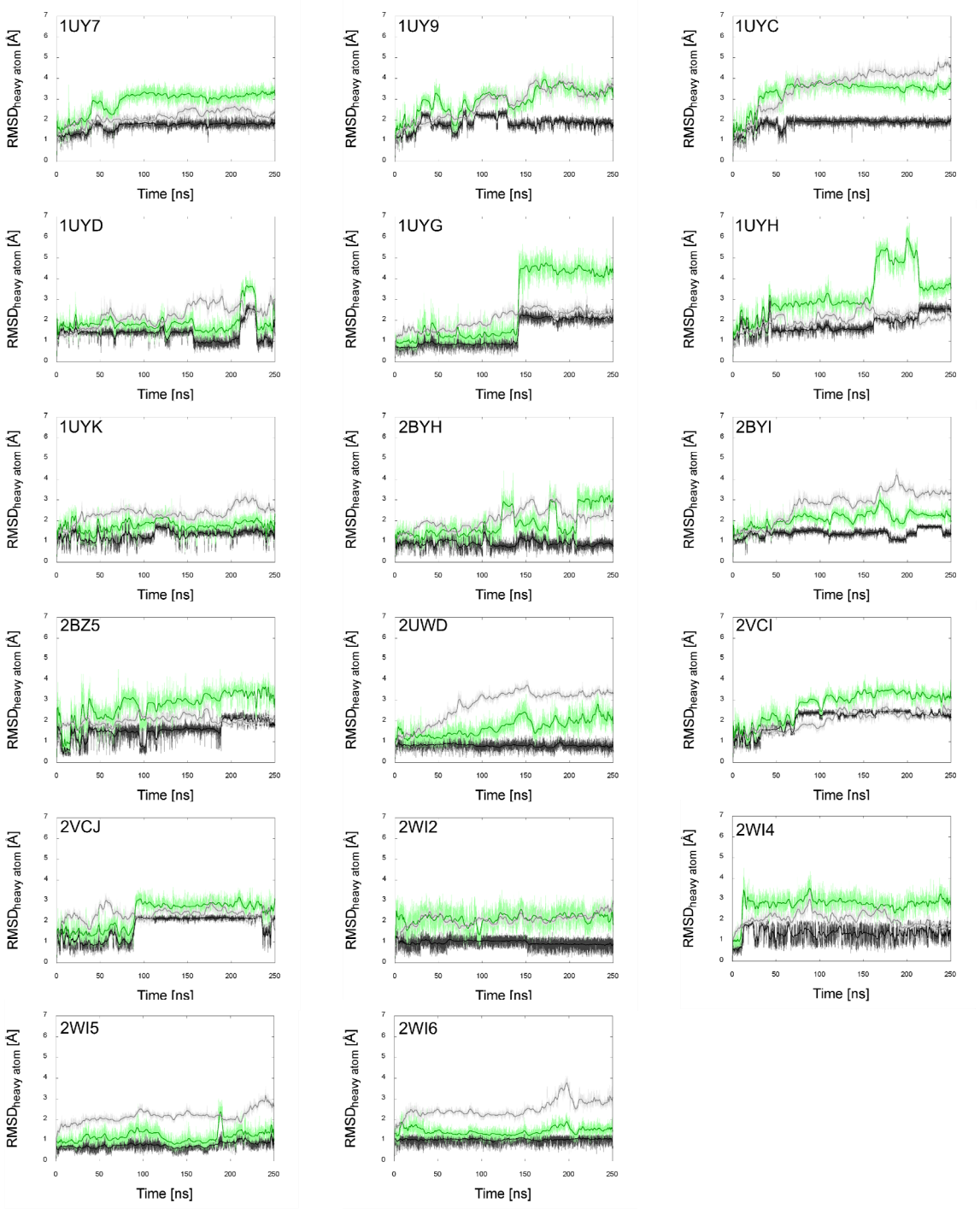

Figure S7: RMSD profile over 250 ns MD simulations for each of the protein-ligand complexes in the Hsp90 dataset. RMSD is calculated with respect to the starting structure. Results over the C_α atoms of the protein are shown in grey, those over the ligand heavy atoms in black, those over the ligand heavy atoms after superimpositioning only the protein in green.

8 Curriculum Vitae

Ido Ben-Shalom

Friedrichstr. 112 40217 Düsseldorf, Germany | ☏ +49 176 251 483 12
idoshalom@yahoo.com

ACADEMIC EDUCATION

05/10 - present	**Heinrich-Heine-University Düsseldorf, institute of Pharmaceutical and Medicinal Chemistry** (under the supervision of Prof. Holger Gohlke) PhD Student *Thesis: Protein-ligand binding entropy in lead optimization.* Developed innovative Binding Entropy Estimation of Rotation & Translation method (BEERT) that significantly improves affinity estimates for lead-optimization. Operational program in use at *Sanofi-Aventis*.
10/06 - 02/10	**The Hebrew University (Jerusalem), department of Medicinal Chemistry** (under the supervision of Prof. Amiram Goldblum) MSc student (GPA: 95.3/100) *Thesis: In-silico screening for novel HSP90 inhibitors.* Novel improved HSP90 inhibitors predicted by multi-method virtua screening.
10/03 - 09/06	**The Hebrew University (Jerusalem), Faculty of Sciences** BSc in Life Sciences (GPA: 91.2/100)

PROFESSIONAL AND ACADEMIC EXPERIENCE

01/08 - 10/08	**Synergix Ltd. (Jerusalem)**, Program Developer. Educational software *Molecular Conceptor* for Medicinal Chemistry, Drug Design, Molecular Modeling, and Cheminformatics
05/05 - 10/06	**The Hebrew University (Jerusalem), Department of Molecular Genetics** (under the supervision of Prof. Adam Friedmann and Prof. Moshe Soller) Research Assistant *Analyzing DNA-level polymorphisms by estimating the allele frequencies of quantitative trait loci of genes for in bovine milk production and quality.*

ADDITIONAL SKILLS

Languages	Hebrew (native speaker), English (fluent), German (fluent)
Programming	Writing operational programs in Python and Shell Application knowledge in C, C++, and Perl
Modeling	AutoDock, PyMol, Maestro, MOE, AMBER, Gnuplot, R, FlexX, Discovery Studio, Sybyl, LigandScout, and InsightII

TEACHING EXPERIENCE

10/10 - 09/15	**Heinrich-Heine-University Düsseldorf**, Organic Chemistry
02/09 - 04/10	**College of Engineering (Jerusalem)**, Molecular Biology & Biochemistry
02/09 - 07/09	**School of Marine Sciences, Ruppin Academic Center (Michmoret)**, General Chemistry
10/06 - 02/08	**The Hebrew University (Jerusalem)**, Organic Chemistry

AWARDS

02/13	Best poster, 27th Molecular Modeling workshop, Erlangen
02/10	MSc, graduation with excellence
10/09	Best lecture, the inauguration of the Institute of Drug Research, Jerusalem
09/08	Prize for excellence from the School of Pharmacy, Hebrew University

LECTURES & POSTER PRESENTATIONS

Ben-Shalom I., Pfeiffer-Marek S., and Gohlke H. *BEERT: A tool for the fast prediction of change in external ligand entropy*, Israel. **16th Israeli Bioinformatics Symposium**, Ramat-Gan**, 2014.** (Poster presentation)

Ben-Shalom I., Pfeiffer-Marek S., and Gohlke H. *Improving entropy prediction in drug binding*, Israel. **12th MCS-ICS**, Rehovot**, 2014.** (Poster presentation)

Ben-Shalom I. and Gohlke H. *Protein-ligand binding entropy in lead optimization*, Germany. **27th molecular modeling workshop**, Erlangen, 2013. (poster presentation)

Ben-Shalom I., *In Silico Screening for Novel HSP90*, Israel. **The inauguration of the institute of drug research**, Jerusalem, 2009.

Ben-Shalom I., Marcus D., Rayan A., and Goldblum A. *In Silico Screening for Novel HSP90 inhibitors*, Israel. **11th Israeli Bioinformatics Symposium**, Tel Aviv, 2008. (poster presentation)

SEMINARS/WORKSHOPS

06/14	*Schrödinger Workshop*, Heinrich-Heine-University, Düsseldorf
05/14	*Success Stories of Preclinical Research between Academia and Industry,* Philipp University of Marburg
09/13	Good Scientific Practice for Doctoral Researchers, Heinrich-Heine-University, Düsseldorf
10/13	*Optimizing Writing Strategies for Publishing Research in English*, Heinrich-Heine-University, Düsseldorf
11/13	*Fundamentals of Project Management*, Heinrich-Heine-University, Düsseldorf
11/13	*Get into Teaching*, Heinrich-Heine-University, Düsseldorf
03/12	*Introductory Seminar for Setup Membrane Simulations with Desmond*, Heinrich-Heine-University, Düsseldorf
04/10	*Conducting Molecular Dynamics Simulations with the AMBER11 Modeling Suite. Introductory and advanced seminars*, Heinrich-Heine-University, Düsseldorf

PUBLICATION

Ben-Shalom I. Y., Pfeiffer-Marek S., Baringhaus K. H, and Gohlke H. *Efficient approximation of ligand rotational and translational entropy changes upon binding for use in MM-PBSA calculations,* submitted.

Gohlke H., Ben-Shalom I.Y., Kopitz H., Pfeiffer-Marek S., and Baringhaus K.H. *Rigidity theory-based approximation of changes in vibrational entropy upon binding to biomolecules*.

9 References

1. Gilson, M. K.; Zhou, H. X., Calculation of protein-ligand binding affinities. *Annu Rev Bioph Biom* **2007**, 36, 21-42.

2. Chen, W.; Chang, C. E.; Gilson, M. K., Calculation of cyclodextrin binding affinities: Energy, entropy, and implications for drug design. *Biophys J* **2004**, 87, 3035-3049.

3. Bennaim, A.; Marcus, Y., Solvation Thermodynamics of Nonionic Solutes. *J Chem Phys* **1984**, 81, 2016-2027.

4. Marcus, Y., A Simple Empirical-Model Describing the Thermodynamics of Hydration of Ions of Widely Varying Charges, Sizes, and Shapes. *Biophys Chem* **1994**, 51, 111-127.

5. Goldstein, H.; Poole, C.; Safko, J.; Addison, S. R., Classical mechanics. *Am J Phys* **2002**, 70, 782-783.

6. Skjaerven, L.; Hollup, S. M.; Reuter, N., Normal mode analysis for proteins. *J Mol Struc-Theochem* **2009**, 898, 42-48.

7. Levitt, M.; Sander, C.; Stern, P. S., Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J Mol Biol* **1985**, 181, 423-47.

8. Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A., The experimental uncertainty of heterogeneous public K(i) data. *J Med Chem* **2012**, 55, 5165-73.

9. Gohlke, H.; Klebe, G., Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew Chem Int Edit* **2002**, 41, 2645-2676.

10. Houk, K. N.; Leach, A. G.; Kim, S. P.; Zhang, X. Y., Binding affinities of host-guest, protein-ligand, and protein-transition-state complexes. *Angew Chem Int Edit* **2003**, 42, 4872-4897.

11. Reynolds, C. H.; Holloway, M. K., Thermodynamics of Ligand Binding and Efficiency. *ACS Med Chem Lett* **2011**, 2, 433-437.

12. King, N. M.; Prabu-Jeyabalan, M.; Bandaranayake, R. M.; Nalam, M. N. L.; Nalivaika, E. A.; Ozen, A.; Haliloglu, T.; Yilmaz, N. K.; Schiffer, C. A., Extreme Entropy-Enthalpy Compensation in a Drug-Resistant Variant of HIV-1 Protease. *ACS Chem Biol* **2012**, 7, 1536-1546.

13. Dunitz, J. D., Win Some, Lose Some - Enthalpy-Entropy Compensation in Weak Intermolecular Interactions. *Chem Biol* **1995**, 2, 709-712.

14. Lyne, P. D., Structure-based virtual screening: an overview. *Drug Discov Today* **2002**, 7, 1047-1055.

15. Lill, M., Virtual screening in drug design. *Methods Mol Biol* **2013**, 993, 1-12.

16. Gane, P. J.; Dean, P. M., Recent advances in structure-based rational drug design. *Curr Opin Struct Biol* **2000**, 10, 401-4.

17. Velec, H. F. G.; Gohlke, H.; Klebe, G., DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J Med Chem* **2005**, 48, 6296-6303.

18. Muegge, I.; Martin, Y. C.; Hajduk, P. J.; Fesik, S. W., Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. *J Med Chem* **1999**, 42, 2498-503.

19. Jacobsson, M.; Karlen, A., Ligand bias of scoring functions in structure-based virtual screening. *J Chem Inf Model* **2006**, 46, 1334-1343.

20. Pan, Y. P.; Huang, N.; Cho, S.; MacKerell, A. D., Consideration of molecular weight during compound selection in virtual target-based database screening. *J Chem Inf Comp Sci* **2003**, 43, 267-272.

21. Searle, M. S.; Westwell, M. S.; Williams, D. H., Application of a Generalized Enthalpy-Entropy Relationship to Binding Cooperativity and Weak Associations in Solution. *J Chem Soc Perk T 2* **1995**, 141-151.

22. Williams, D. H.; Stephens, E.; O'Brien, D. P.; Zhou, M., Understanding noncovalent interactions: Ligand binding energy and catalytic efficiency from ligand-induced reductions in motion within receptors and enzymes. *Angew Chem Int Edit* **2004**, 43, 6596-6616.

23. Gohlke, H.; Case, D. A., Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J Comput Chem* **2004**, 25, 238-50.

24. van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glattli, A.; Hunenberger, P. H.; Kastenholz, M. A.; Ostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F. A.; Yu, H. B., Biomolecular modeling: Goals, problems, perspectives. *Angew Chem Int Edit* **2006**, 45, 4064-4092.

25. Meirovitch, H.; Cheluvaraja, S.; White, R. P., Methods for Calculating the Entropy and Free Energy and their Application to Problems Involving Protein Flexibility and Ligand Binding. *Curr Protein Pept Sc* **2009**, 10, 229-243.

26. Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S. H.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham,

T. E., Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Accounts Chem Res* **2000**, 33, 889-897.

27. Huang, N.; Kalyanaraman, C.; Irwin, J. J.; Jacobson, M. P., Physics-based scoring of protein-ligand complexes: enrichment of known inhibitors in large-scale virtual screening. *J Chem Inf Model* **2006**, 46, 243-53.

28. Liu, H. Y.; Kuntz, I. D.; Zou, X. Q., Pairwise GB/SA scoring function for structure-based drug design. *J Phys Chem B* **2004**, 108, 5453-5462.

29. Majeux, N.; Scarsi, M.; Apostolakis, J.; Ehrhardt, C.; Caflisch, A., Exhaustive docking of molecular fragments with electrostatic solvation. *Proteins Struct Funct Genet* **1999**, 37, 88-105.

30. Maple, J. R.; Cao, Y. X.; Damm, W. G.; Halgren, T. A.; Kaminski, G. A.; Zhang, L. Y.; Friesner, R. A., A polarizable force field and continuum solvation methodology for modeling of protein-ligand interactions. *J Chem Theory Comput* **2005**, 1, 694-715.

31. Shoichet, B. K.; Leach, A. R.; Kuntz, I. D., Ligand solvation in molecular docking. *Proteins Struct Funct Genet* **1999**, 34, 4-16.

32. Zhou, R. H.; Friesner, R. A.; Ghosh, A.; Rizzo, R. C.; Jorgensen, W. L.; Levy, R. M., New linear interaction method for binding affinity calculations using a continuum solvent model. *J Phys Chem B* **2001**, 105, 10388-10397.

33. Zou, X. Q.; Sun, Y. X.; Kuntz, I. D., Inclusion of solvation in ligand binding free energy calculations using the generalized-born model. *J Am Chem Soc* **1999**, 121, 8033-8043.

34. Page, M. I.; Jencks, W. P., Entropic Contributions to Rate Accelerations in Enzymic and Intramolecular Reactions and Chelate Effect. *P Natl Acad Sci USA* **1971**, 68, 1678-&.

35. Finkelstein, A. V.; Janin, J., The Price of Lost Freedom - Entropy of Bimolecular Complex-Formation. *Protein Eng* **1989**, 3, 1-3.

36. Tidor, B.; Karplus, M., The Contribution of Vibrational Entropy to Molecular Association - the Dimerization of Insulin. *J Mol Biol* **1994**, 238, 405-414.

37. Amzel, L. M., Loss of translational entropy in binding, folding, and catalysis. *Proteins Struct Funct Genet* **1997**, 28, 144-149.

38. Yu, Y. B.; Privalov, P. L.; Hodges, R. S., Contribution of translational and rotational motions to molecular association in aqueous solution. *Biophys J* **2001**, 81, 1632-1642.

39. Lazaridis, T.; Masunov, A.; Gandolfo, F., Contributions to the binding free energy of ligands to avidin and streptavidin. *Proteins Struct Funct Genet* **2002**, 47, 194-208.

40. Steinberg, I. Z.; Scheraga, H. A., Entropy Changes Accompanying Association Reactions of Proteins. *J Biol Chem* **1963**, 238, 172-&.

41. Carlsson, J.; Aqvist, J., Calculations of solute and solvent entropies from molecular dynamics simulations. *Phys Chem Chem Phys* **2006**, 8, 5385-5395.

42. Zhou, H. X.; Gilson, M. K., Theory of Free Energy and Entropy in Noncovalent Binding. *Chem Rev* **2009**, 109, 4092-4107.

43. Chang, C. E. A.; Chen, W.; Gilson, M. K., Ligand configurational entropy and protein binding. *P Natl Acad Sci USA* **2007**, 104, 1534-1539.

44. Moghaddam, S.; Yang, C.; Rekharsky, M.; Ko, Y. H.; Kim, K.; Inoue, Y.; Gilson, M. K., New Ultrahigh Affinity Host-Guest Complexes of Cucurbit[7]uril with Bicyclo[2.2.2]octane and Adamantane Guests: Thermodynamic Analysis and Evaluation of M2 Affinity Calculations. *J Am Chem Soc* **2011**, 133, 3570-3581.

45. Seeliger, D.; de Groot, B. L., Ligand docking and binding site analysis with PyMOL and Autodock/Vina. *J Comput Aided Mol Des* **2010**, 24, 417-22.

46. Carlsson, J.; Aqvist, J., Absolute and relative entropies from computer simulation with applications to ligand binding. *J Phys Chem B* **2005**, 109, 6448-6456.

47. Chang, C. E.; Gilson, M. K., Free energy, entropy, and induced fit in host-guest recognition: Calculations with the second-generation mining minima algorithm. *J Am Chem Soc* **2004**, 126, 13156-13164.

48. Muddana, H. S.; Gilson, M. K., Calculation of Host-Guest Binding Affinities Using a Quantum-Mechanical Energy Model. *J Chem Theory Comput* **2012**, 8, 2023-2033.

49. Ruvinsky, A. M.; Kozintsev, A. V., New and fast statistical-thermodynamic method for computation of protein-ligand binding entropy substantially improves docking accuracy. *J Comput Chem* **2005**, 26, 1089-1095.

50. Ruvinsky, A. M.; Kozintsev, A. V., Novel statistical-thermodynamic methods to predict protein-ligand binding positions using probability distribution functions. *Proteins* **2006**, 62, 202-208.

51. Ruvinsky, A. M., Calculations of protein-ligand binding entropy of relative and overall molecular motions. *J Comput Aid Mol Des* **2007**, 21, 361-370.

52. Ruvinsky, A. M., Role of binding entropy in the refinement of protein-ligand docking predictions: Analysis based on the use of 11 scoring functions. *J Comput Chem* **2007**, 28, 1364-1372.

53. Karplus, M.; Ichiye, T.; Pettitt, B. M., Configurational entropy of native proteins. *Biophys J* **1987**, 52, 1083-5.

54. Karplus, M.; Kushick, J. N., Method for Estimating the Configurational Entropy of Macromolecules. *Macromolecules* **1981**, 14, 325-332.

55. Frederick, K. K.; Marlow, M. S.; Valentine, K. G.; Wand, A. J., Conformational entropy in molecular recognition by proteins. *Nature* **2007**, 448, 325-U3.

56. Frauenfelder, H.; Parak, F.; Young, R. D., Conformational substates in proteins. *Annu Rev Biophys Biophys Chem* **1988**, 17, 451-79.

57. Elber, R.; Karplus, M., Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science* **1987**, 235, 318-21.

58. Kitao, A.; Hayward, S.; Go, N., Energy landscape of a native protein: jumping-among-minima model. *Proteins* **1998**, 33, 496-517.

59. Gō, N.; Scheraga, H. A., Analysis of the contribution of internal vibrations to the statistical weights of equilibrium conformations of macromolecules. *J Chem Phys* **1969**, 51, 4751-4767.

60. Bahar, I.; Lezon, T. R.; Bakan, A.; Shrivastava, I. H., Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. *Chem Rev* **2010**, 110, 1463-97.

61. Case, D. A., Normal-Mode Analysis of Protein Dynamics. *Curr Opin Struc Biol* **1994**, 4, 285-290.

62. Hughes, J. P.; Rees, S.; Kalindjian, S. B.; Philpott, K. L., Principles of early drug discovery. *Brit J Pharmacol* **2011**, 162, 1239-1249.

63. Algahtani, M. S.; Scurr, D. J.; Hook, A. L.; Anderson, D. G.; Langer, R. S.; Burley, J. C.; Alexander, M. R.; Davies, M. C., High throughput screening for biomaterials discovery. *J Control Release* **2014**, 190, 115-26.

64. Blount, K. F.; Breaker, R. R., Riboswitches as antibacterial drug targets. *Nat Biotechnol* **2006**, 24, 1558-1564.

65. Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M., DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* **2008**, 36, D901-6.

66. Detering, C.; Varani, G., Validation of automated docking programs for docking and database screening against RNA drug targets. *J Med Chem* **2004**, 47, 4188-201.

67. Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L., Opinion - How many drug targets are there? *Nat Rev Drug Discov* **2006**, 5, 993-996.

68. Hopkins, A. L.; Groom, C. R., The druggable genome. *Nat Rev Drug Discov* **2002**, 1, 727-30.

69. Lybrand, T. P., Ligand-protein docking and rational drug design. *Curr Opin Struct Biol* **1995**, 5, 224-8.

70. Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A., The maximal affinity of ligands. *P Natl Acad Sci USA* **1999**, 96, 9997-10002.

71. Cohen, N. C.; Blaney, J. M.; Humblet, C.; Gund, P.; Barry, D. C., Molecular modeling software and methods for medicinal chemistry. *J Med Chem* **1990**, 33, 883-94.

72. Joseph-McCarthy, D., Computational approaches to structure-based ligand design. *Pharmacol Ther* **1999**, 84, 179-91.

73. Marshall, G. R., Computer-aided drug design. *Annu Rev Pharmacol Toxicol* **1987**, 27, 193-213.

74. Ou-Yang, S. S.; Lu, J. Y.; Kong, X. Q.; Liang, Z. J.; Luo, C.; Jiang, H., Computational drug discovery. *Acta Pharmacol Sin* **2012**, 33, 1131-40.

75. Hornak, V.; Dvorsky, R.; Sturdik, E., Receptor-ligand interaction and molecular modelling. *Gen Physiol Biophys* **1999**, 18, 231-48.

76. Alvarez-Garcia, D.; Seco, J.; Schmidtke, P.; Barril, X., Druggability Prediction. *Protein-Ligand Interactions, First Edition* **2012**, 265-282.

77. Henrich, S.; Salo-Ahen, O. M.; Huang, B.; Rippmann, F. F.; Cruciani, G.; Wade, R. C., Computational approaches to identifying and characterizing protein binding sites for ligand design. *J Mol Recognit* **2010**, 23, 209-19.

78. Craig, I. R.; Pfleger, C.; Gohlke, H.; Essex, J. W.; Spiegel, K., Pocket-space maps to identify novel binding-site conformations in proteins. *J Chem Inf Model* **2011**, 51, 2666-79.

79. Schmidtke, P.; Barril, X., Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J Med Chem* **2010**, 53, 5858-67.

80. Sotriffer, C. A.; Gohlke, H.; Klebe, G., Docking into knowledge-based potential fields: A comparative evaluation of DrugScore. *J Med Chem* **2002**, 45, 1967-1970.

81. Vajda, S.; Guarnieri, F., Characterization of protein-ligand interaction sites using experimental and computational methods. *Curr Opin Drug Discov Devel* **2006**, 9, 354-362.

82. Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S., Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol* **2007**, 25, 71-5.

83. Sousa, S. F.; Ribeiro, A. J.; Coimbra, J. T.; Neves, R. P.; Martins, S. A.; Moorthy, N. S.; Fernandes, P. A.; Ramos, M. J., Protein-ligand docking in the new millennium--a retrospective of 10 years in the field. *Curr Med Chem* **2013**, 20, 2296-314.

84. Sousa, S. F.; Cerqueira, N. M.; Fernandes, P. A.; Ramos, M. J., Virtual screening in drug design and development. *Comb Chem High Throughput Screen* **2010**, 13, 442-53.

85. van Dijk, A. D.; Boelens, R.; Bonvin, A. M., Data-driven docking for the study of biomolecular complexes. *The FEBS journal* **2005**, 272, 293-312.

86. Kubinyi, H., Strategies and recent technologies in drug discovery. *Die Pharmazie* **1995**, 50, 647-62.

87. Leach, A. R.; Shoichet, B. K.; Peishoff, C. E., Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J Med Chem* **2006**, 49, 5851-5.

88. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **1996**, 261, 470-489.

89. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R., Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **1997**, 267, 727-48.

90. Kontoyianni, M.; McClellan, L. M.; Sokol, G. S., Evaluation of docking performance: comparative data on docking algorithms. *J Med Chem* **2004**, 47, 558-65.

91. Sotriffer, C., Scoring functions for protein–ligand interactions. *Protein-Ligand Interactions, First Edition* **2012**, 237-263.

92. Gohlke, H.; Klebe, G., Statistical potentials and scoring functions applied to protein-ligand binding. *Curr Opin Struc Biol* **2001**, 11, 231-235.

93. Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M., BLEEP - Potential of mean force describing protein-ligand interactions: I. Generating potential. *J Comput Chem* **1999**, 20, 1165-1176.

94. Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Forster, M. J.; Thornton, J. M., BLEEP - Potential of mean force describing protein-ligand interactions: II. Calculation of binding energies and comparison with experimental data. *J Comput Chem* **1999**, 20, 1177-1185.

95. Muegge, I., PMF scoring revisited. *J Med Chem* **2006**, 49, 5895-902.

96. Ishchenko, A. V.; Shakhnovich, E. I., SMall Molecule Growth 2001 (SMoG2001): an improved knowledge-based scoring function for protein-ligand interactions. *J Med Chem* **2002**, 45, 2770-80.

97. Gohlke, H.; Hendlich, M.; Klebe, G., Predicting binding modes, binding affinities and 'hot spots' for protein-ligand complexes using a knowledge-based scoring function. *Perspect Drug Discov* **2000**, 20, 115-144.

98. Gohlke, H.; Hendlich, M.; Klebe, G., Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol* **2000**, 295, 337-356.

99. Kruger, D. M.; Jessen, G.; Gohlke, H., How Good Are State-of-the-Art Docking Tools in Predicting Ligand Binding Modes in Protein-Protein Interfaces? *J Chem Inf Model* **2012**, 52, 2807-2811.

100. Kazemi, S.; Kruger, D. M.; Sirockin, F.; Gohlke, H., Elastic Potential Grids: Accurate and Efficient Representation of Intermolecular Interactions for Fully Flexible Docking. *Chemmedchem* **2009**, 4, 1264-1268.

101. Gohlke, H.; Klebe, G., DrugScore meets CoMFA: adaptation of fields for molecular comparison (AFMoC) or how to tailor knowledge-based pair-potentials to a particular protein. *J Med Chem* **2002**, 45, 4153-4170.

102. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res* **2000**, 28, 235-242.

103. Hsieh, J. H.; Yin, S.; Wang, X. S.; Liu, S.; Dokholyan, N. V.; Tropsha, A., Cheminformatics meets molecular mechanics: a combined application of knowledge-based pose scoring and physical force field-based hit scoring functions improves the accuracy of structure-based virtual screening. *J Chem Inf Model* **2012**, 52, 16-28.

104. Yin, S.; Biedermannova, L.; Vondrasek, J.; Dokholyan, N. V., MedusaScore: an accurate force field-based scoring function for virtual drug screening. *J Chem Inf Model* **2008**, 48, 1656-62.

105. Hsieh, M. J.; Luo, R., Physical scoring function based on AMBER force field and Poisson-Boltzmann implicit solvent for protein structure prediction. *Proteins* **2004**, 56, 475-86.

106. Grzybowski, B. A.; Ishchenko, A. V.; Shimada, J.; Shakhnovich, E. I., From knowledge-based potentials to combinatorial lead design in silico. *Acc Chem Res* **2002**, 35, 261-9.

107. Lee, M. C.; Duan, Y., Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized born solvent model. *Proteins* **2004**, 55, 620-634.

108. Petrella, R. J.; Lazaridis, T.; Karplus, M., Protein sidechain conformer prediction: a test of the energy function. *Fold Des* **1998**, 3, 353-77.

109. MacKerell, A. D.; Brooks, B.; Brooks, C. L.; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M., CHARMM: the energy function and its parameterization. *Encyclopedia of computational chemistry* **1998**.

110. Shivakumar, D.; Williams, J.; Wu, Y. J.; Damm, W.; Shelley, J.; Sherman, W., Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. *J Chem Theory Comput* **2010**, 6, 1509-1519.

111. Udier-Blagovic, M.; De Tirado, P. M.; Pearlman, S. A.; Jorgensen, W. L., Accuracy of free energies of hydration using CM1 and CM3 atomic charges. *J Comput Chem* **2004**, 25, 1322-1332.

112. Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J., Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc* **1996**, 118, 11225-11236.

113. Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M., Validation and use of the MM-PBSA approach for drug discovery. *J Med Chem* **2005**, 48, 4040-4048.

114. Weis, A.; Katebzadeh, K.; Soderhjelm, P.; Nilsson, I.; Ryde, U., Ligand affinities predicted with the MM/PBSA method: dependence on the simulation method and the force field. *J Med Chem* **2006**, 49, 6596-606.

115. Metz, A.; Pfleger, C.; Kopitz, H.; Pfeiffer-Marek, S.; Baringhaus, K. H.; Gohlke, H., Hot Spots and Transient Pockets: Predicting the Determinants of Small-Molecule Binding to a Protein-Protein Interface. *J Chem Inf Model* **2012**, 52, 120-133.

116. Hou, T. J.; Wang, J. M.; Li, Y. Y.; Wang, W., Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *J Chem Inf Model* **2011**, 51, 69-82.

117. Swanson, J. M. J.; Henchman, R. H.; McCammon, J. A., Revisiting free energy calculations: A theoretical connection to MM/PBSA and direct calculation of the association free energy. *Biophys J* **2004**, 86, 67-74.

118. Liu, J.; Wang, R., Classification of current scoring functions. *J Chem Inf Model* **2015**, 55, 475-82.

119. Bohm, H. J., The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J Comput Aided Mol Des* **1994**, 8, 243-56.

120. Korb, O.; Stutzle, T.; Exner, T. E., Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J Chem Inf Model* **2009**, 49, 84-96.

121. Pham, T. A.; Jain, A. N., Customizing scoring functions for docking. *J Comput Aided Mol Des* **2008**, 22, 269-86.

122. Bohm, H. J., Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J Comput Aided Mol Des* **1998**, 12, 309-23.

123. Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P., Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des* **1997**, 11, 425-45.

124. Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T., Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem* **2006**, 49, 6177-96.

125. Wang, R. X.; Lai, L. H.; Wang, S. M., Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aid Mol Des* **2002**, 16, 11-26.

126. Jain, A. N., Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J Comput Aided Mol Des* **2007**, 21, 281-306.

127. Jain, A. N., Effects of protein conformation in docking: improved pose prediction through protein pocket adaptation. *J Comput Aid Mol Des* **2009**, 23, 355-374.

128. Jain, A. N., Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem* **2003**, 46, 499-511.

129. Bleicher, K. H.; Bohm, H. J.; Muller, K.; Alanine, A. I., Hit and lead generation: beyond high-throughput screening. *Nat Rev Drug Discov* **2003**, 2, 369-78.

130. Gao, H.; Katzenellenbogen, J. A.; Garg, R.; Hansch, C., Comparative QSAR analysis of estrogen receptor ligands. *Chem Rev* **1999**, 99, 723-44.

131. Headley, A. D.; Mcmurry, M. E.; Starnes, S. D., Effects of Substituents on the Acidity of Acetic-Acids. *J Org Chem* **1994**, 59, 1863-1866.

132. Johnson, M. A.; Maggiora, G. M., Concepts and applications of molecular similarity. **1990**.

133. Perkins, R.; Fang, H.; Tong, W.; Welsh, W. J., Quantitative structure-activity relationship methods: perspectives on drug discovery and toxicology. *Environmental toxicology and chemistry / SETAC* **2003**, 22, 1666-79.

134. Durrant, J. D.; McCammon, J. A., NNScore 2.0: a neural-network receptor-ligand scoring function. *J Chem Inf Model* **2011**, 51, 2897-903.

135. 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, 1865-71.

136. Ballester, P. J.; Schreyer, A.; Blundell, T. L., Does a more precise chemical description of protein-ligand complexes lead to more accurate prediction of binding affinity? *J Chem Inf Model* **2014**, 54, 944-55.

137. Ballester, P. J.; Mitchell, J. B., A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, 26, 1169-75.

138. Zilian, D.; Sotriffer, C. A., SFCscore(RF): a random forest-based scoring function for improved affinity prediction of protein-ligand complexes. *J Chem Inf Model* **2013**, 53, 1923-33.

139. Li, G. B.; Yang, L. L.; Wang, W. J.; Li, L. L.; Yang, S. Y., ID-Score: a new empirical scoring function based on a comprehensive set of descriptors related to protein-ligand interactions. *J Chem Inf Model* **2013**, 53, 592-600.

140. Nishikaw.K; Ooi, T.; Saito, N.; Isogai, Y., Tertiary Structure of Proteins .1. Representation and Computation of Conformations. *J Phys Soc Jpn* **1972**, 32, 1331-&.

141. Levitt, M., Simplified Representation of Protein Conformations for Rapid Simulation of Protein Folding. *J Mol Biol* **1976**, 104, 59-107.

142. Maiorov, V. N.; Crippen, G. M., Significance of Root-Mean-Square Deviation in Comparing 3-Dimensional Structures of Globular-Proteins. *J Mol Biol* **1994**, 235, 625-634.

143. Merlin, J. C.; Cornard, J. P., A pictorial representation of normal modes of vibration using vibrational symmetry coordinates. *J Chem Educ* **2006**, 83, 1393-1398.

144. Hinsen, K., Normal mode theory and harmonic potential approximations. *Boca Raton: Chapman & Hall/CRC* **2006**, 1-16.

145. Hub, J. S.; de Groot, B. L., Detection of functional modes in protein dynamics. *Plos Comput Biol* **2009**, 5, e1000480.

146. Dykeman, E. C.; Sankey, O. F., Normal mode analysis and applications in biological physics. *J Phys Condens Matter* **2010**, 22, 423202.

147. Hayward, S.; Kitao, A.; Berendsen, H. J., Model-free methods of analyzing domain motions in proteins from simulation: a comparison of normal mode analysis and molecular dynamics simulation of lysozyme. *Proteins* **1997**, 27, 425-37.

148. Gibrat, J. F.; Go, N., Normal mode analysis of human lysozyme: study of the relative motion of the two domains and characterization of the harmonic motion. *Proteins* **1990**, 8, 258-79.

149. Marques, O.; Sanejouand, Y. H., Hinge-bending motion in citrate synthase arising from normal mode calculations. *Proteins* **1995**, 23, 557-60.

150. Ma, J.; Karplus, M., The allosteric mechanism of the chaperonin GroEL: a dynamic analysis. *Proc Natl Acad Sci U S A* **1998**, 95, 8502-7.

151. Thomas, A.; Field, M. J.; Perahia, D., Analysis of the low-frequency normal modes of the R state of aspartate transcarbamylase and a comparison with the T state modes. *J Mol Biol* **1996**, 261, 490-506.

152. Thomas, A.; Field, M. J.; Mouawad, L.; Perahia, D., Analysis of the low frequency normal modes of the T-state of aspartate transcarbamylase. *J Mol Biol* **1996**, 257, 1070-87.

153. Thomas, A.; Hinsen, K.; Field, M. J.; Perahia, D., Tertiary and quaternary conformational changes in aspartate transcarbamylase: a normal mode study. *Proteins* **1999**, 34, 96-112.

154. Kidera, A.; Inaka, K.; Matsushima, M.; Go, N., Normal mode refinement: crystallographic refinement of protein dynamic structure. II. Application to human lysozyme. *J Mol Biol* **1992**, 225, 477-86.

155. Kidera, A.; Go, N., Normal mode refinement: crystallographic refinement of protein dynamic structure. I. Theory and test by simulated diffraction data. *J Mol Biol* **1992**, 225, 457-75.

156. Bruschweiler, R.; Case, D. A., Collective NMR relaxation model applied to protein dynamics. *Phys Rev Lett* **1994**, 72, 940-943.

157. Matsumoto, A.; Tomimoto, M.; Go, N., Dynamical structure of transfer RNA studied by normal mode analysis. *Eur Biophys J Biophy* **1999**, 28, 369-379.

158. Duong, T. H.; Zakrzewska, K., Calculation and analysis of low frequency normal modes for DNA. *J Comput Chem* **1997**, 18, 796-811.

159. Matsumoto, A.; Go, N., Dynamic properties of double-stranded DNA by normal mode analysis. *J Chem Phys* **1999**, 110, 11070-11075.

160. Stern, P. S.; Chorev, M.; Goodman, M.; Hagler, A. T., Computer simulation of the conformational properties of retro-inverso peptides. I. Empirical force field calculations of rigid and flexible geometries of N-acetylglycine-N'-methylamide, bis(acetamido) methane, and N,N'-dimethylmalonamide and their corresponding C alpha-methylated analogs. *Biopolymers* **1983**, 22, 1885-900.

161. Ma, J., Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure* **2005**, 13, 373-80.

162. Kottalam, J.; Case, D. A., Langevin modes of macromolecules: applications to crambin and DNA hexamers. *Biopolymers* **1990**, 29, 1409-21.

163. Lamm, G.; Szabo, A., Langevin Modes of Macromolecules. *J Chem Phys* **1986**, 85, 7334-7348.

164. Jacobs, D. J.; Rader, A. J.; Kuhn, L. A.; Thorpe, M. F., Protein flexibility predictions using graph theory. *Proteins Struct Funct Genet* **2001**, 44, 150-165.

165. Thorpe, M. F., Continuous Deformations in Random Networks. *J Non-Cryst Solids* **1983**, 57, 355-370.

166. Jacobs, D. J.; Thorpe, M. F., Generic rigidity percolation in two dimensions. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* **1996**, 53, 3682-3693.

167. Jacobs, D. J.; Thorpe, M. F., Generic rigidity percolation: The pebble game. *Phys Rev Lett* **1995**, 75, 4051-4054.

168. Farrell, D. W., Generating Stereochemically Acceptable Protein Pathways *Thesis* **2010**.

169. Jacobs, D. J.; Thorpe, M. F., Generic rigidity percolation in two dimensions. *Phys Rev E* **1996**, 53, 3682-3693.

170. Jacobs, D. J., Generic rigidity in three-dimensional bond-bending networks. *J Phys a-Math Gen* **1998**, 31, 6653-6668.

171. Chubynsky, M. V.; Thorpe, M. F., Algorithms for three-dimensional rigidity analysis and a first-order percolation transition. *Phys Rev E* **2007**, 76.

172. Thorpe, M. F.; Lei, M.; Rader, A. J.; Jacobs, D. J.; Kuhn, L. A., Protein flexibility and dynamics using constraint theory. *J Mol Graph Model* **2001**, 19, 60-9.

173. Briggs, J. A.; Grunewald, K.; Glass, B.; Forster, F.; Krausslich, H. G.; Fuller, S. D., The mechanism of HIV-1 core assembly: insights from three-dimensional reconstructions of authentic virions. *Structure* **2006**, 14, 15-20.

174. Kaplan, A. H., Assembly of the HIV-1 core particle. *AIDS reviews* **2002**, 4, 104-11.

175. Zhang, S.; Kaplan, A. H.; Tropsha, A., HIV-1 protease function and structure studies with the simplicial neighborhood analysis of protein packing method. *Proteins* **2008**, 73, 742-53.

176. Debouck, C., The HIV-1 protease as a therapeutic target for AIDS. *AIDS research and human retroviruses* **1992**, 8, 153-64.

177. Wlodawer, A.; Erickson, J. W., Structure-Based Inhibitors of Hiv-1 Protease. *Annu Rev Biochem* **1993**, 62, 543-585.

178. Katoh, I.; Yasunaga, T.; Ikawa, Y.; Yoshinaka, Y., Inhibition of retroviral protease activity by an aspartyl proteinase inhibitor. *Nature* **1987**, 329, 654-6.

179. Schechter, I.; Berger, A., On the size of the active site in proteases. I. Papain. *Biochem Biophys Res Commun* **1967**, 27, 157-62.

180. Schechter, I., Mapping of the active site of proteases in the 1960s and rational design of inhibitors/drugs in the 1990s. *Curr Protein Pept Sci* **2005**, 6, 501-12.

181. Brik, A.; Wong, C. H., HIV-1 protease: mechanism and drug discovery. *Org Biomol Chem* **2003**, 1, 5-14.

182. Babine, R. E.; Bender, S. L., Molecular recognition of protein-ligand complexes: Applications to drug design. *Chem Rev* **1997**, 97, 1359-1472.

183. Yang, H. L.; Nkeze, J.; Zhao, R. Y., Effects of HIV-1 protease on cellular functions and their potential applications in antiretroviral therapy. *Cell Biosci* **2012**, 2.

184. Hertzberg, M., Biochemistry of factor X. *Blood reviews* **1994**, 8, 56-62.

185. Freer, S. T.; Kraut, J.; Robertus, J. D.; Wright, H. T.; Xuong, N. H., Chymotrypsinogen: 2.5-angstrom crystal structure, comparison with alpha-chymotrypsin, and implications for zymogen activation. *Biochemistry US* **1970**, 9, 1997-2009.

186. Nar, H., The role of structural information in the discovery of direct thrombin and factor Xa inhibitors. *Trends Pharmacol Sci* **2012**, 33, 279-288.

187. Lin, Z.; Johnson, M. E., Proposed cation-pi mediated binding by factor Xa: a novel enzymatic mechanism for molecular recognition. *FEBS Lett* **1995**, 370, 1-5.

188. Hauptmann, J.; Sturzebecher, J., Synthetic inhibitors of thrombin and factor Xa: from bench to bedside. *Thrombosis research* **1999**, 93, 203-41.

189. Li, J.; Buchner, J., Structure, function and regulation of the hsp90 machinery. *Biomed J* **2013**, 36, 106-17.

190. Taldone, T.; Sun, W. L.; Chiosis, G., Discovery and development of heat shock protein 90 inhibitors. *Bioorgan Med Chem* **2009**, 17, 2225-2235.

191. Young, J. C.; Agashe, V. R.; Siegers, K.; Hartl, F. U., Pathways of chaperone-mediated protein folding in the cytosol. *Nat Rev Mol Cell Biol* **2004**, 5, 781-91.

192. Hartl, F. U.; Hayer-Hartl, M., Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* **2002**, 295, 1852-8.

193. Southworth, D. R.; Agard, D. A., Species-dependent ensembles of conserved conformational states define the Hsp90 chaperone ATPase cycle. *Mol Cell* **2008**, 32, 631-40.

194. Lavery, L. A.; Partridge, J. R.; Ramelot, T. A.; Elnatan, D.; Kennedy, M. A.; Agard, D. A., Structural asymmetry in the closed state of mitochondrial Hsp90 (TRAP1) supports a two-step ATP hydrolysis mechanism. *Mol Cell* **2014**, 53, 330-43.

195. Blagosklonny, M. V.; Fojo, T.; Bhalla, K. N.; Kim, J. S.; Trepel, J. B.; Figg, W. D.; Rivera, Y.; Neckers, L. M., The Hsp90 inhibitor geldanamycin selectively sensitizes Bcr-Abl-expressing leukemia cells to cytotoxic chemotherapy. *Leukemia* **2001**, 15, 1537-43.

196. Pratt, W. B.; Toft, D. O., Steroid receptor interactions with heat shock protein and immunophilin chaperones. *Endocr Rev* **1997**, 18, 306-60.

197. Neckers, L.; Workman, P., Hsp90 Molecular Chaperone Inhibitors: Are We There Yet? *Clin Cancer Res* **2012**, 18, 64-76.

198. Chiosis, G., Targeting chaperones in transformed systems - a focus on Hsp90 and cancer. *Expert Opin Ther Targets* **2006**, 10, 37-50.

199. Whitesell, L.; Lindquist, S. L., HSP90 and the chaperoning of cancer. *Nat Rev Cancer* **2005**, 5, 761-72.

200. Nanbu, K.; Konishi, I.; Mandai, M.; Kuroda, H.; Hamid, A. A.; Komatsu, T.; Mori, T., Prognostic significance of heat shock proteins HSP70 and HSP90 in endometrial carcinomas. *Cancer Detect Prev* **1998**, 22, 549-55.

201. Liu, X. L.; Xiao, B.; Yu, Z. C.; Guo, J. C.; Zhao, Q. C.; Xu, L.; Shi, Y. Q.; Fan, D. M., Down-regulation of Hsp90 could change cell cycle distribution and increase drug sensitivity of tumor cells. *World J Gastroenterol* **1999**, 5, 199-208.

202. Neckers, L.; Mimnaugh, E.; Schulte, T. W., Hsp90 as an anti-cancer target. *Drug Resist Updat* **1999**, 2, 165-172.

203. Yano, M.; Naito, Z.; Yokoyama, M.; Shiraki, Y.; Ishiwata, T.; Inokuchi, M.; Asano, G., Expression of hsp90 and cyclin D1 in human breast cancer. *Cancer Lett* **1999**, 137, 45-51.

204. Whitesell, L.; Shifrin, S. D.; Schwab, G.; Neckers, L. M., Benzoquinonoid ansamycins possess selective tumoricidal activity unrelated to src kinase inhibition. *Cancer Res* **1992**, 52, 1721-8.

205. Sidera, K.; Patsavoudi, E., HSP90 inhibitors: current development and potential in cancer therapy. *Recent Pat Anticancer Drug Discov* **2014**, 9, 1-20.

206. Mimnaugh, E. G.; Chavany, C.; Neckers, L., Polyubiquitination and proteasomal degradation of the p185c-erbB-2 receptor protein-tyrosine kinase induced by geldanamycin. *J Biol Chem* **1996**, 271, 22796-801.

207. Roe, S. M.; Prodromou, C.; O'Brien, R.; Ladbury, J. E.; Piper, P. W.; Pearl, L. H., Structural basis for inhibition of the Hsp90 molecular chaperone by the antitumor antibiotics radicicol and geldanamycin. *J Med Chem* **1999**, 42, 260-6.

208. Page, M. J.; Di Cera, E., Serine peptidases: classification, structure and function. *Cell Mol Life Sci* **2008**, 65, 1220-36.

209. Huber, R.; Bode, W., Structural Basis of the Activation, Action and Inhibition of Trypsin. *Physiol Chem* **1979**, 360, 489-489.

210. Polgar, L., The catalytic triad of serine peptidases. *Cell Mol Life Sci* **2005**, 62, 2161-72.

211. Lin, C. Y.; Anders, J.; Johnson, M.; Sang, Q. A.; Dickson, R. B., Molecular cloning of cDNA for matriptase, a matrix-degrading serine protease with trypsin-like activity. *J Biol Chem* **1999**, 274, 18231-6.

212. Ajay; Murcko, M. A., Computational methods to predict binding free energy in ligand-receptor complexes. *J Med Chem* **1995**, 38, 4953-4967.

213. Baron, R.; McCammon, J. A., Molecular recognition and ligand association. *Annu Rev Phys Chem* **2013**, 64, 151-75.

214. Freire, E.; Biltonen, R. L., Statistical Mechanical Deconvolution of Thermal Transitions in Macromolecules. 1.Theory and Application to Homogeneous Systems. *Biopolymers* **1978**, 17, 463-479.

215. Hill, T. L., An Introduction to Statistical Thermodynamics. *New York: Courier Corporation* **2012**.

216. Laurendeau, N. M., *Statistical thermodynamics: fundamentals and applications*. Cambridge University Press: 2005.

217. Levy, R. M.; Karplus, M.; Kushick, J.; Perahia, D., Evaluation of the Configurational Entropy for Proteins - Application to Molecular-Dynamics Simulations of an Alpha-Helix. *Macromolecules* **1984**, 17, 1370-1374.

218. Neff, R. O.; Mcquarri.Da, Statistical Mechanical Theory of Solubility. *Journal of Physical Chemistry* **1973**, 77, 413-418.

219. Attard, P.; Jepps, O. G.; Marcelja, S., Information content of signals using correlation function expansions of the entropy. *Phys Rev E* **1997**, 56, 4052-4067.

220. Matsuda, H., Physical nature of higher-order mutual information: intrinsic correlations and frustration. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* **2000**, 62, 3096-102.

221. Killian, B. J.; Yundenfreund Kravitz, J.; Gilson, M. K., Extraction of configurational entropy from molecular simulations via an expansion approximation. *J Chem Phys* **2007**, 127, 024107.

222. Killian, B. J.; Kravitz, J. Y.; Somani, S.; Dasgupta, P.; Pang, Y. P.; Gilson, M. K., Configurational entropy in protein-peptide binding: computational study of Tsg101 ubiquitin E2 variant domain with an HIV-derived PTAP nonapeptide. *J Mol Biol* **2009**, 389, 315-35.

223. Lim, M. S.; Johnston, E. R.; Kettner, C. A., The solution conformation of (D)Phe-Pro-containing peptides: implications on the activity of Ac-(D)Phe-Pro-boroArg-OH, a potent thrombin inhibitor. *J Med Chem* **1993**, 36, 1831-8.

224. Testa, B.; Carrupt, P. A.; Gaillard, P.; Billois, F.; Weber, P., Lipophilicity in molecular modeling. *Pharmaceut Res* **1996**, 13, 335-343.

225. Zavodszky, M. I.; Kuhn, L. A., Side-chain flexibility in protein-ligand binding: the minimal rotation hypothesis. *Protein Sci* **2005**, 14, 1104-14.

226. Najmanovich, R.; Kuttner, J.; Sobolev, V.; Edelman, M., Side-chain flexibility in proteins upon ligand binding. *Proteins Struct Funct Genet* **2000**, 39, 261-268.

227. Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A., The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys J* **1997**, 72, 1047-1069.

228. Schafer, H.; Mark, A. E.; van Gunsteren, W. F., Absolute entropies from molecular dynamics simulation trajectories. *J Chem Phys* **2000**, 113, 7809-7817.

229. Mcquarri.Da, Statistical Mechanics. *Phys Today* **1965**, 18, 74-&.

230. Gilson, M. K.; Irikura, K. K., Symmetry Numbers for Rigid, Flexible, and Fluxional Molecules: Theory and Applications. *J Phys Chem B* **2010**, 114, 16304-16317.

231. Wei, J., Molecular symmetry, rotational entropy, and elevated melting points. *Ind Eng Chem Res* **1999**, 38, 5019-5027.

232. Strang, G., Linear algebra and its applications. *New York: Academic Press* **1976**.

233. Bar-Itzhack, I. Y., New method for extracting the quaternion from a rotation matrix. *J Guid Control Dynam* **2000**, 23, 1085-1087.

234. Berner, P.; Toms, R.; Trott, K.; Mamaghani, F.; Shen, D.; Rollins, C.; Powell, E., Technical Concepts Orientation, Rotation, Velocity and Acceleration, and the SRM. *TENA (Test & Training Enabling Architecture) project by SEDRIS* **2008**, 21.

235. Kubo, R., Statistical Physics: An Advanced Course with Problems and Solutions. *Amsterdam: North Holland* **1988**.

236. Pflug, A.; Johnson, K. A.; Engh, R. A., Anomalous dispersion analysis of inhibitor flexibility: a case study of the kinase inhibitor H-89. *Acta Crystallogr F* **2012**, 68, 873-877.

237. Lewis, P. J.; de Jonge, M.; Daeyaert, F.; Koymans, L.; Vinkers, M.; Heeres, J.; Janssen, P. A. J.; Arnold, E.; Das, K.; Clark, A. D.; Hughes, S. H.; Boyer, P. L.; de Bethune, M. P.; Pauwels, R.; Andries, K.; Kukla, M.; Ludovici, D.; De Corte, B.; Kavash, R.; Ho, C., On the detection of multiple-binding modes of ligands to proteins, from biological, structural, and modeling data. *J Comput Aid Mol Des* **2003**, 17, 129-134.

238. Uytterhoeven, K.; Sponer, J.; Van Meervelt, L., Two 1 : 1 binding modes for distamycin in the minor groove of d(GGCCAATTGG). *Eur J Biochem* **2002**, 269, 2868-2877.

239. Wojtczak, A.; Cody, V.; Luft, J. R.; Pangborn, W., Structure of rat transthyretin (rTTR) complex with thyroxine at 2.5 angstrom resolution: first non-biased insight into thyroxine binding reveals different hormone orientation in two binding sites. *Acta Crystallogr D* **2001**, 57, 1061-1070.

240. Dohnalek, J.; Hasek, J.; Duskova, J.; Petrokova, H.; Hradilek, M.; Soucek, M.; Konvalinka, J.; Brynda, J.; Sedlacek, J.; Fabry, M., A distinct binding mode of a hydroxyethylamine isostere inhibitor of HIV-1 protease. *Acta Crystallogr D* **2001**, 57, 472-476.

241. Badger, J.; Minor, I.; Kremer, M. J.; Oliveira, M. A.; Smith, T. J.; Griffith, J. P.; Guerin, D. M. A.; Krishnaswamy, S.; Luo, M.; Rossmann, M. G.; Mckinlay, M. A.; Diana, G. D.; Dutko, F. J.; Fancher, M.; Rueckert, R. R.; Heinz, B. A., Structural-Analysis of a Series of Antiviral Agents Complexed with Human Rhinovirus-14. *P Natl Acad Sci USA* **1988**, 85, 3304-3308.

242. Young, R. D.; Bowne, S. F., Conformational Substates and Barrier Height Distributions in Ligand-Binding to Heme-Proteins. *J Chem Phys* **1984**, 81, 3730-3737.

243. Gilson, M. K., Multiple-site titration and molecular modeling: two rapid methods for computing energies and forces for ionizable groups in proteins. *Proteins* **1993**, 15, 266-82.

244. 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.

245. Head, M. S.; Given, J. A.; Gilson, M. K., "Mining minima": Direct computation of conformational free energy. *Biophys J* **1997**, 72, Wp437-Wp437.

246. Vieth, M.; Kolinski, A.; Skolnick, J., A Simple Technique to Estimate Partition-Functions and Equilibrium-Constants from Monte-Carlo Simulations. *J Chem Phys* **1995**, 102, 6189-6193.

247. Vieth, M.; Kolinski, A.; Brooks, C. L.; Skolnick, J., Prediction of Quaternary Structure of Coiled Coils - Application to Mutants of the Gcn4 Leucine-Zipper. *J Mol Biol* **1995**, 251, 448-467.

248. Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Freer, S. T.; Rose, P. W., Complexity and simplicity of ligand-macromolecule interactions: the energy landscape perspective. *Curr Opin Struc Biol* **2002**, 12, 197-203.

249. Held, M.; Imhof, P.; Keller, B. G.; Noe, F., Modulation of a Ligand's Energy Landscape and Kinetics by the Chemical Environment. *J Phys Chem B* **2012**, 116, 13597-13607.

250. Janshoff, A.; Steinem, C., Energy landscapes of ligand-receptor couples probed by dynamic force spectroscopy. *ChemPhysChem* **2001**, 2, 577-579.

251. Da, C.; Kireev, D., Structural protein-ligand interaction fingerprints (SPLIF) for structure-based virtual screening: method and benchmark study. *J Chem Inf Model* **2014**, 54, 2555-61.

252. Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S., A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application. *J Chem Inf Model* **2007**, 47, 279-94.

253. Brooks, B. R.; Janezic, D.; Karplus, M., Harmonic-Analysis of Large Systems. 1.Methodology. *J Comput Chem* **1995**, 16, 1522-1542.

254. McQuarrie, D. A., Statistical thermodynamics. *New York: Harper and Row* **1973**.

255. Ponder, J. W.; Case, D. A., Force fields for protein simulations. *Adv Protein Chem* **2003**, 66, 27-85.

256. Kirkwood, J. G., The dielectric polarization of polar liquids. *J Chem Phys* **1939**, 7, 911-919.

257. Keating, P., Effect of invariance requirements on the elastic strain energy of crystals with application to the diamond structure. *Phys Rev* **1966**, 145, 637.

258. Pfleger, C.; Rathi, P. C.; Klein, D. L.; Radestock, S.; Gohlke, H., Constraint Network Analysis (CNA): A Python Software Package for Efficiently Linking Biomacromolecular Structure, Flexibility, (Thermo-)Stability, and Function. *J Chem Inf Model* **2013**, 53, 1007-1015.

259. Rathi, P. C.; Pfleger, C.; Fulle, S.; Klein, D. L.; Gohlke, H., Statics of biomacromolecules. *Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA* **2011**, 281-299.

260. Maxwell, J. C., XLV. On reciprocal figures and diagrams of forces. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1864**, 27, 250-261.

261. Lee, A.; Streinu, I., Pebble game algorithms and sparse graphs. *Discrete Math* **2008**, 308, 1425-1437.

262. Duxbury, P. M.; Jacobs, D. J.; Thorpe, M. F.; Moukarzel, C., Floppy modes and the free energy: Rigidity and connectivity percolation on Bethe lattices. *Phys Rev E* **1999**, 59, 2084-2092.

263. Thorpe, M. F.; Jacobs, D. J.; Chubynsky, M. V.; Phillips, J. C., Self-organization in network glasses. *J Non-Cryst Solids* **2000**, 266, 859-866.

264. Radestock, S.; Gohlke, H., Exploiting the Link between Protein Rigidity and Thermostability for Data-Driven Protein Engineering. *Eng Life Sci* **2008**, 8, 507-522.

265. Rader, A. J., Thermostability in rubredoxin and its relationship to mechanical rigidity. *Phys Biol* **2009**, 7, 16002.

266. Taverna, D. M.; Goldstein, R. A., Why are proteins marginally stable? *Proteins* **2002**, 46, 105-9.

267. Gohlke, H.; Kuhn, L. A.; Case, D. A., Change in protein flexibility upon complex formation: analysis of Ras-Raf using molecular dynamics and a molecular framework approach. *Proteins* **2004**, 56, 322-37.

268. Mamonova, T.; Hespenheide, B.; Straub, R.; Thorpe, M. F.; Kurnikova, M., Protein flexibility using constraints from molecular dynamics simulations. *Phys Biol* **2005**, 2, S137-S147.

269. Homeyer, N.; Gohlke, H., Free Energy Calculations by the Molecular Mechanics Poisson-Boltzmann Surface Area Method. *Mol Inform* **2012**, 31, 114-122.

270. Rader, A. J.; Hespenheide, B. M.; Kuhn, L. A.; Thorpe, M. F., Protein unfolding: rigidity lost. *Proc Natl Acad Sci U S A* **2002**, 99, 3540-5.

271. Hespenheide, B. M.; Rader, A. J.; Thorpe, M. F.; Kuhn, L. A., Identifying protein folding cores from the evolution of flexible regions during unfolding. *J Mol Graph Model* **2002**, 21, 195-207.

272. Radestock, S.; Gohlke, H., Protein rigidity and thermophilic adaptation. *Proteins* **2011**, 79, 1089-108.

273. Livesay, D. R.; Jacobs, D. J., Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair. *Proteins* **2006**, 62, 130-43.

274. Dahiyat, B. I.; Gordon, D. B.; Mayo, S. L., Automated design of the surface positions of protein helices. *Protein Sci* **1997**, 6, 1333-7.

275. Rathi, P. C.; Radestock, S.; Gohlke, H., Thermostabilizing mutations preferentially occur at structural weak spots with a high mutation ratio. *J Biotechnol* **2012**, 159, 135-44.

276. Rathi, P. C.; Jaeger, K. E.; Gohlke, H., Structural Rigidity and Protein Thermostability in Variants of Lipase A from Bacillus subtilis. *Plos One* **2015**, 10, e0130289.

277. Pfleger, C.; Gohlke, H., Efficient and Robust Analysis of Biomacromolecular Flexibility Using Ensembles of Network Topologies Based on Fuzzy Noncovalent Constraints. *Structure* **2013**, 21, 1725-1734.

278. Dick, M.; Weiergraber, O. H.; Classen, T.; Bisterfeld, C.; Bramski, J.; Gohlke, H.; Pietruszka, J., Trading off stability against activity in extremophilic aldolases. *Sci Rep* **2016**, 6, 17908.

279. Rader, A. J.; Anderson, G.; Isin, B.; Khorana, H. G.; Bahar, I.; Klein-Seetharaman, J., Identification of core amino acids stabilizing rhodopsin. *Proc Natl Acad Sci U S A* **2004**, 101, 7246-51.

280. Tan, H. P.; Rader, A. J., Identification of putative, stable binding regions through flexibility analysis of HIV-1 gp120. *Proteins* **2009**, 74, 881-894.

281. Berman, H., The protein data bank: A retrospective and prospective. *Biophys J* **2000**, 78, 267a-267a.

282. Brough, P. A.; Barril, X.; Borgognoni, J.; Chene, P.; Davies, N. G.; Davis, B.; Drysdale, M. J.; Dymock, B.; Eccles, S. A.; Garcia-Echeverria, C.; Fromont, C.; Hayes, A.; Hubbard, R. E.; Jordan, A. M.; Jensen, M. R.; Massey, A.; Merrett, A.; Padfield, A.; Parsons, R.; Radimerski, T.; Raynaud, F. I.; Robertson, A.; Roughley, S. D.; Schoepfer, J.; Simmonite, H.; Sharp, S. Y.; Surgenor, A.; Valenti, M.; Walls, S.; Webb, P.; Wood, M.; Workman, P.; Wright, L., Combining hit identification strategies: fragment-based and in silico approaches to orally active 2-aminothieno[2,3-d]pyrimidine inhibitors of the Hsp90 molecular chaperone. *J Med Chem* **2009**, 52, 4794-809.

283. Brough, P. A.; Aherne, W.; Barril, X.; Borgognoni, J.; Boxall, K.; Cansfield, J. E.; Cheung, K. M.; Collins, I.; Davies, N. G.; Drysdale, M. J.; Dymock, B.; Eccles, S. A.; Finch, H.; Fink, A.; Hayes, A.; Howes, R.; Hubbard, R. E.; James, K.; Jordan, A. M.; Lockie, A.; Martins, V.; Massey, A.; Matthews, T. P.; McDonald, E.; Northfield, C. J.; Pearl, L. H.; Prodromou, C.; Ray, S.; Raynaud, F. I.; Roughley, S. D.; Sharp, S. Y.; Surgenor, A.; Walmsley, D. L.; Webb, P.; Wood, M.; Workman, P.; Wright, L., 4,5-diarylisoxazole Hsp90 chaperone inhibitors: potential therapeutic agents for the treatment of cancer. *J Med Chem* **2008**, 51, 196-218.

284. Rowlands, M. G.; Newbatt, Y. M.; Prodromou, C.; Pearl, L. H.; Workman, P.; Aherne, W., High-throughput screening assay for inhibitors of heat-shock protein 90 ATPase activity. *Anal Biochem* **2004**, 327, 176-83.

285. Howes, R.; Barril, X.; Dymock, B. W.; Grant, K.; Northfield, C. J.; Robertson, A. G.; Surgenor, A.; Wayne, J.; Wright, L.; James, K.; Matthews, T.; Cheung, K. M.; McDonald, E.; Workman, P.; Drysdale, M. J., A fluorescence polarization assay for inhibitors of Hsp90. *Anal Biochem* **2006**, 350, 202-13.

286. Brough, P. A.; Barril, X.; Beswick, M.; Dymock, B. W.; Drysdale, M. J.; Wright, L.; Grant, K.; Massey, A.; Surgenor, A.; Workman, P., 3-(5-Chloro-2,4-dihydroxyphenyl)-pyrazole-4-carboxamides as inhibitors of the Hsp90 molecular chaperone. *Bioorg Med Chem Lett* **2005**, 15, 5197-201.

287. Wright, L.; Barril, X.; Dymock, B.; Sheridan, L.; Surgenor, A.; Beswick, M.; Drysdale, M.; Collier, A.; Massey, A.; Davies, N.; Fink, A.; Fromont, C.; Aherne, W.; Boxall, K.; Sharp, S.; Workman, P.; Hubbard, R. E., Structure-activity relationships in purine-based inhibitor binding to HSP90 isoforms. *Chem Biol* **2004**, 11, 775-85.

288. Burlingham, B. T.; Widlanski, T. S., An intuitive look at the relationship of Ki and IC50: A more general use for the Dixon plot. *J Chem Educ* **2003**, 80, 214-218.

289. Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R., PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **2015**, 31, 405-12.

290. Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S., The PDBbind database: methodologies and updates. *J Med Chem* **2005**, 48, 4111-9.

291. Wang, R.; Fang, X.; Lu, Y.; Wang, S., The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J Med Chem* **2004**, 47, 2977-80.

292. Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A., Binding MOAD (Mother Of All Databases). *Proteins* **2005**, 60, 333-40.

293. Ahmed, A.; Smith, R. D.; Clark, J. J.; Dunbar, J. B., Jr.; Carlson, H. A., Recent improvements to Binding MOAD: a resource for protein-ligand binding affinities and structures. *Nucleic Acids Res* **2015**, 43, D465-9.

294. Chen, X.; Lin, Y.; Gilson, M. K., The binding database: overview and user's guide. *Biopolymers* **2001**, 61, 127-41.

295. Chen, X.; Lin, Y.; Liu, M.; Gilson, M. K., The Binding Database: data management and interface design. *Bioinformatics* **2002**, 18, 130-9.

296. Chen, X.; Liu, M.; Gilson, M. K., BindingDB: a web-accessible molecular recognition database. *Comb Chem High Throughput Screen* **2001**, 4, 719-25.

297. Liu, T. Q.; Lin, Y. M.; Wen, X.; Jorissen, R. N.; Gilson, M. K., BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* **2007**, 35, D198-D201.

298. *Maestro, version 9.3, Schrödinger, LLC, New York, NY, 2009.*

299. Hofmann, T.; Hodges, R. S.; James, M. N., Effect of pH on the activities of penicillopepsin and Rhizopus pepsin and a proposal for the productive substrate binding mode in penicillopepsin. *Biochemistry US* **1984**, 23, 635-43.

300. Ferguson, D. M.; Radmer, R. J.; Kollman, P. A., Determination of the relative binding free energies of peptide inhibitors to the HIV-1 protease. *J Med Chem* **1991**, 34, 2654-9.

301. Chatfield, D. C.; Brooks, B. R., Hiv-1 Protease Cleavage Mechanism Elucidated with Molecular-Dynamics Simulation. *J Am Chem Soc* **1995**, 117, 5561-5572.

302. Greenwood, J. R.; Calkins, D.; Sullivan, A. P.; Shelley, J. C., Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *J Comput Aided Mol Des* **2010**, 24, 591-604.

303. Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M., Epik: a software program for pK(a) prediction and protonation state generation for drug-like molecules. *J Comput Aided Mol Des* **2007**, 21, 681-91.

304. Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C., Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol* **1999**, 285, 1735-1747.

305. Schuttelkopf, A. W.; van Aalten, D. M., PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr D Biol Crystallogr* **2004**, 60, 1355-63.

306. Marek, L.; Hamacher, A.; Hansen, F. K.; Kuna, K.; Gohlke, H.; Kassack, M. U.; Kurz, T., Histone deacetylase (HDAC) inhibitors with a novel connecting unit linker region reveal a selectivity profile for HDAC4 and HDAC5 with improved activity against chemoresistant cancer cells. *J Med Chem* **2013**, 56, 427-36.

307. Diedrich, D.; Hamacher, A.; Gertzen, C. G.; Alves Avelar, L. A.; Reiss, G. J.; Kurz, T.; Gohlke, H.; Kassack, M. U.; Hansen, F. K., Rational design and diversity-oriented synthesis of peptoid-based selective HDAC6 inhibitors. *Chem Commun* **2016**.

308. Hartigan, J. A., *Clustering algorithms*. New York, NY: John Wiley & Sons, Inc.: 1975.

309. Mcquitty, L. L., Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data. *Educ Psychol Meas* **1966**, 26, 825-&.

310. Mcquitty, L. L., Expansion of Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data. *Educ Psychol Meas* **1967**, 27, 253-&.

311. Team, R. C. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013*, 2014.

312. Allen, M. P.; Tildesley, D. J., *Computer simulation of liquids*. Oxford university press: 1989.

313. Darden, T.; York, D.; Pedersen, L., Particle Mesh Ewald - an N*Log(N) Method for Ewald Sums in Large Systems. *J Chem Phys* **1993**, 98, 10089-10092.

314. Verlet, L., Computer Experiments on Classical Fluids. I.Thermodynamical Properties of Lennard-Jones Molecules. *Phys Rev* **1967**, 159, 98-&.

315. Cleves, A. E.; Jain, A. N., Knowledge-guided docking: accurate prospective prediction of bound configurations of novel ligands using Surflex-Dock. *J Comput Aided Mol Des* **2015**, 29, 485-509.

316. Spitzer, R.; Jain, A. N., Surflex-Dock: Docking benchmarks and real-world application. *J Comput Aided Mol Des* **2012**, 26, 687-99.

317. Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D., Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, 57, 225-42.

318. Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M. *Amber 11*; University of California: 2010.

319. Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *J Comput Chem* **2005**, 26, 1668-1688.

320. Wang, J. M.; Cieplak, P.; Kollman, P. A., How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J Comput Chem* **2000**, 21, 1049-1074.

321. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, 65, 712-725.

322. Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J Comput Chem* **2004**, 25, 1157-1174.

323. Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A., Application of the Multimolecule and Multiconformational Resp Methodology to Biopolymers - Charge Derivation for DNA, RNA, and Proteins. *J Comput Chem* **1995**, 16, 1357-1377.

324. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C., Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J Comput Phys* **1977**, 23, 327-341.

325. Miyamoto, S.; Kollman, P. A., Settle - an Analytical Version of the Shake and Rattle Algorithm for Rigid Water Models. *J Comput Chem* **1992**, 13, 952-962.

326. Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R., Molecular-Dynamics with Coupling to an External Bath. *J Chem Phys* **1984**, 81, 3684-3690.

327. Brooijmans, N.; Sharp, K. A.; Kuntz, I. D., Stability of macromolecular complexes. *Proteins* **2002**, 48, 645-53.

328. Jorgensen, W. L., Quantum and Statistical Mechanical Studies of Liquids .24. Revised Tips for Simulations of Liquid Water and Aqueous-Solutions. *J Chem Phys* **1982**, 77, 4156-4163.

329. Gohlke, H.; Kiel, C.; Case, D. A., Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. *J Mol Biol* **2003**, 330, 891-913.

330. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995). *J Am Chem Soc* **1996**, 118, 2309-2309.

331. Homeyer, N.; Stoll, F.; Hillisch, A.; Gohlke, H., Binding Free Energy Calculations for Lead Optimization: Assessment of Their Accuracy in an Industrial Drug Design Context. *J Chem Theory Comput* **2014**, 10, 3331-3344.

332. Genheden, S.; Ryde, U., How to Obtain Statistically Converged MM/GBSA Results. *J Comput Chem* **2010**, 31, 837-846.

333. Lu, Q.; Luo, R., A Poisson-Boltzmann dynamics method with nonperiodic boundary condition. *J Chem Phys* **2003**, 119, 11035-11047.

334. Sitkoff, D.; Sharp, K. A.; Honig, B., Accurate Calculation of Hydration Free-Energies Using Macroscopic Solvent Models. *Journal of Physical Chemistry* **1994**, 98, 1978-1988.

335. Kopitz, H.; Cashman, D. A.; Pfeiffer-Marek, S.; Gohlke, H., Influence of the solvent representation on vibrational entropy calculations: generalized born versus distance-dependent dielectric model. *J Comput Chem* **2012**, 33, 1004-13.

336. Page, C. S.; Bates, P. A., Can MM-PBSA calculations predict the specificities of protein kinase inhibitors? *J Comput Chem* **2006**, 27, 1990-2007.

337. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T., Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J Am Chem Soc* **1990**, 112, 6127-6129.

338. Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G., Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *Journal of Physical Chemistry* **1996**, 100, 19824-19839.

339. Onufriev, A.; Case, D. A.; Bashford, D., Effective Born radii in the generalized Born approximation: The importance of being perfect. *J Comput Chem* **2002**, 23, 1297-1304.

340. Verkhivker, G.; Appelt, K.; Freer, S. T.; Villafranca, J. E., Empirical free energy calculations of ligand-protein crystallographic complexes. I. Knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Eng* **1995**, 8, 677-91.

341. Guan, R.; Ho, M. C.; Brenowitz, M.; Tyler, P. C.; Evans, G. B.; Almo, S. C.; Schramm, V. L., Entropy-driven binding of picomolar transition state analogue inhibitors to human 5'-methylthioadenosine phosphorylase. *Biochemistry US* **2011**, 50, 10408-17.

342. Yin, F.; Cao, R.; Goddard, A.; Zhang, Y.; Oldfield, E., Enthalpy versus entropy-driven binding of bisphosphonates to farnesyl diphosphate synthase. *J Am Chem Soc* **2006**, 128, 3524-5.

343. Rucker, C.; Rucker, G.; Meringer, M., y-Randomization and its variants in QSPR/QSAR. *J Chem Inf Model* **2007**, 47, 2345-2357.

344. Canty, A. J.; Davison, A. C.; Hinkley, D. V.; Ventura, V., Bootstrap diagnostics and remedies. *Can J Stat* **2006**, 34, 5-27.

345. Mikulskis, P.; Genheden, S.; Rydberg, P.; Sandberg, L.; Olsen, L.; Ryde, U., Binding affinities in the SAMPL3 trypsin and host-guest blind tests estimated with the MM/PBSA and LIE methods. *J Comput Aid Mol Des* **2012**, 26, 527-541.

346. Silverman, M. P.; Strange, W.; Lipscombe, T. C., The distribution of composite measurements: How to be certain of the uncertainties in what we measure. *Am J Phys* **2004**, 72, 1068-1081.

347. Holmes, D. T.; Buhr, K. A., Error propagation in calculated ratios. *Clin Biochem* **2007**, 40, 728-734.

348. Kar, P.; Lipowsky, R.; Knecht, V., Importance of Polar Solvation and Configurational Entropy for Design of Antiretroviral Drugs Targeting HIV-1 Protease. *J Phys Chem B* **2013**, 117, 5793-5805.

349. Ali, A.; Reddy, G. S.; Cao, H.; Anjum, S. G.; Nalam, M. N.; Schiffer, C. A.; Rana, T. M., Discovery of HIV-1 protease inhibitors with picomolar affinities incorporating N-aryl-oxazolidinone-5-carboxamides as novel P2 ligands. *J Med Chem* **2006**, 49, 7342-56.

350. Altman, M. D.; Ali, A.; Reddy, G. S.; Nalam, M. N.; Anjum, S. G.; Cao, H.; Chellappan, S.; Kairys, V.; Fernandes, M. X.; Gilson, M. K.; Schiffer, C. A.; Rana, T. M.; Tidor, B., HIV-1 protease inhibitors from inverse design in the substrate envelope exhibit subnanomolar binding to drug-resistant variants. *J Am Chem Soc* **2008**, 130, 6099-113.

351. Nalam, M. N.; Ali, A.; Altman, M. D.; Reddy, G. S.; Chellappan, S.; Kairys, V.; Ozen, A.; Cao, H.; Gilson, M. K.; Tidor, B.; Rana, T. M.; Schiffer, C. A., Evaluating the substrate-envelope hypothesis: structural analysis of novel HIV-1 protease inhibitors designed to be robust against drug resistance. *J Virol* **2010**, 84, 5368-78.

352. Ali, A.; Reddy, G. S.; Nalam, M. N.; Anjum, S. G.; Cao, H.; Schiffer, C. A.; Rana, T. M., Structure-based design, synthesis, and structure-activity relationship studies of HIV-1 protease inhibitors incorporating phenyloxazolidinones. *J Med Chem* **2010**, 53, 7699-708.

353. Reddy, G. S.; Ali, A.; Nalam, M. N.; Anjum, S. G.; Cao, H.; Nathans, R. S.; Schiffer, C. A.; Rana, T. M., Design and synthesis of HIV-1 protease inhibitors incorporating

oxazolidinones as P2/P2' ligands in pseudosymmetric dipeptide isosteres. *J Med Chem* **2007**, 50, 4316-28.

354. Maignan, S.; Guilloteau, J. P.; Pouzieux, S.; Choi-Sledeski, Y. M.; Becker, M. R.; Klein, S. I.; Ewing, W. R.; Pauls, H. W.; Spada, A. P.; Mikol, V., Crystal structures of human factor Xa complexed with potent inhibitors. *J Med Chem* **2000**, 43, 3226-32.

355. Guertin, K. R.; Gardner, C. J.; Klein, S. I.; Zulli, A. L.; Czekaj, M.; Gong, Y.; Spada, A. P.; Cheney, D. L.; Maignan, S.; Guilloteau, J. P.; Brown, K. D.; Colussi, D. J.; Chu, V.; Heran, C. L.; Morgan, S. R.; Bentley, R. G.; Dunwiddie, C. T.; Leadley, R. J.; Pauls, H. W., Optimization of the beta-aminoester class of factor Xa inhibitors. Part 2: Identification of FXV673 as a potent and selective inhibitor with excellent In vivo anticoagulant activity. *Bioorg Med Chem Lett* **2002**, 12, 1671-4.

356. Senger, S.; Convery, M. A.; Chan, C. E.; Watson, N. S., Arylsulfonamides: A study of the relationship between activity and conformational preferences for a series of factor Xa inhibitors. *Bioorg Med Chem Lett* **2006**, 16, 5731-5735.

357. Chan, C.; Borthwick, A. D.; Brown, D.; Burns-Kurtis, C. L.; Campbell, M.; Chaudry, L.; Chung, C. W.; Convery, M. A.; Hamblin, J. N.; Johnstone, L.; Kelly, H. A.; Kleanthous, S.; Patikis, A.; Patel, C.; Pateman, A. J.; Senger, S.; Shah, G. P.; Toomey, J. R.; Watson, N. S.; Weston, H. E.; Whitworth, C.; Young, R. J.; Zhou, P., Factor Xa inhibitors: S1 binding interactions of a series of N-{(3S)-1-[(1S)-1-methyl-2-morpholin-4-yl-2-oxoethyl]-2-oxopyrrolidin-3-yl}sulfonamides. *J Med Chem* **2007**, 50, 1546-1557.

358. Young, R. J.; Brown, D.; Burns-Kurtis, C. L.; Chan, C.; Convery, M. A.; Hubbard, J. A.; Kelly, H. A.; Pateman, A. J.; Patikis, A.; Senger, S.; Shah, G. P.; Toomey, J. R.; Watson, N. S.; Zhou, P., Selective and dual action orally active inhibitors of thrombin and factor Xa. *Bioorg Med Chem Lett* **2007**, 17, 2927-30.

359. Young, R. J.; Borthwick, A. D.; Brown, D.; Burns-Kurtis, C. L.; Campbell, M.; Chan, C.; Charbaut, M.; Convery, M. A.; Diallo, H.; Hortense, E.; Irving, W. R.; Kelly, H. A.; King, N. P.; Kleanthous, S.; Mason, A. M.; Pateman, A. J.; Patikis, A. N.; Pinto, I. L.; Pollard, D. R.; Senger, S.; Shah, G. P.; Toomey, J. R.; Watson, N. S.; Weston, H. E.; Zhou, P., Structure and property based design of factor Xa inhibitors: biaryl pyrrolidin-2-ones incorporating basic heterocyclic motifs. *Bioorg Med Chem Lett* **2008**, 18, 28-33.

360. Maignan, S.; Guilloteau, J. P.; Choi-Sledeski, Y. M.; Becker, M. R.; Ewing, W. R.; Pauls, H. W.; Spada, A. P.; Mikol, V., Molecular structures of human factor Xa complexed

with ketopiperazine inhibitors: preference for a neutral group in the S1 pocket. *J Med Chem* **2003**, 46, 685-90.

361. Matter, H.; Defossa, E.; Heinelt, U.; Blohm, P. M.; Schneider, D.; Muller, A.; Herok, S.; Schreuder, H.; Liesum, A.; Brachvogel, V.; Lonze, P.; Walser, A.; Al-Obeidi, F.; Wildgoose, P., Design and quantitative structure-activity relationship of 3-amidinobenzyl-1H-indole-2-carboxamides as potent, nonchiral, and selective inhibitors of blood coagulation factor Xa. *J Med Chem* **2002**, 45, 2749-69.

362. Nazare, M.; Will, D. W.; Matter, H.; Schreuder, H.; Ritter, K.; Urmann, M.; Essrich, M.; Bauer, A.; Wagner, M.; Czech, J.; Lorenz, M.; Laux, V.; Wehner, V., Probing the subpockets of factor Xa reveals two binding modes for inhibitors based on a 2-carboxyindole scaffold: a study combining structure-activity relationship and X-ray crystallography. *J Med Chem* **2005**, 48, 4511-25.

363. Barril, X.; Brough, P.; Drysdale, M.; Hubbard, R. E.; Massey, A.; Surgenor, A.; Wright, L., Structure-based discovery of a new class of Hsp90 inhibitors. *Bioorg Med Chem Lett* **2005**, 15, 5187-91.

364. Sharp, S. Y.; Prodromou, C.; Boxall, K.; Powers, M. V.; Holmes, J. L.; Box, G.; Matthews, T. P.; Cheung, K. M.; Kalusa, A.; James, K.; Hayes, A.; Hardcastle, A.; Dymock, B.; Brough, P. A.; Barril, X.; Cansfield, J. E.; Wright, L.; Surgenor, A.; Foloppe, N.; Hubbard, R. E.; Aherne, W.; Pearl, L.; Jones, K.; McDonald, E.; Raynaud, F.; Eccles, S.; Drysdale, M.; Workman, P., Inhibition of the heat shock protein 90 molecular chaperone in vitro and in vivo by novel, synthetic, potent resorcinylic pyrazole/isoxazole amide analogues. *Mol Cancer Ther* **2007**, 6, 1198-211.

365. Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., Assessing scoring functions for protein-ligand interactions. *J Med Chem* **2004**, 47, 3032-3047.

366. Kim, R.; Skolnick, J., Assessment of programs for ligand binding affinity prediction. *J Comput Chem* **2008**, 29, 1316-31.

367. Gao, C.; Park, M. S.; Stern, H. A., Accounting for ligand conformational restriction in calculations of protein-ligand binding affinities. *Biophys J* **2010**, 98, 901-10.

368. Hopkins, A. L.; Groom, C. R.; Alex, A., Ligand efficiency: a useful metric for lead selection. *Drug Discov Today* **2004**, 9, 430-1.

369. Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D., Improved protein-ligand docking using GOLD. *Proteins Struct Funct Genet* **2003**, 52, 609-623.

370. Cole, J. C.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R., Comparing protein-ligand docking programs is difficult. *Proteins* **2005**, 60, 325-332.

371. Radestock, S.; Bohm, M.; Gohlke, H., Improving binding mode predictions by docking into protein-specifically adapted potential fields. *J Med Chem* **2005**, 48, 5466-5479.

372. McInnes, C., Virtual screening strategies in drug discovery. *Curr Opin Chem Biol* **2007**, 11, 494-502.

373. Liang, J.; Edelsbrunner, H.; Woodward, C., Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Sci* **1998**, 7, 1884-1897.

374. Cai, Y.; Myint, W.; Paulsen, J. L.; Schiffer, C. A.; Ishima, R.; Kurt Yilmaz, N., Drug Resistance Mutations Alter Dynamics of Inhibitor-Bound HIV-1 Protease. *J Chem Theory Comput* **2014**, 10, 3438-3448.

375. Shi, S. H.; Hu, G. D.; Zhang, X. M.; Wang, J. H., A study of the interaction between HIV-1 protease and C-2-symmetric inhibitors by computational methods. *J Mol Model* **2014**, 20.

376. Abdel-Azeim, S.; Oliva, R.; Chermak, E.; De Cristofaro, R.; Cavallo, L., Molecular dynamics characterization of five pathogenic Factor X mutants associated with decreased catalytic activity. *Biochemistry US* **2014**, 53, 6992-7001.

377. Teo, R. D.; Dong, S. S.; Gross, Z.; Gray, H. B.; Goddard, W. A., Computational predictions of corroles as a class of Hsp90 inhibitors. *Mol Biosyst* **2015**, 11, 2907-2914.

378. Yang, T. Y.; Wu, J. C.; Yan, C. L.; Wang, Y. F.; Luo, R.; Gonzales, M. B.; Dalby, K. N.; Ren, P. Y., Virtual screening using molecular simulations. *Proteins* **2011**, 79, 1940-1951.

379. Genheden, S.; Ryde, U., Comparison of end-point continuum-solvation methods for the calculation of protein-ligand binding free energies. *Proteins* **2012**, 80, 1326-42.

380. Genheden, S.; Nilsson, I.; Ryde, U., Binding affinities of factor Xa inhibitors estimated by thermodynamic integration and MM/GBSA. *J Chem Inf Model* **2011**, 51, 947-58.

381. Steinbrecher, T. B.; Dahlgren, M.; Cappel, D.; Lin, T.; Wang, L.; Krilov, G.; Abel, R.; Friesner, R.; Sherman, W., Accurate Binding Free Energy Predictions in Fragment Optimization. *J Chem Inf Model* **2015**.

382. Erickson, H. P., Co-Operativity in Protein-Protein Association - the Structure and Stability of the Actin Filament. *J Mol Biol* **1989**, 206, 465-474.

383. Horton, N.; Lewis, M., Calculation of the free energy of association for protein complexes. *Protein Sci* **1992**, 1, 169-81.

384. Hecht, P., The developing practice of comparative molecular field analysis. *3D QSAR in Drug Design: Volume 1: Theory Methods and Applications* **1993**, 1, 443.

385. Golbraikh, A.; Tropsha, A., Beware of q2! *J Mol Graph Model* **2002**, 20, 269-76.

386. Roy, K., On some aspects of validation of predictive quantitative structure-activity relationship models. *Expert Opin Drug Dis* **2007**, 2, 1567-1577.

387. Hawkins, D. M.; Basak, S. C.; Mills, D., Assessing model fit by cross-validation. *J Chem Inf Comp Sci* **2003**, 43, 579-586.

388. Wang, R. X.; Lu, Y. P.; Wang, S. M., Comparative evaluation of 11 scoring functions for molecular docking. *J Med Chem* **2003**, 46, 2287-2303.

389. Wang, W. J.; Huang, Q.; Zou, J.; Li, L. L.; Yang, S. Y., TS-Chemscore, a Target-Specific Scoring Function, Significantly Improves the Performance of Scoring in Virtual Screening. *Chem Biol Drug Des* **2015**, 86, 781-788.

390. Massova, I.; Kollman, P. A., Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies. *J Am Chem Soc* **1999**, 121, 8133-8143.

391. Kortemme, T.; Kim, D. E.; Baker, D., Computational alanine scanning of protein-protein interfaces. *Sci STKE* **2004**, 2004, pl2.

392. Wichmann, C.; Becker, Y.; Chen-Wichmann, L.; Vogel, V.; Vojtkova, A.; Herglotz, J.; Moore, S.; Koch, J.; Lausen, J.; Mantele, W.; Gohlke, H.; Grez, M., Dimer-tetramer transition controls RUNX1/ETO leukemogenic activity. *Blood* **2010**, 116, 603-13.

393. Fischer, S.; Smith, J. C.; Verma, C. S., Dissecting the vibrational entropy change on protein/ligand binding: Burial of a water molecule in bovine pancreatic trypsin inhibitor. *J Phys Chem B* **2001**, 105, 8050-8055.

394. Moorman, V. R.; Valentine, K. G.; Wand, A. J., The dynamical response of hen egg white lysozyme to the binding of a carbohydrate ligand. *Protein Sci* **2012**, 21, 1066-73.

395. Kuhn, B.; Kollman, P. A., Binding of a diverse set of ligands to avidin and streptavidin: an accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models. *J Med Chem* **2000**, 43, 3786-91.

396. Matter, H.; Defossa, E.; Heinelt, U.; Blohm, P. M.; Schneider, D.; Muller, A.; Herok, S.; Schreuder, H.; Liesum, A.; Brachvogel, V.; Lonze, P.; Walser, A.; Al-Obeidi, F.;

Wildgoose, P., Design and quantitative structure-activity relationship of 3-amidinobenzyl-1H-indole-2-carboxamides as potent, nonchiral, and selective inhibitors of blood coagulation factor Xa. *J Med Chem* **2002**, 45, 2749-2769.

397. Maignan, S.; Guilloteau, J. P.; Choi-Sledski, Y. M.; Becker, M. R.; Ewing, W. R.; Pauls, H. W.; Spada, A. P.; Mikol, V., Molecular structures of human factor Xa complexed with ketopiperazine inhibitors: Preference for a neutral group in the S1 pocket. *J Med Chem* **2003**, 46, 685-690.

398. Watson, N. S.; Brown, D.; Campbell, M.; Chan, C.; Chaudry, L.; Convery, M. A.; Fenwick, R.; Hamblin, J. N.; Haslam, C.; Kelly, H. A.; King, N. P.; Kurtis, C. L.; Leach, A. R.; Manchee, G. R.; Mason, A. M.; Mitchell, C.; Patel, C.; Patel, V. K.; Senger, S.; Shah, G. P.; Weston, H. E.; Whitworth, C.; Young, R. J., Design and synthesis of orally active pyrrolidin-2-one-based factor Xa inhibitors. *Bioorg Med Chem Lett* **2006**, 16, 3784-8.