Applications of integral equation theory to biological systems

Dissertation zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. Nat.)

Die Dissertation wurde im Zeitraum vom 10.2016 bis zum 10.2021 an der Fakultät für Chemie und Chemische Biologie der Technischen Universität Dortmund angefertigt

M. Sc. Yannic Alber

Bergisch Gladbach, 2021

Erstgutachter:	Prof. Dr. Stefan M. Kast
Zweitgutachter:	Prof. Dr. Paul Czodrowski

All models are wrong, some are useful.

George Box

Danksagung

Zunächst möchte ich mich bei all denen bedanken, die mich während der Entstehung dieser Dissertation durch ihre fachliche und persönliche Unterstützung begleitet und motiviert haben.

Mein größter Dank gebührt Herrn Prof. Dr. Stefan M. Kast für seine umfangreiche Betreuung und Unterstützung bei der Umsetzung der gesamten Arbeit. Die vielen Gespräche auf intellektueller und persönlicher Ebene bleiben mir immer als bereichernder und konstruktiver Ausstausch in Erinnerung.

Auch möchte ich Herrn Prof. Dr. Paul Czodrowski meinen Dank für die Übernahme des Zweitgutachtens aussprechen.

Der gesamten Arbeitsgruppe möchte ich für die gute Zusammenarbeit und das angenehme Arbeitsklima danken.

Insbesondere möchte ich mich bei Florian Mrugalla und Leonhard Henkes für ihre Unterstützung, den vielen wertvollen Anregungen und produktiven Gesprächen bedanken, die wesentlich zum Gelingen dieser Arbeit beigetragen haben. Ohne ihre Freundschaft wäre dieser Weg ungleich schwerer gewesen.

Bei dieser Gelegenheit sei zudem Herrn Dr. Thomas Mrziglod sowie Herrn Dr. Georg Mogk für ihre Unterstützung gedankt. Außerdem möchte ich meinen Kollegen der Angewandten Mathematik für ihre offenen Ohren und ihren Zuspruch danken.

Ein herzliches Dankeschön gilt meiner Familie, insbesondere meinen Eltern und meiner Schwester, für ihre stetige Motivation sowie Ermutigungen während meines gesamten Studiums.

Mein größter Dank geht abschließend an meine Tochter Katharina und meine Frau Andrea, die mir während der ganzen Promotion Kraft, Trost und Geborgenheit schenkten. Ohne ihre nicht endende Geduld und Unterstützung wäre mir dieser Weg nicht möglich gewesen.

Diese Arbeit entstand zwischen 12.2016 und 10.2021 unter der Leitung von Herrn Prof. Dr. Stefan M. Kast. Sie wurde von M. Sc. Yannic Alber in Fakultät für Chemie und chemische Biologie der Technischen Universität Dortmund im Arbeitskries für theoretische physikalische Chemie angefertigt.

Die hier vorgestellten Ergebnisse stammen vom Autor. An den folgenden Stellen wurde auf die Arbeit anderer zurückgegriffen.

- Die konzeptionelle Erfindung der „Localized Free Energies" (LFE) und „Free Energy Derivatives" (FED) geht auf Prof. Dr. Stefan M. Kast zurück.
- Teile der in 3.1 vorgestellten Ergebnisse sind bereits in Blech, M., Hörer, S., Kuhn, A. B., Kube, S., Göddeke, H., Kiefer, H., Zang, Y., Alber, Y., Kast, S. M., Westermann, M., Tully, M. D., Schäfer L. V. & Garidel, P., *Biophysical Journal* **116,** 1637–1649; (2019) veröffentlicht.
- Die in 3.2 verwendeten Konformationen wurden von Dr. Nicolas Tielker erstellt und sind in Tielker, N., Tomazic, D., Heil, J., Kloss, T., Ehrhart, S., Güssregen, S., Schmidt, F. & Kast, S. M. *J Comput Aided Mol Des* **30**, 1035–1044 (2016) veröffentlicht.
- Die in 3.2 verwendeten EC-RISM Partialladungen wurden von Dr. Nicolas Tielker erstellt und sind in Tielker, N., Eberlein, L., Güssregen, S. & Kast, S. M. *J Comput Aided Mol Des* **32**, 1151–1163 (2018) veröffentlicht.

Zusammenfassung

Das „three-dimensional reference interaction site model“ (3D RISM) erlaubt es die Solvensverteilung, und somit die damit verbundenen thermodynamischen Eigenschaften, um ein gegebenes Solvat zu berechnen. Dies kann ein kleines, wirkstoffartiges Molekül sein oder ein Protein mit tausenden Atomen. Zusammen mit Methoden, wie Molekulardynamik-(MD) Simulationen und Kraftfeldern, ist es möglich, die Unterschiede in der freien Energie zwischen Konformeren, Molekülen und Komplexen in biologisch relevanten Systemen zu bestimmen.

In dieser Arbeit werden durch Kombination von 3D RISM und MD Simulationen freie Energiedifferenzen zwischen zwei Konformeren eines Antikörpers berechnet und durch Tests mit verschiedenen Wassermodellen und Fehlerkorrekturen validiert. Allerdings entstehen durch starke strukturelle Fluktuationen während der Simulationen häufig große statistische Fehler, was die Anwendungsfelder solcher Methoden limitiert.

Um das Problem abzuschwächen und um auf explizite Simulationen verzichten zu können, werden sogenannte „Localized Free Energies” (LFE) verwendet. Mit ihnen ist es möglich, die freie Energie auf ein atomweises Niveau herunter zu brechen, wo angenommen werden kann, dass besagte Fluktuationen einen geringeren Einfluss haben. Da eine solche Partitionierung rein virtuell ist, gibt es keinen experimentellen Weg, die LFEs zu validieren. Aus diesem Grund wird ihre Plausibilität durch Anwendung als Eingabeinformation für Methoden des maschinellen Lernens (ML) überprüft, indem der Verlust ihrer Vorhersagekraft durch ansteigende Störung der LFEs beobachtet wird.

Mit bestätigter Plausibilität werden die LFEs beispielhaft auf eine Serie von Thrombin-Inhibitoren angwendet, um ihr Potential in der Medikamentenentwicklung zu zeigen. Darüberhinaus wird der Einfluss von experimentellen Unsicherheiten in den Kristallstrukturen sowie die Limitationen des Ansatzes selbst untersucht.

Von der gleichen formalen Basis, wie sie auch bei den LFEs genutzt wurde, lassen sich auch die so genannten „Free Energy Derivatives” (FED) sehr effizient bestimmen. Diese beschreiben auf atomarer Ebene, wie sich die freie Energie in Abhängigkeit von Kraftfeldparametern verändert. Die LFEs werden ebenfalls anhand eines Thrombin Komplexes näher beleuchtet und ihr prädiktiver Einsatz wird anhand eines auf Literaturdaten basierenden in-silico Experiments demonstriert.

Abstract

The three-dimensional reference interaction site model (3D RISM) allows to compute the solvent distribution, and therefore the associated thermodynamic properties, around a given solute. This can be a small, drug-like molecule or a protein with several thousand atoms. Combined with other tools like molecular dynamics (MD) simulations and force fields, it is possible to study the differences in free energy of conformations, molecules, and complexes in biological relevant systems.

By combining 3D RISM with MD simulations, the free energy difference between two structural conformers of an antibody is calculated, and the results are verified by tests with different water models and error corrections. However, due to strong structural fluctuation during the simulations, the statistical errors are often high, which limits the field of applications of such studies.

To alleviate this problem and to be able to do without explicit simulations, so so-called localized free energies (LFE) are employed. With them it is possible to break down free energies to an atom-wise level, where said fluctuations can be assumed to have less influence on the results. Since such a partitioning is purely virtual, there is no experimental way to validate the LFEs. For this reason, their plausibility is checked by using them as input for machine learning (ML) models, analyzing the drop in predictive power upon increasing levels of perturbation in the LFE input.

With the plausibility of the method established, the LFEs are applied to an exemplary series of thrombin inhibitors to illustrate their potential in a drug discovery context. Here they are used to identify the most relevant interactions between host and guest. Furthermore, the influence of experimental uncertainties in crystal structures and the limitations of the approach get explored.

Coming from the same formal basis as it was used for the LFEs, it is possible to calculate so-called free energy derivatives (FED) very efficiently. They describe how the free energy changes with respect to the non-bonded force field parameters on an atomistic level. The FEDs are also applied to thrombin complex, exploring the capabilities of the approach and investigating the predictive applicability of the FEDs by performing an in-silico experiment on literature data.

1 Introduction

1.1 Motivation

The development of a genuinely new drug has multiple phases in its research or preclinical stage, which can be loosely categorized into three tasks: target identification, lead discovery, and lead optimization. It is a loose categorization since all three tasks have some overlap to each other, and findings made in one step can influence the other two. So are insights made during the target identification step, primarily searching for key biological functions and molecules to address specific diseases, also informing the search for an optimal compound for them. Additionally, the border between lead discovery and optimization is fluent since both tasks require detailed knowledge about the biological interactions of the compounds in question. Such interactions may be the activation of an enzyme[1] or triggering a signal cascade, leading to the preprogrammed cell death called apoptosis[2].

Before a candidate molecule can be structurally improved towards optimal activity, absorption, distribution, metabolism, excretion, and toxicology (ADMET)[3], it must first be found, which is in itself a very cost and time-intensive undertaking. Potential candidates can be drawn from multiple sources like already published literature on the topic, known active natural products, off-target experience[4] or, since the late 20th century, from so-called high throughput screening (HTS)[5,6]. Substantial progress in automatization and assay development has made it possible to screen vast libraries of compounds against intended targets, and with increased efficiency also against several off-targets, tackling fundamental toxicologic issues right from the beginning of the drug development process[7]. Despite numbers in the order of 100 thousand tested compounds per day for simple assays, the available chemical space and libraries are still too big for a complete screening in most cases and need to be narrowed down. This is the domain of virtual screening, which uses algorithms based on the idea of a structure-activity relation (SAR) to enrich a set of screening candidates with potential hit molecules in silico. Methods like quantitative structure-activity relation (QSAR) models try to connect structural elements of a set of molecules with their physicochemical and biological properties. Since these models are usually comparably simple and therefore fast to evaluate, they are primarily used to filter down large datasets[4].

More elaborated, but also costly, approaches use a combination of docking (placement of compounds in binding sites) and scoring (evaluating placements and compounds) as well as 3D QSAR (screening methods utilizing 3D informed pharmacophores) where molecules are placed in three-dimensional space into an active binding site of a target protein and are then rated regarding their fitness[8–10]. Experimental and virtual screenings are often performed in an alternating fashion, each informing and optimizing the other. Once a set of suitable candidates is found, the lead optimization can start.

It is often performed over several iterations of a cyclic workflow, employing methods and knowledge from many different disciplines. Here an initial model of the interaction of the candidate molecule and the previously identified target leads to a first hypothesis regarding potentially beneficial structural elements and corresponding molecules. Those are then synthesized and experimentally tested, which results are then used to refine the model of interaction[4]. This very general approach is not new to pharmaceutical research and was already used in the late 19th century leading to drugs like chloral hydrate[11,12] and acetylsalicylic acid[13,14], even though the assumptions about the mode of action were not always correct. Nonetheless, the basic idea of formulating a relation between structure and activity kept the same and is still the basis for the modern structure-based drug design (SBDD)[3].

The necessary knowledge to draw such a connection is gained from targeted compound synthesis and screening of preselected molecules. Typical elements to look for are hydrogen bond donors and acceptors, halogenic and other decorations of aromatic rings, charged groups, heteroatoms and cycles, and steric fragments filling voids and thereby influencing the water network in the binding site or stabilizing specific conformations.[3] The patterns arising from the analysis of experimental findings like binding affinities and ADMET parameters are used to modify the candidate molecules such that they are believed to perform better as their parent generation did before. At the end of the last century, this process was, and in parts still is, characterized by chemical intuition, a few rules, like the famous Lipinski rule of five[15,16], and also not the least by some serendipity.[4] Obviously, this is not a reliable nor efficient way, so methods from computational chemistry become more and more widespread in the workflow, aiding the design and test process and creating the field of computer-aided drug design (CADD)[17]. Methods like quantum mechanics (QM), giving insights in extreme detail and accuracy, and molecular mechanics (MM), making even large proteins and short genes accessible for theoretical descriptions, became, among others, essential in modern drug development processes.[18,19]

Especially molecular dynamics (MD) simulations have seen a tremendous boost in recent years by being implemented on powerful computation hardware like graphics processing units (GPUs)[20]. With them, it is possible to model the behavior of target molecules and drug candidates and their interactions on an ever-growing scale. With advanced simulation protocols like free energy perturbation[21,22] (FEP) and thermodynamic integration[22,23] (TI), it is possible to calculate the binding free energy of a compound upon binding a specific target. In general, there are two methods currently used for that task; a physical transformation where the binding process is directly simulated in a physically possible way and an alchemical transformation where a physically impossible process is calculated.[24–26,26]. An example for an application of such simulations can be found in the publication of Plenker et al.[27] where the binding affinity of two compounds to the RET kinase is investigated via TI simulations. Another in the context of this work important application of free energy simulations is the localization of solvation and binding free energies onto individual atoms of molecules. Like it was done in the publication of Irwin et al.[28]. Here the localization was performed on the basis of intricate FEP simulations, modeling the protein-ligand system on a high level, including conformational variations and entropic contributions. This is a clear advantage over the in this work used rigid-body approximation, which is discussed in more detail later. Unfortunately, such simulation-based methods are relatively expensive and require substantial compute power to achieve reasonable accuracy. This limits their field of application to a small set of molecules and forces the practitioner to choose between high accuracy or high throughput.[29]

Another popular and computationally efficient method is the three-dimensional reference interaction site model (3D RISM)[30–33]. With the significantly increased cost-effective compute power of the last two decades[34], RISM based approaches became feasible for usage in the drug discovery process. By coupling 3D RISM with MD simulations[35,36], the conformational stability of proteins can be estimated, and the binding free energies of protein-ligand complexes be inferred[36–39]. 3D RISM can also be used to investigate the protein-ligand interactions from a solvation perspective in great detail. One of the more obvious fields of application in this regard for a granular solvent model is the elucidation of the water network involved in the binding process [40–45]. Here, several methods for extracting and characterization of water sites from the calculated solvent site densities were introduced in the past. One of the most straightforward ways to find discrete water sites is via an analysis of the maxima and minima of the continuous solvent distribution obtained by 3D RISM. This approach was used in various studies, such as those of Hirano et al.[43] and Güssregen et al.[44], and is the core mechanic in the Placevent algorithm by Sindhikara et al.[45,46]. The GAsol algorithm by Fusani et al.[41] follows the same principle to generate initial hydration sites but solves the issue of competing maxima in close proximity to eachother with a genetic algorithm. Generally speaking, the accuracy in finding crystallographic resolved water molecules of such 3D RISM based methods is on par with molecular dynamic simulation-based methods[47–50] while having the advantage of usually requiring fewer computing resources. Furthermore, they also generate reliable results for buried or occluded binding sites where an analysis derived from explicitly simulated water suffers from statistical uncertainties[49]. The placement and thermodynamic characterization of individual water sites can be used to inform SAR studies and to manipulate the binding relevant water network by specifically targeted compound modifications[44,47,48,51].

Additionally, fast and reliable 3D RISM based water placement methods can also be used to fit ligands in binding sites algorithmically, so-call docking protocols. So could Huang et al.[52] and Hinge et al.[53] show that water site informed docking can lead to reliable poses in complex binding situations. Another ansatz to the same problem, followed for example by Imai et al.[54], Nikolic et al. with 3D-RISM-Dock[55], and recently Sugita et al.[56], is to use a ligand-solvent mixture in 3D RISM to identify regions of high ligand or ligand-fragment density and orientate the compounds accordingly. Due to convergence issues in the 1D RISM solvent calculation, those approaches are limited to small ligands. A way to circumvent this problem is the solute-solute, or *uu*-3D-RISM[57] extension. Here both binding partners, protein and ligand, are treated as solutes, allowing for a more accurate physical description and following interesting applications in structure-based drug design.[58,59] An interesting example for this was given in a publication by Mrugalla et al.[59] and the thesis of Mrugalla[60] where free energy derivatives (FED), based on the potential of mean force (PMF), gets explored in complex formation scenarios. The concept of these FEDs is also an important part of this work. Here an alternative way of calculating them is employed, and its suitability towards drug design gets tested on protein-ligand complexes.

Further applications of RISM like the embedded cluster extension EC-RISM[61] can be used to calculate molecular properties of small molecules, e.g., solvation free energy, acidity, or water-octanol coefficient[62,63]. Such RISM based approaches often stand between expensive but comparably accurate methods, like explicit simulation protocols as full atom FEP and TI, and the much faster but more error-prone methods like implicit solvent simulations such as MM/PBSA[64] and MM/GBSA[65]. Although they also suffer from approximations and intrinsic errors like the overestimation of the solvent- cavity formation[66,67], much of these can be corrected, in parts even with very little additional compute costs[66,68–70]. However, such methods are more than mere compromises between speed and accuracy regarding the targeted task, as they also generate valuable insights into the role of the solvent and thereby deepen the understanding of the involved physical processes. This may not be unique to RISM and can also be achieved by other methods, but often at much higher costs. Exploring this intrinsic potential is a key focus of this work.

One way of doing so are methods from the field of machine learning, and in particular the subfield of deep learning. While by far not new to the field of drug discovery[71,72], they found extreme, even hype-like interest in recent years[73–75]. Generative machine learning methods are increasingly used to explore the vastness of the chemical space efficiently, and the novel approaches to the problem of de novo design are opening new regions in this space.[3,76] The high speed and accuracy, sometimes even comparable to physics-based methods, of deep learning models make them promising tools for predicting protein-ligand affinity and molecular properties.[77,78] The strengths of these models lay in their directly training on experimental data and thereby, in theory, implicitly capturing all necessary properties of the underlying problem. Specifying those explicitly is, in the best case, cumbersome and, in the worst case, impossible to do when they are not fully known. Nevertheless, this key advantage also brings two major downsides. The first is that predictions made by deep learning models are, with a few exceptions, cumbersome to explain, and even when they are, there is only seldom a chance to learn about the problem itself from them. The second problem, even more severe, is the very basis of the training process, the experimental data itself. Deep learning models usually require many samples for training and testing, and a notorious lack of coherent data leads to datasets that are often compiled from a multitude of sources over different labs, scientists, and time, introducing various uncertainties.[79] This is especially true for physical-chemical properties and biological assays, which have a priori a high experimental uncertainty, as they are often measured indirectly or are influenced by hard-to-control variances. This circumstance limits the generalizability and applicability of deep learning to regions of the chemical space where enough data exists. Currently, this restriction cannot be overcome by architectural modeling, and it is doubtful that this is conceptionally possible, as every model is limited by the amount of information the training data holds. The approach to alleviating the problem taken in this work uses physical knowledge in a hybrid modeling fashion, increasing robustness, accuracy, and generalizability.

1.2 Aim of this work

This work focuses primarily on the application of 3D RISM[30–33] based methods to thermodynamically characterize biological relevant systems. Those range from small drug-like molecules in solution, over protein-ligand complexes, to full conformationally flexible antibodies.

By combining 3D RISM with MD simulations, conformational changes in the anti-NPRA IgG4 antibody are modeled and studied, which would only be possible with an extreme expenditure of resources in a pure simulation approach. The learned lessons, especially regarding the limitations of such a combination, lead to other methods, focusing on the localization of free energies and their derivatives with respect to force field parameters.

Especially the concept of localized free energies (LFE), localizing the excess chemical potential μ^{ex} on individual solute sites, is investigated in detail. Due to the novelty and the fact that such a localization cannot be measured by physical means prompts the question of how reliable such a separation is, and therefore demonstrating the validity of the LFEs is the first goal of this work. This question is here answered in two ways. The first is the mathematical derivation and description of the method itself, demonstrating a sound mathematical foundation based on established physical concepts. The second is the utilization of LFEs as model input in two different deep learning approaches, predicting solvation free energies of small molecules. By increasing the level of perturbation of the localization, the plausibility of the by the LFE method calculated partitioning is verified.

One can envision multiple-use cases for the LFE and the free energy derivatives (FED) method, but the major one discussed here is their application in drug discovery and optimization. As a demonstration, they are evaluated on a ligand series of thrombin inhibitors[80], elucidating the protein-ligand interaction from multiple perspectives, opening insights useful for lead discovery, and exploring the potential of the FEDs for lead optimization. The added dimensionality of LFEs and FEDs introduces some challenges regarding a fast and intuitively comprehendible visualization. This, however, is crucial for a successful implementation of the introduced methods in the drug discovery workflow. To make the results more accessible and to move towards this goal, this work also makes suggestions for suitable visualizations of the results.

2 Theory

2.1 Reference interaction site model

The reference interaction site model[81–83] (RISM) is grounded in the classical density functional theory[84], which itself is based on the idea to describe the properties of liquid matter in terms of individual particle density, often referred to as site-density, distributions. The grand canonical potential

$$\Omega = A - \mu N = A[\rho(\mathbf{r})] - \mu \int \rho(\mathbf{r}) d\mathbf{r} \tag{2.1}$$

with the free energy A as functional of the local density $\rho(\mathbf{r})$, the chemical potential μ, and the total number of particles N reflects this idea in terms of statistical thermodynamics. Following this line of reasoning, a specific density distribution ρ_{eq} can be defined, at which the grand canonical potential has its minimum and

$$\left.\frac{\partial \Omega}{\partial \rho(\mathbf{r})}\right|_{(\rho=\rho_{\mathrm{eq}})} = 0 \tag{2.2}$$

is satisfied. Both equations (2.1) and (2.2) are fundamental to classical density functional theory and are the starting point from which its essential equations are derived. This, however, is beyond the scope of this work and the reader may be referred to the very comprehensive works of Hansen and McDonnald[84], Hirata[85], and Montroll and Lebowitz[86] for a more detailed mathematical explanation.

Nonetheless, it is evident from the equations above that the density function is central for any such derived theory. The function itself can be expressed as the probability $p^{(2)}(\mathbf{r}, \mathbf{r}')$ to find a pair of two particles (denoted from here on as (2)) in a homogeneous and isotropic fluid of N particles through:

$$\rho^{(2)}(\mathbf{r}, \mathbf{r}') = N(N-1)p^{(2)}(\mathbf{r}, \mathbf{r}'). \tag{2.3}$$

By considering the distribution of one particle only relative to the position of the other one and normalizing it by the bulk density ρ, the pair distribution function

$$g(\mathbf{r}) = \frac{\rho(\mathbf{r})}{\rho} \tag{2.4}$$

can be derived.

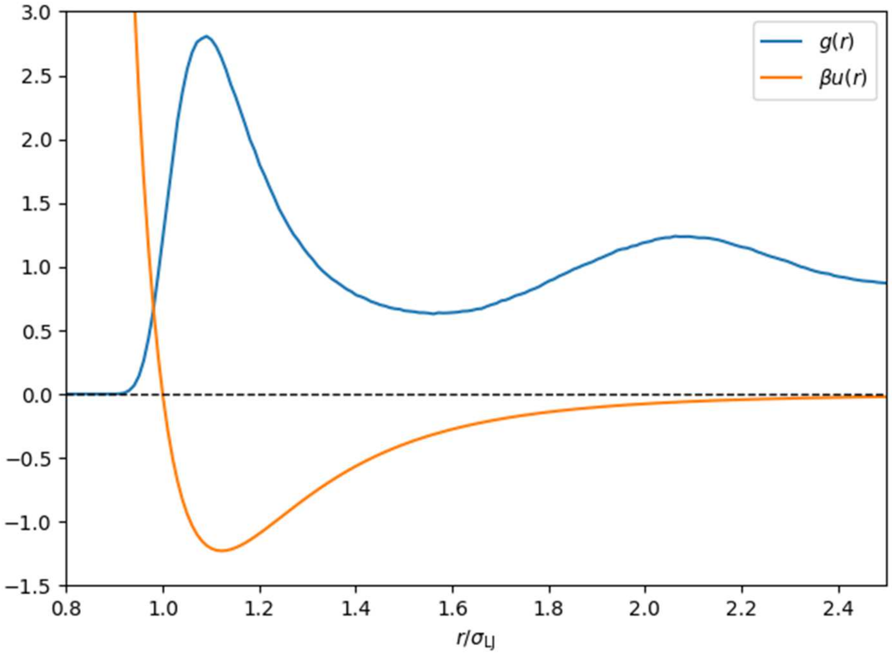

Figure 2.1 Illustrative pair distribution $g(r)$ and and the pair potential function $\beta u(r)$ of a Lennard Jones fluid with a reduced density of $\rho^* = 0.8$ and a reduced temperature of $T^* = 0.81$ obtained by a MonteCarlo simulation (periodic boundary conditions, maximal coordinates shift: 0.002 Å, simulation steps: 1×10^7) plotted against the normalized distance r/σ_{LJ}.

It is one of the key functions of this work and is visualized in **Figure 2.1** for a Lennard-Jones fluid. Here its property of representing the discrete phenomenon of solvation shells as a continuous and smooth function becomes clear. It can be interpreted as the probability of a particle being present at a distance r relative to an undisturbed system, where values greater than one stand for high-density regions and those below one for low-density regions. From the pair distribution function, the total correlation function

$$h(r) = g(r) - 1 \tag{2.5}$$

can be defined, which converges towards 0 instead of 1 at an infinity large distance r, making it behave mathematically more elegant by not having a diverging integral.

In the early 20th century, Ornstein and Zernike[87] formulated a theoretical connection of the total correlation function to the direct correlation function c from experimental measurements, condensed[84] in

$$h(r) = c(r) + \rho \int c(|\mathbf{r} - \mathbf{r}'|)h(r')d\mathbf{r}', \tag{2.6}$$

known as the Ornstein-Zernike equation (OZ). Next to the total correlation function, it is also using the direct correlation function, which may be defined as

$$c(\mathbf{r}, \mathbf{r}') = -\beta \frac{\partial^2 A^{\mathrm{ex}}[\rho]}{\partial\rho(\mathbf{r})\partial\rho(\mathbf{r}')}, \tag{2.7}$$

showing its characteristic of modeling the response of the excess free energy to changes in the local density distribution. It reflects the direct relation between particles and is shorter ranged than the pair distribution function and of simpler structure (see **Figure 2.1**). From (2.7) it is clear that the OZ cannot be solved on its own by its two unknowns. For mathematical modeling, it is usually paired with a second equation that connects both correlation functions, yielding in a closed relation. Such closure relations have the general form of

$$h(r) + 1 = \exp\left[-\beta u(r) + \rho \int c(r')h(|\mathbf{r} - \mathbf{r}'|)d\mathbf{r}' + B\right], \tag{2.8}$$

with $u(r)$ as the pairwise potential and $\beta = 1/k_{\mathrm{B}}T$ as inverse temperature. The integral term describes all correlations up to second-order, and the bridge function B is a substitute for all higher correlations as those are not be solved analytically. There are three general approaches to the bridge function to deal with this problem. First is a numerical approximation[88], for example, via extraction from simulations[89], the second option is an analytical approximation, as shown below, and the third is the neglection entirely, which leads to the hypernetted-chain closure HNC[90–93] consequently written as:

$$h(r) + 1 = \exp\left[-\beta u(r) + \rho \int c(r')h(|\mathbf{r} - \mathbf{r}'|)d\mathbf{r}'\right]. \tag{2.9}$$

This second equation can be solved iteratively by starting with an initial guess of either the direct or total correlation function and using the result as input for the OZ equation (2.6). Repeating this process will eventually lead to a converging solution, providing a sound, density-based thermodynamic description of the fluid.

Still, this holds only for fluids of spherical particles, limiting the applicability of the theory greatly. For molecules that cannot be described or reasonably approximated by spheres, another relation must be found. The first approach to this is the molecular Ornstein-Zernike equation (MOZ)[84,85,94], which encompasses Euler angles Ω_α and Ω_γ and is given by

$$h\left(\mathbf{r}_{\alpha\gamma}, \boldsymbol{\Omega}_\alpha, \boldsymbol{\Omega}_\gamma\right) = c\left(\mathbf{r}_{\alpha\gamma}, \boldsymbol{\Omega}_\alpha, \boldsymbol{\Omega}_\gamma\right) + \frac{\rho}{8\pi^2} \iint c\left(\mathbf{r}_{\alpha\gamma'}, \boldsymbol{\Omega}_\alpha, \boldsymbol{\Omega}_{\gamma'}\right) h\left(\mathbf{r}_{\gamma'\gamma}, \boldsymbol{\Omega}_{\gamma'}, \boldsymbol{\Omega}_\gamma\right) d\mathbf{r}_{\gamma'} d\boldsymbol{\Omega}_{\gamma'}. \tag{2.10}$$

The subscripts α and γ denote different particles in the solution where the first is usually referring to solute sites, while the latter is used for solvent sites. The resulting high dimensionality and angle dependency prevent an analytical or numerical solution, except for a few edge cases[95,96]. For this reason, the reference interaction site model[81–83] (RISM) was developed, which solves the problem by making the key assumption that spatial and angular dimensions can be treated separately from each other. This opens the way to a further assumption, treating the solvent as being made of spherical particles, which allows the approximation of the direct correlation function as a sum over individual site-wise contributions

$$c(r) = \sum_\alpha \sum_\gamma c_{\alpha\gamma}(r_{\alpha\gamma}). \tag{2.11}$$

This approach eliminates the angular dependence but also removes every information about the molecular orientation of the solute[84].

The treatment of the solvent as a mixture of interaction sites itself requires a suitable molecular representation. The intramolecular correlation function (here written for the solvent)

$$\omega_{\gamma\gamma'}(r) = \frac{\delta\left(\left|r_{\gamma\gamma'} - l_{\gamma\gamma'}\right|\right)}{4\pi l_{\gamma\gamma'}^2} \tag{2.12}$$

with the Dirac delta function δ, accommodates the mentioned separation while retaining the molecular structure by encoding interatomic distances l. With this, the 1D-RISM function[85] can be written as

$$h_{\alpha\gamma}(r) = \sum_{\alpha'} \sum_{\gamma'} \iint \omega_{\alpha\alpha'}(|\mathbf{r_1} - \mathbf{r}'|) c_{\alpha'\gamma'}(|\mathbf{r}' - \mathbf{r}''|) \chi_{\gamma\gamma'}(|\mathbf{r}'' - \mathbf{r}_2|) d\mathbf{r}' d\mathbf{r}''. \tag{2.13}$$

The reduced solvent susceptibility[97]

$$\chi_{\gamma\gamma'}(r) = \omega_{\gamma\gamma'}(r) + \rho h_{\gamma\gamma'}(r) \tag{2.14}$$

itself is also calculated from 1D RISM by treating the solute α as one specific solvent particle, a procedure known as the Percus trick[84]. With this in place and together with an appropriate closure, it is possible to calculate the one-dimensional total correlation function of a molecular solute in a molecular solvent. Nonetheless, there are also drawbacks to the 1D RISM approach. For once, its numerical stability is rather poor, and it is difficult to achieve converging solutions for complex solutes. Secondly, by averaging over all solute sites, any site-specific solvent structure is lost, yielding an insufficient description of local solvation effects[67]. In the mid-1990s, multiple groups set out to solve these issues, developing the 3D RISM approach[30–33]. The core idea is to approximate the 6D MOZ by integrating over the orientational dimensions and solving the 3D-dependent relation for each solvent site[67]. Usually, those equations are solved for each point of a 3D grid in which the molecule of interest is embedded. This approach still leads to a loss of any information about angular orientation and a superposition of solvent site correlation functions, but in contrast to 1D RISM, it gives access to the local solvent distribution in 3D and makes calculations for large molecules like proteins possible. While theoretically possible, such can most often not be brought to convergence in 1D RISM.

For 3D RISM, the total correlation function is given by

$$h_\gamma(\mathbf{r}) = \rho^{-1} \sum_{\gamma'} \int c_{\gamma'}(\mathbf{r} - \mathbf{r}')\chi_{\gamma\gamma'}(|\mathbf{r}'|)\, d\mathbf{r}' \tag{2.15}$$

where the solvent site–site susceptibility $\chi_{\gamma\gamma'}$ is once again calculated from 1D RISM. The corresponding HNC closure can be written as

$$h_\gamma(\mathbf{r}) = \exp\Big(-\beta u_\gamma(\mathbf{r}) + h_\gamma(\mathbf{r}) - c_\gamma(\mathbf{r})\Big) - 1. \tag{2.16}$$

The intermolecular potential u_γ is calculated as the sum over all solute sites and taking most often the shape of

$$u_\gamma(\mathbf{r}) = \sum_\alpha \frac{q_\alpha q_\gamma}{4\pi\epsilon_0|\mathbf{r} - \mathbf{r}_\alpha|} + 4\epsilon_{\alpha\gamma}\left(\left(\frac{\sigma_{\alpha\gamma}}{|\mathbf{r} - \mathbf{r}_\alpha|}\right)^{12} - \left(\frac{\sigma_{\alpha\gamma}}{|\mathbf{r} - \mathbf{r}_\alpha|}\right)^{6}\right), \tag{2.17}$$

with the Coulomb potential term on the left and the Lennard-Jones potential term on the right. The partial charges q_α and q_γ as well as the Lennard-Jones parameter $\epsilon_{\alpha\gamma}$ and $\sigma_{\alpha\gamma}$ can be taken from typical force fields[98,99] or from appropriate quantum mechanical calculations in the case of the partial charges for small molecules.

Because of the diverging nature of its exponent for short distances, the HNC closure can be rather unstable, and bringing it to convergence is therefore difficult, especially in its application for 3D RISM. A more stable closure is its partial series expansion of order n (PSE-n)[100] developed by Stefan Kast and Thomas Kloss in the form of

$$h_\gamma(\mathbf{r}) = \begin{cases} \sum_{i=0}^{n} \left(t_\gamma^{\mathrm{R}}(\mathbf{r})\right)^i \Big/ i! - 1 \Leftrightarrow & t_\gamma^{\mathrm{R}}(\mathbf{r}) > 0 \\ \exp\left(t_\gamma^{\mathrm{R}}(\mathbf{r})\right) - 1 \quad \Leftrightarrow & t_\gamma^{\mathrm{R}}(\mathbf{r}) \leq 0 \end{cases} \tag{2.18}$$

with the renormalized indirect correlation function $t_\gamma^{\mathrm{R}}(\mathbf{r}) = h_\gamma(\mathbf{r}) - c_\gamma(\mathbf{r}) - \beta u_\gamma(\mathbf{r})$. Since the excess chemical potential can be written as[101–103]

$$\mu^{\mathrm{ex}} = \sum_{\gamma} \rho_{\gamma} \int_0^1 d\lambda \int d\mathbf{r}\, u(\mathbf{r}, \lambda) g(\mathbf{r}, \lambda), \tag{2.19}$$

(2.15) together with (2.18) after analytical integration of the coupling parameter λ finally yields the closed-form expression

$$\begin{aligned}\mu^{\mathrm{ex}} = \beta^{-1} \sum_{\gamma} \rho_{\gamma} \int d\mathbf{r} \Bigg[& \frac{1}{2} h_{\gamma}^{2}(\mathbf{r}) - c_{\gamma}(\mathbf{r}) - \frac{1}{2} h_{\gamma}(\mathbf{r}) c_{\gamma}(\mathbf{r}) + \\ & - \Theta\Big(h_{\gamma}(\mathbf{r})\Big) \frac{\left(t_{\gamma}^{\mathrm{R}}(\mathbf{r})\right)^{n+1}}{(n+1)!} \Bigg],\end{aligned} \tag{2.20}$$

using the Heaviside step function Θ, being 1 where $h_{\gamma}(\mathbf{r})$ is greater than 0 and 0 everywhere else. The free energy of a molecule in solution can be written as

$$G_{\mathrm{solv}} = E_{\mathrm{sol}} + \mu^{\mathrm{ex}}, \tag{2.21}$$

with E_{sol} being its electronic energy. For most parts of this work, the approximation of unpolarizable molecules is being made, effectively neglecting the electronic term. However, with the embedded-cluster reference interaction site model[61] (EC-RISM), this term can be calculated using an iterative cycle of quantum mechanical and 3D RISM calculations. Here solute's wavefunction is first calculated in vacuum. The resulting electronic potential is then used to approximate the solvent distribution via 3D RISM, which in turn is used to inform a renewed calculation of the wavefunction of the solute. This circle is self consistently repeated until a convergence criterium is reached and E_{sol}. and μ^{ex} are gained.

Besides the excess chemical potential, the effective size of a solute in solvation, the so-called partial molar volume (PMV), can also be calculated from solvent distribution[104,105]. It is accessible via the total correlation functions with

$$V_{m,h} = \beta^{-1}\kappa - \int h_{\gamma}(\mathbf{r}) d\mathbf{r}\ , \tag{2.22}$$

and from the direct correlation function[104] with

$$V_{m,c} = \beta^{-1}\kappa\left(1 - \rho\sum_{\gamma}\int c_{\gamma}(\mathbf{r})d\mathbf{r}\right), \tag{2.23}$$

where κ is the isothermal compressibility[106]. The partial molar volume can be used to correct the intrinsic RISM error of overestimating the contribution of the cavity formation to the excess chemical potential.

2.2 Applications of functional derivative of the free energy

2.2.1 LFE: localization of the free energy onto individual sites

The formalism to localize individual contributions to the free energy on solute sites α is one of two key relations for this work. The fundamental equation is derived from functional integration, which can be found throughout the field of statistical thermodynamics. The general approach leading to (2.32) can be found in a similar form in *Molecular Theory of Solvation*[85], but additional derivations can be found in publications of Kast[107] and Kast et al.[100]. To illustrate the procedure, an excursion in the mathematical background of functional derivatives is necessary.

To this end, a linear functional F may be defined as

$$\mathrm{F} = F[f(x)] \tag{2.24}$$

with a functional derivative given by

$$\frac{\delta \mathrm{F}}{\delta f(x)} = z(x), \tag{2.25}$$

where z is also a functional of the form

$$z(x) = z[x, f(x)]. \tag{2.26}$$

Since F is defined as linear (hence the form of its derivative in Eq. (2.25)) it can also be written as

$$\mathrm{F} = \int \frac{\delta \mathrm{F}}{\delta f(x)} f(x) dx. \tag{2.27}$$

To introduce the Kirkwood coupling parameter[108] λ, the definition of F is expanded to give $\mathrm{F}_\lambda = F[\lambda f(x)]$ with λ ranging between 0 and 1, yielding the respective derivative

$$\frac{\delta \mathrm{F}_\lambda}{\delta \lambda} = \int z_\lambda(x) \frac{\delta \lambda f(x)}{\delta \lambda} dx, \tag{2.28}$$

where $z_\lambda(x) = z[x, \lambda f(x)]$ consequently. Despite being mathematically correct, this linear λ-scaling of $f(x)$ has severe drawbacks when it comes to solving the resulting equation numerically, in most cases rendering it even impossible. To avoid these issues later, the λ in the second term under the integral is not eliminated and instead, the functions f and z become λ-dependent. This opens the way for more elaborated scaling schemes such as softcore scaling[109–111], which is discussed later. Applying the mentioned modifications to Eq. (2.28) yields

$$\frac{\partial \mathrm{F}_\lambda}{\partial \lambda} = \int z(x, \lambda) \frac{\partial f(x, \lambda)}{\partial \lambda} dx, \tag{2.29}$$

The integration of Eq. (2.29) over λ is part of the Kirkwood formalism[108] and finally leads to

$$\mathrm{F} = \mathrm{F}_0 + \int_0^1 d\lambda \int z(x, \lambda) \frac{\partial f(x, \lambda)}{\partial \lambda} dx. \tag{2.30}$$

To make the connection back to statistical mechanics, the Helmholtz free energy A can be interpreted as functional of the intermolecular potential between the solute sites α and the solvent sites γ which gives the derivative with respect to the pair-wise potential as[84,85,112]

$$\frac{\delta A}{\delta u_{\alpha\gamma}(r)} = \rho_0 g(r). \tag{2.31}$$

Starting from (2.25), $z(x)$ is substituted for $\rho_0 g(r)$, $f(x)$ for $u_{\alpha\gamma}(r)$ which itself is given in (2.17), and F for A one can follow the outlined mathematics above, which will eventually lead to[84,85,112]

$$\Delta A = \rho_0 \sum_{\alpha} \sum_{\gamma} \int_0^1 d\lambda \int g_\gamma(\mathbf{r}, \lambda) \frac{\partial u_{\alpha\gamma}(\mathbf{r}, \lambda)}{\partial \lambda} d\mathbf{r}. \tag{2.32}$$

The resulting equation is a general approach to calculate the difference in the Helmholtz free energy ΔA between the start ($\lambda = 0$) and end-point ($\lambda = 1$) of a thermodynamic process and is a universal connection between the pair distribution function g and ΔA.

Finally, localizing individual contributions to ΔA on atoms is achieved by simply not evaluating the sum over sites α in (2.32) which gives an individual free energy for each solute site α

$$\Delta A_\alpha = \rho_0 \sum_{\gamma} \int_0^1 d\lambda \int \mathrm{g}_\gamma(\mathbf{r}, \lambda) \frac{\partial u_{\alpha\gamma}(\mathbf{r}, \lambda)}{\partial \lambda} d\mathbf{r}, \tag{2.33}$$

resulting in the localized free energies (LFE). The connection to the excess chemical potential is given by (2.19).

Equation (2.33) is agnostic regarding the origin of the pair distribution function, and it can be calculated from any method modeling the solvent explicitly like molecular dynamics or Monte Carlo simulations. Since such approaches would take extensive sampling for each λ-step to achieve a sufficiently smooth function, simulation methods are impractical for the task. The 3D RISM method gives access to a more efficient way of calculating the three-dimensional pair distribution function, as described in 2.1. This, however, means that the 3D RISM inherent limitations like the superposition approximation of the solvent sites and the over estimation of the cavity formation[66,67] also apply to the LFEs. Especially the latter can be compensated for whole molecules via a linear regression using the partial molar volume (PMV),[66,68–70] but how this could be extended to a localization approach is still unknown.

To introduce a dependence on λ in $g_\gamma(\mathbf{r})$, the potential term in Eq. (2.18) is being modified such that it gives 0 for $\lambda = 0$, decoupling the solute completely from the solvent, and yields to standard pair-wise potential, as it is given in (2.17), for $\lambda = 1$. Although this can be done, in theory, by scaling it linearly with λ so that $\partial u_{\alpha\gamma}(\mathbf{r}, \lambda) \,/\, \partial \lambda$ in Eq. (2.33) shortens to $u_{\alpha\gamma}(\mathbf{r})$[112], applying this approach on Lennard-Jones- and Coulomb-potentials could cause numerical issues near $\lambda = 0$ and $\lambda = 1$ (depending on the process). This problem is well known in the context of TI simulations and is sometimes called end-point catastrophe[110,111,113–117]. Borrowed from this community, the issue is avoided by implementing softcore scaling, which can be written for the Lennard-Jones potential as[109–111]

$$u_{\alpha\gamma}^{\mathrm{LJ_{sc}}} = 4\varepsilon_{\alpha\gamma}\lambda\left[\frac{1}{\left(\alpha_{\mathrm{LJ}}(1-\lambda)+\left(r_{\alpha\gamma}\,/\,\sigma_{\alpha\gamma}\right)^{6}\right)^{2}} - \frac{1}{\alpha_{\mathrm{LJ}}(1-\lambda)+\left(r_{\alpha\gamma}\,/\,\sigma_{\alpha\gamma}\right)^{6}}\right], \qquad (2.34)$$

where the parameter α_{LJ} is adjusting the harshness of the function. When van der Waal and electrostatic interactions are treated separately, and in this order, the Coulomb term can stay unmodified and is scaled linearly as described above. Eventhough it is a rather stable and robust way to scale the potential $u_{\alpha\gamma}(\mathbf{r},\lambda)$ with respect to λ, it is not the only one and other schemas are possible, which can have an influence on the LFE values themselves. In this work, however, only the here shown approach and the, in the corresponding sections for computational methods detailed, scalings are used, ensuring consistency between LFEs.

2.2.2 Localization of derivatives of the free energy with respect to force field parameters

The approach taken above, to localize the contributions to the free energy on solute sites, can be extended to derivatives in a straightforward manner. Again, starting from equation (2.31) it is written in its general integral form[100,118–120]

$$\delta A = \rho_0 \int g(r)\delta u_{\alpha\gamma}(r)dr, \qquad (2.35)$$

and in analogy to (2.32) the equations for the derivatives with respect to the non-bonded force field parameters ϵ, σ, and q can be written as

$$\frac{\partial A_\alpha}{\partial \epsilon_\alpha} = \rho_0 \sum_\gamma \int g_\gamma(\mathbf{r}) \frac{\partial u_{\alpha\gamma}(\mathbf{r})}{\partial \epsilon_\alpha} d\mathbf{r}\,, \tag{2.36}$$

$$\frac{\partial A_\alpha}{\partial \sigma_\alpha} = \rho_0 \sum_\gamma \int g_\gamma(\mathbf{r}) \frac{\partial u_{\alpha\gamma}(\mathbf{r})}{\partial \sigma_\alpha} d\mathbf{r}\,, \text{and} \tag{2.37}$$

$$\frac{\partial A_\alpha}{\partial q_\alpha} = \rho_0 \sum_\gamma \int g_\gamma(\mathbf{r}) \frac{\partial u_{\alpha\gamma}(\mathbf{r})}{\partial q_\alpha} d\mathbf{r}. \tag{2.38}$$

The derivatives of the pair-wise potential $u_{\alpha\gamma}(\mathbf{r})$ with respect to the corresponding parameters of a single solute site are given by

$$\frac{\partial u_{\alpha\gamma}(\mathbf{r})}{\partial \epsilon_\alpha} = \frac{q_\alpha q_\gamma}{4\pi\epsilon_0 |\mathbf{r}-\mathbf{r}_\alpha|} + 2\frac{\epsilon_\gamma}{\epsilon_{\alpha\gamma}}\left(\left(\frac{\sigma_{\alpha\gamma}}{|\mathbf{r}-\mathbf{r}_\alpha|}\right)^{12} - \left(\frac{\sigma_{\alpha\gamma}}{|\mathbf{r}-\mathbf{r}_\alpha|}\right)^{6}\right), \tag{2.39}$$

$$\frac{\partial u_{\alpha\gamma}(\mathbf{r})}{\partial \sigma_\alpha} = \frac{q_\alpha q_\gamma}{4\pi\epsilon_0 |\mathbf{r}-\mathbf{r}_\alpha|} + 4\epsilon_{\alpha\gamma}\left(\frac{3\left(\sigma_\alpha+\sigma_\gamma\right)^{11}}{1024|\mathbf{r}-\mathbf{r}_\alpha|^{12}} - \frac{3\left(\sigma_\alpha+\sigma_\gamma\right)^{5}}{32|\mathbf{r}-\mathbf{r}_\alpha|^{6}}\right), \tag{2.40}$$

$$\frac{\partial u_{\alpha\gamma}(\mathbf{r})}{\partial q_\alpha} = \frac{q_\gamma}{4\pi\epsilon_0 |\mathbf{r}-\mathbf{r}_\alpha|} + 4\epsilon_{\alpha\gamma}\left(\left(\frac{\sigma_{\alpha\gamma}}{|\mathbf{r}-\mathbf{r}_\alpha|}\right)^{12} - \left(\frac{\sigma_{\alpha\gamma}}{|\mathbf{r}-\mathbf{r}_\alpha|}\right)^{6}\right) \tag{2.41}$$

assuming the mixing rules $\epsilon_{\alpha\gamma} = \sqrt{\epsilon_\alpha \epsilon_\gamma}$ and $\sigma_{\alpha\gamma} = \sigma_\alpha + \sigma_\gamma/2$ for the Lennard-Jones parameters.

The core concept of calculating the derivative of the free energy with respect to force field parameters like ϵ, σ, and q was already introduced by Mrugalla et al.[59,60]. The novelty of the method outlined above is the calculation via analytical derivatives, contrasting the previous numerical, *uu*-3D RISM based approach, and thereby gained computational efficiency.

To illustrate the behavior of the relevant functions and to help with the interpretation of the results, the Coulomb- and Lennard-Jones potential and their derivatives are shown in **Figure 2.2** as functions of r.

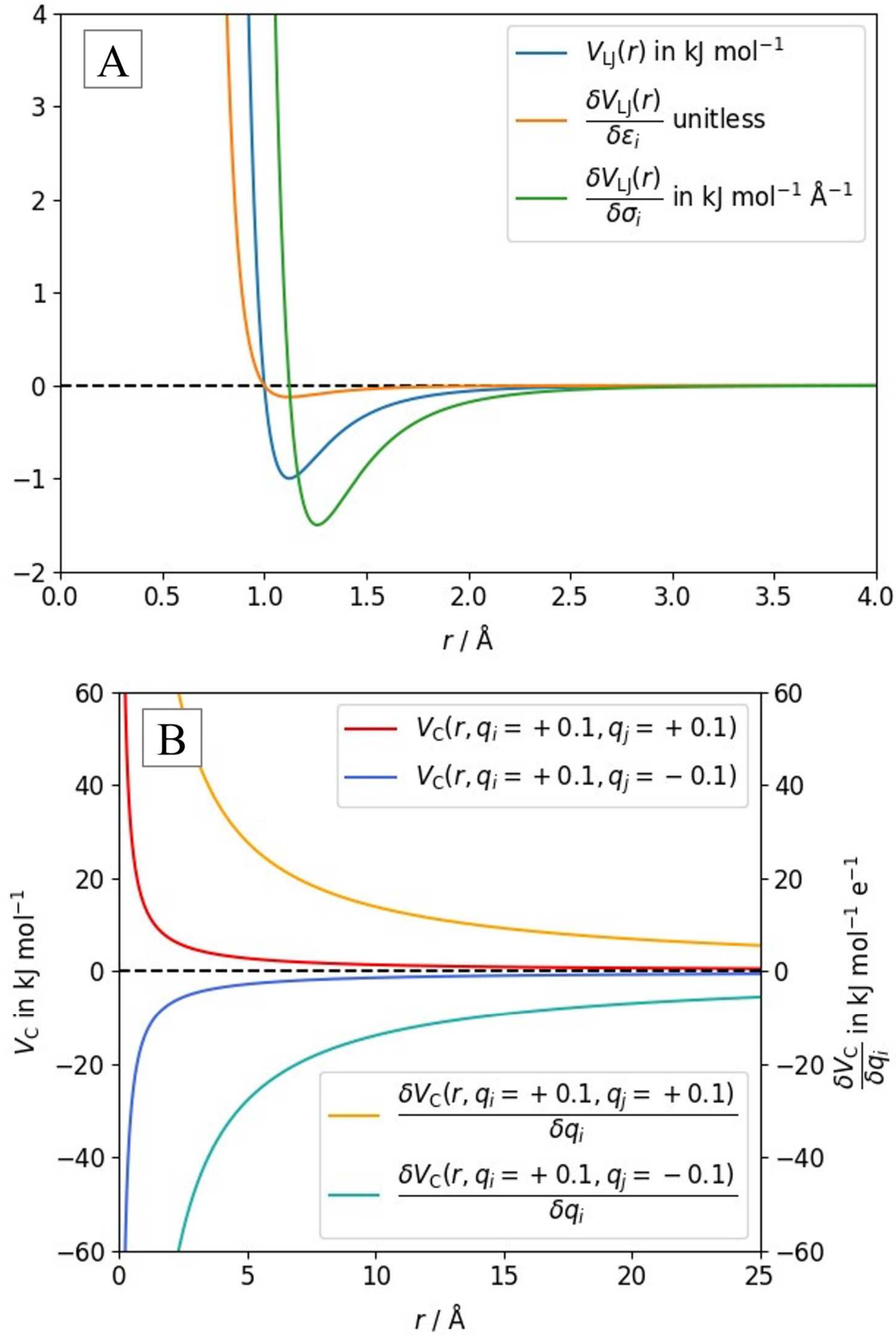

Figure 2.2 A: Lennard-Jones potential energy dependent on distance r for σ (in Å) and ϵ (in kJ/mol) = 1 for both particles i and j as well its end-point derivatives with respect to σ_i and ϵ_i. B: Coulomb potential energy for equal and opposite charged particles as well as their derivatives with respect to q_i (in e).

The derivatives of the Lennard-Jones potential energy show very different behavior relative to each other while still retraining the core characteristics of the function of the potential itself. The minimum of the derivatives with respect to ϵ_i is much lower than the one of the derivative with respect to σ_i but shares its zero-crossing with the parent function. This is in line with the fact that interactions via induced dipoles, which are scaled by the ϵ-parameter, are relatively weak, and extrapolated to the FEDs one can expect that the free energy will change only slightly with varying ϵ_i. This is in strong contrast to the derivative with respect to σ_i, as not only its minimum is deeper, but also the zero-crossing occurs at a larger distance r. Both characteristics are important for the interpretation of the derivatives with respect to σ_i but especially the shift in the zero-crossing makes the free energy sensitive for close contacts.

For the Coulomb potential energy, shown in **Figure 2.2** B, two cases must be distinguished. For charges with the same sign, the potential energy is always positive, and at small distances, it diverges towards positive infinity. For charges with opposite signs, it is always negative and diverges towards negative infinity. The derivatives behave similarly, but their sign is set by the partner charge, which remains in the equation. Especially in comparison to the Lennard-Jones potential energy, the absolute magnitude and the long-ranged nature of the Coulomb potential energy and its derivatives become apparent. Since there is always a summation step over multiple charged sites in the calculation of FEDs (and LFEs) involved, most contributions cancel each other out, especially over long distances. Nonetheless, the derivatives of the Coulomb potential energy with respect to the partial charge of a single site most often dominates the FEDs, which must be taken into considerations when it comes to interpretation of any results.

2.3 Deep learning

The field of deep learning can be seen as a sub-field of machine learning, which itself is the science of developing methods and algorithms that can find and exploit patterns from data distributions by statistical means and without explicitly defined rules. In the case of supervised algorithms, this is usually done by learning the parameters of a function from samples of the provided training distribution. Such a function can then be used to make predictions from unseen samples and thereby solve a predefined task. Those tasks most often fall in one of two categories, classification and regression, of which only the latter is used in this work.

Deep learning itself refers to a specific type of model, so-called artificial (deep) neural networks. Attempts to create artificial neural networks date back to the very beginning of the computer age in the early 1950ths when scientists modeled neurons and synapsis in semi-mechanical-semi digital machines, driven by the question of how to make machines learn. This led, for example, to the Stochastic Neural Analog Reinforcement Calculator (SNARC)[121] by Minsky, Miller, and Edmonds, which was trained to solve mazes or the Perceptron[122,123] by Rosenblatt classifying images. Those early inventions and innovations were often heavily inspired by biological learning processes and cognition in general, which gave many concepts in the field their names. However, modern deep learning algorithms resemble only vaguely biological neural networks and are optimized fundamentally differently.

The architecture of a general multi-layer neural network, also called multi-layer perceptron[124] (MLP), can be expressed as a graph of nodes, standing for simple computational operations and connecting directional edges, visualized in **Figure 2.3**.

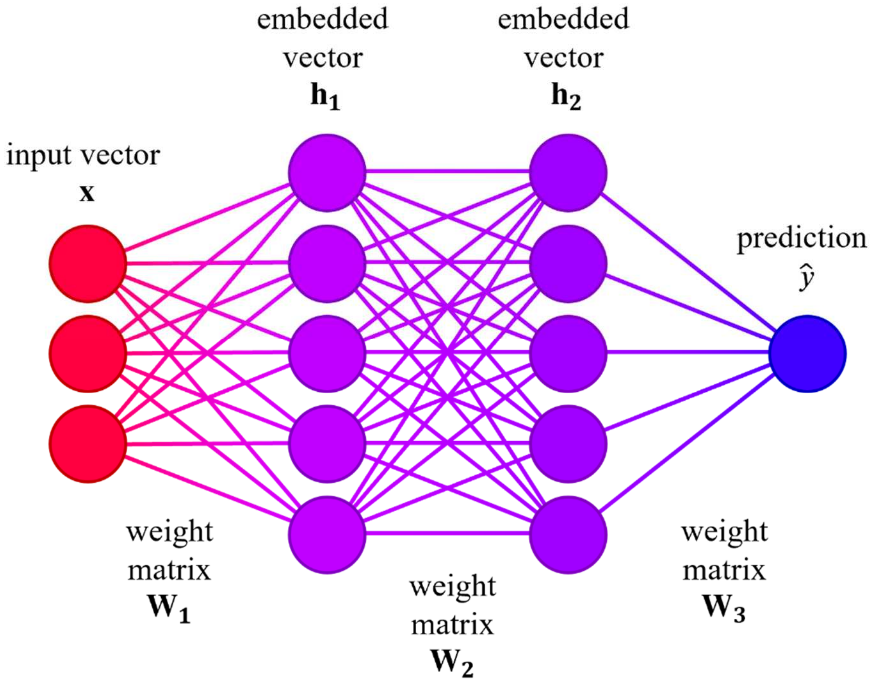

Figure 2.3 General representation of a two-layer MLP The nodes in the graph represent the input layer, accepting the input vector $\boldsymbol{x}$, the two hidden layers with their embedded vectors $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$, as well as the output layer with the prediction $\hat{y}$. The edges represent the weight matrices, connecting the different layers.

In a forward path through the compute graph, the input vector $\mathbf{x} = (x_1 \dots x_K)^T$ first gets multiplied by each weight vector $\mathbf{w}_i = (w_{i1} \dots w_{iK})$ for every node in the first hidden layer. This can be done very efficiently, even for very large weight vectors when they are arranged as a matrix, making the operation a matrix-vector dot product for which highly optimized algorithms and hardware (graphics processing units (GPUs)) can be used. The resulting vector serves as input for a non-linear activation function a for which some examples are shown in **Figure 2.4**. Both operations are written together as

$$\mathbf{h} = a(\mathbf{W}\mathbf{x}) + \boldsymbol{\beta}, \tag{2.42}$$

giving, together with the optional bias vector $\boldsymbol{\beta}$, the embedded vector $\mathbf{h}$. This process is repeated for every hidden layer in the graph. The final prediction $\hat{y}$ is obtained from the last layer, the output layer, whose activation function depends on the network's task. In the case of classification, a softmax or sigmoidal activation function are usual choices[125,126]; for regression tasks, it is usually can be omitted completely, making the output a linear function.

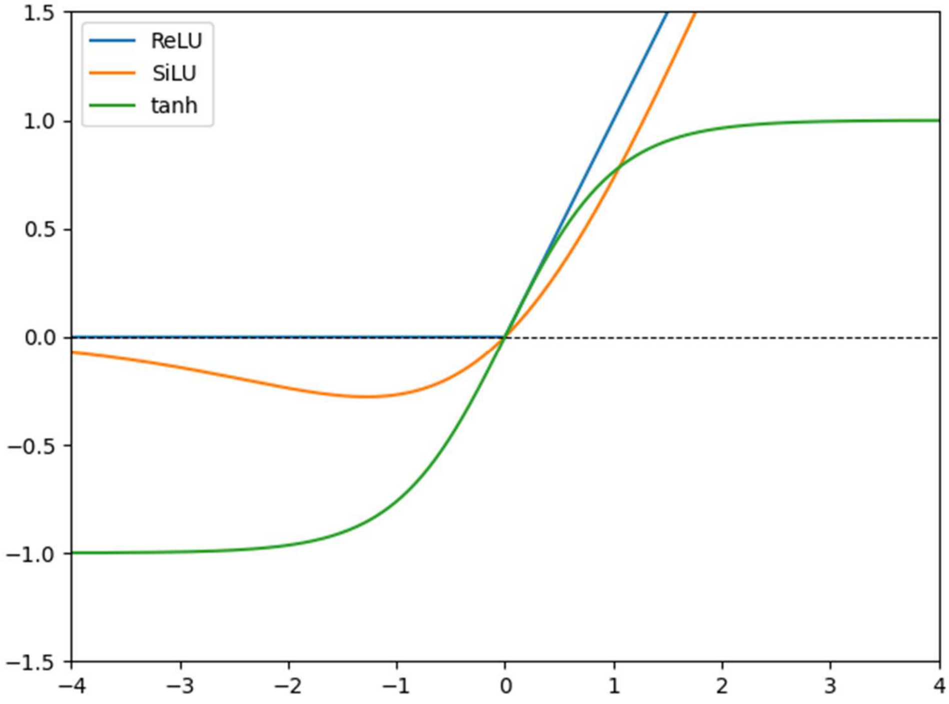

Figure 2.4 Examples for activation functions typically used on hidden layers in deep neural networks.

The weights, as well as the biases if used, are learnable parameters that give deep neural networks their broad field of applications. For a network with a single hidden layer and unlimited depth, they can be fitted such that the network can approximate any arbitrary function connecting input and output within Euclidian space[127–129]. To do so, the neural network must be trained on a set of samples taken from the input distribution, often called the training set. The training process is divided into three separate steps, including the calculation of a loss from $\hat{y}$ and the ground truth y, finding the derivatives of it with respect to each learnable parameter in the network, and lastly the optimization of the parameters such that the error is reduced.

The calculation of the loss function (in older literature, sometimes called cost function) depends mainly on the task required of the model. So is the Cross-entropy loss, derived from maximum likelihood[130], for C classes and N samples

$$E(\hat{\mathbf{y}}, \mathbf{y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\left(\hat{y}_{i,c}\right) , \tag{2.43}$$

where $\hat{\mathbf{y}}$ and $\mathbf{y}$ are vectors of length N is a common choice for classification problems. Whereas the mean absolute error (MAE)

$$E(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} |y_i - \widehat{y_\imath}|, \tag{2.44}$$

and the mean squared error (MSE)

$$E(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \widehat{y_\imath})^2, \tag{2.45}$$

on the other hand are often used for regression tasks. For the optimization of the learnable parameters the backpropagation algorithm[131,132] is used almost exclusively and can be written as four separate equations, revolving around the question how to change the weights and biases so that the loss changes in a specific way. The first expression

$$\boldsymbol{\delta}^L = \frac{\partial E}{\partial \mathbf{h}^L} a'\left(\mathbf{u}^L\right) = \nabla_{h^L} E \circ a'\left(\mathbf{u}^L\right), \tag{2.46}$$

computes $\boldsymbol{\delta}^L$, commonly referred to as the error vector for the output layer L. Where the element-wise product, or Hadamard product, is denoted by $\circ$ and the weighted input to the activation function is given by $\mathbf{u}^L$. In a similar fashion, this approach is now followed backward through the network for each layer l, considering the error vector of the next layer in forward direction $l + 1$ according to

$$\boldsymbol{\delta}^l = \mathbf{w}^{l+1} \delta^{t+1} \circ a'\left(\mathbf{u}^l\right). \tag{2.47}$$

With the error vectors to each layer at hand, the derivative of the loss with respect to each weight in the network can be written as

$$\frac{\partial E}{\partial w_{ij}^l} = h_j^{l-1} \delta_i^l \tag{2.48}$$

and to each bias as

$$\frac{\partial E}{\partial \beta_i^l} = \delta_i^l. \tag{2.49}$$

These gradients can now be used to optimize the learnable parameters accordingly, for which a multitude of algorithms are available[133,134]. In the simplest case, the weights and biases are updated by

$$\Delta w_{ij}^l = -\alpha \frac{\partial E}{\partial w_{ij}^l} \text{ respectivly } \Delta \beta_i^l = -\alpha \frac{\partial E}{\partial \beta_i^l} \tag{2.50}$$

with a learning rate α.

From the outline of the backpropagation algorithm just given, one can see that finding the necessary differentiations for any arbitrary compute graphs is a crucial task and is usually done through automatic differentiation algorithms. Two of the most popular software libraries providing the required and other algorithms are TensorFlow[135] and PyTorch[136].

Since most deep learning applications in chemistry and the natural sciences, like molecular property prediction or chemical reaction-related questions, revolve around molecules, it is necessary to find a suitable Euclidean input representation for the models. This, however, is a non-trivial task, as the very concept of a molecule is not within this space and must therefore first be abstracted to be used in a neural network. Multiple concepts have been developed in the past to do so, but only the most common shall be mentioned here.

A very high level of abstraction is the representation of a molecule via its physical properties and other, more general descriptors. This approach is comparably long known and was popularized by QSAR studies in the 1980s and 90s. Commonly used general features are the number of atoms and elements in a molecule or the presence of specific structural elements and alike as well as more physical ones like the acidity constant pK_a and the n-octanol-water partition coefficient, expressed as log$P$. Structural and conformational information can also be captured, as shown by Behler and Parrinello[137], but only to a limited extend. After calculating these molecular representations, they can then be fed as an input vector to a neural network. An advantage of this approach is the need for only small training datasets, due to the information provided by the physical features. Nevertheless, this also brings the drawback of the potentially high preprocessing costs. Moreover, the calculation of some of the properties, especially the physical ones, can be rather expensive when done towards high accuracy, which often makes cheaper approximations unavoidable not to spoil the speed advantage of DL, introducing undesired noise to the input.

It is often far more efficient to let the model learn everything necessary directly from the input, utilizing both strong points of deep learning approaches, high versatility, and high predictive speed. One way of achieving this is the deployment of molecular fingerprints, which encode the molecular structure as a vector algorithmically. There are a vast variety of different fingerprints available[138] but most commonly used are circular fingerprints[138,139], often referred to as Morgan fingerprints, and, to a lesser degree, MACCS keys[140]. Very recent approaches are transformer models[141,142], taken from natural language processing research, which are applied to text-based encodings of molecules like SMILES[143,144] and SELFIES[145,146]. The strong performance on various tasks, the mostly agnostic input structure, and only little preprocessing make this model type a promising approach for chemical tasks in general. Nonetheless, known downsides are a comparably high need for large numbers of training samples, which are often unavailable, and the typically large size of transformer models, making for longer prediction times.

One of the two in this work used ways to encode molecules for usage in deep learning regression tasks is their representation as three-dimensional point clouds, used as input for three-dimensional convolutional neural networks (3D CNN). This has the advantage of the direct utilization of conformational information, and different atom-wise input feature can be mapped onto different channels of such grids, similar to the color channels of 2D images. However, bond-related information is difficult to represent with this approach since it cannot be projected onto these channels. The other here used method solves this issue by using undirected (in this case equivalent to bi-directed) graphs of nodes and edges to represent molecules in a so-called message passing neural network (MPNN)[147]. Here the nodes are loaded with atom information while the edges hold the bond information consequently. Both methods are explained in more detail below.

2.3.1 3D CNN

Convolutional neural networks in their modern form were pioneered by Yann LeCun[148,149] and inspired by the Neocognitron of Kunihiko Fukushima[150], itself derived from the general architecture of the visual cortices of mammals[151,152]. The core idea is to apply multiple kernels with comparably small dimensionality to the given input, filtering it to extract the relevant features for the task at hand. Mathematically this can be expressed as

$$y_i = b_i + \sum_{j=1}^{N_{\text{channels}}} k_{i,j} * x_j \ , \tag{2.51}$$

where the feature map y_i is calculated from the jth input channel x_j and N_{channels} kernels, denoted by the learnable parameters $k_{i,j}$ as well as the bias b_i. A kernel can be imagined as a single, small layer with learnable parameters, which gets applied stepwise to the full grid. The step width is referred to as stride. This is repeated for each channel (or feature map when applied to hidden layers) i. The expression “channel” originates here from image analysis, referring to the different (color-) channels on a pixel. In practice, the convolution is achieved by moving the kernels stepwise over a 2D slice of the volume input with a predetermined stride. The dot product with the input is computed on each step, and the result is put through a non-linearity function (see above). This process can be repeated over multiple layers with the previous layer's output as input for the next. In this way, the network can detect important features beginning with relatively simple ones in the first layers like edges in images, increasing in complexity throughout the network towards objects and faces in the higher layers. To get the final prediction, most applications, like regression and classification, have fully connected readout layers at the end of the network, which are fed with the linearized output of the convolutional part of the network.

This general idea of a kernel-based transformation is extremely efficient compared to a direct gridpoint-wise computation as the input dimension would be immense (). This way the resulting models stay within computationally feasible limits while still generalizing well and being applicable to different kinds of input data, given sufficient training data and model capacity.

Most modern CNN architectures[153] like ResNet[154], EfficientNet[155], and UNet[156] perform dimensional reduction, or downsampling, to improve translational invariance and to reduce the number of learnable parameters. This is usually done architecturally by either dimension reducing convolution layers or by using so-called Pooling layers, unifying cells by mathematical operations. Two types are used almost exclusively, max pooling and average pooling. Like the convolutional kernels, both variants are also moved over the feature maps and are applied in a stepwise fashion, reducing the dimensionality of the feature maps and increasing the generalizability of the model.

The extension of the 2D CNN to three dimensions, as they are used in this work, is straightforward and simply adds an extra dimension to all operations.

2.3.2 MPNN

Message passing neural networks (MPNN)[147] are a chemistry-specific member of the broader family of graph convolutional networks (GCN), which use a graph-based representation of molecules. Here atoms and atom-based information are represented by the nodes of the graph, or vertices, and bond information is stored on the edges, whereas the adjacent matrix gives the connectivity. This approach is very versatile since all-atom- and bond-based features can be used as input, and therefore the input can be adapted to the model task through feature engineering by presenting the model only specific atom- and edge-based information.

A full path through an MPNN consists of an embedding phase, where the molecular graph gets embedded into an internal representation through a iterating process, and a readout phase where the latent representation is used to make predictions according to the task of the model. The first atom-wise embeddings h_v^0 are computed for the input features x_v by a pass through an initial, learnable function U^0,

$$h_v^0 = U^0\left(x_v, m_v^0\right), \tag{2.52}$$

which also includes in the implementation used in this work the calculation of an optional initial message m_v^0. The general approach to a message m is given by

$$m_v^{t+1} = \sum_{w \in N(v)} M^t(h_v^t, h_w^t, e_{vw}), \tag{2.53}$$

where a message function M^t is evaluated for each connecting edge w of node v, incorporating the embedding vector h_v^t of the current node at iteration step t as well as the ones of its direct neighbors h_w^t. Since the usage of the MPNN algorithm is focused on a specific atom-wise descriptor in this work, no edge-information e_{vw} is needed, and to avoid any further self-correlation, the message function shortens to $M^t = h_w^t$ (assigning rule based bond orders, for example, could also introduce ambiguities and/or redundancies in the case here). With $M^t = h_w^t$ as message function, all messages can be calculated at once by multiplying $\mathbf{h}^t$ with the adjacency matrix $\mathbf{A}$. A message-passing step is completed by updating the hidden states via

$$h_v^{t+1} = U^t\left(h_v^t, m_v^{t+1}\right), \qquad (2.54)$$

where the update function U^t is once again learnable but the same for every iteration. Equations (2.53) and (2.54) are evaluated in an integrative manner, leading finally to an embedded vector for each node of the input graph. The readout function R, most often a simple MLP, takes the sum over those vectors h^t as input and yields the final predictions via $\hat{y} = R(h)$.

3 Results

3.1 3D RISM based estimation of conformational free energy differences

Antibodies (Ab), also referred to as immunoglobulin (Ig), are part of the immune system and can be classified into the five functional groups IgA, IgE, IgD, IgG, and IgM for humans. In addition, those classes can be broken down into isotopes like IgG into IgG1, IgG2, IgG3, and IgG4. Despite their variance, all antibodies share the same principle structure, having two heavy chains, linking the fragment crystallizable (Fc) region with the two antigen-binding fragments (Fab) and two variable light chains. This structure gives them their core functionality of binding to specific target molecules, so-called antigens. Once an antigen is bound to an antibody, like the IgG-type, they trigger different immune responses. Such are, for example, the immobilization of pathogens via agglutination or the activation of the complement system, cascading into different pathways of which one is an attack on the pathogen membrane through small proteins. Others are the antibody-dependent cellular cytotoxicity, where natural killer cells (NK) bind to the Fc region of the antibody, releasing cytotoxic factors and killing the target cell, and antibody-dependent cellular phagocytosis, effectively marking the pathogen for ingestion by specialized cells, so-called phagocytes.[157–160]

In this chapter, two conformations of the anti-NPRA (natriuretic peptide receptor A) IgG4 antibody are studied. IgG4 is special in the family of IgG antibodies, as it does not activate the complement system and is noninflammatory.[161–165] NPRA, on the other hand, is a receptor for the atrial natriuretic peptide (ANP) and B-type natriuretic peptide (BNP), both elevated in patients with heart failure[166], making the anti-NPRA IgG4 monoclonal antibody (mAb) an interesting target for therapeutic treatments.

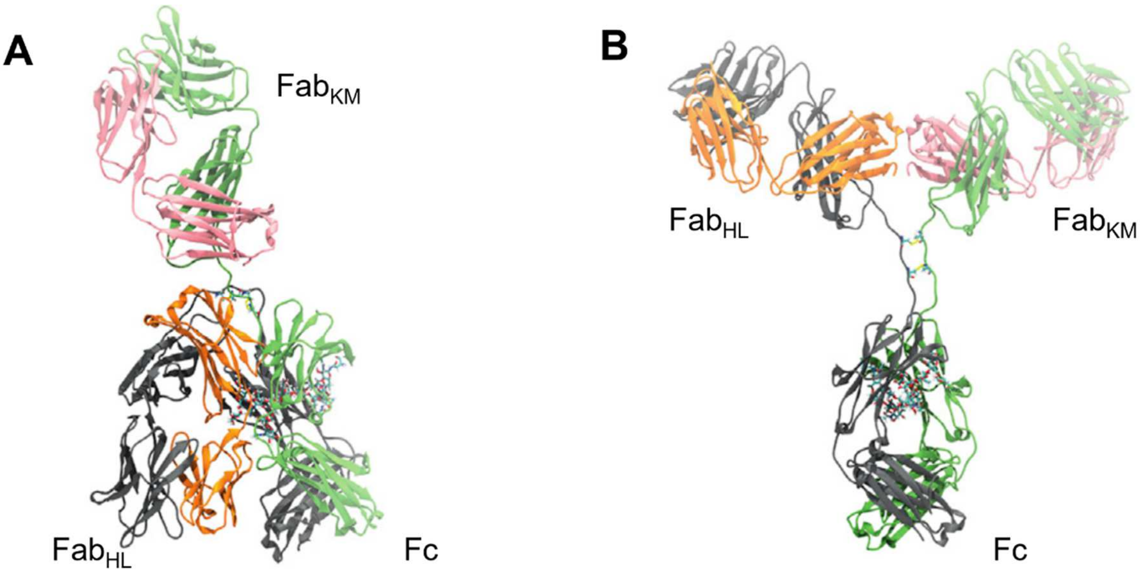

Figure 3.1 Anti-NPRA IgG4 mAb in the λ- (A) and Y- (B) conformation. The anti-NPRA fragments are indicated with the labels Fc, Fab$_{KM}$, and Fab$_{HL}$, while the indices KM and HL stand for the involved heavy chains (colored accordingly). This figure was taken from Blech and Hörer et al.[167] and is not the work of the author.

In x-ray crystallographic experiments, an unusual conformer of anti-NPRA IgG4 can be found. As illustrated in **Figure 3.1** (A) does the overall structure significantly differ from the usual Y-conformer found in antibodies. Here the Fab-Fc orientation displays a distorted λ-shape where the Fab$_{HL}$ region is oriented towards the Fc region. To foster a better understanding of this conformational flexibility, both structures are investigated in solution via MD simulations and 3D RISM calulations in the following.

3.1.1 Computational details

The structures used in the 3D RISM calculations were taken from 63 snapshots of 500 ns MD simulations in the NpT-ensemble of the anti-NPRA IgG4 mAb in the λ- and Y-conformation, generated in the group of Prof. Dr. Lars Schäfer. They were sampled every 8 ns and are therefore considered uncorrelated. The error estimation of the potential force field energy was performed by block averaging[168] (implemented in g_analyze of GROMACS). These simulations were not performed by the author. Further computational details can be found in Blech and Hörer et al.[167].

A specially adapted TIP3P water model with non-zero Lennard-Jones parameters for hydrogen and a dielectric constant of 78.4 at T=298.15 K and a density of 0.0333 Å^{-3} was used as the basis for a DRISM/HNC calculation, yielding the $\chi_{\gamma\gamma'}$-function. The 3D-RISM calculations were performed on a regular grid with a spacing of 0.3 Å and 600 grid points on each axis. The PSE closure[100] of order two was used and the calculations were performed with a convergence criterion of a difference in the residual norm of the direct correlation functions of less than 10^{-5} between two steps. Using the Amber99sb-ILDN/Glycam06j force field parameters in 3D RISM ensured consistency with the MD simulations. A particle-mesh-Ewald approach with Lagrangian interpolation polynomials of order eight was used for the long-range reciprocal-space Coulomb potential. Furthermore, for periodic correction of the long-range electrostatics Kovalenko and Hirata's compensating background charge[169] was used. The excess chemical potential was calculated entirely in reciprocal space. The partial molar volume (PMV) was calculated from the direct correlation function[105] with an isothermal compressibility of $0.450183 \times 10^{9}\ \mathrm{Pa}^{-1}$. The inputs and raw data to each frame and both conformers can be found in the electronic appendix under 3.1/Y_conf/RISM/ and 3.1/lambda_conf/RISM/, respectively.

To test the solvent dependency of the free energy difference between the λ- and Y-conformation of the anti-NPRA, the 3D RISM calculations were repeated in the Schäfer-group with a modified SPC/E water model[170,171], including 150 mM NaCl[172,173]. The density was again 0.0333 Å^{-3} and the dielectric constant 78.4. The employed three-dimensional grid had a spacing of 0.5 Å and 384 grid points on each axis. In this experiment the KH-closure[174] (equivalent with the PSE closure[100] of order one) was used and a convergence criterion of a difference in the residual norm of the direct correlation functions of less than 10^{-5} between two steps was employed. In the 3D RISM calculations utilizing the SPC/E water model the chemical excess potential was calculated in real space and were not performed by the author.

3.1.2 Thermodynamic characterization of IgG4 conformations

The difference in free energy between the λ- and Y-conformer, defined as $\Delta G = G_\lambda - G_\mathrm{Y}$, can be estimated via

$$\Delta G = \Delta H_\mathrm{c} - T\Delta S_\mathrm{c} + \Delta G_\mathrm{hyd}. \tag{3.1}$$

The conformational enthalpies H_c can be calculated by averaging the potential energy E_pot over the MD simulations of the conformers. The potential energy itself consists of the bonded and non-bonded force field terms of the full antibody, including the glycan moiety.

Neglecting the pV contribution does not affect the results in any meaningful way as it is with less than 0.1 kJ/mol comparable small in the condensed phase.[175]

Analyzing the development of the potential energy and hydration free energy (assuming equivalence of μ^{ex} and G_{hyd}) over the simulation time and snapshots respectively in **Figure 3.2**, a clear distinction between both conformers can be seen. The running averages over the potential energies appear to be stable over the 500 ns with no visible drift. Only G_{hyd} of the Y-conformer assumes higher values towards the end of the simulation, approaching G_{hyd} of the λ-conformer. This could be the result of the tendency of the Y-conformer to adopt a third, T-shaped conformer in the simulations. Since the number of snapshots for which this is the case is limited, one can nevertheless assume a significant difference in G_{hyd} between the conformers.

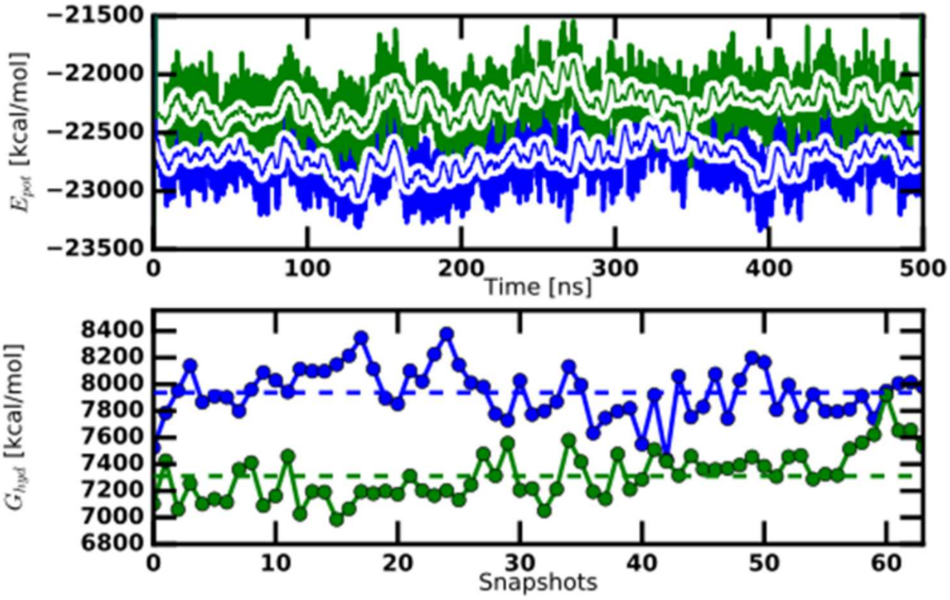

Figure 3.2 The upper plot shows the, from MD simulations extracted, potential energies (bonded and non-bonded solute force field energies) of the λ-conformer (blue) and the Y-conformer (green) of the anti-NPRA IgG4 antibody together with a 5 ns running average. The lower plot shows hydration free energies calculated with 3D RISM (PSE-2/TIP3P) of 63, from the MD simulations extracted snapshots of the λ- (blue) and Y-conformer (green). The MD simulations, and consequently also the snapshots and potential energies, are not the work of the author. This figure was taken from Blech and Hörer et al.[167]. The raw data for G_{hyd} can be found in the electronic appendix under 3.1/Results/.

The difference in the conformational enthalpy between λ- and Y-conformer, as shown in **Table 3.1** and also visible in **Figure 3.2**, is with -521 ± 50 kcal mol^{-1} in clear favor of the λ-

conformer. Apparently, the Fab-Fc interaction and more compact packing of the λ-conformer results in a lower conformational enthalpy as the more spread out Y-conformer.

To estimate the configurational entropy S_c of (3.1), the quasi-harmonic approximation

$$S_\mathrm{c} = \frac{k_\mathrm{B}}{2} \ln\left(\det\left(1 + k_\mathrm{B} T e^2 \hbar^{-2} \mathbf{MC}\right)\right) \tag{3.2}$$

by Schlitter[176], with the Boltzmann constant k_B, reduced Planck constant $\hbar$, and the 3N-dimensional diagonal mass matrix **M** can be used. The elements of the particle position covariance matrix **C** are defined as

$$c_{ij} = \langle (x_i - \langle x_i \rangle)\left(x_j - \langle x_j \rangle\right) \rangle. \tag{3.3}$$

The differences in $-TS_\mathrm{c}$, as shown in **Table 3.1**, is with -19 ± 12 kcal mol^{-1} small compared to ΔH_c and ΔG_hyd but also shows an advantage towards the λ-conformation.

Table 3.1 Conformational enthalpy, H_c, and entropy, $-TS_c$, (MD simulations) as well as the hydration free energy contribution G_{hyd} (3D RISM) to the difference in conformational free energy $\Delta G_{Y \to \lambda}$. To check the robustness of the the results, G_{hyd} was calculated for the TIP3P water model with a minimum (-0.12 kcal mol^{-1} Å^{-3})[177] and maximum (-0.2 kcal mol^{-1} Å^{-3})[177] PMV correction and without PMV correction. In addition to those checks, G_{hyd} was also calculated for a SPC/E water model including 150 mM NaCl. The SPC/E 3D RISM calculations, as well as the MD simulation and the extracted properties, are not the work of the author. This table was taken from Blech and Hörer et al.[167]. The raw data for G_{hyd} for the TIP3P experiments can be found in the electronic appendix under 3.1/Results/.

	H_c / kcal mol^{-1}	G_{hyd} / kcal mol^{-1}	$-TS_c$ / kcal mol^{-1}	$\Delta G_{\mathrm{Y}\to\lambda}$ / kcal mol^{-1}
		PSE-2/TIP3P		
λ-conformer	-22703 ± 32	7937 ± 23	-3134 ± 11	
Y-conformer	-22182 ± 38	7311 ± 23	-3115 ± 5	
$\Delta(\mathrm{Y} \to \lambda)$	-521 ± 50	+626 ± 33	-19 ± 12	+86 ± 61
		PSE-2/TIP3P/PMV$_{\mathrm{min}}$		
λ-conformer	-22703 ± 32	-7479 ± 24	-3134 ± 11	
Y-conformer	-22182 ± 38	-8098 ± 24	-3115 ± 5	
$\Delta(\mathrm{Y} \to \lambda)$	-521 ± 50	+619 ± 34	-19 ± 12	+79 ± 61
		PSE-2/TIP3P/PMV$_{\mathrm{max}}$		
λ-conformer	-22703 ± 32	-17756 ± 24	-3134 ± 11	
Y-conformer	-22182 ± 38	-18370 ± 24	-3115 ± 5	
$\Delta(\mathrm{Y} \to \lambda)$	-521 ± 50	+614 ± 34	-19 ± 12	+74 ± 61
		KH/SPC/E		
λ-conformer	-22703 ± 32	19159 ± 23	-3134 ± 11	
Y-conformer	-22182 ± 38	18584 ± 24	-3115 ± 5	
$\Delta(\mathrm{Y} \to \lambda)$	-521 ± 50	+575 ± 33	-19 ± 12	+35 ± 61

The difference in hydration free energy, however, is for all four 3D RISM experiments strongly positive, overcompensating ΔH_{c} and $-T\Delta S_{\mathrm{c}}$, yielding a positive ΔG. The reproducibility of this result underlines the robustness of the result that the Y-conformer is being favored in solution. The differences in the conformational free energy range between +35 ± 61 to +86 ± 61 kcal mol$^{-1,}$ with the SPC/E-NaCl solution having the lowest and the uncorrected TIP3P having the highest difference.

The PMV correction compensates for the well-known RISM inherent error of overestimating the cavity formation term of the free energy. (This is also a key factor in chapter 3.2, where this error gets compensated for by deep learning models.). Since the correcting parameters are usually gained by fitting on the solvation free energy of small molecules[177,178], the correction for a large protein like an antibody is approximated by a lower (-0.12 kcal mol^{-1} Å^{-3})[177] and upper (-0.2 kcal mol^{-1} Å^{-3})[177] bound for the corresponding parameter. With increased PMV correction, the delta between the conformers decreases slightly but remains positive, even considering the significant errors assigned to ΔG. The dependency of the result on the chosen solvent model, as shown in **Table 3.1**, is with a factor of about two not small, but since the experiment with the SPC/E-NaCl solution also yields a positive ΔG, the Y-conformer can confidently be considered favored in solution.

This finding, however, stands in contrast to the x-ray crystal structure determination in which the anti-NPRA mAb was found to crystallize in the λ-conformer. A possible explanation is indicated by the negative ΔH_c, suggesting a thermodynamic advantage of the λ-conformer in the crystallized form. While in solution the equilibrium between λ- and Y-conformer is heavily shifted towards the Y-form, upon crystallization of the antibody the λ-form is predominantly removed from the solution and is replenished from the Y-form according to the principle of Le Chatelier.

3.1.3 Summary of the subchapter

The results outlined in this subchapter show that it is possible to calculate free energy differences between distinct conformational states of large proteins on a qualitative level with 3D RISM when paired with MD simulations. The employed method of running regular simulations of only the end-states and sampling from them to calculate the hydration free energy separately via 3D RISM, makes usually hard to model substantial conformational changes accessible. This finding is also well in line with previous studies[36–39], successfully using similar methods like MM/3D RISM-KH[35] to calculate binding free energies of protein-ligand complexes.

However, the shown results also demonstrate the limitations of the method. The large statistical errors narrow its field of application to either stable, rigid systems, where fluctuations in potential energy and hydration free energy are minimal, or to situations where the free energy difference between the endpoints is expected to be sufficiently large to overcome the statistical and systematic errors.

For calculations of the binding free energies between compounds in protein-ligand complexes, the second requirement is usually not satisfied, as the differences in the binding free energy tend to be not substantial enough. While there are complexes that are sufficiently stable as a whole[36–39], it is also not uncommon to find some structural flexibility in loops and other semi-free-moving parts of the protein. However, those are often not part of the actual binding site and are only seldomly directly involved in the ligand-binding itself. Nonetheless, for methods like the one used in this subchapter, this is still a problem since such flexibilities still affect the full complex's potential energy and hydration free energy.

In this work, the problem is approached by the localization of interaction energies and solvation free energies on an atom-wise level, focusing on identifying and characterizing key interactions between protein and ligand. The for this task employed localized free energies (LFE) are discussed in detail in the following chapters.

3.2 Free energy localization on small molecules

3.2.1 Computational details

Calculation of the localized free energies

The structures used in this chapter were taken from the SAMPL5 publication[178] as well as the set of EC-RISM partial charges from the SAMPL6 publication[179] of Tielker et al.. Further, the Lennard-Jones (LJ) parameters were taken from the general Amber force field (GAFF) 1.5[98], and the calculation of the AM1BCC- and RESP-partial charges were performed with AMBER14[180] and Gaussian16[181], respectively.

To provide the necessary $\chi_{\gamma\gamma'}$-function, a specially adapted[182] SPC/E water model[170] at T=298.15 K and a density of 0.0333 Å^{-3} was used as the basis for a DRISM/HNC calculation. The λ-dependent 3D-RISM calculations were performed on a regular grid with a spacing of 0.3 Å and 128 grid points on each axis using the PSE closure[100] of order two and a convergence criterion of a difference in the residual norm of the direct correlation functions of less than 10^{-5} between two steps. The Lennard-Jones and Coulomb potentials were in full real space and scaled separately, increasing the numerical robustness while implementing softcore scaling for the Lennard-Jones potential avoids catastrophic divergences at endpoints. For the latter, a value of 0.5 for α_{LJ} was used, as it is well established in TI-simulations[109,110]. A particle-mesh-Ewald approach with Lagrangian interpolation polynomials of order eight was used for the long-range reciprocal-space Coulomb potential. Furthermore, for periodic correction of the long-range electrostatics, a monopole renormalization[183,184] was used. As for λ itself, each potential was scaled with a step size of 0.1, totaling 22 independent steps. The preceding results were not carried over to the next step since file I/O would have nullified most time savings. The actual LFE calculations were done separately and in succession to the 3D-RISM runs, and the partial results were integrated over a cubic spline interpolation with the python package SciPy 1.1.0[185], using the Fortran based quad function. The structures used for the calculations can be found in the electronic appendix under 3.2/Struc/. The force field parameters used for the LFE calculations can be found in 3.2/AM1BCC/, 3.2/RESP/, 3.2/ECRISM/, and in mnsol_atmInf.csv. Latter also contains the LFEs themselves.

Both deep learning models, 3D-CNN and MPNN, are coded in python and make extensive use of the PyTorch 1.0.1 package[136]. The Adam optimizer[186] was used during training, and the relevant hyperparameters for the two in the following used architectures were optimized with HpBandSter[187], a with Bayesian optimization extended variation of the hyperband algorithm[188].

3D CNN

The architecture of the 3D-CNN model is summarized in **Table 3.2** and was trained on the experimental solvation free energies of 502 molecules taken from the MNSol database[189] and was kept simple compared to modern computer vision (2D)[190] and concurs in this regard with those of other publications concerned with similar tasks[191,192]. An extended series of experiments showed a strong tendency to overfit for more complex models probably due to the limited size of the available dataset, which also leads to a rather aggressive usage of dropout[193]. The model was trained on 32^3 grid points over 256 full iterations over the training splits of a 5-fold-cross validation, so-called epochs, with a batch size of 64 and a learning rate of 3 10^{-4}.

Table 3.2 Architecture of the 3D-CNN model. The number of channels for the convolution steps, units in the hidden fully connected layer, the number of these layers as well as the dropout rate were subject to a hyperparameter tuning.

type	channels	kernel	stride	padding	activation	dropout
input	4					
3D-conv	8	3	1	1	relu	0.00
max pooling		2	2	0		
3D-conv	72	3	1	1	relu	0.10
max pooling		2	2	0		
3D-conv	56	3	1	1	relu	0.15
max pooling		2	2	0		
3D-conv	42	3	1	1	relu	0.10
max pooling		2	2	0		
3D-conv	16	3	1	1	relu	0.00
max pooling		2	2	0		
fully connected	384				relu	0.40
fully connected	1					

Mapping 3D coordinates to a regular grid that can be used with 3D CNN's was done using Gaussian function, effectively spreading the information of each input channel onto the grid. In contrast to assigning it to the nearest grid point, this has the advantage that the information is not carried by only a single of thousands (sometimes millions) points, which is bound to invoke problems with the 3D-CNN[192]. Mathematically this procedure is equivalent to a convolution operation typically defined as

$$(f * g)(\mathbf{r}) := \int f(\mathbf{r}')\, g(\mathbf{r} - \mathbf{r}') d\mathbf{r}'. \tag{3.4}$$

The function *f* holds here the LFE information and can be written as

$$f(\mathbf{r}) = \sum_{\alpha} \mathrm{LFE}(\mathbf{r}_{\alpha}) \delta(\mathbf{r} - \mathbf{r}_{\alpha}) \tag{3.5}$$

where δ stands for the Dirac function, which ensures that f is defined everywhere and therefore becoming integrable. The function

$$g(\mathbf{r} - \mathbf{r}_{\alpha}) = \frac{1}{\sqrt{(2\pi)^3 \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{r} - \mathbf{r}_{\alpha})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{r} - \mathbf{r}_{\alpha})\right) \tag{3.6}$$

is consequently the three-dimensional gaussian f gets folded with. Here $\boldsymbol{\Sigma}$ denotes the covariance matrix, which is one on the principal diagonal and zero anywhere else, giving the Gaussian a dimension of 1/Å^{-3}. An unnormalized Gaussian radial basis function (RBF), as used by Kuzminykh et al.[192], was explicitly not chosen since such would shift the total over the LFE. For 3D CNN models, this does not matter in most cases, since moderate additive shifts, which do not alter the relation between the original points, can be easily adapted for. However, for consistency reasons and later usage (see chapter 3.3.3), the computationally more demanding option was chosen here nonetheless.

Additionally, four rotations of 90° on each axis were applied to the grid, resulting in 64 replicas. While having some translational invariance, 3D-CNNs are not rotationally invariant and therefore require this data augmentation.

MPNN

The used implementation of the MPNN network was, like the 3D CNN model trained on solvation free energies taken from the MNSol database[189]. It uses only two message-passing cycles to keep the model complexity low. Models with more such steps were also tested exploratory but tended to overfit, again caused by the limited dataset size. The readout function was composed of four fully connected hidden layers (units: 512, 128, 256, 64), each with a ReLU activation function and a fully connected linear layer giving the final estimation of the solvation free energies. The number (bounds: 1 to 6) and dimensions (bounds: 2^6 to 2^{10}) of the individual layers were found through hyperparameter optimization with the bounds given in parentheses. Experiments implementing more complex features in the model, like a learnable message function $\boldsymbol{M_t}$ or multiple layers in the update function $\boldsymbol{U_t}$, were tested exploratory and did not improve and, to some extent, lead to even worse results (instability during training and overfitting). As before with the 3D-CNN, this is a direct effect of the limited dataset size. The resulting models were trained over 256 epochs with a batch size of 16. All learnable parameters were optimized with the Adam algorithm with enabled AMSGrad[194], a learning rate of $5\ 10^{-4}$, and a weight decay of $1 \cdot 10^{-8}$. Batch size (bounds: 8 to 128), learning rate (bounds: $1\ 10^{-5}$ to $1\ 10^{-3}$), AMSGrad[194] (True or False), and weight decay (bounds: 0 to $1\ 10^{-5}$) were subject to hyperparameter optimization with the bounds given in parentheses.

As for the data preparation, the node matrices and the adjacency matrix were padded to dimensions of 64 x 64 to have access to batch-wise training. The samples (input features, solvation free energy), were normalized to zero mean and standard deviation of one, except for the LFE and partial charges, which retained their mean.

The very limited number of samples and the simultaneously diverse nature of the dataset lead to various challenges during training and evaluation. To couple with them, three main measures were implemented. A split into training- validation- and test-set was not feasible, so a 5-fold-cross-validation (CV) approach was chosen. Seeing a high fluctuation between the individual train-test-compilations early on, the partitioning was repeated three times with different random seeds to decrease the influence of the inhomogeneous composition and increase the statistical reliability. To strengthen the statistical robustness even more and to overcome problems associated with some instabilities during training, each experiment was repeated five times with different random seeds. To compensate for rather rough training curves, a special form of early stopping was deployed, where the models were still trained for all epochs, but only those predictions of the test CV split were taken, with the lowest corresponding RMSE on the training set, respectively. For the calculation of the error metrics, the mean over the five predictions for each molecule was taken, and the three 5-fold CV splits were averaged via a quadratic mean.

3.2.2 Plausibility check of the localized free energy

An important part of the scientific method is the independent and rigorous testing of any new approach or methodology. While being derived from solid mathematics, as shown in the theory part of this work (2.2.1), experimental validation of the LFE and their partitioning is not possible. This poses a fundamental problem since the main means of testing theoretical methods, especially in chemistry, are still experiments, and unusual but potentially fruitful findings would always cast the shadow of doubt on results obtained by a method, judged solely on its plausibility. Therefore, to assess the amount and usefulness of the information carried by the LFE, they get used as an input feature for deep learning models, which were trained on experimental solvation free energies ($\Delta_{\mathrm{solv}}G^{\circ}$) of 502 small molecules. Here it is important to note that the by 3D RISM calculated excess chemical potential, and therefore the localized free energy, is formally not equivalent to the solvation free energy electronic polarization energy $\Delta_{\mathrm{solv}}E$, as shown by[178]

$$\Delta_{\mathrm{solv}}G^{\circ} = \mu^{\mathrm{ex}} + \Delta_{\mathrm{solv}}E. \tag{3.7}$$

Furthermore, μ^{ex} and the LFEs are also suffering from inherent errors of the RISM methodology, and here predominantly from the overestimation of the cavity formation.[66,67] Any model trained on LFEs as input feature to predict solvation free energies must correct those errors and compensate for the missing terms. From this, a hypothesis regarding the LFEs can be formulated: if the assumption of a correct distribution holds, it should be possible to account for inherent errors and missing terms more effectively than with any other. The idea behind it is that since the individual terms and errors are connected through the physics of the unlaying system of solute and solvent, they can be approximated best when the LFE input reflects this system most truthfully.

3.2.2.1 The LFE as input for deep learning models

To prove this hypothesis, a suitable ML model has to be selected. Since the partial charge- and LFE-calculations already require a 3D-conformation of each molecule, the utilization of three-dimensional convolutional neural networks (3D CNN) is a natural choice regarding the selection of a suitable model. Analog to the extremely popular and successful 2D version, which is used extensively in modern video and picture analysis[190], 3D-CNNs work on a regular grid and retain the spatial information of their input.

To illuminate their influence and to ensure that any observed trend is a general phenomenon and not unique to a particular method, each experiment is performed on three individual datasets, containing AM1BCC-, RESP- and EC-RISM-partial charges. A dependency on the 3D-RISM μ^{ex} by the method for partial charge calculation is undoubtedly expected. However, it is unclear how the different sets of partial charges affect the localization and the prediction capabilities of the models.

Table 3.3 The predictive capabilities of the 3D convolutional neural network (3D-CNN) is evaluated on the MNSol data set[189], posterior separated in neutral and single positive and negative charged molecules, treated with three different methods for partial charge calculation. The used metrics are **R**oot **M**ean **S**quare **E**rror (RMSE), **M**ean **A**bsolute **E**rror (MAE), coefficient of determination (R^2), as well as the slope (m) and y-intercept (b) of a linear fit on the predicted solvation free energies in kcal/mol. All numbers are a (where applicable quadratic) average over three different split five-fold cross-validations, each repeated five times. The raw data can be found in the electronic appendix under 3.2/mnsol_3DCNN_modelResults.csv and collectively under 3.2/mnsol_molInf.csv. The values for the three different CV-splits are also listed in **Table 6.1**.

	3D-CNN				
ML model	RMSE	MAE	R^2	m	b
AM1BCC					
All	2.66±0.32	1.50±0.23	0.99±0.00	0.98±0.00	-0.02±0.05
Neutral	1.41±0.18	0.93±0.15	0.89±0.00	0.90±0.01	-0.29±0.05
Anions	5.30±0.62	3.71±0.54	0.77±0.01	0.88±0.01	-7.68±0.74
Cations	3.56±0.43	2.40±0.37	0.78±0.01	0.81±0.01	-11.54±0.69
RESP					
All	2.42±0.32	1.45±0.25	0.99±0.00	0.97±0.00	-0.11±0.07
Neutral	1.41±0.22	0.96±0.19	0.88±0.00	0.91±0.01	-0.35±0.06
Anions	4.47±0.55	3.12±0.45	0.84±0.01	0.90±0.01	-5.54±0.87
Cations	3.62±0.45	2.55±0.38	0.78±0.01	0.83±0.01	-10.25±0.76
EC-RISM					
All	2.80±0.29	1.50±0.20	0.99±0.00	0.98±0.00	-0.11±0.03
Neutral	1.35±0.16	0.92±0.13	0.90±0.00	0.90±0.01	-0.43±0.04
Anions	4.87±0.56	3.53±0.46	0.78±0.00	0.84±0.01	-11.36±0.67
Cations	5.35±0.44	2.72±0.34	0.66±0.01	1.01±0.02	2.02±1.10

The data in **Table 3.3** shows an advantage of RESP-charges in 3D CNN models over both other tested methods for all observed metrics, but most prominent in the RMSE between experimental and predicted solvation free energies. Performance-wise they are followed by the semi-empirical quantum-mechanic AM1BCC- and only then by the high-level EC-RISM-charges. While the trend between the partial charge models is mostly reproduced by the MPNN architecture (discussed later, see **Table 3.4**), the assigned errors should kept in mind for the comparison.

A sub-par performance of AM1BCC-charges is, to a certain extent, expected, especially in comparison with RESP-charges, as their drawbacks in molecular dynamics simulations[195–

[197] and poor recreation of high-level QM observables[198] were shown before. With EC-RISM having the highest level of theory of the here tested partial charge models, explaining the reduced predictive capabilities of those models trained with EC-RISM-charges must have different reasons. As described earlier, the task of the ML models is to compensate for missing terms and, more importantly, to correct for RISM inherent errors from the provided input features like partial charges and LFEs. In the EC-RISM-based data, this connection is presumably more complex than in the other two partial charge models due to the high level of theory with which the underlying system is characterized. The data in **Table 3.3** suggest that the capabilities of the neural networks to draw this more complex connection from the input over the correcting terms to the solvation free energy is limited, yielding suboptimal results compared to RESP-charges-based models.

For example, a higher polarization of the molecules, like it is the case for EC-RISM-charges, leads to a lower μ^{ex} and higher electronic energies, partly canceling each other out in their effects. But more importantly, the more complex electronic structure influences the solvent distribution and thereby the cavity formation accordingly. As discussed in chapter 3.1, this plays a significant part in the RISM inherent errors, usually expressed and corrected via the partial molar volume (PMV), as shown among others by Tielker et al.[66,68–70]. Since the input for the ML-models contains only very limited information about the underlying quantum mechanics, being able to estimate these terms in addition to the already provided μ^{ex} is a critical factor for an accurate prediction. As EC-RISM employs high-level QM for the calculations of the $\Delta_{\mathrm{solv}}\boldsymbol{G}°$ underlying terms, one can assume their estimation is the most challenging compared side by side with the other methods and is, at least, for the 3D CNN models, a bottleneck.

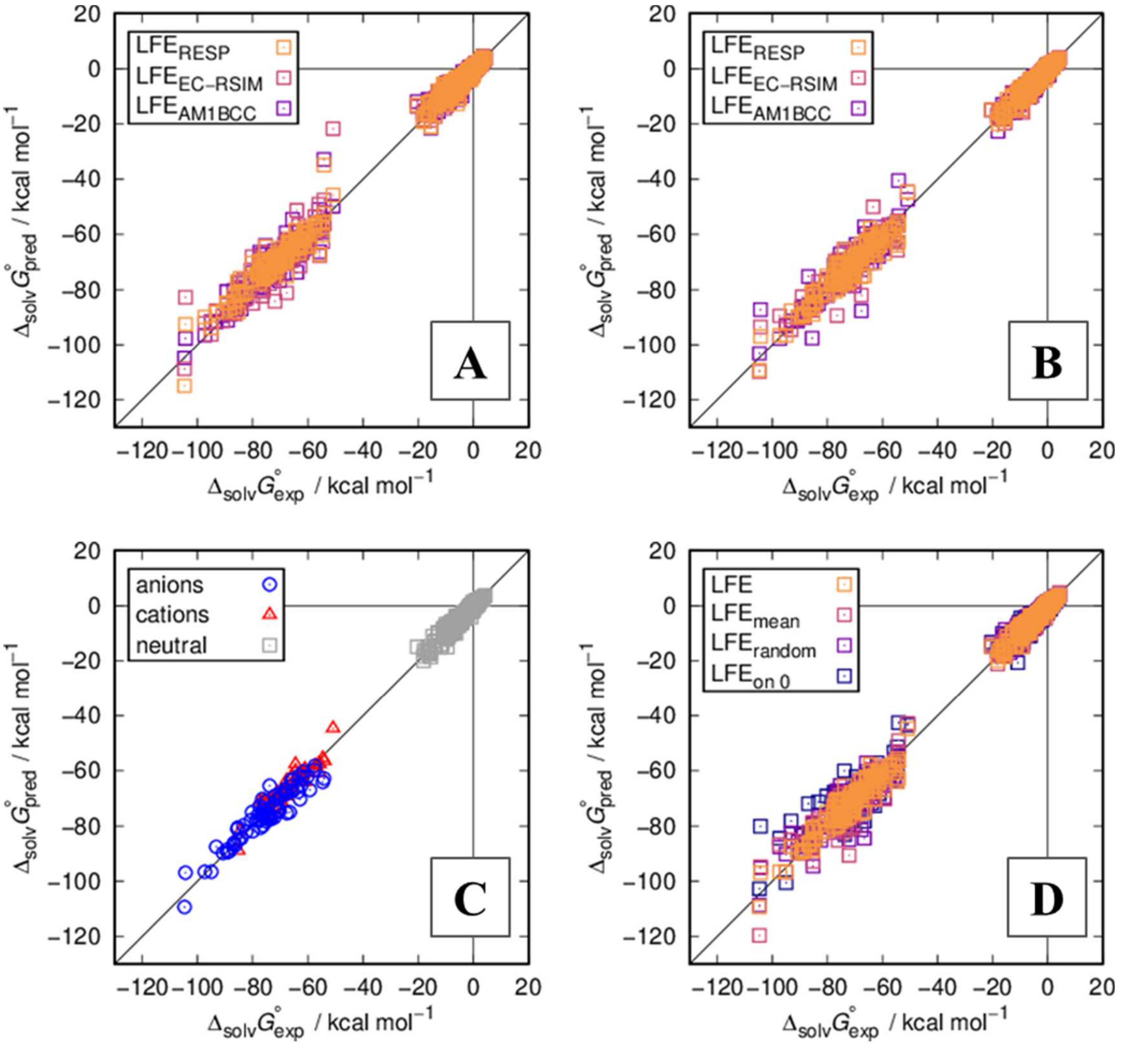

Figure 3.3 All four plots show predicted versus experimentally measure solvation free energies from the MNSol database for water as solvent (exemplary CV). A: Comparison between different sets of partial charges used in LFE calculation and as part of the input for the 3D-CNN model. B: Comparison between different sets of partial charges used in LFE calculation and as part of the input for the MPNN model. C: Results of the MPNN model with full LFE data and RESP charges, colored with respect to molecular charges. D: Comparison between MPNN models trained with varying LFE data. The raw data can be found in the electronic appendix under (A) 3.2/mnsol_3DCNN_modelResults.csv, (B) 3.2/mnsol_MPNN_modelResults.csv, (C) 3.2/mnsol_MPNN_modelResults.csv, and (D) 3.2/mnsol_MPNN_LFEtests_modelResults.csv. They can also be found collectively under 3.2/mnsol_molInf.csv.

From **Figure 3.3** panel C as well as from **Table 3.3**, it is evident that the predictions of $\Delta_{solv}G^{\circ}$ for neutral molecules are far better than for charged ones. Especially the substantial discrepancy between RMSE and MAE for anions and cations compared to neutral molecules displays a high variance in all models and therefore a considerable challenge. The most likely explanation is the strong imbalance between charged (112) and neutral (390) molecules in the used dataset, which is especially tough for the ML-models. Also, the unavoidable experimental uncertainty for charged species is increasing the problem. Another reason for the higher values in both error metrics is the significantly higher absolute values for the solvation free energies for charged molecules. As the high errors for charged molecules are most likely a direct cause of the data set, one could come to the conclusion to omit them and turn to bigger datasets like FreeSolv[199] containing only neutral molecules. However, since charged species are not unusual in aqueous solutions and therefore are indeed relevant (for example, in drug discovery) and the LFE are supposed to be tested also for those cases, it is important not to neglect them.

After all, the obtained models for all three partial charge sets predict the solvation free energy of a given molecule with acceptable accuracy. Nonetheless, during the preprocessing and training phase, some distinct drawbacks and disadvantages of 3D-CNNs become obvious. Due to the three-dimensionality of the grids and the rather intricate strategy to increase the statistical robustness of our results, it becomes quite cumbersome and time-consuming to work with the many large data files.

This raised the question, whether the native encoding of 3D-conformations of the 3D CNN is a significant advantage or approaches omitting this can lead to similar results. Especially when some input features already contain information about the three-dimensional structure of the molecule like partial charges or LFEs themselves, this seems increasingly plausible. So-called message passing neural networks (MPNNs) are proven to cope well with a variety of property prediction tasks[79,147,200,201] and are therefore a plausible choice to complement the experiments with the 3D CNN models.

Table 3.4 The predictive capabilities of the message passing neural network (MPNN) is evaluated on the MNSol data set[189], posterior separated in neutral and single positive and negative charged molecules, treated with three different methods for partial charge calculation. The used metrics are **R**oot **M**ean **S**quare **E**rror (RMSE), **M**ean **A**bsolute **E**rror (MAE), coefficient of determination (R^2), as well as the slope (m) and y-intercept (b) of a linear fit on the predicted solvation free energies in kcal/mol. All numbers are a (where applicable quadratic) average over three different split five-fold cross-validations, each repeated five times. The raw data can be found in the electronic appendix under 3.2/mnsol_MPNN_modelResults.csv and collectively under 3.2/mnsol_molInf.csv. The values for the three different CV-splits are also listed in **Table 6.2**.

			MPNN		
ML model	RMSE	MAE	R^2	*m*	*b*
AM1BCC					
All	2.43±0.30	1.28±0.21	0.99±0.00	1.00±0.00	-0.11±0.02
Neutral	1.34±0.17	0.82±0.14	0.90±0.00	0.93±0.01	-0.30±0.04
Anions	5.02±0.59	3.23±0.46	0.77±0.02	0.83±0.01	-12.56±1.09
Cations	2.59±0.35	1.72±0.30	0.87±0.01	0.89±0.01	-7.02±0.70
RESP					
All	**1.94±0.24**	**1.19±0.19**	1.00±0.00	1.00±0.00	-0.10±0.04
Neutral	1.28±0.16	0.82±0.14	0.91±0.00	0.94±0.00	-0.31±0.03
Anions	3.52±0.40	2.54±0.35	0.88±0.00	0.86±0.00	-10.56±0.33
Cations	2.51±0.38	1.92±0.33	0.88±0.00	0.95±0.01	-3.29±0.77
EC-RISM					
All	2.23±0.27	1.26±0.19	0.99±0.00	1.00±0.00	-0.09±0.04
Neutral	1.29±0.16	0.76±0.13	0.91±0.00	0.95±0.01	-0.26±0.03
Anions	4.31±0.51	3.09±0.38	0.83±0.00	0.89±0.01	-8.65±0.42
Cations	2.96±0.35	2.22±0.30	0.84±0.01	0.91±0.01	-5.89±0.70

As before, the RESP-charges seem to be the sweet spot for polarization and complexity as they once again outperform both other methods. Where the 3D CNN models struggled to make use of the high-level EC-RISM data due to its complexity and even fall behind the low-level AM1BCC charges, eventhough not by much, the MPNN models are able to compensate and reverse the trend. The effect is far less dramatic in the MAE where both methods lead to similar numbers with 3D-CNN as well as with MPNN models. This speaks for a reduced variance and an improved handling of outliers, i.e., under represented molecule types, or in short, a higher generalization capability. This, in turn, could be a cause of the more efficient usage of pre-calculated information in the input.

While it cannot be ruled out that the improved performance of the MPNN-models could be a result of a better-suited choice of architecture and hyperparameters, it nonetheless appears that the explicit encoding of the 3D-conformations is not a significant advantage in this particular instance, under the light of the described trends. Paired with the easier data handling and faster training of the MPNN approach, the 3D CNN models were omitted for further analysis for those reasons. Regarding the influence of the partial charge model, it has to be noted that, although the performance differences between the charge models can not be neglected, they are, in general, not large in the here presented experiments. Therefore, all three will be included in the following plausibility check of the LFEs, outlining trends and increasing the robustness of the finidings.

3.2.2.2 Plausibility of the LFE-distribution

Despite the elucidating results from the investigation of the two deep learning approaches, the validation of the earlier formulated hypothesis is still pending. Therefore, a series of experiments with different amounts of LFE information was conducted. Four sets of input data from each of the three sets of partial charges and the corresponding LFE are trained with the MPNN models, following the same protocol as before. The following four categories describe the range of test scenarios:

- LFE: This is the reference and the LFE as it is computed; There is no reason to assume that there is more than one correct distribution. Therefore, models trained with this input should perform best.
- LFE_{mean}: The mean over all LFE-values if a molecule is assigned to each atom. This is the equivalent to not localize at all and carries the information of a standard 3D RISM calculation uniformly spread over the molecule. This set of input data is expected to perform better than no thermodynamic input[202], but worse than the standard LFE.
- $\text{LFE}_{\text{random}}$: The calculated LFE-values are allocated to randomly chosen atoms. Similar to LFE_{mean}, the sum over all atoms equals the standard 3D RISM free energy. Following the hypothesis, models trained with this input should perform worse than those trained with the correct LFE distribution as well as the LFE_{mean} data.
- $\text{LFE}_{\text{on 0}}$: All values in the LFE input channel are set to 0. This is the baseline model, which demonstrates how much performance originates from the model itself and the remaining input features. Without any thermodynamic information and only little knowledge about the conformation via the partial charges, this is expected to perform the worst.

Table 3.5 To assess the quality of inherent information of the LFE, the MPNN model was fitted with diminishing amounts of data on three levels: the mean over the full molecule is assigned to each atom (LFE mean), the calculated LFE values are assigned to random atoms (LFE randomized), the LFE channel is set to zero for each molecule (LFE set to 0). As a reference, the results utilizing the original data (LFE as calculated) are given once again. The raw data can be found in the electronic appendix under 3.2/mnsol_MPNN_LFEtests_modelResults.csv and collectively under 3.2/mnsol_molInf.csv. The values for the three different CV-splits are also listed in **Table 6.3**.

	RMSE	MAE	R^2	*m*	*b*
AM1BCC					
LFE as calculated	2.43±0.30	1.28±0.21	0.99±0.00	1.00±0.00	-0.11±0.02
LFE mean	2.65±0.26	1.47±0.21	0.99±0.00	1.00±0.00	-0.14±0.05
LFE randomized	2.78±0.33	1.65±0.25	0.99±0.00	0.99±0.00	-0.11±0.03
LFE set to 0	3.29±0.48	1.88±0.33	0.99±0.00	0.98±0.00	-0.23±0.05
RESP					
LFE as calculated	1.94±0.24	1.19±0.19	1.00±0.00	1.00±0.00	-0.10±0.04
LFE mean	2.44±0.26	1.44±0.21	0.99±0.00	1.00±0.00	-0.12±0.03
LFE randomized	2.75±0.32	1.65±0.25	0.99±0.00	1.00±0.00	-0.10±0.03
LFE set to 0	3.32±0.49	1.89±0.33	0.99±0.00	0.98±0.00	-0.23±0.05
EC-RISM					
LFE as calculated	2.23±0.27	1.26±0.19	0.99±0.00	1.00±0.00	-0.09±0.04
LFE mean	3.41±0.40	1.55±0.22	0.99±0.00	0.99±0.00	-0.16±0.05
LFE randomized	2.83±0.36	1.60±0.25	0.99±0.00	0.99±0.00	-0.16±0.04
LFE set to 0	3.43±0.63	1.83±0.36	0.98±0.00	0.98±0.00	-0.28±0.05

The corresponding results can be found in **Table 3.5** and in panel D of **Figure 3.3**. The slope and intersect of a regression line, fitted on calculated vs. experimental values, is almost equal for all experiments. Usually, such uniform results are not overly conclusive; here however, they lead to a valuable insight. In situations where the information in the input data is insufficient or the model itself cannot find a suitable function connecting input and output, often a slope of less than one and an intersect approaching the mean of y in the training set can be found. Given that even the fits on the $\text{LFE}_{\text{on }0}$-results are almost perfect, the MPNN-model can draw this general connection even without thermodynamic input. Nevertheless, these experiments show the largest RMSE values within the series, which are particularly sensitive to outliers. This gives the impression that those MPNN-models, trained without thermodynamic input, are valid estimators of $\Delta_{\text{solv}}G°$, but struggle with chemistry under-represented in the training set. This observation is underlined by the significantly higher statistical errors for the experimetns with reduced thermodynamic information. This is an unsurprising limitation and is usually overcome by adding more data. However, this is often hard to come by, and the lack of it is in chemistry-related, real-world deep learning applications a significant, sometimes insurmountable obstacle, which is already hindering progress and narrows the paths of advancements. The improved handling of outliers of those models trained with thermodynamic input shows that physical input can reduce this need and increase generalizability.

The above introduced experiments demonstrate this in three gradations, where the individual LFE-values of the input modes LFE$_{\text{random}}$ and LFE$_{\text{mean}}$ share two imported properties. On the one hand, they do correspond to a solvent distribution not found by 3D RISM, and on the other, their sum is still equal to μ^{ex} calculated with 3D-RISM. The latter is probably responsible for the improved RMSE and MAE values compared to models trained with no LFE information. The former indicates how important the correctness of the presented physics is to the models. Since the force field parameters are unchanged in both cases, the $\delta u_{\alpha\gamma}(\mathbf{r}, \lambda)/\delta\lambda$ term in (2.33) is also untouched, which leaves only the g-function as modulating and differing term between both cases. By trying to imagine a g-function generating the permuted/altered LFE-values, one must conclude that it cannot be a single function for all atoms but rather a unique one for each. This stands in strong contrast to the physical basis, which explains not only the worse performance compared to the standard LFE-distribution but also the discrepancy between LFE$_{\text{random}}$ and LFE$_{\text{mean}}$. For most molecules, the hypothetical g-functions in the LFE$_{\text{mean}}$-set are closer to each other as those in the LFE$_{\text{random}}$-set, and therefore nearer to a physically plausible representation, which leads to more accurate predictions (except for the RMSE in the EC-RISM partial charge set for unknown reasons).

This thought experiment shows that an arbitrary distribution of the free energy onto atoms will, in almost all cases, break the underlying physics. The numbers in **Table 3.5** illustrate the significant drawbacks associated with this for all three partial charge models and confirm the hypothesis of the calculated LFE being a plausible distribution. The fact that models trained with those datasets outperform all other models confirms the assumption that the in this work introduced LFE method indeed yields the correct distribution of the full free energy.

3.2.2.3 Summary of the subchapter

This subchapter introduced the LFE method as a new way of localizing the free energy obtained by 3D-RISM onto individual sites. In the absence of a direct experimental method to measure the contribution of individual atoms to the total free energy, a deep learning approach was used to validate the LFEs and thereby could also demonstrate the benefits of physical input for such models. It could be shown that carefully designed input features can improve the performance of (comparably) simple models trained with challenging small and heterogeneous datasets. This can be a good solution to the notorious problem in chemistry of having too little data to train models with higher complexity and predictive power. The described findings show a clear advantage of those models trained with LFE values over those trained with the reference input. Taken together with the sound mathematical derivation and the clear trend in the deep learning experiments, it can be said that the localization of the free energy is indeed valid. This subchapter demonstrates only a small set of possible use cases for the LFE, but the concept can be of great use in many other fields of application. Since the approach has no theoretical limit to the size of the treated molecules, other than computational resources, it can even be applied to protein-ligand complexes and help to rationalize experimental trends in compound series and thereby guide the further development and design of new drugs. The plausibility being demonstrated, the application of the LFEs to such systems is the concern of the next chapter.

3.3 Applying FED and LFE to protein-ligand complexes

Thrombin is produced from prothrombin in the blood coagulation pathway and activates fibrinogen (I) to fibrin (Ia) which in turn, together with the also activated XIIIa, forms the cross-linked fibrin clot. It is a common part of the intrinsic pathway and the extrinsic pathway, why it is of particular interest for therapeutic anti-clogging drugs, as inhibition of it can stop both. This is also true for factor X, respectively Xa, for which several inhibiting drugs like Apixaban, Edoxaban, and Rivaroxaban are available.

Commonly used drugs for treatment for patients with an increased risk of strokes, pulmonary embolism, deep vein thrombosis, atrial fibrillation, and other thrombosis are heparin and warfarin, influencing the thrombin availability indirectly. Unfortunately, those drugs suffer from multiple issues. For example, a particularly severe side effect of heparin is the Heparin-induced thrombocytopenia (HIT), causing thrombosis due to an immune reaction to a complex formed with the platelet factor 4[203,204]. The therapeutic usage of warfarin, on the other hand, is problematic for its drug and food interactions, and dosage must be monitored strictly[205,206]. Therefore, several alternative drugs, so-called Direct thrombin inhibitors (DTI)[207] are becoming more widely used lately[208]. This including, among others, Dabigatran[209], Bivalirudin[210], and Argatroban[211], some of which are even reversible in their therapeutic effect. The binding site of thrombin is shown in **Figure 3.4**.

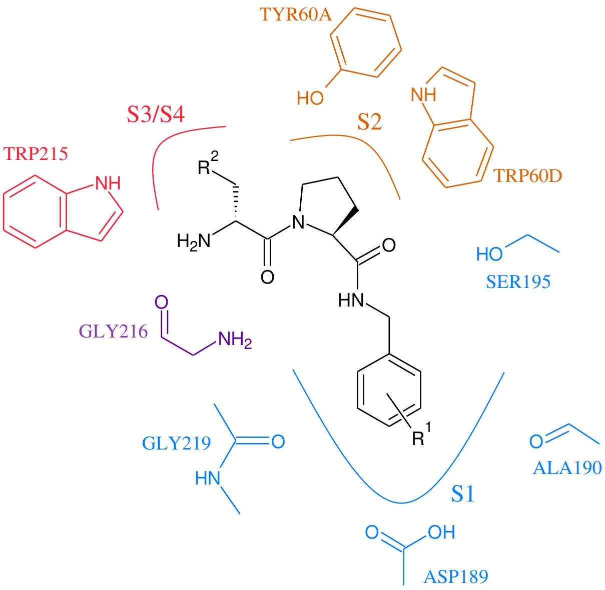

Figure 3.4 Schematic representation of the thrombin binding site after Baum et al.[212]. Substituents R^1 and R^2 refer to the variations in the compound series[80] discussed in this work.

Nonetheless, these drugs are also not without issues. For example, they suffer from narrow therapeutic windows, fast degradation, liver[213] and kidney toxicity[214], and unwanted side effects like bleedings and gastritis[215,216]. For those reasons, the search for new DTIs is still ongoing, and the development of new candidate compounds is of outstanding academic and industrial interest.

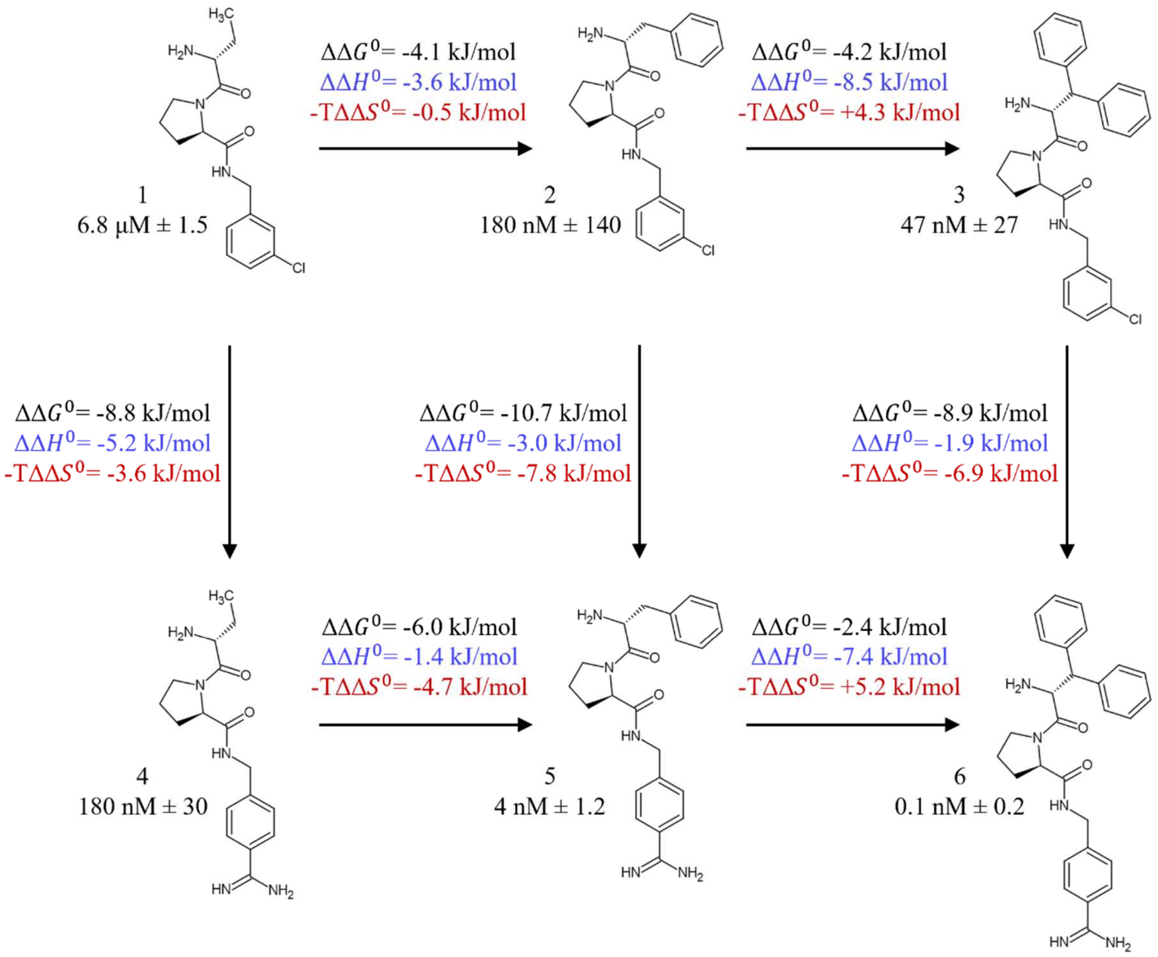

Figure 3.5 Structure of all six thrombin inhibitors[80] used in this work. The inhibitory potency of the compounds towards human thrombin is given as kinetic inhibition constant K_i. The corresponding relative thermodynamic properties, listed over the arrows, were measured via isothermal titration calorimetry (ITC). The figure is modeled after Baum et al.[80].

The following chapter applies the localized free energy and free energy derivatives methods to a series of thrombin inhibitors as an example of their applicability in computer-aided drug discovery and design, focusing on their potential to identify important interaction sites between protein and ligand as well as promising features for optimization. First, the concept of atom-wise LFEs is explored, and its contributions are discussed in the context of protein-ligand complexes. Leading from the learned lessons, an alternative visualization method gets proposed and its key features explained. With a solid understanding of the LFEs and their interpretation, they get applied to the full series of thrombin inhibitors. The chapter is concluded considering likely experimental uncertainties in crystallographic complex structures and their effect on the LFEs.

3.3.1 Thermodynamic background

The formation of complex AB can formally be written as

$$\mathrm{A_{aq}} + \mathrm{B_{aq}} \rightarrow \mathrm{AB_{aq}} \tag{3.8}$$

where A and B are placeholders for the host and guest molecules, and AB stands for the complex of both consequently. The subscript "aq" specifies the medium in which the binding occurs as an aqueous environment. While this is the defacto standard for biologically relevant processes like protein-ligand binding, the here described scheme is not limited to that. It is also valid for any other solvent that may be of interest in different research fields like for example, chemical synthesis and polymer production.

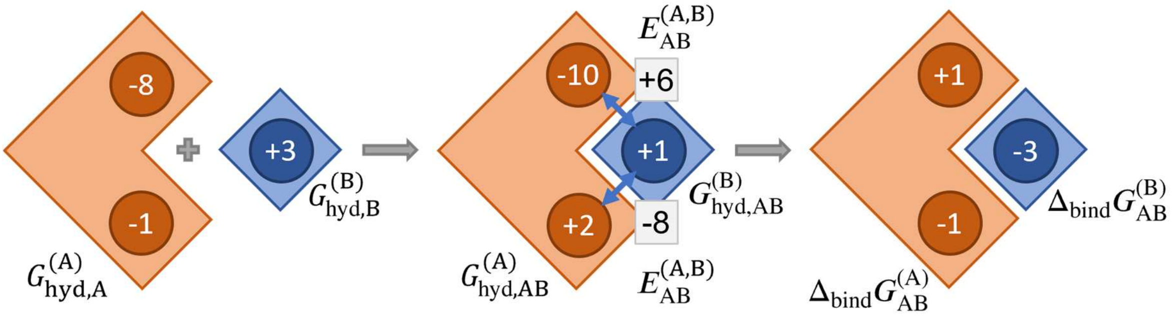
-8
-1
$G^{(A)}_{\mathrm{hyd,A}}$
+3
$G^{(B)}_{\mathrm{hyd,B}}$
$E^{(A,B)}_{\mathrm{AB}}$
-10
+6
+1
+2
-8
$G^{(B)}_{\mathrm{hyd,AB}}$
$G^{(A)}_{\mathrm{hyd,AB}}$
$E^{(A,B)}_{\mathrm{AB}}$
+1
-3
-1
$\Delta_{\mathrm{bind}}G^{(B)}_{\mathrm{AB}}$
$\Delta_{\mathrm{bind}}G^{(A)}_{\mathrm{AB}}$

Figure 3.6 Idealized approach to calculate localized binding free energies on protein (orange; denoted A) and ligand (blue; denoted B). The circles in both binding partners stand for individual sites, while the in the diagram shown numbers give example localized free energies and interaction energies. The labels below the pictograms show to which terms the numbers are associated (left: localized hydration free energy of both binding partners in free solution, middle: localized hydration free energy of the complex and interaction energy, right: localized binding free energy).

The equation used in this chapter to calculate the binding free energy $\Delta_{\mathrm{bind}}G$, defined as the difference in the total Gibbs free energy occurring during complex formation, is given by

$$\Delta_{\mathrm{bind}}G_{\mathrm{AB}} = G_{\mathrm{hyd,AB}} - G_{\mathrm{hyd,A}} - G_{\mathrm{hyd,B}} + E_{\mathrm{AB}}. \tag{3.9}$$

The hydration free energy G_{hyd}, is here treated as equivalent to the excess chemical potential μ^{ex} as calculated by 3D RISM (see chapter 3.1). The interaction energy E_{AB} is modeled as the non-bonded force field terms. This is also a standard, non-localized version of the in the following used equations for the localized binding free energy. In contrast to chapter 3.1 the complex and both binding partners are here treated as rigid bodies, which means in this case that only one structure is used for the calculations of all contributing terms. With this, all solute entropy terms are neglected, and conformational fluctuations cannot be represented. Furthermore, by assuming the same ligand and protein conformer in free solution as in complex, through relaxation process induced changes in the intramolecular energy are not considered. The same potential discrepancy between conformers in solution and in complex also affects the hydration free energy for the unbound molecules. Those approximations inevitably influence the results discussed in this chapter. However, they are made as a compromise in the benefit of speed and computational requirements. Especially with the goal of the localization of $\Delta_{\text{bind}}G$ in mind, considering those terms would increase the time and compute demands significantly (see the approach taken in chapter 3.1).

For a full localization of the binding free energy, each of the contributing terms must be localized. For G_{hyd} this is done with the LFE approach, discussed in chapter 3.2. The atom-wise calculation of E_{AB} is straight forward, as it is simply the sum over all non-bonded force field interaction of an atom with all atoms of the binding partner. The actual localized binding free energy calculation is done separately for A and B but follows the same rules as the standard, non-localized equation and are given by

$$\Delta_{\text{bind}}G_{\text{AB}}^{(\text{A})} = G_{\text{hyd,AB}}^{(\text{A})} - G_{\text{hyd,A}}^{(\text{A})} + \frac{E_{\text{AB}}^{(\text{A})}}{2} \tag{3.10}$$

and

$$\Delta_{\text{bind}}G_{\text{AB}}^{(\text{B})} = G_{\text{hyd,AB}}^{(\text{B})} - G_{\text{hyd,B}}^{(\text{B})} + \frac{E_{\text{AB}}^{(\text{B})}}{2} \tag{3.11}$$

correspondingly. The superscripts (A) and (B) denote the localization of the property on the binding partners. The connection of (3.10(3.10) and (3.11) to (3.9) is given by

$$G_{\mathrm{hyd,A}} = \sum_i G^{(\mathrm{A})}_{\mathrm{hyd,A},i}, \tag{3.12}$$

$$G_{\mathrm{hyd,B}} = \sum_j G^{(\mathrm{B})}_{\mathrm{hyd,B},j}, \tag{3.13}$$

$$G_{\mathrm{hyd,AB}} = \sum_i G^{(\mathrm{A})}_{\mathrm{hyd,AB},i} + \sum_j G^{(\mathrm{B})}_{\mathrm{hyd,AB},j}\text{, and} \tag{3.14}$$

$$\Delta_{\mathrm{bind}} G_{\mathrm{AB}} = \sum_i \Delta_{\mathrm{bind}} G^{(\mathrm{A})}_{\mathrm{AB},i} + \sum_j \Delta_{\mathrm{bind}} G^{(\mathrm{B})}_{\mathrm{AB},j}. \tag{3.15}$$

Only the contributions of both $E^{(\mathrm{A})}_{\mathrm{AB,loc}}$ and $E^{(\mathrm{B})}_{\mathrm{AB,loc}}$ must be treated differently as

$$E_{\mathrm{AB}} = \sum_i E^{(\mathrm{A})}_{\mathrm{AB},i} \text{ and } E_{\mathrm{AB}} = \sum_j E^{(\mathrm{B})}_{\mathrm{AB},j}. \tag{3.16}$$

To avoid double-counting a factor of ½ is introduced in (3.10) and (3.11).

3.3.2 Computational details

Preparation of the protein-ligand complex structures

The following discussed structures are taken from Baum et al.[80] and can be found in the Protein Data Base (PDB) under the codes 2ZDA, 2ZO3, 3DHK, 2ZFP, 2ZGX, and 2ZC9. With a resolution between 1.58 Å (2ZDA) and 1.80 Å (2ZGX) (the only exception is 2ZFP with 2.25 Å), they can be considered of high quality, especially within the requirements of this work.

The structure of all six in more detail discusses complexes show only little variance, as shown in **Figure 3.7**. There is only little disagreement between the cartoon representations, and even key amino acids in the binding site are crystallized in almost the same position in all six complexes, despite structural differences in the corresponding ligands.

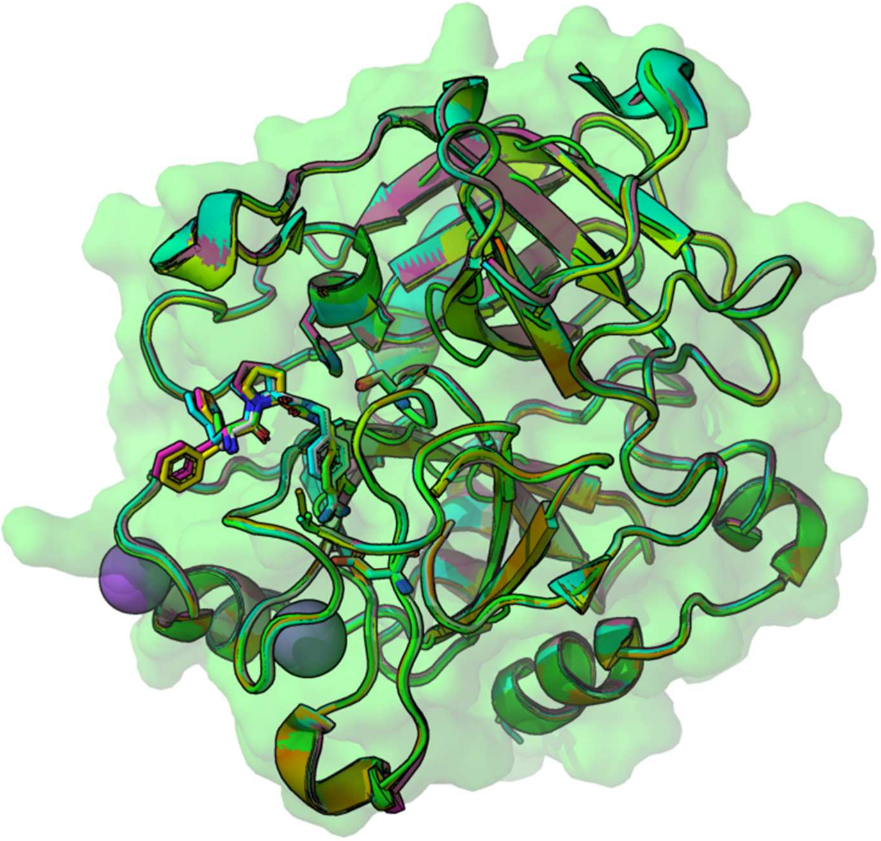

Figure 3.7 Superposition of cartoon representations of the original PDB structures of 2ZDA, 2ZO3, 3DHK, 2ZFP, 2ZGX, and 2ZC9.[80] Ligands and selected amino acids are shown as sticks and the semi-transparent surface is calculated from the 2ZFP structure.

Nonetheless, as it is common in x-ray structures, the most flexible parts of the proteins could not be resolved and create gaps within the structure. For molecular dynamics simulations, such unwanted gaps within a structure are a severe problem, which must be resolved in preprocessing. 3D RISM is not as strict in this regard, and calculations are still possible even without a complete protein even though missing amino acids (AA) influence the solvent distribution and, therefore, the obtained results. Luckily, the observed gaps are far from the binding site and are unlikely to hinder a reliable characterization of the protein-ligand binding process. Nonetheless, for comparison between the individual complexes, the defects have to be unified. This could have been achieved easily by deleting those amino acids not present in all structures or, more cumbersome, adding all missing residues. Yet, the strategy used here was to model only those residues, for which a template was present in at least one of the other here used structures. This can be seen as a compromise between the two extremes and reduces the risk of introducing potential artifacts, which could occur by modeling additional amino acids from homology models or widening already existing gaps. This approach had only one exception of asparagine 14L of chain L in 2ZDA being deleted from the sequence instead of being modeled in the other structures, as it was only present in 2ZDA.

The modeling was done with the software package MODELLER version 9.14[217,218]. A summary of the sequence provided by the x-ray structures as well as the used reference sequence to which missing amino acids were modeled is shown in **Table 3.6**. Each structure was modeled with all remaining structures used as templates to inform the modeling process with as much experimental evidence as possible. This, and the strategy not to model structurally unknown amino acids, rendered an additional force field optimization unnecessary and potentially even contra-productive. Such an optimization would alter not only the coordinates of the added residues but also those of the already experimentally resolved ones. However, as the approach in this chapter uses only one conformation per complex, as described earlier, it relies on the latter. The distance between the nearest ligand atom and the last C-α carbon before the closest gab in the protein is about 11.6 Å, affecting the analysis only minimally. Furthermore, since the gabs are the same for all discussed complexes such artifacts are mostly canceled out in direct comparison. For unknown reasons, the modeling of single amino acids also introduced artifacts in some of the presumably not modeled amino acids, leading to clustered oxygen and hydrogen atoms of sidechains. Those artifacts did not occur in direct contact with any of the ligands. The raw data and modeling results can be found in the electronic appendix under 3.3/Param/Model_struc/.

Table 3.6 Sequences used for modeling the protein structures with the PDB codes listed under Struct. The with Ref. titled sequence lists all residues of the used thrombin variant. Residues not present in the original structure, but part of the set of those to be modeled are highlighted in red, while the one deleted from 2ZDA is shown in green. Chain breaks are indicated by forward slashes. The raw data and modeling results can be found in the electronic appendix under 3.3/Param/Model_struc/.

Struct.	Sequence
Ref.	TFGSGEADCGLRPLFEKKSLEDKTERELLESYIDGR/IVEGSDAEIGMSPWQVMLFRKSPQELLCGASLISDRWV
2ZDA	-----EADCGLRPLFEKKSLEDKTERELLESYI---/IVEGSDAEIGMSPWQVMLFRKSPQELLCGASLISDRWV
2ZO3	------ADCGLRPLFEKKSLEDKTERELLESYI---/IVEGSDAEIGMSPWQVMLFRKSPQELLCGASLISDRWV
2ZC9	-----EADCGLRPLFEKKSLEDKTERELLESYI---/IVEGSDAEIGMSPWQVMLFRKSPQELLCGASLISDRWV
3DHK	------ADCGLRPLFEKKSLEDKTERELLESYI---/IVEGSDAEIGMSPWQVMLFRKSPQELLCGASLISDRWV
2ZFP	-----EADCGLRPLFEKKSLEDKTERELLESYI---/IVEGSDAEIGMSPWQVMLFRKSPQELLCGASLISDRWV
2ZGX	------ADCGLRPLFEKKSLEDKTERELLESYI---/IVEGSDAEIGMSPWQVMLFRKSPQELLCGASLISDRWV
Ref.	LTAAHCLLYPPWDKNFTENDLLVRIGKHSRTRYERNIEKISMLEKIYIHPRYNWRENLDRDIALMKLKKPVAFSD
2ZDA	LTAAHCLLYPPWDKNFTENDLLVRIGKHSRTRYERNIEKISMLEKIYIHPRYNWRENLDRDIALMKLKKPVAFSD
2ZO3	LTAAHCLLYPPWDKNFTENDLLVRIGKHSRTRYERNIEKISMLEKIYIHPRYNWRENLDRDIALMKLKKPVAFSD
2ZC9	LTAAHCLLYPPWDKNFTENDLLVRIGKHSRTRYERNIEKISMLEKIYIHPRYNWRENLDRDIALMKLKKPVAFSD
3DHK	LTAAHCLLYPPWDKNFTENDLLVRIGKHSRTRYERNIEKISMLEKIYIHPRYNWRENLDRDIALMKLKKPVAFSD
2ZFP	LTAAHCLLYPPWDKNFTENDLLVRIGKHSRTRYERNIEKISMLEKIYIHPRYNWRENLDRDIALMKLKKPVAFSD
2ZGX	LTAAHCLLYPPWDKNFTENDLLVRIGKHSRTRYERNIEKISMLEKIYIHPRYNWRENLDRDIALMKLKKPVAFSD
Ref.	YIHPVCLPDRETAASLLQAGYKGRVTGWGNLKETWTANVGKGQPSVLQVVNLPIVERPVCKDSTRIRITDNMFCA
2ZDA	YIHPVCLPDRETAASLLQAGYKGRVTGWGNLKET-------GQPSVLQVVNLPIVERPVCKDSTRIRITDNMFCA
2ZO3	YIHPVCLPDRETAASLLQAGYKGRVTGWGNLKE--------GQPSVLQVVNLPIVERPVCKDSTRIRITDNMFCA
2ZC9	YIHPVCLPDRETAASLLQAGYKGRVTGWGNLKET-------GQPSVLQVVNLPIVERPVCKDSTRIRITDNMFCA
3DHK	YIHPVCLPDRETAASLLQAGYKGRVTGWGNLKE--------GQPSVLQVVNLPIVERPVCKDSTRIRITDNMFCA
2ZFP	YIHPVCLPDRETAASLLQAGYKGRVTGWGNLKET-------GQPSVLQVVNLPIVERPVCKDSTRIRITDNMFCA
2ZGX	YIHPVCLPDRETAASLLQAGYKGRVTGWGNLKE--------GQPSVLQVVNLPIVERPVCKDSTRIRITDNMFCA
Ref.	GYKPDEGKRGDACEGDSGGPFVMKSPFNNRWYQMGIVSWGEGCDRDGKYGFYTHVFRLKKWIQKVIDQFGE/GDF
2ZDA	GYKPDEGKRGDACEGDSGGPFVMKSPFNNRWYQMGIVSWGEGCDRDGKYGFYTHVFRLKKWIQKVIDQFG-/GDF
2ZO3	GYKPDEGKRGDACEGDSGGPFVMKSPFNNRWYQMGIVSWGEGCDRDGKYGFYTHVFRLKKWIQKVIDQFG-/GDF
2ZC9	GYKPDEGKRGDACEGDSGGPFVMKSPFNNRWYQMGIVSWGEGCDRDGKYGFYTHVFRLKKWIQKVIDQFG-/GDF
3DHK	GYKPDEGKRGDACEGDSGGPFVMKSPFNNRWYQMGIVSWGEGCDRDGKYGFYTHVFRLKKWIQKVIDQFG-/-DF
2ZFP	GYKPDEGKRGDACEGDSGGPFVMKSPFNNRWYQMGIVSWGEGCDRDGKYGFYTHVFRLKKWIQKVIDQFG-/-DF
2ZGX	GYKPDEGKRGDACEGDSGGPFVMKSPFNNRWYQMGIVSWGEGCDRDGKYGFYTHVFRLKKWIQKVIDQFG-/-DF
Ref.	EEIPEEYL
2ZDA	EEIPEEYL
2ZO3	EEIPEEY-
2ZC9	EEIPEEY-
3DHK	EEIPEEY-
2ZFP	EEIPEEYL
2ZGX	EEIPEEYL

Due to a lack of proper parameters, an additional alteration of the published structures was performed on the modified O-sulfo-L-tyrosine amino acids (TYS), which were transformed into regular tyrosine by removing the sulfonyl group. This was followed by the removal of all resolved water molecules, whereas the two sodium ions, which are present in all eight structures, were kept.

Protonation and parameterization of the modeled protein structures were done with the program teLeap of the AmberTools18 from Amber2018[219] using the FF14SB force field[99].

This also includes the determination of the protonation state of basic and acidic amino acids to their default values at pH 7.4. The preparations of the ligands comprised of a protonation step using openBabel[220], also at pH 7.4, followed by the calculation of RESP charges with Gaussian 16[221] and teLeap on the basis of Hartree-Fock calculations with the 6-31G* basis set[222,223]. The Lennard-Jones non-bonded parameters were taken from the GAFF[98] force field. The raw data and resulting parameters can be found in the electronic appendix under 3.3/Param/.

Free energy calculations and visualizations

For the calculation of the LFE values, most of the computational parameters described in 3.2.1 were reused in this chapter and are therefore not explained in further detail here. This includes those regarding the 3D RISM calculation as well as the methods and parameters controlling the integration. Only the grid dimension had to be extended to 240 by 250 by 290 with an equidistant grid point spacing of 0.25 Å to accommodate the significantly larger molecules. The parameterization of the proteins was done with the FF14SB[99] force field. The raw data and resulting parameters can be found in the electronic appendix under 3.3/Param/. The non-localized calculation of the excess potential was done via a 3D version of the Morita-Hiroike formula[101,102].

The individual solvation free energies of the unbound ligand and protein and the one of the complex were calculated and on the corresponding sites localized in separate computational steps in a semi-trivial parallel way. Since all λ-steps are independent of each other, they were calculated simultaneously, reducing the calculation time drastically by increasing the demand for computational resources. The values for the intermolecular energy were taken from the $\lambda = 1.0$ step of the electrostatic scaling run of the complex. The 3D RISM results and LFEs can be found in the electronic appendix under 3.3/3.3.3/.

The FED values were calculated on the basis of the derivatives and g-function of the $\lambda = 1.0$ steps of the electrostatic scaling run of each LFE calculation. The raw data and FEDs can be found in the electronic appendix under 3.3/3.3.4/Calc/.

The grid-based three-dimensional convolution of LFE and FED with a Gaussian for the volumetric visualizations as well as the calculation of the difference maps were performed in Python 3.6[224] using the NumPy[225] library. All values on the main diagonal of the covariance matrix $\mathbf{\Sigma}$ of the gaussian were one while all other elements were zero. To keep the computational burden of the convolution manageable on such big grids as they were used here, a cutoff of 5 Å (20 grid points) was implemented.

3.3.3 Protein-ligand binding from the perspective of LFE

3.3.3.1 Naive atom-wise visualization and interpretation

The key idea of applying the LFE method on protein-ligand binding free energy is the separation of $\Delta_{\text{bind}}\boldsymbol{G}$ into its components to get a more in-depth insight into the relation between host and guest and identify those structures in both molecules having the most decisive influence on the binding free energy. The in 3.3.1 introduced approach allows for a hierarchical ordering in which each layer brings a higher level of detail. The coarsest layer is the separation in protein- and ligand-perspective. It is the uppermost layer since both resulting components can be interpreted independently, and it separates both lower layers with it. The separation in perspectives is also a unique feature of the here employed application of LFE to binding free energies. The second layer of this scheme is the separation of the binding free energy into the intermolecular energy and a solvation-driven part. The latter is the difference in $\boldsymbol{G}_{\text{hyd}}$ between bound and unbound state and will be referred to as desolvation penalty, describing the thermodynamic penalty arising from suboptimal solvation distribution of protein and ligand in the bound state. While positive values correspond to unfavorable changes in the solvation situation regarding binding, regions with negative values stabilize the binding due to removal of destabilizing solvent density (in this case, one could also speak of a desolvation gain). The third layer is the localization of the binding free energy and its contributing terms on the atom-wise basis of the LFEs. As an example for this interpretation approach, a few key features of the ligand-binding shown in **Figure 3.8** are discussed in the following.

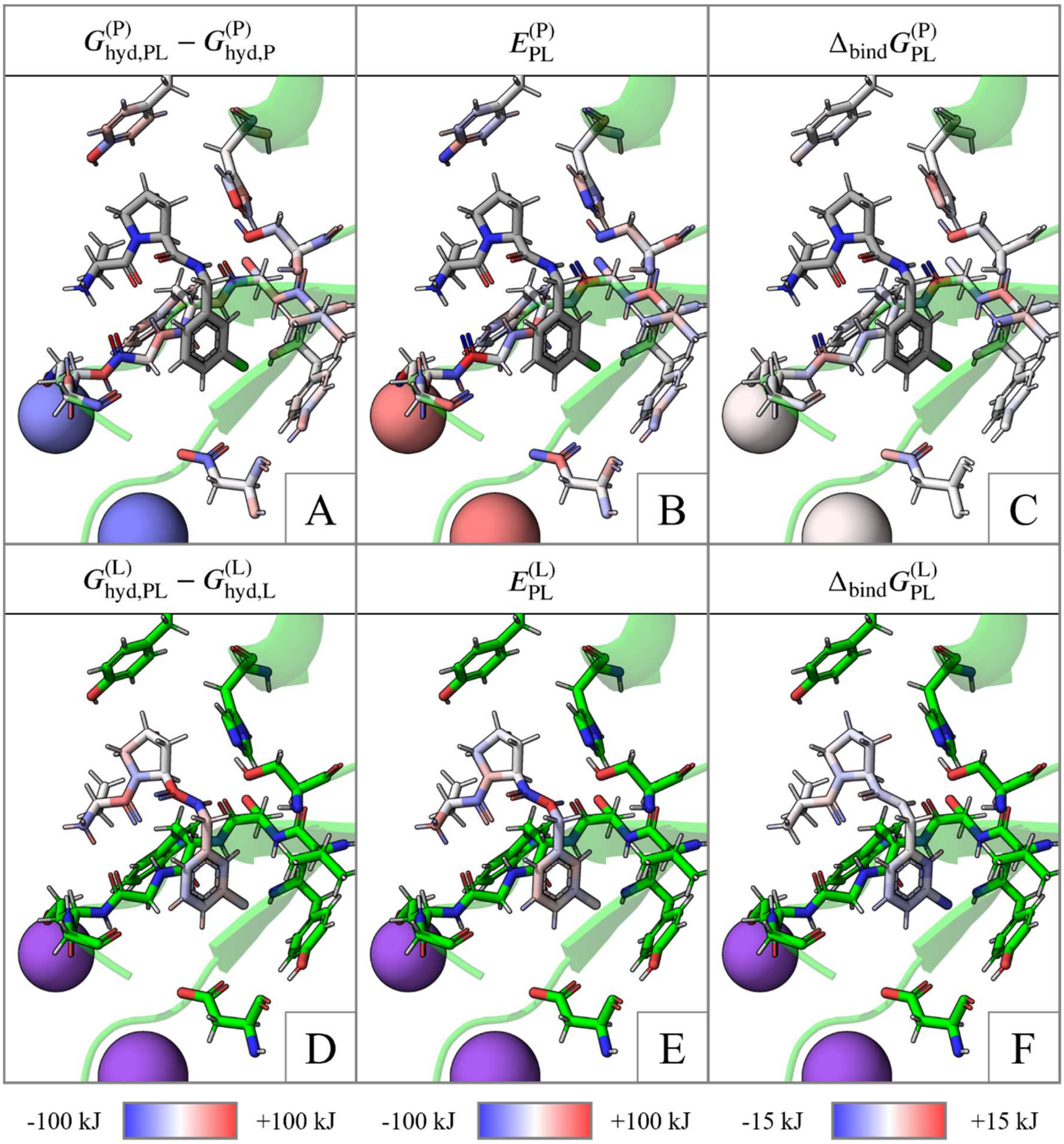
$G^{(P)}_{hyd,PL} - G^{(P)}_{hyd,P}$
$E^{(P)}_{PL}$
$\Delta_{bind}G^{(P)}_{PL}$
A
B
C
$G^{(L)}_{hyd,PL} - G^{(L)}_{hyd,L}$
$E^{(L)}_{PL}$
$\Delta_{bind}G^{(L)}_{PL}$
D
E
F
-100 kJ
+100 kJ
-100 kJ
+100 kJ
-15 kJ
+15 kJ

Figure 3.8 Intermolecular energy as well as desolvation and binding free energies mapped on protein and ligand of the thrombin complex 2ZFP. Figures A to C show the protein (P) perspective of the solvation (A) and intermolecular energy (B) term as well as the binding free energy (C) itself. The bottom row (figures D to F) shows the ligand (L) perspective, respectively. The coloring of the LFE and intermolecular energies share the same scale, while the binding free energy is scaled in a much narrower range to adjust for the compensational effect between the former named two terms. The protein structure is shown in excerpts only to improve clarity. The spheres represent the experimentally found sodium ions. The source data for the LFE-color coding can be found in **Table 6.4** and the electronic appendix under 3.3/3.3.3/2ZFP/Loc/.

The chlorine in the meta position of the terminal phenyl is such a feature. Halogen decorations are a common and versatile option in the ligand design toolbox. They are usually easy to introduce in synthesis and are simultaneously comparatively stable against metabolic breakdown pathways like oxidation through cytochrome p450[226]. Their tendency to increase the ligand's hydrophobicity is advantageous for the binding affinity since it reduces the desolvation penalty caused by the environmental change of the ligand under binding. This general rule is also confirmed here and can be seen in **Figure 3.8** D by the chlorine's blue color and, in fact, of almost the complete phenyl ring. One of the proposed modes of action of the specific chlorine in the here discussed ligand is the replacement of a thermodynamically unstable water[227]. (The term refers to water molecules trapped or forced into cavities within the protein where they have a higher free energy than their counterparts in bulk phase[228–230].) Designing a ligand to replace those such water molecules is usually highly desirable since releasing them into the bulk is associated with an entropy gain, and the potentially newly formed protein-ligand interactions could give an enthalpic advantage.

Table 3.7 List of atoms of the amino group and their corresponding intermolecular energy, desolvation, and binding free energies of the thrombin complex 2ZFP. The hydrogen atoms are listed clockwise from the viewer's perspective in **Figure 3.8**, starting with H1 being the closest one to the oxygen of GLY216. Additionally, the sum over all four atoms of the amino group is also given. A full list can be found in **Table 6.4**

Atom	$\Delta\Delta_{\mathrm{solv}}G^{(\mathrm{L})}_{\mathrm{PL-L,loc}}$ / kJ mol^{-1}	$E^{(\mathrm{L})}_{\mathrm{PL,loc}}$ / kJ mol^{-1}	$\Delta_{\mathrm{bind}}G^{(\mathrm{L})}_{\mathrm{PL,loc}}$ / kJ mol^{-1}
Cl	-10.07	0.48	**-9.60**
N	-39.07	40.19	1.12
H1	24.14	-27.07	-2.93
H2	39.84	-40.13	-0.29
H3	14.92	-16.96	-2.04
Amine group	39.83	-43.97	**-4.14**

The second mode of action discussed in the literature for chlorine in this position is a potential Cl-π interaction formed with TYR228[80,231–235]. Such an interaction cannot be found in the here presented results and is even contradicted by a positive $E^{(\mathrm{L})}_{\mathrm{PL}}$ on the chlorine atom (colored white in **Figure 3.8** E; see **Table 3.7**). This finding is to some degree to be expected in the light of the here for the ligand used force field. The GAFF has no representation of the sigma-hole effect, causing the interaction between halogen and aromatic ring[234–236], and therefore neglects all halogen-bond effects. Nonetheless, the desolvation term (-10.07 kJ/mol) overcompensates the only very slightly positive intermolecular energy strongly enough to form a clear picture for the binding free energy where the chlorine atom is one of the most beneficial within the ligand, being in good agreement with the literature on this topic[227,233,237,238].

Another important structural feature in the ligand series analyzed in this chapter is the amino group, which protonated form is assumed to be the dominant one[80]. While introducing a net positive charge into the ligand is problematic for the oral bioavailability[239] and current research tries to avoid any such charges[227,239], it is a key component in many reported compounds[238,240–242] and it plays an essential role in the complex at hand.

The desolvation free energy is, as one would expect for a polar residue, positive, as reported in **Table 3.7**. The outward-pointing hydrogen atoms show a strong desolvation penalty as they form the amine group's solvent-accessible surface. Hence, they dominate the solvent structure in their direct vicinity in the unbound state by accumulating oxygen density around them. The nitrogen, in contrast, now freed from the electrostatic unfavorable oxygen density, shows a strong negative desolvation penalty $\Delta G^{(\mathrm{L})}_{\mathrm{hyd,PL-L}}$. Nonetheless, the hydrogen atoms combined are overcompensating the negative contribution of the nitrogen by 39.83 kJ/mol, being in line with the rational expectation.

In the literature, amino groups in this position are reported to interact with the oxygen of GLY216 of the protein backbone[80,231,233,239]. This is very clearly reproduced by the intermolecular energy, as indicated by the strong blue color of the oxygen and hydrogens in **Figure 3.8** B and E, as well as by the sum over $E^{(\mathrm{L})}_{\mathrm{PL}}$ in **Table 3.7** and **Table 6.4**. Due to this strong, charge assisted H-bond, the intermolecular energy can compensate for the severe desolvation penalty and resulting in a negativ $\Delta G^{(\mathrm{L})}_{\mathrm{hyd,PL-L}}$.

The in **Table 3.7** shown numbers are exemplary for a common trend throughout the protein and ligand. There is a noticeable pattern of alternating blue and red-colored atoms in both perspectives, only very seldomly broken. Since this effect originates from differences between the atoms' partial charges, it is more pronounced for atom pairs with a high dipole moment, making the discussed amino group and the carboxyl group of GLY216 good examples. In many cases, the signs of $\Delta G^{(\mathrm{X})}_{\mathrm{hyd,PL-X}}$ and $E^{(\mathrm{X})}_{\mathrm{PL}}$ on a given atom are mutually inverse, resulting in less drastic numbers for $\Delta_{\mathrm{bind}} G^{(\mathrm{X})}_{\mathrm{PL}}$. There is a strong tendency for compensating each other's maxima and minima in the relation between the difference in solvation free energy and the intermolecular energy. This can be seen most prominently for the sodium ions in **Figure 3.8**, for which the solvation free energy decreases significantly under binding (blue color in A of Fig. 3.2) but have a positive intermolecular energy due to the net positive charge of the ligand. Both contributions cancel each other nearly entirely so that the sodium ions appear white in the binding free energy. Nonetheless, the mentioned alternating pattern is present for most atoms in the localization of the binding free energy. This makes the interpretation of color-coded figures rather difficult, as it not clear which color is dominating in a particular part at first sight. The summation of the different terms over individual atoms and interpreting the localization solely on those numbers is also not reliable. The same pattern, which is supposed to be overcome in the first place, introduces a decision problem, usually not to solve in a straightforward manner. Extending or retracting the sphere of summation by just one atom can easily flip the result's sign and invert the interpretation. Furthermore, the grouping of atoms in chemically logical groups cannot be automated and must be done manually, as such an assignment is always subjective and depends on the overarching intent of the study at hand.

3.3.3.2 Volumetric visualization and interpretation approach

The just described problems make it clear that a more fitting visualization and interpretation method is required. An ideal solution should cover the following points:

- Correctness: A suitable visualization method should map the LFE and intermolecular energy information as truthfully as possible. For example, essential cornerstones are to retain the sum over all atom-wise energy contributions to preserve the correct relation between the perspectives and energetic contributions as well as the spatial properties to ensure a correct interpretation.
- Grouping: As explained above, defining meaningful sets of atoms is inherently error-prone due to the alternating pattern throughout the molecules and generalizes poorly for its subjective nature. A new visualization approach should avoid any such human interference but still provide a logical grouping and should smoothen out the omnipresent pattern
- Perspectives: A major drawback in the previously used method is the inability to present the localization results while also retain the full information about the underlying chemical structure since the element color code gets overwritten. An optimized visualization should be capable of displaying the ligand and protein perspective separately as well as combined as complex.
- Comprehension: A replacement visualization method must be more intuitive and faster to interpret than the direct atom-wise labeling and color-coding explored above without losing any crucial details.

In this work, a solution is proposed, considering all four of the listed points. Key is a convolution operation with a three-dimensional gaussian, effectively spreading the localized information on a 3D grid. This is the same operation already used to create the input for the 3D convolutional neural networks discussed in 3.2.1.

Since the LFE values are folded with a normalized Gaussian, the sum over the atom-wise localized energies is equal to the integral over the corresponding 3D grid. A numerical evaluation of this is shown for the 2ZFP complex in **Table 3.8**.

Table 3.8 The table gives the excess chemical potential for the protein, ligand, and complex of 2ZFP (see **Figure 3.8** and **Figure 3.10**) in aqueous solution calculated with 3D RISM as a reference as well as the sum over all atom-wise contributions and the integral over the 3D grid. The numbers in the round brackets give the error relative to the 3D RISM reference in percent. For the complex, no integral is given since the ligand bound to the complex is treated separately from the bound protein and therefore, the grid of the complete complex is never calculated for the evaluation. The numbers show clearly that the convolution with the 3D gaussian is not introducing a significant error, even with a seemingly harsh cutoff at 5 Å. The 3D RISM data as well as the corresponding LFE values can be found in the electronic appendix under 3.3/3.3.3/2ZFP/.

	$\mu^{\text{ex}}_{\text{Protein}}$ / kJ mol^{-1}	$\mu^{\text{ex}}_{\text{Ligand}}$ / kJ mol^{-1}	$\mu^{\text{ex}}_{\text{Complex}}$ / kJ mol^{-1}
3D RISM reference	626.63	-70.47	851.07
Sum over LFE	613.86 (2.04%)	-70.63 (0.23%)	836.90 (1.66%)
Integrated over grid	613.87 (2.04%)	-70.63 (0.23%)	-

As an example of this visualization method, the localized energies of the previously introduced 2ZFP complex are shown in **Figure 3.10** as differently sized and colored volumes. The method requires a minimum and maximum cutoff, excluding the range in between from plotting and showing only those grid points where the localized energy is below the lower cutoff (blue) or above the upper (red). This is a critical task to keep the resulting figure interpretable and be able to find and focus on the most important regions. A too narrow excluded range leads to a confusing abundance or meaningless large volumes (see **Figure 3.9**), while a too-wide range may hide less significant but still interesting areas.

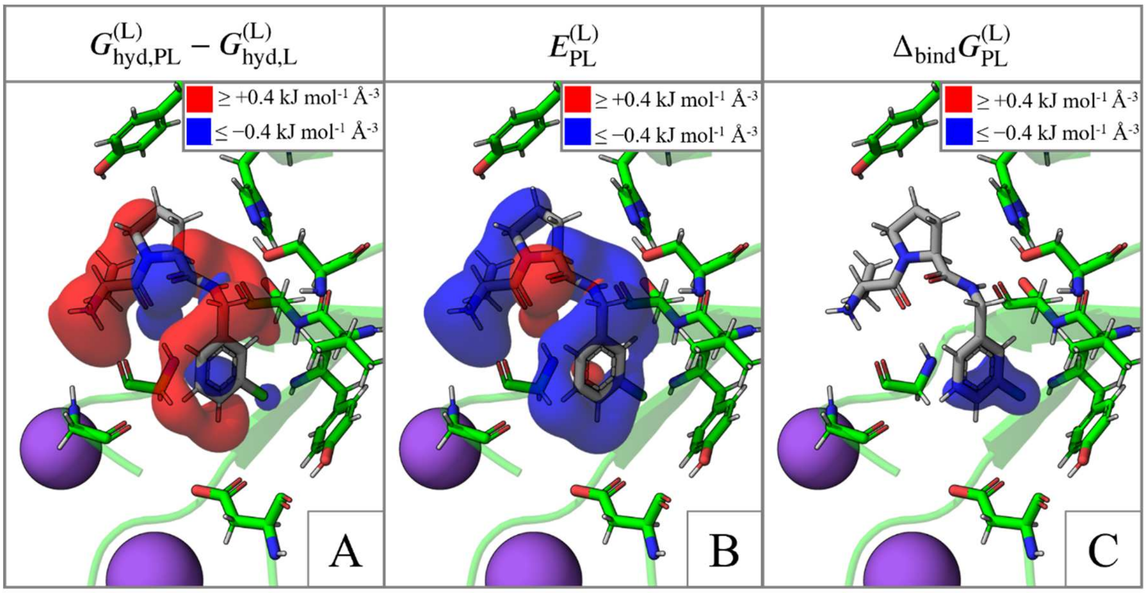
$G^{(L)}_{\mathrm{hyd,PL}} - G^{(L)}_{\mathrm{hyd,L}}$
$E^{(L)}_{\mathrm{PL}}$
$\Delta_{\mathrm{bind}}G^{(L)}_{\mathrm{PL}}$
🟥 ≥ +0.4 kJ mol-1 Å-3
🟦 ≤ −0.4 kJ mol-1 Å-3
A
B
C

Figure 3.9 Demonstration of a homogeneous LFE cutoff choice for desolvation, intermolecular energy, and binding free energy.

To find suitable values for these cutoffs, those values marking the 0.005% and 0.995% quantiles of all gridpoints can be used as initial values for the minimum and maximum, as it was done in the following examples. This gives a vague idea of where to find the best fitting values, but a manual adaptation is recommended. The in **Figure 3.10** shown example demonstrates the significance of adjusted cutoffs for the different contributions, by using the binding free energy cutoff for all three. The resulting volumes in A and B are so extensively big that the localization aspect is mostly lost and they become too unspecific to learn much from them. The necessity to adjust the cutoffs manually can be considered as a clear downside to this approach. However, this drawback is less severe than it may seem on paper, since the intended primary usage of the method is to inspect and interpret the results in a graphical computer program like PyMol[243]. Here, the cutoffs and other parameters like the view can be easily manipulated, revealing even more detail than can be shown by images in this work.

One of the major benefits of the proposed visualization method is its implicit grouping mechanism arising from the blurring effect of the convolution with a gaussian itself. On the one hand, atoms located close to each other, like covalently bonded ones, influence one another strongly and automatically group together. On the other hand, the sphere of influence decays sufficiently fast due to the 1 Å standard deviation in all three dimensions, which retains the localization effects. Furthermore, contributions of the same sign stack, to create more emphasized regions, while those with opposing sign cancel each other out.

An example of this mechanism is the interaction between the protonated amino group and GLY216 of the backbone of thrombin. With the naïve atom-wise color coding, it is not clear from visual inspection whether the negative or positive contributions within a group of atoms dominate. One must perform a summation over manually selected atoms with its previously discussed drawbacks to get a clear answer. With the volumetric representation, however, the desolvation penalty and the strong beneficial intermolecular energy of both groups becomes clear at first glance. This is also true for the analogous interaction between the amid group of the ligand (atoms O0A,C09,N0E,H13, see **Table 6.4**) and the backbone of SER214. The volume around the amino group in **Figure 3.10** E is an example of the convolution's blurring property and illustrates that the whole group interacts with the protein backbone. Nonetheless, the localization is not lost either, demonstrated by the small but intense volume around the carboxylic oxygen backbone atom.

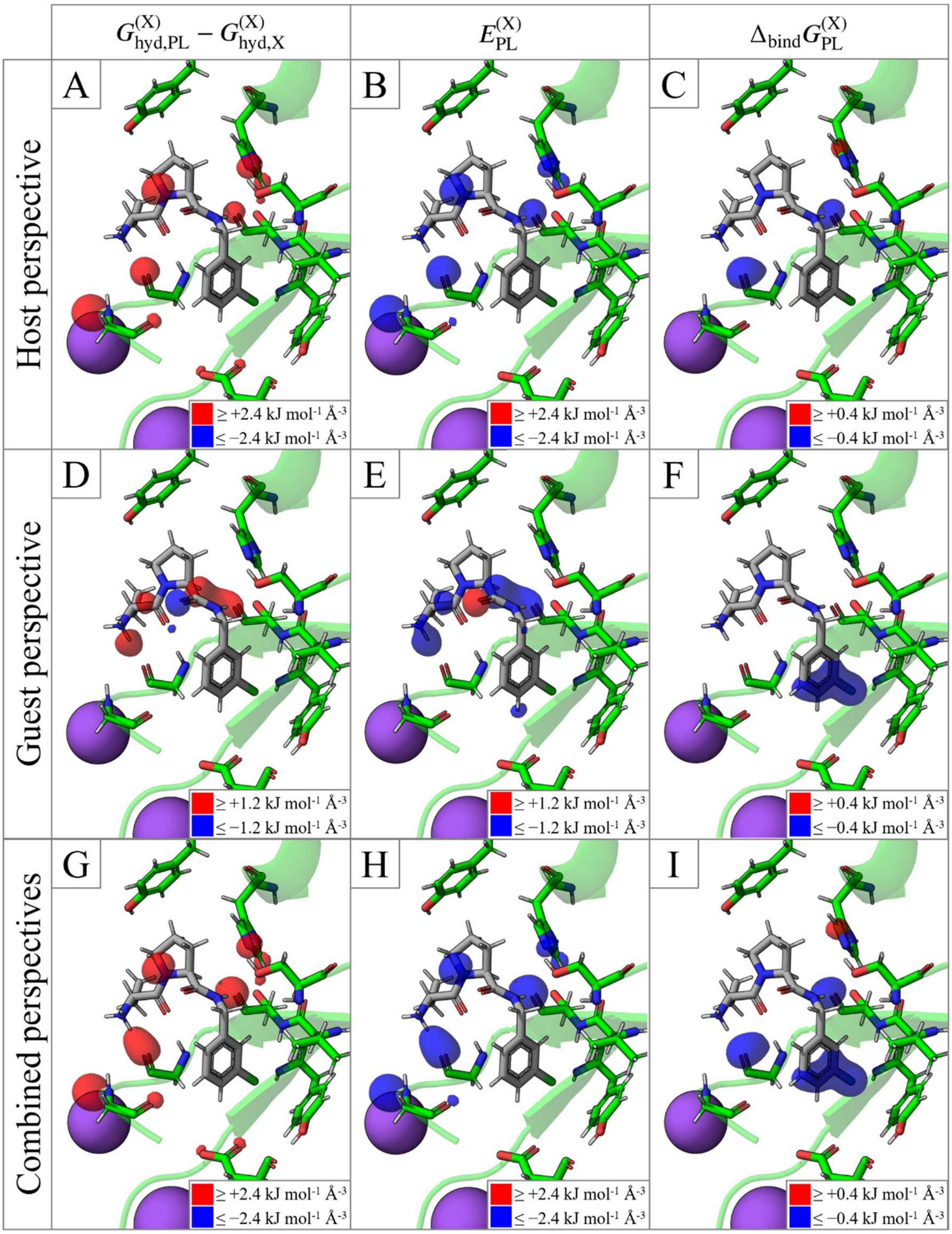

Figure 3.10 The first column from the left shows the desolvation penalty from the perspective of the protein of the thrombin complex 2ZFP (top), the ligand (mid), and a combination of both (bottom). The same is the for the intermolecular energy (mid column) and the binding free energy (right column). All properties are spread onto three-dimensional grids and displayed as volumes, where blue-colored volumes show those gird cells with an energy lower or equal to the cutoff given in each figure individually. For red-colored volumes applies the same but with higher or equal numbers than the given cutoff. The protein structure is shown in excerpts only to improve clarity, and only those residues of highest interest or importance for orientation in the binding site are drawn explicitly.

The volumetric representation is also helpful for comparing the different perspectives. For example, the protein perspective indicates the biggest contributions to the binding free energy mostly in regions that are in direct contact with structures of the ligand, one would initially expect to show up in the ligand perspective, namely the amino- and amid groups. But in contrast to this belief, they are not particularly highlighted in the ligand perspective. A plausible explanation could be, that the groups in question are not ideally solvated in the apo structure to begin with, and taking solvent density away by the binding process has a less pronounced effect on them, as on their ligand counterparts. The solvent site density distribution in the binding site is always the result of the interaction of an abundance of surrounding solute sites. This, in turn, leaves the protein solute sites with a compromise regarding the solvent density distribution and, thus, a less than optimal solvation free energy, dampening the desolvation penalty. This effect is much less pronounced for the ligand, leading to a stronger desolvation penalty which cancels out the negative intermolecular energy on groups like the protonated amino and amid groups.

In general, the canceling effect between $\Delta G^{(\mathrm{X})}_{\mathrm{hyd,PL-X}}$ and $E^{(\mathrm{X})}_{\mathrm{PL}}$ is much more obvious in the volumetric representation due to the similar shaped, sized, and positioned volumes (compare **Figure 3.10** A vs. B, D vs. E, G vs. H). As a result, the included range of $\Delta_{\mathrm{bind}}G^{(\mathrm{X})}_{\mathrm{PL}}$ must be narrower to properly visualize the most important regions. On the other side, adjusting the cutoffs of both contributions can also generate valuable inside, as demonstrated for the ligand perspective in **Figure 3.9**. The region around the chlorine atom and phenyl ring differs between $\Delta G^{(\mathrm{L})}_{\mathrm{hyd,PL-X}}$ and $E^{(\mathrm{L})}_{\mathrm{PL}}$, indicating that the negative contribution to $\Delta_{\mathrm{bind}}G^{(\mathrm{L})}_{\mathrm{PL}}$ of the chlorine atom origins from its desolvation while the hydrogen atoms of the phenyl ring are contributing via a negative $E^{(\mathrm{L})}_{\mathrm{PL}}$. A in the literature reported major mode of action is the replacement of a thermodynamically unstable water, located in the S1 pocket[80,227,237]. Therefore, one would expect to see some blue-colored volumes in this region for the protein perspective too. After all, such a process affects first and foremost the protein. The absence of those volumes can be explained by another perspective switch, this time to the water in question. Observed on its own, its thermodynamic properties may stand out among other water molecules[227], marking it as thermodynamic unstable. Nonetheless, in the case of the LFEs the thermodynamic information of the water is projected onto the surrounding protein atoms, eventually spreading it so thin that it blends with the background.

In a first assessment of the volumetric visualization of the localized energies, it can be concluded that the requirement of a correct representation is indeed fulfilled. The total overall contributions are preserved and the spatial and perspective relations are truthfully mapped. Furthermore, the major drawback of the atom-wise color-coding, the necessity of manual grouping, is solved by the implicit blurring of the convolution operation, which balances generalization and focus nicely, as shown in the previous examples. This also enables a quick and effortless comparison between the different perspectives, which in turn gives more profound insights into the protein-ligand binding relation. Whether the volumetric representation is indeed an improvement over the atom-wise coloring regarding comprehension and ease of interpretation is a subjective matter and probably varies from person to person. A combination of both would most likely prove most beneficial in many research applications since both representations have their individual strong points. Nonetheless, in the context of this work, the volumetric representation is favored for its comprehensibility and the possibility of manipulating the three-dimensional fields, as described in more detail below in 3.3.3.3.

3.3.3.3 Application of LFE analysis and difference maps to a compound series

In the following, the concepts introduced and discussed above shall be applied to the ligand series published by Baum et al.[80], and shown in **Figure 3.5**. A collection of all six compounds in their bound state can be found in **Figure 3.11**, together with their volumetric visualized localized binding free energy. The series introduces step-wise changes to the ligands at the S1 pocket pointing part by replacing the *m*-chlorobenzyl with a benzamidine and at the S3/S4 facing part by introducing benzyl-groups. Each of these modifications increases the binding affinity to thrombin between 2.4 and 10.7 kJ/mol, as determined by ITC measurements, where the introduction of the amidine group is the more potent change.

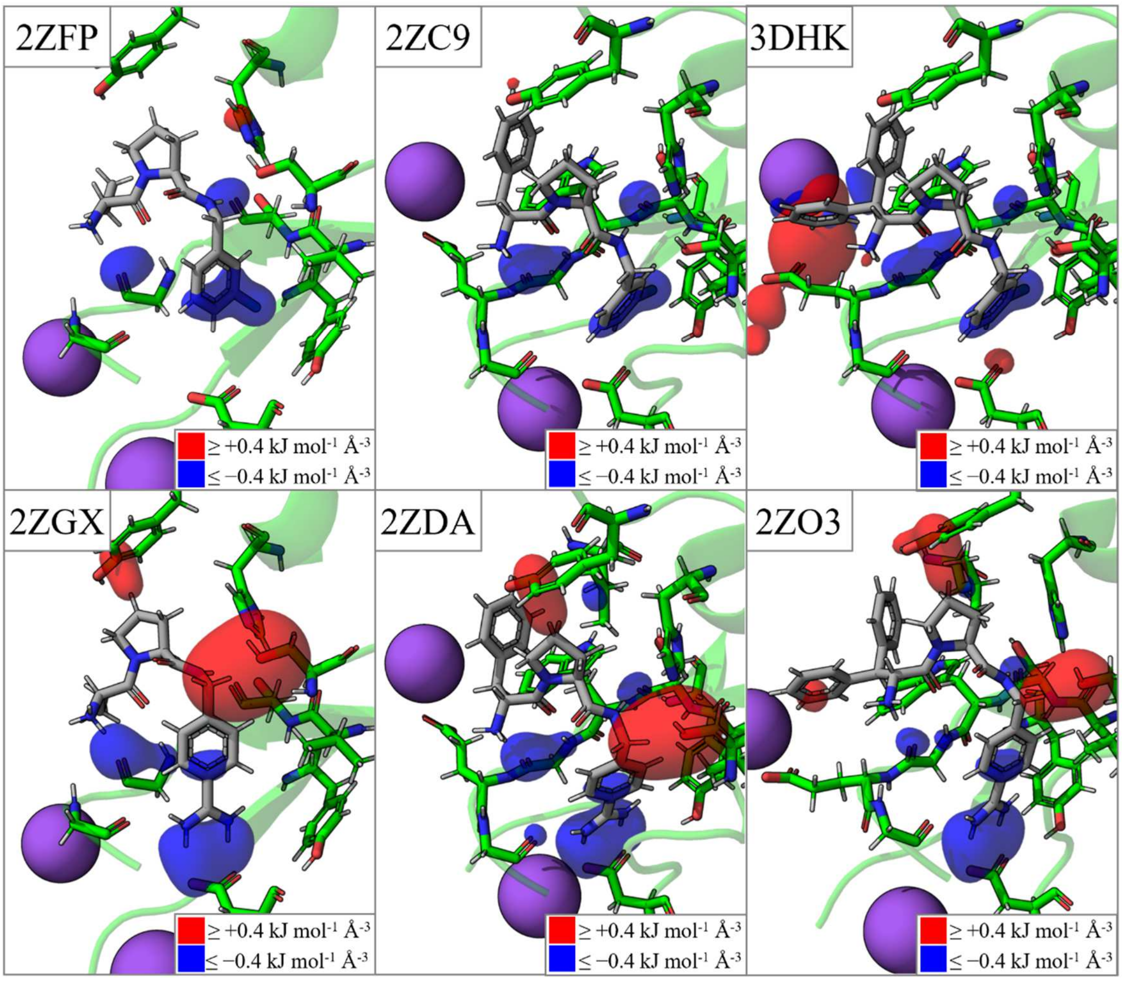

Figure 3.11 Collection of all compounds used in the ligands series published by Baum et al.[80] depicted with their corresponding localized binding free energy, shown as red and blue volumes. The upper row shows the *m*-chlorobenzyl based compounds and the step-wise addition of benzene to the compound expanding in the S3 moiety of thrombin. The bottom row shows the same step-wise additions but for the benzamidine base version of the inhibitor.

To better grasp the changes caused by these modifications, a further advantage of the volumetric representation is being used in addition to the calculations shown in **Figure 3.11**. The underlying three-dimensional fields can be easily manipulated mathematically too. Possible operations could be, for example, combining multiple fields by a weighted sum to generate a superposition of different host and guest states or visualizing differences between compounds by subtracting fields from each other. In cases where all involved fields share the same orientation and dimensionality, like it is the case in the here described examples, all operations can be done without any prior preparations. In cases where this is not given, an alignment and potentially an interpolation between grid points has to be performed. This is something to have in mind before any calculation, as it is easy to enforce proper alignment and dimensions in the preprocessing, but not so much during analysis.

The introduction of the protonated and, therefore, positively charged amidine group has dramatic effects on the distribution of the LFE, as can be seen in **Figure 3.12**. The most noticeable and obvious is the newly formed salt bride to the deprotonated, negatively charged ASP189 residue. In all three cases, the interaction follows the principles outlined above very truthfully, by having a strong positive desolvation penalty on both participating groups in the ligand and protein. This effect is overcompensated by the strong negative intermolecular energy, resulting in a negative binding free energy volume, encompassing the entire interaction site. Similar effects occur at the protein-backbone (GLY216, GLY219), lowering the local binding free energy by increased coulomb interaction to the additional positive charge without an additional desolvation penalty. Far more interesting and less expected are the consequences of the formation of this salt bride on the rest of the ligand.

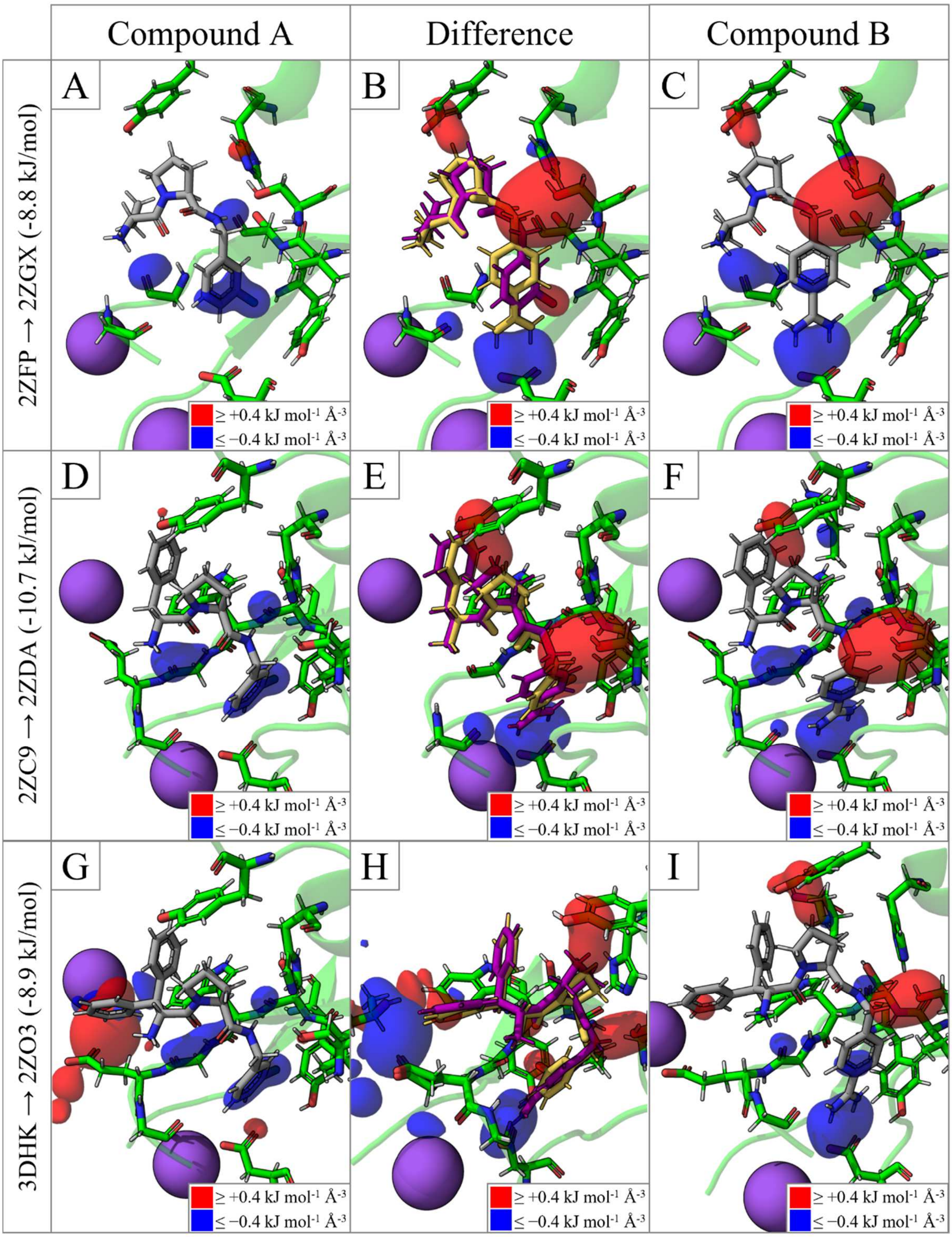

Figure 3.12 Comparison between the *m*-chlorobenzyl based and the benzamidine compounds, each with their corresponding benzyl decoration. In the most left column, labeled 'Compound A', the reference structures are shown from which the alchemical mutation to the structures in the most right column, labeled 'Compound B', accrues. The middle column depicts the volumetric difference maps, overlayed with both structures (compound A in yellow, compound B in violet, both protein structures in green). The shown processes are labeled by the PDB code (Baum et al.[80]) of the start and end structures on the left side of the panel. The via ITC measure $\Delta\Delta G°$ for the transition is given in brackets and was taken from Baum et al.[80].

By pushing the ligand a bit out of the S1/S2 pocket into the protein, it forces the ligand in a slightly altered binding conformation, which causes significant changes in the LFEs in turn. This rearrangement is mainly limited to the ligand itself, and there are almost no changes in the protein conformation and the orientation of sidechains. This is somewhat curious since the new binding mode changes the intermolecular energy pattern, which should lead to a mutual adaption of both binding partners in 3D space. This becomes most obvious in the interaction between one of the ligand hydrogen atoms and the SER195 sidechain oxygen atom. For the *m*-chlorobenzyl based compounds of the series, the distance between both is 2.3 Å, which is already within contact distance regarding the σ-parameter ($\sigma_{\mathrm{mix}} = (\sigma_{\mathrm{H}} + \sigma_{\mathrm{O}})/2 = 2.77$ Å). Here the moderately positive Lennard-Jones potential is compensated by the negative coulomb potential resulting in a negative localized interaction energyfor the hydrogen atom in the 2ZFP compound (-9.32 kJ/mol). For the amidine-based compounds with an H-O distance of around 1.8 Å, however, the repulsion is significant, leading to a considerable positive intermolecular energy contribution on the hydrogen atom for the 2ZGX compound (127.24 kJ/mol). As one can observe in the B figures of **Figure 6.3**, **Figure 6.5**, and **Figure 6.7**, the oxygen does not generate a corresponding red volume, as its coulomb energy is dominated by the +2 net charge of the full ligand.

Nonetheless, this result raises the question, why protein and ligand crystallized consistently in this specific configuration and neither of the participated structures moved out of the way. Especially for the rather mobile serine sidechain, a slight rotation along the CA-CB bond axis seems to be easily possible and collision-free. It is hard to say whether this is an artifact deriving from uncertainties in the process of fitting the protein sequence to the experimentally determined electron density (despite the general high quality; the used structures have a limited resolution between 1.58 and 2.25 Å), or a natural effect modeling the reality truthfully. To avoid over-interpretation, the discussion of this particular host-guest-interaction is best limited to the level of differences between structures, as here most potential artifacts should cancel out.

A very similar but far less pronounced effect of the altered binding mode by the introduction of amidine group is visible around the hydrogen-carbon contact between TYR60A and the proline part of the ligand. Here the distance between both groups is also lowered, leading to repulsion in the Lennard-Jones potential, but this time without the benefit of any compensation by the Coulomb potential since both atoms carry positive partial charges. In this particular case, the protein behaves more plausible. Even though the tyrosine sidechain has far fewer options to turn away, the crystal structures show a slight bending in CA-CB-CG angle of the amino acid to avoid even closer contact with the ligand.

In addition to the direct effects of the introduction of a group, there are also consequences of removing atoms from the compounds, in this case the in the S1 pocket binding chlorine. As discussed above it is one of the main actors in the binding process of the *m*-chlorobenzyl compounds by its negative desolvation penalty and the replacement of a thermodynamically unstable water molecule from the binding site. The amidine-based compounds do not have any residue, which could fulfill this task and are even rotated away from the water-bearing moiety due to the planarity of the benzylamidine, enforced by the π-electron system. For those reasons, one can find the aforementioned unstable water in the experimental structures and the chlorine is indeed highlighted by a red volume in the difference maps, shown in **Figure 3.12**. Nonetheless, the experimental ITC measurements of Baum et al., which show a clear entropic gain for the transformation from *m*-chlorobenzyl towards benzylamidine, do not fully reflect this line of reasoning, since one would expect a positive $-T\Delta\Delta S^{\circ}$ for the reintroduction of a thermodynamically unstable water. The original paper, from which the structures and number were taken, does not give an explanation for this phenomenon and it is unclear how much this has to do with the mentioned water or if it is caused by another, unknown effect, detached from the water placement in the S1 pocket. It is possible, and judging from the magnitude of the number also plausible, that the change in entropy is caused by an increase of the degrees of movement, vibration, and rotation anywhere in complex, which cannot be reflected by the LFE calculation, as it only takes solvents related entropic terms into account.

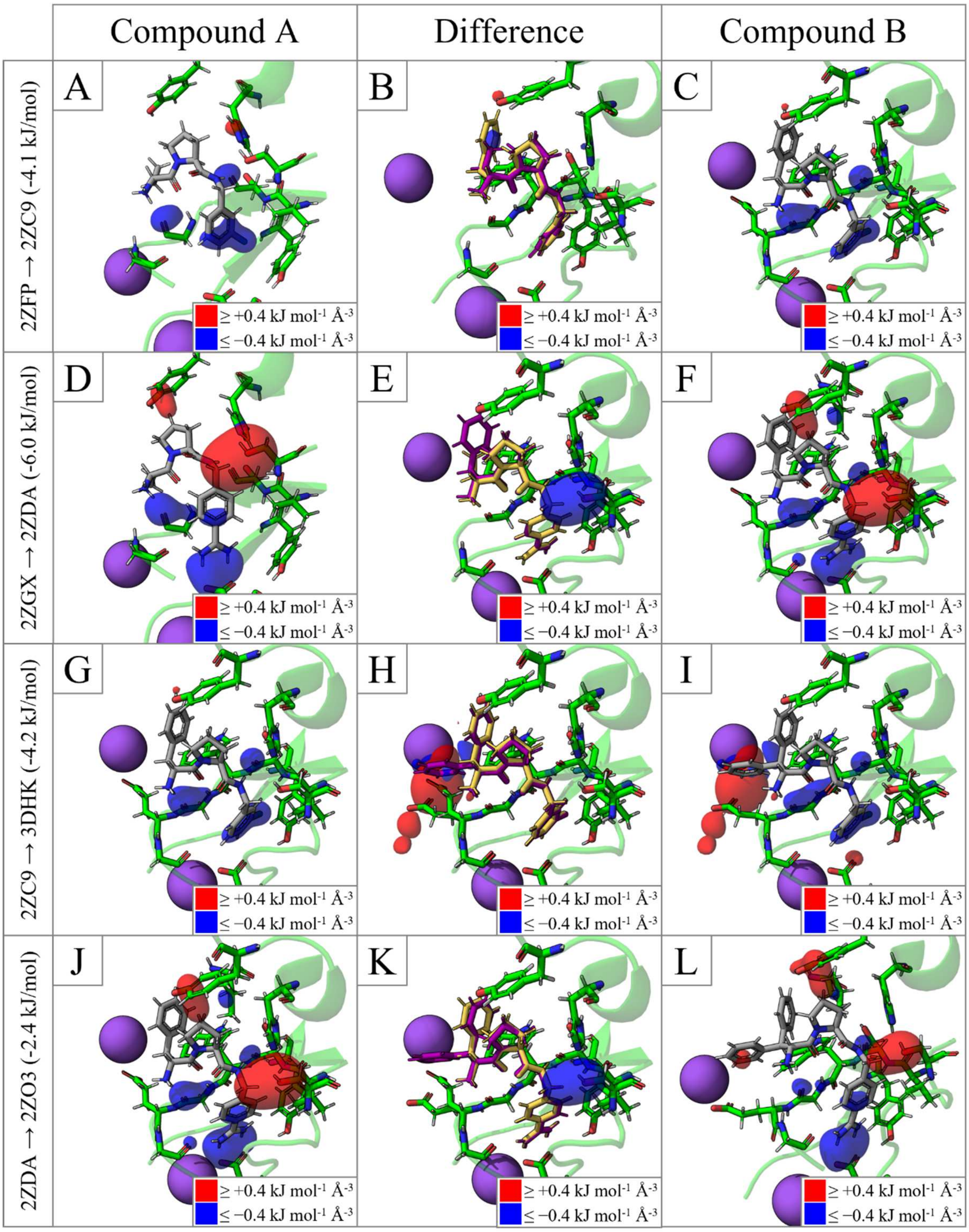
Compound A
Difference
Compound B
2ZFP → 2ZC9 (-4.1 kJ/mol)
2ZGX → 2ZDA (-6.0 kJ/mol)
2ZC9 → 3DHK (-4.2 kJ/mol)
2ZDA → 2ZO3 (-2.4 kJ/mol)
A
B
C
D
E
F
G
H
I
J
K
L
≥ +0.4 kJ mol-1 Å-3
≤ −0.4 kJ mol-1 Å-3

Figure 3.13 Comparison between different grades of benzylation of the *m*-chlorobenzyl based (upper two rows) and the benzamidine based compounds (lower two rows). In the leftmost column, labeled 'Compound A', the reference structures are shown from which the transition towards the structures in the rightmost column, labeled 'Compound B', starts. The middle column depicts the volumetric difference maps, overlayed with both structures (compound A in yellow, compound B in violet, both protein structures in green). The shown processes are labeled by the PDB code (Baum et al.[80]) of the start and end structures on the left side of the panel. The via ITC measure $\Delta\Delta G°$ for the transition is given in brackets and was taken from Baum et al.[80].

The benzylation of the ethyl group in the S3 facing part of the compounds is, like the replacement of *m*-chlorobenzyl by benzamidine too, associated with a gain in the binding affinity to thrombin. However, the modifications also increase the experimental uncertainty to a point where the assigned errors to the K_d are overlapping between the mono and bis(phenyl)methane *m*-chlorobenzyl inhibitors (see **Figure 3.5**).

For the first benzylation, three major modes of action can be identified: altered distribution of partial charges throughout the ligand, a T-shaped π-π stacking interaction with a tryptophane, and the replacement of a thermodynamically unstable water.

The change in the partial charge distribution is mainly visible in **Figure 3.13** E (and K for the second benzylation) by mitigating the clashing ligand-hydrogen with the oxygen of the SER195 sidechain. The introduction of the additional aromatic ring polarized the C-H bond more, leading to a more positively charges hydrogen and ultimately to a less severe positive localized binding free energy on both atoms, visualized by the blue volume in the difference map.

Indicated by a small blue volume, the π-π stacking interaction between the phenyl ring and TRP215 is at the chosen cutoff of 0.4 kJ $Å^{-3}$ only visible in the difference map of the 2ZFP→2ZC9 transformation (**Figure 3.13** B) and not even in corresponding amidine transformation of 2ZGX→2ZDA. This observation is backed by the experimental ITC measurements, where enthalpic contribution to $\Delta\Delta G^{\circ}$ is more negative for the transformation of the m-chlorobenzyl inhibitors (-3.6 kJ mol^{-1} vs. -1.4 kJ mol^{-1}; see **Figure 3.5**).

The entropic contribution to $\Delta\Delta G^{\circ}$ is for the insertion of the first phenyl ring in both cases negative (-0.5 kJ mol^{-1} and -4.7 kJ mol^{-1}), which hints towards the aforementioned replacement of water in the binding process, although as described above, this line of reasoning is not without doubt. Nonetheless, the crystallographic structures support the replacement and seem plausible, given the hydrophobic surrounding of ILE174, TRP215, and LEU99. As discussed in 3.3.3.2, only the LFE contributions on the host perspective can show a reaction to such a replacement, which is usually too far spread out over the protein residues to generate a visible volume at a sensible displaying cutoff[*].

The way the second benzylation of the compounds affects binding is less well understood than the first one and explanations from experimental results and the LFE approach are inconclusive. The positioning of the phenyl ring is already quite solvent-exposed as it is not as deeply buried in the binding pocket, resulting in a high B-factor and movement compared to the rest of the ligand. Nonetheless, the corresponding transformations show a positive $-T\Delta\Delta S^{\circ}$ (see **Figure 3.5**), again raising questions about a not yet understood mode of action. In the original paper, Baum et al. propose the rotation of GLU217 towards LSY224 and the so potential formation of a salt bridge as the main cause for the increased binding affinity. While this is indeed in line with the negative $\Delta\Delta H^{\circ}$ in both transformations, it is hard to verify the theory from any crystal structure-based approach, as the lysine in question is most likely affected by crystal packing effects, and thereby also hampering the expressiveness of the LFE. With this in mind, the contributions to the binding free energy around GLU217 have to be interpreted with caution.

Here most noticeable is the intricate pattern in volumes of the localized binding free energy of the 3DHK complex. It arises from an unusual, incomplete cancelation between the desolvation penalty and the intermolecular energy and is not nearly as fragmented in either one of those. This is even more interesting, since the normal compensation between both contributors to the binding free energy does occur in the amidine counterpart, as expected. Analyzing the different perspectives and contributions in **Figure 6.6** the observed results can be rationalized with the explanations outlined above, but why the magnitudes of the involved fields do not match as well as in other regions and complexes is not entirely clear. In the 2ZO3 complex, where the amidine group introduces an additional positive charge, the intermolecular energy lines up much more closely, leaving only a small volume around the phenyl ring itself, introduced by the guest perspective (see **Figure 3.13**).

[*] An exception to this can be found in **Figure 3.12** and **Figure *3.13*** F as well as in **Figure *6.5*** C and I. Here a small volume around the propyl group of LEU99 can be spotted, actually indicating the positive effect of releasing the unstable water into the bulk.

For the benzamidine-based compound, the most prominent change occurs once again around the clashing hydrogen and oxygen atom, like before with the first benzylation. The explanation is also the same as before, where the additional aromatic ring alters the partial charge distribution within the ligand, polarizing the C-H-bond even more and thereby reducing the energetic penalty introduced by the close contact.

As one of the main goals of this chapter is to introduce a new tool to investigate biological binding processes and help design new ligands, the following subchapters are dedicated to the investigation of various roadblocks and vagueness which one might encounter in a real-world application of the method.

3.3.3.4 Influence of experimental uncertainties in the complex structures

Crystallography is often referred to as “voodoo” and "black magic"[244], and even the most carefully laid out experiments performed by well-trained experts lack reproducibility to a certain degree. It is common and by no means a sign of bad practice, that atoms, amino acids and smaller fragments of the complexes are missing in some structures within a ligand series, investigated by x-ray crystallographic experiments. This is often the case for parts that are in between to be too flexible and dislocated to be resolved from the crystal at all and stable enough to be modeled from the electron density field. Several examples can be found in **Table 3.6**, which had to be added to the structures beforehand to make them comparable to each other. For an accurate interpretation of the localized energies and, even more important, for a fair comparison between different ligands, such variations have to be controlled and aligned. This exceptionally true for ionic amino acids (e.g., arginine, histidine, lysine, aspartic acid, glutamic acid), as their erratic appearance and disappearance in the structure, would vary the net charge of the protein, which in turn has implications for the LFE and intermolecular energy as will be shown in this subchapter. The same is true for co-crystallized and resolved salt ions originating from buffers and crystallization agents. Depending on the protein and the crystallization conditions, those ions are prone to this unsteady behavior within a ligand series let alone between different publications, researchers, and labs. This is actually also the case for one of the sodium ions in the here used ligand series. In the structure 3DUX (not part of the ligand series and therefor not used in this work), one of the two sodium ions which can be found in all of the other structures is not resolved and instead, a Na^+ is located in the vicinity of the thrombin light chain. (This does not apply to catalytic metal ions and other functional ions as they are usually strong enough bound to be resolved and their absence most often qualifies for a failed experiment which in turn is not published.)

Effects of different net charges of the host system on localized energies

The effect of a changed net charge of the host system is here demonstrated by adding a sodium ion to generate a net positively charged host and a chlorine ion for a net negatively charged host. To find suitable coordinates for the extra ions, a 3D RISM calculation (same parameters as described in 3.3.2) with a 1M aqueous sodium chloride solution[245] as solvent was performed and the additional ions were placed at the coordinates with the highest density of the corresponding species (see **Figure 6.1**). To explore the influence of the ion positioning, one of the experimentally resolved sodium ions gets deleted and replaced by the mentioned additional ion, effectively moved to different coordinates. For consistency reasons and facilitating easy comparison of structural features to the neutral reference host system, the already introduced thrombin complex was chosen to investigate both influence factors. However, sodium is discussed in the literature to influence the activity of thrombin via conformational changes.

Sodium is bound to thrombin in the allosteric Na^+-binding site[246–249], located between the 186-loop and the 220-loop, in which for all the here treated complexes a sodium ion is bound. This specific ion is kept untouched in all of the here described experiments, to avoid the introduction of artifacts. Nonetheless, this subchapter should mainly be seen as a methodical demonstration of the concept and not as a biochemical assessment of the role of sodium ions for the functionality of thrombin.

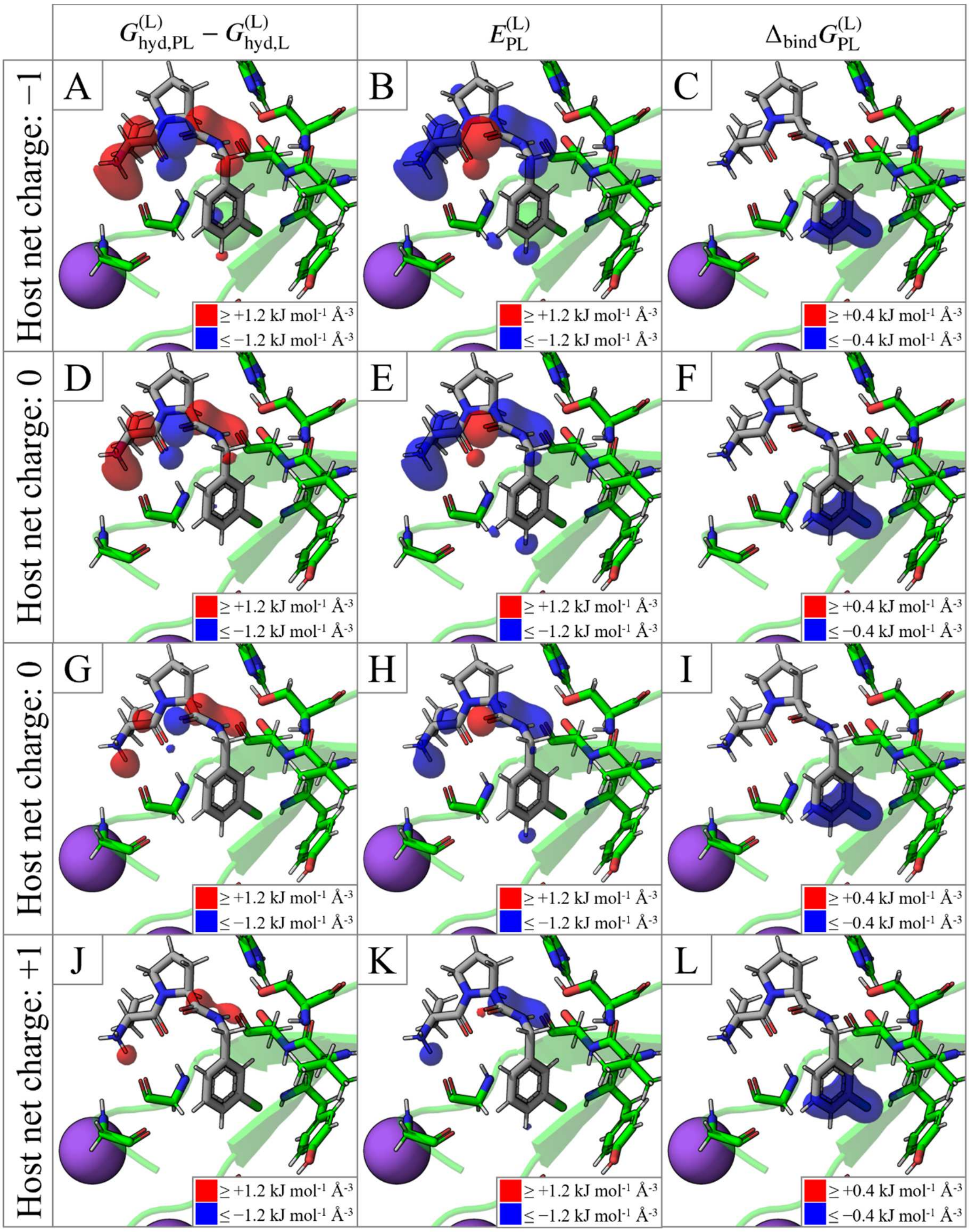

Figure 3.14 Comparison of different net charges of the host system as well as the effect of a relocated charge. The three columns show from left to right the on the ligand localized desolvation penalty, the intermolecular energy, and binding free energy. The first row shows the 2ZFP complex with an additional chlorine ion (green), demonstrating a host net charge of -1. The second row shows the complex with a relocated sodium ion, while the third row shows the unchanged 2ZFP complex as a reference. The bottom row shows the binding process for a net positive charged host system with one additional sodium ion (not visible). The raw data can be found in the electronic appendix under 3.3/3.3.3/IonPlacement/.

The effect of a changed net charge of the host system is shown in **Figure 3.14**, which shows the ligand perspective for a net negatively and net positively charged host system as well as the already introduced neutral variant for direct comparison. The focus is here on the ligand for the simple reason, that almost all changes are located on it while the protein perspective stays unchanged, regardless of its net charge. This does not come as a surprise, since from the protein perspective, only little changes throughout this experiment series, as the ligand stays the same after all and the addition and removal of ions is sufficiently far from the binding site.

Upon visual inspection of **Figure 3.14** two, seemingly contradictory observations stand out. On the one hand, the desolvation free energy and intermolecular energy show a clear trend of decreasing volume size by constant cutoffs with growing charge and, on the other hand, visually indistinguishable volumes in the binding free energy figures of the panel.

The first phenomenon is especially, but by far not exclusively, present around the NH_3- and amide-groups and can also be found in the significantly higher absolute LFE and intermolecular energy values on the atoms of those groups, listed in **Table 6.4**. It is important to note that this trend is only limited to the intensity and does not change the signs of the energy values nor the relations between atoms and groups, visualized by the coherent positioning and arrangement of the corresponding volumes. In this regard, the trend is remarkably consistent throughout the molecule, as it affects all atoms and even follows a linear function through observed charge range for almost all of them. Due to the fact that this is true for both the desolvation free energy and the intermolecular energy, only in opposite directions, it becomes clear why the atom wise binding free energy values and consequently also the aforementioned volumes are as consistent for all host charges as they are depicted in **Figure 3.14**, explaining the second phenomenon.

However, the very reason for the more dramatic response of the ligand upon binding in host systems with lower charge is not yet explained. To do so, it is worthwhile to have a look on the way the desolvation free energy is calculated. In principle, it is referenced to the net positive charge ligand and describes the difference in the solvation free energy upon binding by subtracting the reference from the complex state. The more different those states are, the higher the magnitude of the desolvation energy will become. For example, the environment for the ligand bound to a net negatively charged host system is, farther away from its reference state as a net neutrally charged one, resulting in stronger differences resulting in a higher desolvation penalty. The effect is reversed for net positively charged hosts, which brings the complex environment nearer to the ligand reference state.

Regarding the second term of binding free energy, it is important to note that the intermolecular energy is negative for the binding process in all three charge states of the host. From a global point of view, the introduction of a net charge into the host system by an additional ion just shifts these values either up in case of a positive charge or down in case of a negative one. This is fairly expected, as the ligand itself is positively charged and for a sufficiently distanced interaction site, it can be approximated as a single point charge, since the relative differences in the distance to the ion between the ligand atoms become comparably small and can therefore not overcome any general, molecule wide trend. On an atomistic-level the in **Table 6.4** and **Figure 3.14** observed trends are also quite comprehensible, since it again comes down to a simple interaction with a single additional site. Once again, the effect is mostly limited to a simple offset to each atom, depending on the sign and magnitude of its partial charge, while its special position plays only a minor role for the reason just explained.

Here one can also find a second, more technical motivated explanation for the very efficient cancelation effect between $\Delta G^{(\mathrm{L})}_{\mathrm{hyd,PL-L}}$ and $E^{(\mathrm{L})}_{\mathrm{PL}}$, in addition to the physical one outlined above. The mentioned offset generated by an additional charge occurs not only in the interaction between host and guest but also eventually affects the interaction between solvent and solute and therefore influences the with 3D RISM calculated μ^{ex} and consequently the localized free energies. A sufficiently far distanced ion, so that it has only a very limited influence on the solvent structure itself, shifts the LFE in the same way it does it with the interaction energy by giving each atom a simple positive or negative offset, depending on its partial charge. This, in turn, leads to only very minor differences in the localized and total binding free energy as both effects are opposing each other (see **Table 6.4**), leading eventually to indistinguishable volumetric representations throughout the charge experiments.

Importance of ion positioning in the preprocessing

As already mentioned, the positioning of additional ions is of some significance. If they are placed too close to the binding site, they could alter the solvent structure in relevant areas, which would affect $\Delta G^{(\mathrm{L})}_{\mathrm{hyd,PL-L}}$ and $E^{(\mathrm{L})}_{\mathrm{PL}}$ in such ways that the relations between atoms could be changed, introducing unwanted artifacts. Thus, to limit their influence to mere offsets, they have to be placed sufficiently far away from the binding site, effectively normalizing the energies and making comparisons possible. This shall here be demonstrated by effectively moving one of the sodium ions experimentally found in 2ZFP to the far side of the protein. This new position is taken from 3DUX, in which the second sodium ion not bound to the sodium binding site is not crystallized in the usual place for this experimental series but on the other side of the protein. This setup was chosen to emphasize that this kind of experimental hiccup is indeed a common part of the scientific process and to demonstrate that such considerations are of importance for the applicability of the method.

One of the most noticeable features of the results shown in **Figure 3.14** D to F is their resemblance to the experiment conducted with the net negatively charged host system. Especially, $\Delta G^{(\mathrm{L})}_{\mathrm{PL-L}}$ and $E^{(\mathrm{L})}_{\mathrm{PL}}$ are here of special interest since their similarity is, other than the virtual identicalness of the binding free energy for reasons explained above, somewhat unexpected. After all, the net charge of the host system is still zero and the naïve expectation would have been a close resemblance to the experimentally found standard. The nonetheless present differences are once again limited to the size of the volumes and therefore to the intensity of the LFEs while the relations between the atoms are unchanged, compared to the net negative host experiment and therefore to all three others too.

The explanation for the described observations is also very similar to the previous line of argument for the differences between the charge experiments. The removal of a cation near to the binding site introduces an offset to $\Delta G^{(\mathrm{L})}_{\mathrm{PL-L}}$ and $E^{(\mathrm{L})}_{\mathrm{PL}}$ as seen before, which in turn can not be fully compensated by the now more distant ion, ultimately having the same effect as the introduction of an additional anion. With this in mind, it becomes clear why the localized binding free energies are yet again indistinguishable compared to the other experiments, at least in the case of the volumetric representation. Furthermore, even though some small changes certainly occur due to the long-ranged nature of the Coulomb energy, it appears that the relocation of the ion does not significantly change the solvent structure of the binding site. There are some smaller variations in the LFE values (see **Table 6.4**), but since there is usually not more room in the binding site for more than two, maximum three solvent shells, the solvent distribution function in the binding site is mostly dominated by the atoms in the immediate surrounding and only very little by ions further out, like it is the case here.

3.3.3.5 Summary of the subchapter

This subchapter focused on a practical application of the LFE method to a realistic example, in this case, a thrombin inhibitor series.

Especially in the context of medicinal chemistry and protein-ligand binding the visual inspection and interpretation of results is of great importance. Therefore, the first question asked in this subchapter was how the LFEs are best visualized such that the provided information can be naturally and quickly comprehended. Next to the naïve color-coded atom-wise method, this search led to a volumetric representation, which checked all the predefined requirements and gave easily interpretable figures.

With a suitable visualization method established, the protein-ligand interactions occurring throughout the compound series could be discussed in full detail. In doing so, the binding situation was investigated from the protein, ligand, and combined perspective, as well as for each contribution to the binding free energy separately. Separating and observing the host-guest binding on multiple, gradually finer layers, right down to the atomic level, demonstrated the significant knowledge gain possible with the LFE approach. This knowledge could, for example, be used to identify promising interactions for compound optimization or to understand potential drug resistance. Nonetheless, the presented method has its limitations and how much the neglection of the solute entropic terms and the rigid body assumption influences the final results is not easily to be determined. However, the only little correlation between the LFEs and the experimental trend regarding the addition of phenyl rings to the S3/S4 pocket can be seen as an indication. One could speculate that the lack of conformational diversity in those regions limits the expressiveness of the LFEs on these rings. While no solution to the missing entropic terms or PMV correction of the hydration free energy, the coupling with MD simulations could alleviate some of these problems, as described previously.

Since the basis for calculations as they are presented here, are usually experimental structures of the complexes in question and therefore suffer from inevitable uncertainties, the robustness of the visualization and interpretation against variations in the net charge of the host was also checked in this subchapter. While a change in this regard does indeed lead to a global shift in the LFEs, it does not alter the relations between groups. This means that as long as all members of a series are treated the same, the net charge and ion positioning is of small significance to the overall interpretation.

3.3.4 Free energy derivatives with respect to the non-bonded force field parameters applied to protein-ligand binding

In the development of new protein inhibitors, one often does not have to start from the ground up without any knowledge about the complexed state but can sample from a wide variety of compounds already binding to the target. Those can come from previous research on the topic, in the form of published structures (as was the case in the compound series used here) or already registered and buyable drugs. Alternatively, they come from nature, for example, as natural substrates of the target protein or known pharmacological active substances from plants and fungi[250]. Especially in the latter case, where no previous targeted optimization in the area of selective inhibition or activation other than natural selection via evolutionary processes was performed in the past, the obvious question arises, how the compound can be modified to make it bind even stronger and more selective towards the target.

In an attempt to answer such questions, often so-called structure-activity-relation (SAR) studies are performed, in which several structurally similar compounds are screened against the target (sometimes also against some off-targets[251,252]) to derive a relationship between molecular features and the affinity/selectivity towards the target protein. However, such studies are extremely expensive in means of time requirements, and for the high monetary costs involved in the synthesis of the test compounds, the acquisition of the protein by either buying it from a supplier or producing it through biochemical means, and the measurements themselves, for which highly trained personal and specialized equipment is required. For those reasons, it is of utmost importance for industry and academia alike to select the test compounds very carefully before entering this costly process and perform the task iteratively in small batches to make informed decisions from round to round.

The here proposed method of calculating the derivative of the binding free energy and its terms with respect to (w.r.t.) the non-bonded forcefield parameter can help with such design decisions by driving exploration through sparking new ideas for optimization and steer exploitation by giving directions and reasoning. Those strengths combined can help scientists develop optimal binders in fewer iterations and with fewer failing compounds along the way.

3.3.4.1 Introduction to the concept and interpretation of free energy derivatives

To demonstrate the concept and explain the interpretation of FEDs, there are all three types of derivatives of the solvation free energy shown in **Figure 3.15** for the 2ZFP compound in free solution. Each column shows one type, and the upper and lower row shows both visualization methods, already introduced in 3.3.3.1 and 3.3.3.2. For depicting negative and positive derivatives, the same color code as the one for the LFE was deployed, red for positive, blue for negative numbers. However, the meaning of them is wildly different and even from derivative to derivative, they cannot be interpreted the same. The alternative, using three new and independent color ranges for each derivative would have been quite confusing and hindering in the assessment of the results. Since the sensitivity of the solvation free energy is also vastly different between the nonbonded force field parameters, the range in which the resulting values are distributed is also parameter specific and therefore needs its own range, respectively cutoff for the color-coding to give a useful visualization.

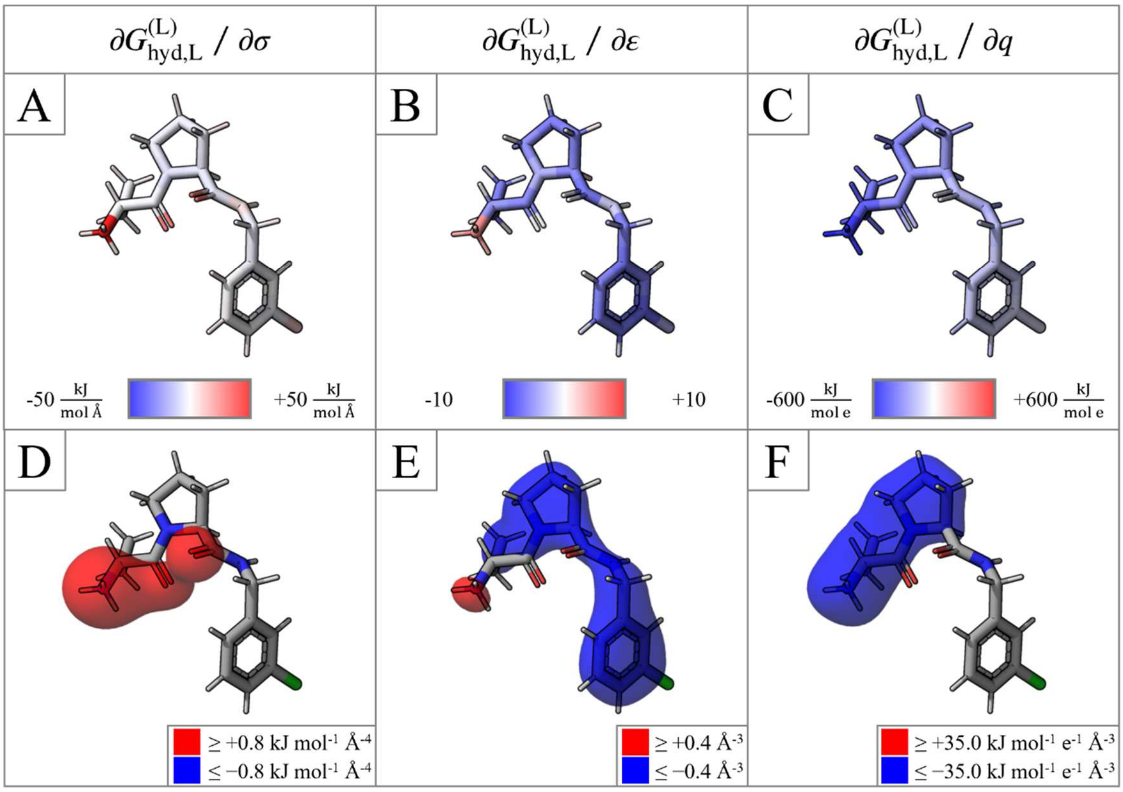
$\partial G^{(L)}_{\mathrm{hyd,L}} / \partial\sigma$
$\partial G^{(L)}_{\mathrm{hyd,L}} / \partial\varepsilon$
$\partial G^{(L)}_{\mathrm{hyd,L}} / \partial q$
A
B
C
$-50 \frac{\mathrm{kJ}}{\mathrm{mol\,Å}}$
$+50 \frac{\mathrm{kJ}}{\mathrm{mol\,Å}}$
-10
+10
$-600 \frac{\mathrm{kJ}}{\mathrm{mol\,e}}$
$+600 \frac{\mathrm{kJ}}{\mathrm{mol\,e}}$
D
E
F
🟥 $\geq +0.8$ kJ mol^{-1} Å^{-4}
🟦 ≤ -0.8 kJ mol^{-1} Å^{-4}
🟥 $\geq +0.4$ Å^{-3}
🟦 ≤ -0.4 Å^{-3}
🟥 $\geq +35.0$ kJ mol^{-1} e^{-1} Å^{-3}
🟦 ≤ -35.0 kJ mol^{-1} e^{-1} Å^{-3}

Figure 3.15 Derivatives wrt. to the nonbonded force field parameters of the solvation free energy of the 2ZFP compound in its binding conformation. The upper row shows the atom-wise visualization of the derivatives with cutoffs chosen to include the highest and lowest values and are centered around zero. The bottom row shows the volumetric visualization for which cutoffs were chosen, to demonstrate the most important aspects of FEDs.

All derivatives with respect to the non-bonded force field parameter were calculated for the ones of the solute, and therefore the correct annotation would be σ_α, ϵ_α, and q_α, where the subscript α denotes solute sites. For easier readability, those subscripts are omitted in the following, and since the derivatives with respect to the solvent sites are not of interest here, there is no chance for any confusion.

Derivative of the solvation free energy with respect to the σ-parameter

The σ-parameter of the Lennard-Jones potential can be seen as a scaling factor for the size of an atom, with bigger values for σ corresponding to larger atoms. The derivative $\partial G_{\mathrm{hyd}}/\partial\sigma$ can therefore be interpreted in which way the solvation free energy would change, if one would change the size of the atom at hand. In the 2ZFP compound the most extreme values for $\partial G_{\mathrm{hyd}}^{(\mathrm{L})}/\partial\sigma$ are focused on the chlorine (10.04 kJ mol^{-1} Å^{-1}), amid oxygen atoms (22.17 and 24.92 kJ mol^{-1} Å^{-1}), and on the nitrogen of the amino group (42.46 kJ mol^{-1} Å^{-1}), the rest of the molecule is compared to those almost neutral and values range around -5 and 5 kJ mol^{-1} Å^{-1}. The fact that it is mostly the nitrogen and not the outward-facing hydrogen atoms of the amino group lighting up is explained by the size difference of those atoms. With its σ of 3.25 Å the nitrogen atom simply swallows the smaller and tightly bond (~1 Å) hydrogen atoms with their σ of 1.07 Å. Positive values indicate close contacts between solute and solvent, and thereby high solvent density in close proximity to those atoms. In the current example, the most positive derivative values can be found on atoms strongest polarized throughout the compound with high absolute partial charges. This leads to strong Coulomb interactions with the polar solvent, drawing the corresponding site density closer towards the atom until the Lennard-Jones repulsion term negates any further energy gain, finally leading to positive values for $\partial G_{\mathrm{hyd}}^{(\mathrm{L})}/\partial\sigma$. The strong influence of charge on density functional theory methods like 3D RISM is well documented[68,178,253], and the effective volumetric contraction found for net charged particles is in line with experimental findings, such as the higher density of water with higher salinity.

Regarding the interpretation of $\partial G_{\mathrm{hyd}}^{(\mathrm{L})}/\partial\sigma$ one can find general observations. The first is, that negative values are limited in their magnitude while positives are not restricted in that regard thanks to the diverging and very steep repulsive term of the Lennard-Jones potential. The second, somewhat related observation, is that, in theory, one can always lower the solvation free energy for a charged particle in polar solvents by shrinking its size. Followed through, this would lead to so-called point charges, which have the lowest theoretical solvation free energy (for 3D RISM, this would be minus infinity). Obviously, this does not have real-world correspondence, but for interpreting the derivatives and understanding their meaning, it is useful to have this tendency in mind.

Derivative of the solvation free energy with respect to the ε-parameter

The ε parameter of the Lennard-Jones potential can be interpreted as the capability to form an induced dipole or polarizability of an atom. The higher this value gets, the deeper is the energy well of the Lennard-Jones potential, but since the parameter applies to both, the attractive and repulsive term, it also scales the energy barrier.

In the currently discussed example, both visualization methods let the molecule appear blue, suggesting a mostly negative derivative. However, the apparent size of the hydrogen atoms is quite small in this particular drawing style, making this a bit misleading. In fact, many of them have positive derivatives, since they are closest to the solvent and therefore in close contact, shielding most of the heavy atoms from repulsive short-ranged interactions. Those are within the energy well of the LJ potential and would profit from a higher ε parameter, carrying mostly negative derivatives. An exception to this are the heteroatoms already known from the σ-derivative, especially the amino group, which have a solvent density in the repulsive zone. Here an increased ε would yield a higher solvation free energy.

In contrast to the σ-parameter, there is no good analogy like the size of an atom for the interpretation of $\partial G_{\text{hyd}}^{(\text{L})} / \partial \varepsilon$. Besides this, the solvation free energy is also the least sensitive to changes in the ε-parameter, and any potential change would have only marginal effects. This is to keep in mind for the interpretation of protein-ligand complexes, where this phenomenon is also occurring. Although this seems a bit underwhelming at first glance, it is indeed a great advantage for any optimization based on the derivatives. Modification of the molecules cannot be done for each parameter independently but have to be done through adding, subtracting, or replacing atoms and since each atom type brings its own combination of the σ- and ε-parameter, it is, therefore, easier to change the size of certain parts of the molecule based on $\partial G_{\text{hyd}}^{(\text{L})} / \partial \sigma$ without worrying too much about the consequences in regard to $\partial G_{\text{hyd}}^{(\text{L})} / \partial \varepsilon$.

Derivative of the solvation free energy with respect to the partial charges

The free energy derivative with respect to the partial charge does not behave as the other two parameters already introduced. The sensitivity of $G_{\mathrm{hyd}}^{(\mathrm{L})}$ is much higher compared to the LJ parameters, and consequently, the necessary range and cutoff for the visualization are orders of magnitude greater. Furthermore, here every atom of the compound carries a negative derivative, coloring the whole molecule in blue. The most negative values are found around the amine group, with a relatively gradual distribution towards it. To explain this, one can come back to the theoretical model of point charges. The stronger a sizeless particle is charged, the lower its solvation free energy will be. Extrapolating from this back to the compound at hand and keeping in mind that $\partial G_{\mathrm{hyd}}^{(\mathrm{L})} / \partial q$ is independent from the LJ potential, it makes sense that a higher partial charge would decrease the solvation free energy.

For a more phenomenological explanation, one can imagine that each solute will arrange its surrounding solvent density so that the resulting system has the lowest possible energy under the given conditions. For the given net positively charged molecule in water, this will involve a high oxygen density in near proximity. And since the here discussed derivatives are calculated for one such static state, increasing the charge would foster this arranged interaction even more, albeit not for every atom equally, focusing on those already charged the strongest.

Interpreting $\partial G_{\mathrm{hyd}}^{(\mathrm{L})} / \partial q$ is less straightforward than it is with the LJ counterparts. One reason for that is, that partial charges are a less well-defined concept and the true distribution is difficult to determine (assuming that there is a true distribution). For common amino acids, most for the task adequate force fields provide a given set of charges for each atom type, regardless of the environment and conformation. This is fundamentally different from those for small molecules, where partial charges get calculated for every compound individually. To do so, various different methods can be used, often utilizing quantum mechanics in one form or the other (see 3.2.2.1 for an in-detail analysis on how the choice of method influences LFE values). In addition to the used method for calculating the final set of charges, they also depend on the exact conformation used as input. Furthermore, potential changes done to the molecule in one position, like they are proposed by the FEDs, likely change the partial charge distribution everywhere throughout the compound. Thus, modifications in structure do affect the rest of the molecule not only via an altered solvent distribution, as is the case for the LJ parameter but also via intramolecular effects. These effects make it hard to derive clear optimization pathways from $\partial G_{\mathrm{hyd}}^{(\mathrm{L})} / \partial q$. Promising strategies most likely focus on broader trends and dismiss local effects, as they are likely to change with slightly altered input conformations or parameters.

3.3.4.2 In-depth analysis of the binding free energy derivatives of a thrombin inhibitor

As seen in the last subchapter, the atom-wise visualizations of the FEDs are, in general, more comprehendible than the ones of the corresponding LFE counterparts. They tend, even though not completely, much less to the notorious alternating pattern and are much easier to interpret. Nonetheless, the volumetric visualization is still a useful tool since the involved terms, especially the Coulomb potential, are acting over distances, and potential changes in the molecular structure often affect full regions rather than points in space. However, since the in a drug development context most interesting effects likely affect single sites, it is the default visualization method for this subchapter.

The derivatives of the binding free energy can be calculated with

$$\frac{\partial \Delta_{\mathrm{bind}} G_{\mathrm{AB}}^{(\mathrm{X})}}{\partial P} = \frac{\partial G_{\mathrm{hyd,AB}}^{(\mathrm{X})}}{\partial P} - \frac{\partial G_{\mathrm{hyd,X}}}{\partial P} + \frac{\partial E_{\mathrm{AB}}^{(\mathrm{X})}}{\partial P}, \tag{3.17}$$

where X is a placeholder for the localization either on the host (A) or guest (B) and P can be one of the LJ-parameter σ and ϵ, or the partial charge q. The equation follows the same logic as it was used for the localized binding free energy and also yields a host-guest separated localization.

While the resulting branches were not independent from each other in the LFE approach, meaning the total binding free energy could only be derived from summing both branches, there are so in the FED approach. The derivatives on one of both partners stand for the tendency of the full system to react according to sign and magnitude of the particular value.

Due to the nature of the data source, only the ligand perspective is discussed in the following. Nonetheless, a similar analysis could also be made for the protein. Such could be, for example, of interest for mutation-induced drug resistance or protein editing. This is, however, is beyond the scope of this work.

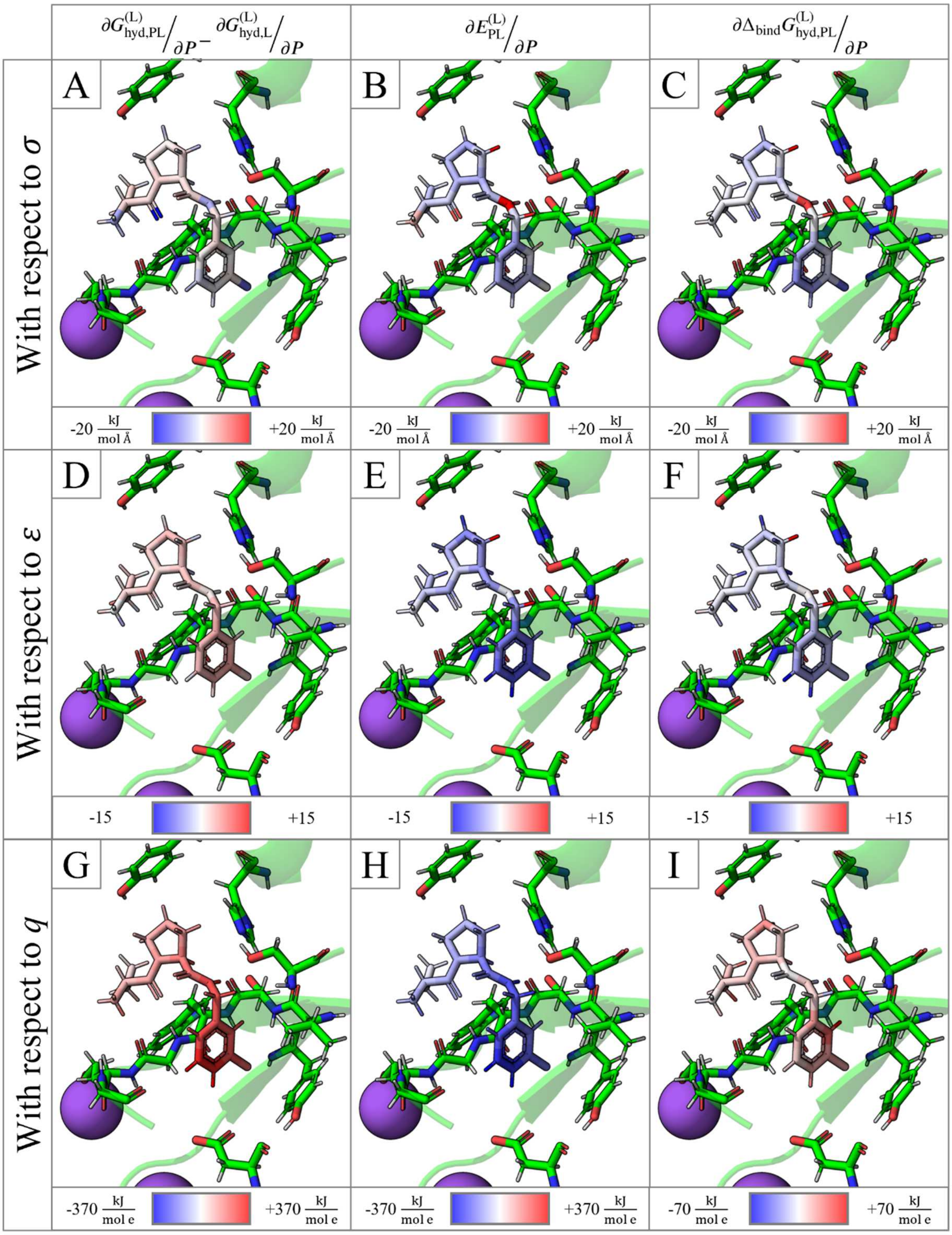
$\partial G^{(L)}_{hyd,PL} / \partial P - \partial G^{(L)}_{hyd,L} / \partial P$
$\partial E^{(L)}_{PL} / \partial P$
$\partial \Delta_{bind} G^{(L)}_{hyd,PL} / \partial P$
With respect to σ
With respect to ε
With respect to q
A
B
C
D
E
F
G
H
I
-20 $\frac{kJ}{mol\,Å}$
+20 $\frac{kJ}{mol\,Å}$
-15
+15
-370 $\frac{kJ}{mol\,e}$
+370 $\frac{kJ}{mol\,e}$
-70 $\frac{kJ}{mol\,e}$
+70 $\frac{kJ}{mol\,e}$

Figure 3.16 Free energy derivatives (FED) of the ligand atoms of the 2ZFP thrombin complex. The first row (A-C) shows the derivatives with respect to the Lennard-Jones σ-parameter, the second (D-F) those with respect to the ϵ-parameter, and the third (G-I) with respect to the partial charge q. The first column shows the derivatives of the desolvation penalty, the second the derivative of the intermolecular energy, and the third the derivatives of the binding free energy. The coloring ranges are adapted to the minimum and maximum values of the particular derivatives but clip outliers (B, C, E, and F). The raw data can be found in **Table 6.5** and in the electronic appendix under 3.3/3.3.4/Loc/2ZFP/.

Derivatives with respect to partial charges: desolvation penalty

When calculating the derivative of the desolvation penalty w.r.t. the partial charges, every atom carries a positive value. This can be interpreted in the sense that a more negatively, or in other words, less positively charged compound would lead to a lower desolvation penalty. In itself, this is a relatively direct result because, generally spoken, neutral molecules have a smaller desolvation penalty to pay since they tend to have a higher solvation free energy. To achieve neutrality, the net charge of the compound hand must be decreased by one, following the direction of the derivatives. The lower desolvation penalty is one, but often not the only reason, for modern drug design to focus on non-charged compounds for their often advantageous ADME properties.[227] The red coloring in **Figure *3.16*** G is most pronounced on the S1-pocket binding part of the ligand, which initially seems quite curious since the corresponding, most blueish derivatives on the free ligand can be found around the amino group, far from the S1-pocket. The unexpected shift is most likely caused by the negatively charged ASP189, which is dominating the S1-pocket. Its negative charge attracts hydrogen density between protein and ligand, where is indeed enough room for a water molecule, confirmed by crystal structures from multiple studies (PDB-structures: 2ZFP, 2ZC9, 3DHK[80], 6ZUG[227], 4UDW[254]) and shown in **Figure *3.17***.

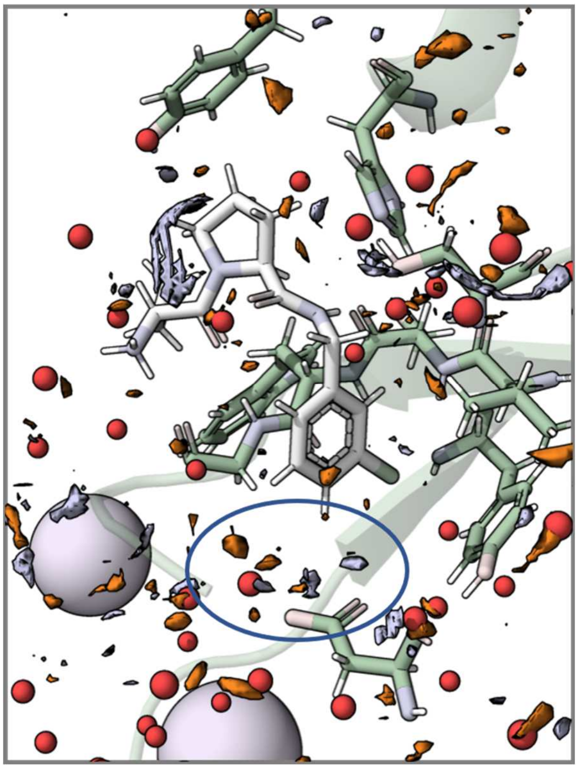

Figure 3.17 Oxygen (orange) and hydrogen (blue-white) distribution around the binding site of the thrombin complex 2ZFP. The complex is shown in desaturated colors. The experimentally resolved water molecules from the x-ray crystal structure of 2ZFP are shown as red spheres. The solvent density presumably causing the positive derivative of the desolvation penalty with respect to partial charges is highlighted by a blue circle.

The positive derivatives of the solvation free energy of the ligand in complex with thrombin is likely reflecting this interaction. In contrast, the amino group carries strong negative derivatives, just like the free ligand, because of its solvent accessibility, partially compensating the derivatives of the free ligand, finally leading to the observed gradient.

Derivatives with respect to partial charges: intermolecular energy

The derivative of the intermolecular energy between ligand and protein w.r.t charge (**Figure *3.16*** H) shows a clear trend and is therefore straightforward interpretable. The net negative charged aspartic acid in the S1 pocket leads to a negative derivative on every atom of the ligand with increased intensity in its direct proximity. A more positive charge, especially in this region, leads without a doubt to a stronger interaction, which was also experimentally proven[237,255] and theoretically shown in this work through the benzamidine based compounds (see **Figure 3.12** and **Figure 3.13**).

Derivatives with respect to partial charges: binding free energy

Knowing that compounds introducing a positive charge in the S1 pocket increase the binding affinity by a significant margin, one is naturally compelled to look for negative derivatives w.r.t q in this general area (**Figure 3.16** I). However, all derivatives found in this part of the ligand are positive, not confirming this very reasonable assumption. This is because of the same effect causing $\partial\Delta G^{(\mathrm{L})}_{\mathrm{hyd,PL-L}}/\partial q$ to be exclusively positive. 3D RISM assigns hydrogen density in the area between ligand and ASP189, giving this part of the ligand a net positive derivative. This example testifies to a fundamental limitation of the FED approach to compound optimization. Interactions that are not yet established or are at least formed in some form of protostage (a thinkable example would be an alcohol group, forming a hydrogen bond with ASP189. Some other candidates are explored by Lumma et al. [242]) will most likely not be found by analyzing FEDs. They should be thought of as a means to find potential point-wise changes, strengthening preexisting constructive effects and weakening deconstructive ones.

Such points are, for example, the ortho-hydrogen and carbon in the *m*-chlorobenzyl, which stand out as significantly more strongly red-colored within the ring. A possible explanation for their unusual high positive derivatives can be found in the apo form of thrombin (2UUF[256], 3D49[257], 2GP9[258]). In the absence of a binding ligand, one can repeatedly find a water molecule in the same position within the binding site, in which the positive $\partial\Delta_{\mathrm{bind}}G^{(\mathrm{L})}_{\mathrm{PL}}/\partial q$ values are located. This water is not flag as particularly unstable in the literature[227,255], but gets replaced anyway, thereby increasing the desolvation penalty. So it seems that an atom or a group, being more locally polarized in this area, making it effectively more similar to a water-oxygen, could decrease the binding free energy. (A similar argument could be made for the γ-carbon and hydrogen of the proline in the S2 pocket, but much less pronounced since the replaced water is in this case known to be unstable[227,255]). Indications towards this initial idea are the orientation of the nitrogen, oxygen, and sulfur in several ligand fragments in the S1 pocket, which orient themselves in the described way[259–261]. However, to verify this hypothesis, there are more and deeper investigations needed.

This concept of using the FEDs points-wise changes is demonstrated later in this chapter, comparing the through FEDs predicted trends with experimental ligand series.

Derivatives with respect to the ϵ-parameter: desolvation penalty

The derivative of the desolvation penalty w.r.t. the ϵ-parameter (**Figure 3.16** D) is on almost all atoms positive and reflects the findings from the discussion about the derivative of the solvation free energy of the free ligand quite accurately. To decrease the desolvation penalty, either the interaction with the solvent in the bound state must be strengthened or weakened in the unbound state. And since the magnitude of those interactions is much higher in the unbound state, the derivative is following this for the desolvation penalty.

Derivatives with respect to the ϵ-parameter: intermolecular energy

One can assume that the ligand will settle in the binding site during crystallization in the conformation with the, under consideration of solvation effects, lowest possible intermolecular energy, which means most atoms end up in the attractive part of the LJ-potential. And since the derivative of the LJ potential w.r.t the ϵ-parameter shares its zero-crossing with the original potential (see **Figure 2.2**), $\partial E_{\mathrm{PL}}^{(\mathrm{L})}/\partial\epsilon$ shows directly which atom has a total negative or positive LJ energy. With this in mind, it is easy to see why most of the ligand in **Figure 3.16** E is colored blue. Nonetheless, in the S2 pocket, two red-colored hydrogens are clashing with the protein. This is something already observed and discussed in the LFE analysis and is therefore not explained in further detail here.

The overlap between $\partial E_{\mathrm{PL}}^{(\mathrm{L})}/\partial\epsilon$ and $\partial E_{\mathrm{PL}}^{(\mathrm{L})}/\partial\sigma$ in contrast, and especially the absence of it, is worth mentioning. Because the positive part of the $\partial E_{\mathrm{PL}}^{(\mathrm{L})}/\partial\epsilon$-function is already within contact distance (see **Figure 2.2**) of two atoms, every positive derivative w.r.t the ϵ-parameter has a corresponding positive derivative w.r.t the σ-parameter on the same atom, always leading to a positive derivative w.r.t. the σ-parameter. Conversely, a positive value in $\partial E_{\mathrm{PL}}^{(\mathrm{L})}/\partial\sigma$ does not have to have a positive correspondence in $\partial E_{\mathrm{PL}}^{(\mathrm{L})}/\partial\epsilon$. The zero-crossing of its function is at a larger distance r and is, therefore, more sensitive to such close contacts.

Derivatives with respect to the ϵ-parameter: binding free energy

The most noticeable atoms regarding the derivative of the binding free are already known from the last paragraph, with two exceptions. The meta- and para-hydrogen of the *m*-chlorobenzyl carry comparably high negative values in **Figure 3.16** F. They underline the argument made in the part of the derivatives of the desolvation penalty and binding free energy w.r.t the charge, where the high absolute values in the ring were found to possibly be caused by accumulated solvent-hydrogen density in between ligand and protein, with which the two atoms likely interact.

Derivatives with respect to the σ-parameter: desolvation penalty

Regarding the derivative of the desolvation penalty w.r.t. the σ-parameter, blue colored atoms in **Figure 3.16** A indicate that a group or an atom with a larger radius than the status quo would decrease the desolvation penalty by either increasing the solvation free energy of the ligand or decreasing it for the complex. Red-colored atoms are indicating the inverse.

Analyzing the coloring in **Figure 3.16** A, the ligand appears mostly neutral with a slight tendency towards smaller radii. At first glance, this may seem relatively unexciting, but at least for the in the S2 pocket binding proline-like substructure of the ligand, this is actually noticeable. In the apo form of the protein, here one can find a thermodynamically unstable water[227,255], which gets replaced by the ligand upon binding. The fact that $\partial\Delta G^{(\mathrm{L})}_{\mathrm{hyd,PL-L}}/\partial\sigma$ is near zero on atoms in this region, shows that the ligand is replacing the water almost perfectly regarding the difference in solvation free energy and the spacial requirements of it. In the discussion of $\partial\Delta_{\mathrm{bind}}G^{(\mathrm{L})}_{\mathrm{PL}}/\partial q$ it is shown that this is not necessarily the case for all non-bonded force field parameters.

Derivatives with respect to the σ-parameter: intermolecular energy

The mostly blueish coloring in the visualization of $\partial \mathrm{E}^{\mathrm{inter}}/\partial\sigma$ in **Figure *3.16*** B is somewhat misleading at first glance. Due to the licorice drawing style of the molecule, heavy atoms have a higher visual weight than hydrogen atoms do have. However, those hydrogen atoms are the ones, in most cases at least, in direct contact with the solvent or binding partner. A negative derivative on a carbon atom can easily be accompanied by a hydrogen atom already interacting in the repulsive part of the LJ-potential and thereby carrying a high positive derivative w.r.t σ. The following red coloring is visually underrepresented, though, giving the impression of a general tendency towards higher atom radii. This shows once again the importance of how results are visualized, especially for methods like LFE and FED. In this particular case, a volumetric visualization might have been better suited, but with the problem in mind, a reliable analysis is still possible.

Nonetheless, there are exceptions to the just described phenomenon. Some atom types in the GAFF force field have σ-parameters large enough to swallow their accompanying hydrogen atoms. This can be seen for example on one of the amid nitrogen in the compound, lighting up in a bright red color, indicating a very high $\partial \mathrm{E}_{\mathrm{PL}}^{(\mathrm{L})}/\partial\sigma$ and thereby a very close contact. This is also true for the amine nitrogen, although to a lesser extend. The S2 pocket, in general, is known to cause steric clashes between protein and ligand, as investigated in detail by Hillisch et al.[227]. By optimizing compounds in this regard, the binding affinity was increased, which suggests that the derivative of the intermolecular energy w.r.t the σ-parameter (and ϵ-parameter) could indeed be useful for lead structure optimization.

Derivatives with respect to the σ-parameter: binding free energy

The in the S1 pocket binding parts of thrombin inhibiting ligands are intensely studied and discussed in the literature and explored by multiple experimental compound series. This makes the distinctive blue color in **Figure 3.16** C of the derivative of the binding free energy w.r.t the σ-parameter of the chlorine atom especially interesting. As explained before, one of its functions is to replace a thermodynamically unstable water molecule and it is therefore considered crucial for the binding affinity of compounds equipped with this particular feature.

The apparent negative $\partial\Delta_{\mathrm{bind}}G_{\mathrm{PL}}^{(\mathrm{L})}/\partial\sigma$ suggest, that the binding free energy would be decreased even further by a slightly bigger atom or increased by a smaller atom in this position.

To test this hypothesis, there are experimental binding constants from multiple studies collected in **Table 3.9**. Additionally, there are derivatives for alternative decoration listed to check whether the trend is consistent within the FED method. These structures were artificially generated with the software Avogadro[262], and the Carbon-X distance is adapted to the corresponding bound element (H, F, Cl, and Br). Everything else was kept the same as for the regular 2ZFP complex treated above.

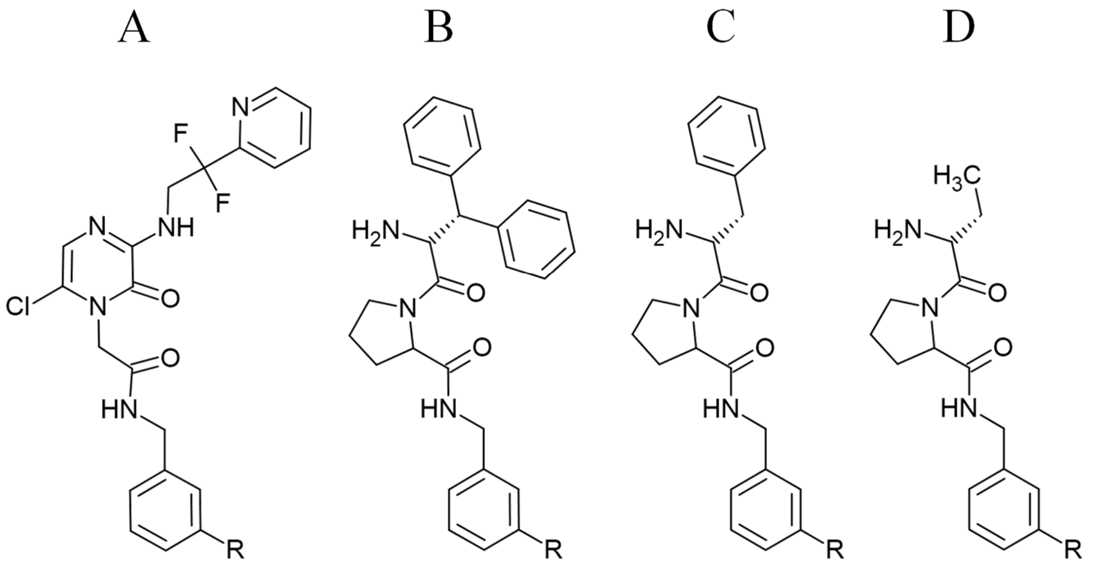

Figure 3.18 Base structures used in the experimental essays of Burgey et al.[259] (A), Lumma et al.[242] (B), Baum et al.[237] (C). Theoretical calculations were performed on compounds derived from base structure D, taken from Baum et al.[80]

Table 3.9 The table gives the experimental binding affinities of compounds from three different publications towards the thrombin serine protease. In the columns for structure D there are the derivatives of the binding free energy with respect to the Lennard-Jones σ-parameter and the corresponding LFE value given.

R	Struc. A[259] (K_i)	Struc. B[242] (IC_{50})	Struc. C[237] (K_i)	Struc. D	Struc. D
				($\partial\Delta_\mathrm{bind}G^\mathrm{(L)}_\mathrm{PL}/\partial\sigma$)	($\Delta_\mathrm{bind}G^\mathrm{(L)}_\mathrm{PL}$)
H	7.0 nM	130 nM	11.21 µM ± 6.70	1.22 kJ mol^{-1} Å^{-1}	2.58 kJ mol^{-1}
F	7.3 nM	-	3.99 µM ± 2.12	-13.94 kJ mol^{-1} Å^{-1}	-2.02 kJ mol^{-1}
Cl	0.26 nM	12 nM	180 nM ± 140	-10.02 kJ mol^{-1} Å^{-1}	-9.64 kJ mol^{-1}
Br	0.19 nM	10 nM	560 nM ± 147	26.19 kJ mol^{-1} Å^{-1}	-9.28 kJ mol^{-1}

Despite that, the base-structure A is structurally significantly different to B and C, and their data origins from three different publications, more than ten years apart, the trend in the experimental findings is mostly uniform. Compared to hydrogen and fluorine, chlorine and bromine are by an order of magnitude better-suited decorations concerning the binding affinity of the corresponding compound. However, within these two groups, the trends differ between the base structures and publications. For base-structure A, benzyl and *m*-fluorobenzyl are virtually indistinguishable from each other based on their K_i. At the same time, the binding affinity is increased by the fluorine decoration in base-structure C. Similarly, for A and B *m*-chlorine- and *m*-brominebenzyl are fairly similar, whereas they differ in C by threefold. The discrepancy between A and C could very well be explained by the quite large structural distance, while the different experimental methods (IC_{50} vs. K_i) and ten years of scientific progress may account for B and C. Either way, the experimental binding affinity seems to increase with the size of the atom in meta position of the in the S1pocket binding benzyl, with a possible, but also debatable, sweet spot at chlorine.

Nevertheless, the in silico calculated derivatives agree well with the reported values from base-structure C. Same is true for the also given LFE values. The derivative on the meta hydrogen is slightly positive, seemingly breaking the trend of decreasing values with shrinking atom size. However, this is indeed in line with theoretical and experimental results in the literature. The unchanged benzyl group does not replace the aforementioned described unstable water from the binding site, as it can be seen in the PDB structures of the in this work used amidine-based protein-ligand complexes and was proven by Abel et al.[255] through simulations and crystal structures. With this water still in place, the S1 pocket becomes quite crowded, making the presents of the water molecule even more disadvantageous, which explains the positive derivative on the hydrogen. Furthermore, the non-halogenic variant carries the highest LFE value of the here tested compounds, underlining the analysis of the derivative and being in line with the corresponding experimental results.

The fluorine atom carries the most negative derivative of the four test substituents, suggesting that the replacement of the unstable water takes place, although sub-optimally. The corresponding LFE value fits this explanation too. By releasing the water molecule into the bulk phase, a cavity is created in the S1 pocket, into which the ligand can be expanded. It is worth noting that the potential energy gain originates not primarily from the protein-ligand interaction, but from the increased solvation free energy and thereby decreased desolvation penalty of the ligand, caused by a larger volume while keeping the dipole moment of the ligand more or less the same.

The good agreement with the literature shows, when the calculated derivative is taken literally and the $\Delta\Delta_{\mathrm{bind}}G$ is calculated from it. Treating the *m*-fluorobenzyl as start and the *m*-chlorobenzyl as end state,

$$\begin{aligned}\Delta\Delta_{\mathrm{bind}}G_{\mathrm{approx.}} &= \frac{\partial\Delta_{\mathrm{bind}}G}{\partial\sigma_{\mathrm{start}}}(\sigma_{\mathrm{end}} - \sigma_{\mathrm{start}}) \\ &= -13.94\ \frac{\mathrm{kJ}}{\mathrm{mol\ Å}}\left(3.4709\ \mathrm{Å} - 3.1181\ \mathrm{Å}\right) \\ &= -4.88\ \frac{\mathrm{kJ}}{\mathrm{mol}}\end{aligned} \tag{3.13}$$

can be formulated. This is only a quick and rough estimation of the difference in binding free energy, but even so, it comes close to the experimental value of -7.68 kJ mol^{-1} ($\Delta\Delta_{\mathrm{bind}}G_{\mathrm{exp}} = -RT \ln K_{\mathrm{i,end}} / K_{\mathrm{i,start}}$), especially considering the relatively high standard deviations of the experimental values. The difference in the LFE values for the aromatic substituent recreates the experimental value even more truthfully with

$$\begin{aligned} \Delta\Delta_{\mathrm{bind}}G_{\mathrm{approx.,LFE}} &= \Delta_{\mathrm{bind}}G_{\mathrm{PL}}^{(\mathrm{Cl})} - \Delta_{\mathrm{bind}}G_{\mathrm{PL}}^{(\mathrm{F})} \\ &= -9.64\ \frac{\mathrm{kJ}}{\mathrm{mol}} - \left(-2.02\ \frac{\mathrm{kJ}}{\mathrm{mol}}\right) \\ &= -7.62\ \frac{\mathrm{kJ}}{\mathrm{mol}}. \end{aligned} \tag{3.13}$$

The negative derivative on the chlorine in the original compound was already mentioned before and initiated this more detailed case study. From this, two hypotheses can be formulated; for once, that a compound with a smaller σ-parameter would lead to a weaker binding affinity, and secondly, that a higher σ-parameter would lead to a stronger binding affinity. The first statement did hold true, backed by all three experimental studies. The second, that an atom with a bigger radius would strengthen the binding affinity does not have an explicit experimental confirmation, as described above. However, the derivative of the binding free energy is a theoretical concept, which is not constrained by the incremental nature of the periodical table and so the optimal size may not be the one of bromine but somewhere in-between. The calculated derivative on the bromine atom is 26.19 kJ mol^{-1} Å^{-1}, a strongly positive value, indicating that the theoretical, but physically impossible, sweet spot is already surpassed by its size[†]. Once again, the LFE value for bromine in this position confirms the FED, but this time only by a slight margin. This makes chlorine the best physical possible substituent of the halogens from a FED point of view, which is indeed confirmed by the experimental results described above.

[†] The considerable absolute magnitude is presumably caused by the artificial placement of the bromine and the following missing conformational relaxation of protein and bound ligand.

3.3.4.3 Summary of the subchapter

Since free energy derivatives are rather seldomly encountered in computational chemistry, this subchapter featured an introduction to the concept, visualization, and interpretation. The application of them to the already known thrombin complexes demonstrated a potential usecases in drug development, underlined by experimentally validated results from an in-silico experiment. This short excursion in a hypothetical compound optimization shows the potential of the FEDs for optimizing compounds.

While there are clear limitations to the method, for example, the relative blindness towards categorically new interactions, the FEDs can be interpreted quite intuitively. They can help rationalize the protein-ligand binding, being at times even more sensitive to problematic interactions than the LFE approach. The primary application is to guide compound design, including in-silico approaches as well as traditional optimization, done by experts, but the potential use cases do not end here. Other possibly profiting fields include, for example, pharmacophore-based docking, virtual screening, or machine learning, informing statistical models with an even more detailed physical model. Especially the latter could profit from FEDs (also of molecules in free solution), informing generative models and enabling them to propose compounds with better ADMET properties. Furthermore, the protein atom FEDs, a topic not touched here, can also be of great value. Pressing questions in toxicology and drug resistance could be helped to answer by analyzing tendencies within the protein structure, thereby identifying specificity issues and potentially dangerous mutation spots.

Depending on the intended field of application, there are also possible extensions to the here demonstrated method of utilization of FED. As they were applied to only one conformation of the complex and the rigid-body assumption was used in this chapter, the results suffer from the same problems with this approach as LFE results do and could therefore profit from a combination with MD simulations. Furthermore, FED calculations are much less computational demanding as it is the case with the LFE, since here no integration over λ-steps is required, making this extension more feasible.

4 Summary and conclusion

This work focused on applying 3D RISM based methods to thermodynamically characterizing biologically relevant systems on multiple scales, regarding not only the size of the studied molecules but also the level of detail.

The combination of 3D RISM with MD simulations enables the study of large systems like the anti-NPRA IgG4 antibody and the calculation of the free energy difference between its λ- and Y-conformer.[167] Running simulations of only these endpoints and estimating the hydration free energy from snapshots with 3D RISM could be demonstrated to be a valid way of treating such large systems, previously only shown for smaller proteins and complexes[35–39]. Tests with an alternative water model and PMV corrections all yielded the same trend between the conformers, underlining the robustness of the method. Nonetheless, the presented results also revealed the limitations of this approach through significant statistical errors due to structural fluctuation in the simulations. The approach taken in this work to cope with this problem was to redirect the focus towards a finer detailed description of the systems of interest. This was done by employing the so-called localized free energies (LFE), breaking down thermodynamic properties like the hydration free energy and, together with force field energies, binding free energy onto individual sites.

To demonstrate the plausibility of the localized free energies beyond mathematical soundness, they were used as input features for deep learning models to predict solvation free energies. To explore their applicability and limitations in a real-world scenario, they were used to explain experimental trends in a thrombin ligand series retrospectively. Complexes taken from this series were also used to investigate the influence of some key factors on the results and their potential interpretation. The second method introduced in this work calculates so-called free energy derivatives, describing the derivative of the excess chemical potential with respect to the non-bonded force field parameters. The FEDs were also applied to the thrombin ligand series, explaining the experimental shifts caused by changes between ligands. Their potential for compound design guidance was also demonstrated by correlating the in-silico suggestions with experimental results from multiple studies.

To verify the plausibility of the LFE partitioned excess chemical potential, two deep learning approaches were implemented and trained on solvation free energies, taken from the MNSol dataset[189]. The LFE values were used as input for the machine learning models and were tested against a random and equal distribution. It could be shown that the LFE distribution indeed gives the best results compared to all other partitionings of the excess chemical potential tried in this work. Even an equal distribution of the calculated total excess chemical potential on all sites gives only subpar results. This means, that the models can differentiate between correct and incorrect LFE distributions, which not only confirms the plausibility of LFE approach itself, but also hints towards the possibility that the models can learn about the physics behind solvation from the LFE distribution, even from such a small dataset as it was used here. Such a learned connection could be used to infer LFEs directly from pair distribution functions, without the necessity of calculating and integrating over multiple λ-steps, providing a, probably less accurate, but much faster access to LFEs as the in this work introduced full method. Further, the inverse direction could also be an interesting way to infer an approximate pair distribution function. In either case, more research has to be conducted in this direction. The conducted machine learning experiments also showed the advantages coming from the utilization of first principle methods like 3D RISM in machine learning approaches. This so-called hybrid modeling is especially useful when experimental data is limited or sparse, as it is often the case with chemistry-related tasks like the prediction of solvation free energies. The results shown in this work, underline this mainly for the strongly underrepresented ionic species in the MNSol dataset, where RMSE and MAE values, comparable with those of high-level methods[178], could only be achieved with the physics-based input like the excess chemical potential and LFEs.

The plausibility confirmed, the LFE approach was then applied to protein-ligand complexes on a ligand series of thrombin inhibitors. On the basis of a suitable visualization strategy, the potential knowledge gain from the LFE method and its plausibility in a protein-ligand binding context was demonstrated by in-depth discussions of the results. It could be shown that the separation of the binding free energy in its contributions and the separate interpretation of the protein and ligand perspective can bring additional insights. This also extends to the comparisons between the ligands within the series and observations relative to each other, where the direct and indirect influence of structural changes was illuminated. With the introduced method, a rationalization of interactions between host and guest and their contribution to the binding is possible right down to an atomic level. Nonetheless, there are still limiting factors to the method. The calculation of the localized binding free energy neglected not only the solute entropic terms, but also could not make use of the PMV corrections, as both terms could not be localized. (Approaches for the localization of entropic terms were studied by Fabian Sendzik in his masters thesis[263] during time of writing of this work. Furthermore, the LFEs were applied to only the minimally modeled crystal structures in a sacrifice for speed, treating them as rigid bodies. In addition, the conformations of ligand and protein were kept the same for both bound and unbound states, which is also only a first approximation. To overcome both problems, MD simulations could be employed in future extensions, similar to the procedure described for the antibody. However, with this, the computational burden would increase drastically, as not only the MD simulations would have to be performed but also the localization of the free energy for multiple snapshots be done. Concentrating on only the ligand and neglecting the protein perspective could therefore be an acceptable compromise.

The investigation of influence factors was focused on the net charge of the host as well as the positioning of free ions in the system, as they are often found in structures from crystallographic experiments. The findings suggest that some thought regarding net charge and ion potioning is required in the preprocessing, but also that the effect is mostly limited to global shifts and is not altering the relations between sites in meaningful ways.

Sticking with drug discovery as a general example application for the introduced methods the, from analytical derivatives, calculated FEDs, were also applied to the thrombin inhibitor series. By investigating the derivative of the binding free energy once again from multiple perspectives and the contributions to it individually, the key properties of structural elements occurring in the series could be identified. This led to an in-silico experiment in which multiple halogenic substituents were tested for one of the ligands in the studied series, sparked by the derivatives with respect to the σ-parameter of a chlorine atom. The correlation of the calculated derivatives with findings from multiple experimental studies could confirm the sign of the derivative and its magnitude, suggesting, although physically not possible, that an optimal substituent would be in between chlorine and bromine.

The examples in this work demonstrate the potential application of the LFEs and FEDs in a drug development environment by identifying the most important protein-ligand interactions and the significant properties of these structural elements. Furthermore, the splitting into different perspectives and contributions to the binding free energy grants deep insights into the interaction. With these options in mind, both methods could be used to boost virtual screening applications starting from a few initial binders by first identifying important structures to look for in large databases with the LFE method and also hinting towards relevant properties in potential hits extracted from the FEDs. Docking and 3D QSAR are also applications that can benefit from, especially the volumetric LFEs. With them, pharmacophores could be defined from LFEs of the protein binding site, to which compounds could then be aligned and ultimately be scored. An initial exploration of LFEs and FEDs as input for machine learning models was conducted with the plausibility check of the LFEs, but possibilities in this field go much further than this. For example, generative models could be informed by atom-wise localized thermodynamic properties to improve properties like binding affinities and ADMET parameters of the generated molecules.

5 References

1. Hameed, A., Al-Rashida, M., Alharthy, R. D., Uroos, M., Mughal, E. U., Ali, S. A. & Khan, K. M. Small molecules as activators in medicinal chemistry (2000-2016). *Expert Opin. Ther. Pat.* **27,** 1089–1110; 10.1080/13543776.2017.1349103 (2017).
2. Nicholson, D. W. From bench to clinic with apoptosis-based therapeutic agents. *Nature* **407,** 810–816; 10.1038/35037747 (2000).
3. Schneider, G. *De novo molecular design* (Wiley-VCH, Weinheim, Germany, 2014).
4. Klebe, G. *Wirkstoffdesign. Entwurf und Wirkung von Arzneistoffen.* 2nd ed. (Spektrum Akad. Verl., Heidelberg, 2009).
5. Attene-Ramos, M. S., Austin, C. P. & Xia, M. High Throughput Screening. In *Encyclopedia of toxicology,* edited by P. Wexler & M. Abdollahi. 3rd ed. (Academic Press, Amsterdam, 2014), pp. 916–917.
6. Inglese, J. & Auld, D. S. High Throughput Screening (HTS) Techniques: Applications in Chemical Biology. In *Wiley encyclopedia of chemical biology,* edited by T. P. Begley (Wiley, [Hoboken, NJ], 2007-).
7. Pereira, D. A. & Williams, J. A. Origin and evolution of high throughput screening. *Br. J. Pharmacol.* **152,** 53–61; 10.1038/sj.bjp.0707373 (2007).
8. G. Papadopoulos, M., K. Shukla, M., Kaczmarek-Kedziera, A., Leszczynski, J., Puzyn, T. & Reis, H. (eds.). *Handbook of Computational Chemistry.* 2nd ed. (Springer International Publishing; Imprint: Springer, Cham, 2017).
9. Todeschini, R. & Consonni, V. *Molecular descriptors for chemoinformatics.* 2nd ed. (Wiley-VCH; Chichester : John Wiley [distributor], Weinheim, 2009).
10. Oprea, T. I. & Matter, H. Integrating virtual screening in lead discovery. *Curr. Opin. Chem. Biol.* **8,** 349–358; 10.1016/j.cbpa.2004.06.008 (2004).
11. Butler, T. C. The introduction of chloral hydrate into medical practice. *Bull. Hist. Med.* **44,** 168–172 (1970).
12. Snelders, S., Kaplan, C. & Pieters, T. On Cannabis, Chloral Hydrate, and Career Cycles of Psychotropic Drugs in Medicine. *Bull. Hist. Med.* **80,** 95–114 (2006).
13. Sneader, W. The discovery of aspirin: a reappraisal. *BMJ (Clinical research ed.)* **321,** 1591–1594; 10.1136/bmj.321.7276.1591 (2000).
14. Montinari, M. R., Minelli, S. & Caterina, R. de. The first 3500 years of aspirin history from its roots - A concise summary. *Vasc.Pharmacol.* **113,** 1–8; 10.1016/j.vph.2018.10.008 (2019).
15. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **23,** 3–25; 10.1016/S0169-409X(96)00423-1 (1997).

16. Hamilton, H. W., Steinbaugh, B. A., Stewart, B. H., Chan, O. H., Schmid, H. L., Schroeder, R., Ryan, M. J., Keiser, J., Taylor, M. D. & Blankley, C. J. Evaluation of physicochemical parameters important to the oral bioavailability of peptide-like compounds: implications for the synthesis of renin inhibitors. *J. Med. Chem.* **38,** 1446–1455; 10.1021/jm00009a005 (1995).

17. van Drie, J. H. Computer-aided drug design: the next 20 years. *J. Comput. Aided Mol. Des.* **21,** 591–601; 10.1007/s10822-007-9142-y (2007).

18. Vivo, M. de, Masetti, M., Bottegoni, G. & Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **59,** 4035–4061; 10.1021/acs.jmedchem.5b01684 (2016).

19. Salo-Ahen, O. M. H., Alanko, I., Bhadane, R., Bonvin, A. M. J. J., Honorato, R. V., Hossain, S., Juffer, A. H., Kabedev, A., Lahtela-Kakkonen, M., Larsen, A. S., Lescrinier, E., Marimuthu, P., Mirza, M. U., Mustafa, G., Nunes-Alves, A., Pantsar, T., Saadabadi, A., Singaravelu, K. & Vanmeert, M. Molecular Dynamics Simulations in Drug Discovery and Pharmaceutical Development. *Processes* **9,** 71; 10.3390/pr9010071 (2021).

20. Hollingsworth, S. A. & Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **99,** 1129–1143; 10.1016/j.neuron.2018.08.011 (2018).

21. Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **22,** 1420–1426; 10.1063/1.1740409 (1954).

22. Ballante, F. *Protein-ligand interactions and drug design* (Humana Press, New York, NY, 2021).

23. Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **3,** 300–313; 10.1063/1.1749657 (1935).

24. 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.* **72,** 1047–1069; 10.1016/S0006-3495(97)78756-3 (1997).

25. Boresch, S., Tettinger, F., Leitgeb, M. & Karplus, M. Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. *J. Phys. Chem. B* **107,** 9535–9551; 10.1021/jp0217839 (2003).

26. Mobley, D. L. & Gilson, M. K. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annu. Rev. Biophys.* **46,** 531–558; 10.1146/annurev-biophys-070816-033654 (2017).

27. Plenker, D., Riedel, M., Brägelmann, J., Dammert, M. A., Chauhan, R., Knowles, P. P., Lorenz, C., Keul, M., Bührmann, M., Pagel, O., Tischler, V., Scheel, A. H., Schütte, D., Song, Y., Stark, J., Mrugalla, F., Alber, Y., Richters, A., Engel, J., Leenders, F., Heuckmann, J. M., Wolf, J., Diebold, J., Pall, G., Peifer, M., Aerts, M., Gevaert, K., Zahedi, R. P., Buettner, R., Shokat, K. M., McDonald, N. Q., Kast, S. M., Gautschi, O., Thomas, R. K. & Sos, M. L. Drugging the catalytically inactive state of RET kinase in RET-rearranged tumors. *Sci. Transl. Med.* **9**; 10.1126/scitranslmed.aah6144 (2017).

28. Irwin, B. W. J. & Huggins, D. J. Estimating Atomic Contributions to Hydration and Binding Using Free Energy Perturbation. *J. Chem. Theory Comput.* **14,** 3218–3227; 10.1021/acs.jctc.8b00027 (2018).

29. 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.* **10,** 3331–3344; 10.1021/ct5000296 (2014).

30. Ikeguchi, M. & Doi, J. Direct numerical solution of the Ornstein–Zernike integral equation and spatial distribution of water around hydrophobic molecules. *J. Chem. Phys.* **103,** 5011–5017; 10.1063/1.470587 (1995).

31. Beglov, D. & Roux, B. Solvation of complex molecules in a polar liquid: An integral equation theory. *J. Chem. Phys.* **104,** 8678–8689; 10.1063/1.471557 (1996).

32. Beglov, D. & Roux, B. An Integral Equation To Describe the Solvation of Polar Molecules in Liquid Water. *J. Phys. Chem. B* **101,** 7821–7826; 10.1021/jp971083h (1997).

33. Kovalenko, A. & Hirata, F. Three-dimensional density profiles of water in contact with a solute of arbitrary shape: a RISM approach. *Chem. Phys. Lett.* **290,** 237–244; 10.1016/S0009-2614(98)00471-0 (1998).

34. Roser, M. & Ritchie, H. Technological Progress. *Our World in Data* (2013).

35. Luchko, T., Gusarov, S., Roe, D. R., Simmerling, C., Case, D. A., Tuszynski, J. & Kovalenko, A. Three-dimensional molecular theory of solvation coupled with molecular dynamics in Amber. *J. Chem. Theory Comput.* **6,** 607–624; 10.1021/ct900460m (2010).

36. Roy, D. & Kovalenko, A. Biomolecular Simulations with the Three-Dimensional Reference Interaction Site Model with the Kovalenko-Hirata Closure Molecular Solvation Theory. *Int. J. Mol. Sci.* **22**; 10.3390/ijms22105061 (2021).

37. Genheden, S., Luchko, T., Gusarov, S., Kovalenko, A. & Ryde, U. An MM/3D-RISM approach for ligand binding affinities. *J. Phys. Chem. B* **114,** 8505–8516; 10.1021/jp101461s (2010).

38. Yesudas, J. P., Blinov, N., Dew, S. K. & Kovalenko, A. Calculation of binding free energy of short double stranded oligonucleotides using MM/3D-RISM-KH approach. *J. Mol. Liq.* **201,** 68–76; 10.1016/j.molliq.2014.11.017 (2015).

39. Hasegawa, T., Sugita, M., Kikuchi, T. & Hirata, F. A Systematic Analysis of the Binding Affinity between the Pim-1 Kinase and Its Inhibitors Based on the MM/3D-RISM/KH Method. *J. Chem. Inf. Model.* **57,** 2789–2798; 10.1021/acs.jcim.7b00158 (2017).

40. Sugita, M., Onishi, I., Irisa, M., Yoshida, N. & Hirata, F. Molecular Recognition and Self-Organization in Life Phenomena Studied by a Statistical Mechanics of Molecular Liquids, the RISM/3D-RISM Theory. *Molecules (Basel, Switzerland)* **26**; 10.3390/molecules26020271 (2021).

41. Fusani, L., Wall, I., Palmer, D. & Cortes, A. Optimal water networks in protein cavities with GAsol and 3D-RISM. *Bioinformatics (Oxford, England)* **34,** 1947–1948; 10.1093/bioinformatics/bty024 (2018).

42. Yoshidome, T., Ikeguchi, M. & Ohta, M. Comprehensive 3D-RISM analysis of the hydration of small molecule binding sites in ligand-free protein structures. *J. Comput. Chem.* **41,** 2406–2419; 10.1002/jcc.26406 (2020).

43. Hirano, K., Yokogawa, D., Sato, H. & Sakaki, S. An analysis of 3D solvation structure in biomolecules: application to coiled coil serine and bacteriorhodopsin. *J. Phys. Chem. B* **114,** 7935–7941; 10.1021/jp911470p (2010).

44. Güssregen, S., Matter, H., Hessler, G., Lionta, E., Heil, J. & Kast, S. M. Thermodynamic Characterization of Hydration Sites from Integral Equation-Derived Free Energy Densities: Application to Protein Binding Sites and Ligand Series. *J. Chem. Inf. Model.* **57,** 1652–1666; 10.1021/acs.jcim.6b00765 (2017).

45. Sindhikara, D. J., Yoshida, N. & Hirata, F. Placevent: an algorithm for prediction of explicit solvent atom distribution-application to HIV-1 protease and F-ATP synthase. *J. Comput. Chem.* **33,** 1536–1543; 10.1002/jcc.22984 (2012).

46. Sindhikara, D. J. & Hirata, F. Analysis of biomolecular solvation sites by 3D-RISM theory. *J. Phys. Chem. B* **117,** 6718–6723; 10.1021/jp4046116 (2013).

47. Bodnarchuk, M. S. Water, water, everywhere… It's time to stop and think. *Drug Discov. Today* **21,** 1139–1146; 10.1016/j.drudis.2016.05.009 (2016).

48. Graves, A. P., Wall, I. D., Edge, C. M., Woolven, J. M., Cui, G., Le Gall, A., Hong, X., Raha, K. & Manas, E. S. A Perspective on Water Site Prediction Methods for Structure Based Drug Design. *Curr. Top. Med. Chem.* **17,** 2599–2616; 10.2174/1568026617666170427095035 (2017).

49. Masters, M. R., Mahmoud, A. H., Yang, Y. & Lill, M. A. Efficient and Accurate Hydration Site Profiling for Enclosed Binding Sites. *J. Chem. Inf. Model.* **58,** 2183–2188; 10.1021/acs.jcim.8b00544 (2018).

50. Stumpe, M. C., Blinov, N., Wishart, D., Kovalenko, A. & Pande, V. S. Calculation of local water densities in biological systems: a comparison of molecular dynamics simulations and the 3D-RISM-KH molecular theory of solvation. *J. Phys. Chem. B* **115,** 319–328; 10.1021/jp102587q (2011).

51. Nittinger, E., Gibbons, P., Eigenbrot, C., Davies, D. R., Maurer, B., Yu, C. L., Kiefer, J. R., Kuglstatter, A., Murray, J., Ortwine, D. F., Tang, Y. & Tsui, V. Water molecules in protein-ligand interfaces. Evaluation of software tools and SAR comparison. *J. Comput. Aided Mol. Des.* **33,** 307–330; 10.1007/s10822-019-00187-y (2019).

52. Huang, W., Blinov, N., Wishart, D. S. & Kovalenko, A. Role of water in ligand binding to maltose-binding protein: insight from a new docking protocol based on the 3D-RISM-KH molecular theory of solvation. *J. Chem. Inf. Model.* **55,** 317–328; 10.1021/ci500520q (2015).

53. Hinge, V. K., Blinov, N., Roy, D., Wishart, D. S. & Kovalenko, A. The role of hydration effects in 5-fluorouridine binding to SOD1: insight from a new 3D-RISM-KH based protocol for including structural water in docking simulations. *J. Comput. Aided Mol. Des.* **33,** 913–926; 10.1007/s10822-019-00239-3 (2019).

54. Imai, T., Oda, K., Kovalenko, A., Hirata, F. & Kidera, A. Ligand mapping on protein surfaces by the 3D-RISM theory: toward computational fragment-based drug design. *J. Am. Chem. Soc.* **131,** 12430–12440; 10.1021/ja905029t (2009).

55. Nikolić, D., Blinov, N., Wishart, D. & Kovalenko, A. 3D-RISM-Dock: A New Fragment-Based Drug Design Protocol. *J. Chem. Theory Comput.* **8,** 3356–3372; 10.1021/ct300257v (2012).

56. Sugita, M., Hamano, M., Kasahara, K., Kikuchi, T. & Hirata, F. New Protocol for Predicting the Ligand-Binding Site and Mode Based on the 3D-RISM/KH Theory. *J. Chem. Theory Comput.* **16,** 2864–2876; 10.1021/acs.jctc.9b01069 (2020).

57. Kovalenko, A. & Hirata, F. Potential of Mean Force between Two Molecular Ions in a Polar Molecular Solvent: A Study by the Three-Dimensional Reference Interaction Site Model. *J. Phys. Chem. B* **103,** 7942–7957; 10.1021/jp991300+ (1999).

58. Kiyota, Y., Yoshida, N. & Hirata, F. A New Approach for Investigating the Molecular Recognition of Protein: Toward Structure-Based Drug Design Based on the 3D-RISM Theory. *J. Chem. Theory Comput.* **7,** 3803–3815; 10.1021/ct200358h (2011).

59. Mrugalla, F. & Kast, S. M. Designing molecular complexes using free-energy derivatives from liquid-state integral equation theory. *J. Phys.: Condens. Matter* **28,** 344004; 10.1088/0953-8984/28/34/344004 (2016).

60. Mrugalla, F. Prediction and Optimisation of Protein-Ligand Affinities by Integral Equation Theory. Dissertation. TU Dortmund, 2017.

61. Kloss, T., Heil, J. & Kast, S. M. Quantum chemistry in solution by combining 3D integral equation theory with a cluster embedding approach. *J. Phys. Chem. B* **112,** 4337–4343; 10.1021/jp710680m (2008).

62. Heil, J., Frach, R. & Kast, S. M. Non-continuum solvation using the EC-RISM method applied to predict tautomer ratios, pKa and enantiomeric excess of alkylation reactions. *J. Cheminformatics* **4**; 10.1186/1758-2946-4-S1-O9 (2012).

63. Tielker, N., Eberlein, L., Hessler, G., Schmidt, K. F., Güssregen, S. & Kast, S. M. Quantum-mechanical property prediction of solvated drug molecules: what have we learned from a decade of SAMPL blind prediction challenges? *J. Comput. Aided Mol. Des.* **35,** 453–472; 10.1007/s10822-020-00347-5 (2021).

64. Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., 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. *Acc. Chem. Res.* **33,** 889–897; 10.1021/ar000033j (2000).

65. Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A. & Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate−DNA Helices. *J. Am. Chem. Soc.* **120,** 9401–9409; 10.1021/ja981844+ (1998).

66. Palmer, D. S., Frolov, A. I., Ratkova, E. L. & Fedorov, M. V. Towards a universal method for calculating hydration free energies: a 3D reference interaction site model with partial molar volume correction. *J. Phys.: Condens. Matter* **22,** 492101; 10.1088/0953-8984/22/49/492101 (2010).

67. Ratkova, E. L., Palmer, D. S. & Fedorov, M. V. Solvation thermodynamics of organic molecules by the molecular integral equation theory: approaching chemical accuracy. *Chem. Rev.* **115,** 6312–6356; 10.1021/cr5000283 (2015).

68. Truchon, J.-F., Pettitt, B. M. & Labute, P. A Cavity Corrected 3D-RISM Functional for Accurate Solvation Free Energies. *J. Chem. Theory Comput.* **10,** 934–941; 10.1021/ct4009359 (2014).

69. Huang, W., Blinov, N. & Kovalenko, A. Octanol-Water Partition Coefficient from 3D-RISM-KH Molecular Theory of Solvation with Partial Molar Volume Correction. *J. Phys. Chem. B* **119,** 5588–5597; 10.1021/acs.jpcb.5b01291 (2015).

70. Tielker, N., Tomazic, D., Eberlein, L., Güssregen, S. & Kast, S. M. The SAMPL6 challenge on predicting octanol-water partition coefficients from EC-RISM theory. *J. Comput. Aided Mol. Des.* **34,** 453–461; 10.1007/s10822-020-00283-4 (2020).

71. Terfloth, L. Neural networks and genetic algorithms in drug design. *Drug Discov. Today* **6,** 102–108; 10.1016/S1359-6446(01)00173-8 (2001).

72. Ekins, S., Mestres, J. & Testa, B. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *Br. J. Pharmacol.* **152,** 9–20; 10.1038/sj.bjp.0707305 (2007).

73. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M. & Zhao, S. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18,** 463–477; 10.1038/s41573-019-0024-5 (2019).

74. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **23,** 1241–1250; 10.1016/j.drudis.2018.01.039 (2018).

75. Freedman, D. H. Hunting for New Drugs with AI. *Nature* **576,** S49-S53; 10.1038/d41586-019-03846-0 (2019).

76. Mouchlis, V. D., Afantitis, A., Serra, A., Fratello, M., Papadiamantis, A. G., Aidinis, V., Lynch, I., Greco, D. & Melagraki, G. Advances in de Novo Drug Design: From Conventional to Machine Learning Methods. *Int. J. Mol. Sci.* **22**; 10.3390/ijms22041676 (2021).

77. Wieder, O., Kohlbacher, S., Kuenemann, M., Garon, A., Ducrot, P., Seidel, T. & Langer, T. A compact review of molecular property prediction with graph neural networks. *Drug Discov. Today Technol.*; 10.1016/j.ddtec.2020.11.009 (2020).

78. Shen, J. & Nicolaou, C. A. Molecular property prediction: recent trends in the era of artificial intelligence. *Drug Discov. Today Technol.* **32-33,** 29–36; 10.1016/j.ddtec.2020.05.001 (2019).

79. Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K. & Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9,** 513–530; 10.1039/c7sc02664a (2018).

80. Baum, B., Muley, L., Heine, A., Smolinski, M., Hangauer, D. & Klebe, G. Think twice: understanding the high potency of bis(phenyl)methane inhibitors of thrombin. *J. Mol. Biol.* **391,** 552–564; 10.1016/j.jmb.2009.06.016 (2009).

81. Andersen, H. C. & Chandler, D. Mode Expansion in Equilibrium Statistical Mechanics. I. General Theory and Application to the Classical Electron Gas. *J. Chem. Phys.* **53,** 547–554; 10.1063/1.1674024 (1970).

82. Andersen, H. C. & Chandler, D. Optimized Cluster Expansions for Classical Fluids. I. General Theory and Variational Formulation of the Mean Spherical Model and Hard Sphere Percus-Yevick Equations. *J. Chem. Phys.* **57,** 1918–1929; 10.1063/1.1678512 (1972).

83. Chandler, D. & Andersen, H. C. Optimized Cluster Expansions for Classical Fluids. II. Theory of Molecular Liquids. *J. Chem. Phys.* **57,** 1930–1937; 10.1063/1.1678513 (1972).

84. Hansen, J. P. & McDonald, I. R. *Theory of simple liquids.* 3rd ed. (Elsevier, Amsterdam, Heidelberg, 2009).

85. Hirata, F. *Molecular Theory of Solvation* (Springer Netherlands, Dordrecht, 2004).

86. Versmold, H. E. W. Montroll, J. L. Lebowitz (Eds.): The Liquid State of Matter: Fluids, Simple and Complex, Vol. 8 der Serie „Studies in Statistical Mechanics. North-Holland Publishing Company, Amsterdam, New York, Oxford 1982. 440 Seiten, Preis: US $ 58.25. *Berichte der Bunsengesellschaft für physikalische Chemie* **87,** 455–456; 10.1002/bbpc.19830870527 (1983).

87. Ornstein, L. & Zernike, F. Accidental deviations of density and opalescence at the critical point of a single substance. *KNAW* **17,** 793–806 (1914).

88. Goodall, R. E. A. & Lee, A. A. Data-driven approximations to the bridge function yield improved closures for the Ornstein-Zernike equation. *Soft matter*; 10.1039/d1sm00402f (2021).

89. Puibasset, J. & Belloni, L. Bridge function for the dipolar fluid from simulation. *J. Chem. Phys.* **136,** 154503; 10.1063/1.4703899 (2012).

90. Morita, T. Theory of Classical Fluids: Hyper-Netted Chain Approximation, I. *Prog. Theor. Phys.* **20,** 920–938; 10.1143/PTP.20.920 (1958).

91. Meeron, E. Theory of Potentials of Average Force and Radial Distribution Functions in Ionic Solutions. *J. Chem. Phys.* **28,** 630–643; 10.1063/1.1744204 (1958).

92. Morita, T. Theory of Classical Fluids: Hyper-Netted Chain Approximation. II. *Prog. Theor. Phys.* **21,** 361–382; 10.1143/PTP.21.361 (1959).

93. Morita, T. Theory of Classical Fluids: Hyper-Netted Chain Approximation. III. *Prog. Theor. Phys.* **23,** 829–845; 10.1143/PTP.23.829 (1960).

94. Blum, L. & Torruella, A. J. Invariant Expansion for Two‐Body Correlations: Thermodynamic Functions, Scattering, and the Ornstein—Zernike Equation. *J. Chem. Phys.* **56,** 303–310; 10.1063/1.1676864 (1972).

95. Blum, L., Cummings, P. T. & Bratko, D. A general solution of the molecular Ornstein–Zernike equation for spheres with anisotropic adhesion and electric multipoles. *J. Chem. Phys.* **92,** 3741–3747; 10.1063/1.457832 (1990).

96. Richardi, J., Millot, C. & Fries, P. H. A molecular Ornstein–Zernike study of popular models for water and methanol. *J. Chem. Phys.* **110,** 1138–1147; 10.1063/1.478171 (1999).

97. Casanova, D., Gusarov, S., Kovalenko, A. & Ziegler, T. Evaluation of the SCF Combination of KS-DFT and 3D-RISM-KH; Solvation Effect on Conformational Equilibria, Tautomerization Energies, and Activation Barriers. *J. Chem. Theory Comput.* **3,** 458–476; 10.1021/ct6001785 (2007).

98. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25,** 1157–1174; 10.1002/jcc.20035 (2004).

99. Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E. & Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11,** 3696–3713; 10.1021/acs.jctc.5b00255 (2015).

100. Kast, S. M. & Kloss, T. Closed-form expressions of the chemical potential for integral equation closures with certain bridge functions. *J. Chem. Phys.* **129,** 236101; 10.1063/1.3041709 (2008).

101. Morita, T. & Hiroike, K. A New Approach to the Theory of Classical Fluids. I. *Prog. Theor. Phys.* **23,** 1003–1027; 10.1143/PTP.23.1003 (1960).

102. Singer, S. J. & Chandler, D. Free energy functions in the extended RISM approximation. *Mol. Phys.* **55,** 621–625; 10.1080/00268978500101591 (1985).

103. Chiles, R. A. & Rossky, P. J. Evaluation of reaction free energy surfaces in aqueous solution: an integral equation approach. *J. Am. Chem. Soc.* **106,** 6867–6868; 10.1021/ja00334a080 (1984).

104. Imai, T., Kinoshita, M. & Hirata, F. Theoretical study for partial molar volume of amino acids in aqueous solution: Implication of ideal fluctuation volume. *J. Chem. Phys.* **112,** 9469–9478; 10.1063/1.481565 (2000).

105. Imai. Molecular theory of partial molar volume and its applications to biomolecular systems. *Condens. Matter Phys.* **10,** 343–361; 10.5488/CMP.10.3.343 (2007).

106. Kirkwood, J. G. & Buff, F. P. The Statistical Mechanical Theory of Solutions. I. *J. Chem. Phys.* **19,** 774–777; 10.1063/1.1748352 (1951).

107. Kast, S. M. Free energies from integral equation theories: enforcing path independence. *Phys. Rev. E* **67,** 41203; 10.1103/PhysRevE.67.041203 (2003).

108. Kirkwood, J. G. *Collected works* (CRC Press, New york, 1968).

109. Shirts, M. R. & Pande, V. S. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *J. Chem. Phys.* **122,** 134508; 10.1063/1.1877132 (2005).

110. Steinbrecher, T., Mobley, D. L. & Case, D. A. Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations. *J. Chem. Phys.* **127,** 214108; 10.1063/1.2799191 (2007).

111. Steinbrecher, T., Joung, I. & Case, D. A. Soft-core potentials in thermodynamic integration: comparing one- and two-step transformations. *J. Comput. Chem.* **32,** 3253–3263; 10.1002/jcc.21909 (2011).

112. Montroll, E. W. & Lebowitz, J. L. (eds.). *The liquid state of matter. Fluids, simple and complex* (North-Holland Publ. Co, Amsterdam usw., 1982).

113. Mezei, M. Polynomial path for the calculation of liquid state free energies from computer simulations tested on liquid water. *J. Comput. Chem.* **13,** 651–656; 10.1002/jcc.540130515 (1992).

114. Resat, H. & Mezei, M. Studies on free energy calculations. I. Thermodynamic integration using a polynomial path. *J. Chem. Phys.* **99,** 6052–6061; 10.1063/1.465902 (1993).

115. Simonson, T. Free energy of particle insertion. *Mol. Phys.* **80,** 441–447; 10.1080/00268979300102371 (1993).

116. Beutler, T. C., Mark, A. E., van Schaik, R. C., Gerber, P. R. & van Gunsteren, W. F. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chem. Phys. Lett.* **222,** 529–539; 10.1016/0009-2614(94)00397-1 (1994).

117. Gapsys, V., Seeliger, D. & Groot, B. L. de. New Soft-Core Potential Function for Molecular Dynamics Based Alchemical Free Energy Calculations. *J. Chem. Theory Comput.* **8,** 2373–2382; 10.1021/ct300220p (2012).

118. Sato, H., Hirata, F. & Kato, S. Analytical energy gradient for the reference interaction site model multiconfigurational self-consistent-field method: Application to 1,2-difluoroethylene in aqueous solution. *J. Chem. Phys.* **105,** 1546–1551; 10.1063/1.472015 (1996).

119. Yoshida, N. & Hirata, F. A new method to determine electrostatic potential around a macromolecule in solution from molecular wave functions. *J. Comput. Chem.* **27,** 453–462; 10.1002/jcc.20356 (2006).

120. Miyata, T. & Hirata, F. Combination of molecular dynamics method and 3D-RISM theory for conformational sampling of large flexible molecules in solution. *J. Comput. Chem.* **29,** 871–882; 10.1002/jcc.20844 (2008).

121. Crevier, D. *AI. The tumultuous history of the search for artificial intelligence / Daniel Crevier* (Basic Books, New York, N.Y., 1993).

122. Rosenblatt, F. The Perceptron. A perceiving and recognizing automaton. Cornell Aeronautical Laboratory, 1957.

123. Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* **65,** 386–408; 10.1037/h0042519 (1958).

124. Hastie, T., Tibshirani, R. & Friedman, J. H. *The elements of statistical learning. Data mining, inference, and prediction / Trevor Hastie, Robert Tibshirani, Jerome Friedman.* 2nd ed. (Springer, New York, 2009).

125. Bishop, C. M. *Pattern recognition and machine learning* (Springer, New York, 2006).

126. Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning* (The MIT Press, Cambridge, Massachusetts, 2016).

127. Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **4,** 251–257; 10.1016/0893-6080(91)90009-T (1991).

128. Leshno, M., Lin, V. Y., Pinkus, A. & Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Netw.* **6,** 861–867; 10.1016/S0893-6080(05)80131-5 (1993).

129. Kratsios, A. The Universal Approximation Property. *Ann. Math. Artif. Intell.* **89,** 435–469; 10.1007/s10472-020-09723-1 (2021).

130. Reed, R. D. & Marks, R. J. *Neural smithing. Supervised learning in feedforward artificial neural networks* (MIT Press, Cambridge, Mass., 1999).

131. Kelley, H. J. Gradient Theory of Optimal Flight Paths. *ARS Journal* **30,** 947–954; 10.2514/8.5282 (1960).

132. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323,** 533–536; 10.1038/323533a0 (1986).

133. Choi, D., Shallue, C. J., Nado, Z., Lee, J., Maddison, C. J. & Dahl, G. E. On Empirical Comparisons of Optimizers for Deep Learning, 12.10.2019, https://arxiv.org/pdf/1910.05446.

134. Schmidt, R. M., Schneider, F. & Hennig, P. Descending through a Crowded Valley -- Benchmarking Deep Learning Optimizers, 03.07.2020, https://arxiv.org/pdf/2007.01547.

135. Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Corrado, G. S., Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Jia, Y., Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu & Xiaoqiang Zheng. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Available at http://tensorflow.org/ (2015).

136. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. & Lerer, A. Automatic differentiation in PyTorch. In *NIPS-W* (2017).

137. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98,** 146401; 10.1103/PhysRevLett.98.146401 (2007).

138. Capecchi, A., Probst, D. & Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminformatics* **12,** 43; 10.1186/s13321-020-00445-4 (2020).

139. Glem, R. C., Bender, A., Arnby, C. H., Carlsson, L., Boyer, S. & Smith, J. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **9,** 199–204 (2006).

140. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **42,** 1273–1280; 10.1021/ci010132r (2002).

141. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. Attention Is All You Need, 12.06.2017, http://arxiv.org/pdf/1706.03762v5.

142. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 11.10.2018, https://arxiv.org/pdf/1810.04805.

143. Honda, S., Shi, S. & Ueda, H. R. SMILES Transformer: Pre-trained Molecular Fingerprint for Low Data Drug Discovery, 12.11.2019, https://arxiv.org/pdf/1911.04738.

144. Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C. & Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **5,** 1572–1583; 10.1021/acscentsci.9b00576 (2019).

145. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn.: Sci. Technol.* **1,** 45024; 10.1088/2632-2153/aba947 (2020).

146. Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction, 19.10.2020, https://arxiv.org/pdf/2010.09885.

147. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (JMLR.org2017), pp. 1263–1272.

148. LeCun, Y. Generalization and Network Design Strategies. In *Connectionism in Perspective,* edited by R. Pfeifer, Z. Schreter, F. Fogelman & L. Steels (Elsevier, Zurich, Switzerland, 1989).

149. LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. & Jackel, L. D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* **1,** 541–551; 10.1162/neco.1989.1.4.541 (1989).

150. Fukushima, K. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **36,** 193–202; 10.1007/BF00344251 (1980).

151. Hubel, D. H. & Wiesel, T. N. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* **148,** 574–591; 10.1113/jphysiol.1959.sp006308 (1959).

152. Hubel, D. H. & Wiesel, T. N. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* **195,** 215–243; 10.1113/jphysiol.1968.sp008455 (1968).

153. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25,** 1097–1105 (2012).

154. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition, 10.12.2015, https://arxiv.org/pdf/1512.03385.

155. Tan, M. & Le V, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *International Conference on Machine Learning*.

156. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation, 18.05.2015, https://arxiv.org/pdf/1505.04597.

157. Kellner, C., Derer, S., Valerius, T. & Peipp, M. Boosting ADCC and CDC activity by Fc engineering and evaluation of antibody effector functions. *Methods* **65,** 105–113; 10.1016/j.ymeth.2013.06.036 (2014).

158. Scallon, B. J., Tam, S. H., McCarthy, S. G., Cai, A. N. & Raju, T. S. Higher levels of sialylated Fc glycans in immunoglobulin G molecules can adversely impact functionality. *Mol. Immunol.* **44,** 1524–1534; 10.1016/j.molimm.2006.09.005 (2007).

159. Raju, T. S. Terminal sugars of Fc glycans influence antibody effector functions of IgGs. *Curr. Opin. Immunol.* **20,** 471–478; 10.1016/j.coi.2008.06.007 (2008).

160. Lee, C. C., Perchiacca, J. M. & Tessier, P. M. Toward aggregation-resistant antibodies by design. *Trends Biotechnol.* **31,** 612–620; 10.1016/j.tibtech.2013.07.002 (2013).

161. Davies, A. M., Rispens, T., Ooijevaar-de Heer, P., Gould, H. J., Jefferis, R., Aalberse, R. C. & Sutton, B. J. Structural determinants of unique properties of human IgG4-Fc. *J. Mol. Biol.* **426,** 630–644; 10.1016/j.jmb.2013.10.039 (2014).

162. Davies, A. M. & Sutton, B. J. Human IgG4: a structural perspective. *Immunol. Rev.* **268,** 139–159; 10.1111/imr.12349 (2015).

163. Dbel, S. & Reichert, J. M. *Handbook of therapeutic antibodies* (Wiley-Blackwell, Weinheim, 2014).

164. Saphire, E. O., Stanfield, R. L., Max Crispin, M. D., Parren, P. W.H.I., Rudd, P. M., Dwek, R. A., Burton, D. R. & Wilson, I. A. Contrasting IgG Structures Reveal Extreme Asymmetry and Flexibility. *J. Mol. Biol.* **319,** 9–18; 10.1016/S0022-2836(02)00244-9 (2002).

165. Wilkinson, I. C., Fowler, S. B., Machiesky, L., Miller, K., Hayes, D. B., Adib, M., Her, C., Borrok, M. J., Tsui, P., Burrell, M., Corkill, D. J., Witt, S., Lowe, D. C. & Webster, C. I. Monovalent IgG4 molecules: immunoglobulin Fc mutations that result in a monomeric structure. *mAbs* **5,** 406–417; 10.4161/mabs.23941 (2013).

166. Potter, L. R., Abbey-Hosch, S. & Dickey, D. M. Natriuretic peptides, their receptors, and cyclic guanosine monophosphate-dependent signaling functions. *Endocr. Rev.* **27,** 47–72; 10.1210/er.2005-0014 (2006).

167. Blech, M., Hörer, S., Kuhn, A. B., Kube, S., Göddeke, H., Kiefer, H., Zang, Y., Alber, Y., Kast, S. M., Westermann, M., Tully, M. D., Schäfer, L. V. & Garidel, P. Structure of a Therapeutic Full-Length Anti-NPRA IgG4 Antibody: Dissecting Conformational Diversity. *Biophys. J.* **116,** 1637–1649; 10.1016/j.bpj.2019.03.036 (2019).

168. Hess, B. Determining the shear viscosity of model liquids from molecular dynamics simulations. *J. Chem. Phys.* **116,** 209; 10.1063/1.1421362 (2002).

169. Kovalenko, A. & Hirata, F. Potentials of mean force of simple ions in ambient aqueous solution. I. Three-dimensional reference interaction site model approach. *J. Chem. Phys.* **112,** 10391–10402; 10.1063/1.481676 (2000).

170. Berendsen, H. J. C., Grigera, J. R. & Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **91,** 6269–6271; 10.1021/j100308a038 (1987).

171. Luchko, T., Gusarov, S., Roe, D. R., Simmerling, C., Case, D. A., Tuszynski, J. & Kovalenko, A. Three-dimensional molecular theory of solvation coupled with molecular dynamics in Amber. *J. Chem. Theory Comput.* **6,** 607–624; 10.1021/ct900460m (2010).

172. Åqvist, J. Ion-water interaction potentials derived from free energy perturbation simulations. *J. Phys. Chem.* **94,** 8021–8024; 10.1021/j100384a009 (1990).

173. Smith, D. E. & Dang, L. X. Computer simulations of NaCl association in polarizable water. *J. Chem. Phys.* **100,** 3757–3766; 10.1063/1.466363 (1994).

174. Kovalenko, A. & Hirata, F. Self-consistent description of a metal–water interface by the Kohn–Sham density functional theory and the three-dimensional reference interaction site model. *J. Chem. Phys.* **110,** 10095–10112; 10.1063/1.478883 (1999).

175. Hess, B. & van der Vegt, N. F. A. Hydration thermodynamic properties of amino acid analogues: a systematic comparison of biomolecular force fields and water models. *J. Phys. Chem. B* **110,** 17616–17626; 10.1021/jp0641029 (2006).

176. Schlitter, J. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chem. Phys. Lett.* **215,** 617–621; 10.1016/0009-2614(93)89366-P (1993).

177. Tomazic, D. Optimizing Free Energy Functionals in Integral Equation Theories. Universitätsbibliothek Dortmund, 2016.

178. Tielker, N., Tomazic, D., Heil, J., Kloss, T., Ehrhart, S., Güssregen, S., Schmidt, K. F. & Kast, S. M. The SAMPL5 challenge for embedded-cluster integral equation theory: solvation free energies, aqueous pK a, and cyclohexane-water log D. *J. Comput. Aided Mol. Des.* **30,** 1035–1044; 10.1007/s10822-016-9939-7 (2016).

179. Tielker, N., Eberlein, L., Güssregen, S. & Kast, S. M. The SAMPL6 challenge on predicting aqueous pKa values from EC-RISM theory. *J. Comput. Aided Mol. Des.* **32,** 1151–1163; 10.1007/s10822-018-0140-z (2018).

180. D.A. Case, V. Babin, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu and P.A. Kollman. *AMBER 14* (University of California, San Francisco, 2014).

181. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, Montgomery, Jr., J. A., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman & D. J. Fox. *Gaussian˜16 Revision C.01* (2016).

182. Kast, S. M., Heil, J., Güssregen, S. & Schmidt, K. F. Prediction of tautomer ratios by embedded-cluster integral equation theory. *J. Comput. Aided Mol. Des.* **24,** 343–353; 10.1007/s10822-010-9340-x (2010).

183. Kloss, T. & Kast, S. M. Treatment of charged solutes in three-dimensional integral equation theory. *J. Chem. Phys.* **128,** 134505; 10.1063/1.2841967 (2008).

184. Perkyns, J. S., Lynch, G. C., Howard, J. J. & Pettitt, B. M. Protein solvation from theory and simulation: Exact treatment of Coulomb interactions in three-dimensional theories. *J. Chem. Phys.* **132,** 64106; 10.1063/1.3299277 (2010).

185. Virtanen, P., Gommers, R., Oliphant Travis E., Haberland, M., Reddy, Tyler and Cournapeau, David, Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt Stéfan J., Brett, M., Wilson, Joshua and Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, Eric and Carey, CJ, Polat, Feng, Yu, Moore Eric W., Vand erPlas, J., Laxalde, Denis and Perktold, Josef, Cimrman, R., Henriksen, Ian and Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, Fabian and van Mulbregt, Paul & Contributors, S. 1. 0. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **17,** 261–272; 10.1038/s41592-019-0686-2 (2020).

186. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization, 12/22/2014, https://arxiv.org/pdf/1412.6980.

187. Falkner, S., Klein, A. & Hutter, F. BOHB: Robust and Efficient Hyperparameter Optimization at Scale. In *Proceedings of the 35th International Conference on Machine Learning,* edited by J. Dy & A. Krause (PMLR, Stockholmsmässan, Stockholm Sweden, 2018), Vol. 80, pp. 1437–1446.

188. Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh & Ameet Talwalkar. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *J. Mach. Learn. Res.* **18,** 1–52 (2018).

189. Marenich, A. V., Kelly, C. P., Thompson, J. D., Hawkins, G. D., Chambers, C. C., Giesen, D. J., Winget, P., Cramer, C. J. & Truhlar, D. G. *Minnesota Solvation Database - version 2012* (Minneapolis, 2012).

190. Rawat, W. & Wang, Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Comput.* **29,** 2352–2449; 10.1162/NECO_a_00990 (2017).

191. Sosnin, S., Misin, M., Palmer, D. S. & Fedorov, M. V. 3D matters! 3D-RISM and 3D convolutional neural network for accurate bioaccumulation prediction. *J. Phys.: Condens. Matter* **30,** 32LT03; 10.1088/1361-648X/aad076 (2018).

192. Kuzminykh, D., Polykovskiy, D., Kadurin, A., Zhebrak, A., Baskov, I., Nikolenko, S., Shayakhmetov, R. & Zhavoronkov, A. 3D Molecular Representations Based on the Wave Transform for Convolutional Neural Networks. *Mol. Pharmaceutics* **15,** 4378–4385; 10.1021/acs.molpharmaceut.7b01134 (2018).

193. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **15,** 1929–1958 (2014).

194. Reddi, S. J., Kale, S. & Kumar, S. On the Convergence of Adam and Beyond, 19.04.2019, https://arxiv.org/pdf/1904.09237.

195. Mukherjee, G., Patra, N., Barua, P. & Jayaram, B. A fast empirical GAFF compatible partial atomic charge assignment scheme for modeling interactions of small molecules with biomolecular targets. *J. Comput. Chem.* **32,** 893–907; 10.1002/jcc.21671 (2011).

196. Xu, L., Sun, H., Li, Y., Wang, J. & Hou, T. Assessing the performance of MM/PBSA and MM/GBSA methods. 3. The impact of force fields and ligand charge models. *J. Phys. Chem. B* **117,** 8408–8421; 10.1021/jp404160y (2013).

197. Mobley, D. L., Bayly, C. I., Cooper, M. D. & Dill, K. A. Predictions of hydration free energies from all-atom molecular dynamics simulations. *J. Phys. Chem. B* **113,** 4533–4537; 10.1021/jp806838b (2009).

198. Shulga, D. A., Oliferenko, A. A., Pisarev, S. A., Palyulin, V. A. & Zefirov, N. S. Parameterization of empirical schemes of partial atomic charge calculation for reproducing the molecular electrostatic potential. *Dokl. Chem.* **419,** 57–61; 10.1134/S001250080803004X (2008).

199. Mobley, D. L. & Guthrie, J. P. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *J. Comput. Aided Mol. Des.* **28,** 711–720; 10.1007/s10822-014-9747-x (2014).

200. Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K. & Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **59,** 3370–3388; 10.1021/acs.jcim.9b00237 (2019).

201. Jørgensen, P. B., Jacobsen, K. W. & Schmidt, M. N. Neural Message Passing with Edge Updates for Predicting Properties of Molecules and Materials, 08.06.2018, http://arxiv.org/pdf/1806.03146v1.

202. Subramanian, V., Ratkova, E., Palmer, D., Engkvist, O., Fedorov, M. & Llinas, A. Multisolvent Models for Solvation Free Energy Predictions Using 3D-RISM Hydration Thermodynamic Descriptors. *J. Chem. Inf. Model.*; 10.1021/acs.jcim.0c00065 (2020).

203. Kelton, J. G. & Warkentin, T. E. Heparin-induced thrombocytopenia: a historical perspective. *Blood* **112,** 2607–2616; 10.1182/blood-2008-02-078014 (2008).

204. Ahmed, I., Majeed, A. & Powell, R. Heparin induced thrombocytopenia: diagnosis and management update. *Postgrad. Med. J.* **83,** 575–582; 10.1136/pgmj.2007.059188 (2007).

205. Holbrook, A. M., Pereira, J. A., Labiris, R., McDonald, H., Douketis, J. D., Crowther, M. & Wells, P. S. Systematic overview of warfarin and its drug and food interactions. *Arch. Intern. Med.* **165,** 1095–1106; 10.1001/archinte.165.10.1095 (2005).

206. Ageno, W., Gallus, A. S., Wittkowsky, A., Crowther, M., Hylek, E. M. & Palareti, G. Oral anticoagulant therapy: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest* **141,** e44S-e88S; 10.1378/chest.11-2292 (2012).

207. Di Nisio, M., Middeldorp, S. & Büller, H. R. Direct thrombin inhibitors. *N. Engl. J. Med.* **353,** 1028–1040; 10.1056/NEJMra044440 (2005).

208. Kiser, K. *Oral anticoagulation therapy. Cases and clinical correlation / Kathryn Kiser, editor* (Springer, Cham, Switzerland, 2017).

209. Comin, J. & Kallmes, D. F. Dabigatran (Pradaxa). *AJNR Am. J. Neuroradiol.* **33,** 426–428; 10.3174/ajnr.A3000 (2012).

210. Warkentin, T. E. Bivalent direct thrombin inhibitors: hirudin and bivalirudin. *Best Pract. Res. Clin. Heamatol.* **17,** 105–125; 10.1016/j.beha.2004.02.002 (2004).

211. Dhillon, S. Argatroban: a review of its use in the management of heparin-induced thrombocytopenia. *Am J Cardiovasc Drugs.* **9,** 261–282; 10.2165/1120090-000000000-00000 (2009).

212. Baum, B., Muley, L., Smolinski, M., Heine, A., Hangauer, D. & Klebe, G. Non-additivity of functional group contributions in protein-ligand binding: a comprehensive study by crystallography and isothermal titration calorimetry. *J. Mol. Biol.* **397,** 1042–1054; 10.1016/j.jmb.2010.02.007 (2010).

213. Southworth, H. Predicting potential liver toxicity from phase 2 data: a case study with ximelagatran. *Stat. Med.* **33,** 2914–2923; 10.1002/sim.6142 (2014).

214. Kong, Y., Chen, H., Wang, Y.-Q., Meng, L. & Wei, J.-F. Direct thrombin inhibitors: patents 2002-2012 (Review). *Mol. Med. Rep.* **9,** 1506–1514; 10.3892/mmr.2014.2025 (2014).

215. Abraham, N. S., Singh, S., Alexander, G. C., Heien, H., Haas, L. R., Crown, W. & Shah, N. D. Comparative risk of gastrointestinal bleeding with dabigatran, rivaroxaban, and warfarin: population based cohort study. *BMJ (Clinical research ed.)* **350,** h1857; 10.1136/bmj.h1857 (2015).

216. Blommel, M. L. & Blommel, A. L. Dabigatran etexilate: A novel oral direct thrombin inhibitor. *Am. J. Health Syst. Pharm.* **68,** 1506–1519; 10.2146/ajhp100348 (2011).

217. Šali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234,** 779–815 (1993).

218. Fiser, A., Do, R. K. & Sali, A. Modeling of loops in protein structures. *Protein Sci.* **9,** 1753–1773; 10.1110/ps.9.9.1753 (2000).

219. D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, Y. Huang, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman. *AMBER2018. AmberTools18* (University of California, San Francisco, 2018).

220. O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T. & Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminformatics* **3,** 33; 10.1186/1758-2946-3-33 (2011).

221. Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Petersson, G. A., Nakatsuji, H., Li, X., Caricato, M., Marenich, A. V., Bloino, J., Janesko, B. G., Gomperts, R., Mennucci, B., Hratchian, H. P., Ortiz, J. V., Izmaylov, A. F., Sonnenberg, J. L., Williams-Young, D., Ding, F., Lipparini, F., Egidi, F., Goings, J., Peng, B., Petrone, A., Henderson, T., Ranasinghe, D., Zakrzewski, V. G., Gao, J., Rega, N., Zheng, G., Liang, W., Hada, M.,

Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Throssell, K., Montgomery, Jr., J. A., Peralta, J. E., Ogliaro, F., Bearpark, M. J., Heyd, J. J., Brothers, E. N., Kudin, K. N., Staroverov, V. N., Keith, T. A., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A. P., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Millam, J. M., Klene, M., Adamo, C., Cammi, R., Ochterski, J. W., Martin, R. L., Morokuma, K., Farkas, O., Foresman, J. B. & Fox, D. J. *Gaussian˜16 Revision A.03* (2016).

222. Ditchfield, R., Hehre, W. J. & Pople, J. A. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *J. Chem. Phys.* **54,** 724–728; 10.1063/1.1674902 (1971).

223. Rassolov, V. A., Pople, J. A., Ratner, M. A. & Windus, T. L. 6-31G * basis set for atoms K through Zn. *J. Chem. Phys.* **109,** 1223–1229; 10.1063/1.476673 (1998).

224. VanRossum, G. & Drake, F. L. *The Python language reference.* 3rd ed. (Python Software Foundation; SoHo Books, [Hampton, NH], [Redwood City, Calif.], 2010).

225. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. & Oliphant, T. E. Array programming with NumPy. *Nature* **585,** 357–362; 10.1038/s41586-020-2649-2 (2020).

226. Hernandes, M. Z., Cavalcanti, S. M. T., Moreira, D. R. M., Azevedo Junior, W. F. de & Leite, A. C. L. Halogen atoms in the modern medicinal chemistry: hints for the drug design. *Curr. Drug Targets* **11,** 303–314; 10.2174/138945010790711996 (2010).

227. Hillisch, A., Gericke, K. M., Allerheiligen, S., Roehrig, S., Schaefer, M., Tersteegen, A., Schulz, S., Lienau, P., Gnoth, M., Puetter, V., Hillig, R. C. & Heitmeier, S. Design, Synthesis, and Pharmacological Characterization of a Neutral, Non-Prodrug Thrombin Inhibitor with Good Oral Pharmacokinetics. *J. Med. Chem.* **63,** 12574–12594; 10.1021/acs.jmedchem.0c01035 (2020).

228. Young, T., Abel, R., Kim, B., Berne, B. J. & Friesner, R. A. Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 808–813; 10.1073/pnas.0610202104 (2007).

229. Abel, R., Young, T., Farid, R., Berne, B. J. & Friesner, R. A. Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *J. Am. Chem. Soc.* **130,** 2817–2831; 10.1021/ja0771033 (2008).

230. Bucher, D., Stouten, P. & Triballeau, N. Shedding Light on Important Waters for Drug Design: Simulations versus Grid-Based Methods. *J. Chem. Inf. Model.* **58,** 692–699; 10.1021/acs.jcim.7b00642 (2018).

231. Biela, A., Sielaff, F., Terwesten, F., Heine, A., Steinmetzer, T. & Klebe, G. Ligand binding stepwise disrupts water network in thrombin: enthalpic and entropic changes reveal classical hydrophobic effect. *J. Med. Chem.* **55,** 6094–6110; 10.1021/jm300337q (2012).

232. Straub, A., Roehrig, S. & Hillisch, A. Oral, direct thrombin and factor Xa inhibitors: the replacement for warfarin, leeches, and pig intestines? *Angew. Chem. Int. Ed.* **50,** 4574–4590; 10.1002/anie.201004575 (2011).

233. Biela, A., Khayat, M., Tan, H., Kong, J., Heine, A., Hangauer, D. & Klebe, G. Impact of ligand and protein desolvation on ligand binding to the S1 pocket of thrombin. *J. Mol. Biol.* **418,** 350–366; 10.1016/j.jmb.2012.01.054 (2012).

234. Wilcken, R., Zimmermann, M. O., Lange, A., Joerger, A. C. & Boeckler, F. M. Principles and applications of halogen bonding in medicinal chemistry and chemical biology. *J. Med. Chem.* **56,** 1363–1388; 10.1021/jm3012068 (2013).

235. Sirimulla, S., Bailey, J. B., Vegesna, R. & Narayan, M. Halogen interactions in protein-ligand complexes: implications of halogen bonding for rational drug design. *J. Chem. Inf. Model.* **53,** 2781–2791; 10.1021/ci400257k (2013).

236. Matter, H., Nazaré, M., Güssregen, S., Will, D. W., Schreuder, H., Bauer, A., Urmann, M., Ritter, K., Wagner, M. & Wehner, V. Evidence for C-Cl/C-Br…pi interactions as an important contribution to protein-ligand binding affinity. *Angew. Chem. Int. Ed.* **48,** 2911–2916; 10.1002/anie.200806219 (2009).

237. Baum, B., Mohamed, M., Zayed, M., Gerlach, C., Heine, A., Hangauer, D. & Klebe, G. More than a simple lipophilic contact: a detailed thermodynamic analysis of nonbasic residues in the s1 pocket of thrombin. *J. Mol. Biol.* **390,** 56–69; 10.1016/j.jmb.2009.04.051 (2009).

238. Tucker, T. J., Brady, S. F., Lumma, W. C., Lewis, S. D., Gardell, S. J., Naylor-Olsen, A. M., Yan, Y., Sisko, J. T., Stauffer, K. J., Lucas, B. J., Lynch, J. J., Cook, J. J., Stranieri, M. T., Holahan, M. A., Lyle, E. A., Baskin, E. P., Chen, I. W., Dancheck, K. B., Krueger, J. A., Cooper, C. M. & Vacca, J. P. Design and synthesis of a series of potent and orally bioavailable noncovalent thrombin inhibitors that utilize nonbasic groups in the P1 position. *J. Med. Chem.* **41,** 3210–3219; 10.1021/jm9801713 (1998).

239. Tucker, T. J., Lumma, W. C., Lewis, S. D., Gardell, S. J., Lucas, B. J., Sisko, J. T., Lynch, J. J., Lyle, E. A., Baskin, E. P., Woltmann, R. F., Appleby, S. D., Chen, I. W., Dancheck, K. B., Naylor-Olsen, A. M., Krueger, J. A., Cooper, C. M. & Vacca, J. P. Synthesis of a series of potent and orally bioavailable thrombin inhibitors that utilize 3,3-disubstituted propionic acid derivatives in the P3 position. *J. Med. Chem.* **40,** 3687–3693; 10.1021/jm970397q (1997).

240. Tucker, T. J., Lumma, W. C., Lewis, S. D., Gardell, S. J., Lucas, B. J., Baskin, E. P., Woltmann, R., Lynch, J. J., Lyle, E. A., Appleby, S. D., Chen, I. W., Dancheck, K. B. & Vacca, J. P. Potent noncovalent thrombin inhibitors that utilize the unique amino acid D-dicyclohexylalanine in the P3 position. Implications on oral bioavailability and antithrombotic efficacy. *J. Med. Chem.* **40,** 1565–1569; 10.1021/jm970140s (1997).

241. Tucker, T. J., Lumma, W. C., Mulichak, A. M., Chen, Z., Naylor-Olsen, A. M., Lewis, S. D., Lucas, R., Freidinger, R. M. & Kuo, L. C. Design of highly potent noncovalent thrombin inhibitors that utilize a novel lipophilic binding pocket in the thrombin active site. *J. Med. Chem.* **40,** 830–832; 10.1021/jm960762y (1997).

242. Lumma, W. C., Witherup, K. M., Tucker, T. J., Brady, S. F., Sisko, J. T., Naylor-Olsen, A. M., Lewis, S. D., Lucas, B. J. & Vacca, J. P. Design of novel, potent, noncovalent inhibitors of thrombin with nonbasic P-1 substructures: rapid structure-

activity studies by solid-phase synthesis. *J. Med. Chem.* **41,** 1011–1013; 10.1021/jm9706933 (1998).

243. Schrödinger, L. L.C. The PyMOL Molecular Graphics System, Version 1.8, 2015.

244. Lowe, D. In the Pipeline. Voodoo Nominations. Available at https://blogs.sciencemag.org/pipeline/archives/2015/06/22/voodoo_nominations (2018).

245. Lars Schumann, personal communication.

246. Zhang, E. & Tulinsky, A. The molecular environment of the Na+ binding site of thrombin. *Biophys. Chem.* **63,** 185–200; 10.1016/S0301-4622(96)02227-2 (1997).

247. Lechtenberg, B. C., Freund, S. M. V. & Huntington, J. A. An ensemble view of thrombin allostery. *Biol. Chem.* **393,** 889–898; 10.1515/hsz-2012-0178 (2012).

248. Kahler, U., Kamenik, A. S., Kraml, J. & Liedl, K. R. Sodium-induced population shift drives activation of thrombin. *Sci. Rep.,* 1086; 10.1038/s41598-020-57822-0 (2020).

249. Huntington, J. A. How Na+ activates thrombin--a review of the functional and structural data. *Biol. Chem.* **389,** 1025–1035; 10.1515/BC.2008.113 (2008).

250. Chin, Y.-W., Balunas, M. J., Chai, H. B. & Kinghorn, A. D. Drug discovery from natural sources. *AAPS J.* **8,** E239-53; 10.1007/BF02854894 (2006).

251. Rao, M. S., Gupta, R., Liguori, M. J., Hu, M., Huang, X., Mantena, S. R., Mittelstadt, S. W., Blomme, E. A. G. & van Vleet, T. R. Novel Computational Approach to Predict Off-Target Interactions for Small Molecules. *Front. Big Data* **2,** 25; 10.3389/fdata.2019.00025 (2019).

252. Bender, A., Scheiber, J., Glick, M., Davies, J. W., Azzaoui, K., Hamon, J., Urban, L., Whitebread, S. & Jenkins, J. L. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem* **2,** 861–873; 10.1002/cmdc.200700026 (2007).

253. Misin, M., Vainikka, P. A., Fedorov, M. V. & Palmer, D. S. Salting-out effects by pressure-corrected 3D-RISM. *J. Chem. Phys.* **145,** 194501; 10.1063/1.4966973 (2016).

254. Rühmann, E., Betz, M., Heine, A. & Klebe, G. Fragment Binding Can Be Either More Enthalpy-Driven or Entropy-Driven: Crystal Structures and Residual Hydration Patterns Suggest Why. *J. Med. Chem.* **58,** 6960–6971; 10.1021/acs.jmedchem.5b00812 (2015).

255. Abel, R., Salam, N. K., Shelley, J., Farid, R., Friesner, R. A. & Sherman, W. Contribution of Explicit Solvent Effects to the Binding Affinity of Small-Molecule Inhibitors in Blood Coagulation Factor Serine Proteases. *ChemMedChem* **6,** 1049–1066; 10.1002/cmdc.201000533 (2011).

256. Ahmed, H. U., Blakeley, M. P., Cianci, M., Cruickshank, D. W. J., Hubbard, J. A. & Helliwell, J. R. The determination of protonation states in proteins. *Acta Crystallogr. D Biol. Crystallogr.* **63,** 906–922; 10.1107/S0907444907029976 (2007).

257. Baum, B., Heine, A. & Klebe, G. Thrombin Inhibition (PDB 3D49); 10.2210/pdb3d49/pdb (2009).

258. Pineda, A. O., Chen, Z.-W., Bah, A., Garvey, L. C., Mathews, F. S. & Di Cera, E. Crystal structure of thrombin in a self-inhibited conformation. *J. Biol. Chem.* **281,** 32922–32928; 10.1074/jbc.M605530200 (2006).

259. Burgey, C. S., Robinson, K. A., Lyle, T. A., Sanderson, P. E. J., Lewis, S. D., Lucas, B. J., Krueger, J. A., Singh, R., Miller-Stein, C., White, R. B., Wong, B., Lyle, E. A., Williams, P. D., Coburn, C. A., Dorsey, B. D., Barrow, J. C., Stranieri, M. T., Holahan, M. A., Sitko, G. R., Cook, J. J., McMasters, D. R., McDonough, C. M., Sanders, W. M., Wallace, A. A., Clayton, F. C., Bohn, D., Leonard, Y. M., Detwiler, T. J., Lynch, J. J., Yan, Y., Chen, Z., Kuo, L., Gardell, S. J., Shafer, J. A. & Vacca, J. P. Metabolism-Directed Optimization of 3-Aminopyrazinone Acetamide Thrombin Inhibitors. Development of an Orally Bioavailable Series Containing P1 and P3 Pyridines. *J. Med. Chem.* **46,** 461–473; 10.1021/jm020311f (2003).

260. Sandner, A., Heine, A., Klebe, G. & Abazi, N. Thrombin in complex with 13k (PDB 6Y02); 10.2210/pdb6y02/pdb (2021).

261. Sandner, A., Heine, A., Klebe, G. & Collins, C. Thrombin in complex with D-Phe-Pro-2-chlorothiophen derivative (16e) (PDB 6YHJ); 10.2210/pdb6yhj/pdb (2021).

262. Hanwell, M. D., Curtis, D. E., Lonie, D. C., Vandermeersch, T., Zurek, E. & Hutchison, G. R. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminformatics* **4,** 17; 10.1186/1758-2946-4-17 (2012).

263. Sendzik, F. Absolute Berechnung und lokale Dekomposition der freien Enthalpie von Komplexbildungsreaktionen. Master Thesis. Technische Universität, 2021.

6 Attachment

Table 6.1 Results of the, with the 3D-CNN models, made predictions measured against the MNSol data set[189]. Posterior separated in neutral and single positive and negative charged molecules, treated with three different methods for partial charge calculation. The used metrics are **R**oot **M**ean **S**quare **E**rror (RMSE), **M**ean **A**bsolute **E**rror (MAE), coefficient of determination (R^2), as well as the slope (m) and y-intercept (b) of a linear fit on the predicted solvation free energies in kcal/mol. All numbers are an average over five repetitions and for each of the three five-fold cross-validations (A, B, and C). The raw data can be found in the electronic appendix under 3.2/mnsol_3DCNN_modelResults.csv and collectively under 3.2/mnsol_molInf.csv.

			3D-CNN			
		RMSE	MAE	R^2	m	b
AM1BCC						
All	A	2.49±0.35	1.45±0.24	0.99±0.00	0.98±0.00	0.07±0.06
	B	2.85±0.29	1.53±0.22	0.99±0.00	0.97±0.00	-0.14±0.04
	C	2.64±0.32	1.52±0.23	0.99±0.00	0.98±0.00	0.01±0.04
Neutral	A	1.40±0.18	0.93±0.15	0.89±0.00	0.91±0.01	-0.19±0.05
	B	1.43±0.17	0.93±0.14	0.88±0.00	0.92±0.01	-0.33±0.06
	C	1.40±0.18	0.93±0.15	0.89±0.00	0.88±0.01	-0.36±0.03
Anions	A	4.79±0.68	3.36±0.55	0.80±0.01	0.89±0.01	-7.13±0.77
	B	5.89±0.54	3.95±0.50	0.74±0.01	0.90±0.01	-6.10±0.78
	C	5.15±0.63	3.78±0.57	0.77±0.01	0.85±0.01	-9.82±0.67
Cations	A	3.48±0.48	2.40±0.40	0.79±0.01	0.82±0.01	-10.37±0.76
	B	3.47±0.41	2.34±0.37	0.80±0.01	0.78±0.01	-13.58±0.55
	C	3.74±0.40	2.46±0.33	0.75±0.01	0.82±0.01	-10.67±0.73
RESP						
All	A	2.36±0.34	1.42±0.26	0.99±0.00	0.97±0.00	-0.03±0.11
	B	2.50±0.29	1.46±0.23	0.99±0.00	0.97±0.00	-0.18±0.02
	C	2.41±0.33	1.47±0.25	0.99±0.00	0.97±0.00	-0.13±0.06
Neutral	A	1.40±0.23	0.95±0.20	0.89±0.00	0.91±0.01	-0.28±0.09
	B	1.46±0.21	0.98±0.17	0.88±0.00	0.92±0.01	-0.34±0.02
	C	1.38±0.23	0.96±0.19	0.88±0.00	0.89±0.01	-0.43±0.06
Anions	A	4.24±0.60	2.94±0.46	0.86±0.01	0.91±0.02	-5.18±1.12
	B	4.75±0.50	3.17±0.44	0.83±0.01	0.91±0.01	-4.53±0.66
	C	4.40±0.54	3.24±0.45	0.84±0.01	0.88±0.01	-6.91±0.76
Cations	A	3.69±0.46	2.61±0.38	0.78±0.02	0.84±0.01	-9.23±0.90
	B	3.42±0.39	2.43±0.34	0.81±0.01	0.82±0.01	-10.25±0.67
	C	3.74±0.50	2.59±0.40	0.75±0.02	0.81±0.01	-11.26±0.70
EC-RISM						
All	A	2.83±0.28	1.52±0.20	0.99±0.00	0.98±0.00	-0.08±0.04
	B	2.69±0.29	1.48±0.21	0.99±0.00	0.98±0.00	-0.14±0.02
	C	2.87±0.31	1.50±0.20	0.99±0.00	0.98±0.00	-0.10±0.03
Neutral	A	1.29±0.15	0.92±0.13	0.91±0.00	0.90±0.01	-0.39±0.06
	B	1.33±0.16	0.93±0.14	0.90±0.00	0.91±0.01	-0.43±0.02
	C	1.42±0.16	0.91±0.13	0.89±0.00	0.89±0.01	-0.47±0.04
Anions	A	5.17±0.52	3.66±0.42	0.76±0.00	0.84±0.01	-11.25±0.62
	B	4.66±0.54	3.51±0.47	0.80±0.00	0.83±0.01	-12.33±0.60
	C	4.77±0.62	3.42±0.49	0.79±0.01	0.85±0.01	-10.50±0.76
Cations	A	5.20±0.46	2.74±0.34	0.67±0.01	1.00±0.02	1.47±1.38
	B	5.12±0.42	2.46±0.35	0.68±0.01	1.01±0.01	2.11±0.82
	C	5.71±0.43	2.95±0.32	0.64±0.01	1.02±0.02	2.47±1.01

Table 6.2 Results of the, with the MPNN models, made predictions measured against the MNSol data set[189]. Posterior separated in neutral and single positive and negative charged molecules, treated with three different methods for partial charge calculation. The used metrics are **R**oot **M**ean **S**quare **E**rror (RMSE), **M**ean **A**bsolute **E**rror (MAE), coefficient of determination (R^2), as well as the slope (m) and y-intercept (b) of a linear fit on the predicted solvation free energies in kcal/mol. All numbers are an average over five repetitions and for each of the three five-fold cross-validations (A, B, and C). The raw data can be found in the electronic appendix under 3.2/mnsol_MPNN_modelResults.csv and collectively under 3.2/mnsol_molInf.csv.

		MPNN				
		RMSE	MAE	R^2	m	b
AM1BCC						
All	A	2.33±0.29	1.25±0.21	0.99±0.00	0.99±0.00	-0.16±0.02
	B	2.26±0.29	1.25±0.20	0.99±0.00	0.99±0.00	-0.10±0.02
	C	2.68±0.31	1.34±0.21	0.99±0.00	1.00±0.00	-0.06±0.02
Neutral	A	1.12±0.17	0.77±0.15	0.93±0.00	0.95±0.00	-0.31±0.03
	B	1.31±0.18	0.85±0.14	0.90±0.00	0.97±0.01	-0.15±0.03
	C	1.55±0.17	0.84±0.14	0.86±0.01	0.88±0.01	-0.45±0.06
Anions	A	4.94±0.57	3.15±0.44	0.78±0.01	0.88±0.01	-9.19±0.66
	B	4.60±0.55	2.97±0.44	0.80±0.01	0.88±0.01	-9.22±0.80
	C	5.49±0.64	3.53±0.50	0.72±0.02	0.75±0.02	-19.27±1.57
Cations	A	2.71±0.35	1.90±0.32	0.86±0.01	0.89±0.01	-7.14±0.58
	B	2.35±0.38	1.55±0.30	0.89±0.01	0.91±0.01	-5.89±0.83
	C	2.68±0.33	1.71±0.29	0.86±0.01	0.88±0.01	-8.03±0.66
RESP						
All	A	1.92±0.24	1.21±0.19	1.00±0.00	1.00±0.00	-0.09±0.03
	B	1.89±0.23	1.18±0.19	1.00±0.00	1.00±0.00	-0.17±0.03
	C	2.01±0.25	1.18±0.19	1.00±0.00	1.00±0.00	-0.04±0.05
Neutral	A	1.14±0.16	0.79±0.14	0.92±0.00	0.95±0.01	-0.23±0.03
	B	1.26±0.16	0.83±0.14	0.91±0.00	0.97±0.00	-0.25±0.02
	C	1.43±0.16	0.83±0.14	0.88±0.00	0.89±0.00	-0.43±0.04
Anions	A	3.58±0.38	2.65±0.34	0.88±0.01	0.86±0.01	-10.93±0.33
	B	3.47±0.36	2.54±0.33	0.89±0.00	0.87±0.00	-10.10±0.30
	C	3.49±0.45	2.42±0.37	0.88±0.00	0.86±0.01	-10.64±0.34
Cations	A	2.69±0.41	2.10±0.35	0.86±0.01	0.94±0.01	-3.48±0.84
	B	2.30±0.37	1.68±0.32	0.90±0.00	0.98±0.01	-1.30±0.76
	C	2.52±0.35	1.95±0.32	0.88±0.00	0.93±0.01	-5.10±0.70
EC-RISM						
All	A	2.24±0.26	1.28±0.18	0.99±0.00	0.99±0.00	-0.12±0.04
	B	2.21±0.27	1.27±0.19	0.99±0.00	1.00±0.00	-0.07±0.02
	C	2.25±0.26	1.25±0.19	0.99±0.00	1.00±0.00	-0.08±0.04
Neutral	A	1.10±0.14	0.71±0.12	0.93±0.00	0.96±0.01	-0.23±0.04
	B	1.30±0.15	0.80±0.13	0.91±0.00	0.96±0.00	-0.16±0.03
	C	1.46±0.17	0.78±0.14	0.88±0.00	0.92±0.01	-0.38±0.01
Anions	A	4.49±0.51	3.19±0.38	0.81±0.01	0.87±0.00	-9.70±0.36
	B	4.24±0.55	3.08±0.40	0.83±0.00	0.87±0.00	-9.88±0.30
	C	4.18±0.47	3.01±0.37	0.84±0.00	0.92±0.01	-6.35±0.54
Cations	A	3.18±0.36	2.56±0.30	0.81±0.00	0.91±0.01	-5.41±0.99
	B	2.88±0.33	2.00±0.28	0.85±0.01	0.94±0.01	-4.04±0.57
	C	2.79±0.37	2.06±0.33	0.85±0.00	0.88±0.01	-8.22±0.38

Table 6.3 Results of the MPNN model predictions with diminishing amounts of LFE data on three levels: the mean over the full molecule is assigned to each atom (LFE mean), the calculated LFE values are assigned to random atoms (LFE randomized), the LFE channel is set to zero for each molecule (LFE set to 0). As a reference, the results utilizing the original data (LFE as calculated) are given once again. The labels A, B, and C stand for the three five-fold cross-validations. The values are averaged over five repetitions with different random seeds. The raw data can be found in the electronic appendix under 3.2/mnsol_MPNN_LFEtests_modelResults.csv and collectively under 3.2/mnsol_molInf.csv.

		RMSE	MAE	R^2	m	b
AM1BCC						
LFE as calculated	A	2.33±0.29	1.25±0.21	0.99±0.00	0.99±0.00	-0.16±0.02
	B	2.26±0.29	1.25±0.20	0.99±0.00	0.99±0.00	-0.10±0.02
	C	2.68±0.31	1.34±0.21	0.99±0.00	1.00±0.00	-0.06±0.02
LFE mean	A	2.60±0.27	1.47±0.21	0.99±0.00	1.00±0.00	-0.11±0.04
	B	2.57±0.26	1.43±0.21	0.99±0.00	1.00±0.00	-0.15±0.05
	C	2.77±0.25	1.51±0.20	0.99±0.00	0.99±0.00	-0.15±0.05
LFE randomized	A	2.62±0.34	1.58±0.26	0.99±0.00	0.99±0.00	-0.03±0.03
	B	2.69±0.33	1.62±0.25	0.99±0.00	0.99±0.00	-0.20±0.04
	C	3.01±0.31	1.73±0.24	0.99±0.00	0.99±0.00	-0.09±0.02
LFE set to 0	A	3.08±0.47	1.82±0.34	0.99±0.00	0.99±0.00	-0.21±0.04
	B	3.35±0.53	1.87±0.34	0.99±0.00	0.98±0.00	-0.31±0.04
	C	3.41±0.43	1.95±0.31	0.99±0.00	0.98±0.00	-0.17±0.07
RESP						
LFE as calculated	A	1.92±0.24	1.21±0.19	1.00±0.00	1.00±0.00	-0.09±0.03
	B	1.89±0.23	1.18±0.19	1.00±0.00	1.00±0.00	-0.17±0.03
	C	2.01±0.25	1.18±0.19	1.00±0.00	1.00±0.00	-0.04±0.05
LFE mean	A	2.48±0.24	1.44±0.20	0.99±0.00	1.00±0.00	-0.06±0.03
	B	2.37±0.27	1.42±0.22	0.99±0.00	0.99±0.00	-0.14±0.01
	C	2.49±0.26	1.45±0.20	0.99±0.00	0.99±0.00	-0.18±0.04
LFE randomized	A	2.68±0.31	1.64±0.24	0.99±0.00	1.00±0.00	-0.06±0.04
	B	2.62±0.34	1.59±0.26	0.99±0.00	0.99±0.00	-0.14±0.01
	C	2.95±0.32	1.72±0.25	0.99±0.00	1.00±0.00	-0.10±0.01
LFE set to 0	A	3.08±0.47	1.82±0.34	0.99±0.00	0.99±0.00	-0.21±0.04
	B	3.41±0.51	1.89±0.33	0.99±0.00	0.98±0.00	-0.31±0.02
	C	3.45±0.47	1.97±0.32	0.99±0.00	0.99±0.00	-0.18±0.07
EC-RISM						
LFE as calculated	A	2.24±0.26	1.28±0.18	0.99±0.00	0.99±0.00	-0.12±0.04
	B	2.21±0.27	1.27±0.19	0.99±0.00	1.00±0.00	-0.07±0.02
	C	2.25±0.26	1.25±0.19	0.99±0.00	1.00±0.00	-0.08±0.04
LFE mean	A	3.43±0.37	1.54±0.22	0.99±0.00	0.99±0.00	-0.13±0.04
	B	3.47±0.40	1.58±0.22	0.99±0.00	0.98±0.00	-0.24±0.05
	C	3.34±0.43	1.54±0.23	0.99±0.00	0.99±0.00	-0.12±0.05
LFE randomized	A	2.90±0.38	1.63±0.26	0.99±0.00	0.99±0.00	-0.14±0.02
	B	2.79±0.33	1.59±0.25	0.99±0.00	0.99±0.00	-0.19±0.06
	C	2.79±0.35	1.58±0.25	0.99±0.00	0.99±0.00	-0.15±0.01
LFE set to 0	A	3.39±0.61	1.80±0.34	0.99±0.00	0.99±0.00	-0.16±0.03
	B	3.36±0.62	1.82±0.35	0.99±0.00	0.98±0.00	-0.47±0.03
	C	3.54±0.67	1.87±0.38	0.98±0.00	0.98±0.00	-0.20±0.08

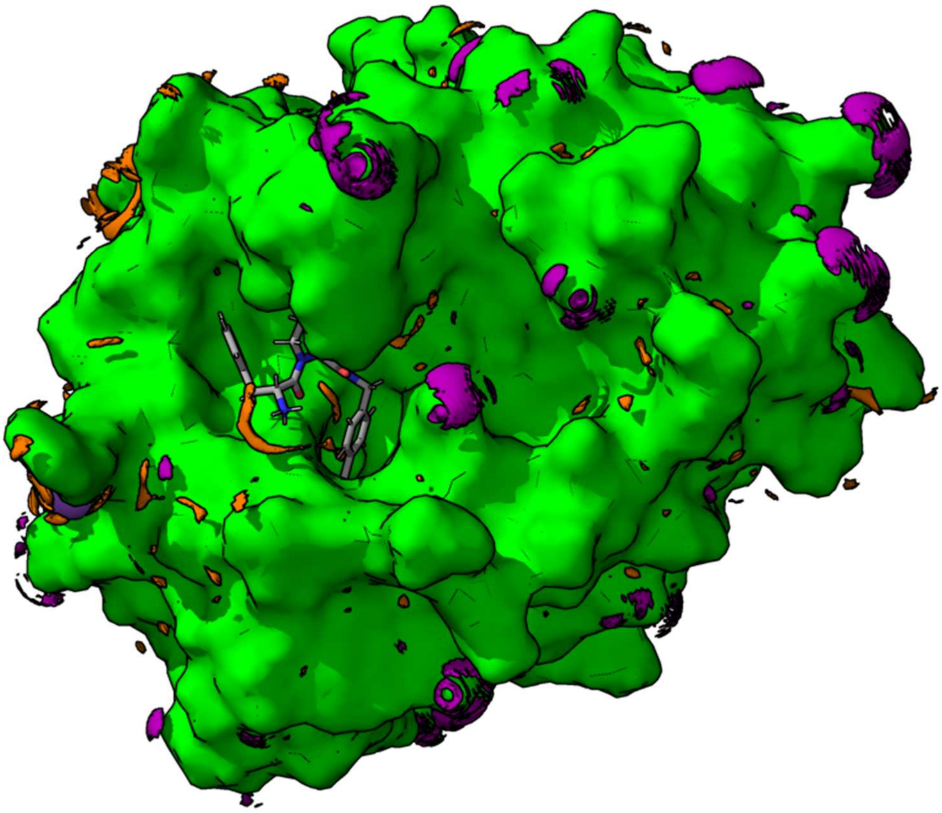

Figure 6.1 Sodium (purpile) and chlorine (orange) densities around the thombin complex 2ZDA (protein surface shown in green; ligand shown in licorice)

Table 6.4 All ligand atoms of the thrombin complex 2ZFP. For each the desolvation penalty (left), intermolecular energy (mid), and binding free energy (right) is given. Each of these columns shows also the corresponding values for the differently charged host systems.

Atom	$\Delta G^{(L)}_{\mathrm{hyd,PL-L}}$ / kJ mol^{-1}			$E^{(L)}_{\mathrm{PL}}$ / kJ mol^{-1}			$\Delta_{\mathrm{bind}} G^{(L)}_{\mathrm{PL}}$ / kJ mol^{-1}		
Host net charge	-1	0	+1	-1	0	+1	-1	0	+1
N01	-66.42	-39.07	-24.39	67.87	40.19	25.46	1.45	1.12	1.07
C02	6.75	4.48	3.29	-6.74	-4.45	-3.25	0.00	0.03	0.04
C03	74.64	44.52	28.27	-73.06	-42.57	-26.05	1.59	1.95	2.22
O04	-61.52	-35.75	-22.11	56.41	30.33	16.45	-5.11	-5.42	-5.66
C05	2.56	2.91	3.08	-2.22	-2.58	-2.74	0.34	0.33	0.33
C06	-8.63	-0.22	3.80	8.40	-0.12	-4.19	-0.23	-0.34	-0.39
N07	-45.89	-27.25	-16.59	43.30	24.44	13.59	-2.59	-2.82	-3.00
C08	-4.61	-2.16	-0.68	3.70	1.21	-0.29	-0.92	-0.95	-0.97
C09	96.49	69.43	52.68	-99.32	-71.94	-54.95	-2.83	-2.50	-2.27
O0A	-68.03	-46.61	-33.11	71.11	49.44	35.75	3.08	2.82	2.64
C0B	-3.32	-1.11	0.32	2.36	0.13	-1.32	-0.96	-0.98	-1.00
C0C	1.92	2.73	3.24	-2.98	-3.81	-4.32	-1.06	-1.07	-1.08
C0D	31.11	19.02	11.98	-30.48	-18.24	-11.08	0.63	0.78	0.90
N0E	-94.05	-71.58	-57.56	93.24	70.50	56.27	-0.81	-1.08	-1.29
C0F	21.90	17.50	14.64	-22.36	-17.90	-15.00	-0.46	-0.40	-0.36
C0G	17.36	14.06	12.05	-16.98	-13.64	-11.60	0.38	0.42	0.45
C0H	-20.80	-15.29	-11.96	19.06	13.48	10.10	-1.75	-1.81	-1.86
C0I	-26.81	-20.99	-17.63	23.66	17.77	14.36	-3.15	-3.22	-3.27
C0J	-11.24	-8.44	-6.89	7.56	4.73	3.15	-3.68	-3.72	-3.74
C0K	16.14	12.54	10.51	-18.32	-14.67	-12.62	-2.18	-2.14	-2.11
C0L	-29.08	-20.83	-15.97	23.13	14.79	9.85	-5.94	-6.04	-6.11
Cl0M	-15.88	-10.07	-6.98	6.35	0.48	-2.67	-9.53	-9.60	-9.64
H0N	39.56	24.14	15.34	-42.68	-27.07	-18.22	-3.12	-2.93	-2.88
H0O	56.59	39.84	31.26	-57.08	-40.13	-31.55	-0.50	-0.29	-0.29
H0P	31.06	14.92	6.08	-33.29	-16.96	-8.03	-2.24	-2.04	-1.95
H0Q	8.05	3.83	1.51	-8.38	-4.11	-1.79	-0.33	-0.28	-0.29
H0R	5.52	2.96	1.74	-4.24	-1.64	-0.42	1.28	1.31	1.32
H0S	3.80	1.10	-0.14	-5.09	-2.36	-1.10	-1.29	-1.26	-1.24
H0T	4.14	1.01	-0.55	-4.18	-1.01	0.57	-0.04	0.00	0.01
H0U	7.00	3.50	1.94	-4.53	-0.99	0.60	2.47	2.51	2.54
H0V	4.80	1.51	-0.11	-5.51	-2.18	-0.54	-0.71	-0.67	-0.64
H0W	13.13	9.09	6.75	-15.05	-10.96	-8.58	-1.92	-1.87	-1.84
H0X	4.61	2.32	0.79	-7.06	-4.75	-3.19	-2.45	-2.42	-2.40
H0Y	3.20	0.81	-0.80	0.60	3.02	4.65	3.79	3.82	3.85
H0Z	0.56	-0.75	-1.54	-2.52	-1.19	-0.39	-1.96	-1.94	-1.93
H10	1.84	0.64	-0.14	-3.85	-2.63	-1.85	-2.01	-2.00	-1.99
H11	-3.10	-1.35	-0.38	2.68	0.91	-0.07	-0.42	-0.44	-0.45
H12	-4.42	-2.82	-1.86	3.22	1.60	0.63	-1.20	-1.22	-1.23
H13	63.43	50.36	42.35	-67.03	-53.80	-45.67	-3.60	-3.44	-3.32
H14	7.27	5.48	4.30	-7.94	-6.13	-4.94	-0.67	-0.65	-0.63
H15	7.45	5.62	4.37	-6.52	-4.66	-3.39	0.94	0.96	0.98
H16	22.88	18.06	15.07	-21.89	-17.01	-13.98	0.99	1.05	1.09
H17	28.77	23.62	20.65	-32.23	-27.02	-24.00	-3.46	-3.39	-3.35
H18	25.95	21.21	18.65	-27.65	-22.84	-20.24	-1.69	-1.63	-1.59
H19	22.18	15.75	11.91	-20.50	-14.00	-10.10	1.68	1.76	1.81
Sum	166.86	128.67	107.18	-217.03	-178.34	-156.7	-50.19	-49.7	-49.52

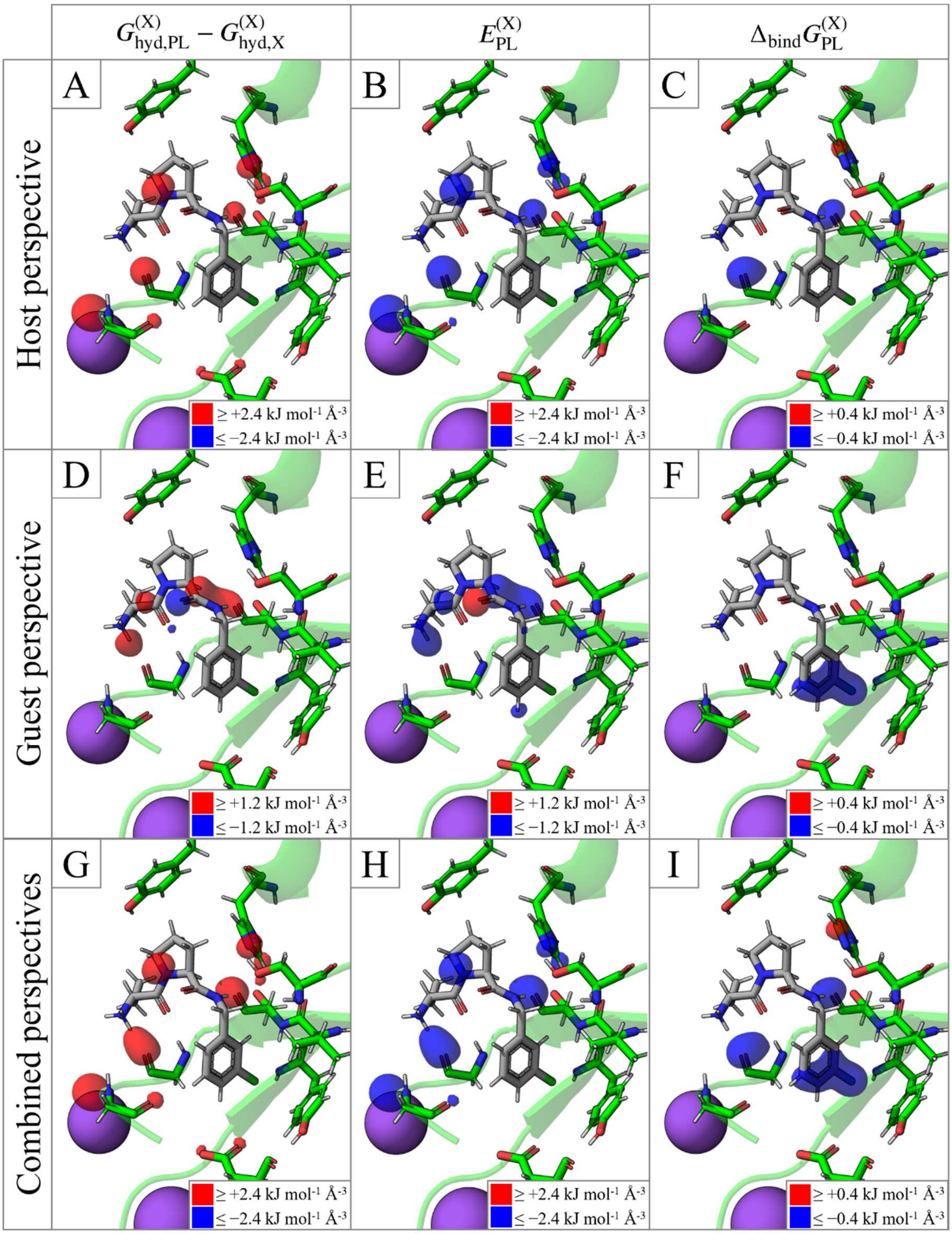

Figure 6.2 Volumetric LFE representation in the protein- (top) and ligand- (mid) perspective of the thrombin complex 2ZFP, as well as their combination (bottom). For each perspective, the desolvation penalty (left column), intermolecular energy (mid column), and binding free energy (right column) is given.

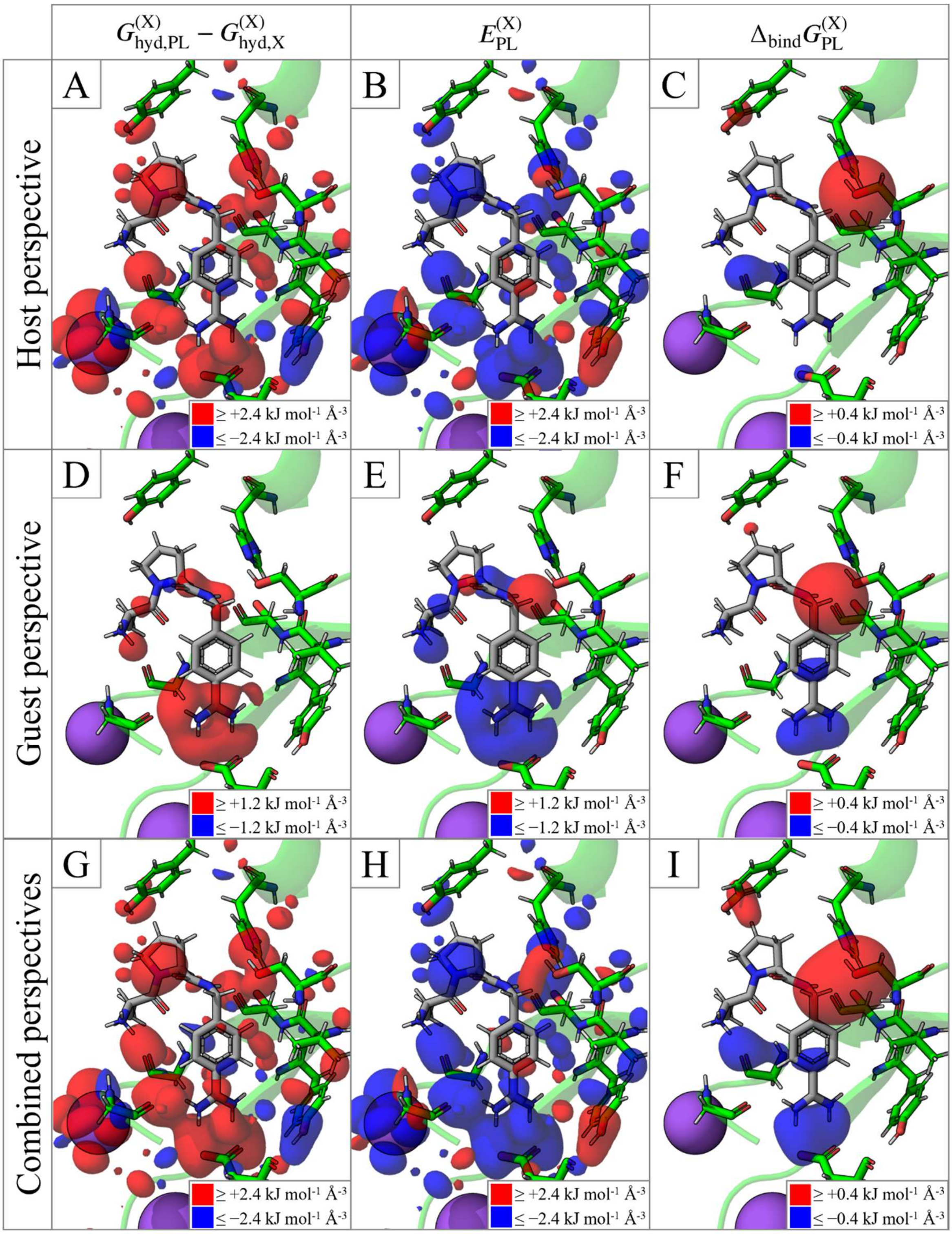

Figure 6.3 Volumetric LFE representation in the protein- (top) and ligand- (mid) perspective of the thrombin complex 2ZGX, as well as their combination (bottom). For each perspective, the desolvation penalty (left column), intermolecular energy (mid column), and binding free energy (right column) is given.

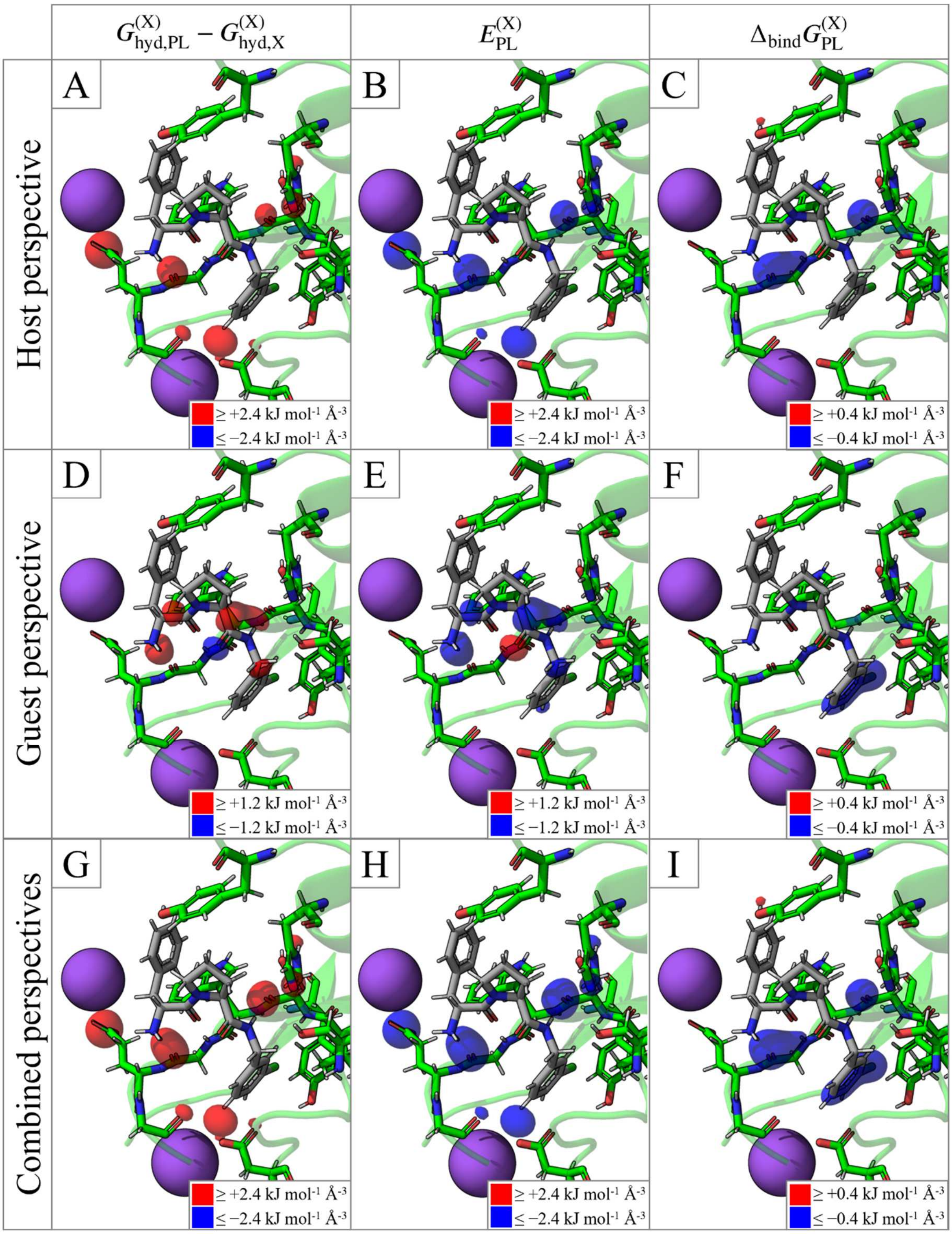

Figure 6.4 Volumetric LFE representation in the protein- (top) and ligand- (mid) perspective of the thrombin complex 2ZC9, as well as their combination (bottom). For each perspective, the desolvation penalty (left column), intermolecular energy (mid column), and binding free energy (right column) is given.

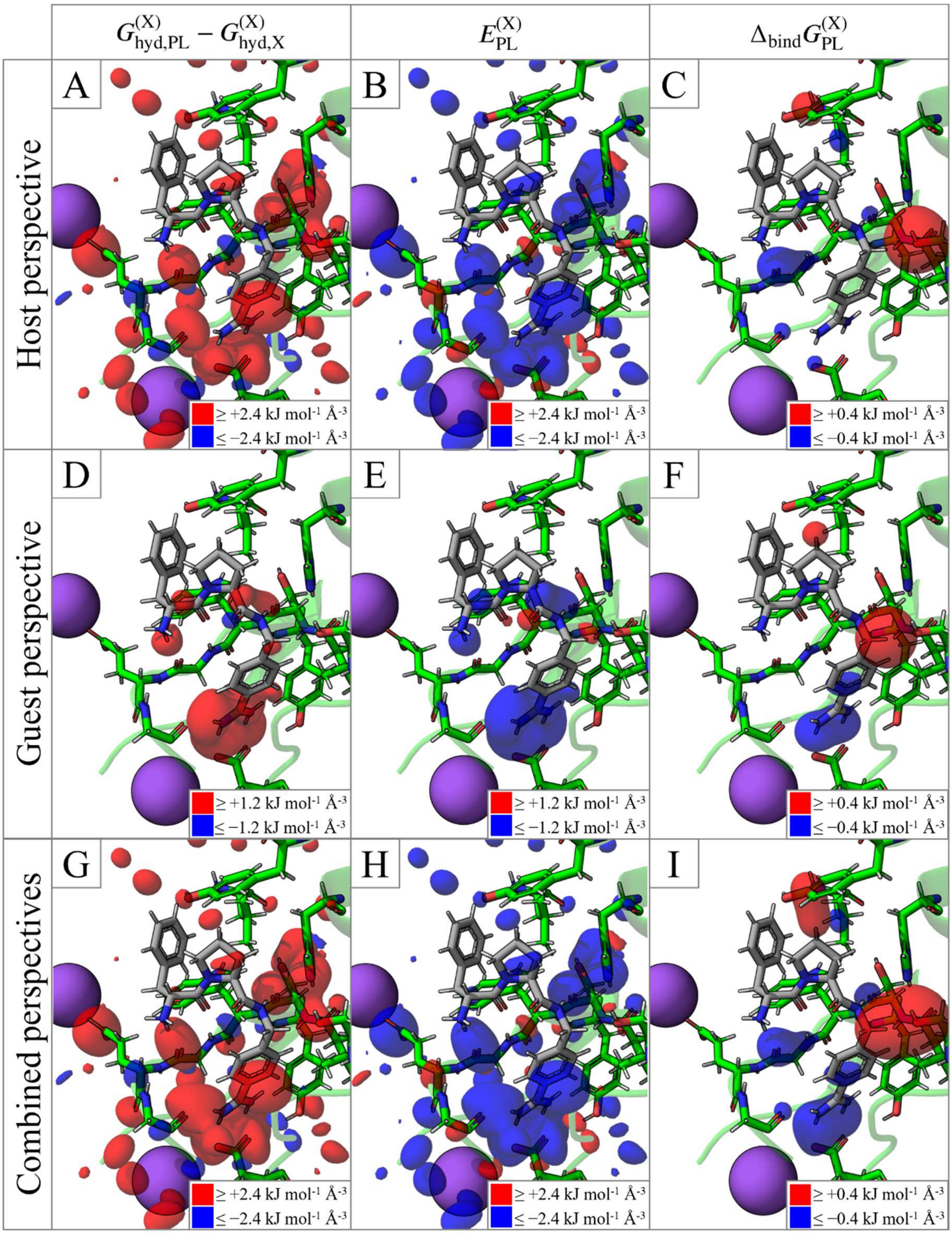

Figure 6.5 Volumetric LFE representation in the protein- (top) and ligand- (mid) perspective of the thrombin complex 2ZDA, as well as their combination (bottom). For each perspective, the desolvation penalty (left column), intermolecular energy (mid column), and binding free energy (right column) is given.

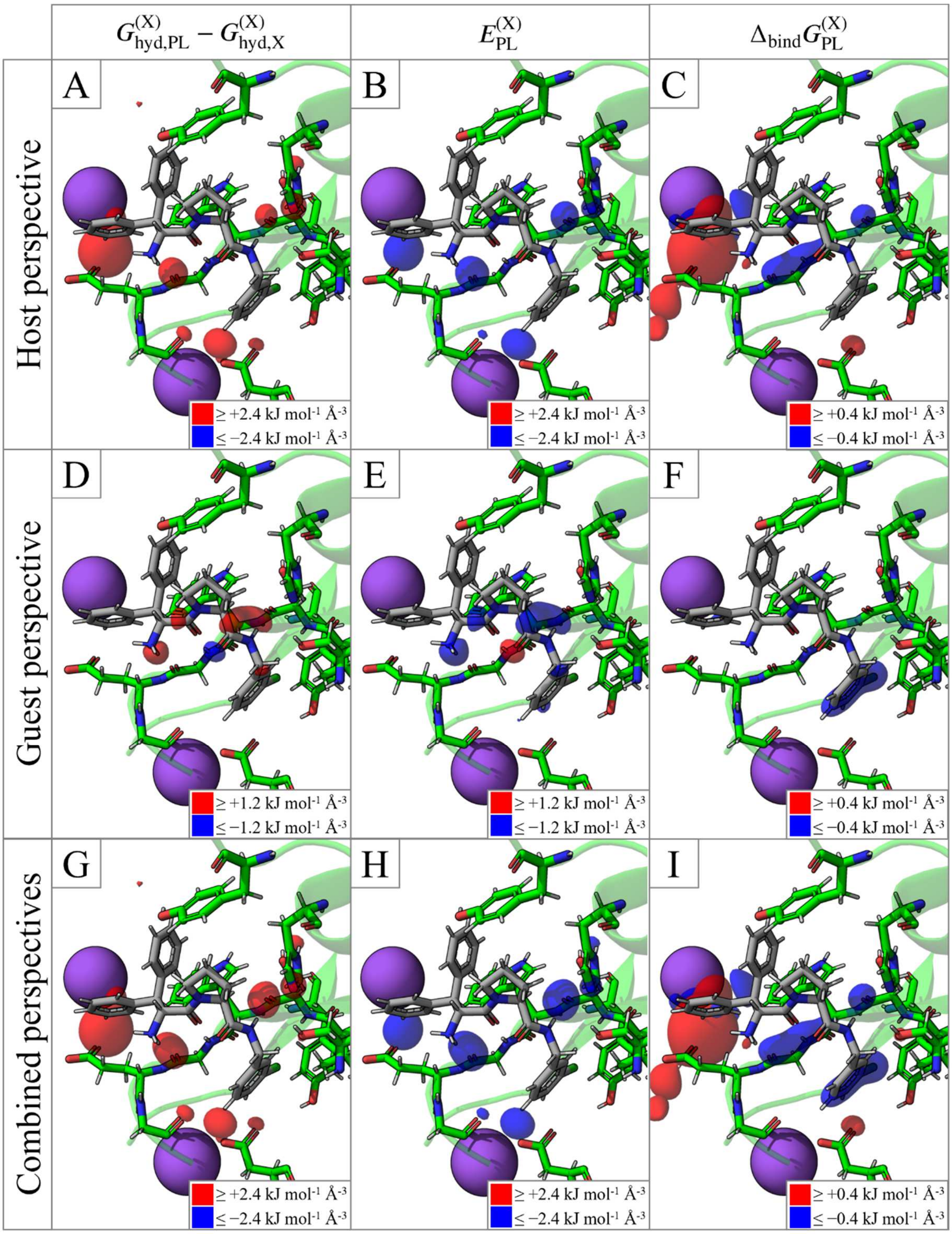

Figure 6.6 Volumetric LFE representation in the protein- (top) and ligand- (mid) perspective of the thrombin complex 3DHK, as well as their combination (bottom). For each perspective, the desolvation penalty (left column), intermolecular energy (mid column), and binding free energy (right column) is given.

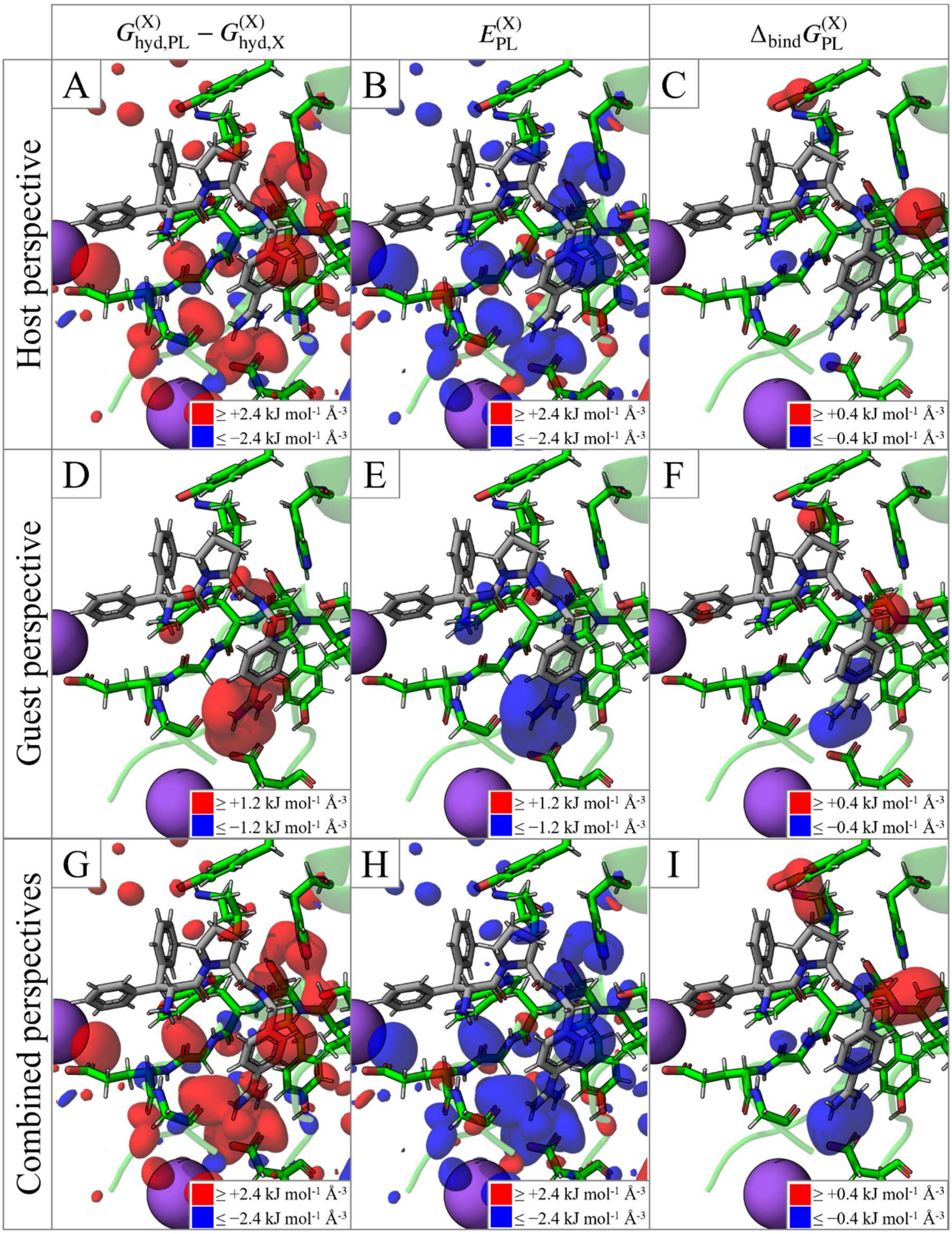

Figure 6.7 Volumetric LFE representation in the protein- (top) and ligand- (mid) perspective of the thrombin complex 2ZO3, as well as their combination (bottom). For each perspective, the desolvation penalty (left column), intermolecular energy (mid column), and binding free energy (right column) is given.

Table 6.5 Ligand atoms of the thrombin complex 2ZFP. For each the derivatives of the desolvation penalty (left), intermolecular energy (mid), and binding free energy (right) with respect to the force field parameters q, ϵ, and σ are given. The unit are of the derivatives w.r.t. q is kJ mol^{-1}e^{-1} and for σ is kJ mol^{-1}Å^{-1}. The derivatives w.r.t. ϵ is unitless. The raw data can be found in the electronic appendix under 3.3/3.3.4/Loc/2ZFP/.

Atom	$\partial\Delta G^{(L)}_{\text{hyd,PL}-\text{L}}/\partial P$			$\partial E^{(L)}_{\text{PL}}/\partial P$			$\partial\Delta_{\text{bind}}G^{(L)}_{\text{PL}}/\partial P$		
Host net charge	$P = q$	$P = \epsilon$	$P = \sigma$	$P = q$	$P = \epsilon$	$P = \sigma$	$P = q$	$P = \epsilon$	$P = \sigma$
N01	119.02	0.01	-6.41	-115.27	-0.96	5.16	3.75	-0.96	-1.25
C02	92.33	2.17	3.05	-81.61	-2.76	-3.34	10.72	-0.6	-0.29
C03	117.25	2.84	2.36	-104.98	-4.23	-3.07	12.27	-1.39	-0.71
O04	106.09	-0.44	-18.82	-101.99	-1.53	11.79	4.1	-1.97	-7.03
C05	72.2	2.58	1.63	-49.45	-3.65	-4.15	22.75	-1.07	-2.52
C06	41.37	2.81	3.2	-25.93	-3.44	-4.31	15.44	-0.62	-1.11
N07	130.83	1.71	2.85	-113.47	-2.62	-5.28	17.36	-0.91	-2.43
C08	177.55	2.91	2.03	-163.78	-5.32	-5.48	13.77	-2.41	-3.45
C09	190.66	3.09	2.45	-188.21	-5.41	-4.31	2.45	-2.32	-1.86
O0A	166.51	1.4	4.57	-170.89	-1.62	-4.43	-4.38	-0.22	0.14
C0B	144.62	3.4	2.32	-133.45	-5.18	-3.29	11.17	-1.78	-0.97
C0C	121.4	3.55	1.07	-92.93	-6.36	-6.16	28.48	-2.81	-5.09
C0D	109.24	2.55	1.23	-93.52	-4.19	-4.76	15.72	-1.63	-3.53
N0E	250.97	0.99	-5.68	-239.50	-0.57	19.88	11.47	0.42	14.2
C0F	238.32	3.47	1.75	-228.85	-5.2	-2.22	9.46	-1.73	-0.46
C0G	233.41	4.27	2.5	-208.51	-7.47	-6.46	24.9	-3.19	-3.96
C0H	249.29	3.63	0.43	-234.64	-7.27	-6.14	14.65	-3.64	-5.7
C0I	303.98	3.54	-2.4	-291.86	-8.63	-5.68	12.12	-5.09	-8.07
C0J	303.57	4.12	-1.13	-284.03	-9.11	-5.95	19.54	-4.98	-7.08
C0K	233.86	4.93	1.47	-210.04	-8.9	-4.36	23.82	-3.98	-2.89
C0L	227.5	4.13	-0.48	-188.03	-8.08	-1.48	39.48	-3.95	-1.96
Cl0M	192.29	2.48	-10.36	-171.93	-5.99	-0.48	20.36	-3.52	-10.84
H0N	119.53	-0.08	-0.15	-126.86	-0.68	-0.21	-7.33	-0.76	-0.35
H0O	199.09	-4.16	-4.34	-188.99	0.71	2.99	10.1	-3.45	-1.35
H0P	75.36	2.5	1.56	-79.35	-0.68	-0.21	-3.99	1.82	1.35
H0Q	72.99	0.94	0.26	-71.28	-1.25	-0.32	1.71	-0.31	-0.06
H0R	47.95	6.85	1.87	-38.76	-4.35	-0.94	9.19	2.5	0.94
H0S	99.95	-1.65	-4	-54.69	-6.5	0.71	45.26	-8.16	-3.29
H0T	22.59	2.42	0.32	-16.88	-3.4	-0.74	5.72	-0.98	-0.41
H0U	15.69	14.29	5.66	-9.15	-5.91	-1.25	6.54	8.38	4.41
H0V	71.37	-0.7	-2.46	-38.00	-6.89	-1.22	33.37	-7.59	-3.68
H0W	225.27	-2.73	-4.35	-202.13	-7.83	0.73	23.14	-10.56	-3.62
H0X	125.58	-0.48	-2.57	-118.83	-7.39	-0.68	6.75	-7.88	-3.24
H0Y	148.9	-2.12	-4.31	-144.28	71.49	39.73	4.61	69.37	35.42
H0Z	121.12	-1.02	-4.19	-71.73	0.17	5.91	49.4	-0.86	1.72
H10	106.96	0.75	-3.24	-87.17	-10.68	-0.34	19.79	-9.93	-3.58
H11	84.5	2.19	-0.31	-64.79	-4.96	-1.03	19.7	-2.77	-1.34
H12	103.29	-0.03	-2.14	-100.22	-5.86	-0.68	3.06	-5.89	-2.82
H13	317.13	-0.07	-1.48	-328.54	6.47	8.47	-11.41	6.4	6.99
H14	215.59	2.09	-0.75	-209.08	-6.83	-1.39	6.5	-4.73	-2.14
H15	256.74	2.45	-2.04	-273.74	21.96	16.35	-16.99	24.41	14.31
H16	243.67	7.85	1.78	-232.65	-8.86	-1.62	11.02	-1.01	0.16
H17	370.26	-1.05	-4.2	-358.61	-13.06	-1.18	11.65	-14.11	-5.38
H18	384.8	2.27	-2.36	-355.55	-14.19	-2.15	29.25	-11.92	-4.51
H19	237.58	0.63	-3.39	-167.70	-6.04	4.52	69.88	-5.42	1.14