Structure Profiling and Geometric Optimization of Protein-Ligand Complexes for the Scoring Function HYDE

Dissertation

with the aim of achieving the degree

Dr. rer. nat.

at the Faculty of Mathematics, Informatics and Natural Sciences
Department of Informatics
Universität Hamburg

Agnes Meyder

Hamburg, February 2020

Gutachter:innen:
Prof. Matthias Rarey
Prof. Johannes Kirchmair
Prof. Christoph Sotriffer

Tag der Disputation: 14.9.2020

Acknowledgements

Ich bedanke mich bei allen Menschen, die diese Promotion ermöglicht haben. Als erstes bedanke ich mich bei meinem Betreuer Prof. Rarey für das Interesse am Thema und die gute Betreuung. Meine Kooperationspartnern Dr. Lange und Dr. Klein bei Bayer sowie dem Team in der BioSolveIT danke ich für die gute Zusammenarbeit im Kontext des Kooperationsprojekts. Mein Dank gilt auch Dr. Schneider, Dr. Schomburg und Dr. Nittinger bei der Durchführung des HYDE-Projekts. Dazu danke ich dem BMBF und Bayer für die Finanzierung des Projekts.

Prof. Kirchmair und N.-O. Friedrich danke ich für die Verwendung meiner Methoden. I am also grateful to Dr. Pearce for the fruitful discussion.

Ich danke der AMD-Gruppe für den immerwährenden Strom an Kuchen, Code-, Diss- und Paper-Reviews, Witzen und Mitleid bei schlechter Datenqualität. Spezifisch danke ich Florian Flachsenberg, Robert Schmidt, Thomas Otto, Nils-Ole Friedrich, Patrick Penner, Jochen Sieg, Rainer Fährrolfes, Stefanie Kampen und Dr. Bietz.

Zu guter Letzt danke ich meiner Familie und Freund*innen, die virtuell oder real nahe waren und über die lange Zeit immer angefeuert haben.

Zusammenfassung

Computergestützter Wirkstoffentwurf verwendet häufig dreidimensionale Proteinstrukturen. Sie sind die Grundlage um Bindetaschen zu analysieren wie auch um neue Ideen für Kleinmolekülmedikamente zu entwickeln. Eins der Ziele dabei ist die Vorhersage der Bindungsaffinität. Sie wird durch die Abschätzung der Beiträge von Interaktionen sowie die Veränderung in der Rigidität der Bindetasche angenähert. Die für diese Aufgabe notwendigen Bewertungsfunktionen müssen entwickelt und auf qualitativ hochwertigen Protein-Ligand-Komplexen mit bekannten Bindungsaffinitäten validiert werden. In dieser Dissertation wird die Bewertungsfunktion HYDE mit einer aktualisierten Version seiner Bewertungsfunktion GeoHYDE zur geometrischen Optimierung ausgestattet. Als Teil der Aktualisierung wurde der *Continuous Torsion Score* entwickelt und die diesem zu grunde liegende Torsionsbibliothek aus dem Jahre 2013 überarbeitet. Da das Ziel von GeoHYDE die lokale Modifikation des Interaktionsprofils zur Maximierung des HYDE Wertes ist, sollte dessen Veränderung als Qualitätsmaßstab betrachtet werden. Zur Messung dieser wird die mittlere quadratische Abweichung (RMSD) zwischen initialer und finaler Atomkoordinaten in Bezug auf die geometrische Optimierung verwendet. Da allerdings das Modell einer Proteinstruktur nur die bevorzugte Interpretation der Elektronendichte ist, sollten kleinste Abweichungen von den initialen Koordinaten des Models weniger relevant sein, so lange sie sich noch im von Elektronendichte bestägten Bereich bewegen. Hierfür wurde in dieser Thesis der *electron density score for individual atoms and molecular fragments* EDIA und EDIA_m für alle Elemente im Periodensystem entwickelt. EDIA_m stellt auch das fehlende Puzzleteil zur automatischen Extraktion qualitativ hochwertiger Proteinstrukturen da. Das daraufhin entwickelte Programm `StructureProfiler` wurde genutzt um den Datensatz ProtFle18 bestehend aus 2386 Taschen zu erstellen, welcher nachfolgend in drei Teile geteilt wurde. Als letzter Teil der Thesis wurde GeoHYDE auf dem Trainingsdatensatz parametrisiert und den zwei Testdatensätzen evaluiert. In 74 zu 79 % aller Fälle stimmen EDIA_m und RMSD bei der Bewertung der geometrisch optimierten Pose als nahe an der kristallienen Pose mit einer mittleren HYDE-Wert Verbesserung von 0.32 kJ überein. Wird Seitenkettenflexibilität auf Proteinseite hinzugenommen, verbessert sich der mittlere HYDE-Wert weiter. Dabei wächst allerdings auch die Rechenzeit um das Vier-

bzw. 15-fache. HYDE in Kombination mit GeoHYDE schneidet im unteren bis mittleren Drittel im Vergleich bei den verschiedenen Testszenarien auf dem externen Validierungsdatensatz CASF-2016 ab.

Abstract

Computational drug design relies heavily on three-dimensional protein structures. They are the foundation for analyzing binding poses as well as developing new ideas for small molecule drugs. One goal in research is the prediction of binding affinity. Such predictions are made by assessing the non-covalent interactions between the small molecule and the protein as well as the change in rigidity of the overall system. Thus, a so called scoring function needs to be defined and validated on high quality protein-ligand complexes with known binding affinity data. In this thesis, the scoring function HYDE is equipped with an updated version of its geometric optimization function GeoHYDE. In the update, the Continuous Torsion Score was newly developed and the underlying Torsion Library of 2013 revised in terms of peaks as well as the torsion rule subset analysis. Since the aim of GeoHYDE is a local revision of the interaction profile to maximize the HYDE score in the given pose, deviations should be observed as a measure for its performance quality. The state of the art metric is the root mean squared deviation (RMSD) between initial and final atom coordinates in regards to the geometric optimization. But since the model of a protein structure is just the most preferred interpretation of electron density observations, slight alterations in the model's coordinates should be less relevant if still confirmed by electron density. Hence, the electron density score for individual atoms (EDIA) and molecular fragments ($EDIA_m$) for any element in the periodic table is proposed in this thesis. With $EDIA_m$, the missing piece in the automatic high quality structure data set assemblage is now present. Thus, the tool `StructureProfiler` was created and the data set ProtFlex18 consisting of 2386 pockets was extracted from the protein data base to consequently analyze the performance of GeoHYDE. As the final part of this thesis, GeoHYDE was parametrized and tested on the training and two test sets extracted from ProtFlex18. In 74 to 79 % of all cases tested, $EDIA_m$ and RMSD both asses the geometrically optimized pose as very close to the crystallized one with a median HYDE score difference of 0.32 kJ. Including side chain flexibility in the pocket, the medians of final HYDE scores further improve but at the cost of at least four times rising computation time. HYDE in combination with GeoHYDE performs in the lower to middle third on the widely used validation data set CASF-2016 depending on the type of the test scheme.

Contents

Chapter 1

Introduction

Numerous species use natural products for treatments. Honeybees collect antimicrobial resin to be included in their nest and in the presence of parasites, fruit flies prefer liquids with a high grade of ethanol to lay their eggs into to ward off parasitic wasps.[10] For a long time, humans have also used medicine to prevent and treat illness. Over the centuries, increasingly targeted pipelines to identify cures were developed. Currently, drug development consists of a multi step process spanning an average development phase of ten years and costing one to two billion dollars until a drug for a specific disease can be released as medicine if successful.[48] While in its initial phases, the target enzyme and possible interacting molecules need to be identified, further 'lead' optimization, tests for possible industrial synthesis, no toxicity to humans and other factors have to be applied. Finally, overall positive treatment effects have to be observed in humans. Over the years, multiple assisting technologies have been developed to further understand the method of actions of medicine in the human body to increase the success of finding a treatment. Especially helpful was the discovery of X-ray radiation in 1895 by Röntgen. Shortly after, protein crystallization and their analysis with X-rays was developed. Since the 1940's, nuclear magnetic resonance (NMR) spectroscopy allows the observation of proteins in solution with increasing resolution.[57] Recently, cryo electron microscopy (EM) has helped to observe membrane proteins which are difficult to be observed with x-ray radiation or NMR.[1] Still, most of the over 450 000 protein-ligand structures up to now have been solved with the help of X-ray radiation.

To determine a structure via X-ray, a solution of the targeted enzyme needs to crystallize in a structured way. Then, the crystal is irradiated with X-ray from multiple directions and the resulting patterns are recorded as intensities I. Through

its regular structure, one section in the crystal can be determined which can reconstruct the whole crystal through symmetric reflection and is thus called unit cell. Subsequently, the atomic model in the unit cell is attempted to be inferred. But because radiation consists of an amplitude, derivable from the measured intensities and their phases, the latter are still missing for full atomic reconstruction. Since they can not be measured, they need to be inferred from I and the atomic model. Via inverse Fourier transformation, electron density ρ at the point (x, y, z) can be retrieved through overlaying sine and cosine waves over the unit cell(Equation 1.1).

$$\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l F_{hkl} \exp^{-2\pi i(hx+ky+lz)} \tag{1.1}$$

$$= \frac{1}{V} \sum_h \sum_k \sum_l |F_{hkl}| \exp^{i\alpha_{hkl}} \exp^{-2\pi i(hx+ky+lz)} \tag{1.2}$$

V denotes the volume of the unit cell and h, k, l are the lattice indices in the reciprocal grid space. [58] F_{hkl} can be further decomposed into $|F_{hkl}|$ as the amplitude which is directly proportional to $\sqrt{I_{hkl}}$ - the actually measured intensities and $\exp^{i\alpha_{hkl}}$ containing the unknown phase per reflection that needs to be determined (Equation 1.2).

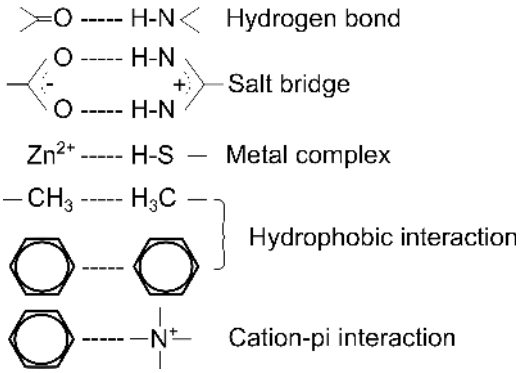
Oversimplified, the so called phase problem is solved by repeatedly suggesting a model, deriving the necessary phases from it and then checking how much the resulting electron density agrees with the suggested model.[33] While solving the phase problem and the overall orientation of the model in the experimental data in the final refinement process, at least two sets of electron density maps are calculated. The experimentally observed electron density with the proposed phases is called fo while the density based on the proposed model is called fc. Those two can be combined to identify errors in the proposed model (missing or surplus atoms, wrongly assigned element) to the $2fo - fc$ and the difference map calculated through computing $fo - fc$. While the first map shows through its contours how much the observed density supports the atomic model, the second map should preferentially have only low electron density.[78] Since electron density is basically a grid with annotated intensities, they can not be displayed in a three dimensional space. Instead, maps are visualized through contour maps an various σ levels. σ in the context of electron density maps denotes the root mean squared value (RMS) of the measured intensities. Since a mean of approximately zero is observed in electron density maps, the abbreviations tend to be used interchangeable. A

contour map at a level of $2\,\sigma$ for example only shows density that has at least an intensity of $2\,\sigma$ above the mean of the map. While in the $2fo-fc$ map a contour level of 0.4 to $1.5\,\sigma$ suggests to an increasing degree an atom[49] in at least one version of the atom model, the difference map is best examined on a contour level of $\pm 3\,\sigma$. Density above and below the interval of $]-3\,\sigma, 3\,\sigma[$ signifies either density not yet explained by the model or atoms without sufficient experimental support. The calculated electron density should further be tailored to the observed one in refining B factor and occupancy per atom. The B factor expresses local disorder due to for example local motion of a loop region or disorder in the crystal while occupancy on the other hands describes the atom's position in multiple conformations. Both are standard values to be optimized per atom in the overall refinement procedure. Consequently, metrics have been developed to estimate the overall agreement of the model with the experimental data. On a global scale, the R measure expresses how closely the amplitudes of the calculated structure factors agree with the observed ones (Equation 1.3).

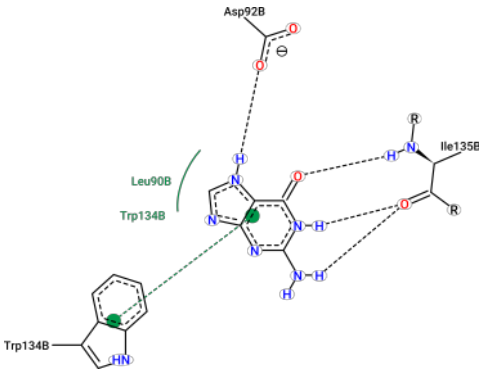
$$R = \frac{\sum ||\mathbf{F}_{obs}| - |\mathbf{F}_{calc}||}{\sum |\mathbf{F}_{obs}|} \tag{1.3}$$

For its computation, the refinement is only run with around 90 % of all reflections. The remaining 10 % are used to compute their correlation with those calculated from the model resulting in R_{free} as an unbiased validation of the agreement between experiment and model.

Depending on the quality of the data, a protein structure model starting from the trace of the peptide backbone in the electron density can be automatically suggested. The identification of cofactors and ligands as well as metals and waters in density has been an area of recent research.[77] Special care is needed when single atoms of an overall ligand are not resolved in otherwise high quality electron density. They might have been eliminated from the overall ligand through the crystallization process. On the other hand, parts of the ligand areas only weakly supported by electron density might be highly flexible. When multiple conformations can be identified, they should each be enriched with a fitting occupancy factor and a B factor to describe their respective movement. If multiple conformers can not be determined, the partial structure needs to receive a higher B-factor to account for the comparatively higher disorder at the position.[11] After understanding the relative positions of the atoms in a protein pocket, the possible driving forces behind the formation of the protein-ligand complex can be evaluated.



(a) Examples for relevant non-covalent protein-ligand interactions (Adapted Fig. 4.3 from Klebe[29])



(b) Phosphorybosyltransferase 1a95 binding to guanine B 304

Figure 1.1: Interactions present in protein-ligand complexes.

1.1 Interactions

Given a three-dimensional protein-ligand complex, protein-ligand interactions can be analyzed (Figure 1.1). Their strength can be expressed by the Gibbs free energy ΔG. On the one hand, ΔG can be mainly described by K_d as the dissociation constant of the protein ligand complex (Eq. 1.4) and on the other hand by the combination of enthalpic and entropic changes upon binding (Equation 1.5).[29]

$$\Delta G = -RT \ln K_d \tag{1.4}$$

$$= \Delta H - T\Delta S \tag{1.5}$$

$$K_d = \frac{[Ligand] \cdot [Protein]}{[Protein - Ligand\ Complex]} \tag{1.6}$$

K_d describes the coefficient of how much unbound protein and ligand in comparison to the protein-ligand complex are in solution (Equation 1.6). R is the gas constant and T the temperature in Kelvin.

Protein-ligand binding can be examined by analyzing the contributions in terms of enthalpy and entropy. When a molecule binds to a complex in aqueous solution, the water hull of both structures need to reorganize and new interactions between the molecule and the complex may form (Figure 1.1). The amount of energy exchanged due to broken and newly created interactions between all components and the bulk water is called enthalpy ΔH. The temperature dependent entropy $-T\Delta S$ is the second component that may explain why two molecular structures bind. Components of a system strive to have similar degrees of freedom. A binding event where in the end some components are more flexible than before

is preferred over one that introduces rigidity into a formerly flexible area. Hence, combining hydrophobic interactions may result in low enthalpic gain but overall, depending on the situation, in an increase of states the system can be in as the hydrophobic surface to the bulk water is reduced. Water molecules then have more options to interact in the bulk. A scoring function to computationally assess the binding affinity of a protein-ligand complex aims to approximate ΔG. The entropic contribution can be roughly estimated through the degree of buriedness of the ligand in the structure pocket after binding. Estimating the enthalpic contribution to ΔG is often estimated in examining polar interactions.

A quantification of such a contribution to the overall binding affinity can be tried through evaluating the solubility in the aqueous solution of the atom's functional group through analyzing experimentally measured octanol-water partition coefficients [66] ($\log P$) in regards to functional groups per ligand. In the following, a brief overview over the scoring function HYDE based on partial $\log P$ increments is given.

1.2 HYDE

Hydrogen bonds are the strongest non-covalent interaction type. Hence, a method for rapidly assessing binding affinity should focus on the possible and actually formed hydrogen bonds between the protein and ligand in the unbound and bound state. Since certain elements such as carbon are known to only interact weakly with polarized groups, they can be treated as apolar. Exposing such apolar atoms to a hydrophilic area can be seen as unfavorable. On the other hand, exposing polar groups to hydrophobic surroundings e.g. through binding can also have a negative effect.

$$\Delta G = \Delta H - T\Delta S \tag{1.7}$$

$$\Delta G_{\text{HYDE}} = G_{\text{HYDE}}(\text{bound}) - G_{\text{HYDE}}(\text{unbound}) \tag{1.8}$$

$$= \sum_{\text{atoms a}} \Delta G^{a}_{\text{saturation}} + \Delta G^{a}_{\text{dehydration}} \tag{1.9}$$

$$\Delta G^{a}_{\text{dehydration}} = \Delta G^{a,\,\text{polar}}_{\text{dehydration}} + \Delta G^{a,\,\text{apolar}}_{\text{dehydration}} \tag{1.10}$$

$$\Delta G^{a,\,\text{apolar}}_{\text{dehydration}} = -2.3RT \cdot p \log P^{a} \Delta acc^{a} \tag{1.11}$$

HYDE aims to quantify the changes in HYdration and DEsolvation to estimate the binding affinity in a protein-ligand complex on the basis of the Gibbs free energy equation 1.4. In the following, its underlying principles are sketched out. More

details can be found in the dissertations of Eva Nittinger and Nadine Schneider [64], [40] with the accompanying publications [65], [66], [67].

For each atom in the binding pocket, the difference between the bound and unbound state of the active site 8 Å around the ligand according to the HYDE theory is accumulated (Equation 1.8). It is split into estimating the change in saturating hydrogen bond functions and the change in exposing hydrophobic areas to the aqueous surrounding (Equation 1.9). Each atomic fraction is then expressed with the prefactor RT resulting from Equation 1.4 multiplied by 2.3 to convert from the natural to the common logarithm combined with a multiple of the atom type's partial $\log P$ function. The exact value depends on the atom being apolar or polar and its surroundings (Equation 1.10).[66] Apolar atoms only add to the HYDE energy if a change in the molecular surface occurs through binding to estimate changes in entropy (Equation 1.11).[67]

$$\Delta G^{a,\,\text{polar}}_{\text{dehydration}} = -2.3RT \cdot p \log P^a \sum_{\text{HBond}} w^h \cdot p^h_{\text{dehyd}} \tag{1.12}$$

$$f^{FSI}_{dev} = f_{dev}(PWP_{best}) \tag{1.13}$$

$$p_{dehyd} = 1 - f^{FSI}_{dev} \tag{1.14}$$

$$\Delta G^{a,\,\text{polar}}_{\text{saturation}} = -\frac{2.3RT}{F_{sat}} \cdot p \log P^a \sum_{\text{HBond}} w^h \cdot f^h_{\text{dev}} \tag{1.15}$$

In contrast, polar atoms contribute if their desolvation and saturation state change when interacting with the modified surroundings.

The desolvation probability for polar atoms describes the penalty for an unsaturated hydrogen bond function (Equation 1.12). The current HYDE model has an updated desolvation detection to compute a free space identification (FSI) [41] The desolvation probability is thus determined by the quality of the hydrogen bond f_{dev} to the implicitly placed water at the first available potential water position (PWP) detected by the FSI (Equations 1.13, 1.14). If no PWP was found, the polar atom is fully desolvated and thus penalized. In practice, all explicit waters are removed before running the geometric optimization. Afterwards, waters can be placed again into the pocket[41] and finally the ligand is scored with HYDE. If a polar atom takes part in a hydrogen bond, the quality of its bond is described by f_{dev} in the HYDE saturation equation 1.15.

In the case, that an atom has multiple interactions, smooth transitioning between them over optimization steps is necessary to always describe a function with a

gradient.

$$w_h = \begin{cases} 1 & \text{if \#IAs = 1} \\ \dfrac{(f_{dev}^h)^2 + \left(\left(\sum_{\text{IAs k}} p_{dehyd}^k\right) - p_{dehyd}^h\right) \cdot 0.0001}{\sum_{\text{IAs k}} \left(f_{dev}^k\right)^2 + (\#\text{IAs} - 1)\left(\sum_{\text{IAs k}} p_{dehyd}^k\right) \cdot 0.0001} & \text{if \# IAs > 1 and } a \text{ is not water} \\ \dfrac{(f_{dev}^h)^2 + \left(\left(\sum_{\text{IAs k}} p_{dehyd}^k\right) - p_{dehyd}^h\right) \cdot \frac{1}{16}}{\sum_{\text{IAs k}} \left(f_{dev}^k\right)^2 + (\#\text{IAs} - 1)\left(\sum_{\text{IAs k}} p_{dehyd}^k\right) \cdot \frac{1}{16}} & \text{if } a \text{ is water} \end{cases} \tag{1.16}$$

Hence, w_h in Equation 1.16 includes on the one hand f_{dev} as well as the desolvation probability for each of the atom's hydrogen bond functions multiplied with a weight of 0.0001. In the HYDE version of 2018, f_{dev} combines four quality factors of a hydrogen bond with the help of the Hoelder mean: the distance between donor and acceptor atom as well as the distance between the donor hydrogen and the acceptor lone pair ([41] Sec B.1.1, here Figure 5.1(a)). Every quality factor is defined by three values: the optimum, the maximum deviation from the optimum, which is still considered acceptable and the maximum deviation where the quality factor is not yet zero (Figure 5.1(a)). The overall interaction geometries were evaluated on crystallographic data and adjusted to the derived interaction schemes.[42] Manual analysis then revealed the necessity to update the cone angle maximum optimum and overall maximum of the generic donor, nitrogen acceptor and water donor interaction geometry. The angle was adjusted for the generic donors from (0°, 15°, 40°) to a relaxed (0°, 30°, 55°), and for the nitrogen acceptor to (0°, 35°, 70°). The cone angle of the water donor was relaxed from (0°, 15°, 45°) to (0°, 30°, 55°) Metal geometries are defined in the absence of a ligand and left unchanged also after binding.

1.3 GeoHYDE

HYDE is a scoring function applicable for any protein-ligand pocket. It prefers interaction geometries which are in accordance to the ones in crystallized complexes. Slight local adjustments may make the difference between a low and a high quality hydrogen bond. Thus, a fast geometric optimization of the protein-ligand pocket should be conducted before scoring with HYDE. The modifications can happen in the overall position of the ligand in the pocket, modification of its atom coordinates but also on the side of the protein in slightly adjusting e.g. amino acid side chains.

HYDE was provided with an optimization function for geometric pose optimization of the ligand in a rigid pocket called GeoHYDE (Equation 1.17).

$$\begin{aligned} GeoHYDE &= w_{sat} \cdot \Delta G_{Saturation} + w_{iLJ} \cdot GeoHYDE_{desolv} \\ &\quad + w_{desolv} \cdot \Delta G_{desolv\ polar\ atoms} + w_t \cdot E_{Torsion} + w_{rLJ} \cdot E_{Lennard-Jones\ intramolecular} \end{aligned} \tag{1.17}$$

In its version from 2012[65], it consisted of a term for the torsion conformation of the ligand and its intra-ligand Lennard-Jones term to safeguard against unusual ligand twists. The terms were combined with a stripped HYDE term without all terms for apolar atoms. Hence, only the degree of desolvation and saturation of polar atoms was part of GeoHYDE. Both terms use a limited set of weights based on the quality of the interaction. Throughout the geometric optimization, the quality of the hydrogen bond may change thus changing its associated weight. This term in its multiple variations over time does not have an analytical gradient which forces GeoHYDE to work with a gradient free optimization procedure. The in HYDE employed approach to calculate the hydrophobic effect through approximating the exposed surface was computationally too expensive. Hence, clash and the hydrophobic effect were approximated through the use of an intermolecular 6-12 Lennard-Jones potential denoted $GeoHYDE_{desolv}$.

The objective function was optimized by a Quasi-Newton method using the numerically estimated derivatives of GeoHYDE. The step size was limited to 1000 in the version of 2012. The overall performance of GeoHYDE was never benchmarked.

The quality of results generated by GeoHYDE could be supposedly improved by introducing protein side chain flexibility. Additionally, GeoHYDE of 2012 has two areas of great concern: the torsion angle scoring is unspecified and optimizing through calculating finite differences with a Quasi-Newton method is outdated. Hence, in both areas recent developments are summarized in the following two sections to lay the foundations for improvement in this thesis.

1.4 Torsion Angle Scoring

Dihedral angles describe in combination with the structure's connectivity the conformation of the structure. Torsion angles as a derivation of dihedral angles are calculated over four connected atoms thus populate the interval [0°, 360°]. Due to sterical effects and those created through orbital hybridization, only limited zones of the interval are most likely populated by torsion angles in a similar environment.

Hence, a limited set of torsion angles can cover the actual conformational range of a molecule with an acceptable accuracy.

After Schrödinger established the foundation for quantum mechanics, the conformational energy and thus the preference per torsion angle is in theory computable. In practice, computation time is a problem. The most precise but also computationally most expensive strategy considers an atom as a many electron wave. The consideration of correlation effects between electrons increases the precision but also the computation time with the Hartree-Fock theory being the simplest strategy. Its algorithmic complexity is at least N^4 where N is the number of spin orbital base functions.[51] The second strategy named density function theory (DFT) only grows close to linearly with an increasing number of atoms. DFT estimates electron density distribution based on the positions of electrons and perturbing potentials.[32] One DFT single point computation can be solved within minutes but to understand preferred molecular conformations, a large number of calculations have to be run. DFT can also be extended with empirically derived parameters which assists in reducing the computational time and but also precision. A force field such as OPLS3[20] on the other hand can determine the conformational energy hyperplane faster but with a considerable error margin. The more than 48,000 torsion angle parameters in OPLS3 were determined by fitting them through using quantum chemical computation of more than 11,000 molecules. Thus, force fields can be used for binding affinity estimation. Statistically derived torsion angle functions normally do not aim to explicitly assist in binding affinity estimation. Statistics can be derived for proteins or small molecules. The rotamer library used in the de-novo folding function ROSETTA is based on high quality amino acid conformations extracted from 3985 protein chains.[70]

The Cambridge Crystallographic Database (CSD) can serve as the base for identifying highly likely torsion angles for small molecules. In a brute force approach in 2006, MIMUMBA[61] created over 52,000 torsion rules consisting of four atoms describing one dihedral angle. Angles per rule over all 20,000 molecules in the training set derived from the CSD were accumulated and the derived highly frequent torsion angles per torsion rule evaluated on the remaining CSD test set of ca. 11,000 molecules. Taylor *et al.* followed suit in 2014 and published a derivative of Mogul to run a similar atomic fragment based exhaustive profile enumeration on the CSD.[76] Schärfer *et al.* have instead created a knowledge based torsion library as part of the software package NAOMI (TorLib13).[62] It describes covalently bound substructures with the graph based molecular pattern language

SMARTS.[9] The torsion rule SMARTS patterns consider recursively described environments as well as additionally attached atoms and lone pairs as nodes in the pattern. Statistics were accumulated over 130,463 molecules (CSD13). They have also evaluated the torsion library on a set of PDB ligands to discuss its applicability on PDB ligands. Taylor *et al.* acknowledged the positive effects of a manually curated torsion library and suggested their library to be a good starting point for a modified knowledge based one. Torsion Libraries and other histogram based statistics can be converted to a continuous potential with the help of a periodic normal distribution - the von Mises function.[35, 70] The in NAOMI present TorLib could hence be combined with a von Mises function resulting in a continuous differentiable scoring method suitable for GeoHYDE to determine the likeliness of torsion angles. But since the HYDE term gradients are not known, the overall objective function needs to be optimized with a gradient free method.

1.5 Gradient Free Optimization

Scoring functions describe their own energy landscape. Scoring a protein-ligand complex without geometrically optimizing it first may miss the close local optimum in the hyper plane. Depending on the scoring function, such an optimization can be executed with or without a gradient. If the gradient is available, the low memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) is currently the method to go.[34] If not, two options are possible. Either, the gradient free optimization function can be locally smoothed [72] to then be optimizable with a gradient based method or algorithms such as the non-stochastic optimization algorithm *bound optimization by quadratic approximation* (BOBYQA) need to be tested on their performance [60]. A software package with a variety of gradient free optimization algorithms is `NLopt` (free and open-source library for non-linear optimization[25]). It can be integrated in e.g. C and C++ based software packages with an easy to use interface that allows switching between optimization algorithms. Formerly, numerical differentiation through approximation by finite differences was often employed. Due to the number of necessary evaluations and its sensitivity to numerical instabilities in the function the method is generally discouraged to be used.[44] Instead, a number of other gradient free methods are available since around 1960. They all have in common the use of a polygon of at least n+1 points when n denotes the number of dimensions. Nelder-Mead-Simplex developed in 1965 [39] spans a simplex of n + 1 points in n dimensions. It moves the simplex through three operations

over the hyperplane to converge on a local minimum without computing any derivatives. `NLopt` allows to use an NMS with bound constraints [6][59]. NMS is known to not be able to converge in some cases so that improvements such as Sbplex are proposed. Sbplex as a reimplementation of Subplex evaluates the NMS repeatedly in sub spaces. This results in needing function evaluations growing only in linear with the number of dimensions in contrast to the NMS. Sbplex was also extended to contain bound constraints as in [6]. `NLopt` also offers the use of PRAXIS[7] as an update to the gradient free optimization over choosing conjugate directions developed by Powell in 1964[52]. PRAXIS resets the search directions not to e.g. the identity matrix but to an orthogonal matrix, related to the function to be optimized related via eigenvalues. Thus, search directions are better spread out and the algorithm has a faster convergence speed than the one by Powell. While this ability can be quite interesting, PRAXIS is superseded by the performance of more recently developed algorithms.[60]

Powell developed COBYLA in 1994[53], NEWUOA in 2007[55] and BOBYQA in 2009[56]. COBYLA as in Constrained Optimization By Linear Approximation optimizes in each step a linear polynomial interpolation of the original function F at the vertexes in a trust region. A trust region is an area where the function approximation is assumed to be highly similar to the underlying objective function. In `NLopt` support for bound constraints and some other improvements were added. But since linear models can not include the curvature per point, the use of a quadratic approximation is advised. More recently, NEWUOA[54] extended in `NLopt` by bound constraints, was added. It creates a quadratic polynomial approximation of F, extended by `NLopt` with the MMA algorithm and bound constraints in a spherical trust region. BOBYQA as in Bound Approximation by Quadratic Approximation is in a nutshell NEWUOA extended with bound constraints [56] in `NLopt` present in the original implementation by Powell translated to C. Following the `NLopt` documentation, BOBYQA performs better than the altered NEWUOA in many cases.[25]

`NLopt` allows two types of termination criteria. If desired, the optimization is stopped when a maximum predefined number of steps is reached or a certain computation time is exceeded. A target score can also be set for which the optimization should stop when reached. Hence, convergence abilities of the algorithms can be compared in benchmarks. The second type of termination criteria are a set of absolute or relative convergence criteria on the function value f or optimization parameter values x. Since the final values are not known, the change in f or x

between steps can instead be observed. Depending on the algorithm, absolute change such as $|\Delta f|/|f|$ is less than the relative function tolerance or $|\Delta f|$ is less than the absolute function tolerance can be a reasonable termination criteria.

Hence, multiple algorithms are available to potentially succeed the Quasi-Newton approach in optimizing GeoHYDE. Their performance in terms of score optimization and computation time need to be analyzed.

The focus of GeoHYDE is letting an interaction converge to the crystallized geometry. Thus, optimization with GeoHYDE on high quality crystallized structures should result in minor deviations from the start structure. One strategy is to measure pure spatial displacement through the root mean squared deviation (RMSD). On the other side, spatial displacement in an area well-defined by electron density is less accepted in contrast to an area with spatial displacement and conspicuous electron density. Hawkins *et al.*[22] have searched for alternative measures but had to note poor correlation of the established metrics with the RMSD. In the following, multiple scoring schemes are examined for their ability to incorporate the use of electron density into the computation of the degree of spatial displacement.

1.6 Evaluation of Spatial Displacement

The root mean square deviation is the measure of choice to determine deviation between two sets of coordinates. There are many chemically more conscious methods available. GARD[3] has been developed as a normalized variant of the RMSD in the interval from zero to 1. It allows weighting e.g. hydrophilic against hydrophobic areas but the weights need to be adapted to the use case thus comparability over different use cases can get lost. A second approach is to compute the general positional uncertainty[18] and use the value to offset RMSD values. But as a global approach local certainty is not reflected in a global measure.

Besides coordinate based approaches, one step back could be to evaluate the expected number of interactions. The interaction-based accuracy classification (IBAC) and related methods assess the interaction pattern of the reference protein-ligand complex and compare its recreation with the different ligand conformation. Non interacting regions are left out. Those should be monitored as well since unnecessary movement should be avoided by GeoHYDE. Interaction changes should be accepted if the proposed conformation allows multiple interaction points. Thus, the pure focus on interaction reproduction seems not to be a good fit.

Electron density measurements allow a very close comparison of a ligand configuration if it is backed up by experimental data. The comparison with electron density allows to automatically capture the degree of a region's rigidity. There are a number of scoring schemes available to determine the deviation of coordinates from electron density. The easiest one is to either globally or locally compute the (squared) sum of errors over $f_o - f_c$ as the one used in Coot. In real space on the $2fo - fc$ map, two main methods are currently in use to check the agreement between model and electron density. In 1991, the real space R factor (RSR) was proposed.[26]

$$\mathrm{RSR}(area) = \frac{\sum |\rho_{obs} - \rho_{calc}|}{\sum |\rho_{obs} + \rho_{calc}|} \tag{1.18}$$

$$\mathrm{RSCC}(area) = CC(\rho_{obs}, \rho_{calc}) \tag{1.19}$$

$$\mathrm{RSR}_n = \frac{\mathrm{RSR}_d}{\mathrm{RSR}_c} \tag{1.20}$$

For a specified area such as a residue, the observed density ρ per atom was compared to the expected one to result in a score in the interval of 0 (good) to 1 (bad correspondence). Since the original publication does not define the radius to be used per atom or the scaling factor between both density components, RSR scores are difficult to compare and implementations of the metric differ. In the PDB, residues can be checked with a normalized RSR (RSR-Z) against the average RSR quality per resolution. Such data ist not made available for small molecules though. The real-space correlation coefficient avoids the need for a scaling factor but still operates on unspecified atom radii. Both also do not account for diverse electron density spacing. Hence, they are not resolution independent. A further advancement to allow the comparison between the crystallized pose and those proposed by docking is RSR_n.[86] Here, the RSR of the crystallized pose is used as the denominator to normalize the RSR of the docking pose. Hence over multiple structures, the ratio between both values can be compared and the best fitting docking pose identified. Neither is an implementation of RSR_n available nor does it handle superfluous density. Recently, the real-space difference density Z score (RSZD) and RSZO were introduced.[78] Both metrics analyze the $fo-fc$ map. RSZO measures the precision of the map through reporting the signal-to-noise rate, which should be above 1σ to allow model building in this area. RSZD reports significant measurement outliers that indicate badly modeled areas in the structure, hence values beyond the range $[-3\sigma, 3\sigma]$ should lead to further examination of the area. In all proposed metrics operating in real-space, each atom's B factor and occupancy influence the shape of

the expected electron density. The metrics operating on the difference map demand for each docking pose a recalculated difference map which makes high throughput screening computationally expensive. Hence, a good method to evaluate poses with regard to the flexibility observed in the crystallization but are not present at the refinement state of the crystal does not yet exist. A data set selected according to such a metric is not available as well.

1.7 High Quality Data Sets

When considering X-ray crystallography data, a sufficient number of reflections and an overall correspondence between atomic model and the experimental data is necessary. A special emphasis is put on the active site, where the position of the ligand should be properly defined. Flexible residues should also be properly identified. Over the years, extensive efforts have been made in manually assembling data sets of various sizes. In 2007, the Astex Diverse Set with 85 protein-ligand complexes was published.[21] It includes numerous tests for the quality of the ligand and some for the overall model. The subsequently released Iridium data set with 207 protein-ligand complexes marked as highly trustworthy further evolved the criteria catalog.[84] It includes more tests for model quality and switched to a well known electron density correlation estimation method followed by manual examination of the structure in its electron density. Both sets include a large set of structures with a resolution of worse than 2 Å. As benchmarking GeoHYDE requires an analysis of the change in quality of the interaction geometries, the position of atoms needs to be highly exact to begin with. A resolution of worse than 2 Å can not guarantee this in all cases. Luckily, the number of structures in public databases have risen tremendously in the past years. After the development of the EDIA (see Chapter 2), a filtering method exists which is stricter than earlier methods and also applicable to geometrically optimized structures. It was used to extract the Platinum data set with 4548 ligands.[16] As Platinum's purpose was conformer generator validation, the quality of the pocket was not controlled for residues well supported by electron density taking part in an interaction. Hence, the pure platinum data set can not be used in a validation scenario where the quality of the interactions is relevant. Also, no automatic tool chain for objectively creating an validation data set was published to be used with ease.

HYDE is also used to predict binding affinity. Thus to fully benchmark HYDE, data sets with highly trustworthy binding affinity data are needed. The correlation

of predicted values with experimental binding affinity can be compared through correlation coefficients that either compare exact values or their rank. An example for the first is the Pearson correlation coefficient $r_{X,Y}$ (Equation 1.21) that analyses the covariance between the two sets of values X and Y divided by the product of each standard deviation σ. r_s over the ranks of the values in X and Y is named Spearman correlation coefficient and abbreviated with r_s (Equation 1.22). Kendall's *tau* on the other hand uses the number of concordant and disconcordant pairs normed by the overall number of possible pairs to compute a rank correlation coefficient. If $x_i = x_j$ and $y_i = y_j$ the pair is excluded from the denominator. If both $x_i < x_j$ and $y_i < y_j$ or reversed with >, the pair is counted as concordant (Equation 1.23).

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})}} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{1.21}$$

$$r_s = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}} \tag{1.22}$$

$$\tau = \frac{\#(concordant\ pairs) - \#(disconcordant\ pairs)}{\binom{n}{2}} \tag{1.23}$$

PI is the predictive index proposed by Pearlman *et al.*[50]. C_{ij} is the ranking difference between a pair of ligands. W_{ij} is the weight extracted from the observed binding affinities of the two ligands. The larger the difference between the binding affinities, the more importantly the ligands should be placed in the correct rank. P_i is the model score and E_i the experimental binding affinity. Overall, PI ranges from -1 to 1 (perfect agreement).

$$PI = \frac{\sum_{j>i}^{n}\sum_{i}^{n} W_{ij}C_{ij}}{\sum_{j>i}^{n}\sum_{i}^{n} W_{ij}} \tag{1.24}$$

$$W_{ij} = \left|E_j - E_i\right| \tag{1.25}$$

$$C_{ij} = \begin{cases} 1 \text{ if } \dfrac{E_j - E_i}{P_j - P_i} > 0 \\ -1 \text{ if } \dfrac{E_j - E_i}{P_j - P_i} < 0 \\ 0 \text{ if } \dfrac{E_j - E_i}{P_j - P_i} = 0 \end{cases} \tag{1.26}$$

The hereby introduced variations of correlation coefficients are utilized in a current benchmark data set named CASF-2016. It consists of 57 target proteins with 285 ligands and their binding affinity data in K_d or K_i.[75] The pockets have been selected from the PDBbind refined set to guarantee a minimum sequence similarity

of 90 % and a wide spread of known binding affinities. They need to differ at least 100-fold in one cluster to be beyond the intrinsic error in reported binding affinity data from different laboratories. Also, the ligands have been checked for uniqueness and to avoid stereo isomers. Four tests are available for the evaluation of a scoring function. The "scoring power" is examined by computing the Pearson correlation coefficient and the standard deviation of the linear correlation between predicted score and annotated binding affinity. With the help of r_s, τ and PI, the ranking ability in each complex cluster has been analyzed. The docking power of the scoring function is assessed by evaluating a ligand with 100 decoys to identify the ligand as the most fitting pose. In addition, the scoring function hyperplane's resemblance to a funnel was analyzed with Spearman's rank correlation coefficient. The last test evaluates the screening ability of the scoring function by analyzing a cross-docking per complex cluster over all CASF ligands. Enrichment factors and the number of highly ranked ligands in the first, five, and tenth percentage are determined. Additionally inverse screening is now also possible with the given data set and quality measurements. In both cases, cross-binders have been identified in ChEMBL and considered in the evaluation. All tests use the bias-corrected and accelerated bootstrapping method to allow the calculation of confidence intervals. Results can be also compared with the posthoc Friedman test with the Shaffer's method to identify statistically relevant performance differences.

There are also other data sets available but they either tend to be very small or not publicly available. As inhouse data set, the cooperation partner of this project BioSolveIT has found a number of 'small series' that offer high quality crystal structures with binding affinities measured in the same lab per series. Another data set is used in Schrödinger's FEP validation.[82, 83, 73] All data sets mentioned in this chapter and their overlap can be found in Tables 3.1, 3.2.

The CASF validation set is too small to use parts of it as training set and parts of it as in- and out of domain test sets. Additionally, the necessary high quality in the pocket needs to be fulfilled according to EDIA (Chapter 2) that can also evaluate the performance of GeoHYDE.

1.8 Motivation and Thesis Content

As outlined, computationally predicting binding affinities is connected to a number of areas of ongoing research. From validation data sets, preferentially annotated with binding affinity to the proper pose optimizing scoring functions bundled with

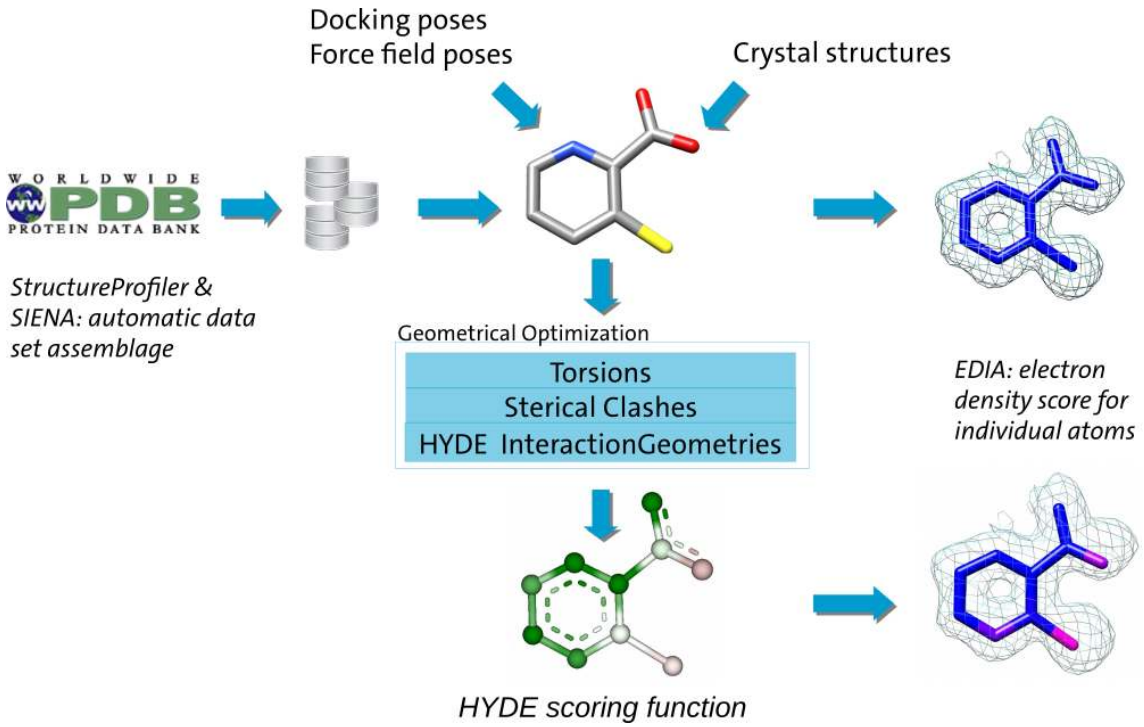


Figure 1.2: Scope of the dissertation

their numerical optimization algorithms to metrics that estimate the closeness of the proposed pose to the actual experimental data - none of these areas have been able to present ready-to-use solutions yet. In this thesis (Figure 1.2), the electron density score for individual atoms (EDIA) and molecular fragments (EDIA_m) was developed. It uses the $2fo - fc$ map to estimate the support of atoms and substructures not present at the refinement phase. With the help of EDIA_m as the missing link in the tool chain for automatically profiling three dimensional protein structures in the search for high quality structural data, the release of `StructureProfiler` combining all necessary tests was possible. Hence, the 2386 pockets large ProtFlex18 data set was extracted from the publicly available database PDB in 2018. The size of the data set has allowed the split into training and in-domain as well as out-of-domain test data sets with additional similarity and protein flexibility analysis with SIENA. The size of the data set also allows sound statistical analysis and makes it possible to find multiple structures with similar characteristics to profoundly analyze trends in behavior in the geometric optimization.

Since GeoHYDE was missing a soundly defined torsion angle potential, the Continuous Torsion Score was subsequently developed. On a side note, multiple corrections were applied on the Torsion Library such as the automatic subset analysis with SMARTScompare. Equipped with the CTS, GeoHYDE was then eval-

uated with a number of gradient free optimization algorithms, parameterized on ProtFlex18, evaluated for its performance on pockets with flexible side chains and finally compared to the external validation data set CASF-2016.

The thesis has partially been conducted as part of the Project P47 in the Cluster BIOKATALYSE2021 and was jointly sponsored by the company Bayer and the German Federal Ministry of Education and Research BMBF under the grant number 031A183B.

All developed software was implemented as an extension of the C++ NAOMI software library. All C++ code was subject to code review, unit, and system tests. The new code has a unit testing coverage of at least 90%. Qt and boost as additional libraries have been used in the standalone tools for e.g. license checks and program option parsing. Multiple Python3 frameworks have also been added to the tools for data analysis. All tools, code libraries and frameworks are presented in Appendix A. Eleven publications, two talks and three posters have been published as part of this thesis and listed in Appendix C.

Chapter 2

Evaluation of Spatial Displacement for GeoHYDE

Initial examinations of the state-of-the-art validation sets have revealed a high number of structures with a resolution of worse than 2Å which makes the determination of the for HYDE necessary interaction geometries difficult. High quality metrics such as the RSZD demand an $fo - fc$ map for each pose to be scored which makes the metric computationally not feasible. Other real-space metrics are incompletely defined. Additionally, the use of atomic B factors and occupancy of e.g. ligands should be avoided. With a version of the electron density score for individual atoms (EDIA) available to analyze the existence of crystallized waters, an incremental improvement suggested itself. Hence, the electron density score for individual atoms (EDIA) and molecular fragments (EDIA_m) was developed as part of this thesis.

2.1 The electron density score for individual atoms and molecular fragments

The idea behind EDIA is the approximation of the gold standard currently in use to evaluate the presence of atoms in a model based on experimental data. Its original design for checking the existence of water oxygens was developed by Eva Nittinger *et al.* [43]. In this thesis, it was extended to be able to handle any element of the periodic table and supplemented with an error analysis. With the help of the power mean, one score for a set of atoms such as a whole molecule could be derived that can guide the automatic identification of high quality pockets for future validation benchmark data sets.

Depending on local disorder captured in a B factor and the resolution of the crystallized complex, atoms of a specific element and charge show a certain expected electron density spread. Below 2Å resolution, this electron density approximates a sphere.[14] Thus EDIA calculates over a resolution dependent sphere centered at each atomic coordinate the weighted electron density. When determining the expected electron density radius for an atom, a resolution dependent average B factor is used to avoid well documented weaknesses [2, 79].

With the help of the structure's connectivity, one can predict areas with and without electron density. Hence, EDIA uses a weighting scheme $w(p, a)$ per atom a considering each grid point p that positively weights electron density in the expected radius and negatively weights it beyond that sphere up to two times the electron density radius called the *sphere of interest*. Additionally, electron density grid points p can be present in multiple *spheres of interests*. We use the term ownership per grid point to determine the distance based degree of ownership of each atom on a specific grid point $o(p, a)$.

The electron density intensity at grid point p named $z(p)$ itself is truncated to the interval of 0 to 1.2σ. σ in the context of electron density maps denotes the root mean squared value (RMS) of the measured intensities. The abbreviations tend to be used interchangeable since a mean of approximately zero tends to be observed in electron density maps. The interval limits stem from properly weighting high electron density intensities in the inner sphere against spotty density observed in the outer sphere area with a radius of $[r, 2r]$.

$$
\begin{aligned}
EDIA(a) &= \frac{\sum^{p \in M_{2fo-fc}} w(p,a)o(p,a)z(p)}{\sum_{p \in M_{2fo-fc} \,|\, w(p,a)>0} w(p,a)} && (2.1)\\
\overline{pa} &= \|p - a\|_2 \text{ (distance)} \\
w(p,a) &: \text{Weight function depending on the distance } \overline{pa} \text{ (see below)} \\
o(p,a) &: \text{Ownership of } p \text{ from } a \text{ (see below)} \\
z(p) &= \begin{cases} 0 & \text{if } \frac{\rho(p)-\mu}{\sigma} < 0.0 \\ \frac{\rho(p)-\mu}{\sigma} & \text{if } 0 \leq \frac{\rho(p)-\mu}{\sigma} \leq \zeta \\ \zeta & \text{if } \frac{\rho(p)-\mu}{\sigma} > \zeta \end{cases} \\
\zeta &= 1.2 \\
\rho(p) &: \text{Density at } p \\
\mu &: \text{Mean of the } 2fo - fc \text{ map} \\
\sigma &: \text{Standard deviation of the } 2fo - fc \text{ map}
\end{aligned}
$$

An additional remark about the use of σ needs to be made: σ and the mean of an electron density map is known to be dependent from the resolution, the B factor of the data and the solvent content of the crystal. Hence, comparing σ levels between different experiments to understand if a structure is supported should be done while considering the three biasing factors as well. Typically, structures with supporting density of at least 0.4σ are increasingly probable to be present. Generally σ level of 1 fully support a modeled atom. EDIA uses this rule of thump of the crystallographic community to allow an automatic identification of highly probable atoms. This may result in flagging inconspicuous models as problematic but we advice to use EDIA values not blindly. Conspicuous areas should instead be examined for the cause of the flagging. After manual examination, the structure could still be of high enough quality to be used in the user's specific scenario.

In the following, components of EDIA are explained in more detail.

Electron Density Radius Determination

EDIA evaluates electron density in a sphere. Its radius is B factor and resolution dependent. Hence, to avoid the dependence on B factors, resolution interval dependent mean B factors were determined with the help of the structures in the PDB. Subsequently, the electron density radii for each element with its various charges are determined and tabulated. The radius for an atom of a specific element and charge in a specific setting is then linearly interpolated based on the tabulated radii values. In the following, the determination of the mean B factors and the computation of the electron density radii are explained in more detail.

The average B factor distribution in the PDB up to a resolution of 3Å was analyzed. The results for the ranges [Å] are: $]0; 0.5]$ with 7, $]0.5; 1.0]$ with 12, $]1.0; 1.5]$ with 18, $]1.5; 2.0]$ with 26, $]2.0; 2.5]$ with 39, and $]2.5; 3.0]$ with 56 Å^2. They were rounded to a multiple of five : 10, 15, 20, 50, and 55 Å. Since the publication of EDIA, diverging metal B factors were observed. Hence, a B factor analysis focused on metals and ions was additionally conducted. It revealed diverging mean B factors especially for the resolution interval $[1.5, 2.0$Å$[$ (Table 2.1). The mean B factor per element was updated in the implementation, if at least ten data points for averaging were available from the PDB.

As described, the electron density of an atom depends on its element and charge,

B factor, resolution, and the amount of data available from the experiment.[78]

$$\mathrm{RI}_{r_{max}} = \int_0^{r_{max}} \rho(r)\mathrm{d}r \tag{2.2}$$

$$\rho(r) = \frac{8}{r} \int_{s_{min}}^{s_{max}} f(s) \exp^{-Bs^2} \sin(4\pi rs) s\mathrm{d}s \tag{2.3}$$

Following the procedure published in Tickle, EDIA uses the electron density radius for which the Radius Integral (RI) of the tested radius r_{max} is 95% of the overall possible RI with r = 3Å (Equation 2.2). For the computation of $\rho(r)$, the atom type depending scattering factor $f(s)$ together with $s_{max} = 0.5 d_m in$ with d_{min} being the resolution present in the observation are necessary. The parameters per atom type to compute $f(s)$ can be looked up in the *International Tables for Crystallography 1999*.

The ratio $RI_r / RI_{3\text{Å}}$ was calculated for the resolutions d_{min} 0.5, 1.0, 1.3, 1.5, 1.8, 2.0, 2.5, and 3.0 and their respective mean B factor. The value of the electron density radius r in the interval $[0, 3]$ with a step size of 0.01 for the respective combination of d_{min} with its B factor was selected, that crossed over the 95% ratio border. The radius is then used as offset to linearly interpolate for a given resolution from a structure the expected electron density radius.

The resulting updated radius offsets for metals and ions can be found in Table B.1. All other radii offsets are published in the original EDIA publication. Subsequently, the electron density grid intensity in the *sphere of interest* needs to be accumulated.

Electron Density Grid Oversampling

With an increasing resolution, the electron density grid spacing increases. As the minimum expected electron density radius is 0.78Å for eg. silicium^{4+}, the grid is oversampled to guarantee a maximum grid spacing of 0.7Å. Hence, the space diagonal d is divided by 0.7Å and rounded up to receive the partitioning factor p. The electron density is calculated by cubic interpolation when demanded.

Grid Point Ownership

Subsequently, each grid point in the sphere of interest of an atom then needs to be examined for its affiliation to neighboring atoms. While Meyder *et al.* give a formal explanation, Figure 2.1(a) shows a visual explanation of the ownership $o(p, a)$. With atom a to be evaluated, grid points beyond its *sphere of interest* are disregarded (P**1**). Points such as P**2** in the *sphere of interest* but outside of the sphere for which density

Element	[0.0, 0.5Å[	[0.5, 1.0Å[	[1.0, 1.5Å[	[1.5, 2.0Å[
Aluminium	-	-	7.74 (4)	16.67 (31)
Barium	-	8.79 (1)	13.27 (16)	27.4 (80)
Beryllium	-	-	8.25 (3)	17.75 (53)
Bromine	-	22.12 (2)	25.32 (129)	32.4 (1118)
Cadmium	-	9.84 (2)	22.68 (258)	33.08 (1501)
Caesium	-	-	24.53 (12)	42.03 (72)
Calcium	-	10.93 (63)	13.74 (1337)	23.02 (8655)
Chlor	-	11.14 (23)	20.04 (1938)	30.12 (10838)
Cobald	-	8.57 (1)	14.97 (165)	26.24 (1046)
Copper	-	6.23 (12)	17.36 (412)	21.12 (1679)
Fluorine	-	-	-	20.01 (1)
Europium	-	-	19.55 (11)	27.05 (29)
Gadolinium	-	-	13.64 (3)	26.54 (106)
Gallium	-	-	-	22.3 (5)
Gold	-	-	56.02 (13)	50.46 (58)
Holmium	-	-	25.44 (3)	25.44 (3)
Iodine	-	-	37.79 (11)	32.48 (109)
Iron	-	4.89 (37)	10.23 (1377)	18.17 (8233)
Kalium	-	10.82 (4)	18.88 (200)	26.71 (1597)
Lead	-	-	-	33.08 (30)
Lithium	-	14.83 (1)	12.1 (31)	15.52 (61)
Lutetium	-	-	-	30.89 (3)
Magnesium	-	9.81 (15)	17.69 (1258)	24.76 (7997)
Manganese	-	11.73 (13)	13.49 (275)	21.91 (2064)
Mercury	-	8.86 (2)	25.37 (59)	32.71 (503)
Nickel	-	21.68 (1)	15.73 (123)	26.93 (731)
Palladium	-	-	22.02 (7)	40.37 (69)
Platinum	-	-	30.03 (12)	44.59 (167)
Praseodynium	-	-	26.13 (8)	25.58 (17)
Rhenium	-	-	21.6 (12)	22.23 (17)
Rhodium	-	-	-	36.85 (42)
Rubidium	-	-	-	40.46 (30)
Ruthenium	-	8.5 (1)	12.3 (7)	35.84 (101)
Samarium	-	-	-	34.87 (7)
Scandium	-	-	7.65 (1)	7.65 (1)
Silver	-	-	16.68 (6)	21.78 (16)
Sodium	-	10.11 (26)	22.51 (875)	27.54 (5437)
Strontium	-	4.42 (5)	16.93 (24)	29.49 (95)

Element	[0.0, 0.5Å[	[0.5, 1.0Å[	[1.0, 1.5Å[	[1.5, 2.0Å[
Tantalum	-	-	-	33.32 (48)
Tellurium	-	-	7.22 (1)	91.8 (20)
Terbium	-	-	-	41.83 (2)
Tin	-	-	-	38.25 (5)
Uranium	-	-	15.52 (2)	19.67 (4)
Vanadium	-	-	22.59 (16)	35.2 (109)
Ytrium	-	-	23.4 (11)	31.46 (44)
Ytterbium	-	-	35.09 (16)	33.85 (68)
Zinc	-	9.62 (21)	13.69 (1236)	24.18 (8430)

Table 2.1: Average B factor ($Å^2$) per metal and ion in the PDB. The number of hits per resolution interval is given in brackets. Values are colored, if they deviate more than 2.5$Å^2$ from the originally determined mean B factor. Green highlights a drop in mean B factor and red marks an increase in mean B factor for the element.

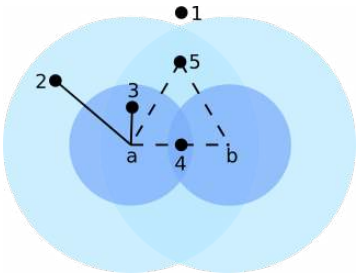
is expected without claims from additional atoms solely belong to a. If the grid point is part of the inner sphere of atom a but only in the outer sphere of any atom b, the second atom's claim is ignored. If both atoms share grid points in either both the outer or inner sphere, both claim ownership. If a is covalently bound to b, both receive an ownership of 1 for the grid point. If not, the atoms share the ownership in accordance to the distance to the respective center so that the total sum of o between all sharing atom is 1. If the set of such atoms is denoted X, the ownership of one point for atom a is calculated as follows:

Point Weighting

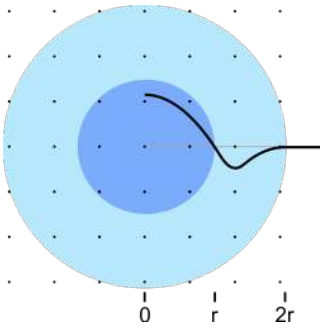
All electron density grid points in the sphere of interest for which the atom a has some degree of ownership are then weighted in accordance to their distance to the atom's center. As shown in Figure 2.1(b) electron density in the sphere with the radius r is scored in the interval $[0, 1]$ while grid points in the outer sphere are scored in the interval $[-0.4, 0]$. The weighting curve consists of three quadratic parabolas. They are parametrized with $r = 1$ to the values in Table 2.2 to achieve an volume integral of zero over the sphere of interest. The supporting material of Meyder *et al.* includes scripts to reproduce the aforementioned calibration of the parabolas.

Error Types Detectable with EDIA

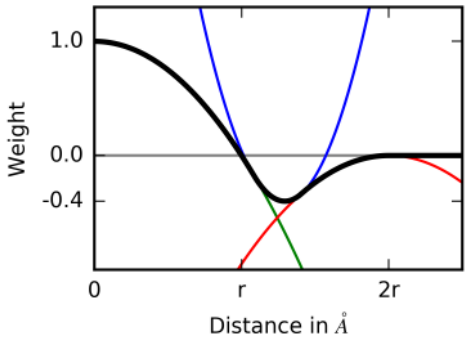
The components of EDIA allow a deeper analysis of the detected problem. When focusing on the information given by $o(p, a)$, overlapping electron density spheres



(a) Ownership visually explained



(b) $w(p,a)$ changes over the spere of interest .



(c) $w(p,a)$ consists of three parabola.

Figure 2.1: Weighting curve $w(p,a)$ over the sphere of interest of atom a of the size of twice the electron density radius r. c reprinted with permission from [37]. Copyright 2017 American Chemical Society.

P	m	c	b	P[i] → P[i+1]
1	-1.0	0	1.0	1.0822
1	5.1177	1.29366	-0.4	1.4043
1	-0.9507	1.0	0.0	2

Table 2.2: Parametrization of $w(p,a)$ with $r = 1$ of the parabola with the form $P(x) = m(x-c)^2+b$. The last column lists the switching points between the parabola.

of non-covalently bound atoms can be identified. In the EDIA error output, a clash (Equation 2.4) for the atoms a and b are reported, when more than 10% of the grid points p in the inner electron density sphere s are shared between atom a and b.

$$clash(a,b) = \frac{2 \cdot \left|\{p \in s(a) \cap s(b)\}\right|}{\left|\{p \in s(a)\}\right| + \left|\{p \in s(b)\}\right|} > 0.1 \tag{2.4}$$

If the weighted sum over all grid points p in the outer electron density sphere is above 0.2, superfluous electron density $EDIA(a)_-$ is reported (Equation 2.5).

$$EDIA(a)_- = \frac{\sum^{p \in M_{2fo-fc} \,|\, w(p,a)<0} w(p,a)o(p,a)z(p)}{\sum_{p \in M_{2fo-fc} \,|\, w(p,a)<0} w(p,a)} > 0.2 \tag{2.5}$$

If on the other hand less than 0.8 is reached with the weighted sum over all grid points in the inner sphere s, missing electron density $EDIA(a)_+$ is reported (Equation 2.6).

$$EDIA(a)_+ = \frac{\sum^{p \in M_{2fo-fc} \,|\, w(p,a)>0} w(p,a)o(p,a)z(p)}{\sum_{p \in M_{2fo-fc} \,|\, w(p,a)>0} w(p,a)} < 0.8 \tag{2.6}$$

EDIA$_\mathrm{m}$

Furthermore, the accumulation of all EDIA scores over a set U of covalently bound atoms such as a residue or a whole ligand with the help of the power mean results in the score named EDIA$_\mathrm{m}$ to rapidly identify inconspicuous components. The correction of +0.1 is a temporary safeguard against an overly strong influence of an EDIA score very close to zero.

$$EDIA_\mathrm{m}(U) \;=\; \left(\frac{1}{|U|} \sum_{a \in U} \left(EDIA(a) + 0.1\right)^{-2} \right)^{-\frac{1}{2}} - 0.1 \tag{2.7}$$

The power mean with an exponent of -2 results in giving single scores close to zero a strong influence on the final EDIA$_\mathrm{m}$ towards the lowest score present in the set of scores. Hence EDIA$_\mathrm{m}$ is a metric suitable to be an indicator for a small set of conspicuous atoms being part of the molecular fragment to be scored. To aid with automatic analysis, EDIA$_\mathrm{m}$ can be annotated with the overall percentage of well-resolved interconnected atoms (OPIA)

Software Update

In contrast to the results reported in our EDIA publication, we have added a set of improvements:

- Atoms in close neighborhood but part of a symmetry copy are now considered thanks to Florian Flachsenberg. NAOMI was also extended to include beforehand not processed ligands.
- A computation error was detected that only half of the negatively weighted space around an atom was analyzed.

While the first and second improvement changes most of the scores by less than 0.1 but metals and ions. We repeated all numerical stability experiments and recomputed all plots based on the PDB ids published as Supporting Information in the original publication but with electron density maps from August 10th, 2018.

Consistency between PDB header and CCP4 density annotation

It was brought to our attention, that certain structures such as 4tmn (CASF-2016) receive low EDIA scores even though they are being used as validation structures in our community. Further examination revealed a well-formed electron density which was falsely oriented. The orientation can be extracted from both the PDB file as well as from the electron density file. In our software, we use the alignment matrix from the electron density file. We scanned the PDBe of August 10th in search for disagreeing H matrices with an epsilon of 0.5 to only detect certain outliers and found 15 structures. We have notified the PDBe about the list of complexes and EDIAScorer now warns the user if a mismatch between the H matrix is detected.

Numerical Stability

The stability of EDIA was tested over multiple artificial examples (Figure 2.3, 2.4). The cases are geared in observing the change of EDIA moving slightly in an discrete electron density grid surrounded by changing levels of electron density support depending on the experiment. More information about the experimental design can be found in [37] and in the Appendix A.2.4. Overall EDIA scores change strongly coupled to the amount of electron density in the vicinity. The update has resulted to further reduce the average EDIA score for experiments with unaccounted density simulated in the sphere of interest beyond one electron density radius r MTS, ATQ, MTM and ATF and ATOF (2.5).

As a result, three score intervals were identified:

- $[0.8, 1.2]$: Atom is highly supported by electron density.
- $[0.4, 0.8[$: Atom is supported by conspicuous electron density.

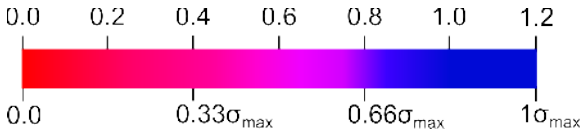


Figure 2.2: EDIA color scheme. Reprinted with permission from [37]. Copyright 2017 American Chemical Society.

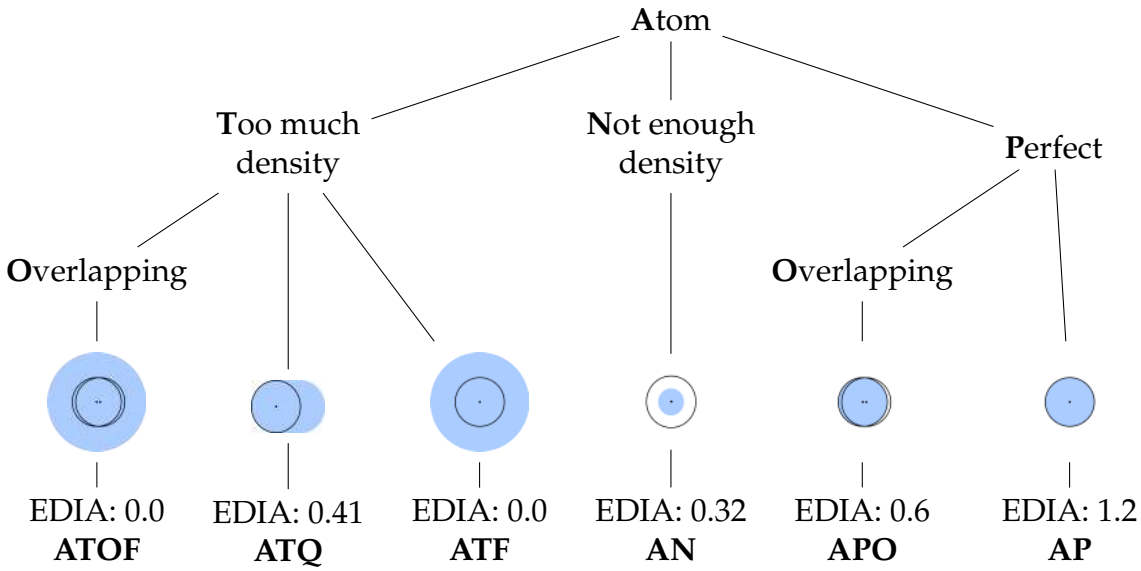


Figure 2.3: All constructed examples with abbreviations for a single atom. Blue denotes the given electron density. Black circles around the atom center have the radius equal to the expected electron density radius. F: fully, Q: quarterly filled $d(a)$ with electron density.

- $[0.0, 0.4[$: The electron density around the atom is highly conspicuous.

which can then be translated into a color scheme ranging from red (0.0) to blue (EDIA of 1.2).

2.2 Results

EDIA is examined in the following sections in various ways following all experiments, already published in the original publication. In every case, changes between the results and the original publication are listed. Supporting examples are included in the chapter where necessary and their scores updated. As a start, the Protein Data Bank is scanned for inconspicuous small molecules bound e.g. by proteins. The analysis of the ligands in the well-known validation data set Astex Diverse Set follows. Then EDIA and EDIA_m are compared to B factor, RSCC,

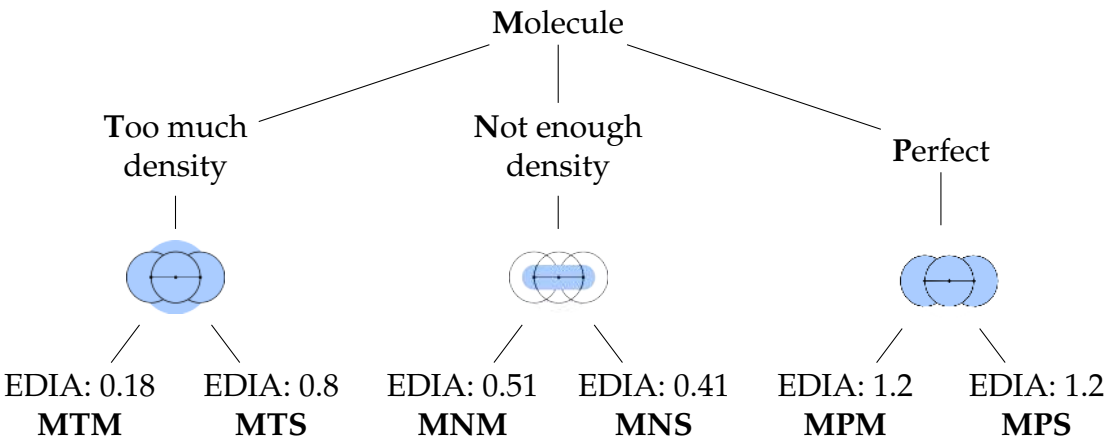
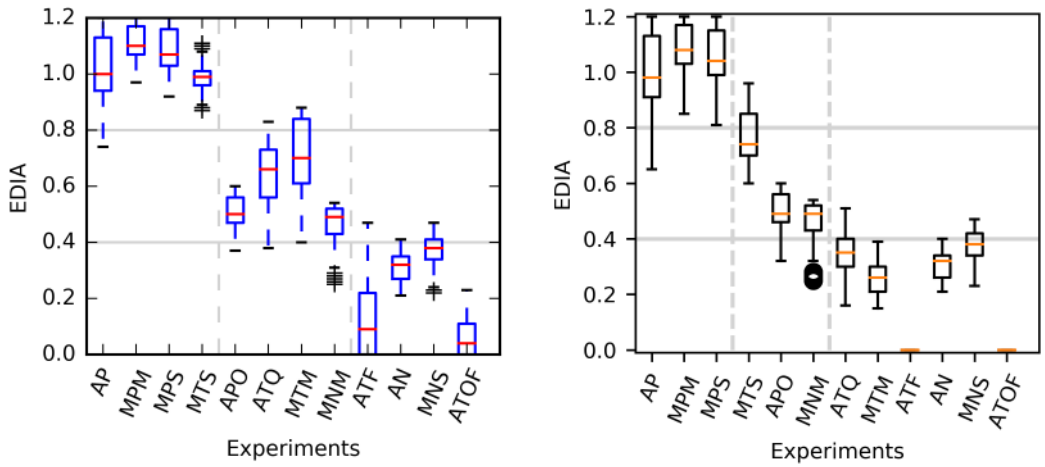


Figure 2.4: All constructed examples with abbreviations for a molecule with three atoms. Blue denotes the given electron density. Black circles around the atom center have the radius equal to the expected electron density radius. M: middle atom, S: atom on the side of the molecule.



(a) Published

(b) Updated

Figure 2.5: Sampling results on the sampled artificial examples. Abbreviations are from Figure 2.3 and 2.4. a reprinted with permission from [37]. Copyright 2017 American Chemical Society.

RSZD, RSZO, and RMSD to understand through a comparative analysis the various aspects of the new metric.

2.2.1 Quality Assessment of Ligands in the PDB

Subsequently, all ligands in the PDB in high quality structures were screened. Since 2017, 41 structures were retracted from the PDBe thus 32803 of originally 32844 structures remained for computation. 66,42 % of 47,712 ligands (originally 76.7 % of 45,113, Figure 2.6) show an $EDIA_m$ of at least 0.8 suggesting high potential for deriving a high quality data set for evaluating protein-ligand interactions. Updated Examples with various EDIAm and OPIA values can be found in Figure 2.7.

2.2.2 Analysis of the Astex Diverse Set with $EDIA_m$

The beforehand introduced Astex Diverse Set of 85 pockets was analyzed with $EDIA_m$. The reevaluation has increased the number of ligands below 0.8 $EDIA_m$ from four to eight. The already in Meyder *et al.* depicted examples are updated and displayed in Figure 2.9. Combined with a resolution cutoff of 2Å 48 pockets remain as a high quality validation data set extracted from the Astex Diverse Set data set (Figure 2.8).

2.2.3 B Factor Comparison

B factor and occupancy are values adjusted in the refinement phase when building the model. As $EDIA_m$ avoids using the such derived B factors per model but instead uses an resolution interval dependent average value, both metrics can be compared to understand their commonalities and their differences. In the following, the original versus updated findings are given. In the updated data set from the PDB, 32,803 structures were analyzed. Initially, 16% disagreement between EDIA and B factor was detected. 5210 residues had a B factor beyond 175% of the expected B factor for the resolution interval while EDIA reports the residue as well-supported (case **1**). On the other side, 36 residues had a B factor of maximally 25% of the expected B factor while EDIA reports a strongly conspicuous (case **2**). After the code update, 2940 structures report an $EDIA_{m]}$ of at least 0.8 (**1**) while 64 structures can be found in case **2**. Overall 9% of all structures report a strongly deviating B factor with an unexpected $EDIA_{m]}$ value (Figure 2.10). As shown in the original publication, structures with case **1** often show stretched out electron density with fuzzy borders. Case **2** structures are often residues with multiple conformations for

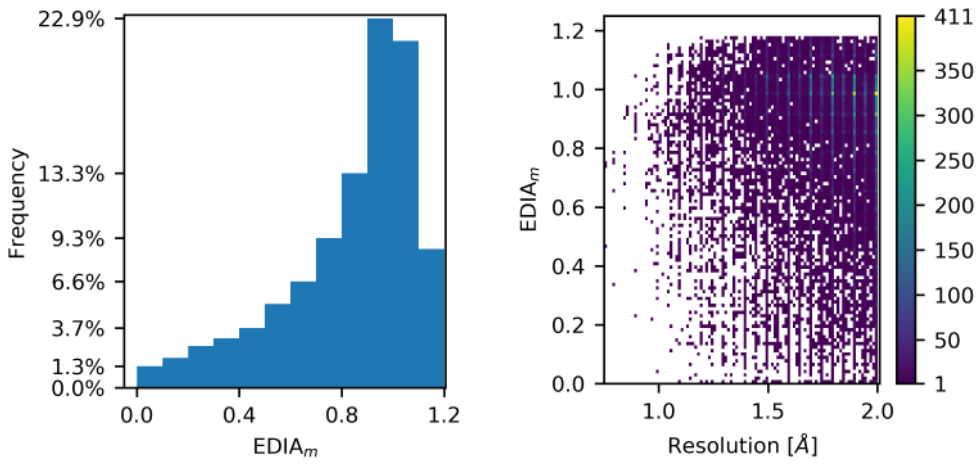
which each conformation has its own set of occupancy and B factor values. Since there is no definite information available from the crystallographic community to understand which level of electron density among other factors has to exist for which degree of occupancy, $EDIA_m$ is currently not able to properly evaluate alternate conformations. In some cases, crystals can also be highly ordered, strongly deviating from the mean B factor determined for the resolution interval. Those cases are wrongly assined to on the up side enforce a high accuracy to identify inconspicuous structures.

2.2.4 Comparison with RSCC

EDIA and $EDIA_m$ assist in similar scenarios where previously the RSCC was used. Hence, an analysis was conducted to further the understanding into both metrics. Mapman[30] was used to calculate an RSCC for the 8283 residues in the Iridium HT closer than 10Å to the ligand. Its atom radius was set to 1.5Å. In the following, results are given and compared to those published in the original EDIA publication. The correlation between $RSCC_{Mapman}$ and $EDIA_m$ show a slightly increasing correlation from 0.62 to 0.68 with 82% of the residues categorized as well-resolved (Figure 2.11(a)). As the RSCC uses the precomputed f_c map with the B factors and occupancies provided by the crystallized structure, weak density is modeled in the map the higher the B factor and the lower the occupancy is. $EDIA_m$ is instead not influenced by both metrics thus marks such areas as conspicuous to suggest further examination (Figure 2.12a) As Mapman reports no atomic RSCC scores, `EDIAscorer` includes an RSCC implementation using the oversampled grid of EDIA and a Gaussian shaped f_c. The previously published correlation coefficient between both metrics drops now from 0.86 to 0.82 over 66009 data points. Further examination showed the sensitivity of the RSCC to the shape of the presented electron density. If a slimmer shape is detected, the RSCC value drops stronger than the EDIA (Figure 2.12) due to the weighting scheme in EDIA allowing blurring density borders.

2.2.5 Comparison with RSZD and RSZO

With the help of EDSTATS, a comparison between $EDIA_m$ and RSZD and RSZO was possible. As both scores are reported for the set of backbone atoms and side chain atoms per residue in the Iridium HT pockets, $EDIA_m$ was adjusted to allow a score comparison on the identical atom sets. The original evaluation was



(a) (b)

Figure 2.6: a) The distribution of all $EDIA_m$ of the 47712 evaluated ligands in the high quality PDB subset. 66.42% are well resolved with an $EDIA_m$ of at least 0.8. b): ligand $EDIA_m$ versus resolution is visualized as a heatmap.

published in Meyder *et al.*. Through the update in EDIA calculation, no substantial changes in the comparison could be detected. $EDIA_m$ agrees for 83% with RSZD (former 85%) and 84% (former 86%) with RSZO in marking the atom sets as well-resolved. $EDIA_m$ is again more sensitive with now 13% of the atom sets in its medium range (before: 11%, Figure 2.11(b)) which are still seen as well-resolved by RSZD and RSZO. Figure 2.13a shows an example for which $EDIA_m$ detects conspicuous electron density. Both RSZD and RSZO do not mark any of the two atom sets in Glutamate 241 I as problematic as the associated high B Factor explains for them the smeared density at this position. Figure 2.13b depicts Leucine 42 A with weak density for which missing data is only reported by RSZO and not by RSZD but again with $EDIA_m$. Hence, $EDIA_m$ summarizes conspicuous areas for which information can also be found in occupancy, B Factor RSZD or RSZO in a single score. After identifying such regions with EDIA, exploration with additional metrics and information can then be able to identify the possible cause to decide if the substructure is still usable for the specific use case.

2.2.6 $EDIA_m$ vs. RMSD

In molecular modeling, RMSD is the metric of choice to analyze deviation from the original structures. Since the RMSD uses the exact atomic coordinates for comparison, neither locally limited areas of motion can be considered nor does the RMSD

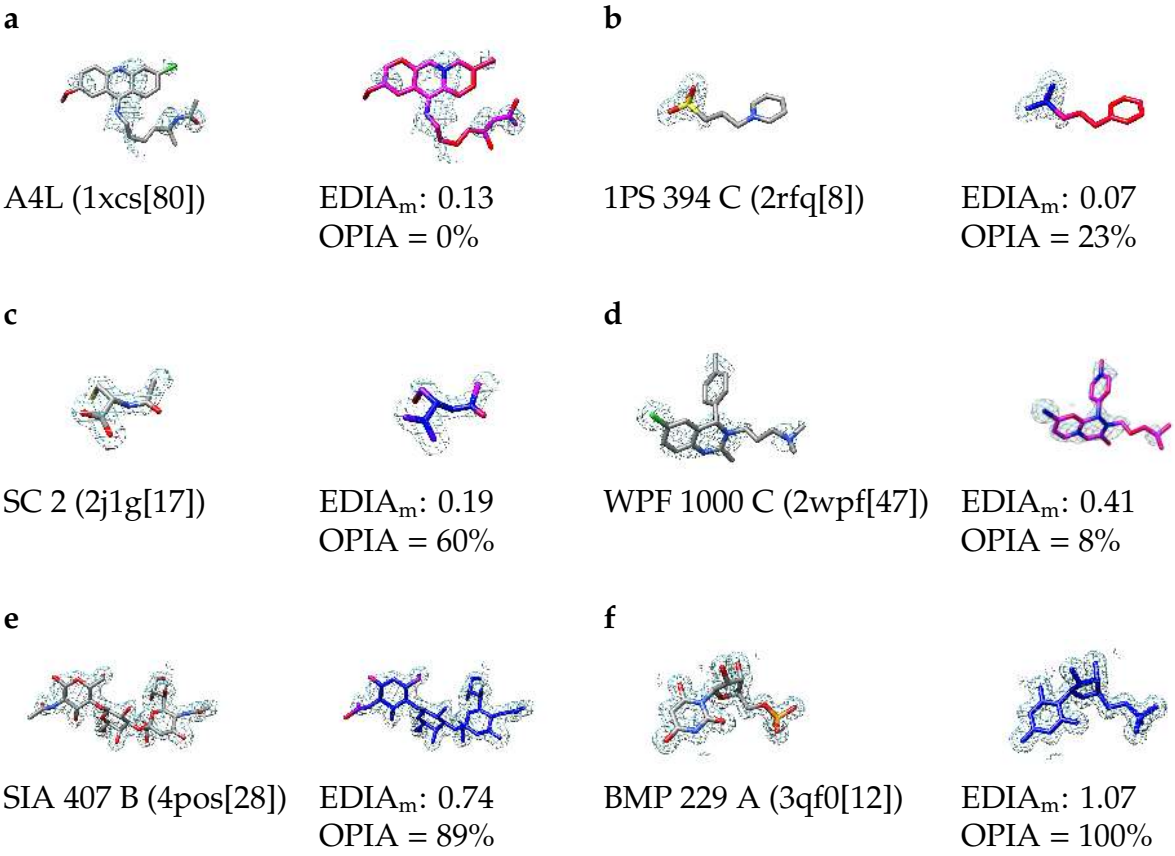


Figure 2.7: The updated set of PDB ligands with various EDIA$_m$ and the rounded percentage of atoms in good substructures (OPIA) values published in Meyder *et al.*. **a-c** show similarly low EDIA$_m$ scores but strongly deviating OPIA values. SC2, 1PS, and A4L partially consists of atoms with an occupancy below 1.

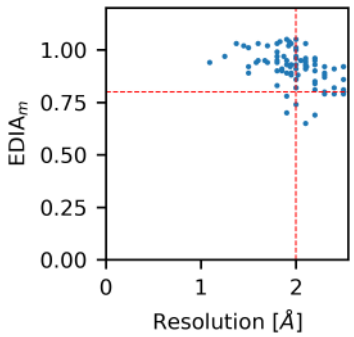


Figure 2.8: EDIA$_m$ of all Astex Diverse Set ligands against their resolution. 45 ligands with a resolution of at least 2Å and an EDIA$_m$ of at lest 0.8

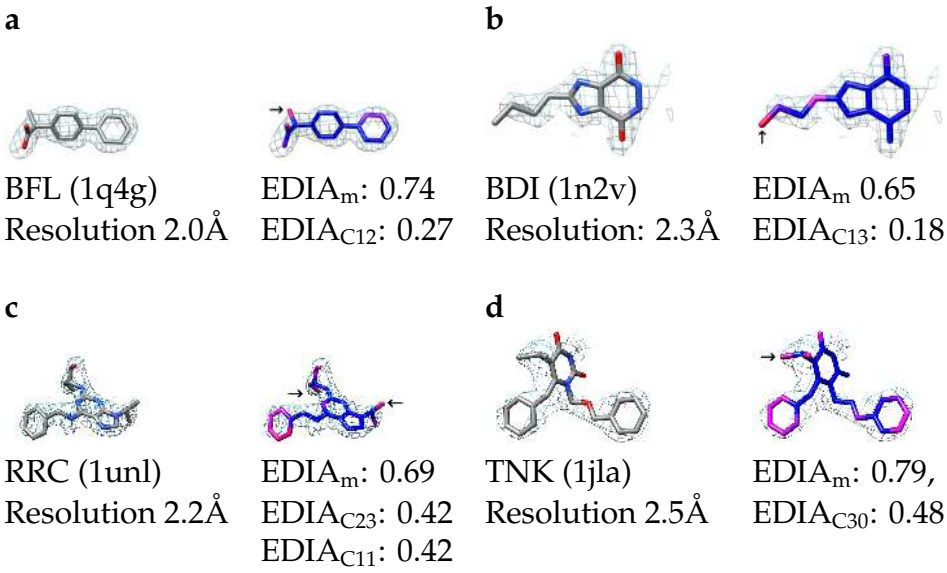


Figure 2.9: A group of four ligands in the Astex Diverse Set with an EDIA$_m$ below 0.8 are shown. The minimal atomic EDIA scores per molecule are annotated and marked in the picture. The $2fo - fc$ map is shown at 1σ.

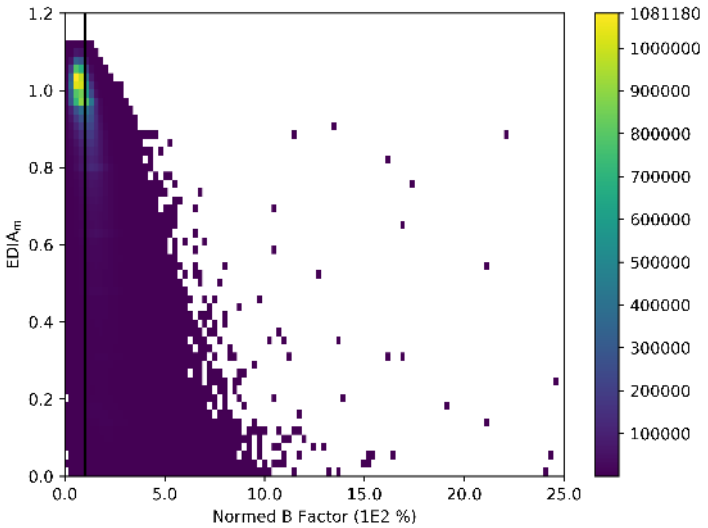
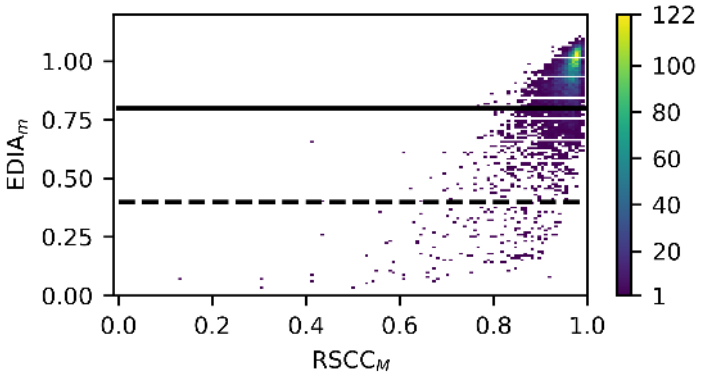
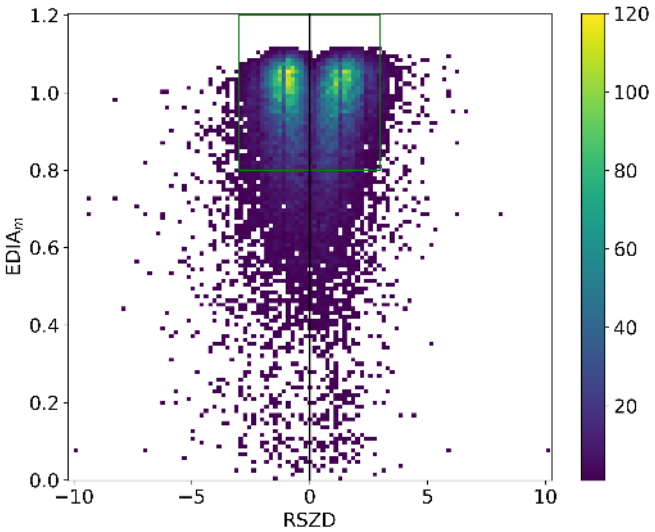


Figure 2.10: Comparison of normed residual B factor with EDIA$_m$ over the PDB displayed as deviating percentage per structure.



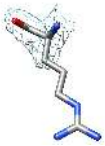
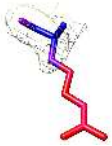
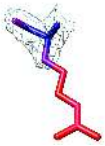
(a) Correlation of $EDIA_m$ with the residual RSCC calculated by Mapman. Pearson correlation coefficient: 0.68



(b) Comparison of $EDIA_m$ with RSZD. Data points in the green box show agreement between both measures. Examples are shown in Figure 2.13.

Figure 2.11: $EDIA_m$ compared with RSZD and RSCC over 8263 binding pocket residues of the Iridium HT set.

a

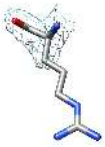
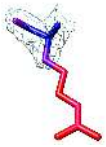




Arginine 191 A
1u4d

fc map at 1σ
Occupancy of six atoms = 0.01

$EDIA_m$: 0.08
$RSCC_M$: 0.917

b

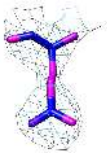



CE Methionine 124 A
1hq2

EDIA: 0.77
RSCC: 0.43

Figure 2.12: Two examples are shown to depict the difference between the RSCC computed by Mapman ($RSCC_M$) and $EDIA_m$ as well as the atomic RSCC and EDIA. Residues are colored in element and EDIA colors. The $2fo - fc$ map is shown at 1σ in blue.

a

Glutamate 241 I (1ml1)
$RSZD_{bb}$: -0.3, +0.3
$RSZO_{bb}$: 5.0

$EDIA_m$: 0.74 | 0.77
$RSZD_s$: -0.8, +0.4
$RSZO_s$: 4.1

b


Leucine 42 A (1d3h)
$RSZD_{bb}$: -0.2, +0.5
$RSZO_{bb}$: 1.8

$EDIA_m$: 0.96 | 0.25
$RSZD_s$: -1.2, +1.4
$RSZO_s$: 0.1

Figure 2.13: To depict the difference between RSZD, RSZO and EDIA scores, two examples are shown. Scores are divided into backbone (bb) and side chain (s) scores. The $2fo - fc$ map is visualized with a contour level of 1σ. The $fo - fc$ map is shown above 3σ in green and below -3σ in red.

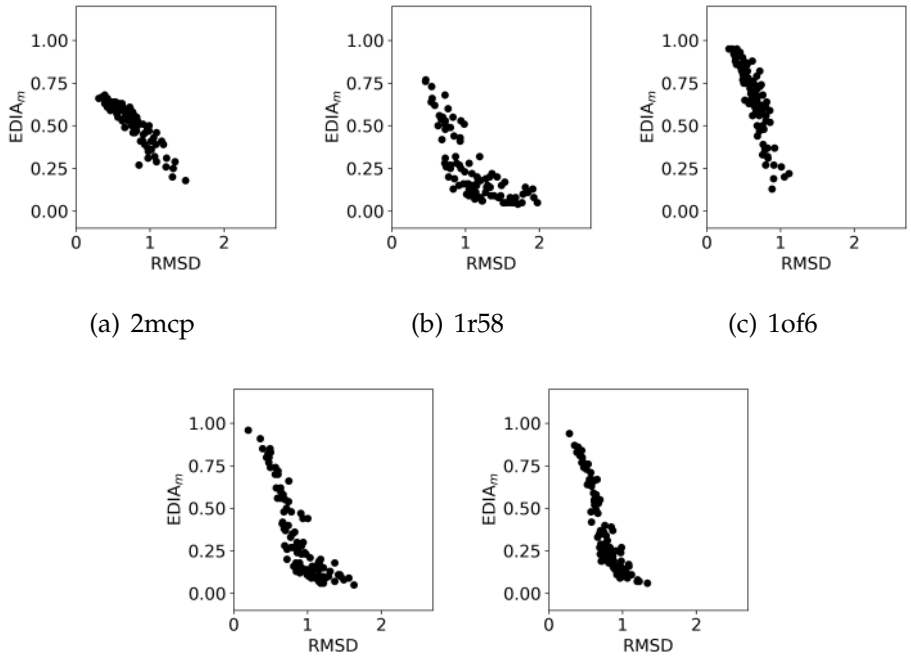
PDB ID	# Data points	$r_{X,Y}$	Maximum RMSD
2br1	100	-0.86	1.63064
1of6	100	-0.89	1.11414
1fh9	100	-0.91	1.33619
1r58	97	-0.76	1.97405
2mcp	100	-0.91	1.47982

Table 2.3: Results of the spatial displacement analysis visualized in Figure 2.14.

generally incorporates the idea of a multi conformation solution of the experimental electron density. On the other hand, RMSD with its value of zero for identical coordinates and its increasingly positive value for increasingly deviating coordinates does not correlate well with the electron density metrics RSCC and RSR [22]. Since $EDIA_m$ should further assists in the validation of methods for molecular modeling, a certain degree of correlation of RMSD should be verifiable. As already shown in the original publication for five complexes from the Iridium HT data set, $EDIA_m$ shows a correlation of maximally -0.93 in the original publication over at least 1764 sampled conformers for the first ligand of Mc/Pc603 Fab-Phosphocholine Complex (2mcp[46]), Methionine Aminopeptidase 2 (1r58[71]), Phosphate Synthase (1of6[31]), Protein Kinase CHK1 (2br1[15]) and Beta-Xylanase (1fh9[45]) from the Iridium HT data set. After the software update, the analysis was repeated on a randomly sampled set of up to 100 conformers per structure. The results underline the findings explained in the publication. (Table 2.3) $EDIA_m$ plotted against RMSD shows a sigmoid shape with the first plateau stretching from 0.0 to 0.4 RMSD dropping to the second plateau around 1.5 Å RMSD (Figure 2.14). The findings underline the ablity of $EDIA_m$ to increase the resolution of spatial deviations in the interval $[0, 0.5]$ Å RMSD. An example to highlight differences in $EDIA_m$ while having the same RMSD can be found in the original publication as well as in Figure 5.9 in this thesis.

2.3 Applications

EDIA as electron density scoring scheme has shown its use in various application scenarios. In its version of 2017 it was used to identify ligands without conspicuous electron density for the Platinum data set. It also served as an additional quality criteria in the NaomiNova software published by Inhester *et al.*[23]. It was used in quality control when creating a data set for mutation analysis[27] as well as controlling the quality of fragments in the PDB [13]. EDIA was used inhouse to



(a) 2mcp (b) 1r58 (c) 1of6

(d) 2br1 (e) 1fh9

Figure 2.14: Correlation of $EDIA_m$ with RMSD over the ligands in five pockets of the Iridium HT.

check the quality of metals and is integrated in an automatic data set assemblage tool `StructureProfiler` that is introduced in the next chapter. It is integrated into an Naomi based GUI tool called `HydeDebugGUI` (see Section A.2.3). The tool colors the selected protein-ligand pocket or the backbone over the whole protein to allow easy identification of conspicuous areas in a structure. As of now, EDIA is cited by at least 18 publications. They can be found for example in the journals *Acta Crystallographica Section D Structural Biology*[81], *PLOS Computational Biology*[69], *Journal of Medicinal Chemistry*[13], *Journal of Chemical Theory and Computation*[87], and *Proceedings of the National Academy of Sciences*[38].

2.4 Conclusion

EDIA is a measure to assist in identifying structures with conspicuous density. The metric is easy to compute as a weighted sum and thus easy to understand. The coloring scheme allows the visualization of parts to be inspected further in the structure to assist the user in focusing his or her attention on that specific part. EDIA is not limited to pure crystallized pockets but can be used on e.g. docking

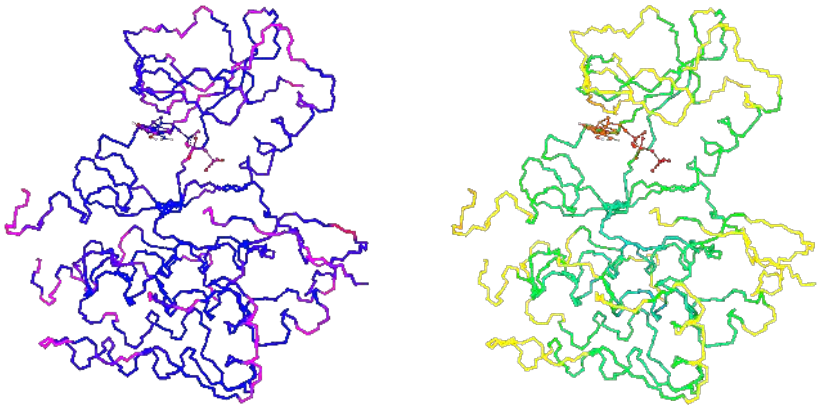


Figure 2.15: 1k3a with $EDIA_m$ and B factor backbone coloring in the `HydeDebugGUI`

poses as well since it does not incorporate B factor and occupancy. Hence, induced effects are not present in the new metric. It has started to be in use in numerous application scenarios on the one hand to identify high quality ligands, to reassess kinase families and on the other hand assist in validating algorithms for geometrical optimization in pockets. The following chapter focuses on using $EDIA_m$ among other quality metrics to determine a high quality data set. Subsequently $EDIA_m$ is employed on the results of GeoHYDE to understand the extend of deviation in the pockets of the newly found data set.

Chapter 3

Data Sets

Data sets for validating methods in structure based drug design currently in use consist of maximally 285 pockets (Table 3.1). The low number makes it difficult to create training and pure test data sets. In each case, the assemblage and quality check strategy was published in writing. A configurable tool chain, easy to set up, for objective structure selection was never published. In the following, the tool `StructureProfiler` is presented to solve that bottle neck. It was used to derive the subsequently introduced ProtFlex18 data set relevant for the validation of GeoHYDE in this thesis. The chapter discusses the similarity of ligands in the ProtFlex18 data set. It closes with applying SIENA, the ensemble analysis tool on ProtFlex18 and gives an overview of the found structure clusters and their flexible residues.

3.1 StructureProfiler: A Tool for Automatic High Quality Benchmark Data Set Assemblage

The `StructureProfiler` is an integrated tool based on NAOMI with seven complex tests, eight test for the active site and 21 tests to profile a ligand.[36] An active site is defined as the area around the ligand up to 8 Ångstrom distance including possible metals, waters and cofactors. The active site is prepared with Protoss before any test is run. Test parameters are configurable and three presets are available to profile structures as closely as possible to the Astex, Iridium or Platinum quality set. The `StructureProfiler` is also integrated in our ProteinsPlus server.

3.1.1 Validation

The tool has been tested on each given data set to control against deviations. Single cases for all three data sets are discussed in the Supporting Information of Meyder *et al.*[36]. Noteworthy on the one hand are ligands in Iridium HT with partially low electron density. They make the Iridium HT not suitable for benchmarking the quality of GeoHYDE. On the other hand, we detected a large list of ligands in the Platinum data set that do not have an $EDIA_m$ of at least 0.8 due to updated electron density maps. Software updates in the electron density refinement tool chain from 2009 to 2017 had modified the maps in a significant amount. This underlines the necessity for easy to use tools for benchmark data set creation to allow regular updates.

3.2 Data Set ProtFlex18

The tool `StructureProfiler` was used in the data set work flow A.1 with the tests listed in Tables B.22-B.23. 2386 ligands in 1598 of initially 63,889 PDB structures passed the 31 active tests. Of those, 1116 ligands are unique (Table B.24) based on stereoisomeric unique SMARTS comparison. The overlap between ProtFlex18 and other validation data sets consists of maximally 28 structures (Table 3.2). Hence, the hereby published data set offers a large, not yet used data set of inconspicuous pockets.

In the following, structures and ligands of the data set are analyzed based on fundamental properties and similarity. Figure 3.1 gives an overview over the properties of the 2386 ligands. The molecular weight stretches from 132 to 596 u with at least ten to 42 heavy atoms. The aLogP computed with NAOMI ranges from -7.5 to 13. The median ligand has two rings, two rotatable bonds, four hydrogen acceptors, and two donors. Oxygens are nearly twice as many present as nitrogens per ligand. Halogenes, phosphor and sulfur are also present in at least 279 ligands. The analysis based on stereo isomer aware unique SMILES identified 1116 unique ligands. Table B.24 shows the 32 ligands present in at least five differing structures with respectively three example PDB ids given. Additionally, all metals in the active sites present in the data set were accumulated (Table 3.2). Occurrence ranged from the most frequent metal magnesium with 226 hits in contrast to vanadium with just one occurrence.

Figure B.32 shows the results over all 1559192 ligands from the PDB on February 5th, 2020 known as LigandExpo for comparison. 32672 unique SMILES have been

Data set	# PDB ids
ProtFlex18	1598
Small Series	263
CASF-2016	285
CASF-2013	197
FEP	21

Table 3.1: Number of PDB ids per validation data set.

Data set	Data set	Overlap
ProtFlex18	CASF-2016	28
	CASF-2013	23
	Small Series	26
	FEP	0
Small Series	CASF-2016	11
	CASF-2013	10
	FEP	8
CASF-2013	CASF-2016	107

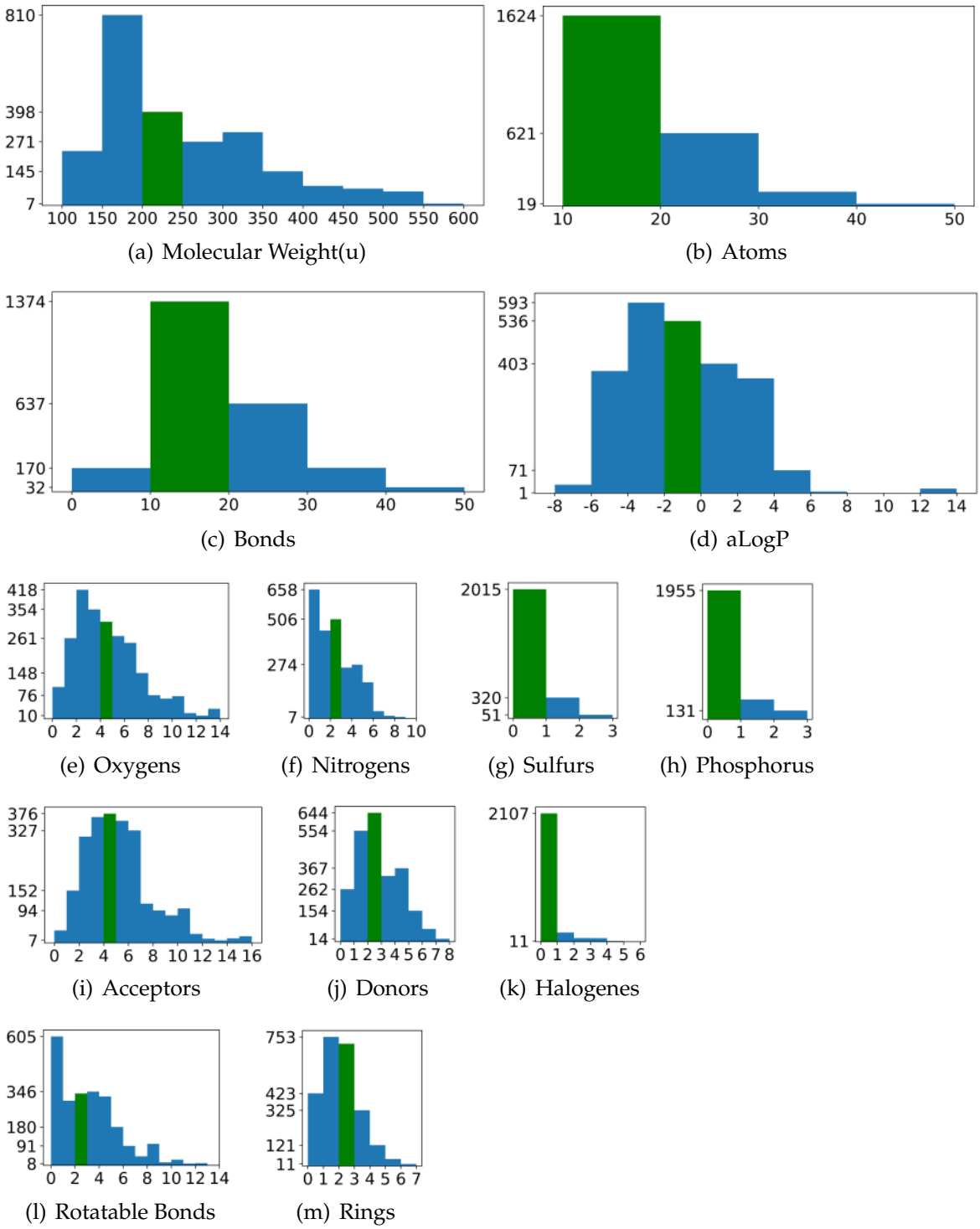
Table 3.2: PDB ID overlap between data sets.

detected. The molecules support the value distributions found in ProtFlex18 with a median molecular weight of 65 u ranging from 1 to 2975 u overall. A median molecule in the LigandExpo has four atoms, three bonds, an aLogP of -0.39, one acceptor, and one oxygen.

3.2.1 Enzyme Clustering with SIENA

To allow the validation of GeoHYDE's protein flexibility mode it was necessary to identify the number of flexible residues in the pocket. SIENA as part of the NAOMI tool suite computes an alignment between the query pocket and its database to identify structurally similar pockets.[4] SIENA allows to use e.g. a user defined cutoff for the maximally deviating backbone RMSD to identify structural similarity as one structure filter. Through complete linkage clustering, aligned residues with differing conformations are identified. The resulting SQLite database can then be used in further validation scenarios including flexible residues.

For the analysis of the ProtFlex18 data set, each of its 2386 pockets was used as query to identify similar pockets in the initial aforementioned 63,889 PDB structures as the database. The first screening with a backbone RMSD of 0.1 Å as similarity cutoff revealed three structure clusters for both the carbonic anhydrase 2 and the transcription attenuation protein MTRB. A further SIENA screen with a backbone RMSD cutoff of 1 Å merged those and revealed an RMSD cutoff of 0.4 Å to be the maximum cutoff to unite both enzyme clusters and the cutoff to gather the most closely related binding pockets into ensembles (Figure 3.3(b)). This resulted in orotidine-5'-phosphate decarboxylate (OMPDC, part of pyrimidine biosynthesis) having two clusters and the heat shock protein 90-alpha with three clusters. Both enzymes are known for a flexible binding site. OMPDC has two distinct binding



(a) Molecular Weight(u) (b) Atoms

(c) Bonds (d) aLogP

(e) Oxygens (f) Nitrogens (g) Sulfurs (h) Phosphorus

(i) Acceptors (j) Donors (k) Halogenes

(l) Rotatable Bonds (m) Rings

Figure 3.1: ProtFlex18 Ligand properties. In all plots, the number of e.g. oxygens per ligand is given on the y-axis. The bin including the median value is colored in green.

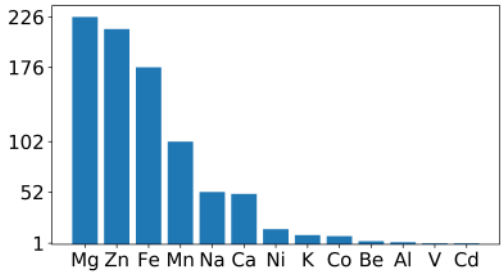


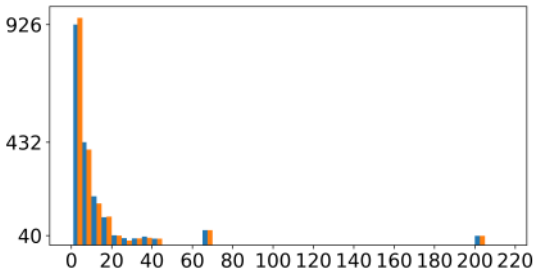
Figure 3.2: Metals in the ProtFlex18 data set. 867 metals in the active site of 797 complexes were found.

modes (Figure 3.4)[12], Hsp 90-alpha's clusters show a first RMSD peak of up to 0.4 Å and a second peak in the interval 0.6 to 1.0 Å RMSD (Figure 3.3(d)). Since the backbone positions deviate too far in these enzymes, the backbone RMSD cutoff of 1.0 Å was rejected. Thus, the final SIENA run was conducted with a backbone RMSD cutoff of 0.4 Å.

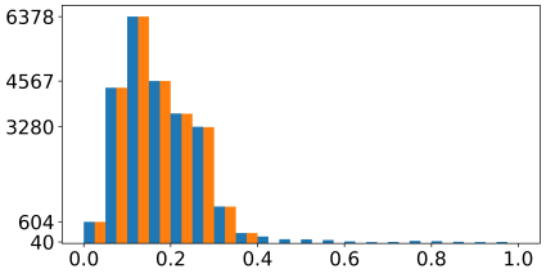
425 ensembles were detected in total. Ensemble sizes range from one to 204 structures (Figure 3.3(a)). If limited to only members of the ProtFlex18 data set, the largest ensemble consists of the carbonic anhydrase 2 and includes 67 high quality pockets given by 66 structures with 60 unique ligands. The largest ensembles sorted by the number of structures from the ProtFlex18 data set are given in Table 3.3. The ensembles are annotated with enzyme classification numbers if relevant and colored in green when well-known to be pharmaceutically relevant. The number of PDB ids from ProtFlex18, the number of their pockets and the number of unique ligands present in the ensemble complete each entry. If flexible residues were detected in the ensembles, the number ranges from one to 18 with a median of two flexible residue in the pocket (Figure 3.3(c)). Overall, 80 ensembles report flexible residues. Example structures are listed in Table B.25.

3.3 Conclusion

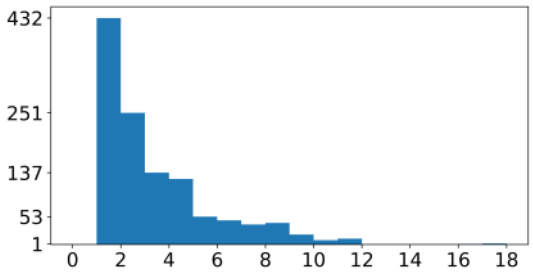
`StructureProfiler` is a new, configurable tool accompanied with a Python framework to easily identify inconspicuous pockets according to the selected tests. With the *GH* filter criteria set, 2386 pockets in the PDB of August 2018 were identified to be of high quality to be used in the validation of GeoHYDE. The pockets come from 1598 structures with 1116 unique ligands and 80 structure ensembles with flexible residues. The data set with the name ProtFlex18 shows hardly any overlap with existing validation data sets so that it can be used in combination with them



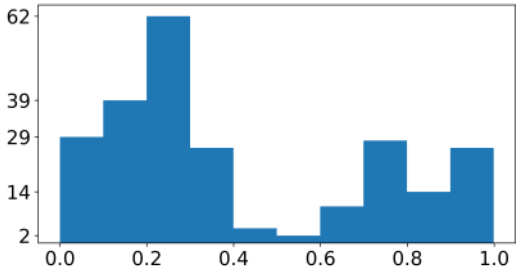
(a) Ensemble sizes



(b) RMSD distribution

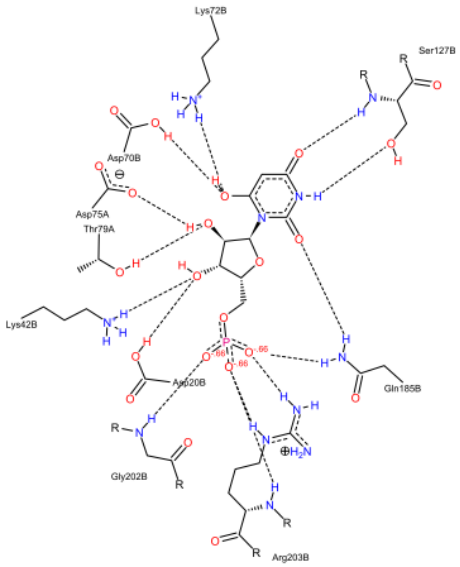


(c) Number of flexible residues (backbone RMSD 0.4Å)

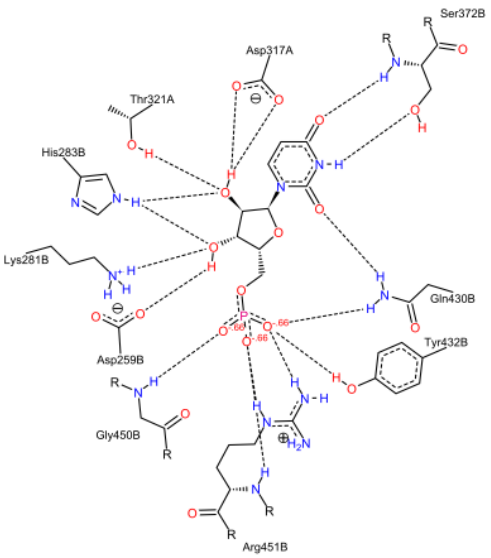


(d) Backbone RMSD distribution of the three hsp 90-alpha clusters (backbone RMSD cutoff 1.0 Å)

Figure 3.3: Statistics over the SIENA clusters with differing backbone RMSD cutoffs (bb RMSD). a,b) blue: bb RMSD of 1.0, orange: bb RMSD of 0.4 Å. Ensembles are created in using the 2386 ligands as starting query for SIENA.



(a) OMPDC complexed with BMP B 229 (3p5z)



(b) OMPDC complexed with U5P B 1479 (2v30)

Figure 3.4: The two binding modes of orotidine-5′-phosphate decarboxylate (OMPDC), resulting in two not mergeable SIENA ensembles.

Enzyme name	EC numbers	# PDB Ids	# Ligands	# Pockets
carbonic anhydrase 2	4.2.1.1	66	60	67
nitric-oxide synthase	1.14.13.39	22	16	34
nicotinamide phosphoribosyltransferase	2.4.2.12	20	18	33
glycogen phosphorylase, muscle form	2.4.1.1	19	18	20
orotidine 5′-phosphate decarboxylase	4.1.1.23	17	5	26
alpha-mannosidase 2	3.2.1.114	16	15	16
thrombin heavy chain	3.4.21.5	15	15	15
tankyrase-2	2.4.2.30	14	14	21
epsp synthase	2.5.1.19	11	4	14
transcription attenuation protein mtrb		11	1	39
endothiapepsin	3.4.23.22	10	9	10
heat shock protein hsp 90-alpha		10	10	11
7,8-dihydro-8-oxoguanine triphosphatase	3.6.1.55 3.6.1.56	8	8	8
transcriptional regulatory repressor protein (tetr-family)		8	8	10
4-hydroxy-3-methylbut-2-enyl diphosphate reductase	1.17.1.2	8	7	11
pteridine reductase 1	1.5.1.33	8	8	14
isopenicillin n synthetase	1.21.3.1	8	8	8
trna (guanine-n(1)-)-methyltransferase	2.1.1.228	7	7	7
cytochrome p450		7	5	11
bromodomain-containing protein 4		7	7	8
glutamate receptor 2		7	3	7
cgmp-dependent 3′,5′-cyclic phosphodiesterase	3.1.4.17	7	7	16
camp-dependent protein kinase catalytic subunit alpha	2.7.11.11	7	7	7
heat shock protein hsp 90-alpha		7	7	7
heat shock protein hsp 90-alpha		7	7	7
orotidine-5′-phosphate decarboxylase dehydrogenase [quinone]	4.1.1.23 1.10.5.1 1.6.99.2	7	6	12
beta-glucosidase a	3.2.1.21	7	7	8
ribosyldihydronicotinamide	1.10.99.2	7	6	9
pantothenate synthetase	6.3.2.1	7	7	7
gamma-enolase	4.2.1.11	6	4	7
dihydroorotase	3.5.2.3	6	3	7
methionine aminopeptidase	3.4.11.18	6	6	6
beta-galactosidase	3.2.1.23	6	2	18

Enzyme name	EC number/s # Ensembles	# PDB Ids	# Ligands	# Pockets
poly [adp-ribose] polymerase 3	2.4.2.30	5	5	5
serine/threonine-protein kinase pim-1	2.7.11.1	5	5	5
anthranilate phosphoribosyltransferase	2.4.2.18	5	5	8
xanthine dehydrogenase/oxidase	1.17.3.2 1.17.1.4	5	4	8
phosphoglycerate kinase 1	2.7.2.3	5	1	5
camp-specific 3′,5′-cyclic phosphodiesterase 4d	3.1.4.53	5	5	6
carbonic anhydrase 12	4.2.1.1	5	5	7
thermolysin	3.4.24.27	5	5	5
neuraminidase	3.2.1.18	5	5	5
liver alcohol dehydrogenase	1.1.1.1	5	3	6
(4)	14	56		78
(3)	36	108		129
(2)	118	236		355
(1)	214	214		564

Table 3.3: The above stated ensembles were detected by SIENA with a maximum backbone RMSD deviation of 0.4 Å. The clusters are listed with their most frequent enzyme name extracted from the PDB and their EC number. All ensembles with at least five different PDB structures present in the dataset of 2386 pockets together with the number of unique ligands and the total number of aligned pockets in the data set structures are presented. The last four entries list the number of ensembles with only four to two pockets and four to one unique PDB ids.

without adding any bias. Hence, ProtFlex18 will be the data set used for training and evaluating GeoHYDE's performance on crystal structures.

Chapter 4

Torsion Angles

The in the next chapter evaluated objective function GeoHYDE for geometrically optimizing a protein-ligand complex allows changes of torsion angles. As GeoHYDE in its 2012 version used an unspecified torsion term, it was decided to harness the knowledge from the in the group developed Torsion Library for better assessing the likeliness of the respective torsion angle. But multiple problems have motivated us to revisit the TorLib13. Longterm evaluation has shown torsion angles marked as unlikely even though they deviated only slightly from highly likely torsion angle. Additionally, when comparing peaks in torsion rules for which substructures differ only by one proton, diverging peaks have been found. Finally, a continuous torsion potential was needed to assist in scoring. In a multi step approach resulting in the creation of the TorLib16 and the TorLib18, we have addressed all issues. Methods, results and conclusions are given in the following sections.

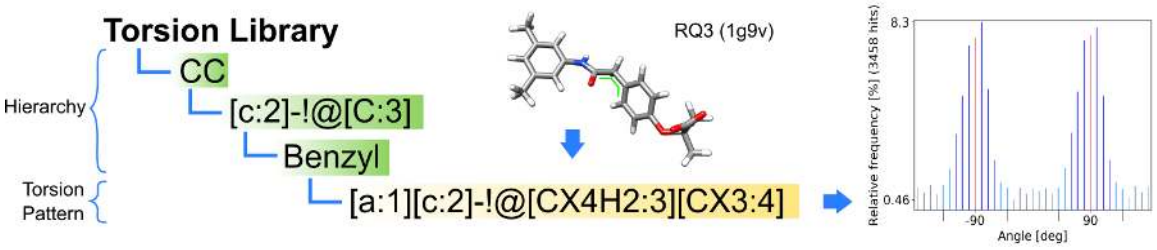


Figure 4.1: The structure of the torsion library.

4.1 Torsion Library Updates

4.1.1 Structure of the Torsion Library

The torsion library created by Schärfer *et al.* consists of a hierarchical collection of hand crafted SMARTS patterns. Each pattern describes the chemical environment around an acyclic bond in necessary detail to avoid inconsistent results in accordance to the assessment of the experts. For rapid matching, the patterns are sorted into six specific and one generic class. Class categorization is done by the elements present at the acyclic bond to be evaluated (CC, CN, CO, CS, NS, SS, Figure 4.1). Most of the classes hold a number of so called sub hierarchy torsion patterns to further bracket types of molecular environments together such as the example 'Benzyl' in Figure 4.1. If none of the handcrafted patterns in the specific classes match, a fitting pattern is searched in the generic class GG. These patterns aim to cover the whole chemically possible space with an acyclic bond in between as a failsafe.
Highly likely torsion angles per torsion pattern in the CSD were subsequently automatically identified. These peak candidates were then analyzed by experts and if confirmed annotated with two sets of tolerances. In rare cases, purely manual peaks not supported by experimental data were set. Tolerance borders were initially automatically identified with the first tolerance interval set to stretch symmetrically over all bins next to the peak with at least 2.5 % of overall hits. Second tolerance interval were set to stretch over bins with at least 1.5 % of overall hits. Torsion angles in between the first tolerance interval are treated as highly likely. They are understood to be less likely if they fall into the second tolerance interval around the peak. A torsion angle outside of any tolerance interval is described as unlikely in the subsequent chapter. The likeliness is a guideline towards seeing the torsion angle in a crystallized structure, not an absolute decision about its existence. Protein binding and interaction with the solvent content can result in effects not accounted for in the torsion rules. The likeliness is displayed in the TorsionAnalyzer as the bond colored in green if highly likely, orange if less likely and red if the torsion angle is statistically found to be unlikely.

4.1.2 Datasets

As first validation set, the original CSD subset from 2013 was used as published in Schärfer *et al.*[62] but with one change. JARNAR was removed due to its unlikely

conformation (Figure 4.2). The data set is named CSD13 from now on.
In 2018, we accessed the CSD through its Python API to retrieve an updated set of molecules. In implementing a `CustomSearch` class, entries with 3D coordinates, no errors and no disorder with at least one carbon atom and a maximum R factor of 10% in accordance to Schärfer *et al.*[62] were kept. Ions, metals, power structures, organometallic or polymeric compounds were filtered out. Subsequently, we controlled in accordance to Schärfer *et al.*[62] to only use molecules with the elements H, C, N, O, F, Cl, Br, I, S, and P with the help of our NAOMI tool suite and excluded JARNAR[19]. Thus in contrast to the original publication in 2013, the data set increased by 56% to 212,250 molecules, called CSD18, while following the original filtering strategy.
We also examined the performance of the torsion lib on ligands resolved with X-ray crystallography and deposited electron density at the PDBe (August 10, 2018). Initial analysis of the performance on all PDB ligands had shown many conspicuous ligands not backed by experimental data. Thus, not relevant ligands following the combined `StructureProfiler` criteria set with an $EDIA_m$ below 0.8, an R factor above 0.4 and a resolution larger than 2.5 Å were removed. Of the initial 115627 complexes with electron density in the PDBe, at least one ligand in 25915 complexes passed. Multiple molecules per complex with an identical name, chain id and infile id were detected describing e.g. parts of organometallic compounds. Those were removed to stay close to the filtering criteria of the CSD. The subsequently derived 49.204 ligands were then filtered with the list of only allowed elements and resulted in 48.873 molecules called PDB18 subsequently.

4.1.3 Torsion Library Validation Strategy

TorLib13, TorLib16 and TorLib18 were validated on the aforementioned CSD and PDB data sets fitting to their release time. TorLib13 and TorLib16 was validated on the CSD13 by Wolfgang Guba with the help of the `torsionchecker` in 2015. The `torsionchecker` as a command line tool analyses the torsion bonds of a given set of molecules into the group pf likely, less likely and unlikely torsion angles. This tool as well as the `TorsionAnalyzer`[62] from 2013 have known problems in speed, code quality, and the SMARTS matching algorithm. This made the creation of the so called `TorsionPatternMiner` in 2018 necessary (see SI A.2.2). It

Figure 4.2: JARNAR is excluded from the CSD set due to its unlikely conformation.

Rotatable Bond Definition	2013[19, 62]	2018
Bond is a single bond	x	x
Bond is not ring bond	x	x
Bond is not delocalized	x	x
Bond is not part of a nitrile.	x	x
No atom in bond has linear geometry.		x
No atom in bond is a terminal atom	x	x
No atom in bond is a heavy atom, connected to only hydrogens.	x	x
Bond does not connect to SF_3		x
Bond does not connect to CF_3	x	x

Table 4.1: Definition of a rotatable bond in the 2013 and 2018 implementation. All conditions need to be fulfilled by a bond to be rotatable.

combines the ability to mine large data sets of the `torsionchecker` with the automatic statistics generation ability of the `TorsionAnalyzer` in a command line tool with a state of the art smart matching which has been extensively tested. In contrast to 2013, we have extended the definition of non-rotatable bonds to include bond atoms with linear geometry as well as bonds connected to -SF_3. A full overview of currently rotatable bonds can be found in Table 4.1. The `TorsionPatternMiner` also allows the exclusion of bonds to any terminal heavy atom as well as limiting the creation of all statistics to only single bonds if necessary.
According to the publications, torsion library statistics are generated in adding any torsion angle to the statistics of a torsion rule if its SMARTS matches. All 511 torsion rules are evaluated per bond. Each torsion angle peak receives an updated score at the end of the analysis. In activating the selective matching in the `TorsionPatternMiner`, the SMARTS matches for the validation according to Guba *et al.* [19] are computed. In this case, matches are only tested on rotatable bonds and only the most specific match is reported. The likeliness of each torsion angle is then accumulated and the relative percentage of all unlikely but observed torsion angles in regards to all observed torsion angles over the data set is computed. The relative percentage of unlikely torsion angles per torsion rule is then plotted sorted by its absolute statistical occurrence. If above 40%, it is colored red, above 20% colored orange and else colored green in the validation plot. With the help of the Intel Threading Building Blocks[24] the `TorsionPatternMiner` can calculate CSD statistics in three hours on an eight core cluster node with 63 GB RAM and openSUSE Leap 42.2.

4.1.4 Analysis of SMARTS

The TorLib16 was controlled against manual errors when creating SMARTS patterns as well as when integrating them into the torsion rule hierarchy. The recently developed SMARTScompare was pivotal for the analysis. The method is based on fingerprint generation and subsequent maximum common subgraph analysis to compare two SMARTS. For each node in the SMARTS expression, a list of possible atom types available in NAOMI is generated. These atom type lists can be pruned in considering the environment around each node. If a SMARTS recursion is present, the pruning is limited to the environment inside the recursive expression. When comparing two SMARTS to understand if SMARTS A is a subset of SMARTS B, atom type lists are compared. Atoms are matched on each other, if the atom type list of a node in A is a subset of the atom type list of the respective nodes in B. A matching to solve the maximum common subgraph problem is searched that matches all nodes from A to nodes from B.

SMARTS Modifications

Hierarchy and torsion rules were adjusted to allow sub set analysis. In a step wise approach, all hierarchy and child hierarchy SMARTS were rewritten to only match single, not ring bonds, expressed with −!@ in the SMARTS language. As second step, all 511 torsion rule SMARTS patterns were transformed to only match single bonds (−) besides the already declared non-ring bond. Hence, the subsequently applied SMARTScompare algorithm [63] was able to align the relevant rotatable bond between two SMARTS patterns.

Resorting through Subset Analysis

SMARTScompare was run multiple times. In the first step, torsion sub hierarchy patterns were sorted from specific to generic. Hence the top most sub hierarchy is more specific than any following sub hierarchy pattern in the same class. Afterwards, the SMARTS of each torsion rule was verified to be correctly sorted into its class hierarchy. In the third step, torsion rules in a higher level hierarchy were moved into the top most fitting lower level child hierarchy. Thus, if a pattern is positioned on level three and thus checked as the first possible pattern in the torsion angle matching process, it was moved into the fitting top most child hierarchy at level four to stay close to its related patterns if possible. The generic hierarchy is excluded from moving patterns into lower level child hierarchies since in this

class, the child hierarchies are evaluated first. As third step, all torsion rules were analyzed so that SMARTS at the bottom of each hierarchy are subsets of SMARTS at the top of the hierarchy to guarantee the ordering from specific to generic per sub hierarchy. If a pattern is more specific than any of its predecessors in the hierarchy, it is sorted in front of the highest more generic predecessor. The analysis can also detect duplicates. The resulting reordered torsion library will be called TorLib18.

Validation of the Sorting Strategy

We identified the minimally invasive resorting strategy in analyzing the movement and the change in angle likeliness of rotatable bonds. Per insertion strategy, the likeliness of the most specific torsion rule per rotatable bond for the torsion library prior and past sorting were computed as validation strategy. Each bond is uniquely identifiable with the help of the atom ids participating in the rotatable bond labeled with '3' and '4' combined with the torsion angle measured over the four labeled atoms in the torsion rule.

4.1.5 Results

To formalize an performance base line, the TorLib13 was evaluated on the CSD13 and the CSD18 in accordance to Guba *et al.*. Subsequently manual and due to using SMARTScompare necessary changes are documented in SMARTS describing sub hierarchies and torsion rules. They are followed by the results of the automatic subset determination and reordering with SMARTScompare on the CSD18. The performance on the newly created TorLib18 is discussed subsequently. As final part, an outlook on necessary future work is given.

Validation of the TorLib13 on the CSD13 and CSD18

The likeliness of the most specific torsion rule per rotatable bond in each dataset was computed. The output file was then parsed to count the amount of unlikely rotatable bonds per torsion rule as the relative percentage over all matched bonds per torsion rule. These percentages are then plotted against the absolute number of matchings in validation mode. Torsion rules with more than 40% unlikely bonds, marked in red in Figure 4.4, have been manually analyzed (Table 4.4). As published in Guba *et al.*[19], the TorLib13 was controlled on the CSD13. Numerous torsion rules were revived. In the end, 112 torsion rules received updated tolerance intervals or updated peaks. In 54 cases, additional torsion rules were introduced and

in 24 cases, the environment was refined. Additionally, peak and tolerance overlaps were automatically removed and environment descriptions were transformed to recursive SMARTS for technical reasons. A slight overestimation of the [175°, 185°] interval was also corrected in the `torsionchecker`. In total, the number of torsion angles flagged as unlikely dropped from 40,453 to 10,678 and no torsion rules with more than 40% unlikely torsion angles in the evaluation scheme were reported in the thus published TorLib16. Figure 4.4 a) reproduces the validation scenario on the TorLib16 with the newly written `TorsionPatternMiner`. In contrast, two torsion rules with more than 40% unlikely torsion angles were detected. In both cases, single case examination revealed an error in the smarts matching algorithm in the old source code. Thus, they have not been evaluated in 2016 in their current state. Figure 4.4 b) then shows the evaluation of the TorLib16 on the CSD18. Three torsion rules are marked in red. While again one of them was subject to the known bug in the old smarts matching algorithm, the other two rules were always correctly matched. [a : 1][c : 2]−!@[O : 3][CX3H0 : 4] has a well-filled statistic with 5952 torsion angles but only one match beyond any peak in the validation scenario (Figure 4.6). Here, structures with small end groups have clouded the statistic, even though they match to more specific patterns in the validation. [cH0 : 1][c : 2]([cH0])−!@[O : 3][!#1 : 4] is also problematic in the evaluation. The bond in question is shown in Figure 4.7. Due to a sterical hindrance, we see the torsion angle as unusual but correct. The difference in unlikely torsion bond per torsion rules is additionally shown in Figure 4.5. When comparing the number of unlikely, red flagged torsion angles on the CSD13 and CSD18 while scanning the molecules with the TorLib16, 268 of the overall matched 395 torsion rules report changes. Seven torsion rules report their first matches on the CSD18 but do not report any matches on the CSD13. In contrast, when comparing the number of red flagged torsion angles against their absolute value of the TorLib16 against the TorLib18 on the CSD18, only 31 torsion rules report a change. Additionally 41 torsion rules in the TorLib18 report being matched the first time after resorting the TorLib.

Manual SMARTS Corrections

Two torsion rules were detected to be corrected manually (Table 4.2). In the first case, the label '4' was wrongly used twice. The second pattern in the table allows two types of elements as first atom [O, S : 1]. Their sub hierarchy was splitted and the pattern duplicated with the first atom only describing one element each.

Changes in SMARTS Due to Subset Relations

Prior to the final reordering with SMARTScompare, 48 patterns were detected to be less specific than their parental hierarchy (see Table B.2). Missing specifications were added to allow the subset relations checks with SMARTScompare for all of them. For three patterns starting with [$([C](= O)([$([NX3H1]), $([NX3H2])]) [NX3H1]) : 1][NX3H1 : 2]−!@, SMARTScompare was not initially able to confirm the subset relation to its subset hierarchy. These patterns were extended with environment information about the nitrogen having a valence of three and not being part in any ring in the recursion ([NX3H1**!Rv3** : 1]). The new specification does not cause changes in the torsion rule statistics but allows SMARTScompare to work correctly. One sub hierarchy was rewritten to be more specific because five patterns in it started with an aromatic carbon while the other three started with any aromatic atom at the first position. Changes in the SMARTS pattern of each torsion rule result in a changed statistic extracted from the CSD while updated sub hierarchies do not affect any statistic. Subsequently, sub hierarchies were analyzed to be correctly ordered. Three sub hierarchies had to be reordered (Figure B.3). As next step, the torsion rules on a higher level were analyzed to fit into a sub hierarchy of a lower level (Table B.4). While 14 torsion rules were successfully moved into the sub hierarchy [CX4][CX3], two patterns had two choices. They were moved into the top most possible sub hierarchy to still allow a comparatively early matching check. 12 patterns are not a subset of any sub hierarchy listed in [C : 2][C : 3] and stayed at their place in the torsion library (Table B.5). Finally, the torsion rules in every sub hierarchy were resorted. 19 torsion rules had to be moved (Table B.6) and four duplicates were found (Table B.7).

In the end, two torsion rules still do not fit to their sub hierarchy SMARTS. The torsion rule [NH2] − [C : 1](= [NH2]) − [NH1 : 2]−!@[CH2 : 3] − [C : 4] in Guanidine II ([NH1 : 2]−!@[C : 3]([N, n]) [N, n]) overlaps but has the torsion rule on differing bonds. Due to its aliphatic carbon as part of the rotatable bond, it can not be moved to the related sub hierarchy Guanidine I . There, the carbon is expected to be aromatic. The torsion rule [cH1 : 1][c : 2]([cH1])−!@[CX3 : 3](−c) = [O : 4] in the sub hierarchy a(−[NH1, NH2, OH1])[c : 2]−!@[CX3 : 3] = O is not able to match its first carbon to any possibility given for it by the sub hierarchy pattern (see Figure 4.3).

Pattern (old)	Pattern (new)	Reason for Change
Path: CN ⇒ O = [C : 2]−!@[NX3 : 3]		
[O, S : 1] = [C : 2]([$([NX3H1]), $([NX3H2])])	[O : 1] = [C : 2] ~	wrong sub hierarchy
−!@[$([NX3]c[nH0]) : 3][H : 4]	[S : 1] = [C : 2] ~	wrong sub hierarchy
Path: CC ⇒ [c : 2][C : 3]		
[$([cH0](F)) : 1][c : 2]([cH1])	~ ([CX3]) = [O : 4]	4th label used twice
!@[CX3 : 3]([CX3 : 4]) = [O : 4]		

Table 4.2: Manual Corrections in SMARTS Pattern.

Validation of the Sorting Strategy

The torsion rules were resorted and then changes in the torsion angle likeliness analyzed. The least invasive resorting strategy was to insert the more specific pattern right above the in relative terms more generic pattern. Due to the reordering, torsion angles of specific bonds changed likeliness (Table 4.3). Overall, changes showed the movement to a more specific pattern. In seven cases, overall increase of angle likeliness was found. In eight cases, an overall decrease in angle likeliness was found. Such patterns need to be observed closely in the next section when validating the overall performance on the CSD18. We highlight the case of 187 bonds moving from the torsion rule [S : 1] = [C : 2]([$([NX3H1]), $([NX3H2])]) −!@[$([NX3](cn)) : 3][H : 4] to the strongly deviating torsion rule [!#1 : 1][CX3 : 2] (= S)−!@[NX3H1 : 3][!#1 : 4] for which 48 reported a likeliness increase and only three bonds a decrease. In this case, the parental sub hierarchies were detected to be in the wrong order (see Table B.3, second entry) and rearranged. The increase in angle likeliness justifies the move in our opinion.

Torsion Rule: Old	New	Bonds	Angel Likeliness	
			Increase	Decrease
[$([c]([NH1, NH2])) : 1][c : 2]	[$(c[NH1, NH2]) : 1][c : 2]	58	22	4
−!@[CX3 : 3]([!O]) = [O : 4]	−!@[CX3 : 3]([NX3H0]) = [O : 4]			
[$(a[OH1]) : 1][c : 2]	[$(c[OH1]) : 1][c : 2]	31	6	21
−!@[CX3 : 3]([NX3H0, CX4H0, c]) = [O : 4]	−!@[CX3 : 3]([NX3H0]) = [O : 4]			
[a : 1][c : 2]([a])−!@[O : 3][CX4H0 : 4]	[nX2H0 : 1][c : 2]([cH0])−!@[O : 3][CX4H0 : 4]	5	0	0
[∗ : 1] [N, n : 2]−!@[S : 3] [∗ : 4]	[∗ : 1] [NX2 : 2]−!@[SX4 : 3] [∗ : 4]	642	0	0
[∗ : 1] [N, n : 2]−!@[S : 3] [∗ : 4]	[∗ : 1] [NX2 : 2]−!@[SX3 : 3] [∗ : 4]	79	9	2
[∗ : 1] [N, n : 2]−!@[S : 3] [∗ : 4]	[∗ : 1] [NX2 : 2]−!@[SX2 : 3] [∗ : 4]	114	0	25
[$(C = O) : 1][NX3 : 2]−!@[c : 3][aH0 : 4]	[$(C = O) : 1][NX3H1 : 2]−!@[c : 3]([cH0])[cH : 4]	1703	526	312
[$(C = O) : 1][NX3 : 2]−!@[c : 3][aH0 : 4]	[$(C = O) : 1][NX3H1 : 2]−!@[c : 3]([cH0])[cH0 : 4]	431	96	39
[$(C = O) : 1][NX3 : 2]−!@[c : 3][aH0 : 4]	[$(C = O) : 1][NX3H0 : 2]−!@[c : 3]([cH0])[cH0 : 4]	381	0	1
[$(C = O) : 1][NX3 : 2]−!@[c : 3][aH0 : 4]	[$(C = O) : 1][NX3H0 : 2]−!@[c : 3]([cH0])[cH : 4]	549	4	46
[$([C](= O)) : 1][NX3H1 : 2]	[$([C](= O)) : 1][NX3H1 : 2]	533	0	162
−!@[$([c]([nH0, o])) : 3][cH1 : 4]	−!@[c : 3]([cH])[nX2H0 : 4]			
[O : 1] = [C : 2]([!$([NH1])]	[O : 1] = [C : 2]([CX4])	28	19	0
−!@[NX3H1 : 3]([H : 4])[$(c([nX2H0])([nX2H0]))]	−!@[$([NX3](c([nX2H0])([nX2H0]))) : 3][H : 4]			
[S : 1] = [C : 2]([$([NX3H1]), $([NX3H2])])	[!#1 : 1][CX3 : 2](= S)	187	48	3
−!@[$([NX3](cn)) : 3][H : 4]	−!@[NX3H1 : 3][!#1 : 4]			
[!#1 : 1][CX4 : 2]−!@[NX3; ”N_lp” : 3]	[!#1 : 1][CX4 : 2]−!@[NX3; ”N_lp” : 3][!#1 : 4]	1965	48	71
[cH1, nX2H0 : 1][c : 2]([cH1, nX2H0])	[cH1, nX2H0 : 1][c : 2]([cH1, nX2H0])	2287	105	268
−!@[NX3r : 3][∗ : 4]	−!@[NX3r : 3][CX4r : 4]			
[C : 1][$(S(= O) = O) : 2]−!@[”N_lp” : 3]	[C : 1][$(S(= O) = O) : 2]−!@[NX3H1 : 3][C : 4]	86	8	20
[c : 1][$(S(= O) = O) : 2]−!@[”N_lp” : 3]	[c : 1][$(S(= O) = O) : 2]−!@[NX3H1 : 3][C : 4]	794	86	79

Table 4.3: Bonds changing torsion rules after resorting. The table gives the number of bonds, that switch to a different torsion rule due to resorting the torsion library. The number of bonds with improved and dropping torsion angle likeliness is also given.

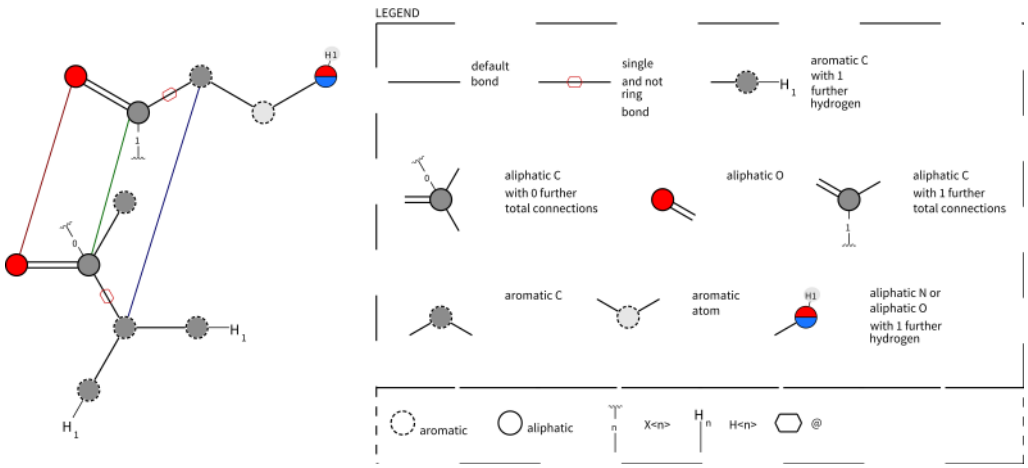
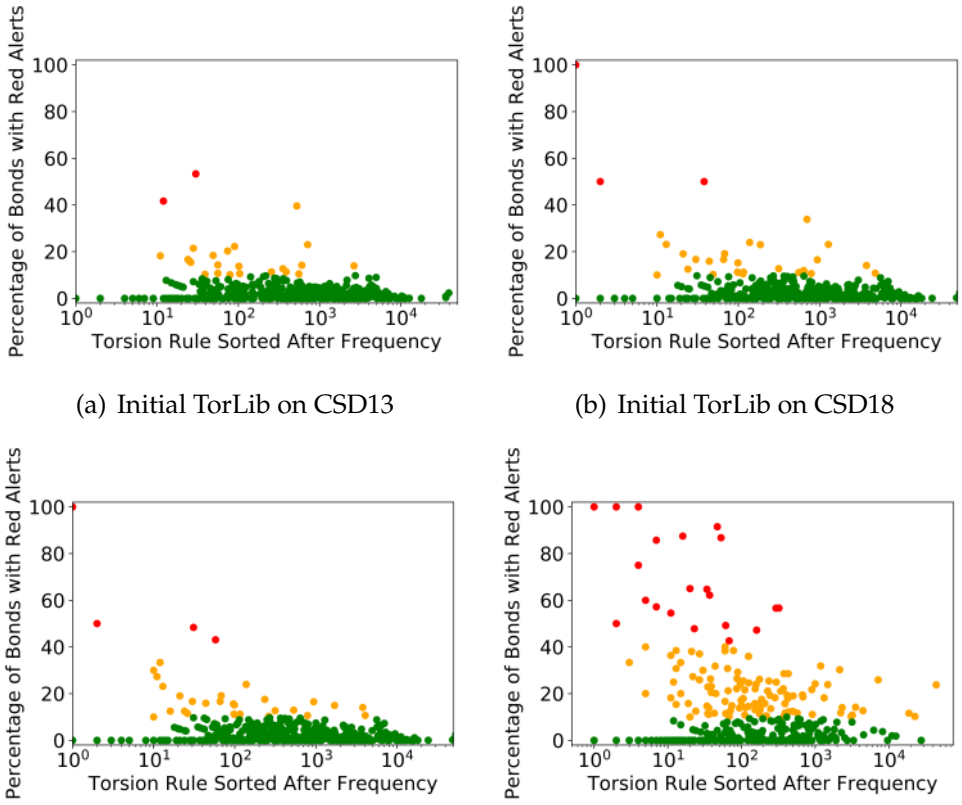


Figure 4.3: The torsion rule [cH1 : 1][c : 2]([cH1])−!@[CX3 : 3](−c) = [O : 4] (top) is not included in its sub hierarchy a(−[NH1, NH2, OH1])[c : 2]−!@[CX3 : 3] = O (bottom). The atom labeled as first atom in the pattern can not be included in the sub hierarchy SMARTS pattern.



(a) Initial TorLib on CSD13

(b) Initial TorLib on CSD18

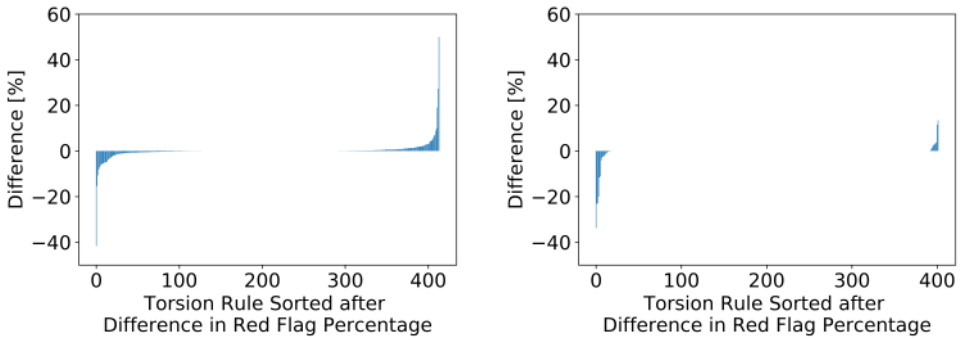
(c) Reordered TorLib on CSD18

(d) TorLib18 on high quality PDB ligands.

Figure 4.4: Torsion rule sorted by frequency in the respective data set versus percentage of red flags in it. Torsion rules with less than 10% red flags are colored in green, with less than 40% are colored in orange. Torsion rules above 40% are colored in red.

Validation of the TorLib18 on the CSD18

A final step, the resorted torsion library (TorLib18) was evaluated on the CSD18 (see Section 4.1.2). [cH0 : 1][c : 2]([cH0])−!@[O : 3][!#1 : 4] and [a : 1][c : 2]−!@[O : 3][CX3H0 : 4] are again problematic. The evaluation of the reordered TorLib18 on the CSD18 has besides the above mentioned two torsion rules ([cH0 : 1][c : 2]([cH0])−!@[O : 3][!#1 : 4] and [a : 1][c : 2]−!@[O : 3][CX3H0 : 4]) two additional rules marked in red. Both torsion rules describe an internal hydrogen bond but do not account for sterically restricted ligands with multiple rings or strongly aliphatic branched parts attached to the third atom in the torsion rule (see Figure 4.8, 4.9).



(a) Difference TorLib16 on CSD14 vs. CSD18

(b) Difference TorLib16 to TorLib18 on CSD18

Figure 4.5: Difference in red flags per observed torsion rule if present in both sets. 10 torsion rules were matched in validation mode with the TorLib13 on the CSD13 while seven torsion rules were only matched on the CSD18 with the same torsion library. 395 rules were matched in both sets. When updating and reordering the torsion library, again 395 torsion rules were matched by both torsion libraries. Additionally, 41 torsion rules were only matched when scanning the CSD18 with the TorLib18 (Table B.8).

Figure 4.4	Torsion rule SMARTS	Total Matches	Examples
a,	[NH2][C : 1](= [NH2])[NH : 2]!@[CH2 : 3][C : 4]	12	
a, b	[O : 1] = [C : 2]([!$([NH1])]) −!@[NX3H1 : 3]([H : 4])[$(c([nX2H0])([nX2H0]))]	30, 38	
b, c	[a : 1][c : 2]−!@[O : 3][CX3H0 : 4]	1	Fig. 4.6
b, c	[cH0 : 1][c : 2]([cH0])−!@[O : 3][!#1 : 4]	2	Fig. 4.7
c	[$(c[OH1]) : 1][c : 2] −!@[CX3 : 3]([NX3H0]) = [O : 4]	31	Fig. 4.8
c	[$(c[NH1, NH2]) : 1][c : 2] −!@[CX3 : 3]([NX3H0]) = [O : 4]	58	Fig. 4.9

Table 4.4: All torsion rules with more than 40% unlikely torsion angles in any of the three validation scenarios from Figure 4.4. Patterns tend to be problematic in multiple scenarios: a denotes the evaluation of the initial torsion library on the CSD13, b marks the performance of the initial torsion library on the CSD18 and c signifies the evaluation of the resorted TorLib18 on the CSD18.

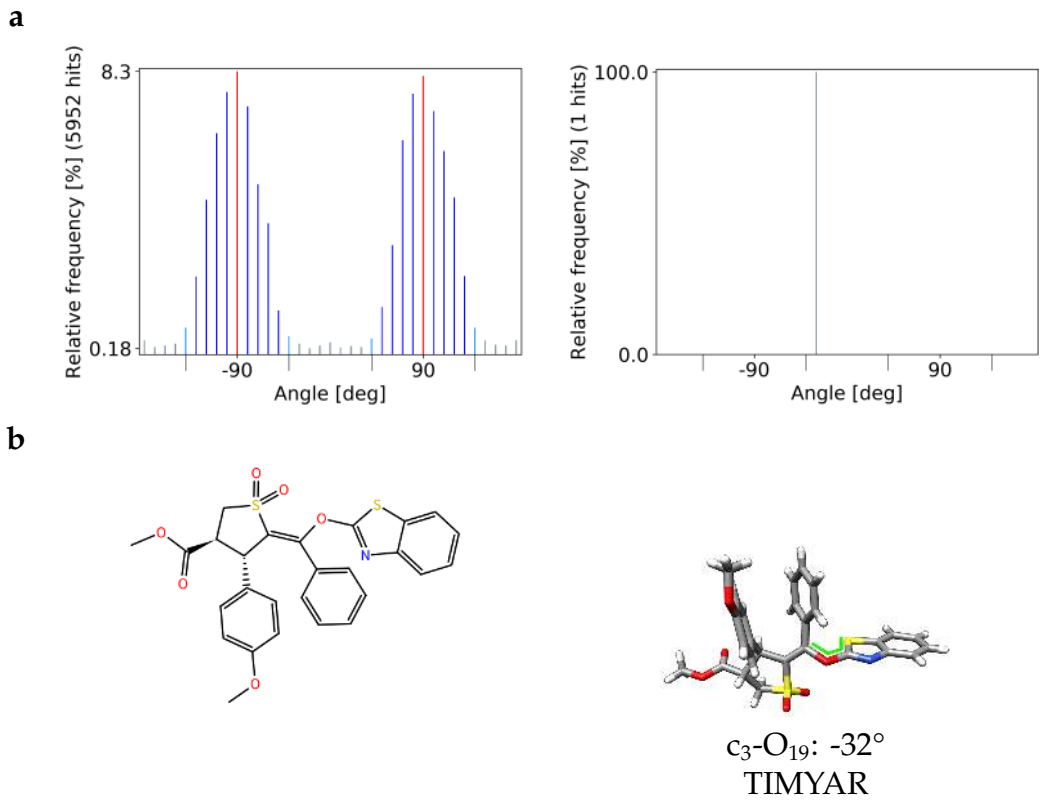


Figure 4.6: Outliers of the torsion rule [a : 1][c : 2]−!@[O : 3][CX3H0 : 4] from the TorLib 18 on the CSD18. While the statistic of the pattern is filled with 5052 hits, the validation shows only one matching bond in TIMYAR. The resulting torsion angle is outside of the second tolerance of any peak.

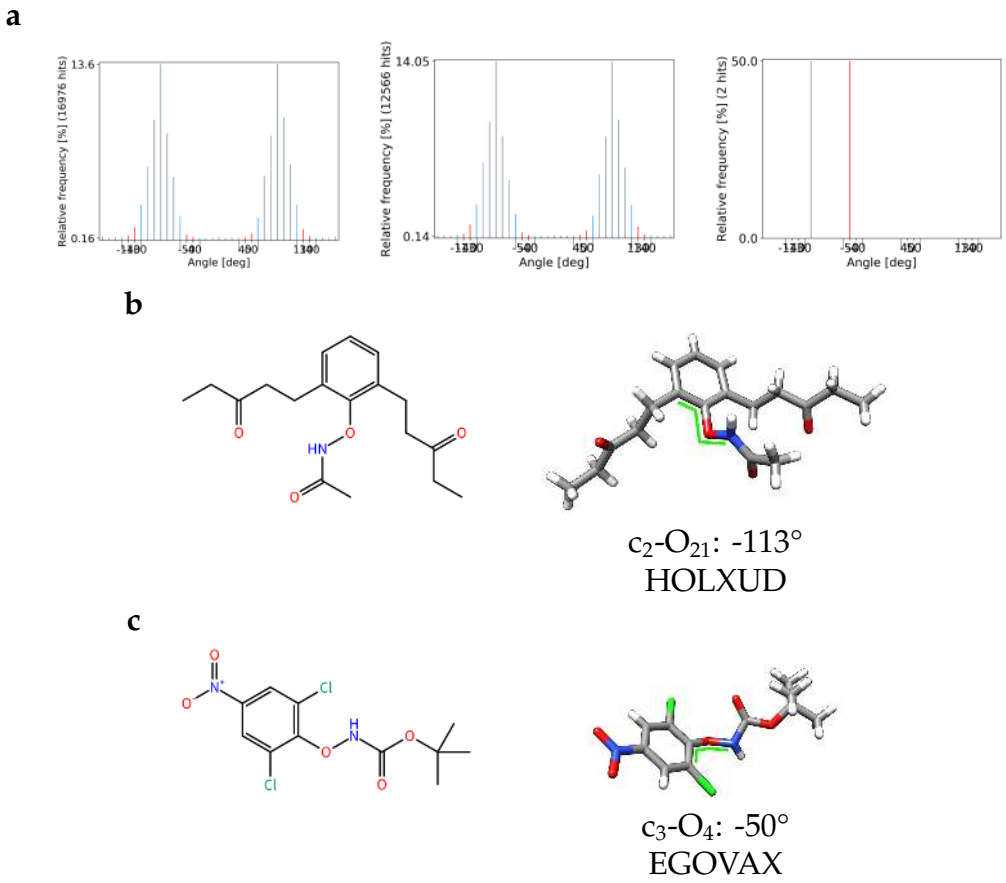


Figure 4.7: The two matching structures of the torsion rule [cH0 : 1][c : 2]([cH0])−!@[O : 3][!#1 : 4] from the TorLib 18 on the CSD18 in validation mode. While the statistic of the pattern is filled with 16976 hits, the validation shows only two matching bonds. The peaks are also not supported by the new matching strategy. The torsion angle in HOLXUD is outside of the second tolerance of any peak due to sterical hindrance.

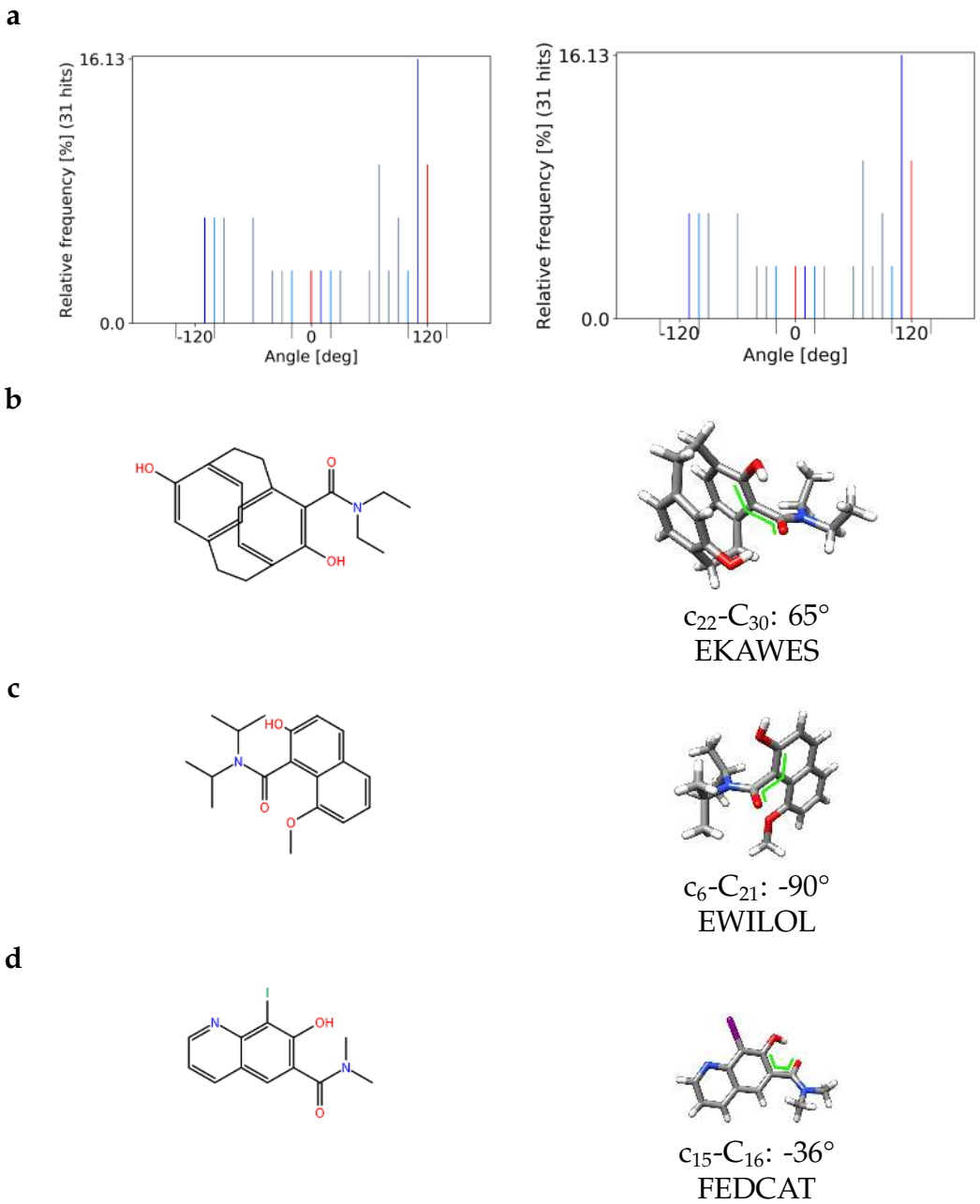


Figure 4.8: Outliers of the torsion rule [$(c[OH1]) : 1][c : 2]−!@[CX3 : 3]([NX3H0]) = [O : 4] from the TorLib 18 on the CSD18. The statistic of the pattern is only filled with 31 hits. The pattern was resorted thus was not controlled in this constellation in 2016.

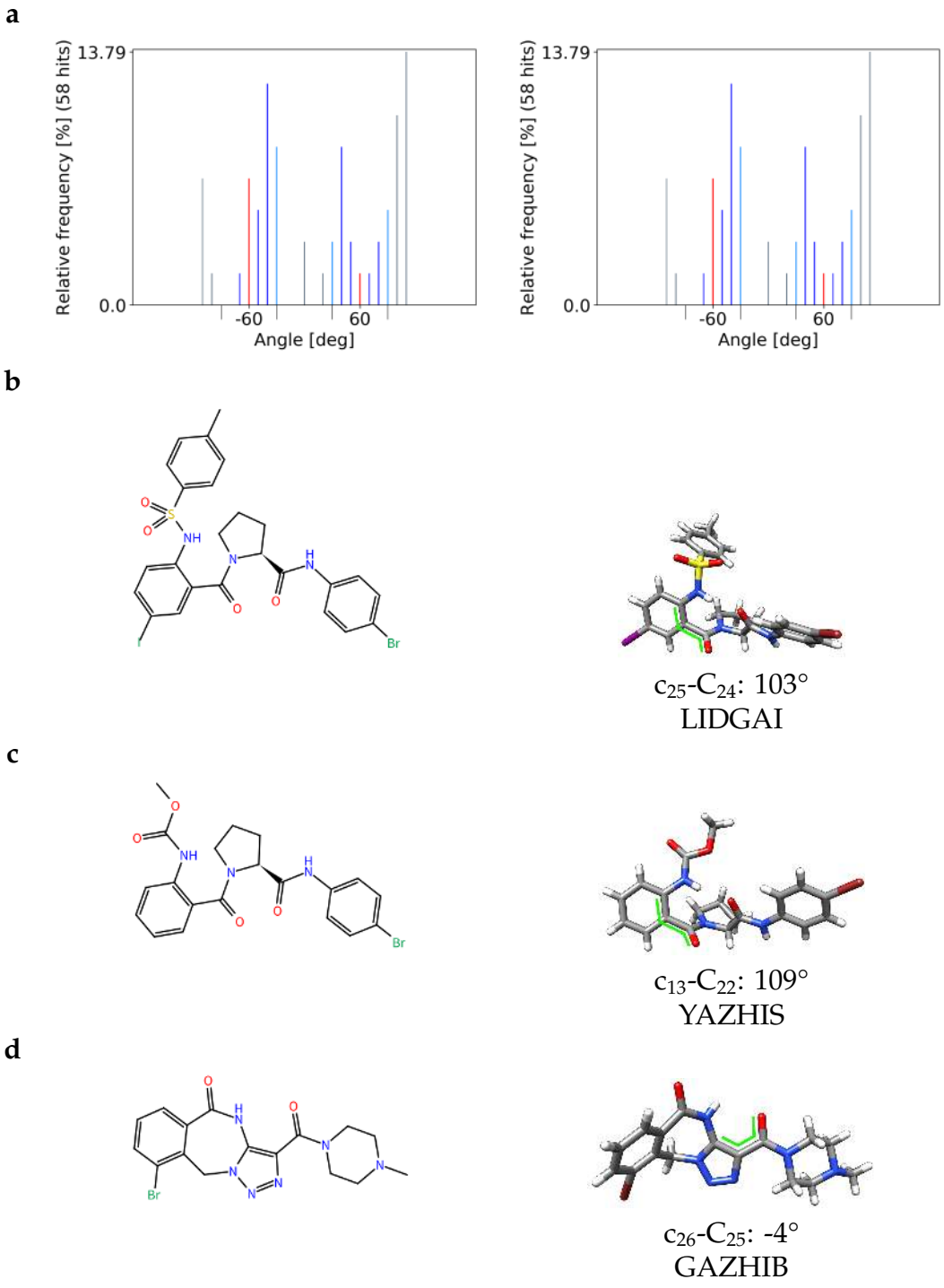


Figure 4.9: Outliers of the torsion rule [$(c[NH1, NH2]) : 1][c : 2]−!@[CX3 : 3]([NX3H0]) = [O : 4] from the TorLib 18 on the CSD18. The statistic of the pattern is only filled with 56 hits. The pattern was resorted thus was not controlled in this constellation in the analysis from 2016.

Comparison to PDB18

The high quality PDB ligand set (see Section 4.1.2) was fed into the TorsionPattern-Miner im combination with the TorLib18 to evaluate its performance. 19% of the torsion rules show more than 40% red flags (Table B.9) . Of these 25 torsion rules, 14 are matched over 10 times as the most specific torsion rule. We examined the three maximally matched torsion rules to search for systematic differences between the two molecule sets (see Figures 4.10 - 4.11).

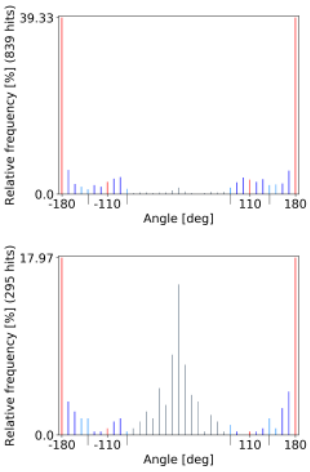
The first example [O : 1] = [CX3 : 2]−!@[CX3 : 3] = [O : 4] (Figure 4.10)**a** shows the population of 0° by ligands in the PDB. The position of fructose-6-phosphate (3t2e, F5R A 3469) in its pocket suggests that the isolating effects of bulk water in combination with the surrounding pocket facilitates the given angle. In the case of orotic acid ([nX3H1 : 1][c : 2]−!@[CX3 : 3] = [O : 4], see Figure 4.10**b**, 1g0x, ORO A 1) displays a case for an echo of the CSD peak at 0°. Ten interactions between pocket and orotic acid stabilize a slightly skewed torsion bond. Due to the mesomeric ability of the carboxylate group, the 180° CSD peak also has a shadow peak around 130° in the PDB ligand histogram. NS3/4A protease inhibitors such as danoprevir own a sulfonamide group with an angle of 180° from the groups nitrogen to the attached cyclopropyl group (Figure 4.11) which is stabilized by two protein-ligand hydrogen bonds. Thus they do not conform to the set of likely torsion angles of -80 and 80°.

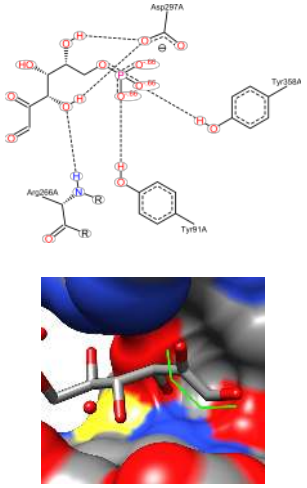
Outlook

Evaluations on three data sets as well as heavy changes in the torsion library and their performance on the evaluation data sets were described. The reordering of the SMARTS has left the torsion library as well performing as before. Through detailed analysis, problems have been detected that should be considered in future work.

A general reevaluation is necessary for the two top most patterns in Table 4.4 due to the old smarts matching error. Torsion rule [a : 1][c : 2]−!@[O : 3][CX3H0 : 4] should be made more specific to account for the sterical limitations created by an attached benzol ring. [cH0 : 1][c : 2]([cH0])−!@[O : 3][!#1 : 4] could be updated with the multi-chain environment to account for the depicted cramped situation at the benzol ring. Besides such specific modifications, two more comprehensive updates should be introduced in the future. Firstly, the matching torsion rule can change based on the protonation of the environment around a rotatable

a

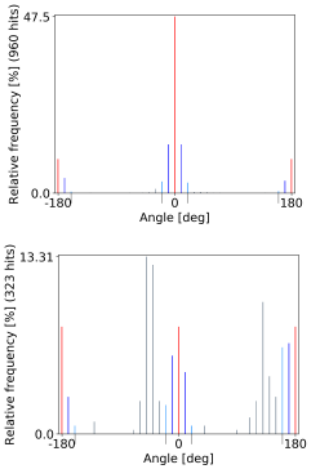


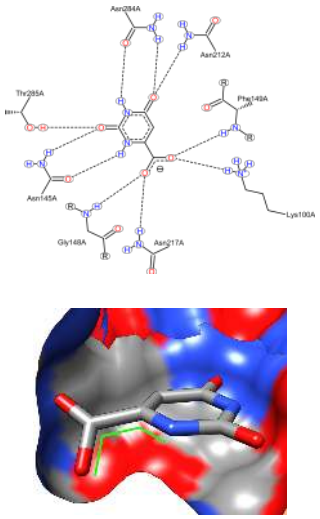


[O : 1] = [CX3 : 2]–!@[CX3 : 3] = [O : 4]

Fructose-6-phosphate C_{11}-C_{12}: 6°
3t2e F6R A 3469

b



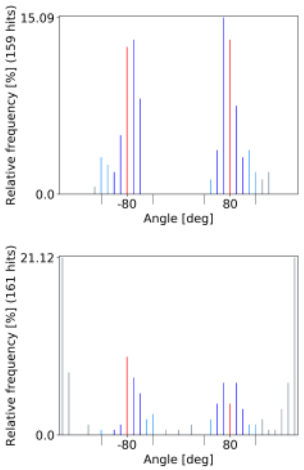


[nX3H1 : 1][c : 2]–!@[CX3 : 3] = [O : 4]

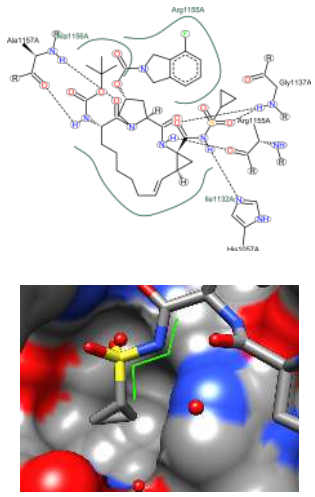
Orotic acid C_7-C_{10}: -46°
3g0x ORO A 1

Figure 4.10: High quality PDB ligand torsion angles in comparison to CSD statistics I. The CSD statistics shows all possible hits of each SMARTS pattern over the whole CSD, while the PDB ligand statistics only counts the most specific SMARTS for each bond. Three cases highlight the noteworthy differences between both data sets.

a



[C : 1][$(S(= O) = O) : 2]−!@[NX3H1 : 3][C : 4]



Danoprevir S_1-N_{14}: 180°
3m5l TSV A 100

Figure 4.11: High quality PDB ligand torsion angles in comparison to CSD statistics II. The CSD statistics shows all possible hits of each SMARTS pattern over the whole CSD, while the PDB ligand statistics only counts the most specific SMARTS for each bond. Three cases highlight the noteworthy differences between both data sets.

bond. For example, a tyrosine treated as ligand in the tyrosyl-T/RNA synthetase (4ts1) with an negatively charged carboxyl group is matched with the most specific torsion rule [O : 1] = [C : 2]([O−])!@[CX4H1 : 3][H : 4] with the peaks at 180, -120 and 120 °. After protonating the oxygen, the most specific torsion pattern is [N : 1][CX4 : 2]!@[CX3 : 3] = [O : 4] with the peaks at 0 and 180 ° which marks the bond in this case as unlikely. The pattern with the negative charged oxygen was matched 734 times in the CSD13 but never in the CSD18. Further analysis revealed negatively charged oxygens in the CSD18 but random molecule samples did not reveal any molecule to be present in the CSD18 responsible for a hit in the CSD13. All statistics can be found in Figure B.1. The divergence in the data sets and the torsion rules involved in scoring protonation states should be analyzed and harmonized.

The second major update is about patterns that use terminal heavy groups as part of their statistic. An especially difficult case are pattern that include terminal hydroxy groups. Those are present in high frequency in the data set but the position of the hydrogen has great flexibility. A bond to a hydroxy group is not seen as rotatable in the subsequent validation as well as in the day-to-day use of the TorLib. Thus, using such groups may cloud the statis-

tic in certain cases. A search for a hydroxy group as labeled part of the torsion rules detected the three patterns [cH0 : 1][c : 2]([cH0])−!@[O : 3][!C; !H : 4], [cH0 : 1][c : 2]([cH1])−!@[O : 3][!C; !H : 4], [cH1 : 1][c : 2]([cH1])−!@[O : 3][!C; !H : 4] (Figure B.2), only varying in the number of hydrogens at the carbon atoms, to be impacted by the clouding effect. The comparison between the pattern's distribution in the TorLib 16 vs TorLib 18 vs. the validation statistic of the TorLib18 show weak backings of the peaks in the first two torsion libraries. Only the number of hits per pattern in validation mode back the marked peaks in two of the three patterns. Pattern [cH0 : 1][c : 2]([cH0])−!@[O : 3][!#1 : 4] is similar to the above mentioned first pattern. We propose to reconsider the existence of [O : 1] = [C : 2]([O−])!@[CX4H1 : 3][H : 4] and check other patterns for their stability against protonation. It would be preferential if a pattern switch due to protonation does not result in a change in the angle likeliness. One possible solution would be changing the SMARTS expression of the fourth node in ~ −!@[O : 3][!C; !H : 4]to [!C; !H; !#1 : 4] to not only exclude carbons with one implicit hydrogens but also explicit hydrogens. Another strategy could be to determine the statistic only based on rotatable bonds.

Overall, the change from CSD13 to CSD18 has shown a rise in the number on unlikely torsion angles (see Figure 4.5a). We advise two steps to counter the development. An automatic strategy for peak detection combined with the help of an expert needs to reevaluate each peak in the torsion rules. It should also be evaluated if a switch to only use the single matching mode for peak detection in certain cases removes the described effects.

While molecules in the CSD are subject to influences by the crystallized content, ligands from the PDB are influenced by interactions to the protein pocket as well as effects from the crystallization process. Torsion rules only based on the covalently bound environment can not integrate exterior forces such as stabilizing interactions that results in breaking up internal hydrogen bonds or stabilizing unlikely torsion angles. The torsion library based on CSD histograms is hence well suited for conformation generation and light, local geometrical optimization. If used on ligands bound in protein binding pockets, the effects of interaction and spatial influences need to be considered additionally.

4.2 Continuous Torsion Score

The torsion rule peaks freely available as part of the torsion library and curated by experts are an attractive base for scoring the relative torsion angle preference. Scoring functions prefer twice differentiable curves over discrete histogram bins. Hence, the so-called kernel density estimation can be used employed. A kernel is a continuous function such as the normal distribution. If per bin a normal distribution is centered on each bin scaled by the bin value, the summation over all these normal distributions results again in a continuous function which then is easy to optimize. Thus, at any bin, not only the curve describing the current position but all other curves have to be computed as well and summed up. Since torsion angles are periodic on 360° the periodic normal distribution, named von Mises function can be employed [35]. Its parameter κ is a measure of concentration for the von Mises distribution (see Figure 4.12 a). A value of zero results in an uniform distribution while an increasingly positive value results in an increasingly concentrated distribution at the peak position.

We have developed equations to compute the von Mises curve width in computing κ as the measure of concentration from a torsion peak. To determine the curve width, the curve peak score (see Equation 4.2) needs to be put in relation to a second point on the curve, here the second tolerance at 1.5% (see Equation 4.3). The resulting Equation 4.4 is then derived to compute κ.

$$
\begin{aligned}
f(x, \mu_m, \kappa, \alpha) &= vonMises(x, \mu_m, \kappa, \alpha) = \alpha \cdot e^{\kappa \cos(x-\mu_m)} && (4.1)\\
f(\mu_m)_{\mu_m,\kappa,\alpha} &= \alpha \cdot e^{\kappa} = s_i && (4.2)\\
\alpha &= \frac{s_i}{e^{\kappa}} \\
f(\mu_m + \tau_{2_i})_{\mu_m,\kappa,\alpha} &= 1.5\% \\
1.5\% &= \alpha \cdot e^{\kappa \cdot \cos(\tau_{2_i})} \\
1.5\% &= \frac{s_i}{e^{\kappa}} \cdot e^{\kappa \cdot \cos(\tau_{2_i})} && (4.3)\\
\kappa(s_i, \tau_{2_i}) &= \frac{\ln \frac{1.5\%}{s_i}}{cos(\tau_{2_i}) - 1.0} && (4.4)
\end{aligned}
$$

The continuous torsion score for a given angle can then be computed in calculating the normalized curvature of each kernel per peak scaled by its relative peak score. The sum over all kernels is then normalized with the sum over all relative peak scores to achieve a surface area of one (see Equation 4.6). To keep the score

in the interval [0, 1] it is finally normalized with the overall maximum score (see Equation 4.7).

$$f_h(x) = \frac{1}{h \cdot n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{4.5}$$

$$f_1(x) = \frac{1}{\sum_{i=1}^{n} s_i} \sum_{i=1}^{n} s_i \cdot \frac{2\pi I_0(\kappa_i)}{e^{\kappa_i \cos 0}} \cdot \frac{e^{\kappa_i \cos(x-\mu_{m_i})}}{2\pi I_0(\kappa_i)} \tag{4.6}$$

$$= \frac{1}{\sum_{i=1}^{n} s_i} \sum_{i=1}^{n} s_i \cdot e^{\kappa_i(\cos(x-\mu_{m_i})-1)}$$

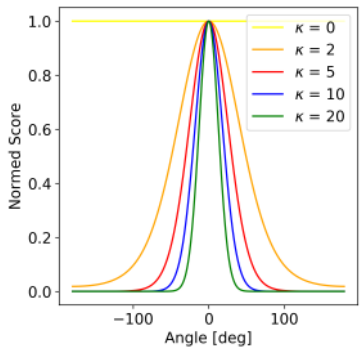
$$f_1(x)_{\text{norm}} = \frac{1}{\max f_1 \cdot \sum_{i=1}^{n} s_i} \sum_{i=1}^{n} s_i \cdot e^{\kappa_i(\cos(x-\mu_{m_i})-1)} \tag{4.7}$$

The torsion library is produced in using expert knowledge. Thus, preferred peaks are set even though these angles are not frequently observed in the current CSD. If the score is below 1.5%, it does not fulfill the prerequisites of the torsion lib for computing κ and κ is set to 20, resulting in a locally concentrated peak (72 cases, Figure 4.12**b**, Tables B.12 - B.13). If the score is zero, the peak is omitted from the estimation and later in the calculation (17 cases, see Table B.11). Due to cumulative effects two problem are possible. The overall curve may be beyond the interval [0, 1] (**1**). Also, neighboring scores may change their relative ranking (**2**). Both cases are tolerated with an epsilon of 2^{-20}. If necessary, the peak with the maximum interference is identified and its κ will be increased in decreasing the second tolerance step wise by 0.05% of the initial second tolerance. 19 patterns were modified due to **2** and none due to **1** (see Figure 4.12**b**, Figures B.3-B.8). 20 torsion rules were not matched at all thus all peaks were set to zero (Table B.10).

4.3 Conclusion

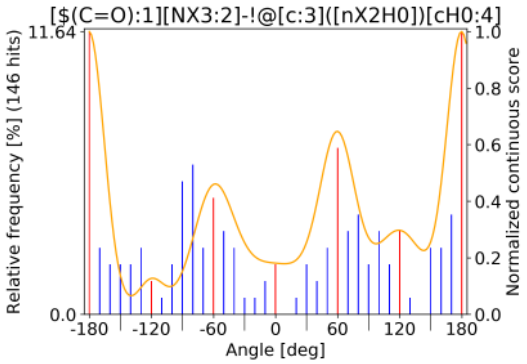
The chapter has covered multiple improvements for the torsion library resulting in the TorLib18. Additionally, the two times differentiable Continuous Torsion Score (CTS) was derived to score the likeliness of torsion angles in e.g. a geometrical optimization. Hence, after defining EDIA, identifying a prober training and validation data set ProtFlex18, and developing the CTS, all missing pieces to evaluate and improve GeoHYDE, the objective function for geometrically optimizing for HYDE, are now assembled.

a



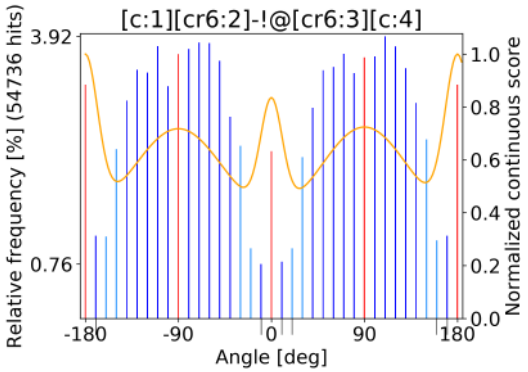
Influence of multiple κ on the normed von Mises distribution

b

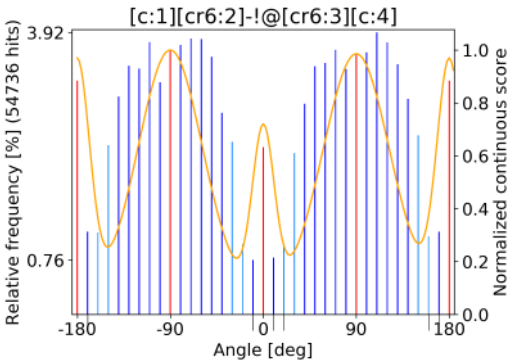


Score at angle -120 is below 1.5%.

c



Original



Corrected.

Figure 4.12: Plots for describing the continuous torsion score. a show [$(C = O) : 1][NX3 : 2]−!@[c : 3]([nX2H0])[cH0 : 4] with the peak score less than 1.5% at 120°, thus with κ of 20. b displays the change in the continuous torsion score when relative peak ranking is achieved through internal tolerance reduction at the peaks -90° and 90° for pattern [c : 1][cr6 : 2]−!@[cr6 : 3][c : 4].

Chapter 5

GeoHYDE: Optimizing HYDE by Geometrically Optimizing the Pocket

Protein-ligand binding can be assessed with the scoring function HYDE. The pockets can result from crystallized structures but also from docking or molecular dynamics simulations. The HYDE chemistry model is not fully in line with those used in other software since its underlying publication for the interaction geometries was published in 2017 [42]. It expects the geometries to be close to those found in crystallized structures. Hence, a strategy is needed to translate between those slightly differing chemistry models.
As such, the overall aim of this thesis is to develop an optimization function that on the one side closely follows the HYDE model and on the other side is fast to calculate and easy to optimize to serve as a translator. The optimization process should be fully integrated into the in NAOMI existing capabilities of preprocessing three dimensional structural models. Also, the introduction of side chain flexibility when optimizing should be tackled. Since staying close to the HYDE interaction model results in not having an analytical gradient available, great care should be taken to guarantee an unknown but existing gradient so that a search algorithm working with approximations finds reliably the local minimum. The following chapter introduces GeoHYDE as the objective function and motivates the adaptations in it used in this thesis. Then, an extensive evaluation over gradient free optimization algorithms with subsequent weight parametrization over the training data set share of ProtFlex18 follows. The chapter ends with the evaluation of GeoHYDE with varying degrees of flexibility in the pocket over the test data set sections of ProtFlex18. As an external validation, GeoHYDE is tested on the aforementioned CASF-2016 data set closing the chapter.

5.1 GeoHYDE

$$GeoHYDE_{sat} = w_s \cdot \Delta G_{sat} + w_{desolv} \cdot \Delta G_{DP} \quad (5.1)$$

$$GeoHYDE_{ds} = GeoHYDE_{sat} + w_{iLJ} \cdot GeoHYDE_{desolv} \quad (5.2)$$

$$GeoHYDE = GeoHYDE_{ds} + w_t \cdot E_{tors} + w_{rLJ} \cdot E_{intra} \quad (5.3)$$

$$GeoHYDE_{prot} = GeoHYDE + w_{tp} \cdot E_{torsp} + w_{rLJp} \cdot E_{intrap} \quad (5.4)$$

GeoHYDE as published by Schneider *et al.* [65] in 2012 consists of HYDE's saturation term (Equation 5.1) with an intermolecular Lennard-Jones potential (LJP) $GeoHYDE_{desolv}$ to describe repulsive effects in close contact but also the attractive forces present as part of the hydrophobic effect (Equation 5.2). To safeguard the in the geometric optimization flexible ligand against unusual torsion angles and clashing atoms, an unspecified torsion score and an intramolecular LJP completes the GeoHYDE equation (see Equation 5.3, E_{intra}, LJ_{intra}). Its weights of 2012 and the empirical ones as of 2018 are listed in Table 5.1. In the optimization, the ligand can change its orientation and can be translated. Additionally, rotatable bonds and single bonds leading to hydrogen donors can be rotated in the ligand.

Due to the move to the then new NAOMI code base in 2012 at the beginning of this thesis, GeoHYDE and the library for handling interactions had to be fully reimplemented. While the general terms have been left unaltered, some implementation details had to be changed to account for the subsequently presented reasons. The project partners Bayer and BioSolveIT identified multiple problems through single case analysis:

1. With the eye not discernible changes in the initial ligand pose resulted in distinct pose and hence score differences after the optimization.

2. In many cases, the ligand was detected to be too close to the residues of the protein.

3. Averaging over three to four hydrogen bond quality factors with a normal mean was found to be too lenient when mediocre interactions were present and should have been penalized.

The problem of diverging poses after optimization suggests, that the objective function consists of a very rough energy landscape. As first step, the optimization of GeoHYDE was changed from numerically determining derivatives with the Quasi-Newton method to BOBYQA as the current gradient-free optimization algorithm.

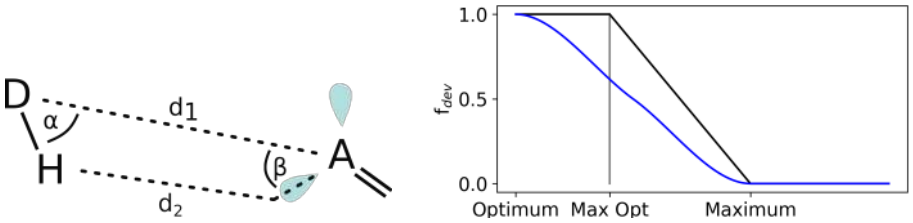
Also, performance can be further improved in guaranteeing the existence of a second derivative over the domain of the function. Hence with the help of our cooperation partner BioSolveIT and all members in our project, the scoring function in the Lennard-Jones potential and in parts of HYDE were adjusted to be in theory two times continuously differentiable and stabilized against differences between operating systems. Additional care was placed on smooth scoring of interaction quality. Also, side chain flexibility was added to GeoHYDE extending GeoHYDE as in Equation 5.4 for the protein side.

The described HYDE-GeoHYDE combination was then used to evaluate mutation effects in the protein on the protein stability.[68] The scoring combination showed overall better results which can be computed in just around a minute in contrast to the alternative MD simulations.

Subsequently, further progress was made in quantifying the quality of interactions for HYDE and in general. Four quality factors now describe the hydrogen bond quality in HYDE 2018 (Figure 5.1(a)). It was found that switching from the arithmetic mean to the power mean with the exponent of 1 to one of -2 would score interactions with at least one low quality factor more closely to the model developer's intention (Equation 5.5).

$$\bar{x} = \left(\frac{1}{m} \sum_{i=1}^{m} x_i^n \right)^{\frac{1}{n}} \tag{5.5}$$

The thus derived score per interaction is one of the goals of GeoHYDE to optimize. In NAOMI, all interactions have an optimal range forming a plateau, called maximum optimum after which the quality estimator drops from one to zero. GeoHYDE has in contrast no plateau between optimum and maximum optimum but instead has the maximum optimum moved to the optimum to allow the optimization function to focus on the actual goal of a good interaction geometry (Figure 5.1(b)). The last problem to be tackled are the close contacts between atoms which is directly linked to the parametrization of the intermolecular LJP. BioSolveIT and Bayer took great care in fine-tuning the LJP to let it mirror the actual observed distances in public crystallographic protein-ligand complexes. Additionally, the positions of zero crossings were identified for a number of non-covalently bound neighboring functional groups. Depending on the atom's functional group, hydrogens are considered for clash control. As final update, the torsion angle potential was changed to the in Chapter developed 4 Continuous Torsion Score (CTS) based on the Torsion Library 2018 on the protein and ligand side. It is accompanied



(a) Hydrogen bond quality factors between donor and acceptor (adapted from Fig. B.1 [40])

(b) Interaction quality factor f_{dev} in HYDE and GeoHYDE (blue)

Figure 5.1: Interactions in HYDE

by an intramolecular Lennard-Jones-Potential to safeguard each in the geometric optimization flexible substructure against clashes. Hence, when GeoHYDE is used to optimize flexible pockets both terms are used for the ligand and each flexible side chain with identical configuration (Equation 5.4).

Evaluation Strategy

An evaluation strategy for GeoHYDE needs to answer the following questions in geometrically optimizing crystal poses in ProtFlex18. Generally, the poses from a high quality crystal dataset should not derivate far from their crystallized poses.

1. Do the partial scores of GeoHYDE perceive the ProtFlex18 data set as high quality as well?
2. Which gradient free optimization algorithm e.g. from `nlopt` can be used for GeoHYDE? Does it reliably terminate the computation and how much time does optimization need?
3. Analyzing the initial to final HYDE and GeoHYDE score shifts, can trends be detected to guide a grid based parameter search for GeoHYDE?
4. What are the optimal parameters for GeoHYDE in this context?

To simulate the more demanding task of handling docking poses, ligands should be sampled with an emphasis on overall changing its position (six degrees of freedom due to allowed rotation around the mass center and translation of the ligand) or changing its internal configuration through rotation around single bonds and bonds to hydrogen donors.

5. Does the above found parameter set perform equally well on the perturbed structures?

When protein flexibility here defined as amino acid side chain flexibility is present in the active site,

6. Does the above found parameter set perform equally well in flexible pockets?
7. Does protein flexibility increase the abilities of GeoHYDE to optimize the final HYDE score?

For analyzing the performance quality of a parameter set, the per cent of structures with an $EDIA_m$ of at least 0.8 combined with low RMSD and the highest HYDE improvements should be observed.

5.2 Methods

In the following, software tools are introduced that were developed to try to answer all of the questions above. Subsequently, the split of ProtFlex18 into training and two test sets is explained. The multi-step parameter search is outlined and the statistical analysis on the data sets is explained as the last part of the section.

GeohydeEvaluator as Benchmarking Tool

Through this thesis, the `HYDE` library was expanded and the `Interactions` library in NAOMI rewritten. For optimization, the `NumOptimization` with the `NumOptimizationHelper` library was created. For a standardized preprocessing, the `HydePreprocessingLib` was added to NAOMI. The graphical tool `HydeDebugGUI` was partially rewritten and extended and two more command line tools implemented. All NAOMI libraries and tools created and modified for this chapter are presented in Section A.2.3.

To examine the stated evaluation questions, the tool `GeohydeEvaluator` based on NAOMI was developed. It accepts as input a PDB file a ligand specification or a multi mol SDF file and a configuration file. For example all partial score terms in GeoHYDE can be activated as wished. The tool can compute the phases sampling, scoring, optimizing, scoring of the pocket in succession or separately. It also accepts a protein flexibility database computed by SIENA which can then be used to identify flexible residues in the binding pocket. This allows a reasonable estimation of pocket flexibility and avoids the computational bottleneck that comes

Weight	Description	HYDE	GeoHYDE$_{old}$	GeoHYDE$_{empirical}$
w_{sat}	GeoHYDE$_{sat}$	1.0	2.0	3.0
w_{desolv}	GeoHYDE$_{sat-desolv,polaratoms}$	1.0	1.0	0.5
	HYDE$_{desolv}$	1.0	0.0	0.0
w_{iLJ}	GeoHYDE$_{desolv}$	0.0	1.0	1.0
w_t	Continuous Torsion Score	0.0	3.0	5.0
w_{rLJ}	intramolecular Lennard-Jones	0.0	1.0	0.5

Table 5.1: HYDE and GeoHYDE parametrizations used in this thesis compared to the parametrization of GeoHYDE of 2012 called GeoHYDE$_{old}$[65].

with full side chain flexibility in the pocket. More information and other tools relevant for HYDE can be found in Section A.2.3. In all cases, each pocket is then preprocessed by the standard NAOMI work flow. It consists of optimizing the hydrogen bond network in the binding pocket of 8 Å with the help of Protoss [5]. Consecutively, all waters are deleted in the pocket to prepare for the implicit water placement technique used in the HYDE version of 2018. The pocket is then scored in default mode with the initially available parametrization (see Table 5.1 of GeoHYDE and HYDE. The parametrization of GeoHYDE can be changed through the tool configuration.

Benchmark Data Sets

Historically, HYDE was developed with the help of the Astex and Iridium data set as well as the 'small series' (see Chapter 3). The low number of pockets available in these data sets and their longterm involvement in the development may have resulted in tuning HYDE and GeoHYDE to the specific cases in the data set.

The through this thesis assembled ProtFlex18 data set with its 2386 pockets provides not yet seen opportunities for sound parameter tuning such as splitting between training and test data for in-domain and out-of-domain generalization tests in accordance to current benchmarking standards. Great care should be used in analyzing similarity in the data set. As mentioned, all pockets were clustered to ensembles with the help of SIENA. To avoid bias per cluster, each unique ligand, defined by its HET code, should only be present once. Thus, only the ligand with the highest EDIA$_{\mathrm{m}}$ of those with an identical HET code per cluster is kept to be used in the further creation of benchmark sets. The data split used in this thesis was initially that of a training set of the 1095 most common structures. If a cluster has less than ten entries, every tenth pocket is send to the test set ProtFlex18$_{id}$. Testing the model for out of domain generalization can be done on the 101 least

common pockets detected by the SIENA analysis (ProtFlex18$_{od}$). No pocket in any test set is present in ProtFlex18$_{train}$. The data set for testing in-domain generalization consisted initially of around 122 representatives from each cluster present in the training set. Subsequent filtering due to proton clashes (Section 5.3.2, 231 pockets) has resulted in ProtFlex18$_{train}$: 997, ProtFlex18$_{id}$: 112, and ProtFlex18$_{od}$ with 101 pockets.

Parameter Search Methods

Three types of parameter search were conducted. First, the relevance of each score part of GeoHYDE have to be determined. Then, different values per weight are tested while keeping all other weights to the empirical values (Tables 5.1, 5.4). Finally, a 'capped' Lennard-Jones potential with a removed attractive part developed by Florian Flachsenberg is evaluated over a range of possible scores. The final EDIA$_{\mathrm{m}}$, RMSD values and the change in HYDE scores of the ProtFlex18$_{train}$ data set are observed. All parameter searches are accompanied by setting the weight to three specific values. The weight set to zero checks for its overall relevance in the optimization. A weight of 100 checks for the score parts influence when other policies are present but not highly relevant. Finally, a weight of one while all other weights are set to zero so that only the respective score part drives the optimization can show the maximum potential of each score part. It is labeled as 'only' in each plot. The ProtFlex18 data set has been assembled through e.g. avoiding strong intramolecular clashes as well as unlikely torsion angles. Score terms in GeoHYDE that safe guard against both parts have not shown any statistically significant reaction to the weight changes tested on the pockets in ProtFlex18$_{train}$. Subsequently, perturbation was applied in both rotating and translating the ligand deterministically around its center of mass as well as rotating around rotatable bonds. The ligand perturbations for the data set are selected once and kept for all further parameter validation runs. More information about the sampling and its configurations can be found in Section A.2.3.

Statistical Analysis

Since 1019 data points allow the use of a statistical test, the Mann-Whitney-Wilcoxon Rank Sum test is used. It compares two sets of data to test for identical underlying distributions. If a parameter change in our analysis thus results in visually differing data distributions but the MWW test reports a high probability of an underlying identical distribution, the deviation should not be taken as relevant.

5.3 Results

Firstly, the performance of multiple gradient free optimization algorithms with the GeoHYDE$_{old}$ parameter set is examined. With the thus selected algorithm, initial scores and their shifts through a geometric optimization on ProtFlex18$_{train}$ are discussed. Finally, the results over the parameter search to define GeoHYDE$_{final}$ are presented.

5.3.1 Optimization Algorithms

In the following, the gradient free optimization algorithms available in `NLopt` have been evaluated with GeoHYDE in terms of computation time and similarity in final scores. `NLopt` suggests six deterministic gradient free optimization algorithms: PRAXIS[52], Nelder-Mead-Simplex[39], COBYLA[55], BOBYQA[56], NEWUOA[54], NEWUOA$_{bound}$[25] and Sbplex[6]. NMS and PRAXIS are superseded and thus not tested. Only BOBYQA (*b*), Sbplex (*s*), and a sufficient number of pockets optimized with the help of NEWUOA (*n*) and NEWUOA$_{bound}$ (*nb*) were able to successfully finish all necessary geometric optimizations in less than four hours per rigid pocket with a flexible pocket to be optimized (BOBYQA: 2155, Splx: 2154, NEWUOA: 2155, NEWUOA$_{bound}$: 1068). For all, the initial step size per parameter was set to 0.4 (radiant and Ångstrom). The criteria to detect convergence were for the change in function value: $|\Delta f| < 10^{-9}$ and for the absolute change in any function parameter $|\Delta x| < 10^{-7}$. Computing the score correlations between the three algorithms show correlation coefficients between 0.96 and 0.98 (Figure B.9). Four to eight per cent of all data points show a score difference of more than 5 units for which BOBYQA returns a less highly optimized score but NEWUOA against NEWUOA$_{bound}$ with box constraints set to the maximum with only 2%. In contrast, the median computation time of *b* with 24 over 64 for *n* to *nb* with 76 and *s* with 170 seconds triples at best (MWW Test: p values below 0.0001 on ProtFlex18, Figure B.9). The large difference in computation time confirm the current ranking in quality and usability of the different gradient free optimization methods[60]. Since the reached scores have a high correlation, further convergence tests were not conducted.

5.3.2 Quality Analysis of the Initial Poses

The ProtFlex18 data set is selected through e.g. removing intra- and intermolecular clash in the binding site. Also, uncommon torsion angles are a criterion for exclusion. Thus, the torsion scoring policy as well as the policies including different kinds of Lennard-Jones potential should be observed to not show any major flaws in the 2386 crystallized poses. In Figure 5.2, the initial scores of the four mentioned components are shown. Comparing GeoHYDE$_{desolv}$ as a Lennard-Jones potential that partially includes protons with the generic intermolecular Lennard-Jones potential only calculated between heavy atoms shows 231 poses with a GeoHYDE$_{desolv}$ ligand score above 0 while the heavy atom LJ score reports one structure with a score above one. This strongly suggests problematic placements of protons which are generated by Protoss (Figure 5.3(a)). Since the aim of the evaluation based on ProtFlex18 is to stay close to the high quality crystal structure, the 231 pockets were marked for exclusion as they are not high quality for GeoHYDE. They should be included in future evaluations nevertheless.

2155 pockets remained. Of them, 11% (241, before 256) still have an CTS above zero and 28% (612, before 678) structures show an intramolecular Lennard-Jones potential including protons above zero. Here, cases show clashing protons but also tightly packed ligands (Figure 5.3(c), 5.3(b)). Since many of these cases can be ameliorated through slightly rotating single bonds, all these structures stay in the data set. In the case of the CTS in the default parametrization, one strained torsion angle has a score of 5. Since the CTS sum over all rotatable bonds is maximally seven, the ligands have been filtered properly and do not need additional adjustment.

5.3.3 Analyzing Score Shifts

The score shift through optimizing with the empirically determined score parametrization for GeoHYDE (Table 5.1) can be examined to develop an initial strategy for the parameter search (Figure B.11). While the overall GeoHYDE score always improves, nine structures decrease the for HYDE scoring relevant GeoHYDE$_{ds}$ part of GeoHYDE (Equation 5.2, Table 5.2) due to a focus on the torsion angle and intramolecular LJP score terms. Overall 386 pockets show a a misaligned score development when comparing GeoHYDE$_{ds}$ with HYDE. While improvements in GeoHYDE$_{sat}$ result for 71% in an improved HYDE$_{sat}$ score, overall 290 pockets (29%) have misaligned score directions between GeoHYDE$_{sat}$ and HYDE$_{sat}$. For

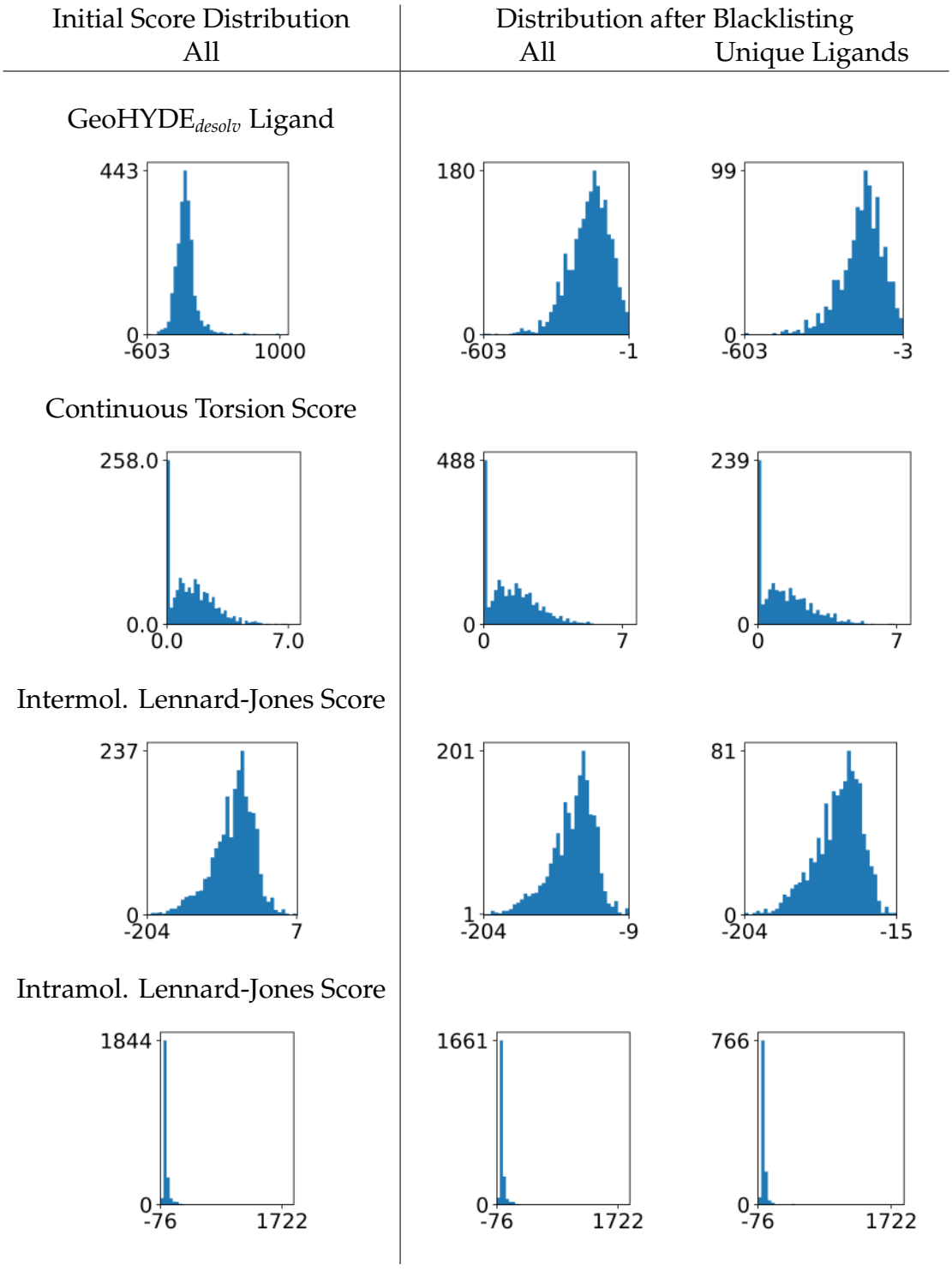
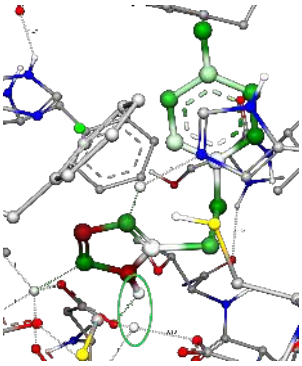
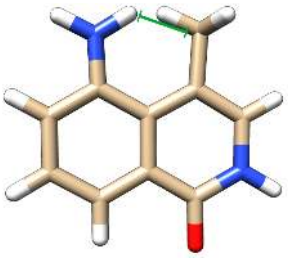


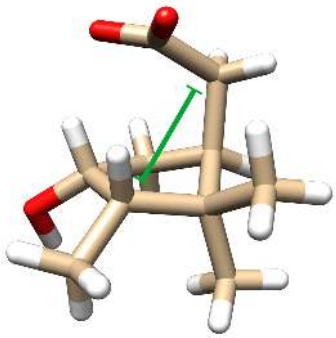
Figure 5.2: Distribution of partial scores of the initial GeoHYDE before and after blacklisting all ligands with a positive ligand GeoHYDE$_{desolv}$ score. The last column shows only the ligands present in the filtered ProtFlex18 data sets. All plots show the minimum and maximum score on the x-axis as well as the maximum frequency per plot on the y-axis. Per row, the score distributions stay similar while the number of ligands is reduced. The plot of GeoHYDE$_{desolv}$ ligand scores in its entirety can be found in Figure B.10



(a) T03 A 289 (3iu8) GeoHYDE$_{desolv}$: 323455



(b) Hydrogens in plane with ring in G1O A 2167 (4uvt), Intramol. LJ: 1722.5 kJ



(c) C3 - C10: 3.1 Å in CAX B 5002 (1szo), Intramol. LJ: 118.5 kJ118.5

Figure 5.3: Structures with protons modeled too close to heavy atoms as well as a tightly packed ligand representing common reasons for a high Lennard-Jones score. As annotated, the two first examples represent the poses generating the maximum inter- and intramolecular Lennard-Jones Scores in Figure 5.2.

	Sat. + Desolv.				Saturation				Desolvation			
GeoHYDE$_{ds}$ diff sign	+	+	-	-	+	+	-	-	+	+	-	-
HYDE diff sign	+	-	+	-	+	-	+	-	+	-	+	-
GeoHYDE$_{empirical}$	602	386	0	9	614	243	47	98	474	494	9	20
GeoHYDE$_{final}$	611	381	1	4	696	266	13	22	369	452	56	120

Table 5.2: Agreement between GeoHYDE and HYDE score components with the empirical and final parametrization.

GeoHYDE$_{desolv}$ in only 49% of all cases, score improvements in GeoHYDE$_{desolv}$ result in score improvements for HYDE$_d$.

In the following, full total score shifts are discussed. As preferred for high quality crystal poses in the ProtFlex18$_{train}$ data set, strong score shifts can not be detected in all histograms when comparing the initial to final score per GeoHYDE scoring term (Figure B.12, B.13). Only GeoHYDE$_{desolv}$ shows the already detected decrease in the already negative initial score through the geometric optimization.

To diminish the strong influence of GeoHYDE$_{desolv}$, two options come to mind:

- Reduce the overall w_{iLJ} to a value below one.
- Reshape the Lennard-Jones curve in its attractive area.

w_{sat} (GeoHYDE$_{\text{sat}}$)
0, 1, 2, **3**, 4, 5, 3.5, 4, 4.5, 5.5, 100
w_{iLJ} (GeoHYDE$_{\text{desolv}}$)
0, 0.001, 0.01, 0.1, 0.3, 0.5,
0.7, **1**, 2, 3, 4, 100
w_{desolv} (GeoHYDE$_{\text{sat-desolv, polar atoms}}$)
0, 0.1, **0.5**, 1, 2, 3, 100
w_t (Torsion weight)
0, 1, 2, 3, 4, **5**, 6, 7, 8, 100
w_{rLJ} (Intramolecular LJ weight)
0, 0.2, **0.5**, 0.8, 1, 2, 3, 100

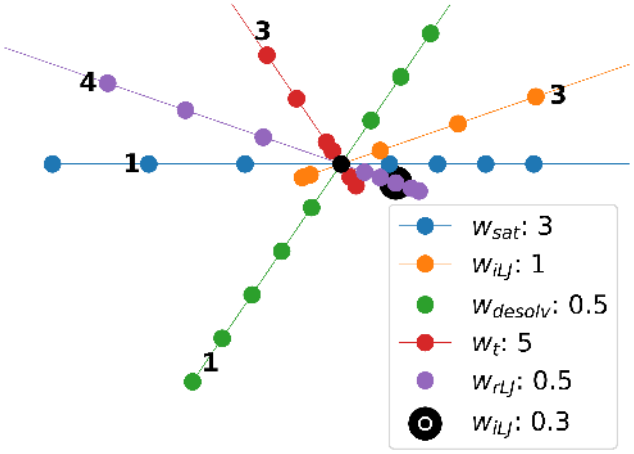


Figure 5.4: GeoHYDE parameter search starting from the empirical parameters marked in bold. As result, a new weight of 0.3 is only determined for w_{desolv}. All other weights stay the same.

5.3.4 Parameter Search

In the following, a variation of the greedy search for parameter tuning is performed. In each test, one weight is changed in agreement with Table 5.4 while all other weights are kept to the empirical ones (Table 5.1). The first parameter to be estimated is w_{desolv}. It scales the desolvation part for polar atoms (Equation 1.12). In HYDE set to one and in GeoHYDE$_{empirical}$ set to 0.5, it was now tested with the values 0.1, 0.5, 1, 2, 3. Additionally, solely using the partial score term as well as overweighting it with a value of 100 and not using it in contrast to the other contributions was tested. The results over ProtFlex18$_{train}$ are shown in Figure B.14. The MWW test only suggests changed base distributions for the entry 'only' for RMSD and the HYDE difference ($p < 1e\text{-}4$). In the case of the changes in EDIA$_{\text{m}}$ nearly all pairwise combinations show a p value below 1e-3 or smaller. While RMSD and EDIA$_{\text{m}}$ control against the deviation from the crystal structure, the optimization over the whole data set should maximize the positive HYDE score difference. Here, the test for a significant difference does not advice to see any parameter change as significantly different. As a consequence for the final parametrization for GeoHYDE (GeoHYDE$_{final}$), w_{desolv} will be kept to its empirical value.

The second parameter to be validated is w_{sat}. It scales the contribution of hydrogen bond functions to the overall GeoHYDE and HYDE equation. In GeoHYDE$_{empirical}$ it is set to three. Besides the initial three parameter configurations, the geometric optimization was evaluated on the values 1, 2, 4, 5, 6, 7, 8, 9 and 10 (Figure B.15). A change in w_{sat} has a strong influence on the optimization result. While optimizing

only with the GeoHYDE$_{sat}$ partial score results in a significant drop in final EDIA$_{\mathrm{m}}$ values to a median of 0.52, its exaggerated weight of 100 only lets the median EDIA$_{\mathrm{m}}$ drop to 0.79. With the empirical weight of 3, the median is 0.88. The weight values of one to eight report a median of 0.9 with a median RMSD between 0.26 and 0.28. As can be seen in B.15, the median of the HYDE differences changes from 0.19 to 0.63 with the empirical weight having a median of 0.53. While the value four seems to be the best, the MWW test reports a p value of 0.96 for the probability of being from the same distribution as the weight value three. As consequence, the empirical value of three for w_{sat} in GeoHYDE$_{final}$ is kept.

The third parameter to be validated is GeoHYDE$_{desolv}$. This weight regulates the contribution of a 12-6 Lennard-Jones potential considering protons and fitted to distances in crystal structures. It is evaluated as the only score contribution, on its empirical weight of one and on its influence to the overall optimization in setting it to zero. Additionally, the values 0.0001, 0.001, 0.01, 0.1, 0.3, 0.5, 0.7, 2, 3, 4 and 100 were tested ((Figure B.15). Over all test runs, the median improvement of the HYDE score was maximally 0.6 for the weight 0.3 followed by 0.53 for both the weight of one and 0.1. Between all three distributions, the MWW p value is at least 0.4 which would not make it necessary to change the weight from the default value of one. Additionally considering the EDIA$_{\mathrm{m}}$ and RMSD spread shows that the weight of 0.3 results in better EDIA$_{\mathrm{m}}$ (0.89 against 0.88 and 0.86) and lower RMSD (0.26 against 0.27 and 0.32). Additionally, the higher whisker spread in RMSD for the weight 0.1 is larger (0.44 to 0.77 against 0.37 to 0.65) with a p value of below 0.0001. As a result, the GeoHYDE$_{desolv}$ weight is changed from one to 0.3.

The fourth parameter to be validated is the weight for the continuous torsion score w_t. Besides the empirical weight of 5 the weights of 0, 1, 2, 3, 4, 6, 7, 8 and 100 are tested as well in using only the CTS as the scoring term guiding the geometric optimization (Figure B.16). The results show very similar behavior over all tested weights but the weight of 100 and when using the CTS on its own. The last two cases are not recommended. A sound decision to choose between the other possible weights does not seem to be possible. Since the ligands have been selected to be high quality, the CTS term may not be a strong influence on the overall geometric optimization from the start. To identify a proper weight, ligands should be sampled based on torsional degrees of freedom in their pocket as a following experiment. The same findings hold true when evaluating the experiments for the intramolecular Lennard-Jones potential w_{rLJ} which guards the ligand against internal clashes through the geometric optimization (Figure B.16).

Pertubation of Ligands to Identify Values for w_t and w_{rLJ}

Through four configurations called GTT, GTTL, T and TS (Table 5.5(a)), the ligands have been perturbed from zero to 3 Å RMSD from their original structure (see Figure 5.6(a)) over ProtFlex18$_{train}$. Every structure has contributed at least one ligand configurations. While the GRTL configuration has in the most cases 20 configurations per structure, the GRT configuration is concentrated on one to ten configurations. A overall different pattern is seen with the T and TS configurations. Both focus on structures with either one, five or ten to 20 configurations 5.6(b). As an example an overlay of all poses of 3VR B 502 in PLP-dependent transaminase (4wyd) is shown in Figure 5.7. A decomposition per sampling strategy can be found in Figure B.17.

Subsequently, optimization with GeoHYDE was carried out with various values for w_t and w_{rLJ} (see Table 5.5(b)). Tests were run with the empirical parameters but w_t set to five and 1 combined with w_{rLJ} to 0.05, 0.1, 0.5 and 1 with GeoHYDE$_{desolv}$ set to 0.3 for all three. The results can be found in Figure B.18 to B.25. No substantial and by the Man-Whitney-Wilcoxon Test change defined as significant was found. Hence, the weights w_t and w_{rLJ} are kept unchanged.

GeoHYDE$_{desolv}$ as a Repulsive Lennard-Jones Potential

In Section 5.3.3, a second option to increase the abilities of GeoHYDE was suggested: to change the intermolecular Lennard-Jones potential in GeoHYDE$_{desolv}$ to a purely repulsive ('capped') one to avoid the accumulation of large negative potentials. The curve is approximated by polynomials up to the degree of four and implemented by Florian Flachsenberg in NAOMI in the `ScoringLib` (more information in Section A.2.3). While too close contacts are still penalized, attractive effects are not considered in this way. Ligand optimization was performed over all weights also used for evaluating GeoHYDE$_{desolv}$ given in Table 5.4. While a weight of 0.3 proved also here to be the best choice for the 'capped' Lennard-Jones potential, the default LJP still performed significantly better for e.g. the weight of 0.3 over HYDE, EDIA$_{\text{m}}$ and RMSD (p values: 0.0054, < 0.0001, < 0.0001, Figure 5.8).

Comparative Analysis of Poorly Optimized Pockets

Experiments with two Lennard-Jones potentials open the door for further comparative analysis. Hence as a start, four randomly picked pockets of the default Lennard-Jones potential (LJP) with a weight of 1 (**D**), a weight of 0.3 (**D03**) and the

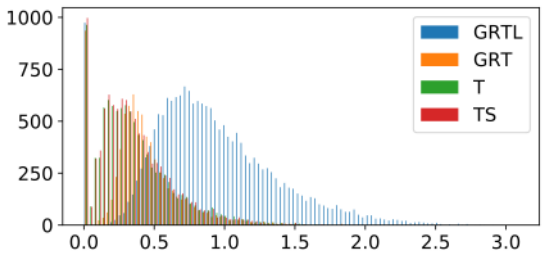
	GRT	GRTL	T	TS
GlobalRotationSamplingMaximum	0.1	0.2	0.05	0.05
GlobalRotationSamplingMinimum	-0.1	-0.2	-0.05	-0.05
GlobalRotationSamplingStepsize	0.1	0.1	0.1	0.1
GlobalTranslationSamplingMaximum	0.1	0.2	0.05	0.05
GlobalTranslationSamplingMinimum	-0.1	-0.2	-0.05	-0.05
GlobalTranslationSamplingStepsize	0.1	0.1	0.1	0.1
TorsionSamplingMaximum	0.1	0.1	0.1	0.01
TorsionSamplingMinimum	-0.1	-0.1	-0.1	-0.01
TorsionSamplingStepsize	0.05	0.05	0.05	0.05
MaxNumberTorsionWobblingPoses	30	30	100	100

(a) Ligand pertubation parametrization in the `GeohydeEvaluator` configuration file. 0.1 radiant is 5.7 degrees.

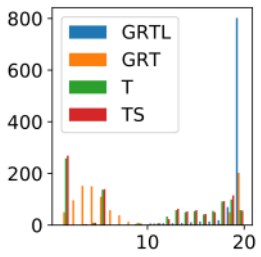
Weight	E	1	2	3	4
w_{sat}	3	3	3	3	3
w_{desolv}	0.5	0.5	0.5	0.5	0.5
w_{iLJ}	1.0	0.3	0.3	0.3	0.3
w_t	5	5	5	5	1
w_{rLJ}	0.5	0.05	0.1	1	0.5

(b) HYDE and GeoHYDE parametrizations used in this thesis compared to the sampling parametrizations.

Figure 5.5: Configuration for the molecule perturbation and GeoHYDE parametrization for the sampling experiments.

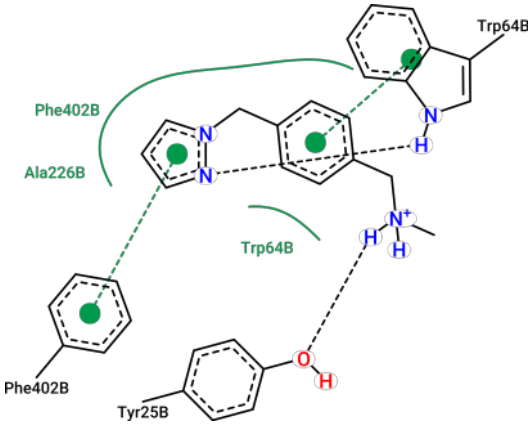


(a) RMSD spread per configuration over ProtFlex18$_{train}$.

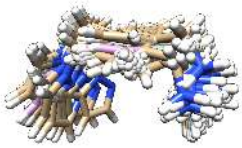


(b) Number of configurations per structure over ProtFlex18$_{train}$. At least one configuration per pocket was present.

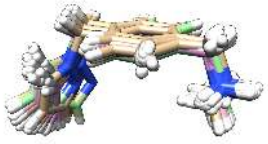
Figure 5.6: RMSD spread and sampled data set size of the ProtFlex18$_{train}$.



(a) Binding pocket of 3VR B 502 in 4wyd (PoseView)



(b) Initial poses



(c) Final poses

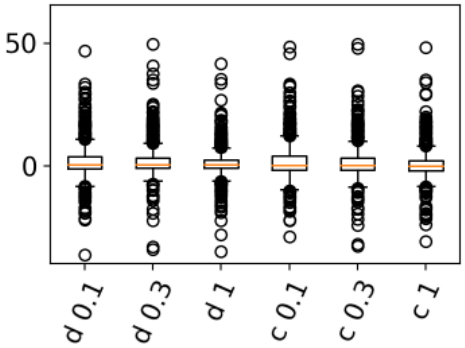
Figure 5.7: Overview of the configurations for 3VR B 502 in 4wyd created through GRTL, GRT, T, TS combined with the description of the pocket in 2D by PoseView.. While three configurations are present with respectively 20 configurations for the ligand, GRT has only resulted in four. The decomposition of the configurations can be found in Figure B.17.

'capped' LJP as the best performing LJP with a weight of 0.3 (**C03**) were analyzed. All of them show in the configuration **D** low performance with an $EDIA_m$ below 0.8 and RMSD above 0.5. Partial scores of 2zzd TLA C 4001, 5edb 5M8 A 201, 5d9y OGA A 2001 and 4c9o CAM A 423 are given in Table B.14 and B.15 and all pockets are documented in Figure 5.3 and B.26. In the case of 2zzd TLA C 4001, the geometric optimization mainly focuses on rotating parts of L-tartaric acid e.g. O3 out of plane to reduce the repulsive intramolecular LJP. The best result for the ligand interacting over two hydrogen bonds with arginine C 117 is for $EDIA_m$ the parametrization **D** and for RMSD **D03**. 6-chloranyl-2-methyl-4-phenyl-quinoline-3-carboxylic acid (5M8 A 201) in 5edb shows in contrast no changes in neither CTS nor intramolecular LJP. Instead, $GeoHYDE_d$ drives the geometric optimization in **D** to further improve its already attractive score resulting from the pyridine ring close to phenylalanine A 17. In the case of **D03**, the term is also scored as attractive but has little effect hence resulting in a pose with the best HYDE score, $EDIA_m$ and RMSD in comparison. In the case of N-oxalylglycine (OGA A 2001) in 5d9y clashes are detected to the metal in the pocket in $GeoHYDE_{desolv}$. Hence, in all variations of the experiment, $GeoHYDE_d$ is improved. **D** performs the worst in reducing $EDIA_m$ to 0.33 while **D03** keeps $EDIA_m$ at 0.82 dropping from 0.93. The last pocket to be compared is camphor (CAM A 423) in 4c9o. All three experiments result in a high quality hydrogen bond to the oxygen of tyrosine A 98 for which **D** creates the lowest RMSD of 0.77 and the best $EDIA_m$ of 0.59. In the other two cases, $GeoHYDE_s$ instead of $GeoHYDE_{desolv}$ strongly drives the optimization.

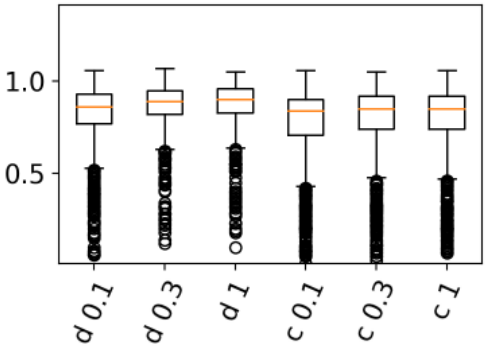
Overall, in three of four cases, the configuration **D03** performs the best. In two cases, either the intramolecular or intermolecular LJP integrated in $GeoHYDE_{desolv}$ detect clashing atoms in ambiguous situations. More evaluation and parametrization are needed so that GeoHYDE correctly assesses such tight configurations. Additionally, multiple pockets such as 2zzd TLA C 4001 show diverging algebraic signs between changes in $HYDE_s$, $HYDE_d$ and $GeoHYDE_s$ and $GeoHYDE_{desolv}$.

5.4 Results with Final Parametrization of GeoHYDE

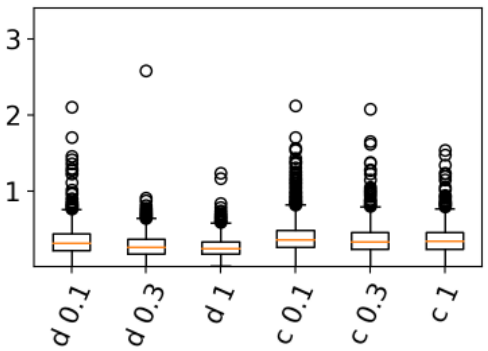
In the following, the $GeoHYDE_{final}$ parametrization is analyzed on the training and the validation data sets $ProtFlex18_{id}$ and $ProtFlex18_{od}$. Subsequently, the performance of the parametrization is compared between the optimization with a rigid pocket, a partially, and a fully flexible pocket. The section closes with an evaluation of $GeoHYDE_{final}$ on the CASF-2016.



(a) HYDE difference



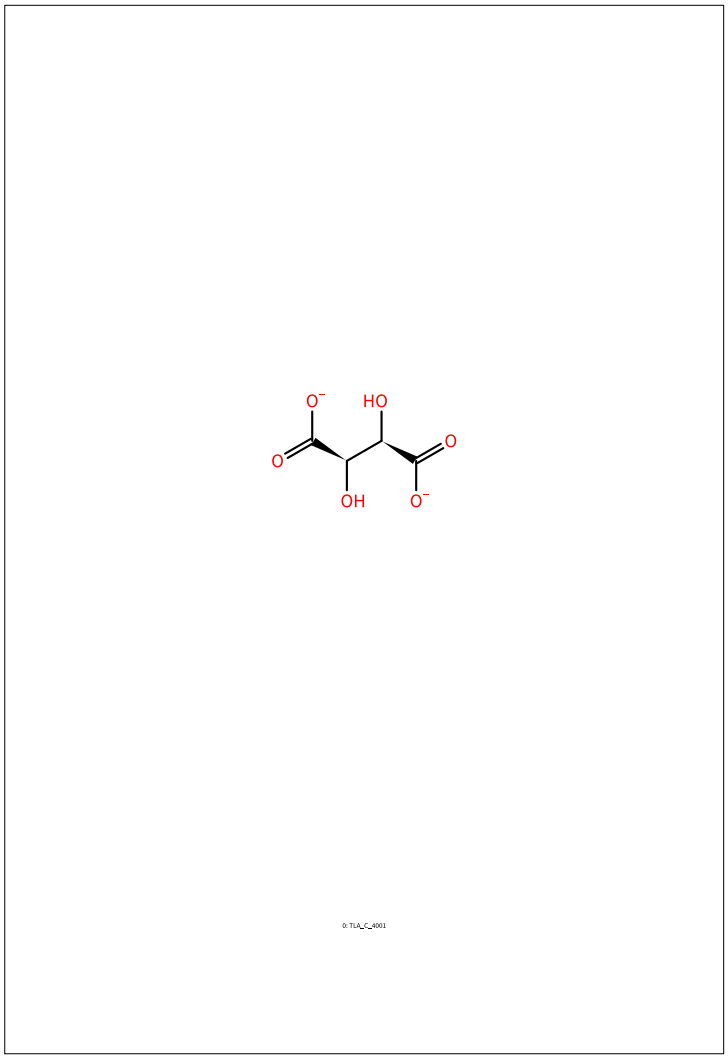
(b) Final $EDIA_m$

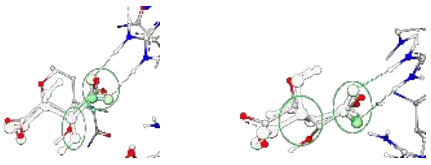


(c) Final RMSD

	Median		
LJP: w_{iLJ}	HYDE diff	$EDIA_m$	RMSD
d: 0.1	0.69	0.86	0.33
d: 0.3	0.73	0.89	0.28
d: 1	0.54	0.89	0.26
c: 0.1	0.44	0.84	0.37
c: 0.3	0.32	0.85	0.34
c: 1	-0.02	0.85	0.35

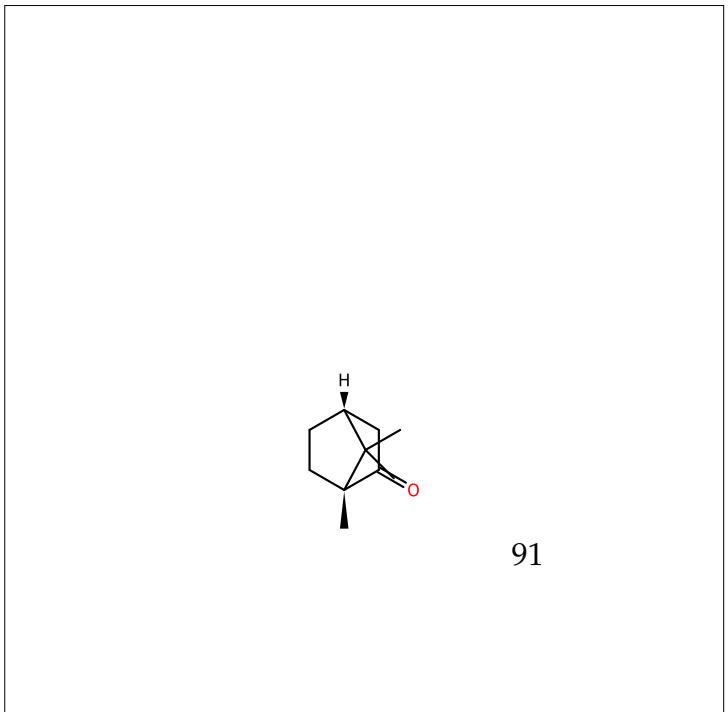
Figure 5.8: Comparison of default Lennard-Jones (d) with 'capped' Lennard-Jones potential (c) on $ProtFlex18_{train}$. Both score terms are tested over the same list of weights given in Table 5.4, the best performing ones of both are compared above.





2zzd TLA C 4001

O4, O41 interact with arginine C 11



5.4.1 Optimizing a Rigid Pocket With a Flexible Ligand

The overall goal of GeoHYDE is to guide a local geometric optimization towards the nearby HYDE score optimum without substantially changing the binding mode. Since HYDE prefers interaction geometries close to those in high quality crystal structures, ligand poses in data sets such as the ProtFlex18 should not be altered strongly. As first step, the degree of score alignments between GeoHYDE$_{ds}$ and HYDE was again examined (Section 5.3.3). Over the whole ProtFlex18 data set, the misalignment between GeoHYDE$_{ds}$ and HYDE score directions was slightly reduced by four structures in comparison to GeoHYDE$_{empirical}$ (see Table 5.2). For the three data sets, each best and worst pose in terms of GeoHYDE$_{sd}$ are reported in Figure 5.5, B.16 and in Table 5.6. Additionally, the pocket with the maximum and minimum change in its HYDE score are reported. All examples are presented with a 2D view of the ligand configuration and a three dimensional overlay of the initial with the final pocket. While in all best performing cases, GeoHYDE$_{sat}$ is the most improved partial score term, in two of four worst performing cases, the intramolecular Lennard-Jones term appears to be the driving force behind the optimization. Besides, in three cases, both GeoHYDE$_{d}$ and GeoHYDE$_{s}$ have diverging algebraic sings and in one case, GeoHYDE$_{d}$ disagrees with HYDE$_{d}$ on the direction of improvement.

Then, the ligand poses in the data sets ProtFlex18$_{train}$, ProtFlex18$_{id}$ and ProtFlex18$_{od}$ have been analyzed with the root mean square deviation to the original crystal structure as well as their initial and final EDIA$_{\mathrm{m}}$. For RMSD, a change of maximally 0.5 Å does not signify a substantial change. EDIA$_{\mathrm{m}}$ values on the other hand should not drop below 0.8. When analyzing the final ligand poses, the results can be divided along both cutoffs to split the data into four sections. The absolute number and percentage per section per data set can be found in Table 5.4 and Figure B.27. More information about e.g. HYDE score changes per data set can be found in Figure B.28. Initially, all ligands ahve an RMSD of zero and an EDIA$_{\mathrm{m}}$ in between 0.8 and 1.2. From the training data set over the in to the out domain test set, between 74 and 79 per cent of the ligand poses have an RMSD of maximally 0.5 and an EDIA$_{\mathrm{m}}$ of at least 0.8 after the geometric optimization. In 7 to 14% of the ligand poses, the two metrics agree in declaring them not close to the initial, crystallized pose anymore. They have an RMSD above 0.5 Å and an EDIA$_{\mathrm{m}}$ below 0.8. While only 9 to zero cases are determined of having an RMSD above 0.5 but still acceptably well supported by electron density (EDIA$_{\mathrm{m}}$), 15 to 8% of the

RMSD - $EDIA_m$	$ProtFlex18_{train}$	$ProtFlex18_{id}$	$ProtFlex18_{od}$
RMSD ≤0.5,$EDIA_m \geq 0.8$	787 (78.94%)	83 (74.77%)	79 (78.22%)
RMSD >0.5,$EDIA_m \geq 0.8$	9 (0.9%)	0 (0.0%)	0 (0.0%)
RMSD ≤0.5, $EDIA_m < 0.8$	126 (12.64%)	17 (15.32%)	8 (7.92%)
RMSD >0.5, $EDIA_m < 0.8$	75 (7.52%)	11 (9.91%)	14 (13.86%)

Table 5.4: RMSD - $EDIA_m$ correlation per quality segment over the three data sets $ProtFlex18_{train}$, $ProtFlex18_{id}$, $ProtFlex18_{od}$. Result visualization can be found in Figure B.27

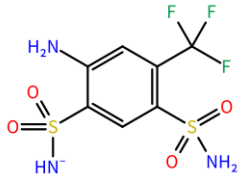
.

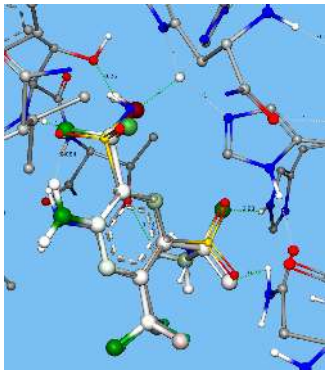
ligands report an RMSD close to the crystal pose but an $EDIA_m$ below 0.8. $EDIA_m$ has thus highlighted ligand poses, that are not supported by electron density but still close to the model coordinates. Additional examination reveals ligands with an RMSD of e.g. 0.4 Å but with $EDIA_m$ spanning from 0.32 to 0.7 (Figure B.18, 5.9, Table B.17). In the case of EXI A 902 in 4ugy with a final $EDIA_m$ of 0.32, an intermolecular hydrogen bond quality is reduced to increase the already negative $GeoHYDE_d$ score. For K66 A 1 in 3kxh with a final $EDIA_m$ of 0.41, the carboxylate group connected to the pyrimidine is shifted to in sum increase the quality of its interaction and further optimize $GeoHYDE_{desolv}$ and reduce the amount of intramolecular clash detected by the intramolecular Lennard-Jones potential. Again, $GeoHYDE_{desolv}$ and $GeoHYDE_{sat}$ do not fully agree with their corresponding HYDE score terms. The same pattern repeats itself in TD6 F 601 in 5eja with a final $EDIA_m$ of 0.48. Additionally, strong intramolecular clash in the ligand is removed through the optimization. Only for 1DC A 601 in 4l6z with a final $EDIA_m$ of 0.7 $GeoHYDE_{sat}$ and $HYDE_{sat}$ agree for the direction of the score improvement.
It is noteworthy, that the optimization is unconstrained but still the maximum RMSD is 2.6 Å with an RMSD median of 0.27, 0.28, and 0.28 Å for the data sets $ProtFlex18_{train}$, $ProtFlex18_{id}$ and $ProtFlex18_{od}$. In contrast, the to my knowledge only other gradient free published geometric optimization algorithm MinimuDS achieves an average RMSD of 0.53 Å on the PDBbind core set with a limited geometric optimization of maximally 2 Å RMSD.[72]

5.4.2 Results on CASF-2016

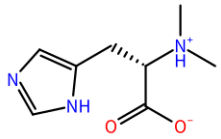
As CASF-2016 is an external validation set also used by others, the publication delivers results for 33 scoring functions combined with ΔSAS as the example for a simplistic scoring function.[75] $GeoHYDE_{empirical}$ and $GeoHYDE_{final}$ have been analyzed on the CASF-2016 in terms of scoring, ranking and docking ability. ΔSAS

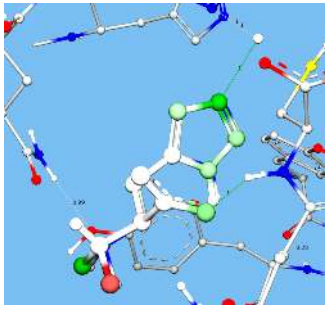
ProtFlex18$_{id}$





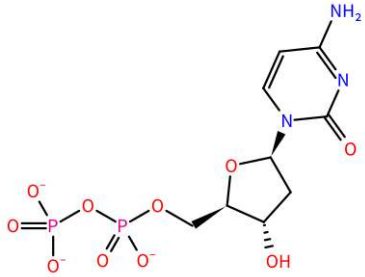
2pow I7C A 1000
HYDE$_{diff}$: -14,79
GeoHYDE$_{ds,diff}$: 11,59

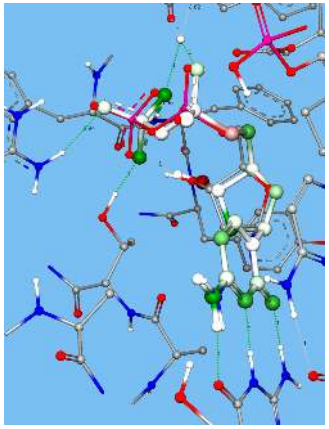




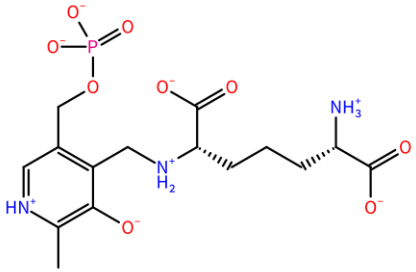
4x8d AVI B 502
HYDE$_{diff}$: 33,74
GeoHYDE$_{ds,diff}$: 43,54

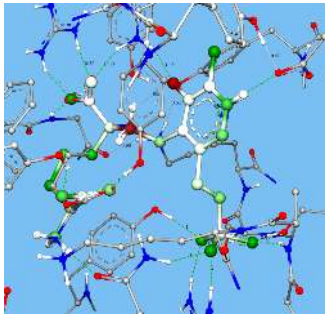
ProtFlex18$_{od}$





4qwb YYY A 401
HYDE$_{diff}$: 17,67
GeoHYDE$_{ds,diff}$: 46,07



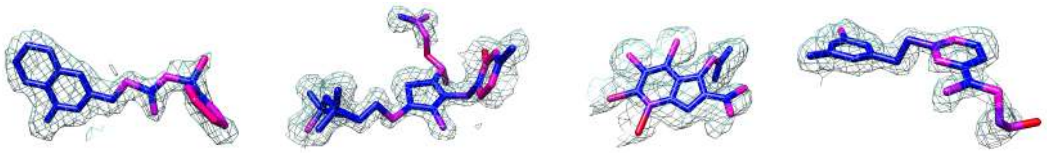


3ei6 PL4 A 434
HYDE$_{diff}$: -12,29
GeoHYDE$_{ds,diff}$: 42,91

Table 5.5: ProtFlex18$_{id}$ and ProtFlex18$_{od}$ single cases with the largest score improvement or worsening for GeoHYDE$_{sd}$

PDB, Mol Id	$EDIA_m$	RMSD	HYDE	$HYDE_s$	$HYDE_d$	GH_{ds}	GH_s	GH_d	CTS	LJ_{intra}	os	time
$ProtFlex18_{train}$												
1qxw M1C A 3001	0.89	0.0	-38.36	-11.93	-26.42	83.78	11.01	72.77	16.81	42.42	143.0	0.0
	0.14	0.88	-4.4	14.82	-19.22	-48.25	-17.3	-30.95	14.54	38.62	4.91	29.46
Diff initial - final	0.75	-0.88	**-33.96**	-26.75	-7.2	**132.02**	28.3	**103.72**	2.27	3.79	138.09	-29.46
4a6v IKY B 1264	0.89	0.0	28.46	55.86	-27.4	182.0	199.81	-17.82	0.0	60.52	242.52	0.0
	0.64	0.51	-21.26	2.87	-24.13	121.32	155.69	-34.38	0.0	60.52	181.84	17.79
Diff initial - final	0.25	-0.51	49.72	52.99	-3.27	60.68	**44.12**	16.56	0.0	0.0	60.68	-17.79
5gmz 6XU F 202	1.02	0.0	-69.62	3.55	-73.17	-3.64	35.88	-39.52	7.54	705.91	709.82	0.0
	0.28	0.66	-57.59	13.35	-70.94	31.47	56.03	-24.56	4.38	17.34	53.19	44.06
Diff initial - final	0.74	-0.66	**-12.03**	-9.8	-2.23	**-35.11**	-20.15	-14.96	3.16	**688.57**	656.63	-44.06
3ucd 2PG A 601	1.01	0.0	11.84	24.63	-12.79	180.57	191.73	-11.16	8.02	-3.87	184.73	0.0
	0.8	0.39	0.04	9.74	-9.7	20.04	104.35	-84.3	10.57	2.52	33.13	43.72
Diff initial - final	0.21	-0.39	11.8	14.89	-3.09	160.53	**87.38**	73.14	-2.54	-6.39	151.6	-43.72
$ProtFlex18_{id}$												
2pow I7C A 1000	0.94	0.0	-6.69	9.96	-16.65	16.2	67.52	-51.33	3.79	43.01	62.99	0.0
	0.89	0.24	8.1	23.52	-15.42	4.61	52.66	-48.05	8.43	1.63	14.67	19.37
Diff initial - final	0.05	-0.24	**-14.79**	-13.57	-1.23	11.59	14.87	-3.28	-4.64	**41.37**	48.32	-19.37
4x8d AVI B 502	0.96	0.0	0.29	11.28	-10.99	35.49	51.28	-15.79	8.1	-7.96	35.64	0.0
	0.92	0.22	-33.44	-22.81	-10.64	-8.05	6.9	-14.95	8.55	-8.14	-7.64	12.7
Diff initial - final	0.04	-0.22	33.74	34.09	-0.35	43.54	**44.38**	-0.84	-0.45	0.19	43.28	-12.7
$ProtFlex18_{od}$												
3ei6 PL4 A 434	0.95	0.0	-52.2	-16.9	-35.3	-56.52	66.08	-122.6	24.1	-15.11	-47.54	0.0
	0.91	0.21	-39.91	-7.32	-32.59	-99.44	53.79	-153.23	26.32	-21.01	-93.58	145.29
Diff initial - final	0.04	-0.21	**-12.29**	-9.58	-2.71	42.92	12.29	**30.63**	-2.22	5.89	46.04	-145.29
4qwb YYY A 401	1.02	0.0	-8.4	16.14	-24.54	33.92	108.11	-74.19	17.67	-23.66	27.93	0.0
	0.93	0.29	-26.07	-1.7	-24.38	-12.15	51.17	-63.31	17.58	-23.02	-17.59	41.45
Diff initial - final	0.09	-0.29	17.67	17.84	-0.17	46.07	**56.94**	-10.87	0.09	-0.64	45.52	-41.45

Table 5.6: Selected pockets of $ProtFlex18_{train}$, $ProtFlex18_{id}$, $ProtFlex18_{od}$ are given. For all data sets, the pocket with the maximum and minimum change in $GeoHYDE_{ds}$ is listed. Additionally, the pockets with the maximum and minimum change in HYDE score in the $ProtFlex18_{train}$ data set are given. All initial and final score terms after the optimization are combined with their difference per score term. The score term with the largest improvement per pocket is marked in bold.



4l6z 1DC A 601 EDIA$_m$: 0.7 | 5eja TD6 F 601 EDIA$_m$: 0.48 | 3kxh K66 A 1 EDIA$_m$: 0.41 | 4ugy EXI A 904 EDIA$_m$: 0.32

Figure 5.9: Ligand poses of ProtFlex18$_{train}$ in EDIA coloring with an RMSD of 0.4 but diverging EDIA$_m$. Changes in the ligands range from slightly tilting the pyridine ring in 1DC to fully moving the methyl from its original position in EXI. The full pockets are depicted in Figure B.18 and score terms can be found in Table B.17.

Scoring Power		cryst		opt	
		$r_{X,Y}$	$\sigma_{X,Y}$	$r_{X,Y}$	$\sigma_{X,Y}$
	c	0.466	1.90	0.483	1.90
GeoHYDE$_{empirical}$	o	0.505	1.86	0.482	1.89
GeoHYDE$_{final}$	o	0.506	1.86	0.496	1.86

Ranking Power		cryst			opt		
		r_s	τ	PI	r_s	τ	PI
	c	0.404	0.337	0.425	0.411	0.347	0.424
GeoHYDE$_{empirical}$	o	0.419	0.340	0.432	0.437	0.351	0.457
GeoHYDE$_{final}$	o	0.461	0.375	0.482	0.391	0.319	0.414

Table 5.7: Results of the CASF-2016 scoring and ranking benchmark are presented. The poses are subdivided into those from the crystal structure (cryst) and those, optimized by the CASF team (opt). Results are given for three types of HYDE scoring: without any optimization with GeoHYDE (c), after optimization with GeoHYDE$_{empirical}$ and after optimization with GeoHYDE$_{final}$ (o).

Docking Power[%]		Top 1	Top 2	Top 3
	c	70.2	79.2	84.2
GeoHYDE$_{empirical}$	o	68.4	79.6	84.6
GeoHYDE$_{final}$	o	66.0	80.0	86.0

Docking Power	r_s	$[0-2]$	$[0-3]$	$[0-4]$	$[0-5]$	$[0-6]$
	c	0.523	0.541	0.510	0.470	0.441
GeoHYDE$_{empirical}$	o	0.479	0.523	0.526	0.510	0.496
GeoHYDE$_{final}$	o	0.486	0.524	0.521	0.507	0.489

	SP	$[0-7]$	$[0-8]$	$[0-9]$	$[0-10]$
	c	0.410	0.381	0.366	0.344
GeoHYDE$_{empirical}$	o	0.480	0.456	0.438	0.417
GeoHYDE$_{final}$	o	0.472	0.451	0.434	0.413

Table 5.8: Results of the CASF-2016 docking benchmark are presented. The performance of HYDE on the poses without any optimization with GeoHYDE (c), after optimization with GeoHYDE$_{empirical}$ and after optimization with GeoHYDE$_{final}$ (o) is given in the top one to three as well as the Spearman correlation coefficient of the funnel shape analysis over various RMSD intervals in Ångstrom (SP).

is ranked in the top third of the scoring functions for the scoring power analysis. In comparison, HYDE without optimization performs on the crystallized and the by the CASF team preoptimized poses with Pearson correlation coefficients of 0.47 and 0.48. After the geometric optimization with GeoHYDE$_{empirical}$, the correlation coefficients of HYDE on the same data changes to 0.51 and 0.48. The optimization with GeoHYDE$_{final}$ results in a slight improvement to 0.51 and 0.50 (Table 5.7). The results place HYDE in the middle of the 34 tested scoring functions, performing less well for example as ΔSAS.

For the ranking power analysis, ΔSAS is again in the first third while both variations of GeoHYDE with subsequent scoring with HYDE result in a performance in the middle of the field. Optimizing the crystallized poses with GeoHYDE$_{final}$ results in the best correlation coefficients (r_s : 0.46, τ : 0.38 and PI : 0.48) which positions HYDE in the lower third of the 34 tested scoring functions (Table 5.7). Scoring the by the CASF team preoptimized poses which are subsequently optimized by GeoHYDE$_{empirical}$ performs better than optimizing with GeoHYDE$_{final}$ but still does not move HYDE out from the lower third.

The docking power analysis evaluates the one, two and three top most ranking poses if the original ligand was found. While HYDE performs in the top third segment already without optimization with 70, 79 and 84%, GeoHYDE$_{final}$ can still

increase the results for the two and three top most ranking poses to 80 and 86% while reducing the results in the best ranked pose to 66% (Table 5.8). The analysis of the funnel shape with the help of the Spearman correlation coefficient r_s shows HYDE with either GeoHYDE parametrization to be weaker in the narrow RMSD interval of 0 to 2 Å and to 3 Å than the unoptimized pockets scored with HYDE. The subsequent correlation coefficients increase above those of the unoptimized pockets but still keep HYDE in the midfield of the 34 evaluated scoring functions (Table 5.8). In contrast, ΔSAS is the second to last scoring function in the overall docking power test.

5.4.3 Optimizing a Pocket With Side Chain and Ligand Flexibility

The newly implemented ability to geometrically optimize not just the ligand but also specific side chains in the active site was evaluated on the 546 flexible pockets in ProtFlex18$_{train}$, 62 pockets in ProtFlex18$_{id}$, and 23 pockets in ProtFlex18$_{od}$. GeoHYDE$_{prot}$ (Equation 5.4) with the weights $w_{tp} = 10$ and $w_{rLJp} = 1$ was used as scoring function in the optimization. The as flexible determined proteins in each data set were optimized in using a flexible ligand in a rigid pocket (**R**) and the fully flexible pocket (**F**) and the flexible residues in the binding pocket of the ligand previously determined by SIENA (**P**) (Chapter 3.2.1). The resulting poses were analyzed based on the ligand's RMSD to the crystal pose, the final EDIA$_m$ and HYDE scores. Overall, it can be said that with increasing flexibility, median HYDE scores improve and RMSD and EDIA$_m$ slightly decrease over all three data sets. For both types of protein flexibility, RMSD and EDIA$_m$ correlation values are between 0.61 to 0.77 for **P** dropping to 0.36 to 0.57 (**F**) even though the median for both values only differ marginally up to 0.04 (Table 5.9 and Figure B.29 - B.31). In both cases, HYDE scores strongly correlate from 0.94 dropping to 0.89 for **F**. HYDE median scores improve for all data sets letting them range from $[-26.5, -29.2]$ to $[-27.6, -30.7]$ (**P**) and even more for **F** to $[-28.5, -32.2]$ kJ. The dropping correlation of the two metrics suggest that in a structural view, **R** convergences in poses different to **P** and **F** while the HYDE scores increase in a similar way. One such case is 1xes with 3IO A 2000 where the HYDE score improves from -27.84 to −30.65 kJ while only resulting in an RMSD of 0.31 Å and an EDIA$_m$ dropping from 0.89 to 0.71. A score change of 2 kJ even though the RMSD is just 0.31 demonstrate the high sensitivity of HYDE for slight changes in the pocket's geometry. Increasing the flexibility of the pocket also increases the number of outliers (Table B.19, B.20). The majority of all outliers report an improvement beyond the respective RMSE

(12.3 to 4.4% in ProtFlex18$_{train}$) and for **F** of 11.5% and 5.9%. 4b4v L34 B 2001 as an outlier presenting the minimum HYDE score improvement for both types of optimization with protein flexibility from the ProtFlex18$_{id}$ shows only a substantial change in GeoHYDE$_{desolv}$ and the protein intramolecular clash score LJ$_{ip}$. The latter causes the movement of Arginine B 8 from an EDIA$_{\mathrm{m}}$ of 0.82 to 0.51 only in **F** even though it is also flexible in **P**. 4qxc OGA A 600 of the ProtFlex18$_{od}$ on the other hand shows the best HYDE score improvement in both flexibility optimizations but does not move the by SIENA determined flexible residues Met A 11 or VAL A 286. Instead, in **F** MLY E 36 is moved and LJ$_{ip}$ reduced (Table B.21). Computation time increases four (**P**) to 15 times (**F**) when optimizing with flexible residue side chains. Further examination revealed that in the case of **P** and especially for **F**, the termination criteria at 10,000 evaluation steps and not any of the termination criteria for convergence of the optimization function was relevant for finishing the computation (Figure 5.10). Such an example is 4qxc with needing 666 steps for **R**, 896 for **P** and terminating at 10,000 steps in **F**. Hence geometrically optimizing a fully flexible protein-ligand pocket in the current set up might demand even more computation time. The last topic to mention is the offset of around 30,000,000 units for the intramolecular Lennard-Jones potential for the protein consistently through the three data sets (see Table B.21). Also, the position of Arginine B 8 in 1xes was modified towards lowering the intramolecular LJP of the protein even though the residue is not relevant for the binding pocket. This shows the need to adequately assess the relevance of each residue for the optimization as well as the shape of the LJP itself to result in a meaningful value. In summary, the evaluation on ProtFlex18 shows promising results with high computational costs and the need for further work.

5.5 Conclusion

In this chapter, GeoHYDE as the objective function to geometrically optimize a pocket in accordance to HYDE was evaluated on the ProtFlex18 data set. For state of the art parameter tuning and subsequent evaluation, ProFlex18 with its 2386 pockets was split into three datasets of 997 pockets in ProtFlex18$_{train}$, 112 in ProtFlex18$_{id}$, and 101 in ProtFlex18$_{od}$. As first step, multiple gradient free optimization algorithms in the software package `NLopt` were tested for their performance and run time requirements with the parametrization GeoHYDE$_{empirical}$. Hence, BOBYQA was selected to be the fastest and in terms of GeoHYDE scores well performing

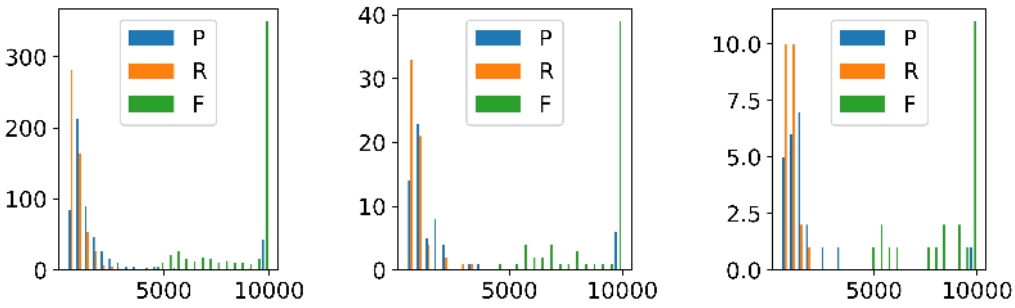
(a) Comparison between poses derived through optimization with GeoHYDE (**R**) against those with GeoHYDE$_{prot}$ with fully flexible residues (**F**). The initial median values are listed in the column marked with I. While GeoHYDE$_{prot}$ shows an improvement in HYDE scores, optimization time increased at least 15 fold

Data set (size) Metric	Median_I	Median_R	Median_F	$r_{X,Y}$	p value	RMSE
ProtFlex18$_{train}$ (546)						
RMSD	0.0	0.27	0.32	0.44	0	0.16
EDIA_m	0.98	0.89	0.87	0.57	0	0.14
HYDE	-24.49	-26.58	-30.63	0.90	0	6.51
Time (s)	0.0	25.60	603.0	0.37	0	
ProtFlex18$_{id}$ (62)						
RMSD	0.0	0.25	0.28	0.49	0	0.12
EDIA_m	0.97	0.90	0.87	0.57	0	0.11
HYDE	-22.70	-26.49	-28.47	0.89	0	7.4
Time (s)	0.0	25.13	617.07	0.56	0	
ProtFlex18$_{od}$ (23)						
RMSD	0.0	0.23	0.30	0.36	0.09	0.14
EDIA_m	0.98	0.92	0.88	0.42	0.05	0.1
HYDE	-28.17	-29.24	-32.20	0.94	0	4.08
Time (s)	0.0	30.58	575.42	0.56	0	

(b) Comparison between poses derived through optimization with GeoHYDE (**R**) against those with GeoHYDE$_{prot}$ with limited flexible residues (**P**). The initial median values are listed in the column marked with I. While GeoHYDE$_{prot}$ shows an improvement in HYDE scores, optimization time increased four times.

Data set (size) Metric	Median_I	Median_R	Median_P	$r_{X,Y}$	p value	RMSE
ProtFlex18$_{train}$ (546)						
RMSD [Å]	0.0	0.27	0.253	0.67	0.0	0.11
EDIA_m	0.98	0.89	0.9	0.68	0	0.1
HYDE [kJ]	-24.49	-26.58	-28.25	0.96	0	4.14
Time [s]	0.0	24.71	94.07	0.18	0	
ProtFlex18$_{id}$ (62)						
RMSD [Å]	0.0	0.25	0.27	0.71	0	0.1
EDIA_m	0.97	0.9	0.90	0.77	0	0.11
HYDE [kJ]	-22.70	-26.50	-28.50	0.94	0	6.33
Time [s]	0.0	25.78	95.73	0.39	0	
ProtFlex18$_{od}$ (23)						
RMSD [Å]	0.0	0.23	0.23	0.61	0	0.1
EDIA_m	0.98	0.92	0.92	0.73	0	0.07
HYDE [kJ]	-28.17	-29.24	-30.67	0.98	0	2.76
Time [s]	0.0	30.34	102.428	0.40	0.06	

Table 5.9: For the three ProtFlex18 data sets with actual flexible residues of the theoretically possible 1164 pockets, medians with Pearson correlation coefficient r and p value are given.



(a) ProtFlex18$_{train}$ (b) ProtFlex18$_{id}$ (c) ProtFlex18$_{od}$

Figure 5.10: Number of steps reached per data set for the optimization with just a flexible ligand (R), flexible ligand and selected flexible residue side chains (P) and flexible ligand and fully flexible side chains in the pocket (F).

algorithm. Subsequently, GeoHYDE score terms were checked in the unoptimized pockets of ProtFlex18$_{train}$ to check if the dataset is also for GeoHYDE high quality. 231 pockets with problems with protons were found. The other noticeable problem was with tightly packed ligands marked as clashing. An additional analysis showed misaligned score development between HYDE and GeoHYDE terms. 29 % for the saturation terms and 51 % of misaligned scores for GeoHYDE$_{desolv}$ versus HYDE$_{desolv}$ show a substantial misalignment.

As second stage in the analysis, a parameter search for the weights in GeoHYDE was run over ProtFlex18$_{train}$. Results on RMSD, HYDE score difference and final EDIA$_{\mathrm{m}}$ have been compared with the help of the Mann-Whitney-Wilcoxon Rank Sum test. None of the partial terms could be removed from GeoHYDE but also a strong overweighting of each term did not contribute positively. Besides their apparently necessary existence and having weights in between 0.3 and ten, only w_{iLJ} showed the need to be specifically adjusted from 1.0 to 0.3. Further tests have been conducted in comparing the GeoHYDE$_{desolv}$ behavior with that of an purely repulsive Lennard-Jones potential. Combined with the single case analysis, the attractive Lennard-Jones potential shows its needfulness. But again, densely packed ligands show an inappropriately configured intramolecular Lennard-Jones potential. Further tests have been conducted on the weight w_t for the Continuous Torsion Score and the weight w_{rLJ} for the intramolecular Lennard-Jones potential. But as both weights are safe guards in place to protect against unusual distortion, the ligand per pocket was perturbed until an RMSD of 2.5 Å with four sampling strategies. The MWW test did not identify substantial changes nonetheless. As

result, GeoHYDE$_{final}$ was derived.

With GeoHYDE$_{final}$, the performance of GeoHYDE was analyzed on the in and out of domain tests sets ProtFlex18$_{id}$, ProtFlex18$_{od}$. Only a slight improvement in the alignment between GeoHYDE and HYDE score terms was achieved. Overall 74 to 79 % of the pockets result in a ligand configuration with an EDIA$_{\mathrm{m}}$ of at least 0.8 and an absolute coordinate deviation of maximally 0.5 Å. The median deflection over the three data sets lies at 0.27 to 0.28 Å. Furthermore, pockets were identified where the ligand deviates less than 0.5 Å from its crystallized position but shows a strong drop in its electron density coverage estimated with EDIA$_{\mathrm{m}}$. GeoHYDE$_{final}$ was then tested on the external validation data set CASF-2016. HYDE before and after optimization performed comparatively in the middle third of of all tested 34 scoring functions for the scoring benchmark. In the ranking benchmark, HYDE unoptimized and optimized with GeoHYDE performed in the lower third. The scoring function passed in the middle range for the docking test. As last test, the newly integrated side chain optimization was then tested on the flexible pockets over the three data sets. In general terms, HYDE scores improved and computation time increased with increasing flexibility.

Overall, GeoHYDE performs well on ProtFlex18 and shows its ability to keep crystal structures close to their original ones while suggesting an improved docking performance for HYDE. But the validation scenarios have also repeatedly highlighted three areas for which future work is necessary. As a problem quite specific for working with approximative functions, studies about partial score misalignments should be integrated into the test consensus in the future. A connected area of great concern is the behavior of the inter- and intramolecular Lennard-Jones potentials. In most of the analysis, a number of outliers showed questionable assessment of the situation by the LJP. This may have assisted in the substantial score misalignment between GeoHYDE and HYDE. As the last problem, computation time needs to be discussed. While BOBYQA needs in median 26 s for the optimization of a flexible ligand within a rigid pocket, the run time with protein flexibility increases at least four times. Since BOBYQA is a sequential algorithm, speed improvements can only be achieved in switching the calculation of GeoHYDE from an absolute to an incremental approach in the future. Since in many cases BOBYQA only proposes changes in a small set of parameters, areas unchanged between evaluation steps may contribute an identical score. Leveraging them may result in computational speed up. After recently finalizing the interaction weighting

scheme in HYDE it may also be possible now to develop an analytical gradient to allow the optimization of GeoHYDE with the BFGS.

With the help of the large and highly diverse data set ProtFlex18, subsequent work should be able to tackle all of the aforementioned problems towards objectively quantifiable improvements.

Chapter 6

Conclusion and Future Directions

This thesis has resulted in improvements in four areas of computational drug design resulting in establishing a sound benchmark routine for GeoHYDE. Firstly, EDIA and $EDIA_m$ were developed to asses the agreement between model and electron density for any element in the periodic table. The metrics were subsequently used to release the first of its kind configurable tool `StructureProfiler` which comprises all state of the art quality checks for protein structures. Thereby, the ProtFlex18 data set was extracted from structures deposited in the PDB. It consists of 2386 pockets which makes it around ten times larger than any other validation data set currently in use. Updates in the Torsion Library were introduced, such as automatically resorting torsion rules with SMARTScompare with subsequent validation with the help of the tool `TorsionPatternMiner`. The Continuous Torsion Score was developed based on the Torsion Library and integrated into GeoHYDE. At last, the objective function GeoHYDE for the optimization towards the interaction model of HYDE was parameterized and evaluated on the optimization of flexible ligands as well as flexible ligands in a flexible pocket. For external comparison, its performance on the CASF-2016 was also analyzed.

EDIA and $EDIA_m$ have shown their usefulness through numerous publications beyond this thesis. It is expected that `StructureProfiler` with the ability to generate benchmark data sets to the liking of the user will have a similar impact in the future. It would be beneficial to be able to automatically annotate high quality protein-ligand complexes with binding affinity if possible to further open the path towards data sets applicable in machine learning.

Future directions for the Torsion Library have been already extensively discussed in Chapter 4. GeoHYDE also has a number of points that should be pursued in the future. Overall, the step width and termination criteria of GeoHYDE when

being optimized by BOBYQA should be evaluated further. The funnel shape of the hyperplane created by GeoHYDE in the RMSD interval of zero to three Å also call for attention. One strategy could be to examine the partial score misalignment between GeoHYDE and HYDE as well as the objective parametrization of parts of the Lennard-Jones potentials. Finally, the thesis has evaluated a first version of GeoHYDE also optimizing flexible side chains. In the future, weight parametrization tests should be conducted and strategies for speed up considered. It may also be wise to change from the CTS to a rotamer based approach for estimating the likeliness of torsion angles on the protein side.

Bibliography

[1] M. Adrian, J. Dubochet, J. Lepault, and A. W. McDowall. Cryo-electron microscopy of viruses. *Nature*, 308(5954):32–36, 1984.

[2] P. V. Afonine, R. W. Grosse-Kunstleve, N. Echols, J. J. Headd, N. W. Moriarty, M. Mustyakimov, T. C. Terwilliger, A. Urzhumtsev, P. H. Zwart, and P. D. Adams. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica Section D*, 68(4):352–367, 2012.

[3] J. C. Baber, D. C. Thompson, J. B. Cross, and C. Humblet. GARD: a Generally Applicable Replacement for RMSD. *Journal of Chemical Information and Modeling*, 49(8):1889–1900, 2009.

[4] S. Bietz and M. Rarey. SIENA: Efficient Compilation of Selective Protein Binding Site Ensembles. *Journal of Chemical Information and Modeling*, 56(1):248–59, 2016.

[5] S. Bietz, S. Urbaczek, B. Schulz, and M. Rarey. Protoss: A holistic approach to predict tautomers and protonation states in protein-ligand complexes. *Journal of Cheminformatics*, 6(12):1–12, 2014.

[6] M. J. Box. A New Method of Constrained Optimization and a Comparison With Other Methods. *The Computer Journal*, 8(1):42–52, 1965.

[7] R. P. Brent. *Algorithms for minimization without derivatives*. Prentice-Hall, Englewood Cliffs, NJ, 1973.

[8] C. Chang, T. Skarina, O. Kagan, A. Savchenko, A. Edwards, and A. Joachimiak. Crystal structure of 3-HSA hydroxylase, oxygenase from Rhodococcus sp. RHA1. *to be published*, 2007.

[9] I. Daylight Chemical Information Systems. SMARTS. `https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html`, 2020.

[10] J. C. De Roode, T. Lefevre, and M. D. Hunter. Self-medication in animals. *Science*, 340(6129):150–151, 2013.

[11] M. C. Deller and B. Rupp. Models of protein-ligand crystal structures: trust, but verify. *Journal of Computer-Aided Molecular Design*, 29(9):817–836, 2015.

[12] B. J. Desai, B. M. K. Wood, A. A. Fedorov, E. V. Fedorov, B. Goryanova, T. L. Amyes, J. P. Richard, S. C. Almo, and J. A. Gerlt. Conformational changes in orotidine 5'-monophosphate decarboxylase: A structure-based explanation for how the 5-phosphate group activates the enzyme. *Biochemistry*, 51(43):8665–8678, 2012.

[13] M. N. M. Drwal, G. Bret, C. Jacquemard, E. Kellenberger, C. Perez, J. Desaphy, C. Jacquemard, J. Desaphy, and E. Kellenberger. Structural Insights on Fragment Binding Mode Conservation. *Journal of Medicinal Chemistry*, 61(14):5963–5973, 2018.

[14] L. F. T. Eyck. Efficient structure-factor calculation for large molecules by the fast Fourier transform. *Acta Crystallographica Section A*, 33(3):486–492, 1977.

[15] N. Foloppe, L. M. Fisher, R. Howes, P. Kierstan, A. Potter, A. G. S. Robertson, and A. E. Surgenor. Structure-based design of novel Chk1 inhibitors: Insights into hydrogen bonding and protein-ligand affinity. *Journal of Medicinal Chemistry*, 48(13):4332–4345, 2005.

[16] N.-O. Friedrich, A. Meyder, C. de Bruyn Kops, K. Sommer, F. Flachsenberg, M. Rarey, and J. Kirchmair. High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators. *Journal of Chemical Information and Modeling*, 57(3):529 – 539, 2017.

[17] V. Garlatti, N. Belloy, L. Martin, M. Lacroix, M. Matsushita, Y. Endo, T. Fujita, J. C. Fontecilla-Camps, G. J. Arlaud, N. M. Thielens, C. Gaboriaud, J. Arlaud, and N. M. Thielens. Structural insights into the innate immune recognition specificities of L- and H-ficolins. *EMBO Journal*, 26(2):623–633, 2007.

[18] J. Goto, R. Kataoka, and N. Hirayama. Ph4Dock: pharmacophore-based protein-ligand docking. *Journal of Medicinal Chemistry*, 47(27):6804–6811, 2004.

[19] W. Guba, A. Meyder, M. Rarey, and J. Hert. Torsion Library Reloaded: A New Version of Expert-Derived SMARTS Rules for Assessing Conformations

of Small Molecules. *Journal of Chemical Information and Modeling*, 56(1):1–5, 2016.

[20] E. Harder, W. Damm, J. Maple, C. Wu, M. Reboul, J. Y. Xiang, L. Wang, D. Lupyan, M. K. Dahlgren, J. L. Knight, J. W. Kaus, D. S. Cerutti, G. Krilov, W. L. Jorgensen, R. Abel, and R. A. Friesner. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *Journal of Chemical Theory and Computation*, 12(1):281–296, 2016.

[21] M. J. Hartshorn, M. L. Verdonk, G. Chessari, S. C. Brewerton, W. T. M. Mooij, P. N. Mortenson, and C. W. Murray. Diverse, high-quality test set for the validation of protein-ligand docking performance. *Journal of Medicinal Chemistry*, 50(4):726–741, 2007.

[22] P. C. D. Hawkins, B. P. Kelley, and G. L. Warren. The application of statistical methods to cognate docking: A path forward? *Journal of Chemical Information and Modeling*, 54(5):1339–1355, 2014.

[23] T. Inhester. *Mining of Interaction Geometries in Collections of Protein Structures*. PhD dissertation, Universität Hamburg, 2017.

[24] Intel. Intel Threading Building Blocks. `https://software.intel.com/en-us/tbb`, 2020.

[25] S. G. Johnson. The NLopt nonlinear-optimization package. `http://github.com/stevengj/nlopt`, 2020.

[26] T. A. Jones, J. Y. Zou, S. W. Cowan, and M. Kjeldgaard. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallographica Section A*, 47:110–119, 1991.

[27] L. Kalinowsky, J. Weber, S. Balasupramaniam, K. Baumann, and E. Proschak. A Diverse Benchmark Based on 3D Matched Molecular Pairs for Validating Scoring Functions. *ACS Omega*, 3(5):5704–5714, 2018.

[28] Z. M. Khan, Y. Liu, U. Neu, M. Gilbert, B. Ehlers, T. Feizi, and T. Stehle. Crystallographic and Glycan Microarray Analysis of Human Polyomavirus 9 VP1 Identifies N-Glycolyl Neuraminic Acid as a Receptor Candidate. *Journal of Virology*, 88(11):6100–6111, 2014.

[29] G. Klebe. *Wirkstoffdesign*. Spektrum Akademischer Verlag, 2 edition, 2009.

[30] G. J. Kleywegt and T. A. Jones. xdlMAPMAN and xdlDATAMAN – Programs for Reformatting, Analysis and Manipulation of Biomacromolecular Electron-Density Maps and Reflection Data Sets. *Acta Crystallographica Section D*, 52(4):826–828, 1996.

[31] V. Koenig, A. Pfeil, G. Heinrich, G. Braus, and T. Schneider. Crystal Structure of the Double Complex of the Tyrosine Sensitive Dahp Synthase from Yeast. *to be published*, 2004.

[32] W. Kohn. Nobel Lecture: Electronic structure of matter—wave functions and density functionals. *Reviews of Modern Physics*, 71(5):1253–1266, 1999.

[33] D. Liebschner, P. V. Afonine, M. L. Baker, G. Bunkoczi, V. B. Chen, T. I. Croll, B. Hintze, L. W. Hung, S. Jain, A. J. McCoy, N. W. Moriarty, R. D. Oeffner, B. K. Poon, M. G. Prisant, R. J. Read, J. S. Richardson, D. C. Richardson, M. D. Sammito, O. V. Sobolev, D. H. Stockwell, T. C. Terwilliger, A. G. Urzhumtsev, L. L. Videau, C. J. Williams, and P. D. Adams. Macromolecular structure determination using X-rays, neutrons and electrons: Recent developments in Phenix. *Acta Crystallographica Section D: Structural Biology*, 75:861–877, 2019.

[34] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989.

[35] P. McCabe, O. Korb, and J. Cole. Kernel density estimation applied to bond length, bond angle, and torsion angle distributions. *Journal of Chemical Information and Modeling*, 54(5):1284–8, 2014.

[36] A. Meyder, S. Kampen, R. Fährrolfes, F. Flachsenberg, J. Sieg, N. O. Friedrich, and M. Rarey. StructureProfiler: An all-in-one tool for 3D protein structure profiling. *Bioinformatics*, 35(5):874–876, 2019.

[37] A. Meyder, E. Nittinger, G. Lange, R. Klein, and M. Rarey. Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-ray Structures. *Journal of Chemical Information and Modeling*, 57(10):2437–2447, 2017.

[38] V. Modi and R. L. Dunbrack. Defining a new nomenclature for the structures of active and inactive kinases. *Proceedings of the National Academy of Sciences of the United States of America*, 116(14):6818–6827, 2019.

[39] J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 1965.

[40] E. Nittinger. *Water Molecules Within the HYDE Scoring Function: Placement, Optimization, and Energetic Contributions*. PhD dissertation, Universität Hamburg, 2018.

[41] E. Nittinger, F. Flachsenberg, S. Bietz, G. Lange, R. Klein, and M. Rarey. Placement of Water Molecules in Protein Structures: From Large-Scale Evaluations to Single-Case Examples. *Journal of Chemical Information and Modeling*, 58(8):1625–1637, 2018.

[42] E. Nittinger, T. Inhester, S. Bietz, A. Meyder, K. T. Schomburg, G. Lange, R. Klein, and M. Rarey. Large-Scale Analysis of Hydrogen Bond Interaction Patterns in Protein-Ligand Interfaces. *Journal of Medicinal Chemistry*, 60(10):4245–4257, 2017.

[43] E. Nittinger, N. Schneider, G. Lange, and M. Rarey. Evidence of water molecules - a statistical evaluation of water molecules based on electron density. *Journal of Chemical Information and Modeling*, 55(4):771–783, 2015.

[44] J. Nocedal and S. Wright. *Numerical optimization, series in operations research and financial engineering*. Springer, 2006.

[45] V. Notenboom, S. J. Williams, R. Hoos, S. G. Withers, and D. R. Rose. Detailed Structural Analysis of Glycosidase/Inhibitor Interactions: Complexes of Cex from Cellulomonas fimi with Xylobiose-Derived Aza-Sugars. *Biochemistry*, 39(38):11553–11563, 2000.

[46] E. Padlan, G. Cohen, and D. Davies. Refined Crystal Structure of the Mc/Pc603 Fab-Phosphocholine Complex at 3.1 Angstroms Resolution. *to be published*, 1984.

[47] S. Patterson, M. S. Alphey, D. C. Jones, E. J. Shanks, I. P. Street, J. A. Frearson, P. G. Wyatt, I. H. Gilbert, and A. H. Fairlamb. Dihydroquinazolines as a novel class of Trypanosoma brucei trypanothione reductase inhibitors: Discovery, synthesis, and characterization of their binding mode by protein crystallography. *Journal of Medicinal Chemistry*, 54(19):6514–6530, 2011.

[48] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht. How to improve RD productivity: The pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9(3):203–214, 2010.

[49] N. M. Pearce, T. Krojer, and F. Von Delft. Proper modelling of ligand binding requires an ensemble of bound and unbound states. *Acta Crystallographica Section D: Structural Biology*, 73:256–266, 2017.

[50] D. A. Pearlman and P. S. Charifson. Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 MAP kinase protein system. *Journal of Medicinal Chemistry*, 44(21):3417–3423, 2001.

[51] J. A. Pople. Nobel Lecture: Quantum chemical models. *Reviews of Modern Physics*, 71(5):1267–1274, 1999.

[52] M. J. D. Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2):155–162, 1964.

[53] M. J. D. Powell. A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in Optimization and Numerical Analysis*, pages 51–67. 1994.

[54] M. J. D. Powell. The NEWUOA software for unconstrained optimization without derivatives. *DAMTP*, 8:255–297, 2006.

[55] M. J. D. Powell. A view of algorithms for optimization without derivatives. *DAMTP*, 3:1–12, 2007.

[56] M. J. D. Powell. The BOBYQA algorithm for bound constrained optimization without derivatives. *DAMTP*, 6:39, 2009.

[57] I. I. Rabi, J. R. Zacharias, S. Millman, and P. Kusch. A new method of measuring nuclear magnetic moment. *Physical Review*, 53(4):318, 1938.

[58] G. Rhodes. *Crystallography Made Crystal Clear 3rd Edition*. Academic Press, 3 edition, 2006.

[59] J. A. Richardson and J. L. Kuester. The Complex Method for Constrained Optimization. *Communications of the ACM*, 16(8):487–489, 1973.

[60] L. M. Rios and N. V. Sahinidis. Derivative-free optimization: A review of algorithms and comparison of software implementations. In *Journal of Global Optimization*, volume 56, pages 1247–1293, 2013.

[61] J. Sadowski and J. Boström. MIMUMBA revisited: Torsion angle rules for conformer generation derived from X-ray structures. *Journal of Chemical Information and Modeling*, 46(6):2305–2309, 2006.

[62] C. Schärfer, T. Schulz-Gasch, H. C. Ehrlich, W. Guba, M. Rarey, and M. Stahl. Torsion angle preferences in druglike chemical space: A comprehensive guide. *Journal of Medicinal Chemistry*, 56(5):2016–2028, 2013.

[63] R. Schmidt, E. S. Ehmki, F. Ohm, H. C. Ehrlich, A. Mashychev, and M. Rarey. Comparing Molecular Patterns Using the Example of SMARTS: Theory and Algorithms. *Journal of Chemical Information and Modeling*, 59(6):2560–2571, 2019.

[64] N. Schneider. *HYDE : Konsistente Bewertung von Protein-Ligand-Komplexen auf der Basis von Wasserstoffbrücken- und Dehydratationsenergie*. PhD dissertation, Universität Hamburg, 2012.

[65] N. Schneider, S. Hindle, G. Lange, R. Klein, J. Albrecht, H. Briem, K. Beyer, H. Claußen, M. Gastreich, C. Lemmen, and M. Rarey. Substantial improvements in large-scale redocking and screening using the novel HYDE scoring function. *Journal of Computer-Aided Molecular Design*, 26(6):701–723, 2012.

[66] N. Schneider, R. Klein, G. Lange, and M. Rarey. Nearly no Scoring Function Without a Hansch-Analysis. *Molecular Informatics*, 31(6-7):503–507, 2012.

[67] N. Schneider, G. Lange, S. Hindle, R. Klein, and M. Rarey. A consistent description of HYdrogen bond and DEhydration energies in protein-ligand complexes: methods behind the HYDE scoring function. *Journal of Computer-Aided Molecular Design*, 27:15–29, 2013.

[68] K. T. Schomburg, E. Nittinger, A. Meyder, S. Bietz, N. Schneider, G. Lange, R. Klein, and M. Rarey. Prediction of protein mutation effects based on dehydration and hydrogen bonding – A large-scale study. *Proteins: Structure, Function and Bioinformatics*, 85(8):1550–1566, 2017.

[69] M. Shapovalov, S. Vucetic, and R. L. Dunbrack. A new clustering and nomenclature for beta turns derived from high-resolution protein structures. *PLoS Computational Biology*, 15(3), 2019.

[70] M. Shapovalov and R. Dunbrack. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure*, 19(6):844–858, 2011.

[71] G. S. Sheppard, J. Wang, M. Kawai, N. Y. BaMaung, R. A. Craig, S. A. Erickson, L. Lynch, J. Patel, F. Yang, X. B. Searle, P. Lou, C. Park, K. H. Kim, J. Henkin, and R. Lesniewski. 3-Amino-2-hydroxyamides and related compounds as inhibitors of methionine aminopeptidase-2. *Bioorganic & Medicinal Chemistry Letters*, 14(4):865–868, 2004.

[72] A. Spitzmüller, H. F. Velec, and G. Klebe. MiniMuDS: A new optimizer using knowledge-based potentials improves scoring of docking solutions. *Journal of Chemical Information and Modeling*, 51(6):1423–1430, 2011.

[73] T. B. Steinbrecher, M. Dahlgren, D. Cappel, T. Lin, L. Wang, G. Krilov, R. Abel, R. Friesner, and W. Sherman. Accurate Binding Free Energy Predictions in Fragment Optimization. *Journal of Chemical Information and Modeling*, 55(11):2411–2420, 2015.

[74] K. Stierand and M. Rarey. Drawing the PDB: Protein-ligand complexes in two dimensions. *ACS Medicinal Chemistry Letters*, 1(9):540–545, 2010.

[75] M. Su, Y. Du, Q. Yang, R. Wang, Z. Liu, G. Feng, and Y. Li. Comparative Assessment of Scoring Functions: The CASF-2016 Update. *Journal of Chemical Information and Modeling*, 59(2):895–913, 2019.

[76] R. Taylor, J. Cole, O. Korb, and P. McCabe. Knowledge-based libraries for predicting the geometric preferences of druglike molecules. *Journal of Chemical Information and Modeling*, 54(9):2500–14, 2014.

[77] T. C. Terwilliger, H. Klei, P. D. Adams, N. W. Moriarty, and J. D. Cohn. Automated ligand fitting by core-fragment fitting and extension into density. *Acta Crystallographica Section D: Biological Crystallography*, 62(8):915–922, 2006.

[78] I. J. Tickle. Statistical quality indicators for electron-density maps. *Acta Crystallographica Section D*, 68(4):454–467, 2012.

[79] W. G. Touw and G. Vriend. BDB: databank of PDB files with consistent B-factors. *Protein Engineering, Design and Selection*, 27(11):457–462, 2014.

[80] N. Valls, R. A. Steiner, G. Wright, G. N. Murshudov, and J. A. Subirana. Variable role of ions in two drug intercalation complexes of DNA. *Journal of Biological Inorganic Chemistry*, 10(5):476–482, 2005.

[81] B. van Beusekom, N. Wezel, M. L. Hekkelman, A. Perrakis, P. Emsley, and R. P. Joosten. Building and rebuilding N-glycans in protein structure models. *Acta Crystallographica Section D: Structural Biology*, 75:416–425, 2019.

[82] L. Wang, Y. Deng, J. L. Knight, Y. Wu, B. Kim, W. Sherman, J. C. Shelley, T. Lin, and R. Abel. Modeling Local Structural Rearrangements Using FEP/REST: Application to Relative Binding Affinity Predictions of CDK2 Inhibitors. *Journal of Chemical Theory and Computation*, 9(2):1282–1293, 2013.

[83] L. Wang, Y. Wu, Y. Deng, B. Kim, L. Pierce, G. Krilov, D. Lupyan, S. Robinson, M. K. Dahlgren, J. Greenwood, D. L. Romero, C. Masse, J. L. Knight, T. Steinbrecher, T. Beuming, W. Damm, E. Harder, W. Sherman, M. Brewer, R. Wester, M. Murcko, L. Frye, R. Farid, T. Lin, D. L. Mobley, W. L. Jorgensen, B. J. Berne, R. A. Friesner, and R. Abel. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society*, 137(7):2695–2703, 2015.

[84] G. L. Warren, T. D. Do, B. P. Kelley, A. Nicholls, and S. D. Warren. Essential considerations for using protein-ligand structures in drug discovery. *Drug Discovery Today*, 17(23-24):1270–1281, 2012.

[85] Z. Yang, K. Lasker, D. Schneidman-Duhovny, B. Webb, C. C. Huang, E. F. Pettersen, T. D. Goddard, E. C. Meng, A. Sali, and T. E. Ferrin. UCSF Chimera, MODELLER, and IMP: An integrated modeling system. *Journal of Structural Biology*, 179(3):269–278, 2011.

[86] D. Yusuf, A. M. Davis, G. J. Kleywegt, and S. Schmitt. An alternative method for the evaluation of docking performance: RSR vs RMSD. *Journal of Chemical Information and Modeling*, 48(7):1411–1422, 2008.

[87] Y. Zhang and M. F. Sanner. Docking Flexible Cyclic Peptides with AutoDock CrankPep. *Journal of Chemical Theory and Computation*, 15(10):5161–5168, 2019.

Appendix A

Software and Workflows

In the following, software tool chains are introduced to create validation data sets and run evaluation schemes. They are followed by the technical description of all relevant tools in C++ and their surrounding Python frameworks that were build for this dissertation. In retrospective, five major and seven minor tools were created. Additionally, four Python frameworks and multiple C++ libraries had to be created or modified. All the below mentioned tools and frameworks are now present in the NAOMI code base fulfilling our internal levels of code quality guaranteed by code review as well as sufficient unit testing and consistent system tests to guard against changes over time. The `Reproducibility` area of the NAOMI library was founded to allow the grouping of the minor tools and the Python frameworks with the respective main tool.

Visualization

Pictures in this thesis are created with the help of Chimera[85], PoseView[74], Python3, and the HydeDebugGUI explained later on.

A.1 Tool Chains

The workflow to create ProtFlex18 and input to run the evaluation of GeoHYDE is shown in Figure A.1. Figure A.2 displays the tool chain to create a new torsion library and calculate a CSD validation on the CSD and high quality PDB ligands.

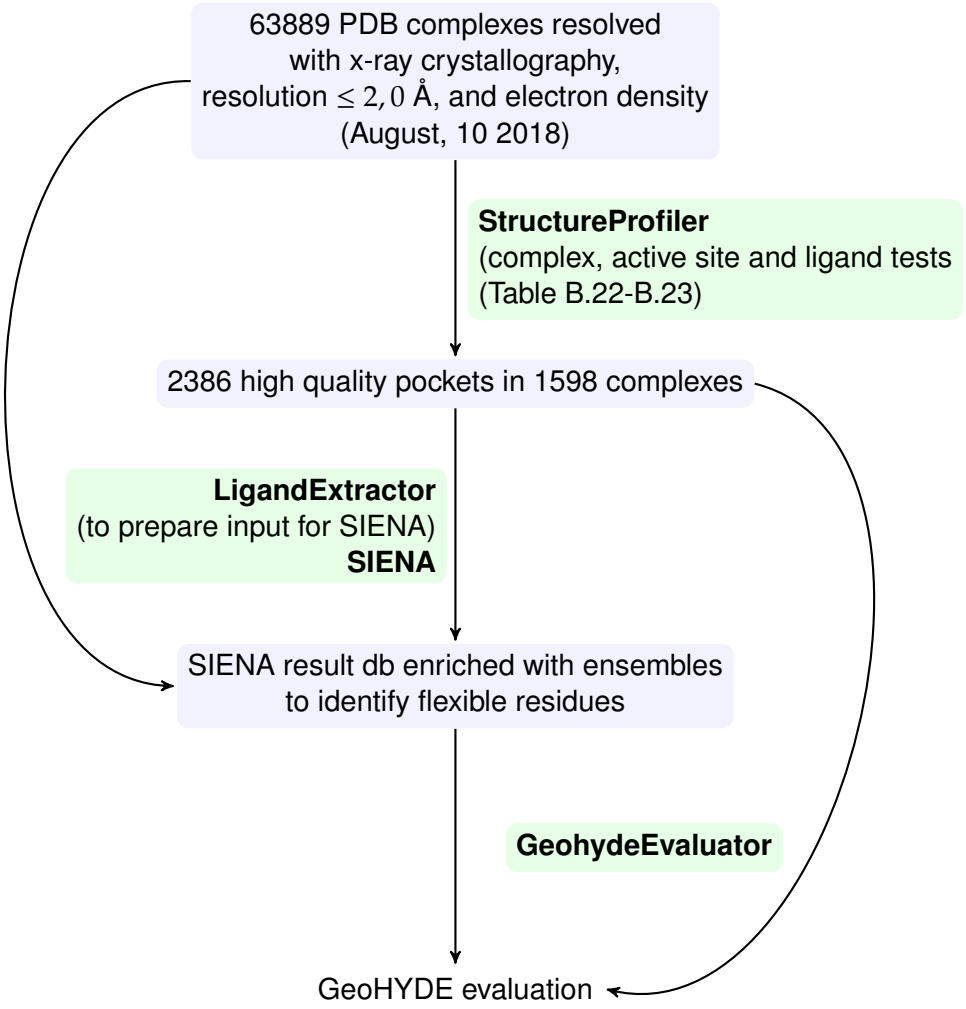


Figure A.1: Workflow to create the validation data set for GeoHYDE and run the evaluation.

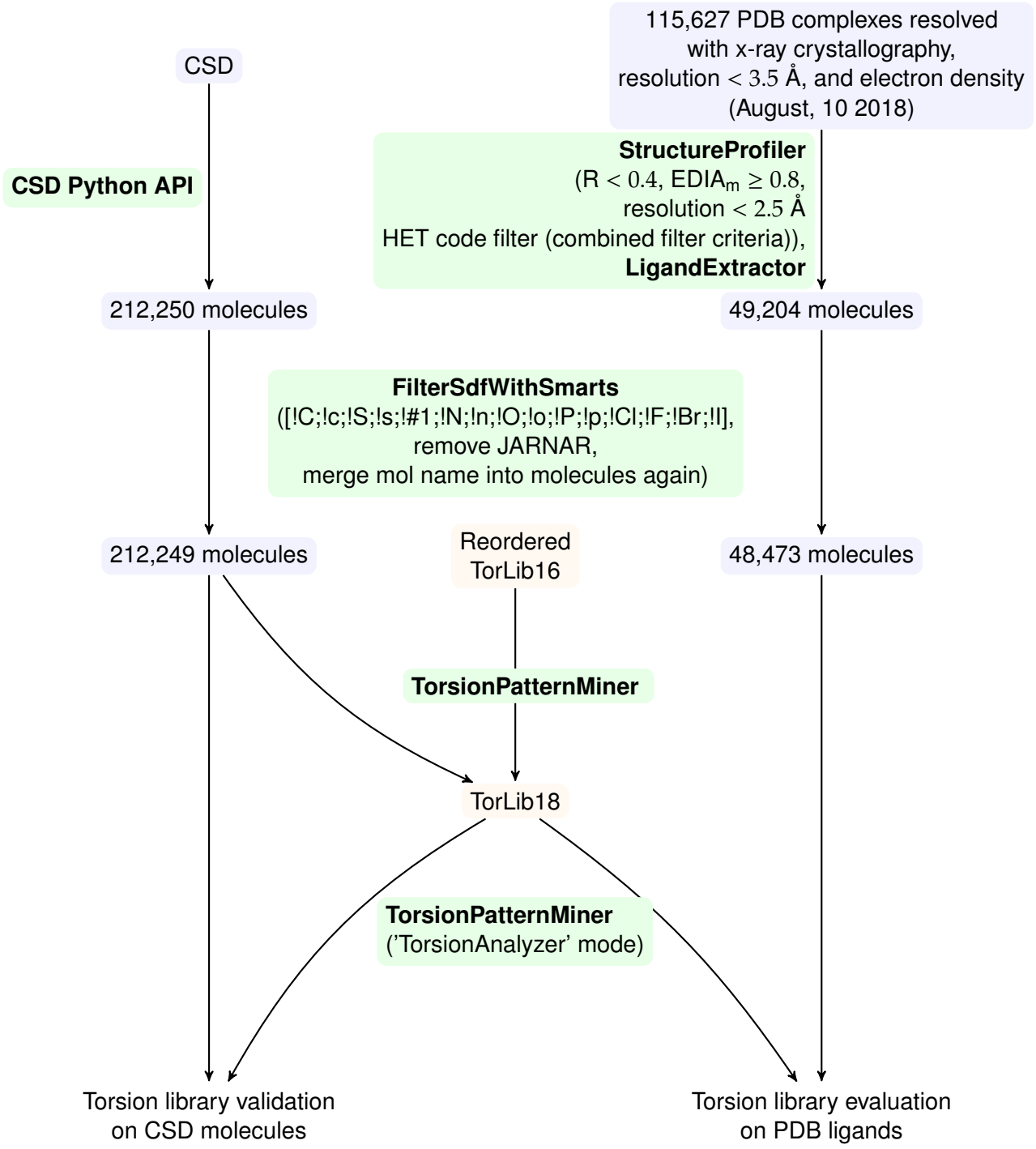


Figure A.2: Workflow to create the TorLib18, validate it on the CSD 2018 and evaluate it on high quality ligands in the PDB.

A.2 Tools and Libraries

In the following, all tools and newly developed or noteworthy adjusted libraries are introduced in short.

In the AMD group lead by Professor Rarey, a multi-step code quality assurance was built up over the time of this thesis. Before publication, C++ and Python code should be published on the internal code review server. The software needs to pass multiple checks:

- code review by two other PhD students
- automatic code style analysis (cpp check)
- unit test coverage should not be reduced
- new code needs to have sufficient unit tests
- each tool should have at least one system test testing for its general activity

After fulfilling all prerequisites, each code is merged and standalone tool packages can subsequently be built to be integrated on the groups server `http://proteins.plus` as well as into the AMD ChemBio Suit. By default, the tools are free for academic use once published.

Detailed documentation for the published tools `EDIAscorer` and the `StructureProfiler` can be found in their respective publication. The can be used online on `https://proteins.plus` (Figure A.3). In the following, all not yet published tools are presented.

In all cases, Python frameworks and tools are accompanied by basic tests to e.g. explain their usage and their long-term operation. The work flow to create ProtFlex18 with the help of `StructureProfiler` and `SIENA` is fully converted into system tests so that it will be available in the future.

A.2.1 Tools for Generating Data Sets

LigandExtractor

According to the PDB, the residue sequence ID is the unique id of an molecular entity in a structure data file. The `LigandExtractor` reads a given PDB file and the identification of a molecule or metal in the form HET_Chain_ResSeqID to write out its coordinates into an SDF file.

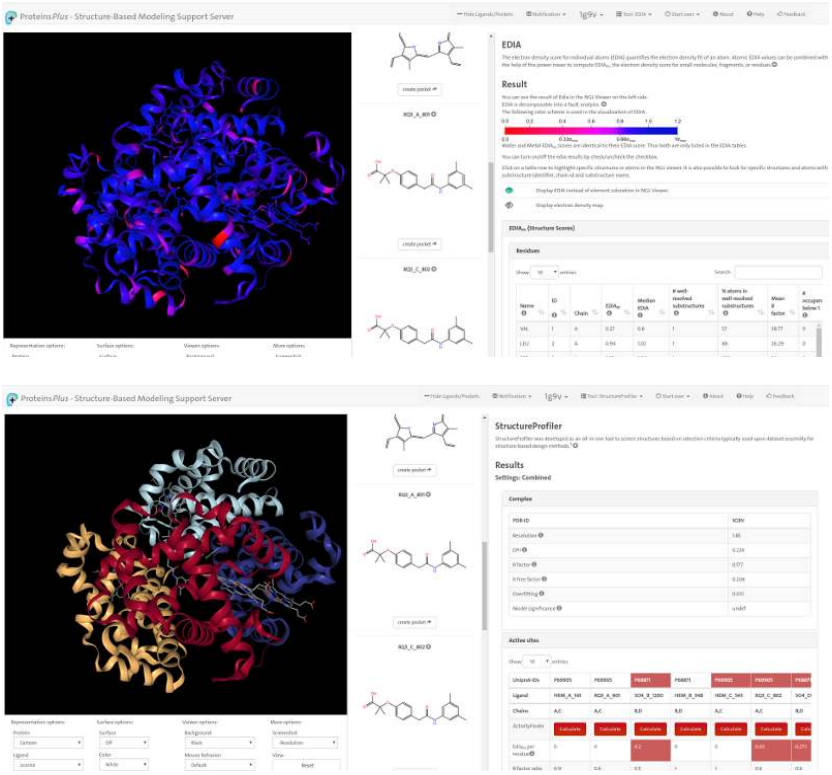


Figure A.3: `EDIAscorer` and the `StructureProfiler` are integrated into `ProteinsPlus`.

```
LigandExtractor -c COMPLEX.PDB -l HET_CHAIN_RESSEQID -o OUTPUTDIR
```

The structures can then be used in extracting high quality ligands identified by the `StructureProfiler` to generate a PDB based benchmark data set for the torsion library. They can also identify the pocket for `SIENA` on which an ensemble search should then be conducted.

StructureProfiler Python Framework

Detailed documentation for the `StructureProfiler` can be found in the Supporting Information in its publication [36]. All available tests and those active when compiling the ProtFlex18 data set can be found in Table B.22 - B.23. To allow the automatic analysis of the results of 100 00 structures an Python based accumulation framework was added to NAOMI.

```
python3 run_structureprofiler_analysis.py -i DIROFSPFOUTPUT
-e DENSITYDIR -d IDFILE -o OUTPUTDIR
```

It can identify the ligands passing all activated tests. It also generates accumulated output over the number of failed tests with example ids to allow single case analysis. The resulting data set file can directly be used in the `SIENA` and `GeohydeEvaluator` Python frameworks.

PDBDataExtractor

`PDBDataExtractor` allows the extraction of information stored in a PDB header about the enzyme classification and the name of the enzyme for each chain present in the PDB file. Its output should be merged into one file and then used by the `SIENA` Python Framework. It is situated in the `Reproducibility` area of the NAOMI library.

```
PdbDataExtractor -c COMPLEX -o OUTPUTDIR
```

SIENA and its Python Framework

`SIENA` comes with the ability to apply various filters on its detected ensembles to reduce their size. In validation mode, it can also store residues in an SQLite database for further use in `GeohydeEvaluator` that are identified as flexible in an ensemble. As part of this thesis, one additional filter that checks for passing

`StructureProfiler` tests under the consideration of electron density was introduced to `SIENA` in extending the `ProteinFlexibilityLib`. Additionally, a Python framework around `SIENA` and its result database was added to NAOMI. It processes the output created by the `PDBDataExtractor` for identifying and naming clusters after their enzyme function.

```
python3 run_siena_analysis.py -e ECINFORMATION -i DIROFSIENAOUTPUT
-d IDFILE -o OUTPUTDIR
```

It also allows e.g. the analysis of the interconnectivity of ensembles. The graph can be stored in an SQLite database for future use by the Python framework of the `GeohydeEvaluator`. For this thesis, further output in e.g. LaTeX with the amount of unique pdb ids and ligands per ensemble (Chapter 3) can also be created.

A.2.2 Tools for Generating a Torsion Library

The workflow to generate a new torsion library lists four tools of which the `LigandExtractor` and `StructureProfiler` have already been introduced. Here, the last two tools `FilterSdfWithSmarts` and `TorsionPatternMiner` are presented.

FilterSdfWithSmarts

The torsion library was extracted from the CSD with among other criteria not consisting of the elements matching the SMARTS string *ExclusionSmart* below. Also, Guba et al. removed the molecule JARNAR from the validation data set [19]. The tool `FilterSdfWithSmarts` filters a given SDF file by the *ExclusionSmart* and removes JARNAR if detected.

```
ExclusionSmarts: [C;!c;!S;!s;!#1;!N;!n;!O;!o;!P;!p;!Cl;!F;!Br;!I]
FilterSdfWithSmarts -i INPUTSDF -o OUTPUTSDF -n NAMES
```

Since the CSD Python framework does not annotate the molecule names in its SDF files, they need to be annotated later on with the help of e.g. `FilterSdfWithSmarts`.

TorsionPatternMiner

We have implemented the tool `TorsionPatternMiner` as part of NAOMI to supersede the `torsionchecker` with additional parts of the `TorsionAnalyzer` [62][19]. It is geared towards the creation and validation of a new torsion library in mining

a specific molecule file and creating output necessary for the by Guba et al. developed validations strategy. The reimplementation is now based on the up to date NAOMI C++ code using e.g. the recently published SMARTScompare algorithm for SMARTS matching [63]. The tool is supposed to be published in the future when the changes in the new torsion library are finished. In the following, all tool options are given and examples are provided below.

- `--outdir` Location to store the output (required)
- `--molfile` File path to mols in sdf, will be stored in given database
- `--database` File path to (new) molecule database
- `--initialtorsionlib` Torsion lib to be analyzed
- `--selectivematching` (=false) Match only the most selective smarts pattern - default mode in NAOMI
- `--useonlysinglebonds` (=false) Only allow single bonds for matching
- `--donotuseterminalbonds` (=true) Do not use bonds to a terminal heavy atom
- `--storeincsdhistograms` (=true) True: store in csd histogr., 1: store in pdb histogrs.
- `--sequential` (=false) Switch to sequential calculation
- `--startfrommol` Start evaluation from specific mol in database
- `--matchpatternwithatleastXhits` (=0) Default: 0
- `--extractmol` Extract specific mol id from database

Update of the TorLib Statistics and Peaks

`TorsionPatternMiner` can update the statistics of a specified torsion library (`--initial torsionlib <TorsionLib>`) with all data present in a multi mol sdf file (`--molfile <multi mol sdf file>`). All molecules will first be stored in the given database file (`--database <Database File>`) and subsequently processed. For future runs, the molecule database can then be reused. A run tarting from a specific molecule is also possible: (`--startfrommol <FilePosition>`).

TorsionPatternMiner uses the Intel Threading Building Blocks [24] for automatic multiprocessing on all available threads of the machine. If this behavior is undesired, the sequential mode should be activated (--sequential true). To update the statistics according to [62], the multimatching needs to be active (--selectivematching false). This means, that per bond, each matching torsion rule will receive an increase in the statistics. If a pattern matches multiple times on the torsion bond e.g. due to leaving the element of the substituing partner on position 1 or 4 undefined, each match will be added to the statistic. TorsionPatternMiner also updates all peak records and adjusts their tolerances automatically if needed.
Since TorsionPatternMiner matches all available torsion rules to any bond in all given molecules, one may want to limit the type of bonds to be used for matching. It is possible to explicitly avoid any non-single bond (--useonlysinglebonds true) as well as all bonds connected to a terminal heavy bond (--donotuseterminalbonds true). The torsion library stores the statistics from the CSD in the histogram and histogram_shifted XML tag. It is possible to store a second statistic per pattern in histogram2 and histogram2_shifted with --storeincsdhistograms false. TorsionPatternMiner updates peaks always based on data in the histogram_shifted tag per pattern.

TorLib Statistics Analysis
Besides the aforementioned command line options, two are relevant for the quality analysis of the derived torsion library. It may be desired to leave out low populated patterns for the single matching (--matchpatternwithatleastXhits 50). As each bond in the output is annotated with the molecule id, this id can be used to extract the specific molecule from the database (--extractmol <Molecule ID>) for a single case analysis with the TorsionAnalyzer.

Torsion Rule Visualization
TorsionPatternMiner uses the parameter --visualizetorlib im combination with a torsion library and an output directory to convert each torsion rule into a text format. This can then be converted into graphics to understand the correspondence of peaks with the underlying histogram data. The conversion code is available in the attached python package to the tool.

Examples
Create a new torsion library based on a given multi mol file and a torsion library hierarchy:

```
TorsionPatternMiner --out DIR --initialtorlib TOR_LIB
    --molfile MULTIMOLFILE --database mols.db
    --selectivematching false --storeincsdhistograms true
```

From the resulting output files, the new tor lib should then be used to control the quality of the peak determination in running the tool in single matching mode on the molecule set.

```
TorsionPatternMiner --out DIR --initialtorlib DIR/newtorlib.xml
    --database mols.db --donotuseterminalbonds true
```

The resulting `bondanglesmatching.csv` can then be analyzed by our python script `createpaperplots.py` to generate the torsion rule - red flags in per cent plot. It is advisable to compare the `bondanglesmatching.csv` file of the initial torsion lib with the one generated by the new tor lib. Likewise, a different molecule set such as the ligand expo can be employed to test its agreement with the presented torsion library.
The python script `sortandcomparetorsionpatterns.py` takes as input two such files and compares each bond, angle data triplet in terms of the matching torsion rule and the determined angle quality. If the data triplet matches a different torsion rule and, or receives a differing quality assessment, it will be quantified in the output files and annotated with examples. This analysis was applied on the resorted TorLib to control against unwanted sorting until only reasonable switches were found.
`visualizeContTorScoreFromPatMiner.py` takes as input the directory with the extracted data and visualizes the given patterns.

A.2.3 Libraries and Tools Connected to HYDE

Over the term of this thesis, two graphical and two command line tools for the development and evaluation of HYDE were developed with my participation.

HydeDebugGUI

The `HydeDebugGUI` (HDG) is a graphical tool to analyze and optimize HYDE scores. It was initially developed by Dr. Schneider and further expanded by Dr. Nittinger and me. A full reimplementation due to HDG's current incompatibility to Qt on Windows was implemented by M. Grössler under my supervision and is in preparation for the merge into the NAOMI mainline. The merge is currently not possible since the ability to compute structures on the command line has not yet been reimplemented.
HDG visualizes the active site with the for HYDE relevant hydrogen bonds and other information such as the position of possible waters (Figure A.4). It also shows residual HYDE scores for e.g. thermostability analysis and protein-protein interface scores in a given structure. Great care has been taken on allowing a full export of the result of a geometrical optimization including proton positions. To identify structural deficits, atomic B factor and EDIA coloring are integrated. Ligands in the score table are marked yellow, when strained torsions are present. They are marked orange when a close heavy atom contact is detected. If both are present, the ligand entry is dyed in red.

The results of e.g. the small series data set given on the command line can then be analyzed with the analysis scripts of the CASF benchmarks and the Python framework called `hyde_evaluator` written by our cooperation partner BioSolveIT. For working graphically with HYDE, the HDG is the center tool.

geohydeoptimizer

`geohydeoptimizer` was written by BioSolveIT to analyze the dispersion of small scale sampled ligand configurations through optimization with GeoHYDE. Through RMSD based cluster analysis, the spread of HYDE scores per RMSD cluster as well as the amount of such clusters and the existence of singletons can be observed. A strong increase in RMSD clusters or in the HYDE score difference per cluster indicate the introduction of an unwanted step function into GeoHYDE. The tool is a derivative and an extension of my first now outdated benchmarking tool called `hydeoptimizer`.

GeohydeEvaluator

`GeohydeEvaluator` is a newly developed, highly configurable tool for benchmarking GeoHYDE that integrates `geohydeoptimizer`'s ligand sampling ability. In its

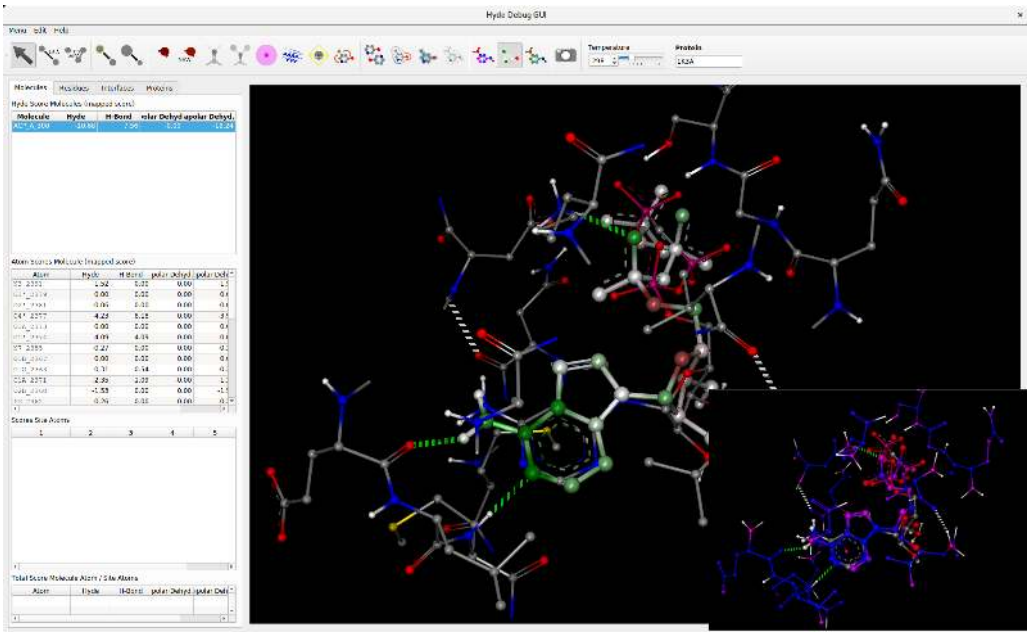


Figure A.4: HydeDebugGUI displays e.g. pocket with HYDE or $EDIA_m$ colors.

normal configuration, it optimizes a specific complex-ligand complex in regards to HYDE and reports partial initial and final scores. The results of a pocket flexibility analysis with SIENA can be supplied to the tool. If such flexible residues are present in the pocket, their side chains will also be geometrically optimized in combination with the ligand. In the following examples as well as necessary details of the implementation are outlined. Available tool options:

- `--resultFolder` Specify result output folder (required)
- `--complex` Complex PDB file (required)
- `--config` Configuration file (required)
- `--ligand` Ligand or conformers of ligand sdf file
- `--density` Density in ccp4 file format
- `--molId` Mol id of the ligand in the PDB file to be analyzed. Format should be HET ID_CHAIN_RESSEQID
- `--waters` Geometrically optimize waters in binding pocket which initially have at least an $EDIA_m$ of a certain value
- `--sienadb` SIENA result DB to extract flexible residues from

- `--printconfig` Write out config file

Examples
Evaluate a specific ligand with `GeohydeEvaluator`:

```
GeohydeEvaluator --resultFolder YOURLOCATION --complex YOURCOMPLEX
    --molID ID_CHAIN_RESSEQID --density DENSITY_PDBID.ccp4
```

An initial tool configuration can be obtained in setting `--printconfig` to True. It can subsequently be fed back into the tool:

```
GeohydeEvaluator --resultFolder YOURLOCATION --complex YOURCOMPLEX
    --molID ID_CHAIN_RESSEQID --density DENSITY_PDBID.ccp4
    --config YOURCONFIG
```

In switching `RunSampling` in the configuration file to `true`, the ligand will be sampled at the beginning. All conformations will then be optimized and evaluated. Be aware, that sampling around torsion bonds should be strongly restricted in setting the maximum number of conformers generated by sampling around torsion bonds with `MaxNumberTorsionWobblingPoses` to a value of 30 for example. The sampling can further be configured for rotation, translation and torsion bond sampling.

```
GeohydeEvaluator --resultFolder YOURLOCATION --complex YOURCOMPLEX
    --molID ID_CHAIN_RESSEQID --density DENSITY_PDBID.ccp4
    --config YOURMODIFIEDCONFIG
```

Optimizing with flexible side chains is possible in giving a SIENA result database to `GeohydeEvaluator` with an entry for the specific PDB structure and switching `FlexibleResidues` to `true`. Please be aware, that only flexible residue in the active site of the specified ligand can be considered. If none of them are close enough to the ligand, `GeohydeEvaluator` automatically switches to a normal optimization without protein flexibility. If flexible residues are detected, their initial and final $EDIA_m$ after the optimization will be reported in an additional column in the output file.

```
GeohydeEvaluator --resultFolder YOURLOCATION --complex YOURCOMPLEX
    --molID ID_CHAIN_RESSEQID --density DENSITY_PDBID.ccp4
    --sienadb SIENARESULTDB
```

The experiment to optimize consecutively all waters in the active site with an initial $EDIA_m$ above e.g. 0.8 is as follows:

```
GeohydeEvaluator --resultFolder YOURLOCATION --complex YOURCOMPLEX
    --molID ID_CHAIN_RESSEQID --density DENSITY_PDBID.ccp4
    --waters 0.8
```

The subsequent passages describe the inner work flow from reading an input file over optimizing to scoring a pocket with HYDE.

Preprocessing

In the following, the preprocessing of a complex with its ligand is explained. A complex can be presented in PDB format and with the help of the `ComplexLib::ComplexFactory` translated into a NAOMI complex. The ligand to be optimized can either be added with the help of an SDF file or in specifying its molecular id in giving the triplet HETcode_Chain_ResSeqId to the executable. In the first case, the SDF file is processed and all entries are seen as the configurations of an identical ligand. With the help of the second method, a infile id and chain match is searched in the complex molecules, ions and waters. The matching structure is then used as the ligand. Subsequently, the active site needs to be prepared. First, the standard HYDE site with the radius of 8 Å around the ligand as well as the big site with the radius of 11.5 Å are created. All waters are then removed in both pockets and Protoss [5] is run for both sites and the ligand. Only if the user supplies precomputed ligand configurations, Protoss is not used on the ligand to avoid changes in its proton configuration. In accordance to the workflow for treating waters as used in `warpp`[41], all waters are removed from the binding site. If a `SIENA` result database is defined, flexible residues are identified in the pocket with functionality, that had to be moved from `SIENA` to the `SIENAToolLib` as part of this thesis.
If waters should be optimized, first, they will be evaluated with $EDIA_m$ and those above the given cutoff will be assembled. Each water to be optimized thus needs to be removed from the complex to not duplicate it while the other waters are kept as part of the pocket.

File Output

After the optimization, poses can be written out in storing the ligand in its own SDF file. Cofactors of the complex are also written into an SDF file while the complex with its annotated protons is written out to a PDB file. Score data is written into a CSV output file. Pockets are available with and without explicitly placed waters for which the recently published tool `warpp` had to be refactored.

Python Framework

The accompanying Python framework reads and stores all scores in SQLite data

bases. It automatically detects the affiliation of the pocket to one of the ProtFlex18 data sets. Connecting to both the score and the SIENA cluster analysis data base, it generates all necessary analysis plots.

Structural Deviations

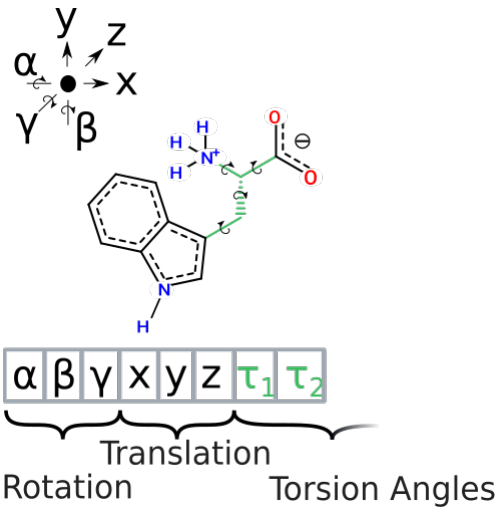
The classes `GlobalRotationTranslationWobbling` (GRTW), `TorsionWobbling` (TW), and both classes combined in `GlobRotTransLocalTorWobbling` (GRTTW) in the NAOMI library `Molecule` allow to perturb the coordinates of a molecule in `GeohydeEvaluator`. GRTW and TW both need range intervals and step sizes to direct the modifications. The first class allows the molecule to be rotated around its center of mass as well as to be translated along the unit vectors in $\mathbb{R}^3$. The torsional perturbation allows the rotation around each rotatable bond while only producing a maximum number of molecule configurations. Hence, a root atom with the minimum distance to any atom is determined in the initialization phase. Then, all rotatable bonds are grouped together by their minimum distance to the root atom. Going from the most distant set of bonds towards the root atom, all rotatable bonds with at least the current distance to the root atom are allowed to be perturbed, when the in total generated number of configuration is not above the number of maximally allowed configurations. Thus, the total number of configurations is as follows:

$$c_{GRTW} = \#step_{rot}^{3} \cdot \#step_{trans}^{3} \tag{A.1}$$

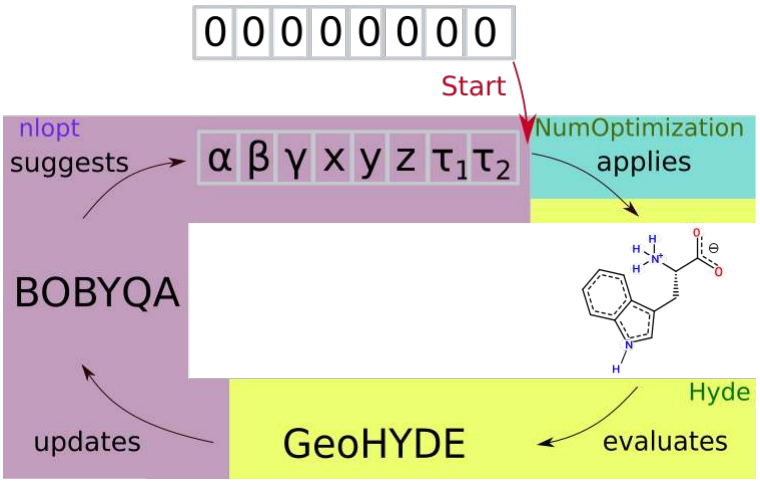
$$c_{TW} = \#step_{tor}^{\#rotbonds} \tag{A.2}$$

$$c_{GRTLTW} = c_{GRTW} \cdot c_{TW} \tag{A.3}$$

A step can also have the value of zero, thus being neutral. Since the number of possible configurations can escalate quickly, the `GeohydeEvaluator` only selects twenty conformers by random from them. Those are not allowed to exceed an RMSD of 2 Å. The ligand configuration is also removed if it has an intramolecular clash higher than those of the original ligand pose. Only slight intermolecular clash is accepted in either maximally three atom contacts or maximally as much contact as the original ligand configuration had with the protein. Contact is identified by analyzing the van der Waals sphere intersections of the protein atoms with the sphere of 0.4 times the van der Waals radius sphere of the heavy atoms in the ligand.



(a) Degrees of freedom in the Optimization in the case of a flexible molecule.



(b) Optimization work flow annotated with the participating code libraries in NAOMI with tryptophan as ligand.

Figure A.5: Optimization workflow with its degrees of freedoms.

Optimization

`NumOptimization` as the library for gradient free optimization for HYDE was created by me. In the middle of my thesis, Florian Flachsenberg extended the software to also allow gradient based optimization. As a result, the libraries `NumOptimization` and `NumOptimizationHelper` now contain all abstract classes necessary for using `NLopt` to optimize a set of atoms, an active site or a ligand. The implementation of these for GeoHYDE can be found in the `Optimization` directory of the `Hyde` library. Geometrically optimizing a ligand in a fixed protein pockets means to allow global rotation and translation for the ligand. Also, rotatable bonds following the criteria of the TorsionLib (Table 4.1) as well as single bonds towards a hydrogen donor should be rotatable throughout the optimization (see Figure A.5(a)). These types of bonds can also be made flexible in a protein side chain. Following the work flow depicted in Figure A.5(b) the initial position of the ligand and other groups are the baseline against which the optimization strategy suggests changes. They are always applied on the original pocket configuration, scored with the active GeoHYDE terms and given to BOBYQA in the external `NLopt` package. The algorithm then integrates the score in its calculations and proposes the next pocket configuration to be tested. The cycle of suggesting, applying changes and scoring them is repeated until termination criteria are met.

Difference between GeoHYDE$_{desolv}$ and the intermolecular Lennard-Jones Score

In Chapter 5, three Lennard-Jones Potentials are monitored. Both the GeoHYDE$_{desolv}$ as an intermolecular and the intramolecular Lennard-Jones potential to identify clashes in the ligand and if necessary protein residues are specially fine tuned potentials including protons if necessary developed by our cooperation partner. The third potential which is not part of GeoHYDE but monitored in the evaluations is a standard 12-6 Lennard-Jones potential only evaluated between heavy atoms in differing components in the pocket.

Integration of the repulsive Lennard-Jones potential

The purely repulsive LJ potential (**C**) from the NAOMI `ScoringLib` has two configuration parameters: The preferred value of the potential when the two atoms fully overlap in their center and the position, for which the potential should reach zero at around twice the sum s of the van der Waals radii of the atoms. It would be preferred if the potential would be highly similar to the repulsive part in GeoHYDE$_d$

for proper comparison. A rough parametrization at $x = 0$ to $LJ_s = 100$ and to $LJ_s = 0$ for $x = s$ performed best in contrast to $x = 2s$ or $x = \sigma$, the position of the original zero crossing of the LJP. Regardless, **C** assesses some atom pairs as clashing while GeoHYDE$_d$ disagrees (see 2zzd TLA C 4001 in Table B.14).

A.2.4 EDIA and other extensions in CrystalGeometry

EDIA and EDIA$_\mathrm{m}$ are computed with the help of the `ElectronDensityScorer` in the NAOMI Library `CrystalGeometry`. It accesses through the `ElectronDensityWeighter` the precomputed electron density radii offsets and returns for each grid point the fitting element and charge dependent weight. The scorer then accumulates over all relevant grid points the EDIA and the detected fault. The result is either directly returned to the user or stored in a given object of the type `ElectronDensityScores`. The score container holds a number of unordered maps to e.g. store the atomic EDIAs as well as residue and molecule EDIA$_\mathrm{m}$. Substructures can also be scored.

Additional utilities such as the B Factor extraction and the computation of the H matrix can be found in the `Utils` area of the `CrystalGeometry` library. Python scripts for computing the electron density radius for each element and charge as well as the weighting curve of EDIA have been attached to the paper published in 2017.

EdiaStabilityAnalyzer

The `EdiaStabilityAnalyzer` computes the analysis of the numerical stability of EDIA as a system test as part of the test suite in NAOMI every time a code commit is merged. More information about the test setup can be extracted from the system test Python file if needed.

ScanHMatrixForErrors

The tool is situated in the `Reproducibility` Section of the NAOMI library. It takes as input a complex, a CCP4 density file and a cutoff epsilon and then compares the H matrix from the PDB with those computed from the density file. If the difference is larger than the given epsilon, both matrices are printed.

molwobbler

The tool is situated in the `Reproducibility` Section of the NAOMI library. It takes as input a complex, a ligand as SDF file and returns up to 100 not clashing perturbed ligand configurations in an multi mol file with the help of the `GlobRotTransLocalTorWobbling` utilities. Configurations are tested against internal clashes, clashes with the protein and are not allowed to be further away from the crystallized ligand position than 2 Å RMSD.

Appendix B

Additional Tables and Figures

Element, Charge	Resolution [Å]:	0.5	1.0	1.5	2.0	2.5	3.0
H		1.08	1.2	1.29	1.41	1.68	1.98
H -1		1.47	1.56	1.68	1.74	1.95	2.16
He		0.93	1.05	1.17	1.32	1.59	1.92
Li		0.9	0.9	0.99	1.23	1.68	2.01
Li +1		0.81	0.81	0.9	1.14	1.53	1.86
Be		1.02	1.2	1.32	1.35	1.71	2.01
Be +2		0.78	0.9	1.05	1.17	1.53	1.86
B		1.05	1.2	1.32	1.44	1.71	1.98
C		1.02	1.14	1.26	1.38	1.65	1.98
N		0.96	1.11	1.23	1.35	1.62	1.95
O		0.93	1.08	1.2	1.32	1.62	1.92
O -1		0.99	1.11	1.23	1.35	1.65	1.95
F		0.9	1.05	1.17	1.32	1.59	1.92
F -1		0.93	1.08	1.2	1.32	1.62	1.92
Ne		0.87	1.02	1.14	1.29	1.59	1.89
Na		0.87	0.99	1.14	1.29	1.59	1.92
Na +1		0.84	0.99	1.11	1.26	1.56	1.89
Mg		0.87	0.84	1.14	1.32	1.62	1.92
Mg +2		0.81	0.81	1.08	1.26	1.56	1.89
Al		0.87	1.02	1.05	1.2	1.62	1.95
Al +3		0.78	0.93	0.99	1.11	1.53	1.86
Si		0.87	1.05	1.17	1.32	1.62	1.95
Si +4		0.78	0.93	1.08	1.23	1.53	1.86
P		0.9	1.05	1.17	1.32	1.62	1.95
S		0.9	1.05	1.17	1.32	1.62	1.92
Cl		0.9	0.9	1.17	1.41	1.62	1.92
Cl -1		0.93	0.9	1.2	1.41	1.62	1.95
Ar		0.9	1.05	1.17	1.32	1.59	1.92

Element, Charge	Resolution [Å]:	0.5	1.0	1.5	2.0	2.5	3.0
K		0.87	1.02	1.17	1.32	1.59	1.92
K +1		0.87	1.02	1.14	1.29	1.59	1.92
Ca		0.87	0.87	1.05	1.32	1.62	1.92
Ca +2		0.87	0.84	1.02	1.29	1.56	1.89
Sc		0.87	1.02	1.17	1.32	1.62	1.92
Sc +3		0.84	0.99	1.11	1.26	1.56	1.89
Ti		0.87	1.02	1.14	1.32	1.59	1.92
Ti +2		0.84	0.99	1.11	1.29	1.56	1.89
Ti +3		0.84	0.99	1.11	1.26	1.56	1.89
Ti +4		0.81	0.96	1.11	1.26	1.56	1.89
V		0.87	1.02	1.26	1.47	1.59	1.92
V +2		0.84	0.99	1.23	1.41	1.56	1.89
V +3		0.84	0.99	1.2	1.41	1.56	1.89
V +5		0.81	0.96	1.2	1.41	1.56	1.89
Cr		0.87	1.02	1.14	1.29	1.59	1.92
Cr +2		0.84	0.99	1.11	1.29	1.56	1.89
Cr +3		0.84	0.96	1.11	1.26	1.56	1.89
Mn		0.87	0.84	1.05	1.23	1.59	1.92
Mn +2		0.84	0.84	1.02	1.2	1.56	1.89
Mn +3		0.84	0.81	1.02	1.2	1.56	1.89
Mn +4		0.81	0.81	0.99	1.2	1.56	1.89
Fe		0.84	0.84	0.93	1.23	1.59	1.92
Fe +2		0.84	0.84	0.93	1.2	1.56	1.89
Fe +3		0.81	0.81	0.9	1.2	1.56	1.89
Co		0.84	0.99	1.02	1.29	1.59	1.92
Co +2		0.84	0.96	1.02	1.26	1.56	1.89
Co +3		0.81	0.96	0.99	1.26	1.56	1.89
Ni		0.84	0.99	1.02	1.29	1.59	1.89
Ni +2		0.81	0.96	0.99	1.26	1.56	1.89
Ni +3		0.81	0.96	0.99	1.26	1.56	1.89
Cu		0.84	0.84	1.02	1.2	1.56	1.89
Cu +1		0.84	0.84	1.02	1.2	1.56	1.89
Cu +2		0.81	0.81	0.99	1.2	1.56	1.89
Zn		0.84	0.84	1.02	1.29	1.56	1.89
Zn +2		0.81	0.81	0.99	1.26	1.56	1.89
Ga		0.84	0.99	1.11	1.29	1.56	1.89
Ga +3		0.81	0.96	1.08	1.26	1.53	1.86
Ge		0.84	0.99	1.11	1.29	1.59	1.89
Ge +4		0.78	0.93	1.08	1.26	1.53	1.86
As		0.84	0.99	1.11	1.29	1.59	1.89
Se		0.84	0.99	1.11	1.29	1.59	1.89
Br		0.84	0.99	1.11	1.29	1.59	1.89
Br -1		0.84	0.99	1.23	1.38	1.59	1.92
Kr		0.84	0.99	1.14	1.29	1.59	1.89

Element, Charge	Resolution [Å]:	0.5	1.0	1.5	2.0	2.5	3.0
Rb		0.84	0.99	1.11	1.5	1.56	1.89
Rb +1		0.84	0.99	1.11	1.5	1.56	1.89
Sr		0.84	0.99	1.02	1.35	1.59	1.89
Sr +2		0.81	0.96	1.02	1.35	1.56	1.89
Y		0.84	0.99	1.23	1.35	1.59	1.92
Y +3		0.81	0.96	1.2	1.32	1.56	1.89
Zr		0.84	0.99	1.11	1.29	1.59	1.92
Zr +4		0.81	0.96	1.08	1.26	1.56	1.89
Nb		0.84	0.99	1.14	1.29	1.59	1.89
Nb +3		0.81	0.96	1.11	1.26	1.56	1.89
Nb +5		0.81	0.96	1.08	1.26	1.53	1.86
Mo		0.84	0.99	1.11	1.29	1.59	1.89
Mo +3		0.81	0.96	1.11	1.26	1.56	1.89
Mo +5		0.81	0.96	1.08	1.26	1.56	1.89
Mo +6		0.78	0.93	1.08	1.26	1.53	1.86
Tc		0.84	0.99	1.11	1.29	1.59	1.89
Ru		0.84	0.99	1.11	1.44	1.56	1.89
Ru +3		0.81	0.96	1.11	1.41	1.56	1.89
Ru +4		0.81	0.96	1.11	1.41	1.56	1.89
Rh		0.84	0.99	1.11	1.44	1.56	1.89
Rh +3		0.81	0.96	1.11	1.41	1.56	1.89
Rh +4		0.81	0.96	1.08	1.41	1.56	1.89
Pd		0.84	0.99	1.11	1.5	1.56	1.89
Pd +2		0.81	0.96	1.11	1.47	1.56	1.89
Pd +4		0.81	0.96	1.08	1.47	1.56	1.89
Ag		0.84	0.99	1.11	1.2	1.56	1.89
Ag +1		0.84	0.96	1.11	1.2	1.56	1.89
Ag +2		0.81	0.96	1.11	1.2	1.56	1.89
Cd		0.84	0.99	1.23	1.41	1.56	1.89
Cd +2		0.81	0.96	1.2	1.41	1.56	1.89
In		0.84	0.99	1.11	1.29	1.56	1.89
In +3		0.81	0.96	1.11	1.26	1.56	1.89
Sn		0.84	0.99	1.11	1.29	1.56	1.89
Sn +2		0.81	0.96	1.11	1.26	1.56	1.89
Sn +4		0.81	0.96	1.08	1.26	1.56	1.89
Sb		0.84	0.99	1.11	1.29	1.56	1.89
Sb +3		0.81	0.96	1.11	1.26	1.56	1.89
Sb +5		0.81	0.93	1.08	1.26	1.53	1.86
Te		0.84	0.99	1.11	1.8	1.8	1.89
I		0.84	0.99	1.11	1.29	1.56	1.89
I -1		0.84	0.99	1.32	1.35	1.59	1.92
Xe		0.84	0.99	1.11	1.29	1.56	1.89
Cs		0.84	0.99	1.23	1.5	1.56	1.89
Cs +1		0.84	0.99	1.23	1.5	1.56	1.89

Element, Charge	Resolution [Å]:	0.5	1.0	1.5	2.0	2.5	3.0
Ba		0.81	0.99	1.02	1.29	1.56	1.89
Ba +2		0.81	0.96	1.02	1.26	1.56	1.89
La		0.84	0.99	1.11	1.29	1.59	1.89
La +3		0.81	0.96	1.11	1.26	1.56	1.89
Ce		0.81	0.99	1.11	1.29	1.56	1.89
Ce +3		0.81	0.96	1.11	1.26	1.56	1.89
Ce +4		0.81	0.96	1.11	1.26	1.56	1.89
Pr		0.81	0.96	1.11	1.29	1.56	1.89
Pr +3		0.81	0.96	1.11	1.26	1.56	1.89
Pr +4		0.81	0.96	1.11	1.26	1.56	1.89
Nd		0.81	0.96	1.11	1.29	1.56	1.89
Nd +3		0.81	0.96	1.11	1.26	1.56	1.89
Pm		0.81	0.96	1.11	1.26	1.56	1.89
Pm +3		0.81	0.96	1.11	1.26	1.56	1.89
Sm		0.81	0.96	1.11	1.26	1.56	1.89
Sm +3		0.81	0.96	1.11	1.26	1.56	1.89
Eu		0.81	0.96	1.11	1.26	1.56	1.89
Eu +2		0.81	0.96	1.11	1.26	1.56	1.89
Eu +3		0.81	0.96	1.11	1.26	1.56	1.89
Gd		0.81	0.96	1.11	1.26	1.56	1.89
Gd +3		0.81	0.96	1.08	1.26	1.56	1.89
Tb		0.81	0.96	1.11	1.26	1.56	1.89
Tb +3		0.81	0.96	1.08	1.26	1.56	1.89
Dy		0.81	0.96	1.11	1.26	1.56	1.89
Dy +3		0.81	0.96	1.08	1.26	1.56	1.89
Ho		0.81	0.96	1.11	1.26	1.56	1.89
Ho +3		0.81	0.96	1.08	1.26	1.56	1.89
Er		0.81	0.96	1.11	1.26	1.56	1.89
Er +3		0.81	0.96	1.08	1.26	1.56	1.89
Tm		0.81	0.96	1.11	1.26	1.56	1.89
Tm +3		0.81	0.96	1.08	1.26	1.53	1.89
Yb		0.81	0.96	1.29	1.41	1.56	1.89
Yb +2		0.81	0.96	1.29	1.41	1.56	1.89
Yb +3		0.81	0.93	1.29	1.38	1.53	1.86
Lu		0.81	0.96	1.11	1.26	1.56	1.89
Lu +3		0.78	0.93	1.08	1.26	1.53	1.86
Hf		0.81	0.96	1.11	1.26	1.56	1.89
Hf +4		0.78	0.93	1.08	1.26	1.53	1.86
Ta		0.81	0.96	1.11	1.41	1.56	1.89
Ta +5		0.78	0.93	1.08	1.38	1.53	1.86
W		0.81	0.96	1.11	1.26	1.56	1.89
W +6		0.78	0.93	1.08	1.23	1.53	1.86

Element, Charge	Resolution [Å]:	0.5	1.0	1.5	2.0	2.5	3.0
Re		0.81	0.96	1.11	1.2	1.56	1.89
Os		0.81	0.96	1.11	1.26	1.56	1.89
Os +4		0.78	0.93	1.08	1.26	1.53	1.86
Ir		0.81	0.96	1.11	1.26	1.56	1.89
Ir +3		0.78	0.93	1.08	1.26	1.53	1.89
Ir +4		0.78	0.93	1.08	1.26	1.53	1.86
Pt		0.81	0.96	1.29	1.56	1.62	1.89
Pt +2		0.81	0.96	1.29	1.53	1.59	1.89
Pt +4		0.78	0.93	1.29	1.53	1.59	1.86
Au		0.81	0.96	1.29	1.62	1.68	1.89
Au +1		0.81	0.96	1.29	1.62	1.65	1.89
Au +3		0.78	0.93	1.29	1.62	1.65	1.86
Hg		0.81	0.96	1.2	1.41	1.56	1.89
Hg +1		0.81	0.96	1.2	1.41	1.56	1.89
Hg +2		0.81	0.96	1.2	1.41	1.56	1.89
Tl		0.81	0.96	1.11	1.26	1.56	1.89
Tl +1		0.81	0.96	1.08	1.26	1.56	1.89
Tl +3		0.78	0.93	1.08	1.26	1.53	1.86
Pb		0.81	0.96	1.11	1.41	1.56	1.89
Pb +2		0.81	0.96	1.08	1.41	1.56	1.89
Pb +4		0.78	0.93	1.08	1.38	1.53	1.86
Bi		0.81	0.96	1.11	1.26	1.56	1.89
Bi +3		0.78	0.93	1.08	1.26	1.53	1.89
Bi +5		0.78	0.93	1.08	1.26	1.53	1.86
Po		0.81	0.96	1.11	1.26	1.56	1.89
At		0.81	0.96	1.11	1.26	1.56	1.89
Rn		0.81	0.96	1.11	1.26	1.56	1.89
Fr		0.81	0.96	1.11	1.26	1.56	1.89
Ra		0.81	0.96	1.11	1.26	1.56	1.89
Ra +2		0.81	0.96	1.08	1.26	1.56	1.89
Ac		0.81	0.96	1.11	1.26	1.56	1.89
Ac +3		0.81	0.96	1.08	1.26	1.56	1.89
Th		0.81	0.96	1.11	1.26	1.56	1.89
Th +4		0.78	0.93	1.08	1.26	1.56	1.89
Pa		0.81	0.96	1.11	1.26	1.56	1.89
U		0.81	0.96	1.11	1.26	1.56	1.89
U +3		0.81	0.96	1.08	1.26	1.56	1.89
U +4		0.78	0.93	1.08	1.26	1.56	1.89
U +6		0.78	0.93	1.08	1.26	1.53	1.86
Np		0.81	0.96	1.11	1.26	1.56	1.89

Element, Charge	Resolution [Å]:	0.5	1.0	1.5	2.0	2.5	3.0
Np +3		0.81	0.96	1.08	1.26	1.56	1.89
Np +4		0.78	0.93	1.08	1.26	1.56	1.89
Np +6		0.78	0.93	1.08	1.26	1.53	1.86
Pu		0.81	0.96	1.11	1.26	1.56	1.89
Pu +3		0.81	0.96	1.08	1.26	1.56	1.89
Pu +4		0.78	0.93	1.08	1.26	1.56	1.89
Pu +6		0.78	0.93	1.08	1.26	1.53	1.86
Am		0.81	0.96	1.11	1.26	1.56	1.89
Cm		0.81	0.96	1.11	1.26	1.56	1.89
Bk		0.81	0.96	1.11	1.26	1.56	1.89
Cf		0.81	0.96	1.08	1.26	1.56	1.89

Table B.1: The updated configuration file for the electron density radius determination is given. All elements with their respective charges are grouped with the resolution interval and b factor dependent electron density radius offsets in Ångstrom.

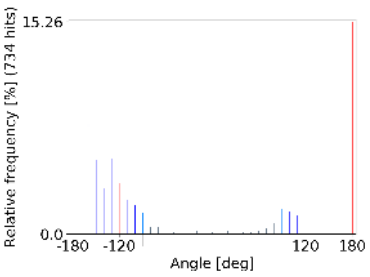
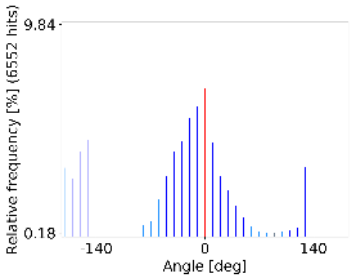
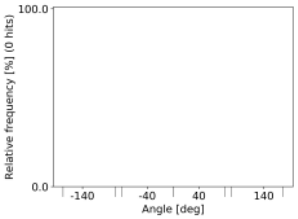
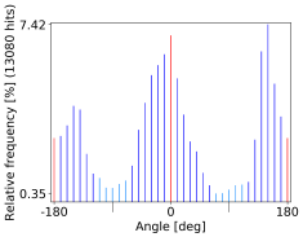
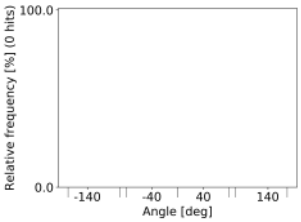
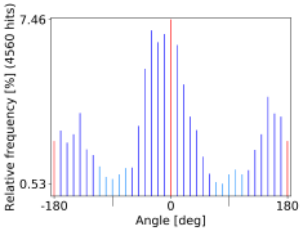
Torsion	[O : 1] = [C : 2]([O−])	[N : 1][CX4 : 2]
Library	!@[CX4H1 : 3][H : 4]	!@[CX3 : 3] = [O : 4]
TorLib16		
TorLib18		
TorLib18 validation		

Figure B.1: A change in protonation results in a change in the matched torsion rule with diverging peaks. Further analysis show the torsion rule [O : 1] = [C : 2]([O−])!@[CX4H1 : 3][H : 4] to not be found in the CSD18.

Torsion Rule SMARTS: Old	Updated
Path: CN ⇒ [c : 2]−!@[NH1 : 3][C, c]([N, n])([N, n])	
[cH0 : 1][c : 2]([cH0])−!@[NX3H1 : 3][C, c : 4]	[cH0 : 1][c : 2]([cH0])−!@[NX3H1 : 3][C, c : 4](~[**N**, **n**])(~ [**N**, **n**])
[cH0 : 1][c : 2]([cH1])−!@[NX3H1 : 3][C, c : 4]	[cH0 : 1][c : 2]([cH1])−!@[NX3H1 : 3][C, c : 4](~ [**N**, **n**])(~ [**N**, **n**])
[cH0 : 1][c : 2]([nX2H0])−!@[NX3H1 : 3][C, c : 4]	[cH0 : 1][c : 2]([nX2H0])−!@[NX3H1 : 3][C, c : 4](~ [**N**, **n**])(~ [**N**, **n**])
[cH0 : 1][c : 2]([nX3H1])−!@[NX3H1 : 3][C, c : 4]	[cH0 : 1][c : 2]([nX3H1])−!@[NX3H1 : 3][C, c : 4](~ [**N**, **n**])(~ [**N**, **n**])
[cH1 : 1][c : 2]([cH1])−!@[NX3H1 : 3][C, c : 4]	[cH1 : 1][c : 2]([cH1])−!@[NX3H1 : 3][C, c : 4](~ [**N**, **n**])(~ [**N**, **n**])
[cH1 : 1][c : 2]([nX3H1])−!@[NX3H1 : 3][C, c : 4]	[cH1 : 1][c : 2]([nX3H1])−!@[NX3H1 : 3][C, c : 4](~ [**N**, **n**])(~ [**N**, **n**])
[nX2H0 : 1][c : 2]([nX2H0])−!@[NX3H1 : 3][C, c : 4]	[nX2H0 : 1][c : 2]([nX2H0])−!@[NX3H1 : 3][C, c : 4](~ [**N**, **n**])(~ [**N**, **n**])
[nX2H0 : 1][c : 2]([nX3H1])−!@[NX3H1 : 3][C, c : 4]	[nX2H0 : 1][c : 2]([nX3H1])−!@[NX3H1 : 3][C, c : 4](~ [**N**, **n**])(~ [**N**, **n**])
[a : 1][**a** : 2]−!@[NH1 : 3][C, c : 4]	[a : 1][**c** : 2]−!@[NH1 : 3][C, c : 4](~ [**N**, **n**])(~ [**N**, **n**])
Path: CN ⇒ O = C[NX3 : 2]−!@[C : 3]	
[$(C = O) : 1][NX3 : 2]−!@[**!#**1 : 3][!#1 : 4]	[$(C = O) : 1][NX3 : 2]−!@[**C** : 3][!#1 : 4]
Path: CN ⇒ O = C[NX3 : 2]−!@[c : 3]	
[$([C](= O)([$([NX3H1]), $([NX3H2])])[NX3H1]) : 1]	[$([C](= O)([$([NX3H1]), $([NX3H2])])[NX3H1!**Rv3**]) : 1]
[NX3H1 : 2]−!@[c : 3]([nH1])[nH0 : 4]	[NX3H1 : 2]−!@[c : 3]([nH1])[nH0 : 4]
[$([C](= O)([$([NX3H1]), $([NX3H2])])[NX3H1]) : 1]	[$([C](= O)([$([NX3H1]), $([NX3H2])])[NX3H1!**Rv3**]) : 1]
[NX3H1 : 2]−!@[c : 3]([nH0])[cH1 : 4]	[NX3H1 : 2]−!@[c : 3]([nH0])[cH1 : 4]
[$([C](= O)) : 1][NX3H1 : 2]−!@[$([**a**]([nH0, o])) : 3][cH1 : 4]	[$([C](= O)) : 1][NX3H1 : 2]−!@[$([**c**]([nH0, o])) : 3][cH1 : 4]
[$([C](= O)([$([NX3H1]), $([NX3H2])])[NX3H1]) : 1]	[$([C](= O)([$([NX3H1]), $([NX3H2])])[NX3H1!**Rv3**]) : 1]
[NX3H1 : 2]−!@[cr6 : 3][nH0r6 : 4]	[NX3H1 : 2]−!@[cr6 : 3][nH0r6 : 4]
[$([C](= O)) : 1][NX3 : 2]−!@[**a** : 3](s)[a : 4]	[$([C](= O)) : 1][NX3 : 2]−!@[**c** : 3](s)[a : 4]
[$(C = O) : 1][NX3 : 2]−!@[**a** : 3]([nX2H0])[cH0 : 4]	[$(C = O) : 1][NX3 : 2]−!@[**c** : 3]([nX2H0])[cH0 : 4]
[$(C = O) : 1][NX3H1 : 2]−!@[**a** : 3]([nX2H0])[cH1 : 4]	[$(C = O) : 1][NX3H1 : 2]−!@[**c** : 3]([nX2H0])[cH1 : 4]
[$(C = O) : 1][NX3 : 2]−!@[**a** : 3]([nX2H0])[cH1 : 4]	[$(C = O) : 1][NX3 : 2]−!@[**c** : 3]([nX2H0])[cH1 : 4]

Torsion Rule SMARTS: Old	Updated
Path: CN ⇒ O = C[NX3 : 2]−!@[c : 3]	
[$(C = O) : 1][NX3 : 2]−!@[**a** : 3][nH : 4]	[$(C = O) : 1][NX3 : 2]−!@[**c** : 3][nH : 4]
[$(C = O) : 1][NX3 : 2]−!@[$([**a**]([cH1])) : 3][$([aH0](−!@O)) : 4]	[$(C = O) : 1][NX3 : 2]−!@[$([**c**]([cH1])) : 3][$([aH0](−!@O)) : 4]
[$(C = O) : 1][NX3 : 2]−!@[**a** : 3][aH0 : 4]	[$(C = O) : 1][NX3 : 2]−!@[**c** : 3][aH0 : 4]
[$(C = O) : 1][NX3H0 : 2]−!@[**a** : 3][a : 4]	[$(C = O) : 1][NX3H0 : 2]−!@[**c** : 3][a : 4]
[$(C = O) : 1][NX3H1 : 2]−!@[**a** : 3][a : 4]	[$(C = O) : 1][NX3H1 : 2]−!@[**c** : 3][a : 4]
Path: CC ⇒ [c : 2]−!@[c : 3]	
[$([cH0]([$([NX3H2]), $([NX3H1])])) : 1][**a** : 2]−!@[**a** : 3][nX2 : 4]	[$([cH0]([$([NX3H2]), $([NX3H1])])) : 1][**c** : 2]−!@[**c** : 3][nX2 : 4]
Path: CC ⇒ [c : 2]−!@[C : 3](= N)(−N)	
[a : 1][c : 2]−!@[C : 3](= [$([NH0][CX4]) : 4])	[a : 1][c : 2]−!@[C$(**CN**) : 3](= [$([NH0][CX4]) : 4])
[cH0 : 1][c : 2]([cH0])−!@[C : 3](= [N : 4])	[cH0 : 1][c : 2]([cH0])−!@[C$(**CN**) : 3](= [N : 4])
[cH0 : 1][c : 2]−!@[C : 3](= [N : 4])	[cH0 : 1][c : 2]−!@[C$(**CN**) : 3](= [N : 4])
[a : 1][c : 2]−!@[C : 3](= [N : 4])	[a : 1][c : 2]−!@[C$(**CN**) : 3](= [N : 4])
Path: CC ⇒ c([NH1, NH2, OH1])[c : 2]−!@[CX3 : 3] = O	
[$(a[OH1]) : 1][**a** : 2]−!@[CX3 : 3]([NX3H0, CX4H0, c]) = [O : 4]	[$(a[OH1]) : 1][**c** : 2]−!@[CX3 : 3]([NX3H0, CX4H0, c]) = [O : 4]
[$(a[NH1, NH2]) : 1][**a** : 2]−!@[CX3 : 3]([NX3H0, CX4H0, c]) = [O : 4]	[$(a[NH1, NH2]) : 1][**c** : 2]−!@[CX3 : 3]([NX3H0, CX4H0, c]) = [O : 4]
[cH0 : 1][c : 2]([cH1])−!@[CX3 : 3](c) = [O : 4]	[cH0 : 1]([**NH1**, **NH2**, **OH1**])[c : 2]([cH1])−!@[CX3 : 3](c) = [O : 4]
[a : 1][**a** : 2]−!@[CX3 : 3](a) = [O : 4]	[a$(**a**[**NH1**, **NH2**, **OH1**]) : 1][**c** : 2]−!@[CX3 : 3](a) = [O : 4]

Torsion Rule SMARTS: Old	Updated
Path: CC ⇒ [c : 2]−!@[CX3 : 3](= O)([NX3])	
[nr6 : 1][cr6 : 2]([nH0r6])−!@[C : 3]([NH1, NH2]) = [O : 4]	[nr6 : 1][cr6 : 2]([nH0r6])−!@[C : 3]([N**X3**H1, N**X3**H2]) = [O : 4]
[nH0r6 : 1][cr6 : 2]([cH1r6])−!@[C : 3]([NH1, NH2]) = [O : 4]	[nH0r6 : 1][cr6 : 2]([cH1r6])−!@[C : 3]([N**X3**H1, N**X3**H2]) = [O : 4]
[s : 1][c : 2]−!@[C : 3]([NH1]) = [O : 4]	[s : 1][c : 2]−!@[C : 3]([N**X3**H1]) = [O : 4]
[$([cH0][OH0]) : 1][c : 2]([cH1])−!@[C : 3](= O)[NH1 : 4]	[$([cH0][OH0]) : 1][c : 2]([cH1])−!@[C : 3](= O)[N**X3**H1 : 4]
[$([cH0][OH1]) : 1][c : 2]([cH1])−!@[C : 3](= O)[NH1 : 4]	[$([cH0][OH1]) : 1][c : 2]([cH1])−!@[C : 3](= O)[N**X3**H1 : 4]
[cH1 : 1][c : 2]([cH1])−!@[C : 3]([NH1, NH2]) = [O : 4]	[cH1 : 1][c : 2]([cH1])−!@[C : 3]([N**X3**H1, N**X3**H2]) = [O : 4]
[a : 1][c : 2]−!@[C : 3]([NH0]) = [O : 4]	[a : 1][c : 2]−!@[C : 3]([N**X3**H0]) = [O : 4]
[a : 1][c : 2]−!@[C : 3]([NH1, NH2]) = [O : 4]	[a : 1][c : 2]−!@[C : 3]([N**X3**H1, N**X3**H2]) = [O : 4]
Path: CC ⇒ [c : 2]−!@[CX3 : 3] = O	
[nX3H1 : 1][**a** : 2]−!@[CX3 : 3] = [O : 4]	[nX3H1 : 1][**c** : 2]−!@[CX3 : 3] = [O : 4]
[nX2H0 : 1][**a** : 2]([nX2H0])−!@[CX3 : 3] = [O : 4]	[nX2H0 : 1][**c** : 2]([nX2H0])−!@[CX3 : 3] = [O : 4]
[a : 1][**a** : 2]−!@[CX3 : 3] = [O : 4]	[a : 1][**c** : 2]−!@[CX3 : 3] = [O : 4]
Path: CC ⇒ [c : 2]−!@[CX3 : 3] = [CX3]	
[a : 1][**a** : 2]−!@[CX3 : 3] = [CX3H2 : 4]	[a : 1][**c** : 2]−!@[CX3 : 3] = [CX3H2 : 4]
[a : 1][**a** : 2]−!@[CX3 : 3] = [CX3H1 : 4]	[a : 1][**c** : 2]−!@[CX3 : 3] = [CX3H1 : 4]

Hierarchy Sub Class SMARTS: Old	Updated
Path: CC	
c([NH1, NH2, OH1])[c : 2]−!@[CX3 : 3] = O	**a**([NH1, NH2, OH1])[c : 2]−!@[CX3 : 3] = O

Table B.2: SMARTS pattern of torsion rules and sub hierarchies transformed. The sub hierarchy is transformed to be more generic, while the SMARTS patterns in the torsion rules are updated to be more specific based on the definition of their sub hierarchy.

Hierarchy Sub Class SMARTS	Position	Hierarchy Sub Class SMARTS	Position
Path: CO			
[a][c : 2]−!@[O : 3]	4 ⇒ 3	[c : 2]−!@[O : 3]	3 ⇒ 4
Path: CN			
S = [C**X3** : 2]−!@[NX3 : 3]	12 ⇒ 9	S = [C : 2]−!@[NX3 : 3]	9 ⇒ 10
[n : 2]−!@[C**X3** : 3]	14 ⇒ 13	[n : 2]−!@[C : 3]	13 ⇒ 14

Table B.3: Reordered sub hierarchies.

Torsion Rule SMARTS	New Parental Hierarchy
Path: CC ⇒ [C : 2]−!@[C : 3]	
[O : 1] = [CX3 : 2]−!@[CX4H1r3 : 3][H : 4]	[CX4][CX3]
[O : 1] = [CX3 : 2]−!@[CX4r3 : 3]−!@[!#1 : 4]	[CX4][CX3]
[CX3 : 1] = [CX3 : 2]−!@[CH2 : 3][!#1 : 4]	[CX4][CX3]
[CX3 : 1] = [CX3 : 2]−!@[CH2 : 3][c : 4]	[CX4][CX3]
[CX3 : 1] = [CX3 : 2]−!@[CH2 : 3][C : 4]	[CX4][CX3]
[CX3 : 1] = [CX3 : 2]−!@[CH1 : 3](C)[C : 4]	[CX4][CX3]
[CX3 : 1] = [CX3 : 2]−!@[CH2 : 3][OX2 : 4]	[CX4][CX3]
[O : 1] = [C : 2]([O−])−!@[CX4H1 : 3][H : 4]	[CX4][CX3]
N[C : 2](= [O : 1])−!@[CH2 : 3][N : 4]	[CX4][CX3]
[N : 1][C : 2](= O)−!@[CX4H2 : 3][CX4H2 : 4]	[CX4][CX3]
[$([CX3]([C])([H])) : 1] = [CX3 : 2]([H])−!@[CH2 : 3][C : 4]	[CX4][CX3]
[$([CX3]([C])([H])) : 1] = [CX3 : 2]([H])−!@[CH1 : 3](C)[C : 4]	[CX4][CX3]
[$([CX3]([C])([H])) : 1] = [CX3 : 2]([C])−!@[CH2 : 3][C : 4]	[CX4][CX3]
[O : 1] = [CX3 : 2]([NH1])−!@[CH2 : 3][C : 4]	[CX4][CX3]
[O : 1] = [CX3 : 2]([NH1])−!@[CH2 : 3][CX3 : 4] = O	[CX4][CX3]

Table B.4: Torsion rules send into a child hierarchy.

Torsion Rule SMARTS	Fitting Child Hierarchies
Path: NC	
[$([CX3] = O) : 1][NX3H1 : 2]−!@[CX4H2 : 3][C : 4]	O = C[NX3 : 2]−!@[C : 3][CX4 : 2][NX3 : 3]
[$([CX3] = O) : 1][NX3H0 : 2](C)−!@[CX4H2 : 3][C : 4]	O = C[NX3 : 2]−!@[C : 3][CX4 : 2][NX3 : 3]
Path: CC ⇒ [C : 2]−!@[C : 3]	
[∗ ∧ 2 : 1] [C ∧ 2 : 2]−!@[C ∧ 2 : 3] [∗ ∧ 2 : 4]	-
[∗ ∧ 2 : 1] [C ∧ 2 : 2]([!H])−!@[C ∧ 2 : 3] [∗ ∧ 2 : 4]	-
[CX3 : 1] = [CX3 : 2]−!@[CX3 : 3] = [CX3 : 4]	-
[CX3H0 : 1] = [CX3H0 : 2]−!@[CX3 : 3] = [CX3 : 4]	-
[CX3H0 : 1] = [CX3 : 2]−!@[CX3H0 : 3] = [CX3 : 4]	-
[CX3H0 : 1] = [CX3H0 : 2]−!@[CX3H0 : 3] = [CX3 : 4]	-
[CX3H0 : 1] = [CX3H0 : 2]−!@[CX3 : 3] = [CX3H0 : 4]’	-
[CX3R : 1] = [CX3R : 2]−!@[CX3 : 3] = [CX3 : 4]	-
[O : 1] = [CX3 : 2]−!@[CX3 : 3] = [O : 4]−!@	-
[O : 1] = [CX3 : 2](O)	
−!@[CX3 : 3]([$([NH1, NH2, CH2])]) = [O : 4]	-
[CX3H2 : 1] = [CX3 : 2]−!@[CX3 : 3] = [C : 4]	-

Table B.5: Torsion rules with problems when sending to lower level child hierarchies. In two cases, more than one possible sub hierarchy is available. For the rest, no matching sub hierarchies are available.

Torsion Rule SMARTS	Position Change	Torsion Rule SMARTS	Position Change
Path: GG			
[∗ : 1] [NX2 : 2]−!@[SX4 : 3] [∗ : 4]	51 ⇒ 50	[∗ : 1] [N, n : 2]−!@[S : 3] [∗ : 4]	50 ⇒ 51
[∗ : 1] [NX2 : 2]−!@[SX3 : 3] [∗ : 4]	52 ⇒ 51	∼	51 ⇒ 52
[∗ : 1] [NX2 : 2]−!@[SX2 : 3] [∗ : 4]	53 ⇒ 52	∼	52 ⇒ 53
Path: CO ⇒ [a][c : 2]−!@[O : 3]			
[nX2H0 : 1][c : 2]([cH0])−!@[O : 3][CX4H0 : 4]	16 ⇒ 13	[a : 1][c : 2]([a])−!@[O : 3][CX4H0 : 4]	13 ⇒ 14
Path: NC ⇒ O = C[NX3 : 2]−!@[c : 3]			
[$([C](= O)) : 1][NX3H1 : 2] −!@[c : 3]([cH])[nX2H0 : 4]	5 ⇒ 2	[$([C](= O)) : 1][NX3H1 : 2] −!@[$([c]([nH0, o])) : 3][cH1 : 4]	2 ⇒ 3
[$(C = O) : 1][NX3H1 : 2]−!@[c : 3]([nX2H0])[cH1 : 4]	10 ⇒ 2	[$([C](= O)) : 1][NX3H1 : 2]−!@[c : 3]([cH])[nX2H0 : 4]	2 ⇒ 3
[$(C = O) : 1][NX3H0 : 2]−!@[c : 3]([cH0])[cH : 4]	19 ⇒ 16	[$(C = O) : 1][NX3 : 2]−!@[c : 3][aH0 : 4]	16 ⇒ 17
[$(C = O) : 1][NX3H1 : 2]−!@[c : 3]([cH0])[cH : 4]	20 ⇒ 17	∼	17 ⇒ 18
[$(C = O) : 1][NX3H0 : 2]−!@[c : 3]([cH0])[cH0 : 4]	21 ⇒ 18	∼	18 ⇒ 19
[$(C = O) : 1][NX3H1 : 2]−!@[c : 3]([cH0])[cH0 : 4]	22 ⇒ 19	∼	19 ⇒ 20
Path: NC ⇒ O = [C : 2]−!@[NX3 : 3]			
[O : 1] = [C : 2]([CX4]) −!@[$([NX3](c([nX2H0])([nX2H0]))) : 3][H : 4]	6 ⇒ 4	[O : 1] = [C : 2]([!$([NH1])]) −!@[NX3H1 : 3]([H : 4])[$(c([nX2H0])([nX2H0]))]	4 ⇒ 5
Path: NC ⇒ S = [C : 2]−!@[NX3 : 3]			
[S : 1] = [C : 2]([$([NX3H1]), $([NX3H2])]) −!@[$([NX3]c[nH0]) : 3][H : 4]	1 ⇒ 0	[S : 1] = [C : 2]([$([NX3H1]), $([NX3H2])]) −!@[$([NX3](cn)) : 3][H : 4]	0 ⇒ 1
Path: NC ⇒ a[a : 2]−!@[N : 3]			
[cH1, nX2H0 : 1][c : 2]([cH1, nX2H0]) −!@[NX3r : 3][CX4r : 4]	3 ⇒ 2	[cH1, nX2H0 : 1][c : 2]([cH1, nX2H0]) −!@[NX3r : 3][∗ : 4]	2 ⇒ 3
[nX2H0 : 1][$(a([!nX2H0])([nX2H0])−!@[NX3H1]) : 2] −!@[NX3H1 : 3][$([CX3]([NX3H1])([NX3H1]) = O) : 4]	19 ⇒ 16	[nX2H0 : 1][$(a([!nX2H0])([nX2H0])−!@[NX3H1]) : 2] −!@[NX3H1 : 3][$([CX3](A)([NX3H1]) = O) : 4]	16 ⇒ 17

Torsion Rule SMARTS	Position Change	Torsion Rule SMARTS	Position Change
Path: NC ⇒ [CX4 : 2]−!@[NX3 : 3]			
[!#1 : 1][CX4 : 2]−!@[NX3;"N_lp" : 3][!#1 : 4]	11 ⇒ 10	[!#1 : 1][CX4 : 2]−!@[NX3;"N_lp" : 3]	10 ⇒ 11
Path: SN			
[c : 1][$(S(= O) = O) : 2]−!@[NX3H1 : 3][C : 4]	11 ⇒ 2	[c : 1][$(S(= O) = O) : 2]−!@["N_lp" : 3]	2 ⇒ 3
[C : 1][$(S(= O) = O) : 2]−!@[NX3H1 : 3][C : 4]	15 ⇒ 1	[C : 1][$(S(= O) = O) : 2]−!@["N_lp" : 3]	1 ⇒ 2
Path: CC ⇒ a([NH1, NH2, OH1])[c : 2]−!@[CX3 : 3] = O			
[$(c[OH1]) : 1][c : 2]−!@[CX3 : 3]([NX3H0]) = [O : 4]	3 ⇒ 1	[$(a[OH1]) : 1][c : 2] −!@[CX3 : 3]([NX3H0, CX4H0, c]) = [O : 4]	1 ⇒ 2
[$(c[NH1, NH2]) : 1][c : 2]−!@[CX3 : 3]([NX3H0]) = [O : 4]	4 ⇒ 0	[$([c]([NH1, NH2])) : 1][c : 2] −!@[CX3 : 3]([!O]) = [O : 4]	0 ⇒ 1

Table B.6: Reordered torsion rules per sub hierarchy. Each pattern on the left side changes its position due to the respective pattern on the right side in moving atop of it as it is detected to be more specific. (Sorting indexes start with 0.). ~ denotes the pattern in the same column in the cell above.

Torsion Rule SMARTS 1	Torsion Rule SMARTS 2
[∗ : 1] [CX4 : 2]−!@[n : 3] [∗ : 4]	[∗ : 1] [CX4 : 2]−!@[n**X3** : 3] [∗ : 4]
[∗ : 1] [CX3 : 2]−!@[n : 3] [∗ : 4]	[∗ : 1] [CX3 : 2]−!@[n**X3** : 3] [∗ : 4]
[∗ : 1] [cX3 : 2]−!@[n : 3] [∗ : 4]	[∗ : 1] [cX4 : 2]−!@[n**X3** : 3] [∗ : 4]
[$([C](= O)) : 1][NX3H1 : 2]	[$(C = O) : 1][NX3H1 : 2]
−!@[c : 3]**([cH])[nX2H0:4]**	−!@[c : 3]**([nX2H0])[cH1:4]**
[nX2H0 : 1][cr6 : 2]([cH0])	[nX2H0 : 1][cr6 : 2]([cH0])
−!@[cr6 : 3]([cH0])[nX2H0 : 4]	−!@[cr6 : 3]([cH0])[nX2H0 : 4]

Table B.7: Torsion rule duplicates. Relevant parts are marked in red.

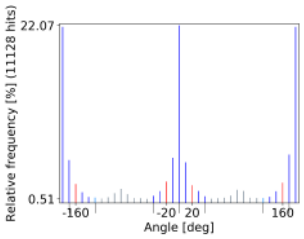
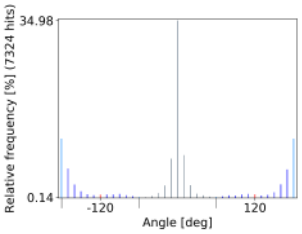
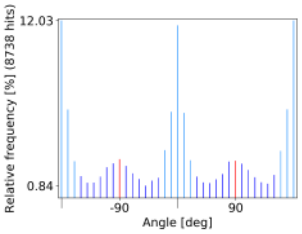
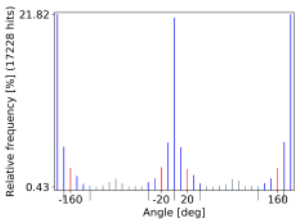
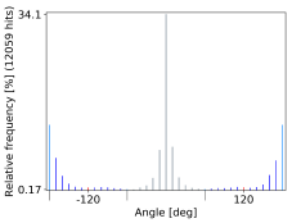
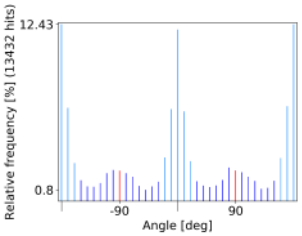
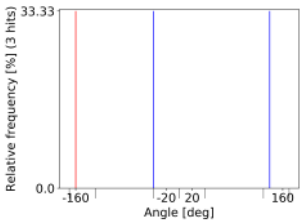
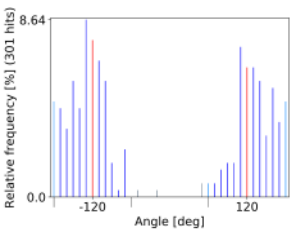
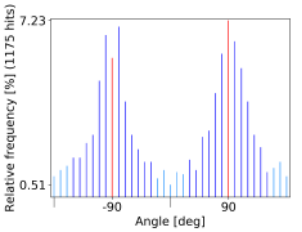
Torsion Library	[cH0 : 1][c : 2]([cH0])	[cH0 : 1][c : 2]([cH1])	[cH1 : 1][c : 2]([cH1])
TorLib16			
TorLib18			
TorLib18 validation			

Figure B.2: Hydroxy patterns in comparison to the original distribution. [cH0 : 1][c : 2]([cH0])−!@[O : 3][!C; !H : 4], [cH0 : 1][c : 2]([cH1]) −!@[O : 3][!C; !H : 4], and [cH1 : 1][c : 2]([cH1]) −!@[O : 3][!C; !H : 4] with statistics on the TorLib16, TorLib18 as well as the statistic from the validation with the TorLib18.

TorLib18	TorLib16
[$(C = O) : 1][NX3 : 2]−!@[c : 3][nH : 4]	[$(C = O) : 1][NX3 : 2]!@[a : 3][nH : 4]
[cH1 : 1][c : 2]([nX2])−!@[CX3 : 3] = [NX2 : 4]	[cH1 : 1][c : 2]([nX2])!@[CX3 : 3] = [NX2 : 4]
[cH1 : 1][c : 2]([cH0])−!@[CX3x0 : 3] = [NX2 : 4]	[cH1 : 1][c : 2]([cH0])!@[CX3x0 : 3] = [NX2 : 4]
[$(C = O) : 1][NX3H0 : 2]−!@[c : 3]([cH1])[cH1 : 4]	[$(C = O) : 1][NX3H0 : 2]!@[c : 3]([cH1])[cH1 : 4]
[$(C = O) : 1][NX3H0 : 2]−!@[c : 3]([s, o])[n : 4]	[$(C = O) : 1][NX3H0 : 2]!@[c : 3]([s, o])[n : 4]
[$(C = O) : 1][NX3H0 : 2]−!@[c : 3]([cH0])[cH0 : 4]	[$(C = O) : 1][NX3H0 : 2]!@[c : 3]([cH0])[cH0 : 4]
[$(c[NH1, NH2]) : 1][c : 2]−!@[CX3 : 3]([NX3H0]) = [O : 4]	[$(c[NH1, NH2]) : 1][c : 2]!@[CX3 : 3]([NX3H0]) = [O : 4]
[cH0 : 1][c : 2]([cH, nX2H0])−!@[NX3H1 : 3][CX4 : 4]	[cH0 : 1][c : 2]([cH, nX2H0])!@[NX3H1 : 3][CX4 : 4]
[$([C]([CX4])(= O)) : 1][NX3H1 : 2]−!@[c : 3]([nX2H0])[nX2H0 : 4]	[$([C]([CX4])(= O)) : 1][NX3H1 : 2]!@[c : 3]([nX2H0])[nX2H0 : 4]
[!#1 : 1][CX4 : 2]−!@[NX3;”N_1p” : 3][!#1 : 4]	[!#1 : 1][CX4 : 2]!@[NX3;”N_1p” : 3][!#1 : 4]
[$(C = O) : 1][NX3H0 : 2]−!@[c : 3]([cH0])[cH : 4]	[$(C = O) : 1][NX3H0 : 2]!@[c : 3]([cH0])[cH : 4]
[$([C](= O)) : 1][NX3H1 : 2]−!@[c : 3]([cH])[nX2H0 : 4]	[$([C](= O)) : 1][NX3H1 : 2]!@[c : 3]([cH])[nX2H0 : 4]
[cH1 : 1][c : 2]([cH1])−!@[CX3 : 3] = [NX2 : 4]	[cH1 : 1][c : 2]([cH1])!@[CX3 : 3] = [NX2 : 4]
[$(C = O) : 1][NX3H1 : 2]−!@[c : 3]([cH0])[cH : 4]	[$(C = O) : 1][NX3H1 : 2]!@[c : 3]([cH0])[cH : 4]
[$(C = O) : 1][NX3 : 2]−!@[$([c]([cH1])) : 3][$([aH0](−!@O)) : 4]	[$(C = O) : 1][NX3 : 2]!@[$([a]([cH1])) : 3][$([aH0](!@O)) : 4]
[$(C = O) : 1][NX3 : 2]−!@[c : 3]([nX2H0])[cH0 : 4]	[$(C = O) : 1][NX3 : 2]!@[a : 3]([nX2H0])[cH0 : 4]
[$(C = O) : 1][NX3H1 : 2]−!@[c : 3]([cH0]Cl)[cH : 4]	[$(C = O) : 1][NX3H1 : 2]!@[c : 3]([cH0]Cl)[cH : 4]
[cH1 : 1][c : 2]([$([cH0][OH1])])−!@[CX3 : 3] = [NX2 : 4]	[cH1 : 1][c : 2]([$([cH0][OH1])])!@[CX3 : 3] = [NX2 : 4]
[cH0 : 1][c : 2]([cH0])−!@[CX3!r : 3] = [NX2!r : 4]	[cH0 : 1][c : 2]([cH0])!@[CX3!r : 3] = [NX2!r : 4]
[$(C = O) : 1][NX3H1 : 2]−!@[c : 3]([cH1])[cH1 : 4]	[$(C = O) : 1][NX3H1 : 2]!@[c : 3]([cH1])[cH1 : 4]
[$(C = O) : 1][NX3H1 : 2]−!@[c : 3]([cH0])[cH0 : 4]	[$(C = O) : 1][NX3H1 : 2]!@[c : 3]([cH1])[cH1 : 4]
[$(C = O) : 1][NX3H1 : 2]−!@[c : 3]([cH0]F)[cH : 4]	[$(C = O) : 1][NX3H1 : 2]!@[c : 3]([cH0]F)[cH : 4]
[cH1 : 1][c : 2]([nX2])−!@[CX3 : 3] = [NX3 : 4]	[cH1 : 1][c : 2]([nX2])!@[CX3 : 3] = [NX3 : 4]
[$([C](= O)) : 1][NX3 : 2]−!@[c : 3](s)[a : 4]	[$([C](= O)) : 1][NX3 : 2]!@[a : 3](s)[a : 4]
[$(c[OH1]) : 1][c : 2]−!@[CX3 : 3]([NX3H0]) = [O : 4]	[$(c[OH1]) : 1][c : 2]!@[CX3 : 3]([NX3H0]) = [O : 4]

TorLib18	TorLib16
[$([C](= O)([$([NX3H1]), $([NX3H2])])[NX3H1!Rv3]) : 1][NX3H1 : 2] −!@[c : 3]([nH0])[cH1 : 4]	[$([C](= O)([$([NX3H1]), $([NX3H2])])[NX3H1]) : 1][NX3H1 : 2] !@[c : 3]([nH0])[cH1 : 4]
[$([cH0](F)) : 1][c : 2]([cH1])−!@[CX3 : 3](a) = [O : 4]	[$([cH0](F)) : 1][c : 2]([cH1])!@[CX3 : 3](a) = [O : 4]
[nX2H0 : 1][c : 2]([cH0])−!@[O : 3][CX4H0 : 4]	[nX2H0 : 1][c : 2]([cH0])!@[O : 3][CX4H0 : 4]
[$([C](= O)([$([NX3H1]), $([NX3H2])])[NX3H1!Rv3]) : 1][NX3H1 : 2]' −!@[c : 3]([nH1])[nH0 : 4]	[$([C](= O)([$([NX3H1]), $([NX3H2])])[NX3H1]) : 1][NX3H1 : 2] !@[c : 3]([nH1])[nH0 : 4]
[∗ : 1] [NX2 : 2]−!@[SX2 : 3] [∗ : 4]	[∗ : 1] [NX2 : 2]!@[SX2 : 3] [∗ : 4]
[$(C = O) : 1][NX3H1 : 2]−!@[c : 3]([s, o])[n : 4]	[$(C = O) : 1][NX3H1 : 2]!@[c : 3]([s, o])[n : 4]
[$(C = O) : 1][NX3 : 2]−!@[c : 3][aH0 : 4]	[$(C = O) : 1][NX3 : 2]!@[a : 3][aH0 : 4]
[∗ : 1] [NX2 : 2]−!@[SX4 : 3] [∗ : 4]	[∗ : 1] [NX2 : 2]!@[SX4 : 3] [∗ : 4]
[$([C](= O)([$([NX3H1]), $([NX3H2])])[NX3H1!Rv3]) : 1][NX3H1 : 2] −!@[cr6 : 3][nH0r6 : 4]	[$([C](= O)([$([NX3H1]), $([NX3H2])])[NX3H1]) : 1][NX3H1 : 2] !@[cr6 : 3][nH0r6 : 4]
[∗ : 1] [NX2 : 2]−!@[SX3 : 3] [∗ : 4]	[∗ : 1] [NX2 : 2]!@[SX3 : 3] [∗ : 4]
[O : 1] = [C : 2]([CX4])−!@[$([NX3](c([nX2H0])([nX2H0]))) : 3][H : 4]	[O : 1] = [C : 2]([CX4])!@[$([NX3](c([nX2H0])([nX2H0]))) : 3][H : 4]
[cH1, nX2H0 : 1][c : 2]([cH1, nX2H0])−!@[NX3r : 3][CX4r : 4]	[cH1, nX2H0 : 1][c : 2]([cH1, nX2H0])!@[NX3r : 3][CX4r : 4]
[$([C](= O)) : 1][NX3H1 : 2]−!@[$([c]([nH0, o])) : 3][cH1 : 4]	[$([C](= O)) : 1][NX3H1 : 2]!@[$([a]([nH0, o])) : 3][cH1 : 4]
[a : 1][c : 2]−!@[NX3H1 : 3][$([CX4r]([C; r])([C; r])) : 4]	[a : 1][c : 2]!@[NX3H1 : 3][$([CX4r]([C; r])([C; r])) : 4]
[$(C = O) : 1][NX3 : 2]−!@[c : 3]([nX2H0])[cH1 : 4]	[$(C = O) : 1][NX3 : 2]!@[a : 3]([nX2H0])[cH1 : 4]
[cH1 : 1][c : 2]([nX3H1])−!@[CX3 : 3] = [NX2 : 4]	[cH1 : 1][c : 2]([nX3H1])!@[CX3 : 3] = [NX2 : 4]

Table B.8: Torsion rules only matched in the CSD18 with TorLib18 and not in their original form and position in the TorLib14

SMARTS	Occurrence	Strained [%]
[nX3H1 : 1][c : 2]−!@[CX3 : 3] = [O : 4]	323	56.66
[O : 1] = [CX3 : 2]−!@[CX3 : 3] = [O : 4]	295	56.61
[C : 1][$(S(= O) = O) : 2]−!@[NX3H1 : 3][C : 4]	161	47.2
[$([cH0][OH0]) : 1][c : 2]([cH1])−!@[C : 3](= O)[NX3H1 : 4]	68	42.65
[∗ : 1]˜[NX2 : 2]−!@[OX2 : 3]˜[∗ : 4]	61	49.18
[∗ : 1][CX4 : 2]−!@[O : 3][$([CX3](= [!O])) : 4]	60	40
[$(c[OH1]) : 1][c : 2]−!@[CX3 : 3]([NX3H0]) = [O : 4]	53	86.79
[$(C = O) : 1][NX3 : 2]−!@[c : 3][nH : 4]	47	91.49
[cH1 : 1][c : 2]([cH1])−!@[O : 3][S : 4]	37	62.16
[$(c[NH1, NH2]) : 1][c : 2]−!@[CX3 : 3]([NX3H0]) = [O : 4]	34	64.71
[$(C = O) : 1][NX3H1 : 2]−!@[CX3 : 3] = [∗H0 : 4]	23	47.83
[$(C = O) : 1][NX3H1 : 2]−!@[CX3 : 3] = [NX2 : 4]	20	65
[nX2H0r6 : 1][cr6 : 2]([cr6])−!@[CX3 : 3]([!O]) = [O : 4]	16	87.5
[cH0 : 1][c : 2]−!@[CX4H2 : 3][!#1 : 4]	11	54.55
[a$(a[NH1, NH2, OH1]) : 1][c : 2]−!@[CX3 : 3](a) = [O : 4]	7	85.71
[$([cH0](F)) : 1][c : 2]([cH1])−!@[CX3 : 3]([O, N]) = [O : 4]	7	57.14
[cH1 : 1][c : 2]([cH0])−!@[CX3x0 : 3] = [NX2 : 4]	5	60
[nX2H0 : 1][c : 2]([!nX2H0])−!@[c : 3]([!nX2H0])[nX2H0 : 4]	5	40
[$(C = O) : 1][NX3H1 : 2]−!@[CX3 : 3] = [∗H2 : 4]	4	100
[$([C](= O)([$([NX3H1]), $([NX3H2])])[NX3H1!Rv3]) : 1] [NX3H1 : 2]−!@[c : 3]([nH1])[nH0 : 4]	4	100
[cH1 : 1][c : 2]−!@[NX2 : 3] = [$(C([NX3])N) : 4]	4	75
[!#1 : 1][CX3 : 2]−!@[SX4 : 3][!#1 : 4]	2	100
[∗ : 1]˜[CX4 : 2]−!@[SX3 : 3]˜[∗ : 4]	2	50
[c : 1][$(S(= O) = O) : 2]−!@[NX3H0 : 3][c : 4]	2	50
[O : 1] = [C : 2]([$([NX3H1]), $([NX3H2])]) −!@[$([NX3](cn)) : 3][H : 4]	2	50
[∗ : 1]˜[OX2 : 2]−!@[SX2 : 3]˜[∗ : 4]	1	100

Table B.9: Torsion rules with number of hits in PDB18 and their percentage of unlikely torsion angles.

[∗ : 1] [CX3 : 2]−!@[NX4 : 3] [∗ : 4]
[∗ : 1] [NX4 : 2]−!@[NX4 : 3] [∗ : 4]
[∗ : 1] [NX4 : 2]−!@[NX3 : 3] [∗ : 4]
[∗ : 1] [NX4 : 2]−!@[OX2 : 3] [∗ : 4]
[∗ : 1] [SX3 : 2]−!@[SX3 : 3] [∗ : 4]
[O : 1] = [C : 2]([O−])−!@[CX4H1 : 3][H : 4]
[O : 1] = [C : 2]([O−])−!@[c : 3][$(aC(= O)(O)) : 4]
[O : 1] = [C : 2]([O−])−!@[c : 3][$(a[CX3] = O) : 4]
[O : 1] = [C : 2]([O−])−!@[c : 3][nX3H1 : 4]
[O : 1] = [C : 2]([O−])−!@[c : 3][nX2H0 : 4]
[O : 1] = [C : 2]([O−])−!@[c : 3]([cH0])[cH0 : 4]
[O : 1] = [C : 2]([O−])−!@[c : 3]([cH1])[$([cH0][NH1, NH2]) : 4]
[O : 1] = [C : 2]([O−])−!@[c : 3]([cH1])[cH0 : 4]
[O : 1] = [C : 2]([O−])−!@[c : 3]([cH1])[cH1 : 4]
[O : 1] = [C : 2]([O−])−!@[c : 3][a : 4]
[c : 1][S : 2](= O)(= O)−!@[NX2H0− : 3] − [∗ : 4]
[cH0 : 1][c : 2]([nX3H1])−!@[NX3H1 : 3][C, c : 4](~ [N, n])(~ [N, n])
[cH1 : 1][c : 2]([nX3H1])−!@[NX3H1 : 3][C, c : 4](~ [N, n])(~ [N, n])
[C : 1][NH : 2]−!@[C : 3](= [NH2 : 4])[NH2]
[NH2][C : 1](= [NH2])[NH : 2]−!@[CH2 : 3][C : 4]

Table B.10: 20 Torsion rules were not hit on the CSD18 with the TorLib18 when creating the statistics.

[∗ : 1] [NX2 : 2]−!@[SX3 : 3] [∗ : 4]
[∗ : 1] [OX2 : 2]−!@[SX3 : 3] [∗ : 4]
[cH0 : 1][c : 2]([cH1])−!@[NX3H1 : 3][C, c : 4]([N, n])([N, n])
[a : 1][c : 2]−!@[NX2 : 3] = [$(C([NX3])n) : 4]
[$(C = O) : 1][NX3H0 : 2]−!@[CX3 : 3] = [∗H2 : 4]
[O : 1] = [C : 2](c)−!@[$([NX3](c([nX2H0])([nX2H0]))) : 3][H : 4]
[cH0 : 1][n : 2]−!@[CX3H0 : 3] [$([n, N](−a)) : 4]
[!#1 : 1][CX3 : 2]−!@[SX3 : 3][!#1 : 4]
[$(c[OH1]) : 1][c : 2]−!@[CX3 : 3]([NX3H0]) = [O : 4]
[a$(a[NH1, NH2, OH1]) : 1][c : 2]−!@[CX3 : 3](a) = [O : 4]
[nr6 : 1][cr6 : 2]([nH0r6])−!@[C : 3]([NX3H1, NX3H2]) = [O : 4]
[$([cH0](F)) : 1][c : 2]([cH1])−!@[CX3 : 3](a) = [O : 4]
[$([cH0](F)) : 1][c : 2]([cH1])−!@[CX3 : 3]([CX3]) = [O : 4]
[$([cH0](Cl)) : 1][c : 2]([cH1])−!@[CX3 : 3]([CX3H]) = [O : 4]
[$([cH0](Cl)) : 1][c : 2]([cH1])−!@[CX3 : 3]([CX2]) = [O : 4]
[$([cH0](Cl)) : 1][c : 2]([cH1])−!@[CX3 : 3](O) = [O : 4]
[$([cH0](Cl)) : 1][c : 2]([cH1])−!@[CX3 : 3]([CX4H2]) = [O : 4]

Table B.11: In 17 torsion rule, at least one peak score is zero.

[∗ : 1] [CX3 : 2]−!@[NX4 : 3] [∗ : 4]
[∗ : 1] [cX3 : 2]−!@[NX4 : 3] [∗ : 4]
[∗ : 1] [CX4 : 2]−!@[NX2 : 3] [∗ : 4]
[∗ : 1] [CX3 : 2]−!@[NX2 : 3] [∗ : 4]
[∗ : 1] [CX3 : 2]−!@[OX2 : 3] [∗ : 4]
[∗ : 1] [cX3 : 2]−!@[SX4 : 3] [∗ : 4]
[∗ : 1] [cX3 : 2]−!@[SX3 : 3] [∗ : 4]
[∗ : 1] [NX4 : 2]−!@[NX4 : 3] [∗ : 4]
[∗ : 1] [NX4 : 2]−!@[NX3 : 3] [∗ : 4]
[∗ : 1] [NX2 : 2]−!@[nX3 : 3] [∗ : 4]
[∗ : 1] [NX4 : 2]−!@[OX2 : 3] [∗ : 4]
[∗ : 1] [NX2 : 2]−!@[SX3 : 3] [∗ : 4]
[∗ : 1] [SX3 : 2]−!@[SX3 : 3] [∗ : 4]
[∗ : 1] [S : 2]−!@[P : 3] [∗ : 4]
[nX2H0 : 1][a : 2]−!@[a : 3]([o])[nX2H0 : 4]
[a : 1][a : 2]−!@[a : 3][$(a−!@a) : 4]
[a : 1][ar5 : 2]−!@[ar5 : 3][a : 4]
[a : 1][ar6 : 2]−!@[ar5 : 3][a : 4]
[C : 1][CH2 : 2]−!@[O : 3][CX4 : 4]
[cH0 : 1][c : 2]([cH1])−!@[O : 3][!C; !H : 4]
[cH0 : 1][c : 2]([cH0])−!@[O : 3][!#1 : 4]
[C : 1][CX4H2 : 2]−!@[OX2 : 3][!#1 : 4]
[cH0 : 1][c : 2]([nX3H1])−!@[NX3H1 : 3][C, c : 4]([N, n])([N, n])
[cH1 : 1][c : 2]([nX3H1])−!@[NX3H1 : 3][C, c : 4]([N, n])([N, n])
[C : 1][NH : 2]−!@[C : 3](= [NH2 : 4])[NH2]
[NH2][C : 1](= [NH2])[NH : 2]−!@[CH2 : 3][C : 4]
[nX2 : 1][c : 2]−!@[NX2 : 3] = [$(C([NX3])N) : 4]
[$(C = O) : 1][NX3H0 : 2]−!@[CX3 : 3] = [∗H0 : 4]
[$(C = O) : 1][NX3H0 : 2]−!@[CX3 : 3] = [∗H1 : 4]
[$(C = O) : 1][NX3H1 : 2]−!@[CX3 : 3] = [∗H0 : 4]
[$(C = O) : 1][NX3 : 2]−!@[c : 3]([nX2H0])[cH0 : 4]
[$(C = O) : 1][NX3 : 2]−!@[c : 3]([nX2H0])[cH1 : 4]
[nX2H0 : 1][a : 2]([nX2H0])−!@[NX3H0 : 3][$([CX3] = O) : 4]
[cH0 : 1][n : 2]−!@[CX3H0 : 3] [$([n, N](−a)) : 4]
[!#1 : 1][CX4H2 : 2]−!@[NX3 : 3][!#1 : 4]
[!#1 : 1][CX4 : 2]−!@[NX3 : 3][!#1 : 4]
[!#1 : 1][$(S(= O) = O) : 2]−!@[”N_lp” : 3]
[c : 1][S : 2](= O)(= O)−!@[NX2H0− : 3] − [∗ : 4]
[!#1 : 1][CX3 : 2]−!@[SX3 : 3][!#1 : 4]
[aH0 : 1][c : 2]([aH1])−!@[SX4 : 3][!#1 : 4]
[CX3R : 1] = [CX3R : 2]−!@[CX3 : 3] = [CX3 : 4]
[CX3H0 : 1] = [CX3H0 : 2]−!@[CX3 : 3] = [CX3H0 : 4]

Table B.12: In 72 torsion rule, at least one peak score is below 1.5%.

[CX3H0 : 1] = [CX3H0 : 2]−!@[CX3H0 : 3] = [CX3 : 4]
[CX3H0 : 1] = [CX3 : 2]−!@[CX3H0 : 3] = [CX3 : 4]
[∗ ∧ 2 : 1] [C ∧ 2 : 2]([!H])−!@[C ∧ 2 : 3] [∗ ∧ 2 : 4]
[O : 1] = [C : 2]([O−])−!@[CX4H1 : 3][H : 4]
[CX3 : 1] = [CX3 : 2]−!@[CH1 : 3](C)[C : 4]
[O : 1] = [CX3 : 2]−!@[CX4H1r3 : 3][H : 4]
[c : 1][CX4H2 : 2]−!@[CX3 : 3] = [O : 4]
[!#1 : 1][CX4H2 : 2]−!@[CX3 : 3] = [O : 4]
[c : 1][CX4 : 2]−!@[CX3 : 3][C : 4]
[c : 1][c : 2]−!@[c : 3][$(c−!@c) : 4]
[nX2H0 : 1][$(c([nX2H0])(a(a)(a))−!@c[nX2H0]) : 2]−!@[c : 3][nX2H0 : 4]
[c : 1][cr5 : 2]−!@[cr5 : 3][c : 4]
[nX2r6 : 1][cH0r6 : 2]([cH1r6])−!@[CX4H2 : 3][O!H : 4]
[cH0 : 1][c : 2]−!@[CX4H0 : 3][N, O, S : 4]
[cH0 : 1][c : 2]([cH0])−!@[C$(CN) : 3](= [N : 4])
[cH0 : 1][c : 2]−!@[C$(CN) : 3](= [N : 4])
[O : 1] = [C : 2]([O−])−!@[c : 3][$(aC(= O)(O)) : 4]
[O : 1] = [C : 2]([O−])−!@[c : 3][$(a[CX3] = O) : 4]
[O : 1] = [C : 2]([O−])−!@[c : 3][nX3H1 : 4]
[O : 1] = [C : 2]([O−])−!@[c : 3][nX2H0 : 4]
[O : 1] = [C : 2]([O−])−!@[c : 3]([cH0])[cH0 : 4]
[O : 1] = [C : 2]([O−])−!@[c : 3]([cH1])[$([cH0][NH1, NH2]) : 4]
[O : 1] = [C : 2]([O−])−!@[c : 3]([cH1])[cH0 : 4]
[O : 1] = [C : 2]([O−])−!@[c : 3]([cH1])[cH1 : 4]
[O : 1] = [C : 2]([O−])−!@[c : 3][a : 4]
[$(a[OH1]) : 1][c : 2]−!@[CX3 : 3]([NX3H0, CX4H0, c]) = [O : 4]
[cH0 : 1]([NH1, NH2, OH1])[c : 2]([cH1])−!@[CX3 : 3](c) = [O : 4]
[a$(a[NH1, NH2, OH1]) : 1][c : 2]−!@[CX3 : 3](a) = [O : 4]
[a : 1][c : 2]−!@[C : 3]([NX3H1, NX3H2]) = [O : 4]
[$([cH0](F)) : 1][c : 2]([cH1])−!@[CX3 : 3]([O, N]) = [O : 4]

Table B.13: In 72 torsion rule, at least one peak score is below 1.5%.

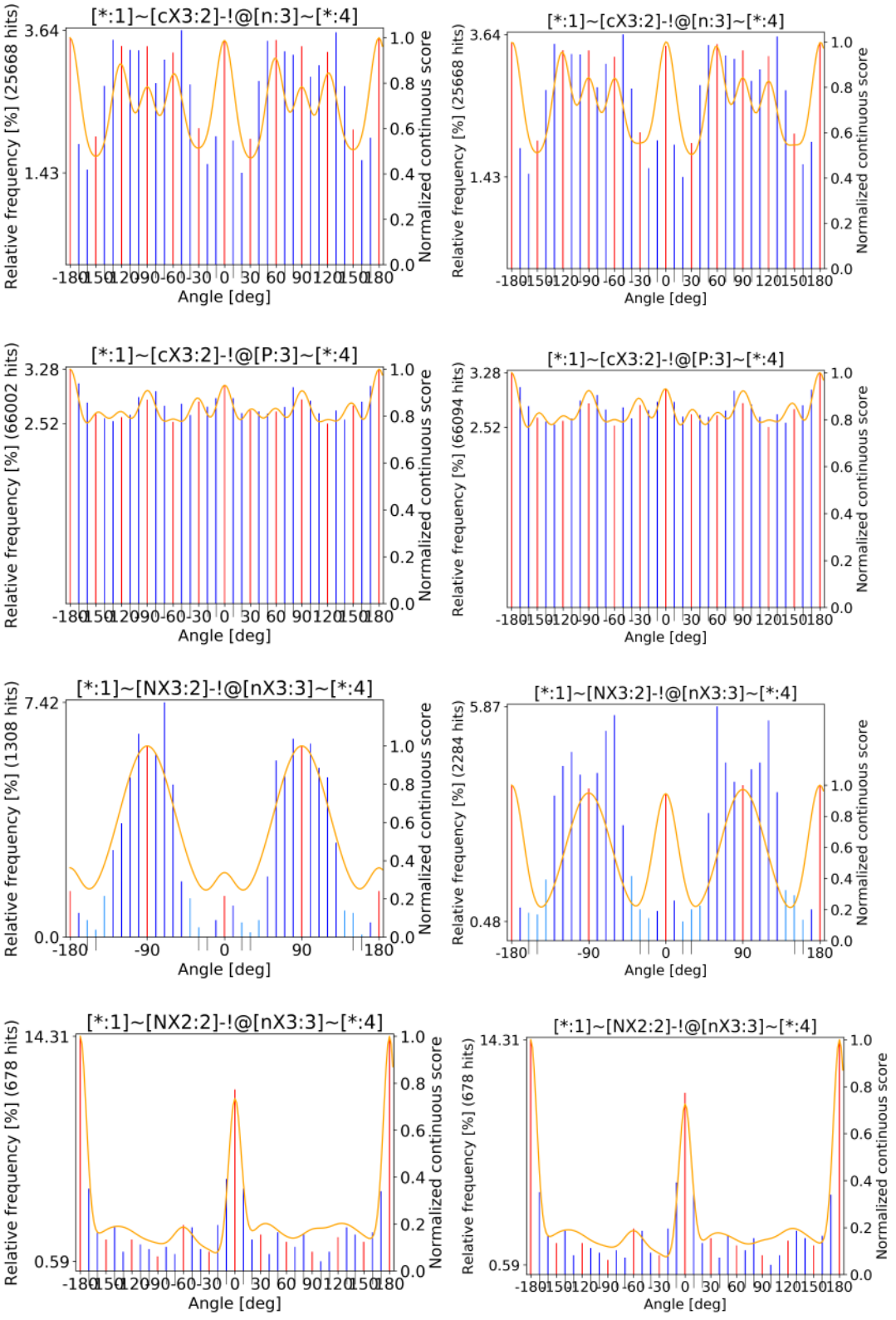


Figure B.3: SMARTS with internally reduced tolerances I

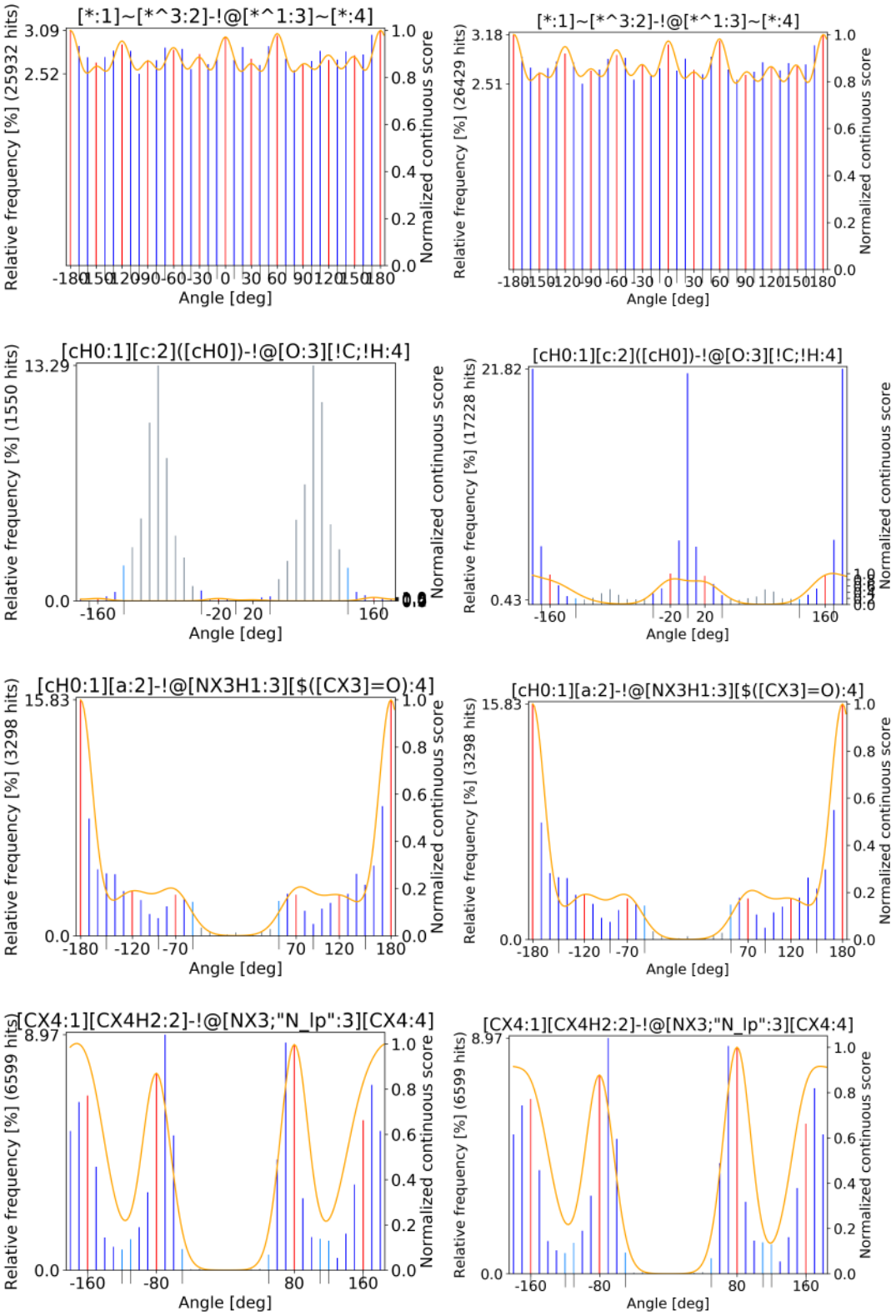


Figure B.4: SMARTS with internally reduced tolerances II

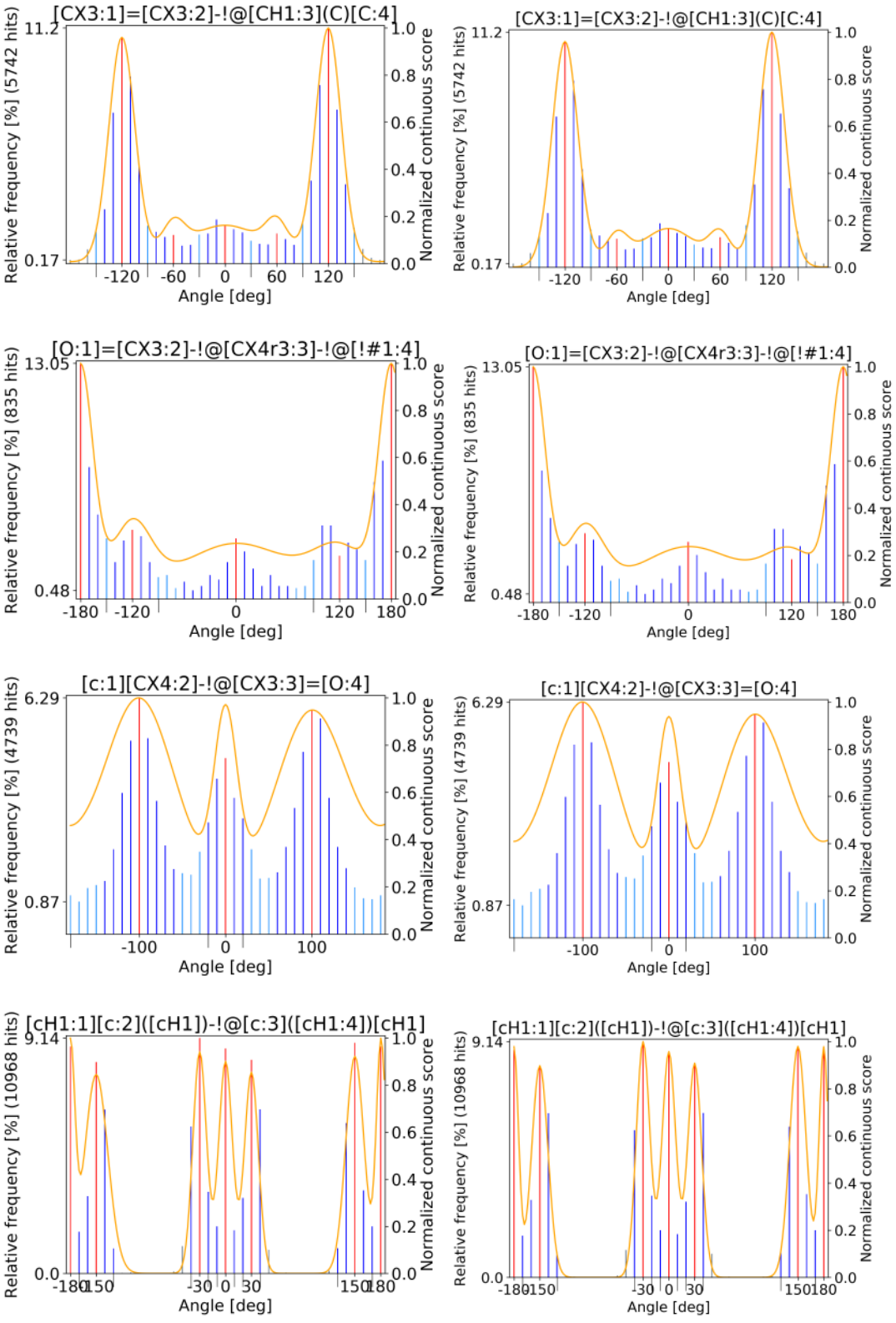


Figure B.5: SMARTS with internally reduced tolerances II

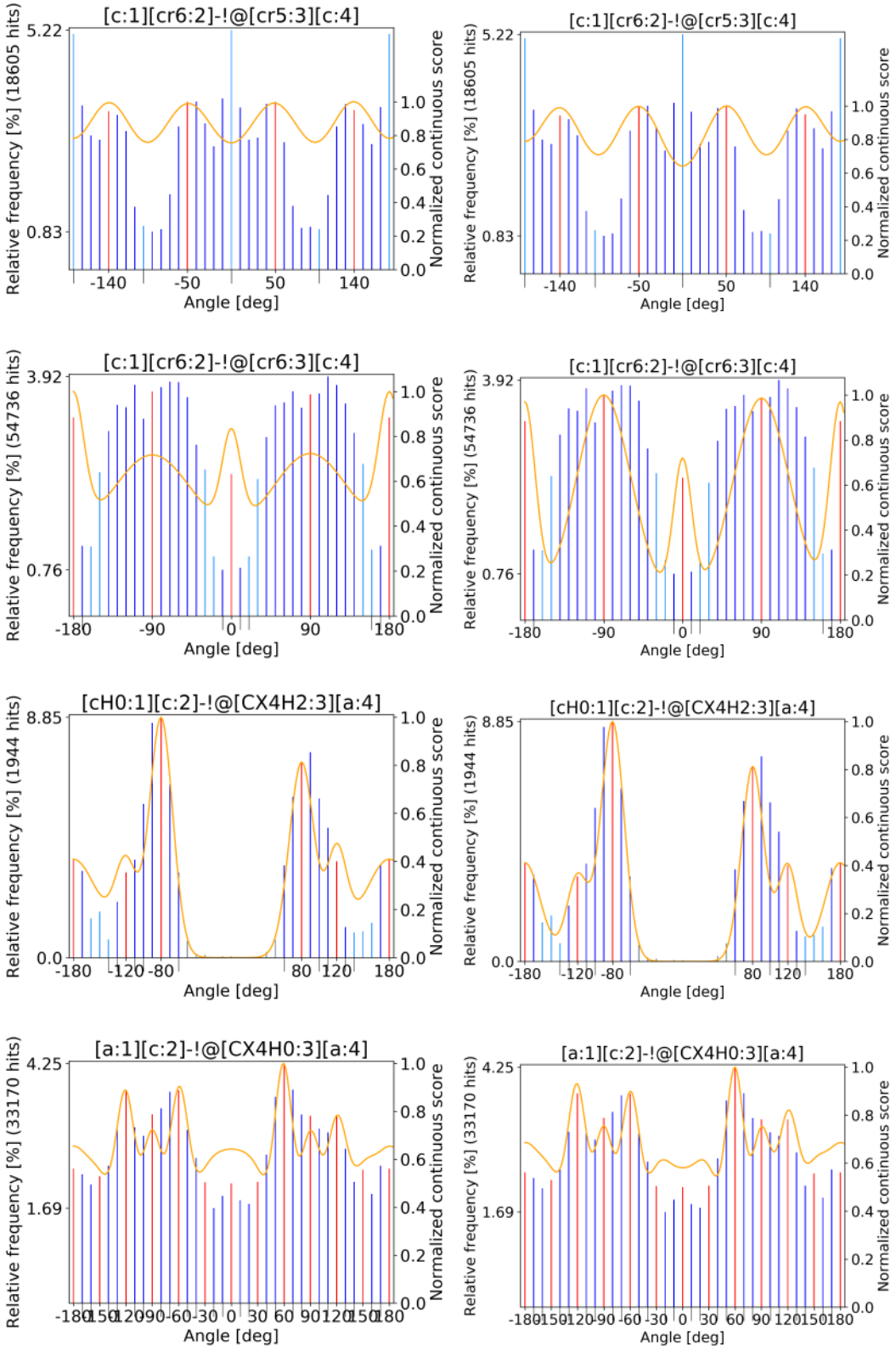


Figure B.6: SMARTS with internally reduced tolerances V

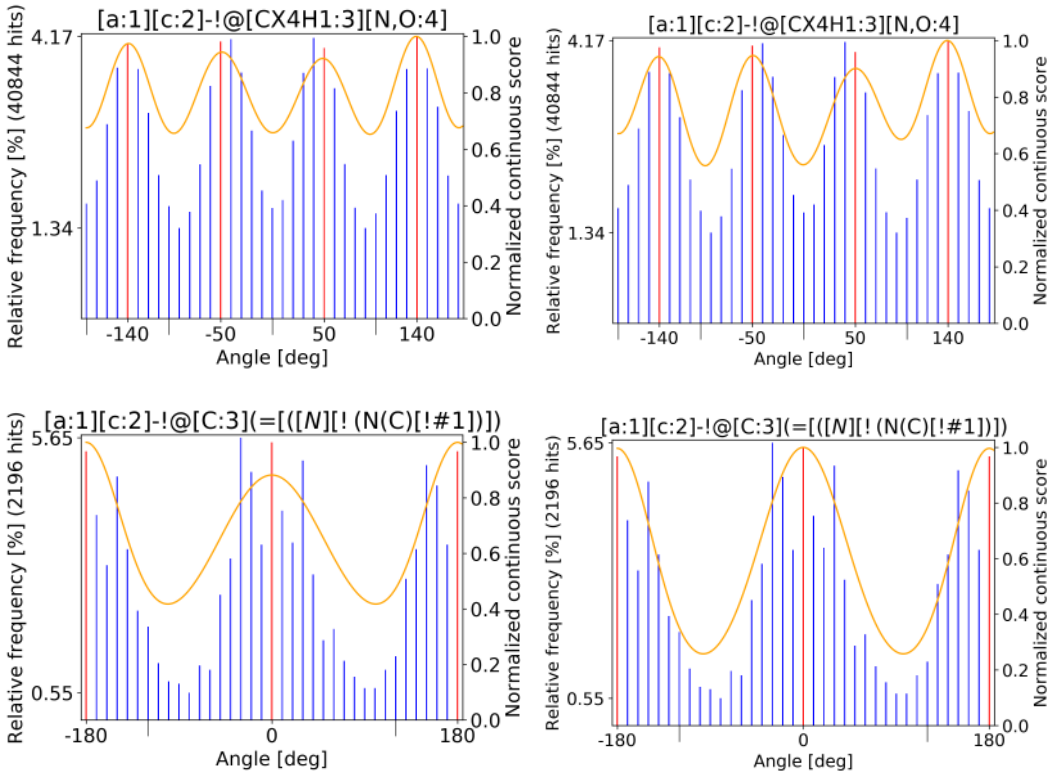


Figure B.7: SMARTS with internally reduced tolerances VI

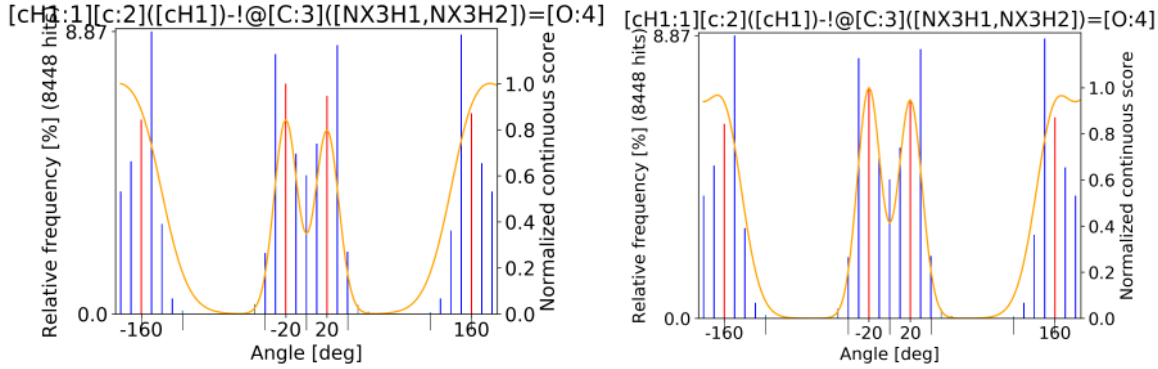
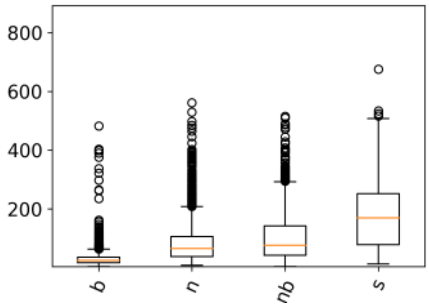
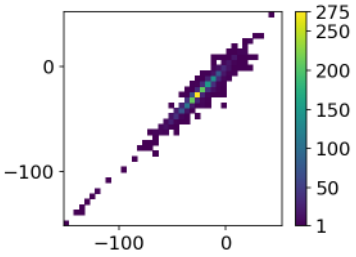


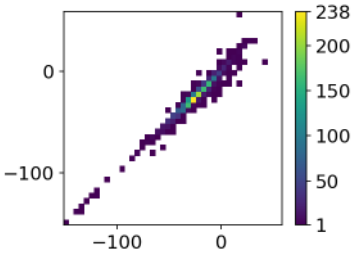
Figure B.8: SMARTS with internally reduced tolerances VII



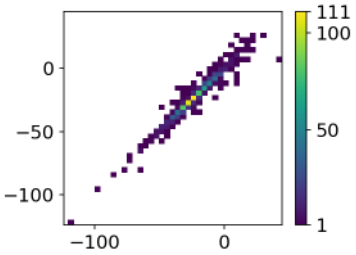
Time [s]



BOBYQA - Sbplx
(2155 points)



BOBYQA - NEWUOA
(2154 points)



BOBYQA - NEWUOA$_{bound}$
(1068 points)

Figure B.9: Computation Time of GeoHYDE optimization and their score correlations.

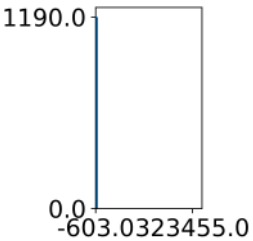


Figure B.10: GeoHYDE$_{desolv}$ Ligand score distribution before blacklisting all ligands with a positive ligand GeoHYDE$_{desolv}$ score and without limiting the x axis. The maximum socre is 323455 kJ/mol. See Figure 5.2 for more information.

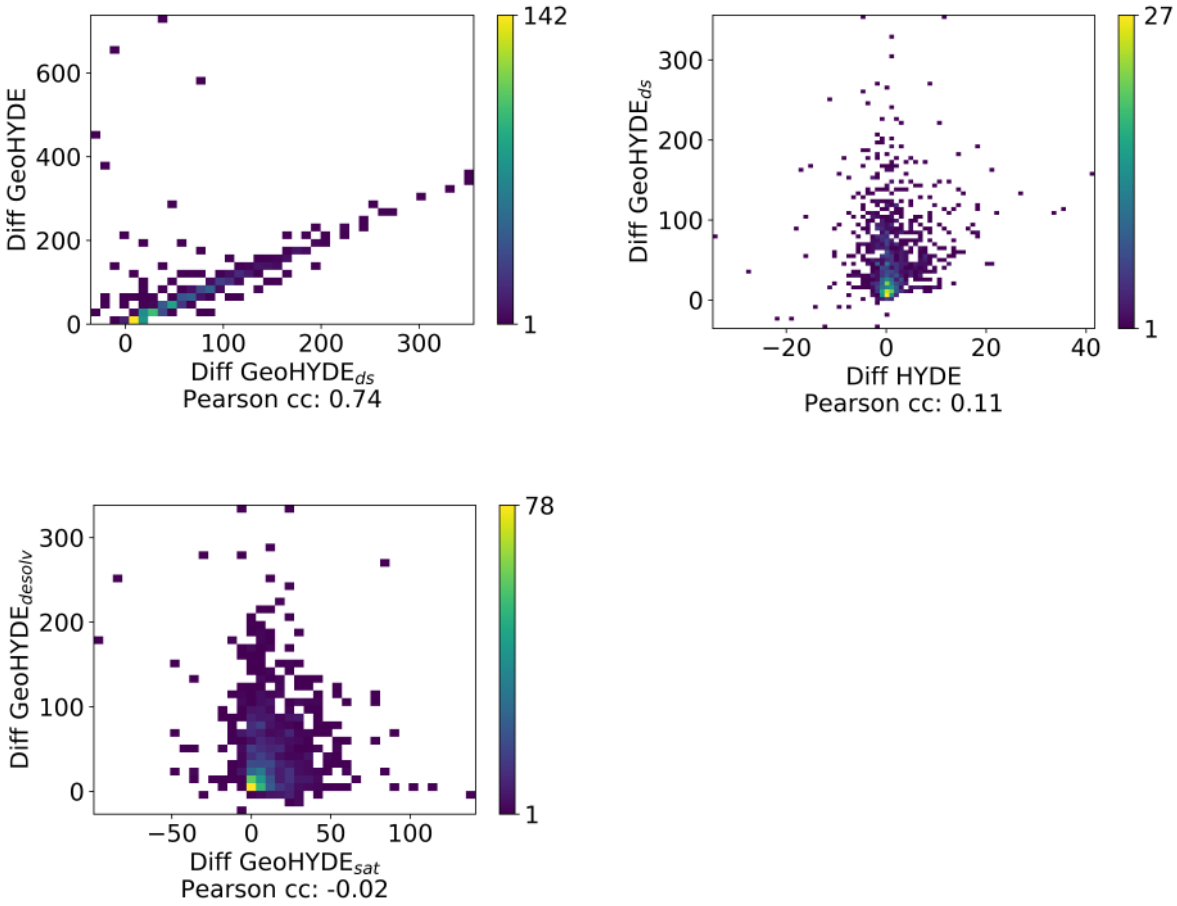
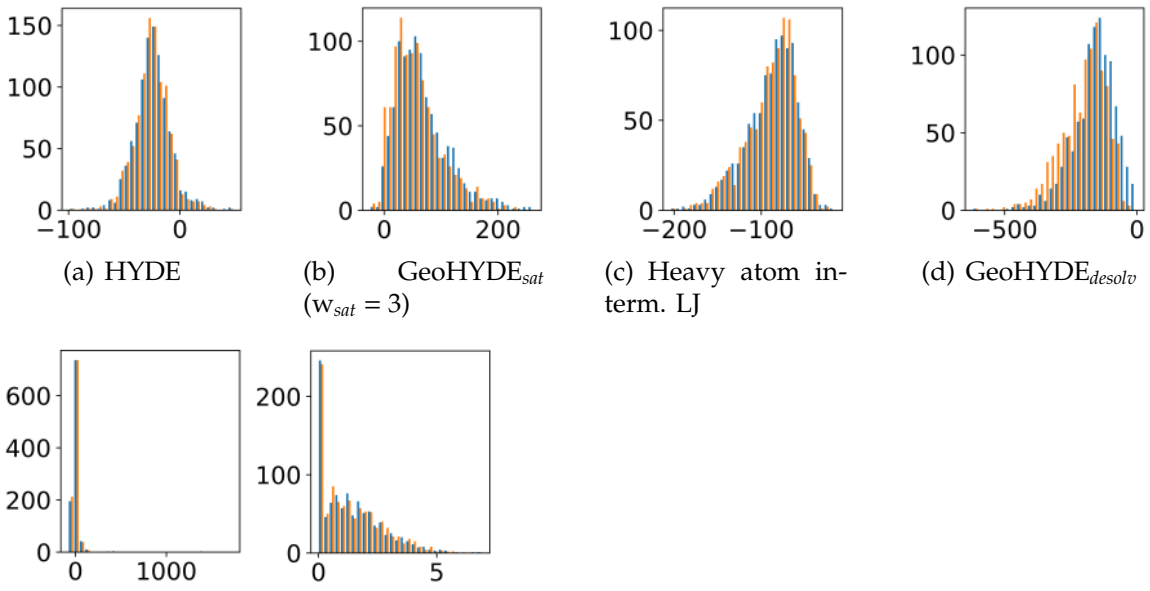


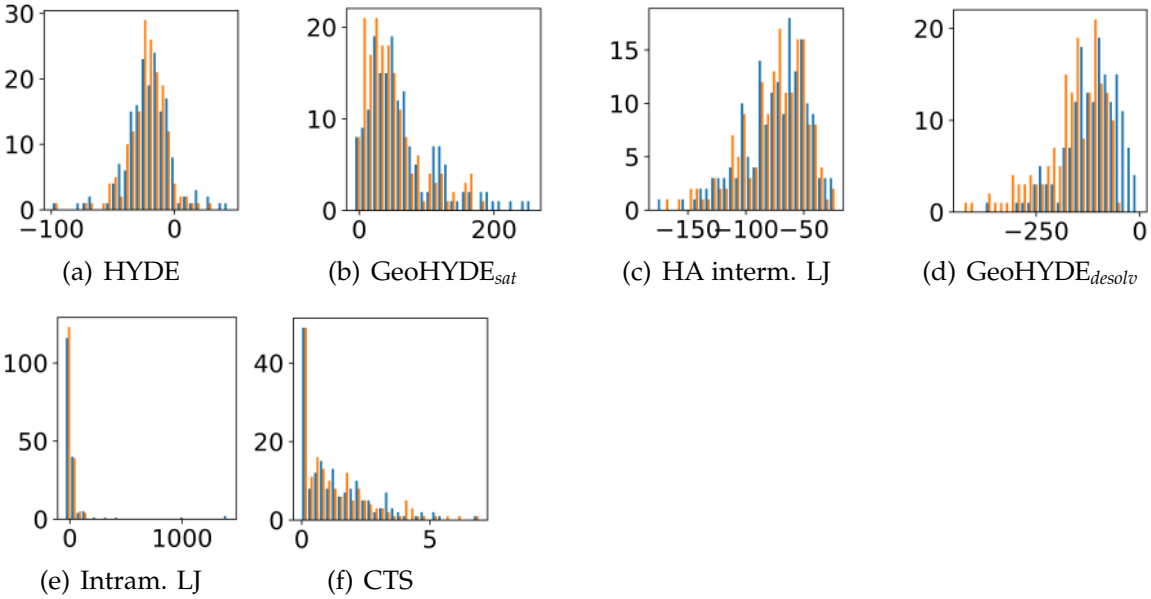
Figure B.11: Correlation of score changes annotated with their Pearson correlation coefficient.



(a) HYDE (b) GeoHYDE$_{sat}$ (w_{sat} = 3) (c) Heavy atom interm. LJ (d) GeoHYDE$_{desolv}$

(e) Intram. LJ (f) CTS

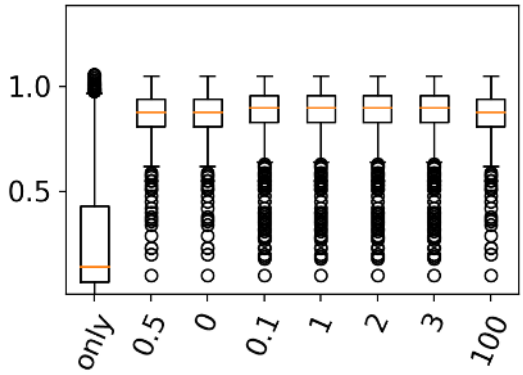
Figure B.12: Partial score shifts when using the empirical parametrization in GeoHYDE on ProtFlex18$_{train}$. Blue bars denote the initial, orange bars the final score on the x-axis while the frequency per bin is given on the y-axis.



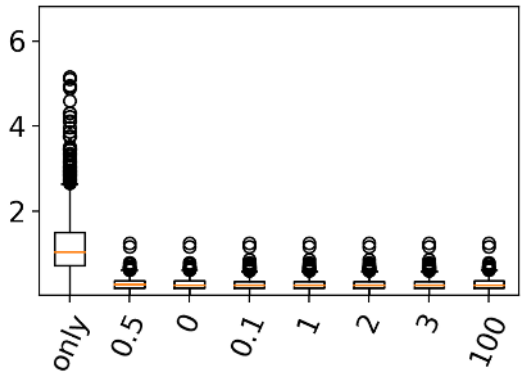
(a) HYDE (b) GeoHYDE$_{sat}$ (c) HA interm. LJ (d) GeoHYDE$_{desolv}$

(e) Intram. LJ (f) CTS

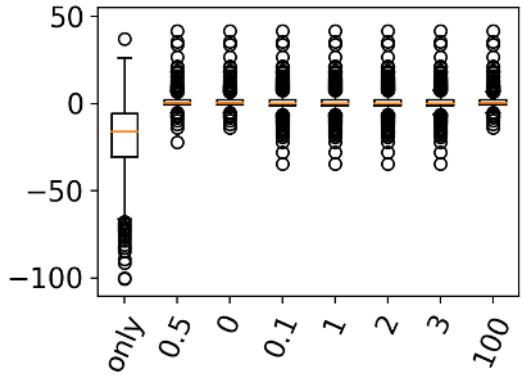
Figure B.13: Partial score shifts when using the empirical parametrization in GeoHYDE on ProtFlex18$_{train}$. Plotted are only those with a final EDIA$_m$ below 0.8.



(a) Final $EDIA_m$

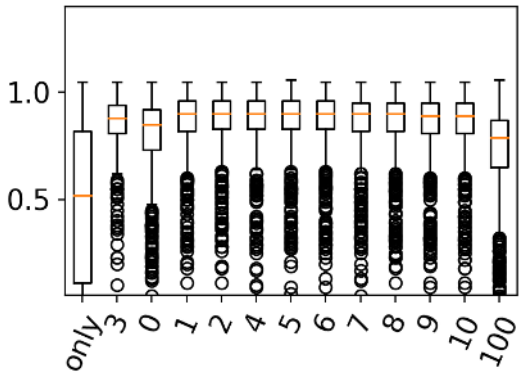


(b) Final RMSD

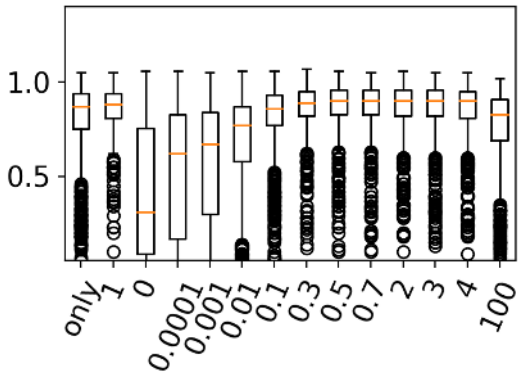


(c) HYDE score difference

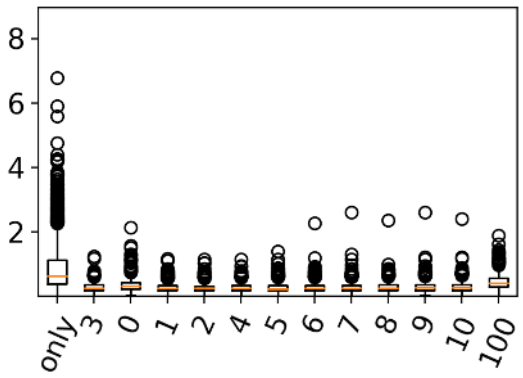
Figure B.14: The results of the parameter search for ProtFlex18$_{train}$ of w_{desolv}. The entry 'only' marks the test where only the score part of w_{desolv} was used for the optimization. The second entry, here 0.5 shows the results with the empirical determined parameter. The following entries show the results on the parameter search from zero to 100.



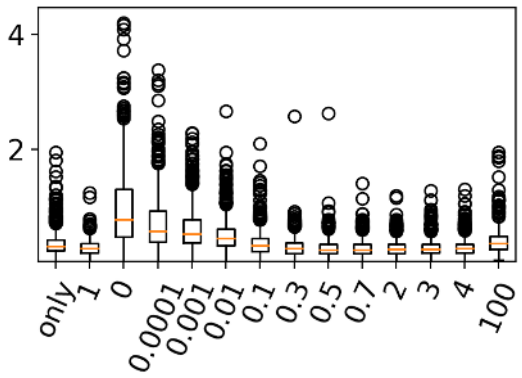
(a) Final $EDIA_m$



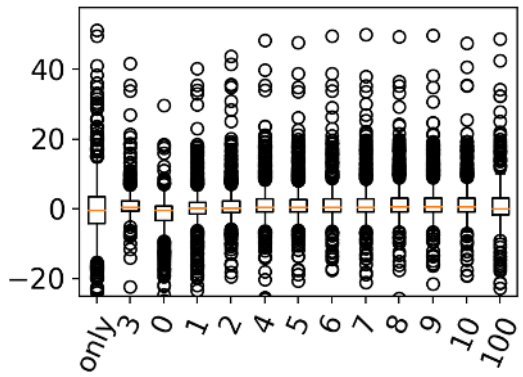
(b) Final $EDIA_m$



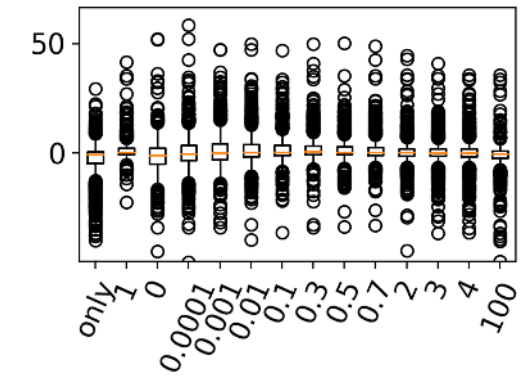
(c) Final RMSD



(d) Final RMSD

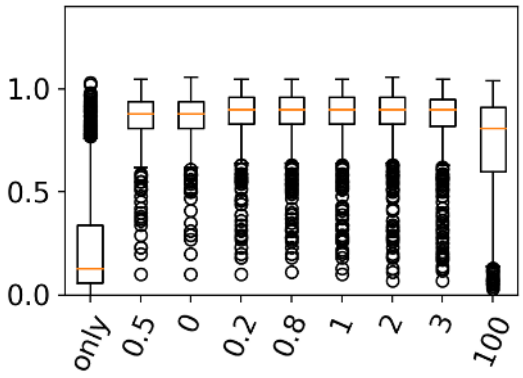


(e) HYDE score difference

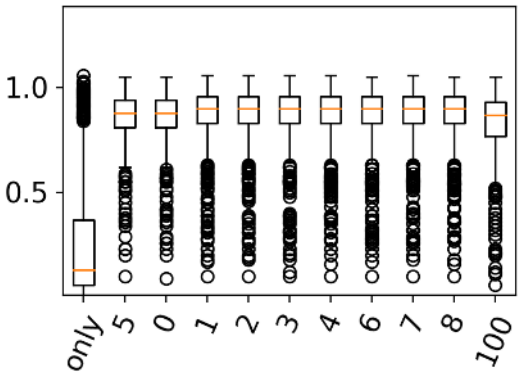


(f) HYDE score difference

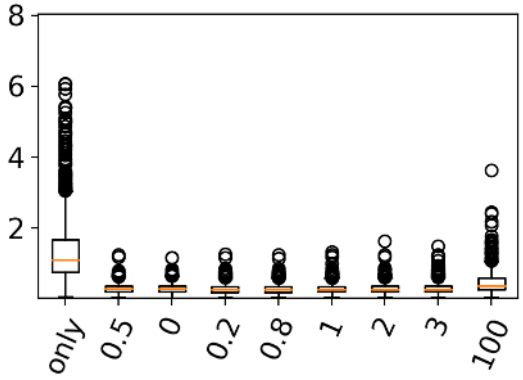
Figure B.15: The results of the parameter search for ProtFlex18$_{train}$ of w_{sat} (left) and w_{iLJ} (right). The entry 'only' marks the test where e.g. only the GeoHYDE$_{sat}$ score part was used for the optimization. The respectively second entries, here 3 and 1 show the results with the empirical determined parameter. The following entries show the results on the parameter search from zero to 100



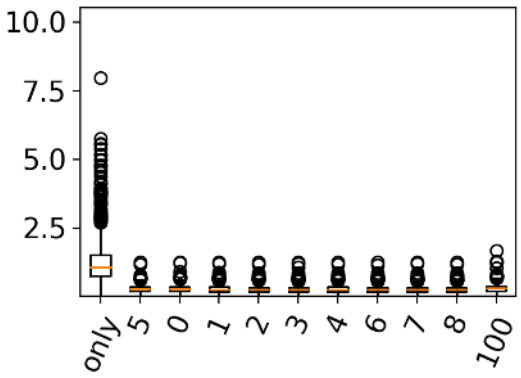
(a) Final $EDIA_m$



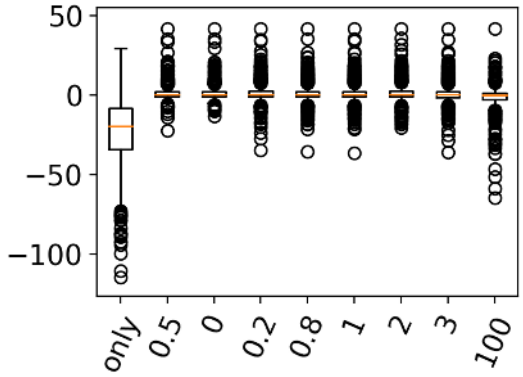
(b) Final $EDIA_m$



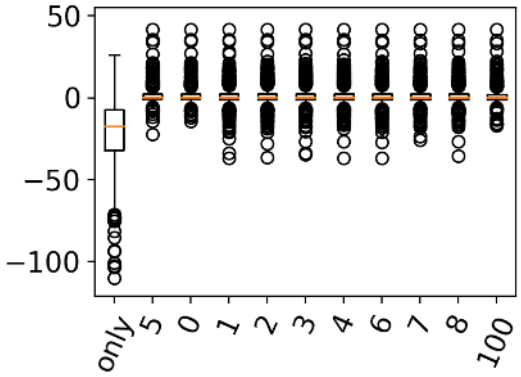
(c) Final RMSD



(d) Final RMSD



(e) HYDE score difference



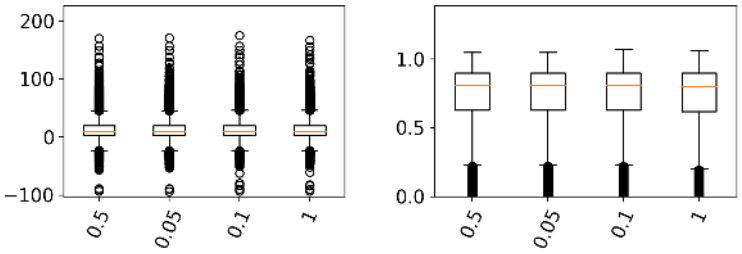
(f) HYDE score difference

Figure B.16: The results of the parameter search for ProtFlex18$_{train}$ of the intramolecular LJ potential for the ligand (w_{rLJ}, left) and CTS (w_t, right). The entry 'only' marks the test where e.g. only the CTS part was used for the optimization. The respectively second entry, here 0.5 and 5 show the results with the empirical determined parameter. The following entries show the results on the parameter search from zero to 100

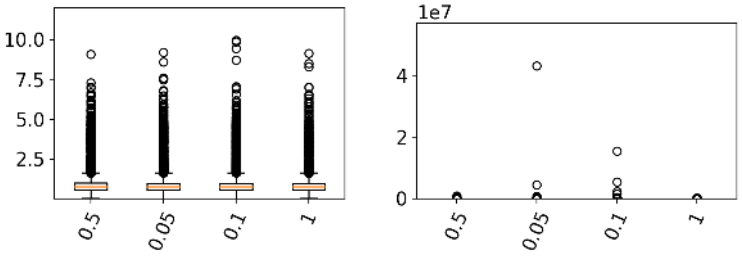
(a) GRTL: initial (b) final (c) GRT: initial (d) final

(e) T: initial (f) final (g) TS: initial (h) final

Figure B.17: Ligand configuration per sampling strategy GRTL, GRT, T and TS. Besides GRT with four configurations, the other sampling strategies resulted in 20 ligand configurations respectively.

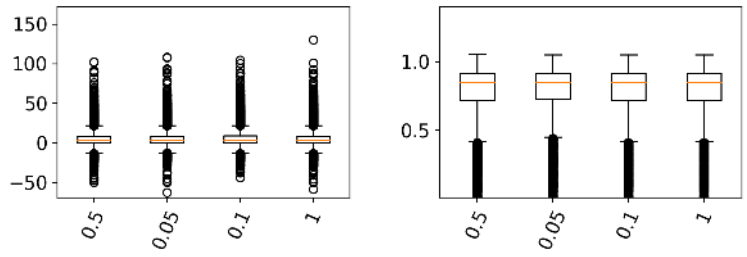


(a) HYDE (b) $EDIA_m$



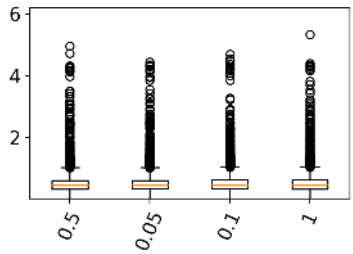
(c) RMSD (d) Intramolecular LJ

Figure B.18: Results of the parameter search for $ProtFlex18_{train}$ of the intramolecular LJ potential w_{rLJ} with the sampling configuration GRTL.

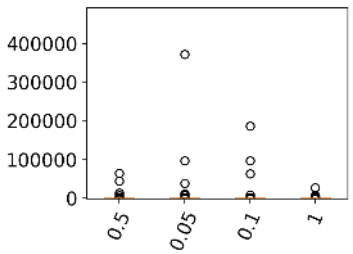


(a) HYDE

(b) EDIA$_{\mathrm{m}}$

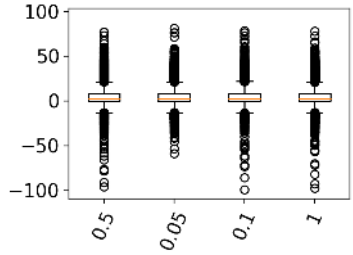


(c) RMSD

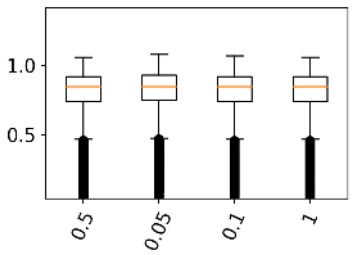


(d) Intramolecular LJ

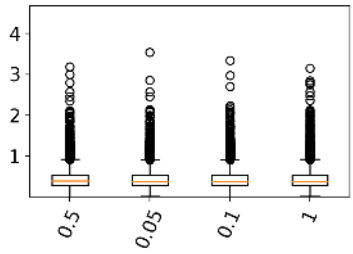
Figure B.19: Results of the parameter search for ProtFlex18$_{train}$ of the intramolecular LJ potential w_{rLJ} with the sampling configuration GRT.



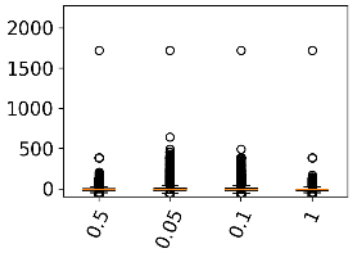
(a) HYDE



(b) EDIA$_{\mathrm{m}}$

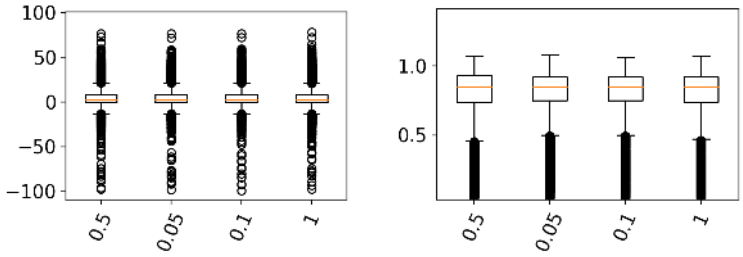


(c) RMSD



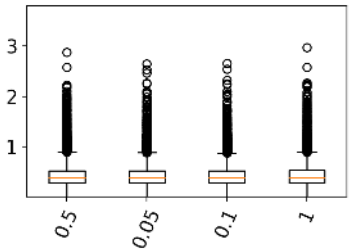
(d) Intramolecular LJ

Figure B.20: Results of the parameter search for ProtFlex18$_{train}$ of the intramolecular LJ potential w_{rLJ} with the sampling configuration T.

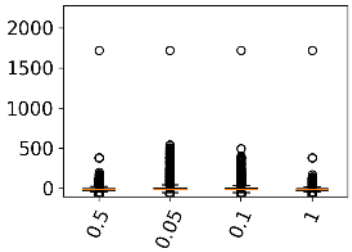


(a) HYDE

(b) $EDIA_m$

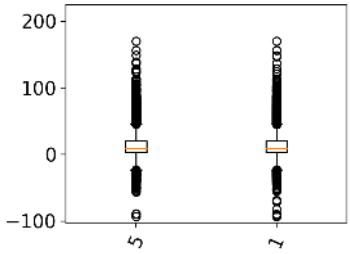


(c) RMSD

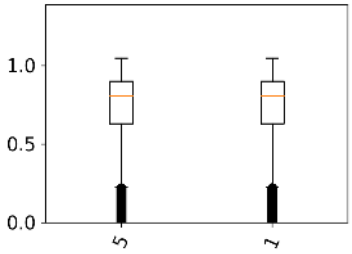


(d) Intramolecular LJ

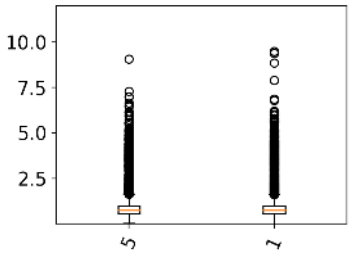
Figure B.21: Results of the parameter search for $ProtFlex18_{train}$ of the intramolecular LJ potential w_{rLJ} with the sampling configuration TS.



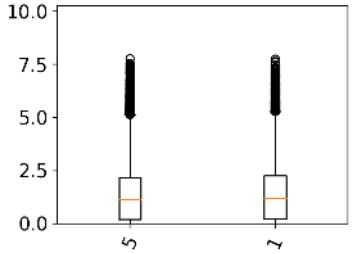
(a) HYDE



(b) $EDIA_m$

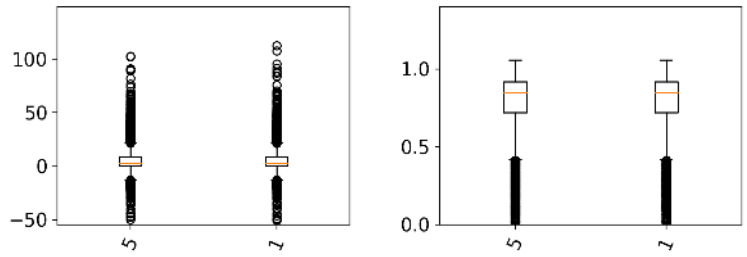


(c) RMSD

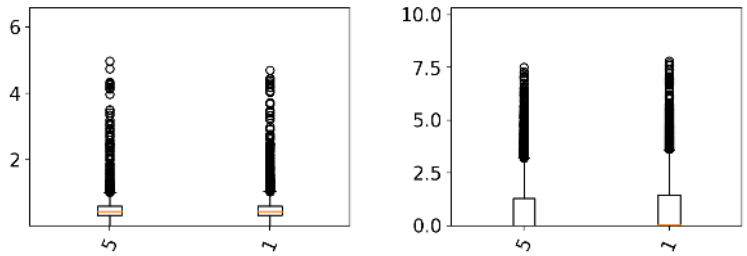


(d) CTS

Figure B.22: Results of the parameter search for $ProtFlex18_{train}$ of the CTS w_t with the sampling configuration GRTL.

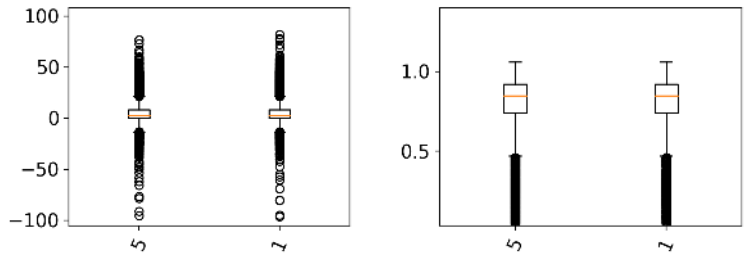


(a) HYDE (b) EDIA$_{\mathrm{m}}$

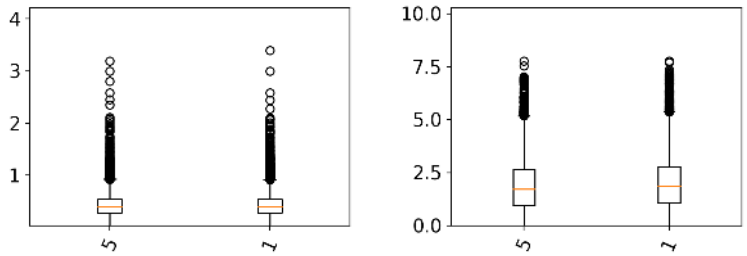


(c) RMSD (d) CTS

Figure B.23: Results of the parameter search for ProtFlex18$_{train}$ of the CTS w_t with the sampling configuration GRT.

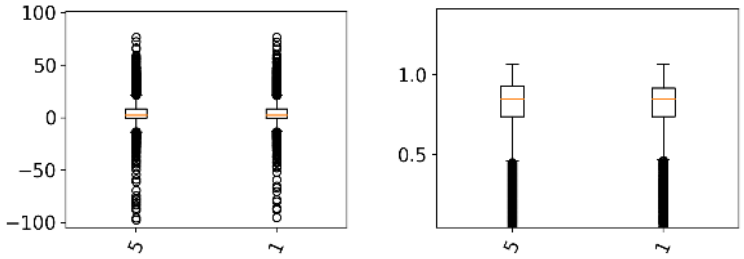


(a) HYDE (b) EDIA$_{\mathrm{m}}$



(c) RMSD (d) CTS

Figure B.24: Results of the parameter search for ProtFlex18$_{train}$ of the CTS w_t with the sampling configuration T.



(a) HYDE (b) $EDIA_m$

(c) RMSD (d) CTS

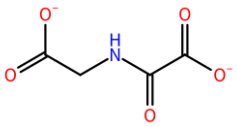
Figure B.25: Results of the parameter search for ProtFlex18$_{train}$ of the CTS w_t with the sampling configuration TS.

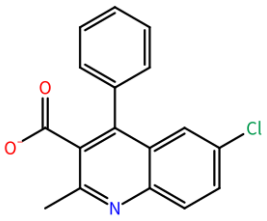
PDB Mol Id	Type	$EDIA_m$	RMSD	HYDE	$HYDE_s$	$HYDE_d$	GH_{ds}	GH_s	GH_d	CTS	LJ_{intra}
2zzd TLA C 4001	**D**	1.0	0.0	-5.39	-2.26	-3.13	-38.95	3.89	-42.85	1.67	38.4
		0.18	0.71	-5.34	-2.26	-3.08	-53.48	0.47	-53.95	3.11	-25.48
Diff initial - final		0.82	-0.71	-0.05	0.0	-0.05	14.52	3.42	11.1	-1.44	**63.89**
	D03	1.0	0.0	-5.39	-2.26	-3.13	-9.22	3.89	-13.11	8.36	19.2
		0.32	0.96	-3.24	-1.08	-2.16	-14.85	-1.1	-13.76	10.78	-9.04
Diff initial - final		0.68	-0.96	-2.15	-1.18	-0.97	5.64	4.99	0.64	-2.42	**28.25**
	C03	1.0	0.0	-5.39	-2.26	-3.13	10.09	3.89	6.2	1.67	60.05
		0.16	1.08	-3.01	-1.08	-1.93	-3.9	-3.9	0.01	2.46	0.37
Diff initial - final		0.84	-1.08	-2.38	-1.18	-1.2	13.98	7.8	6.19	-0.79	**59.68**
4c9o CAM A 423	**D**	0.84	0.0	-36.47	2.79	-39.26	-11.54	20.44	-31.98	0.0	169.77
		0.59	0.77	-41.97	-1.89	-40.08	-46.09	9.0	-55.1	0.0	169.77
Diff initial - final		0.25	-0.77	5.5	4.68	0.82	34.55	11.44	**23.11**	0.0	0.0
	D03	0.84	0.0	-36.47	2.79	-39.26	11.78	20.44	-8.66	0.0	84.88
		0.42	1.09	-39.57	-1.89	-37.68	-9.03	3.18	-12.21	0.0	84.88
Diff initial - final		0.42	-1.09	3.1	4.68	-1.58	20.81	**17.26**	3.55	0.0	0.0
	C03	0.84	0.0	-36.47	2.79	-39.26	31.93	20.44	11.49	0.0	142.66
		0.47	0.87	-40.16	-1.89	-38.27	10.51	5.03	5.48	0.0	142.66
Diff initial - final		0.37	-0.87	3.69	4.68	-0.99	21.42	**15.41**	6.01	0.0	0.0

Table B.14: Score terms for comparative analysis of the Lennard-Jones Potential.

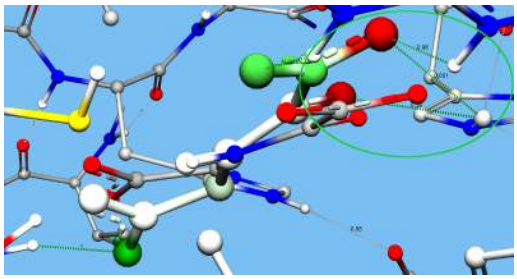
PDB Mol Id	Type	$EDIA_m$	RMSD	HYDE	$HYDE_s$	$HYDE_d$	GH_{ds}	GH_s	GH_d	CTS	LJ_{intra}
5d9y OGA A 2001	**D**	0.93	0.0	8.99	23.05	-14.06	1176.2	184.99	991.22	0.57	-9.17
		0.33	0.76	19.52	33.27	-13.75	113.1	169.64	-56.54	0.58	-9.01
Diff initial - final		0.6	-0.76	-10.53	-10.22	-0.31	1063.1	15.34	**1047.76**	-0.01	-0.16
	D03	0.93	0.0	8.99	23.05	-14.06	412.2	114.44	297.75	2.86	-4.58
		0.82	0.36	19.85	33.81	-13.96	85.02	85.89	-0.87	2.82	-4.63
Diff initial - final		0.11	-0.36	-10.86	-10.76	-0.1	327.17	28.55	**298.62**	0.04	0.05
	C03	0.93	0.0	8.99	23.05	-14.06	406.52	184.99	221.53	0.57	0.0
		0.52	0.5	-5.49	8.69	-14.18	151.53	109.2	42.33	0.56	0.0
Diff initial - final		0.41	-0.5	14.49	14.36	0.12	254.98	75.78	**179.2**	0.02	0.0
5edb 5M8 A 201	**D**	0.9	0.0	-37.74	1.2	-38.94	-92.15	11.45	-103.6	0.0	-9.67
		0.12	1.23	-35.66	-0.71	-34.95	-126.23	-0.66	-125.58	0.0	-9.67
Diff initial - final		0.78	-1.23	-2.09	1.91	-4.0	34.09	12.11	**21.98**	0.0	0.0
	D03	0.9	0.0	-37.74	1.2	-38.94	-17.26	11.45	-28.71	0.0	-4.83
		0.87	0.18	-38.76	0.46	-39.22	-19.57	9.38	-28.94	0.0	-4.83
Diff initial - final		0.03	-0.18	1.02	0.74	0.28	2.31	**2.08**	0.24	0.0	0.0
	C03	0.9	0.0	-37.74	1.2	-38.94	14.86	11.45	3.41	0.0	6.6
		0.21	0.64	-36.15	0.01	-36.16	3.66	0.04	3.61	0.0	6.6
Diff initial - final		0.69	-0.64	-1.59	1.19	-2.78	11.21	**11.41**	-0.2	0.0	0.0

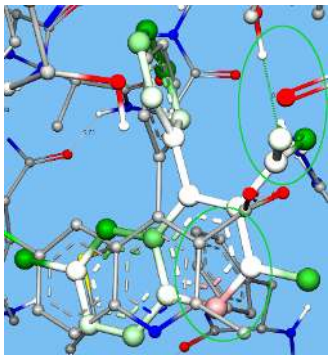
Table B.15: Score terms for comparative analysis of the Lennard-Jones Potential.





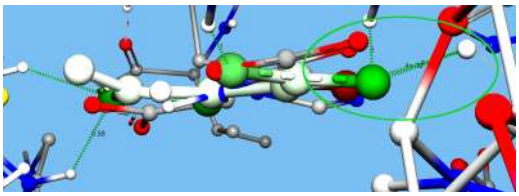
D03

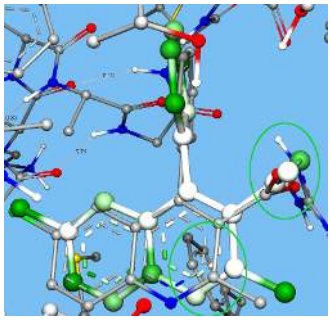




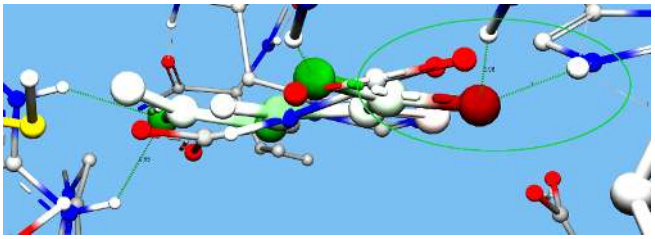
O10 interacts with serine A 54

D

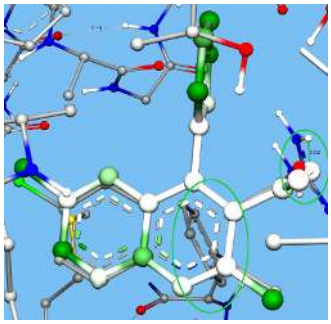




C03



5d9y OGA A 2001
O1, O2 interact with iron A 2002
O4 interacts with O3 of arginine A

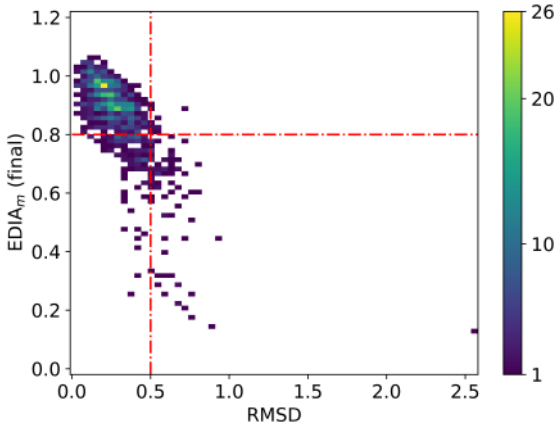


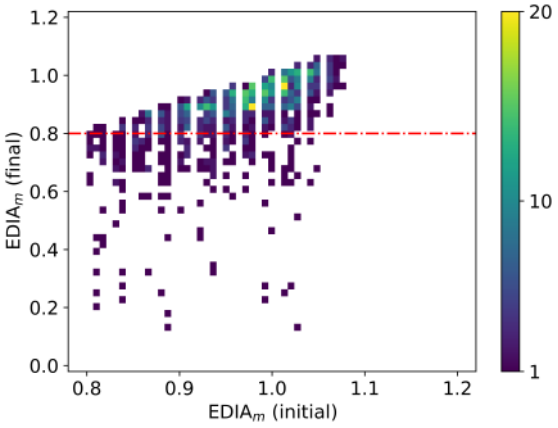
5edb 5M8 A 201
pyridine ring close to phenylalanine A 17
1896 O13 interacts with arginine A 127

Figure B.26: 5d9y OGA A 2001 and 5edb 5M8 A 201 for comparative analysis of the Lennard-Jones Potential. The original ligand is given in 2D and in element coloring in each picture of the pocket. The ligand after optimization is shown in HYDE coloring and with interactions colored in green if relevant for the HYDE score. See Figure 5.3 for the second set of pockets. Partial score terms can be found in Table B.15.

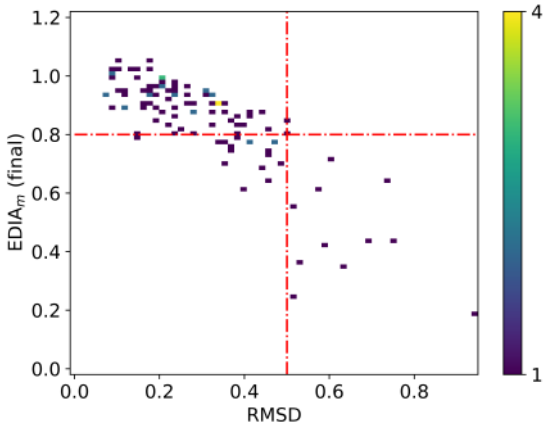
Final RMSD - Final $EDIA_m$ | Initial $EDIA_m$ - final $EDIA_m$

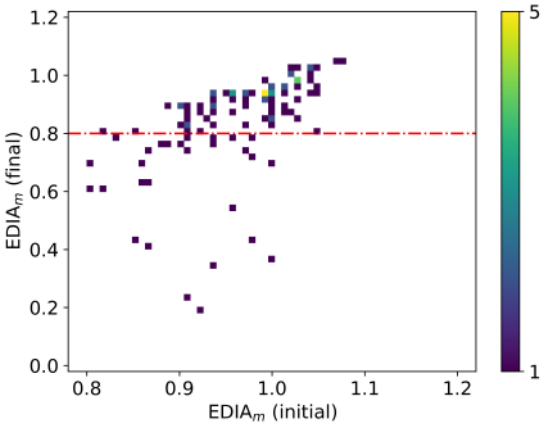
$ProtFlex18_{train}$



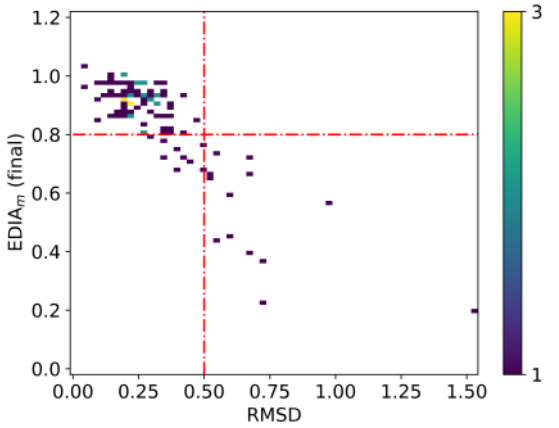


$ProtFlex18_{id}$





$ProtFlex18_{od}$



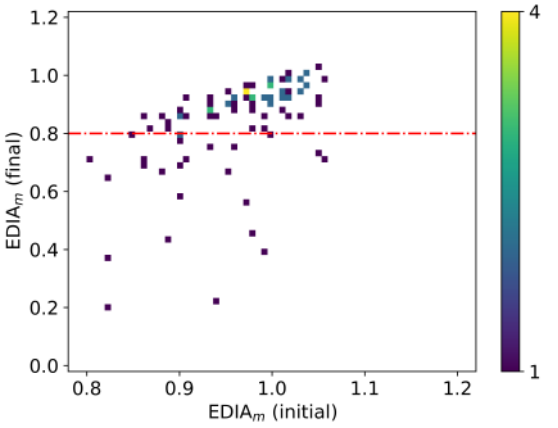
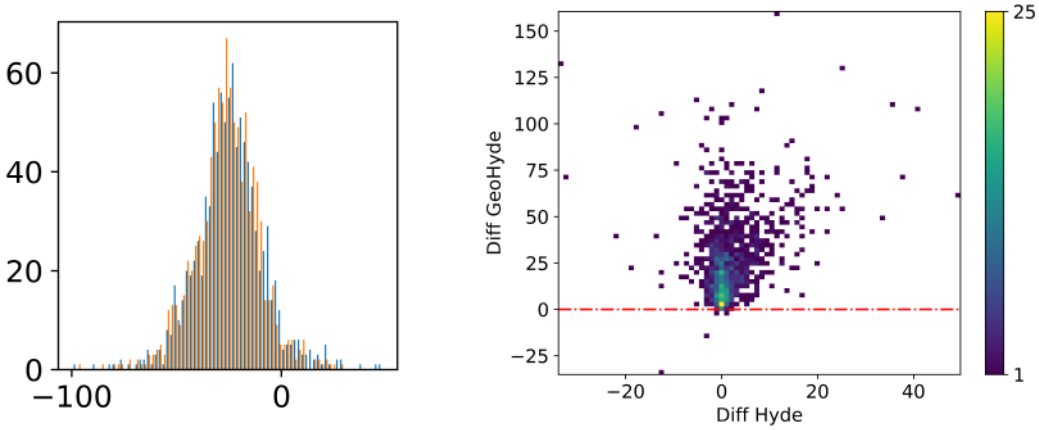


Figure B.27: Final RMSD - Final $EDIA_m$ and initial $EDIA_m$ - final $EDIA_m$ correlation over the three data sets $ProtFlex18_{train}$, $ProtFlex18_{id}$, $ProtFlex18_{od}$. Quality segment analysis can be found in Table 5.4.

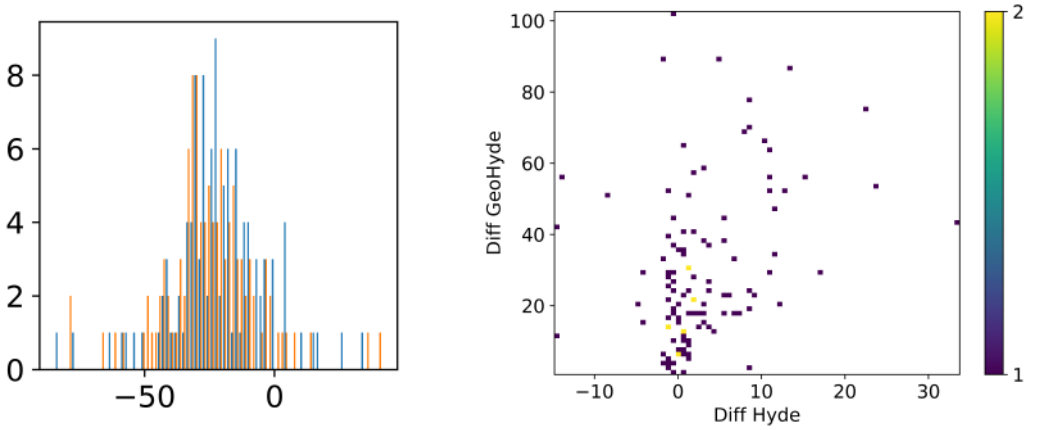
HYDE initial (blue) to HYDE final absolute scores

HYDE score difference - GeoHYDE$_{ds}$ score difference

ProtFlex18$_{train}$



ProtFlex18$_{id}$



ProtFlex18$_{od}$

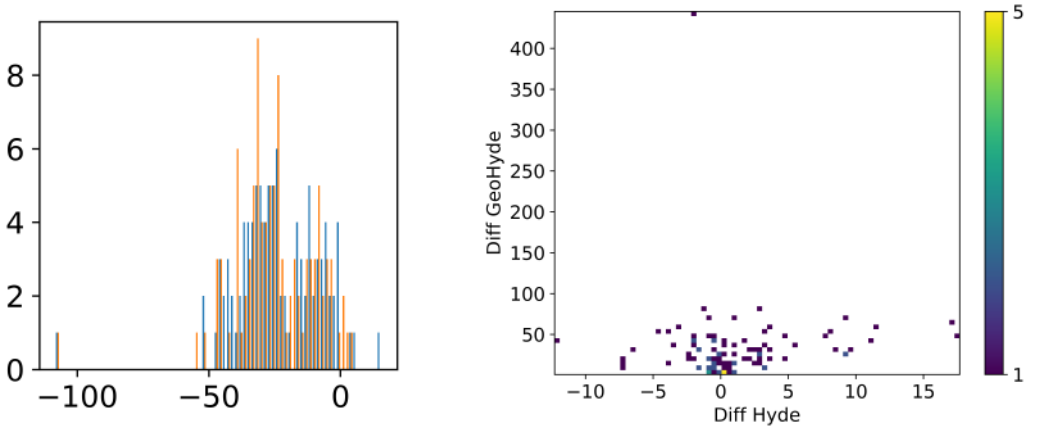
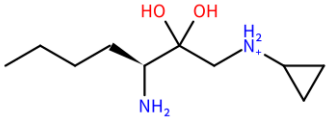
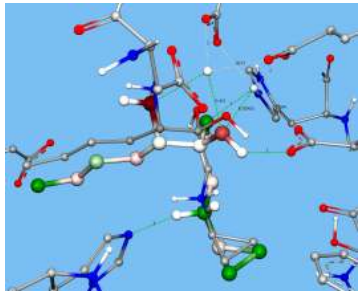
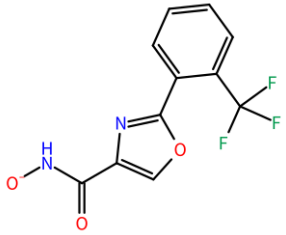


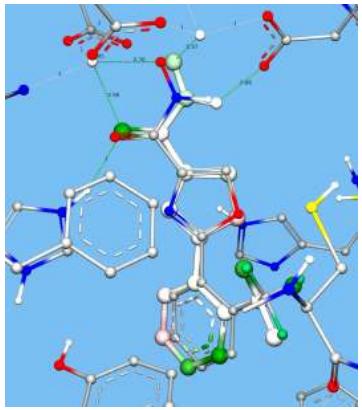
Figure B.28: ProtFlex18$_{train}$ with the final parametrization optimized by GeoHYDE



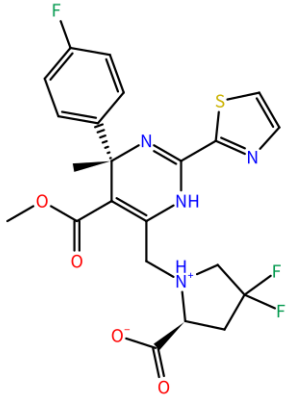


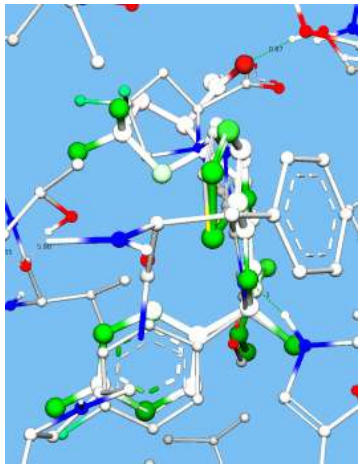
1qxw M1C A 3001
HYDE$_{diff}$: -33.96
GeoHYDE$_{ds,diff}$: 132.02



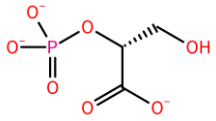


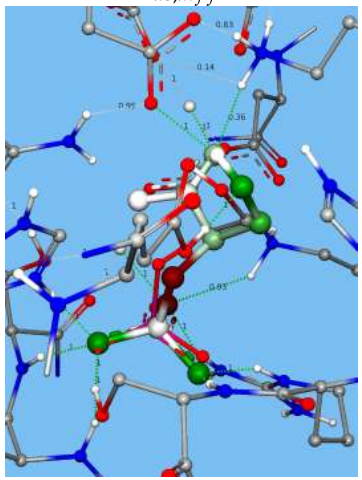
4a6v IKY B 1264
HYDE$_{diff}$: 49.72
GeoHYDE$_{ds,diff}$: 60.68





5gmz 6XU F 202
HYDE$_{diff}$: -12,03
GeoHYDE$_{ds,diff}$: -35,11



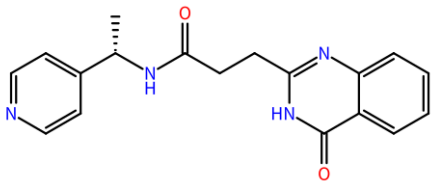


3ucd 2PG A 601
HYDE$_{diff}$: 11,80
GeoHYDE$_{ds,diff}$: 160,53

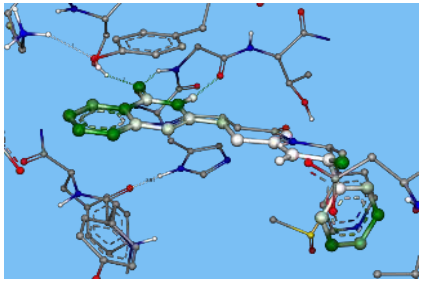
Table B.16: ProtFlex18$_{train}$ single cases with the largest score improvement or worsening for HYDE and GeoHYDE$_{sd}$.

PDB, Mol Id	$EDIA_m$	RMSD	HYDE	$HYDE_s$	$HYDE_d$	GH_{ds}	GH_s	GH_d	CTS	LJ_{intra}	os	time
4l6z 1DC A 601	0.82	0.0	-44.11	-0.75	-43.35	-26.9	27.71	-54.61	3.68	-7.08	-30.3	0.0
	0.7	0.41	-50.6	-6.67	-43.93	-53.25	4.5	-57.75	2.23	-7.46	-58.48	34.97
Diff initial - final	0.12	-0.41	6.5	5.92	0.58	26.35	**23.21**	3.14	1.45	0.37	28.17	-34.97
5eja TD6 F 601	0.83	0.0	-45.3	7.83	-53.14	43.79	151.62	-107.84	34.58	155.0	233.37	0.0
	0.48	0.4	-32.47	19.36	-51.83	31.55	145.75	-114.2	34.02	6.6	72.18	56.45
Diff initial - final	0.35	-0.4	-12.83	-11.53	-1.3	12.23	5.87	6.36	0.56	**148.4**	161.19	-56.45
3kxh K66 A 1	0.82	0.0	-29.33	14.97	-44.3	52.28	71.19	-18.91	2.47	76.07	130.82	0.0
	0.41	0.43	-29.28	14.14	-43.41	32.55	61.16	-28.62	0.97	66.16	99.68	15.52
Diff initial - final	0.41	-0.43	-0.05	0.83	-0.88	19.73	**10.03**	9.71	1.5	9.91	31.14	-15.52
4ugy EXI A 904	0.87	0.0	-33.25	6.15	-39.4	52.14	82.94	-30.79	11.42	-4.04	59.52	0.0
	0.32	0.4	-30.73	7.98	-38.71	22.5	73.29	-50.79	10.3	-8.65	24.16	43.79
Diff initial - final	0.55	-0.4	-2.52	-1.83	-0.69	29.64	9.64	**20.0**	1.12	4.61	35.37	-43.79

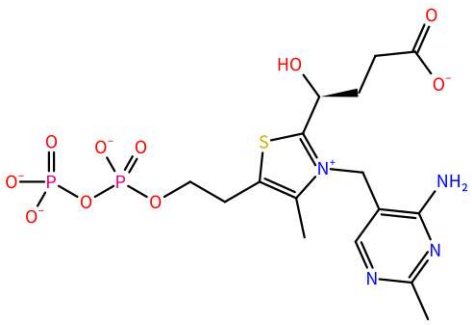
Table B.17: Ligand poses of $ProtFlex18_{train}$ with highly similar RMSD of 0.4 but diverging $EDIA_m$. All initial and final score terms after the optimization are given. The third row per pocket depicts the difference per score term. In bold is marked the score term with the largest improvement per pocket. Negative differences per partial score term should be avoided. In essence, three of four cases show a dropping HYDE score while the GeoHYDE score (GH_{ds}) improves. Visualization of the pockets are shown in Figure 5.9 and Figure B.18.



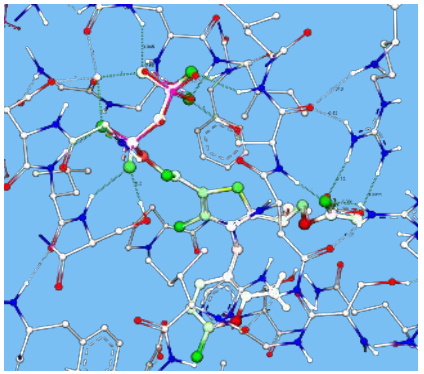
4l6z 1DC A 601



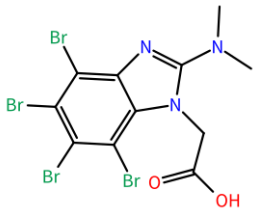
EDIA$_m$: 0.7, RMSD: 0.41



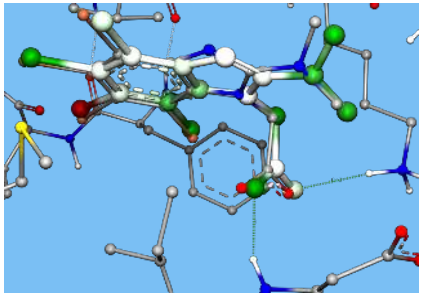
5eja TD6 F 601



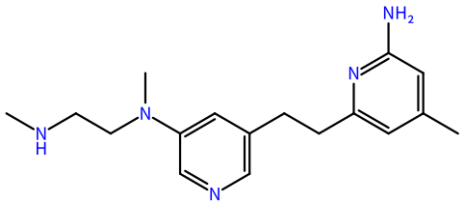
EDIA$_m$: 0.48, RMSD: 0.4



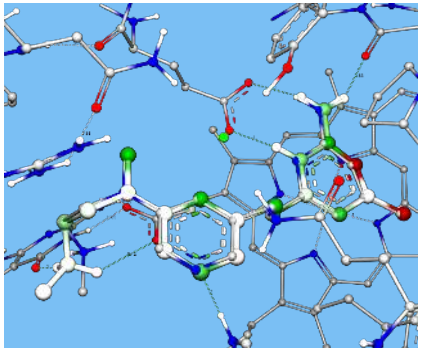
3kxh K66 A 1



EDIA$_m$: 0.41, RMSD: 0.43



4ugy EXI A 904



EDIA$_m$: 0.32, RMSD: 0.4

Table B.18: Ligand poses of ProtFlex18$_{train}$ with highly similar RMSD of 0.4 but diverging EDIA$_m$. Each ligand with EDIA coloring can be found in Figure 5.9 and score terms can be found in Table B.17.

Data set (size) Type	Outlier R upper	lower	Outlier P upper	lower
ProtFlex18$_{train}$ (546)				
RMSD	**0.73 (4)**	0.0 (0)	**12.11 (66)**	10.03 (59)
EDIA$_m$	0.0 (0)	**3.49 (19)**	4.95 (27)	**7.71 (42)**
HYDE	**0.0 (0)**	0.0 (0)	**4.4 (24)**	12.29 (67)
ProtFlex18$_{id}$ (62)				
RMSD	**1.61 (1)**	0.0 (0)	**12.9 (8)**	6.45 (4)
EDIA$_m$	0.0 (0)	**4.84 (3)**	1.61 (1)	**8.06 (5)**
HYDE	**0.0 (0)**	0.0 (0)	**3.23(2)**	8.06(5)
ProtFlex18$_{od}$ (23)				
RMSD	**0.0 (0)**	0.0 (0)	**13.04 (3)**	4.35 (1)
EDIA$_m$	0.0 (0)	**0.0 (0)**	8.7 (2)	**13.04 (3)**
HYDE	0.0 (0)	0.0 (0)	**8.7 (2)**	21.74 (5)

Table B.19: Pockets with a larger difference than their RMSE for three metrics when comparing optimization with and without partial side chain flexibility in the pocket.

Data set (size) Type	Outlier R upper	lower	Outlier F upper	lower
ProtFlex18$_{train}$ (546)				
RMSD	**2.01 (11)**	8.61 (47)	**11.36 (62)**	8.97 (49)
EDIA$_m$	0.0 (0)	**2.38 (13)**	3.66 (20)	**9.52 (52)**
HYDE	**0.73 (4)**	0.0 (0)	**5.86 (32)**	11.54 (63)
ProtFlex18$_{id}$ (62)				
RMSD	**3.23 (2)**	9.68 (6)	**14.52 (9)**	16.13 (10)
EDIA$_m$	0.0 (0)	**4.84 (3)**	4.84 (3)	**9.68 (6)**
HYDE	**3.23 (2)**	0.0 (0)	**9.68 (6)**	12.9 (8)
ProtFlex18$_{od}$ (23)				
RMSD	**4.35 (1)**	4.35 (1)	**4.35 (1)**	4.35 (1)
EDIA$_m$	0.0 (0)	**8.7 (2)**	4.35 (1)	**4.35 (1)**
HYDE	**8.7 (2)**	0,0.0 (0)	**8.7 (2)**	21.74 (5)

Table B.20: Pockets with a larger difference than their RMSE for three metrics when comparing optimization with and without full side chain flexibility in the pocket.

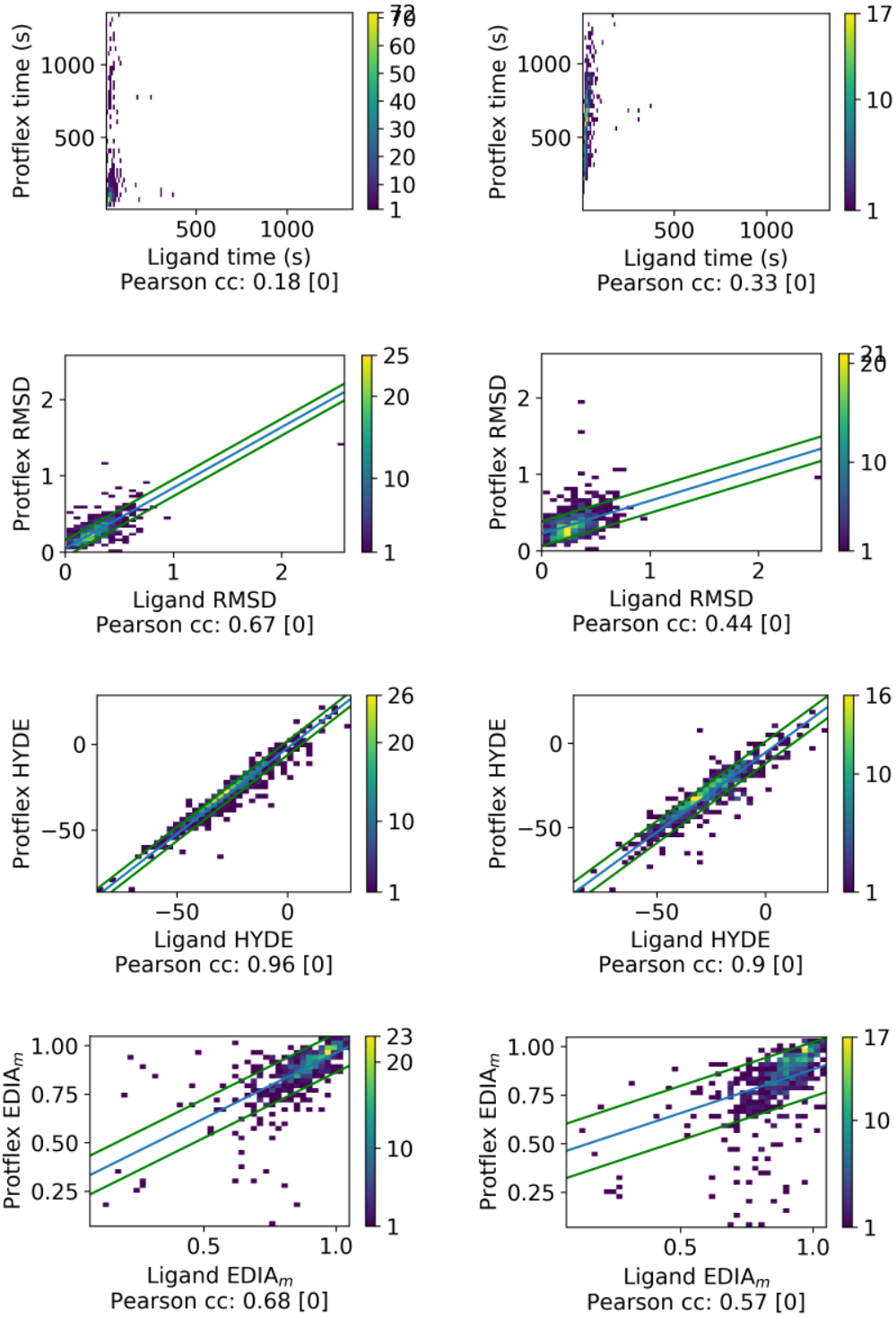


Figure B.29: Results of the optimization with GeoHYDE$_{final}$ on the 546 flexible pockets of ProtFlex18$_{train}$. X axis: optimization of ligands in the rigid pocket. Left row, y axis: optimization of ligands with partial side chain flexibility. Right row, y axis: optimization of ligands with full side chain flexibility. Blue: correlation line, green: line with one RSME distance to correlation line for outlier analysis. Pearson correlation coefficient and p value annotated in brackets.

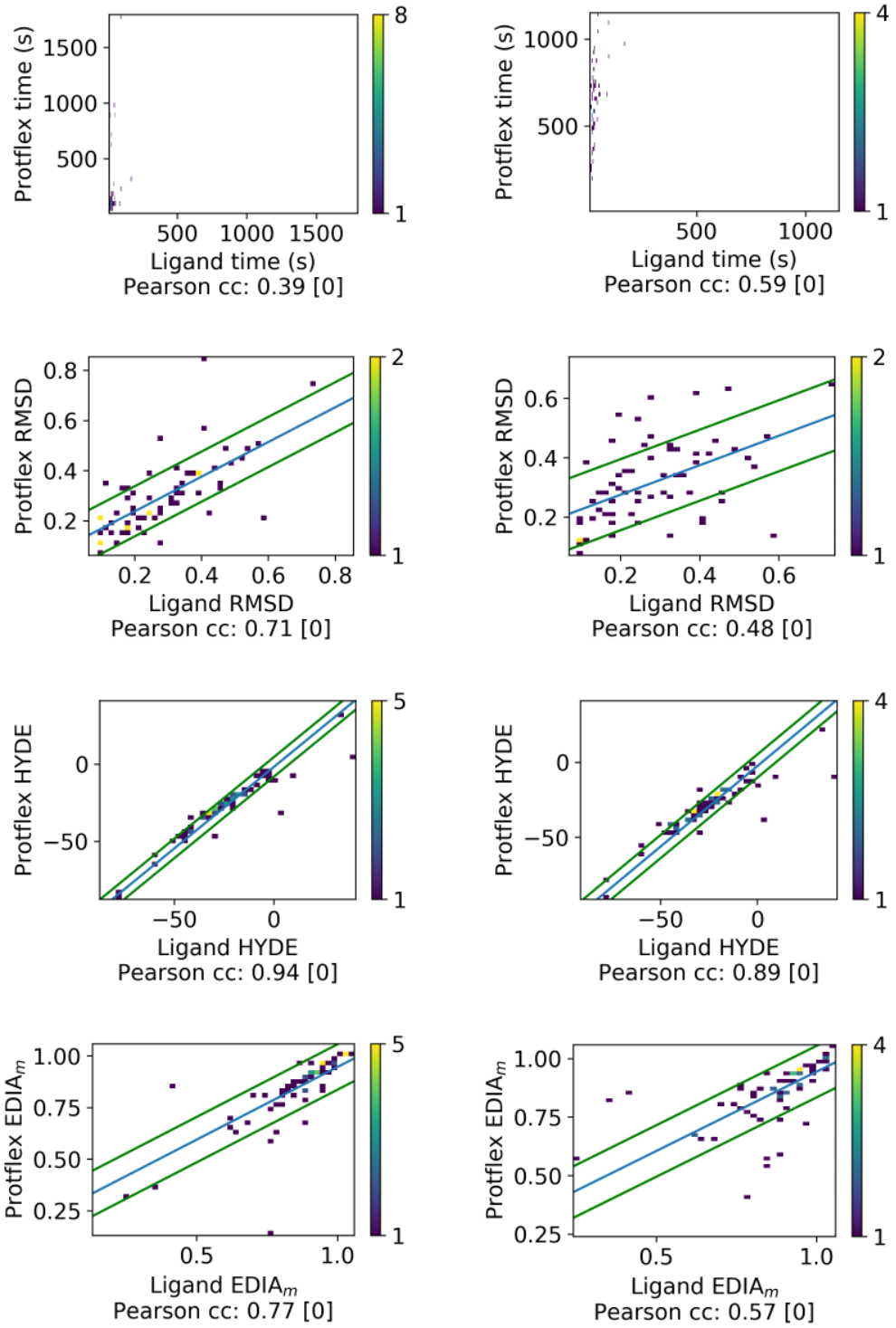


Figure B.30: Results of the optimization with GeoHYDE$_{id}$ on the 62 flexible pockets of ProtFlex18$_{id}$. X axis: optimization of ligands in the rigid pocket. Left row, y axis: optimization of ligands with partial side chain flexibility. Right row, y axis: optimization of ligands with full side chain flexibility. Blue: correlation line, green: line with one RSME distance to correlation line for outlier analysis. Pearson correlation coefficient and p value annotated in brackets.

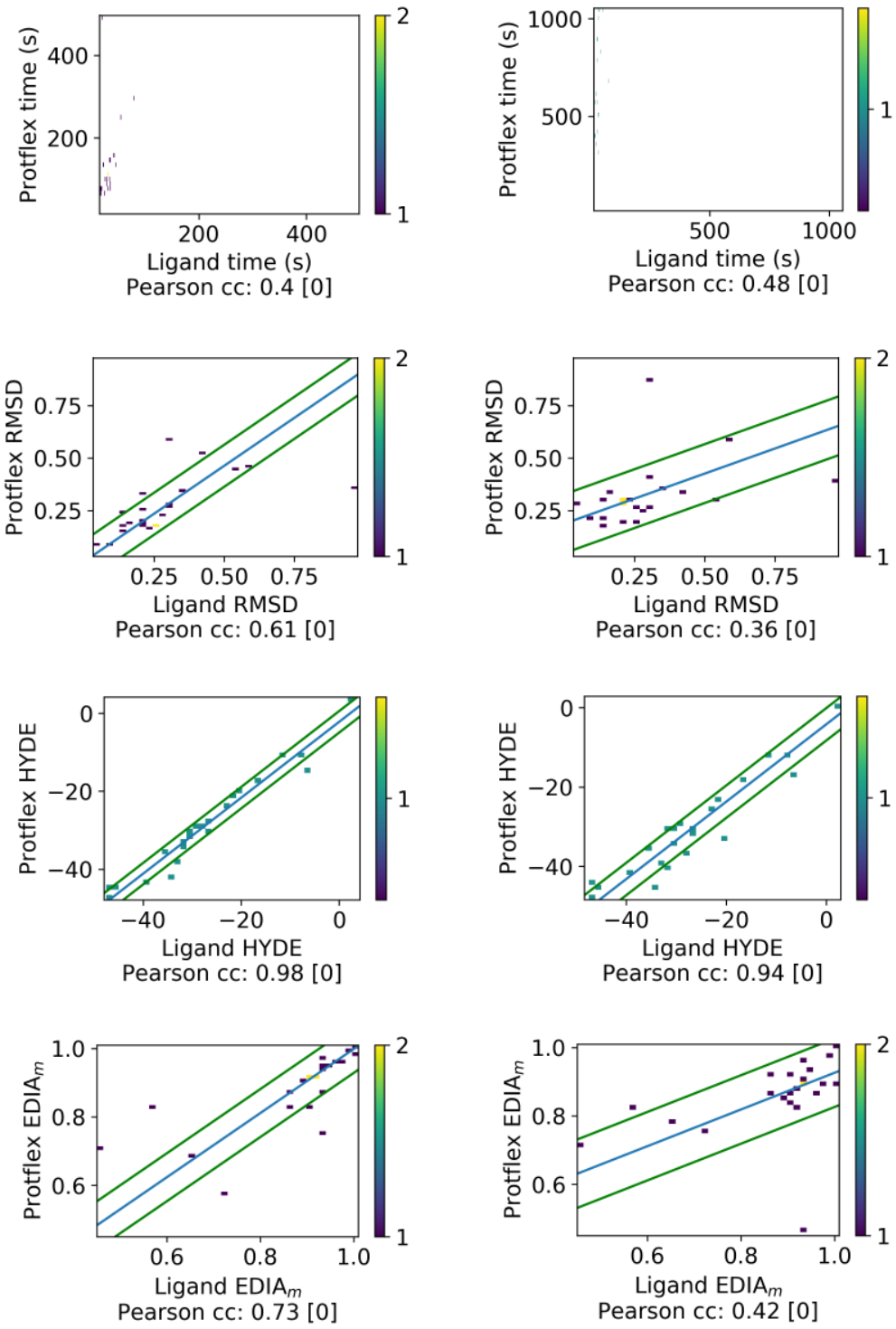


Figure B.31: Results of the optimization with GeoHYDE$_{final}$ on the 23 flexible pockets of ProtFlex18$_{od}$. X axis: optimization of ligands in the rigid pocket. Left row, y axis: optimization of ligands with partial side chain flexibility. Right row, y axis: optimization of ligands with full side chain flexibility. Blue: correlation line, green: line with one RSME distance to correlation line for outlier analysis. Pearson correlation coefficient and p value annotated in brackets.

PDB, Mol Id	Type	$EDIA_m$	RMSD	HYDE	$HYDE_s$	$HYDE_d$	GH_{ds}	GH_s	GH_d	CTS	LJ_{intra}	LJ_{intrap}	CTS_p
4B4V, L34 B 2001	R	1.01	0.0	-44.54	-7.99	-36.55	-9.86	55.19	-65.05	12.24	20.02	27816500.0	32.61
4B4V, L34 B 2001	R	0.94	0.2	-47.5	-8.9	-38.6	-36.05	47.27	-83.33	14.05	16.17	27816500.0	32.61
		0.07	-0.2	2.96	0.91	2.05	26.19	7.92	18.27	-1.81	3.86	0.0	0.0
4B4V, L34 B 2001	P	1.01	0.0	-44.54	-7.99	-36.55	-9.86	55.19	-65.05	12.24	20.02	27816500.0	32.61
4B4V, L34 B 2001	P	0.87	0.25	-44.11	-9.12	-34.99	-33.67	50.91	-84.58	14.31	15.84	27816500.0	32.55
		0.14	-0.25	-0.43	1.13	-1.56	23.82	4.28	19.53	-2.07	4.18	0.0	0.06
4B4V, L34 B 2001	F	1.01	0.0	-44.54	-7.99	-36.55	-9.86	55.19	-65.05	12.24	20.02	27816500.0	32.61
4B4V, L34 B 2001	F	0.91	0.29	-44.63	-10.79	-33.84	-40.39	50.47	-90.86	12.64	16.46	27816300.0	32.08
		0.1	-0.29	0.09	2.81	-2.71	30.53	4.72	25.81	-0.4	3.56	200.0	0.54
4QXC, OGA A 600	R	0.98	0.0	26.33	44.25	-17.92	195.21	207.68	-12.47	4.57	-4.24	37816800.0	38.43
4QXC, OGA A 600	R	0.78	0.48	41.1	58.84	-17.73	153.78	180.65	-26.88	4.85	-4.26	37816800.0	38.43
		0.2	-0.48	-14.78	-14.58	-0.19	41.44	27.03	14.41	-0.28	0.02	0.0	0.0
4QXC, OGA A 600	P	0.98	0.0	26.33	44.25	-17.92	195.21	207.68	-12.47	4.57	-4.24	37816800.0	38.43
4QXC, OGA A 600	P	0.77	0.49	3.65	22.88	-19.23	141.72	166.38	-24.66	4.37	-4.23	37816800.0	38.42
		0.21	-0.49	22.67	21.37	1.3	53.49	41.3	12.19	0.2	-0.01	0.0	0.02
4QXC, OGA A 600	F	0.98	0.0	26.33	44.25	-17.92	195.21	207.68	-12.47	4.57	-4.24	37816800.0	38.43
4QXC, OGA A 600	F	0.4	0.63	-9.21	11.13	-20.34	174.57	208.94	-34.37	4.5	-4.2	37816700.0	37.44
		0.58	-0.63	35.54	33.12	2.41	20.64	-1.26	21.9	0.06	-0.04	100.0	0.99

PDB, Mol Id	Type	$EDIA_m$	GH_{ds}	CTS	LJ_{intra}	LJ_{intrap}	CTS_p	time (s)	steps	Res0	Res1
4B4V, L34 B 2001	R	1.01	-9.86	12.24	20.02	27816500.0	32.61	0.0	0.0		
4B4V, L34 B 2001	R	0.94	-36.05	14.05	16.17	27816500.0	32.61	31.03	1244.0		
		0.07	26.19	-1.81	3.86	0.0	0.0	-31.03	-1244.0		
4B4V, L34 B 2001	P	1.01	-9.86	12.24	20.02	27816500.0	32.61	0.0	0.0	0.82	
4B4V, L34 B 2001	P	0.87	-33.67	14.31	15.84	27816500.0	32.55	75.16	1157.0	0.8	
		0.14	23.82	-2.07	4.18	0.0	0.06	-75.16	-1157.0	0.02	
4B4V, L34 B 2001	F	1.01	-9.86	12.24	20.02	27816500.0	32.61	0.0	0.0	0.82	
4B4V, L34 B 2001	F	0.91	-40.39	12.64	16.46	27816300.0	32.08	411.09	8235.0	0.51	
		0.1	30.53	-0.4	3.56	200.0	0.54	-411.09	-8235.0	0.31	
4QXC, OGA A 600	R	0.98	195.21	4.57	-4.24	37816800.0	38.43	0.0	0.0		
4QXC, OGA A 600	R	0.78	153.78	4.85	-4.26	37816800.0	38.43	19.83	666.0		
		0.2	41.44	-0.28	0.02	0.0	0.0	-19.83	-666.0		
4QXC, OGA A 600p	P	0.98	195.21	4.57	-4.24	37816800.0	38.43	0.0	0.0	0.8	1.0
4QXC, OGA A 600p	P	0.77	141.72	4.37	-4.23	37816800.0	38.42	78.92	896.0	0.8	1.0
		0.21	53.49	0.2	-0.01	0.0	0.02	-78.92	-896.0	0.0	0.0
4QXC, OGA A 600	F	0.98	195.21	4.57	-4.24	37816800.0	38.43	0.0	0.0	0.8	1.0
4QXC, OGA A 600	F	0.4	174.57	4.5	-4.2	37816700.0	37.44	660.27	10000.0	0.79	1.0
		0.58	20.64	0.06	-0.04	100.0	0.99	-660.27	-10000.0	0.01	0.0

Table B.21: Two examples that show comparatively the score differences over optimizing the pockets with the three types of flexibility R, P and F. The first entry always lists the score terms before optimization, the second column always those after the optimization and the third column depicts the difference between both lines. 4b4v L34 B 2001 (Res0: Arg B 8) presents the minimum HYDE score improvement for P and F in $ProtFlex18_{id}$ while 4qxc OGA A 600 (Res0: Met A 11, Res1: VAL A 286) shows the best HYDE score improvement in $ProtFlex18_{od}$. CTS denotes the Continuous Torsion Score and LJ_{intra} the intramolecular Lennard-Jones potential. If added with a p in a subscript, the potentials are evaluated on the protein side.

Component	Test Abbreviation	Test Description with Default Cutoffs	Configurations A	I	P	C	GH
Complex	Resolution	Resolution at most 2.5 Å		3.5			2.0
	DPI	Model diffraction precision index (Goto) at most 0.42					
	R factor	Model R Factor at most 0.4					
	R free factor	Model R_{free} Factor at most 0.45					
	Overfitting	Difference between R and R_{free} Factor at most 0.05					
	Model significance	Model is significant (R_{free} < 0.4 and Resolution < 3.5 Å)					
	Deposition date	PDB deposition date later than 11-AUG-00					
ActiveSite	RSCC	RSCC at least 0.7		(yes)			
	$EDIA_m$ per residue	Percentage of all residues with $EDIA_m$ below 0.7 (tolerated: up to 0%)	(no)	(no)			
	B factor ratio	Absolute active site to ligand B factor ratio at least 0.5 and at most 2					
	Occupancy	Percentage of atoms with occupancy of less than 1 (tolerated: up to 0%)					
	Intramolecular clash	Intramolecular clash for no heavy atom pair (sum vdw - 0.9 Å)					
		Intramolecular clash for no heavy atom pair (0.8·(sum vdw))					
	Intermolecular clash	Intermolecular clash for no heavy atom pair(sum vdw - 0.9 Å)					
		Intermolecular clash for no heavy atom pair (0.8·(sum vdw))					
	VSEPR bond angles	Percentage of bond angles differing by more than 16° (tolerated: up to 0%)					
	Unusual bond lengths	Percentage of bond lengths differing by more than 0.2 Å (tolerated: up to 0%)					

Table B.22: Available tests for the model, and the active site in the StructureProfiler annotated by the primary criteria catalogs and unusual cutoffs if necessary. Table adapted with permission from the original publication SI Table 1 [36]. Copyright 2018 Oxford University Press.

Component	Test Abbreviation	Test Description with Default Cutoffs	Configurations A	I	P	C	GH
Ligand	Maximum atomic B factor	Atomic B Factor is maximally 50 Å^2 for 100% of atoms					
	Occupancy	Percentage of atoms with occupancy of less than 1 (tolerated: up to 0%)					
	Crystal symmetry contacts	At most 0 crystal symmetry contacts closer than 6 Å to the ligand					
	OWAB	OWAB is at most 50 Å^2					
	RSCC	RSCC is at least 0.7		(yes)			
	EDIA$_m$	EDIA$_m$ score at least 0.8	(no)				
		EDIA$_i$ score at least 0.8					
	Intramolecular clash	Intramolecular clash for no heavy atom pair (sum vdw - 0.9 Å)					
		Intramolecular clash for no heavy atom pair (0.8·(sum vdw))					
	Unusual bond lengths	Percentage of bond lengths differing by more than 0.2 Å (tolerated: up to 0%)					
	VSEPR bond angles	Percentage of bond angles differing by more than 16° (tolerated: up to 0%)					
	Torsion angles	Number of torsion angles beyond the second TorLib tolerance interval					
	Aromatic ring planarity	Aromatic rings with the maximum size of 6 differing by more than 20° from planarity					
	Number of heavy atoms	At least 10 heavy atoms present					
	Molecular weight	Molecular weight is at least 100 and at most 600u			(130)		
	Lipinski acceptors	At most 10 Lipinski acceptors					
	Lipinski donors	At most 5 Lipinski donors					
	LogP	aLogP is at most 5					
	Number of peptide residues	At most 3 connected peptide residues are present					
	Number of rotatable bonds	Number of rotatable bonds are at most 16	(10)				
	Number of stereo centers	At most 5 stereo centers present					
	HET code	HET code does not match the exclusion list					
	SMARTS	Ligand matches none of the SMARTS in exclusion and if defined at least one in the inclusion SMARTS list					

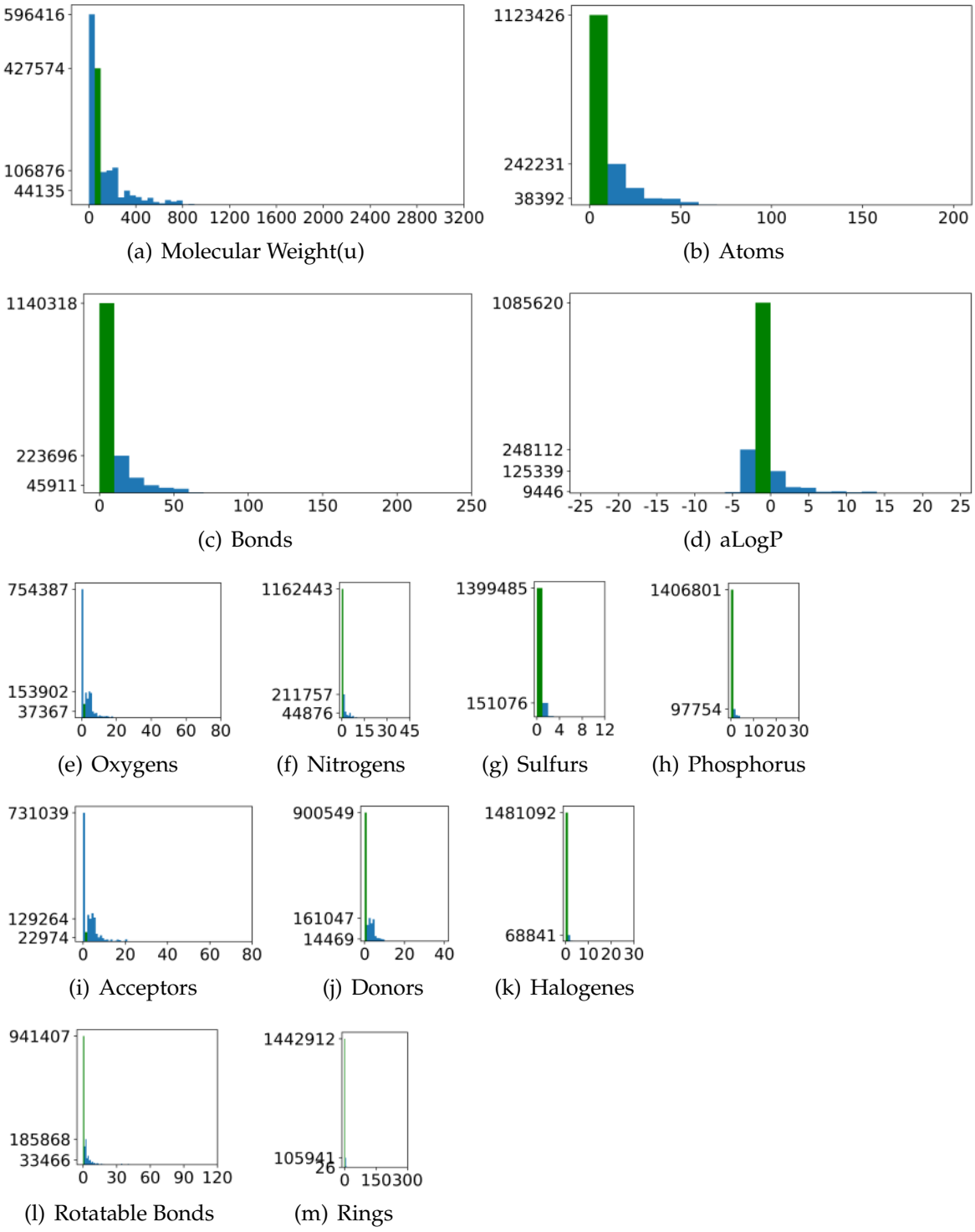
Table B.23: Available tests for the ligand in the StructureProfiler annotated by the primary, the combined, and GeoHYDE criteria catalogs and unusual cutoffs if necessary. Table adapted with permission from the original publication SI Table 1 [36]. Copyright 2018 Oxford University Press.

HET code	present in #PDB ids	found in total	PDB ids
BMQ	34	58	3lhu, 3lhv, 3lhw
TLA	31	64	1nxj, 1smo, 2b13
CAM	26	40	1dz4, 1dz6, 1dz8
ARG	23	33	1m15, 1om4, 2g6h
TRP	15	49	1c9s, 1gtf, 2aqj
OGA	13	16	2qrl, 3avs, 4bg1
S3P	13	13	1g6s, 1g6t, 1mi4
DGL	11	26	1zuw, 2gzm, 2jfy
INS	10	25	3ea2, 4i9t, 4miy
PHE	10	15	2ypo, 3ayj, 3kgf
NOJ	10	14	2jke, 2pwd, 3gbe
GPJ	10	15	1g6s, 1rf6, 2aay
G39	8	15	2ya8, 4k1i, 4k1k
3PG	8	12	2f90, 2h4x, 2vfg
MTA	8	11	1z5o, 2o06, 3fpf
SAL	7	12	2y7k, 3rem, 3twp
DOR	7	8	2e68, 2z25, 2z26
PC	7	7	2bib, 3uj9, 3ujc
TYD	7	10	1lvw, 3evo, 3oti
IPT	7	13	1jyx, 1px4, 2p9h
AZM	7	8	1jd0, 3hs4, 4g7a
2PG	7	7	1eqj, 1o98, 3ucc
TPP	6	13	2ozl, 2pgn, 2pgo
FUL	6	11	1ofz, 1rdj, 4gvx
UP6	6	9	1los, 3g1a, 3g24
MFU	6	9	1kww, 2boi, 2jdm
RIP	6	7	1drk, 2dri, 2gx6
EVF	5	5	5jdv, 5je7, 5jep
BCR	5	14	3wu2, 4ub6, 5b5e
RAM	5	7	2zux, 2zx2, 3w5n
U5P	5	6	1wlj, 2cze, 2v30
PAF	5	10	1n2j, 3guz, 3q12

Table B.24: Ligands present in at least five PDB ids in the ProtFlex18 data set identified by one of their HET codes are listed. There are 1116 unique ligands in total in terms of stereo isomer aware unique SMILES.

Enzyme cluster name	# PDB ids	Example structures
carbonic anhydrase 2	64	1oq5 3dcw 5sz4
nitric-oxide synthase	20	1d0c 4d1o 5agn
nicotinamide phosphoribosyltransferase	20	3dhf 4o28 5wi1
glycogen phosphorylase, muscle form	19	3bd7 5ox3 5ox1
orotidine 5′-phosphate decarboxylase	17	3g1a 3lhy 4nx5
alpha-mannosidase 2	7	3ejr 3ejq 3ddg
thrombin heavy chain	15	2zc9 5lpd 5jzy
tankyrase-2	14	3p0n 5nwc 4tjw
epsp synthase	9	1g6s 2qfu 2qft
transcription attenuation protein mtrb	11	1c9s 5ef1 5eez
endothiapepsin	10	2v00 4y5m 4y57
heat shock protein hsp 90-alpha	9	1yc4 4w7t 4fcp
7,8-dihydro-8-oxoguanine triphosphatase	8	4n1t 6f23 5nhy
transcriptional regulatory repressor protein (tetr-family)	8	3o8g 5ioy 5myn
4-hydroxy-3-methylbut-2-enyl diphosphate reductase	8	3ke8 4mv5 4mv0
pteridine reductase 1	8	3jqa 4cle 4cmk
trna (guanine-n(1)-)-methyltransferase	6	4yqj 4yq8 4ypz
cytochrome p450	7	4dnj 5u6u 5u6t
bromodomain-containing protein 4	5	3u5k 4a9e 6ckr
glutamate receptor 2	6	3rtf 4u1z 5jei
cgmp-dependent 3′,5′-cyclic phosphodiesterase	7	3itu 5u00 5tzz
camp-dependent protein kinase	7	3dne 4ujb 5vhb
heat shock protein hsp 90-alpha	7	2wi3 6eln 5xqd
orotidine-5′-phosphate decarboxylase	7	2qcg 3mi2 3l0n
ribosyldihydronicotinamide	7	1sg0 3nhw 5lbz
pantothenate synthetase	6	4fzj 4ddk 3iub
gamma-enolase	6	3ucc 4zcw 3ujs
dihydroorotase	6	2eg7 3mjm 2z28
methionine aminopeptidase	6	1xnz 4a6w 4a6v
beta-galactosidase	5	1jyx 3t0d 3muz
poly [adp-ribose] polymerase 3	5	4gv0 4l7o 4l70
serine/threonine-protein kinase pim-1	5	3r02 5n4v 5kgg
anthranilate phosphoribosyltransferase	5	3qs8 3uu1 4owo
xanthine dehydrogenase/oxidase	5	3bdj 3unc 3una
carbonic anhydrase 12	5	1jd0 5ll9 4ww8
thermolysin	5	1hyt 3fgd 3fcq
neuraminidase	5	1f8c 4mwq 1l7f

Table B.25: Clusters with flexible side chains identified by SIENA are given. The list is limited to clusters with at least five unique PDB ids. In total, 80 clusters reported flexible side chains.



(a) Molecular Weight(u) (b) Atoms

(c) Bonds (d) aLogP

(e) Oxygens (f) Nitrogens (g) Sulfurs (h) Phosphorus

(i) Acceptors (j) Donors (k) Halogenes

(l) Rotatable Bonds (m) Rings

Figure B.32: Ligand properties of the LigandExpo (Feb. 2020). In all plots, the number of e.g. oxygens per ligand is given on the y-axis. The bin including the median value is colored in green.

B.0.1 ProtFlex18 Data Sets

B.0.1.1 ProtFlex18$_{train}$ Data Set

5A1L S2I B 2267, 3A22 ARA B 751, 3A22 ARA B 761, 3A22 ARA B 781, 2A3B CFF A 1433, 5A4U I3A F 1213, 4A6V IKY B 1264, 1A95 GUN B 304, 1A96 XAN C 303, 4A9E 3PF A 1000, 2AA9 SKM A 501, 5AF9 SJR H 1250, 5AGM WT2 B 800, 5AGN 4JK B 800, 4AJJ 88S A 1332, 4AJL 88W C 1336, 4AJO 88N A 1332, 4ALH A9P A 1185, 5AL3 TGW B 3166, 6ALR ARG A 402, 4AM8 PAO F 402, 5AM9 GLU C 911, 5AMB ILE Q 41, 5ANW 9CQ A 1157, 4AO4 PLK A 1446, 6APS SV2 B 301, 2ATJ BHO B 353, 2AW1 COX A 264, 3AYI HCI B 1907, 4AYU N8P D 499, 6AY3 C3J A 1201, 4AZJ SEP B 500, 6B04 C6J C 401, 6B07 C6M B 401, 3B1D PLS B 501, 3B1E 0JO D 401, 3B28 B2X A 237, 5B2E MQG C 302, 3B3M JI1 B 800, 3B3N JI2 B 800, 6B5E TYD C 303, 3B6H MXD B 551, 4B7R G39 D 801, 4BAM MM9 B 1287, 4BC5 5FX C 1532, 3BDJ 141 A 5101, 3BEX PAU D 248, 3BF3 PAZ B 248, 4BHG C2T A 401, 1BK0 ACV A 351, 3BL0 BL0 A 300, 4BQG 50Q A 1225, 3BTO SSB A 378, 3BWL I3A B 601, 3BXM ACE I 1, 5BX3 NOJ A 901, 5BX4 GIM A 901, 4BZN UGX A 1306, 4C5W OGA A 400, 4C6Z TLE B 1421, 4C73 TLH A 1427, 6C9X VOG B 701, 2CBU CTS A 1447, 5CBS E42 B 301, 2CHN NGT B 1717, 3CHC ZRG B 440, 4CHS GS8 B 1219, 5CI5 T6T B 501, 2CJF RP4 H 2551, 6CJA F0G D 400, 6CKR F5V B 201, 4CLD JUO A 1270, 4CLE JR2 B 1270, 4CLR FDB D 1270, 5CLE ADK B 101, 5CLU S8A A 302, 4CM4 4NR B 1270, 4CM6 AOB D 1270, 1CRU PQQ A 504, 3CTP XLF B 401, 4CTM MIF A 998, 4CTW S71 B 1721, 4CWD 449 A 1385, 4CXR 2BG A 502, 2CYB TYR B 501, 1D0O INE B 761, 4D08 Q2T A 1918, 5D04 PHE D 407, 4D1J DGJ H 600, 6D28 NEC A 401, 5D3U TRP B 502, 3D4L 605 A 1521, 3D4Z GIM A 1048, 2D5Z L35 A 1201, 3D51 GOX A 1048, 6D6P FY1 B 300, 4D7O 0GD B 800, 3D9Z D9Z A 263, 3DA9 44U B 1, 5DB1 58O A 601, 5DB3 58Q A 610, 3DD0 EZL A 301, 3DDG GB7 A 5001, 3DDW CFF B 903, 3DDW NBG A 901, 4DDK 0HN B 401, 4DE0 0JB B 301, 4DE1 0J6 A 301, 4DE3 DN8 B 301, 5DEU OGA A 2001, 4DF1 BMP B 301, 4DGN LU2 A 401, 3DHF NMN B 503, 3DJE FSA B 501, 5DKV T6T D 401, 3DNE LL1 A 351, 4DO4 DJN B 510, 4DO5 DGJ B 509, 4DTS DCP A 1001, 3DUR KDO D 303, 4DUB LDP B 501, 3DX0 MSN A 1049, 3DX3 YTB A 1050, 3DX4 GOO A 1049, 3DYO IPT C 2001, 5E0I 5J6 D 500, 5E2K BX4 A 302, 4E3D GTQ A 303, 2E40 LGC B 2001, 3E5X 3C4 B 504, 2E68 DOR B 2353, 3E7M AT2 B 1906, 4E70 N7I B 1402, 3EA2 INS B 802, 4EAR IM5 C 301, 3EBO 57D A 940, 3EBP CPB A 940, 1ED5 NRG B 2705, 4EGN TWO C 506, 3EHW DUP Y 777, 3EJQ HN3 A 1049, 3EJR HN4 A 1049, 5EJ9 TPP H 602, 5EJA TD6 F 601, 3EKR PY9 B 901, 4EKQ NPO A 202, 5ELO KRS D 602, 1ENU APZ A 400, 4EO6 0S2 B 600,

2EPN NGT B 2650, 3EPW JMQ B 1003, 4EQL SAL B 602, 3EWZ CNU B 484, 3EX2 6CN B 481, 5EXK 5AD K 403, 5EZH 841 A 302, 5F1J 5TO A 301, 6F23 C8Z A 201, 1F3E DPZ A 400, 5F5N 5VD B 302, 2F7R SK3 A 5009, 5F76 MTA C 301, 1F8C 4AM A 4, 1F8D 9AM A 0, 1F8E 49A A 0, 6F8V D0B A 605, 3FAT AMQ A 427, 4FB4 DHC A 401, 1FCY 564 A 450, 1FCZ 156 A 450, 3FCQ M3S A 600, 4FCK GPA B 401, 4FCP 42C B 301, 5FDC 5WN A 302, 3FFP LC1 X 300, 3FGD BYA A 322, 3FH5 24P A 611, 3FH8 27P A 611, 5FIU TLA A 1300, 5FIU Y3J C 1299, 3FJ7 PEQ B 301, 4FJL DGT A 1001, 5FKY 2J4 B 1716, 4FL7 BHO A 304, 4FLI Y16 A 503, 5FLR XCZ A 1266, 5FNM 5O6 A 1262, 6FNQ AVJ B 401, 6FNS DY8 B 401, 2FOQ B15 A 301, 2FPZ 270 C 1002, 1FSG 9DG C 304, 5FSO S76 A 1158, 3FT5 MO8 A 237, 3FTW 11X A 710, 3FUX MTA C 272, 4FUB 4UP A 301, 2FVC 888 B 902, 3FVG MS8 B 902, 3FVK 8DX B 2, 4FVY 3KJ B 804, 6FWH 5LD I 202, 2FYP RDE A 1001, 4FZJ 0W1 B 401, 2G1A 5HG A 700, 3G15 HC6 A 603, 3G1V 5FU B 502, 3G2I RUG A 998, 3G2K SKY A 998, 5G2T UAP D 510, 6G2N O84 B 302, 3G35 F13 B 2, 4G3J VNT A 502, 6G36 EKH A 401, 6G38 TBN A 800, 3G4K ROL A 901, 4G4P GLN A 302, 5G4J EXT A 1441, 1G6C IFP A 2001, 2G6N ARG A 770, 5G66 M5K A 1366, 4G88 API G 401, 1G9V RQ3 A 801, 6G92 ERZ A 404, 2GC0 PAN A 901, 4GC4 BMP B 301, 3GDN MXN B 534, 2GGD GPJ A 601, 2GGD S3P A 501, 4GHD DHY C 403, 5GI7 54W A 302, 5GI8 7DP A 302, 5GI9 7AN A 302, 5GIG 7DP A 302, 5GIH 54W A 302, 5GII 54W A 302, 6GI6 EZB A 501, 4GLW 0XT A 402, 4GLX 0XS A 603, 5GLP ARA B 403, 5GMZ 6XU F 202, 6GO2 LU0 A 407, 4GQN INI B 302, 3GUZ PAF B 177, 4GUI QIC A 301, 4GV0 8ME A 601, 4GV4 MEJ A 601, 4GVX FUL D 303, 3GWC UFP H 260, 3GY4 PBZ A 1, 5H0B OOQ A 601, 5H19 LQF A 501, 1H46 RNP X 1433, 2H4X 3PG B 2408, 4H4E 10G A 402, 5H41 IFM B 1203, 3H5S H5S A 571, 4H5G ARG B 305, 3HAC 361 A 767, 1HB1 OCV A 1332, 2HDU F12 B 1001, 2HF9 GSP B 300, 3HHK 77Z B 564, 5HHY PLR B 401, 2HKJ RDC A 501, 3HKU TOR A 300, 2HNC 1SA A 265, 2HOX P1T B 6002, 1HPU A12 C 1604, 5HQE 64B A 401, 3HS4 AZM A 701, 3HSN HAR B 1770, 3HT3 DCP A 201, 1HWW SWA A 1103, 3HWT D3T A 576, 1HYT BZS A 807, 4HYI 1AO A 304, 2HZY DHJ B 1101, 5I07 NE8 A 405, 1I13 7HP B 810, 1I14 7HP A 800, 5I2C ARG B 401, 4I3B BLR D 201, 2I5N UQ1 L 502, 5I5X 68C A 402, 5I6D AU6 D 402, 4I7N 1DJ B 203, 5I7I 3HB B 401, 4I9T INS A 502, 5IAI RB0 A 501, 5IBQ XXM A 401, 2IEJ FII B 944, 5IED CTS A 1025, 5IEE NOJ A 1023, 4IG3 J94 A 609, 2II6 C5P A 1427, 3IIT D14 A 700, 4IIC IFM B 950, 4IIL RBF A 401, 5III DTP A 601, 5IIJ DCT A 601, 3IJL PRO A 384, 4IJI 99T B 501, 4IKU SHX A 401, 4ILX 1EZ A 303, 3IMC BZ3 B 701, 4IM7 CS2 A 501, 3INL BXB H 1001, 3IOB A4D B 302, 5IOY 6C5 A 304, 3IP8 B85 A 249, 5IP6 6C9 A 301, 3IT3 3AM B 343, 3ITL LRH C 603, 3ITU IBM D 999, 4ITJ 1HX B 301,

5ITP 6DB B 301, 3IU7 FCD A 288, 3IUB FG2 A 302, 4IUO QIC B 301, 2IVI ACW B 1332, 2IVJ BCV A 1332, 3IVD URI B 603, 4IV9 TSR B 602, 5IV3 LRI A 506, 5IVE 6E8 A 601, 5IVV 6EN A 601, 3IX8 TX3 D 174, 4IXE IXE D 301, 1IY7 CXA A 500, 5IZZ 3HP A 401, 5J1U P93 A 301, 2J4D MHF A 1499, 5J42 6FV A 401, 2J78 GOX B 1446, 2J79 GTL A 1446, 2J7B NTZ A 1446, 2J7H AZF A 1446, 4J7H TLO A 501, 5J9O 6H8 A 404, 1JD0 AZM B 2401, 5JDY TOF B 303, 5JE0 AZ8 B 302, 2JFZ 003 B 1256, 5JF6 BB4 A 301, 5JGS EVF B 301, 2JKE NOJ B 1727, 5JMX DZ5 A 304, 5JNA TOR D 302, 3JQA DX4 B 270, 5JTT 6MY A 902, 5JTU 6NE A 902, 4JZB P2H B 402, 5JZY 6OV H 308, 5K0C 6OZ A 304, 5K0F 6P1 A 302, 4K1I G39 A 507, 4K1W CS2 C 502, 4K3N 1OT L 1004, 5K32 6Q2 B 1003, 3K5E K5E B 369, 3K5X P8D A 401, 4K60 1P8 A 703, 3K7S R52 A 160, 4K8K 1PJ B 404, 4K9Y K9Y A 701, 3KCZ 3AB B 1, 5KDY ANN A 502, 5KDZ ANN A 501, 3KE8 EIP B 998, 3KFL ME8 A 801, 3KFX MCY B 502, 4KFN 1QR B 601, 3KGC ZK1 B 263, 5KGG 6SO A 423, 1KHB GCP A 704, 5KIT 6TA B 501, 3KJD 78P A 1, 5KMA 777 A 201, 4KP5 E1F A 302, 5KR1 017 B 101, 3KS9 Z99 B 1, 4KTF 1TM A 406, 5KTO NTM A 402, 3KVL DOR A 399, 1KW6 BPY B 401, 4KWD JF2 B 705, 3KXH K66 A 1, 3KZZ OBG A 181, 4KZB NZ2 B 401, 3L0N S5P B 257, 3L0V 724 B 485, 5L09 482 A 201, 5L4S 6KX A 401, 4L51 HSX B 401, 4L6D VNL G 402, 4L6G CNL B 502, 4L6Z 1DC A 601, 1L7F BCZ A 801, 1L7G BCZ A 801, 1L7H BCZ A 801, 3L79 DKX A 843, 4L70 1V9 A 601, 4L7O 1VD A 601, 5L8A 6RB A 101, 4L91 X29 A 301, 5L9V OGA B 502, 5LBZ 6T3 B 302, 5LCF 6TJ A 504, 1LD7 U66 B 1003, 1LD8 U49 B 1003, 5LE1 6UW A 503, 3LHV BMQ D 229, 3LHW BMQ B 229, 3LHZ BMQ B 229, 3LI1 BMQ B 229, 5LJQ ANV A 302, 1LKD BP6 A 300, 3LLD UP6 B 229, 4LLS IPE A 301, 5LL4 6YH B 305, 5LL9 6YQ D 302, 5LLC V26 A 308, 5LLG VD9 A 302, 1LOS UP6 D 5004, 4LP0 1YM A 301, 4LPB 1YP A 301, 5LR1 72Y A 4000, 5LRC 73E A 902, 5LRD KS2 A 901, 4LS3 HIS B 601, 4LTS LTS A 601, 4LUK PA5 A 202, 4LUU BTM A 704, 4LVB 20N B 601, 4LVD 1EB B 603, 4LVF 20P B 601, 4LWF FJ3 A 301, 4LWI FJ6 A 301, 4LWW LWW B 601, 5LWN PHU A 1201, 4LXQ TYD B 302, 1LZX HAR B 1770, 3M0L PSJ C 603, 3M0X PSJ B 602, 1M15 ARG A 403, 3M14 BEV A 505, 1M5E AM1 C 1702, 1M5F AM1 C 1202, 4M5M DX4 A 401, 4M6P 20R B 601, 5M67 3D1 D 503, 4M7T 25W A 504, 4MCC 21X A 301, 4MCD 22L A 301, 3ME3 3SZ A 540, 4MES 26G A 203, 5MFQ 2J9 A 902, 3MI2 PFU B 1, 4MIY INS D 402, 4MJL CBU D 402, 5MLJ 9ST A 902, 5MLS 22U H 301, 1MMW VIO B 2780, 3MMS Q88 A 231, 5MM6 32U H 311, 5MMN O54 A 301, 4MNC 173 A 401, 4MO8 2VQ A 302, 3MS5 REE A 391, 4MSS 2CZ B 401, 5MT9 SRO W 101, 5MTP 53K D 302, 4MUY 2E5 B 402, 3MVX BHZ B 504, 4MV0 2E6 B 402, 4MV5 2E7 A 402, 3MYZ TFX A 101, 4MYS 164 B 301, 5MYN ZUF A 301, 3MZC S6I A 263, 5MZY 8EZ B 301, 4N1S WM4

A 201, 4N1T 2GD A 201, 4N1U 2GE B 201, 1N2J PAF A 1001, 3N3M NUP B 2001, 3N3X GUN A 247, 3N4B WWZ A 263, 5N4V 8MW A 401, 4N5V FA0 A 404, 4N5V FA0 B 404, 3N62 XFJ B 800, 3N86 RJP R 147, 4N8D 2KS A 815, 4N8G DAL D 402, 4NAE 1GP B 301, 5NAB 8RK A 503, 5NAG 8R5 A 502, 4NBD 9CA C 503, 4NBN 2J7 B 401, 5NEA 8V8 A 302, 1NF0 13P B 1150, 1NF8 ISC A 220, 4NF4 2JK A 301, 5NGT 8WZ A 201, 3NHU M42 B 233, 4NHK PD2 A 702, 5NHY 8XT A 201, 3NI2 AYL A 537, 4NJK 2KA A 303, 4NJM 3PG B 401, 4NJS G08 D 500, 4NJT 017 D 101, 1NNK CE2 A 454, 2NND PRZ A 300, 2NNS M25 A 301, 4NN3 ORO A 403, 5NN4 SC2 A 1016, 5NN5 NOJ A 1016, 5NN6 MIG A 1013, 1NQX RLP D 4201, 3NQ6 UP6 I 2, 3NQC BMP I 2, 3NQE BMP B 229, 3NQF BMP B 229, 4NQ8 PAF B 401, 4NQG CZH A 201, 5NS8 NOJ C 506, 3NT1 NPS A 5, 4NUW U5P A 301, 3NVW GUN L 503, 5NWC 9CE B 1203, 5NWE G39 C 503, 3NXR D2D A 192, 3NXV D2F A 187, 4NX5 UP6 B 301, 5NXO 9HK A 302, 5NZ4 G39 B 503, 5NZE G39 B 503, 5NZF G39 D 503, 4O08 PO6 B 302, 4O10 2QF B 601, 4O15 2P1 A 601, 5O1E 9GT B 402, 4O28 1QS A 601, 3O31 3O3 A 1, 5O38 9JB A 501, 5O4J 9KH C 302, 5O4J PJL A 304, 5O4V 9K2 C 502, 5O50 9L2 A 902, 5O52 9LE A 902, 2O73 2AL F 3001, 2O78 TCA F 701, 2O7D DHC G 701, 2O7S DHK A 4733, 4O7E 2RN B 202, 3O8G O8G A 217, 5OB3 1TU A 101, 3OCC DIH F 500, 3OCZ SRA A 264, 4OCP GN1 A 401, 1ODM ASV A 1332, 3ODG XAN A 288, 4OES EDT A 601, 1OFZ FUL B 1313, 3OGS IPT A 1024, 3OGV PTQ A 1024, 4OGI R78 B 202, 5OGO WWO A 302, 5OHY 9VH D 705, 5OIC 9VQ A 302, 5OLV 9Y2 A 1201, 1OM4 ARG B 771, 5OO4 URI A 201, 5OO5 UUA A 201, 1OPK P16 A 2, 1OQ5 CEL A 701, 3OTI TYD B 377, 2OU3 I3A B 165, 4OVT LFC B 402, 2OW6 NK1 A 4001, 2OW7 NK2 A 6001, 3OWP 2SB A 2001, 4OWO 6F0 B 404, 4OWO 6F0 B 405, 5OWY B0W A 902, 5OWZ B0Z A 903, 4OX2 SPV B 704, 5OX0 B1H A 901, 5OX1 B1K A 901, 5OX3 B1N A 901, 3OYS OYS A 263, 2OZL TPP C 1330, 3P0N BPU C 1163, 1P1W AMQ B 427, 3P1F 3PF B 1198, 1P5Z AR3 B 304, 4P56 RMN A 401, 4P5A 5BU A 302, 3P7Y P7Y A 402, 4P7X CXS A 308, 2P9H IPT B 999, 3P93 CS2 F 407, 5P9R 7JJ B 303, 3PC3 P1T A 702, 1PG4 PRX B 998, 2PGO TPP B 615, 4PGN 3IO D 401, 3PHC IM5 E 501, 4PIO AVI B 402, 3PKE Y10 A 286, 4PML 3AB C 1201, 4PNN JPZ A 1202, 4PNR G18 C 1201, 2POU I7A A 1000, 2POV I7B A 1000, 4POW OP1 B 301, 4PPS ESE B 601, 2PQ9 GG9 A 501, 1PX0 RPN A 1001, 1PX4 IPT B 2001, 2PYW SR1 B 998, 3PYY 3YY B 532, 1Q0N PH2 A 181, 4Q0N 2XD B 801, 1Q11 TYE A 401, 3Q12 PAF B 501, 2Q2A ARG C 902, 3Q23 G2P B 1109, 1Q36 SKP A 600, 1Q6Q LXP A 7301, 1Q6R LX1 B 9301, 4Q6D WW3 A 304, 2Q88 4CS A 501, 4Q83 3FH A 302, 4Q87 4FH A 303, 4Q8Y HQT A 303, 2Q94 A04 A 400, 2Q96 A18 A 400, 3QAX ARG B 600, 2QCG 5BU A 1, 3QEX DGT A 904, 2QFU GPJ A 801, 2QFU

S3P A 701, 4QFP VAL B 601, 3QHD CTN B 165, 3QHD MSR A 166, 2QIS RIS A 901, 1QK5 XMP A 300, 3QMR BMP B 229, 3QMT BMP B 229, 2QOA MAJ A 800, 3QRY DMJ B 427, 3QTO 10P H 1001, 1QV6 24B B 378, 1QXW M1C A 3001, 2QX0 PH2 B 182, 3QX5 02P H 5, 1QY2 IPZ A 300, 4QYG 3DW B 301, 3R02 UNM A 555, 4R07 URI D 901, 3R16 5UN A 1, 3R17 5UM B 1, 4R34 TRP B 505, 1R5L VIV A 301, 2R5E QLP B 430, 3R77 QLI B 500, 1R8Q AFB B 503, 3R8G IZP A 409, 1RF6 GPJ C 1628, 2RFQ 1PS D 392, 3RG9 WRA B 702, 3RHK M97 A 1, 3RIE JFQ C 501, 3RIE MTA C 401, 4RJE FNR C 401, 3RLP 3RP A 901, 3RLR 3RR B 901, 3RLU BMP B 229, 3RLW S28 H 1, 3RLY S29 H 1, 4RLF 4MA B 1001, 3RM0 S54 H 2, 3RM8 RM8 B 417, 3RME RME A 418, 3RMN M41 H 1, 4RN0 L6G A 501, 1RPJ ALL A 291, 4RP9 ASC A 501, 3RQL X2D A 800, 4RRO 3UX A 502, 1RS7 D7P B 798, 3RTF CWD D 800, 4RU1 INS L 401, 4RY8 SR1 D 401, 4RYA MTL A 501, 3S1G ITE A 501, 3S2N P4D A 401, 3S2Z DHC A 259, 4S26 IRN B 703, 4S28 AIR A 702, 3S44 FN5 A 1, 1S63 778 B 3012, 3SBI E90 A 266, 3SCS GLF B 1002, 1SD3 SYM B 999, 1SG0 STL B 502, 3SHA P97 H 1, 3SIZ BMP B 229, 3SJ3 BMP A 229, 3SLH GPJ D 441, 3SV2 P05 H 1, 3SVH KRG A 1, 1SW1 PBE B 302, 5SW3 46L A 405, 5SXT NIZ B 808, 5SYI NIZ B 806, 1SZO CAX J 5010, 3SZU H6P B 998, 5SZ4 72D A 304, 3T0D 149 A 2001, 1T2B CNL A 500, 3T2S AGS B 301, 3T7V MD0 A 993, 1T93 FLV A 431, 3T95 PAV A 400, 3T9V CNI B 400, 1TA8 NMN A 401, 1TC1 FMB A 900, 1TC2 7HP B 810, 3TCF UNK O 1, 3TCY PHE A 302, 5TE2 7B9 B 501, 3THQ NUP A 1000, 5THH TYR A 401, 3TIA LNV A 801, 3TIC ZMR C 1002, 4TJU CNQ C 1202, 4TJW P34 B 1202, 5TJX GBT A 701, 5TKD 7GL A 901, 3TL1 JRO B 160, 3TQ8 TOP A 2001, 3TR9 PT1 C 1001, 5TSQ BDR A 402, 3TWP SAL C 404, 5TWM 7NG A 601, 4TXJ THM D 302, 3TY3 GGG A 363, 5TYA 7QS A 302, 1TZC PA5 B 601, 5TZA 7OG D 1001, 5TZH 7OP D 1001, 5TZW 7P4 D 1001, 5TZZ 7OJ D 1001, 5U00 7OV D 1001, 5U0E 7R4 A 302, 5U0V 7VJ B 302, 3U15 03M D 1, 4U1Z KAI A 301, 4U23 FWD A 401, 5U2M 7T7 B 501, 1U3U BNF B 2378, 5U3B 7TD B 502, 5U3F 7TS B 400, 4U4X 3C2 B 801, 5U5H 7VV A 501, 5U62 7WD B 501, 5U6T 81J A 502, 5U6U 81M A 502, 5U8A 82D A 501, 5U8F 82G A 501, 5U8Z 83D D 602, 5U98 1KX D 301, 3UCD 2PG A 601, 4UCN JRB A 1422, 3UES DFU B 501, 5UER 87P A 501, 1UF5 CDT B 999, 1UF7 CDV B 998, 3UFY NPX A 701, 5UF0 89J A 501, 5UFM AZ8 A 302, 4UGI SKO A 904, 4UGY EXI A 904, 1UHH CZP B 2001, 4UIX TVU B 1170, 5UII BFR A 204, 5UIT 8CD B 501, 3UJC PC A 301, 3UJS 0V5 B 602, 4UJ9 S3N A 1351, 4UJB 8BQ A 1351, 1UMD TDP C 2402, 4UMA GZ3 B 1351, 3UNC SAL B 1338, 5UPF 8HV A 901, 3UR4 0BW A 1000, 5UT3 IK1 A 901, 3UUD EST A 600, 2UVZ GVJ A 1351, 4UVL 32X A 2165, 4UVT G1O A 2167, 4UVZ 5NN C 2165, 1UWT GTL B 1491, 1UWU GOX B 1490, 3UWE VJJ A 701, 3UWO 0DJ

A 800, 2UY5 H35 A 1313, 2UYT LRH A 1481, 1UZ1 IFL A 1446, 1UZ4 IFL A 1432, 2UZ1 TPP D 1556, 3UZ5 0CU A 291, 5UZ0 AMZ D 603, 1V0H SHA X 253, 2V00 V15 A 1336, 3V1P BMP B 229, 4V24 GYR B 1450, 4V2V OGA B 1355, 5V2Z OOG A 402, 1V3E ZMR B 2200, 2V3D NBV B 1504, 2V5Z SAG B 1498, 2V7V 5FD C 1299, 3V7Z GEM A 405, 2VBD V10 A 1333, 2VBF TPP B 1551, 2VBP VB1 A 1333, 2VCZ VC3 B 1200, 3VC1 GST I 303, 3VC1 GST L 303, 3VC3 C6P F 501, 3VD4 IPT C 2001, 3VE7 BMP A 301, 2VFG 3PG D 1249, 5VGY 9AA A 304, 3VHD VHE B 1, 5VHB 9CY A 401, 2VIO L1O A 1246, 3VIG NOJ A 507, 2VJX IFL B 1867, 2VL4 MNM B 1868, 5VLC HSO A 501, 2VNY M02 A 1351, 2VPO 6CS A 1312, 3VRI 1KX A 301, 1VSO AT1 A 258, 3VV5 SLZ B 301, 3VVF ARG B 301, 3VXJ 3DM A 503, 1W1T CHQ B 1512, 1W1Y TYP B 1508, 2W5T GP9 A 1644, 3W51 AJ2 B 1201, 4W5I 3GX A 1204, 3W6O GCP B 801, 4W7T 3JC A 301, 1W8S FBP J 270, 4WCK API A 533, 5WCZ NOJ B 601, 2WEG FBV A 1263, 2WEO FBW A 1263, 3WFG WFG A 1001, 5WFO 5UU N 1101, 2WGG TLM E 1417, 5WGP AUD A 302, 2WI3 ZZ3 A 1225, 5WI1 AOY B 501, 2WJ6 SRT B 1283, 5WJH MHA A 302, 4WKP 3QA C 301, 5WLT 86B A 302, 5WM2 SAL A 601, 2WOG ZZD B 1365, 4WQ6 3TQ B 601, 5WQJ 7N3 B 301, 4WT7 X9X A 401, 3WU2 BCR B 620, 3WUR O4B B 202, 3WVJ B3P B 302, 4WW8 VD9 D 304, 4WYD 3VR B 502, 2WZI 5FN B 1721, 2X0V X0V B 1291, 5X1N DHB B 502, 5X2A 7XO C 1104, 1X38 IDD A 1001, 1X39 IDE A 1001, 3X44 PUS B 401, 5X49 01B B 604, 1X54 4AD A 2001, 1X55 NSS A 3002, 4X5S AZM B 302, 4X6K 3XR A 802, 4X7K 3Z3 A 1101, 1X8D RNS B 1106, 1X8X TYR A 952, 4X8E AVJ B 501, 1XBX 5RP B 502, 1XBX HMS A 501, 1XBY 5RP A 501, 5XBI 81U B 500, 2XDU MT0 A 1228, 4XDA RIB A 401, 5XDE 83R C 504, 5XDG 83U D 503, 2XE8 3PG A 1417, 4XE1 IL5 A 305, 4XEQ PAF C 401, 2XF3 J01 B 600, 2XH9 J01 B 1436, 2XH9 J01 B 1437, 2XII TA9 B 1002, 2XIR 00J A 2169, 4XJ4 3AT A 1006, 4XJ5 GH3 A 1014, 5XJN 88L A 501, 1XKW 188 A 1001, 5XKR BZE D 202, 1XON PIL B 502, 5XQD 8CF A 301, 1XS6 DUT B 2194, 1XTB S6P B 2001, 1XUA HHA A 1001, 2XX2 13C B 1215, 2XXZ 8XQ A 3001, 5XXM LGC B 802, 2XZJ KFN B 503, 4Y14 C0A B 404, 1Y2C 3DE B 1003, 4Y3D 45N A 401, 4Y3G 463 A 401, 4Y3T 46J A 416, 4Y3X 46P A 402, 4Y41 46O A 405, 4Y45 F91 A 405, 4Y47 479 A 401, 4Y57 F63 A 404, 4Y5M 47Y A 401, 4Y5N 487 A 401, 5Y52 AZA D 402, 2YA8 G39 A 1777, 2YA8 G39 B 1777, 4YAB 4CN B 1103, 4YB6 HIS F 302, 1YC4 43P A 301, 2YE2 XQI A 1225, 5YE8 8U3 B 501, 4YGF AZM A 303, 4YI7 BE2 A 400, 4YIA IMN B 401, 2YJX YJX A 1224, 4YJI TYL A 502, 5YJI 8WO A 302, 2YKC YKC A 1224, 2YKV IK2 B 1447, 2YKY PLP A 1446, 4YLA ILV A 401, 4YMX ARG B 301, 2YNE YNE A 1001, 2YPO PHE A 900, 4YPX 4FG A 301, 4YPZ 4FL A 301, 4YQ8 4FV A 301, 4YQJ 4GT A 301, 2YR6 BE2 B 1906, 5YSQ INS B 301, 4YTR TGK D 402,

4YTT PUF B 403, 4YWY PBD C 402, 4YX4 FB2 A 303, 4YXI 4J8 A 303, 4YXU 4JE A 305, 2YYJ 4HP A 550, 2YZB URC F 2307, 2Z1Y LEU B 401, 2Z27 DOR A 1410, 2Z28 NCD B 2410, 2Z29 NCD B 2410, 1Z4K T3P A 4341, 4Z4S FUL B 604, 1Z57 DBQ A 1, 1Z5O MTA B 9233, 1Z82 G3H A 600, 1Z9G RRT E 1006, 1Z9Y FUN A 500, 2Z9X ALA B 2502, 4ZBB GDN B 300, 4ZBB GDN D 300, 2ZC9 22U H 1501, 2ZCZ TRP C 100, 2ZDT 46C A 901, 1ZFQ ZEC A 300, 1ZGE SDA A 300, 3ZGL 10E A 1311, 2ZI6 3D1 D 4302, 5ZJ6 VSE A 601, 1ZL2 ANU B 7016, 3ZMC GPP B 1292, 3ZO1 SIJ A 1351, 2ZP1 IYR A 501, 4ZQT 4QP A 1101, 1ZUW DGL C 3301, 2ZVP NPO X 1202, 4ZVK ET B 301, 4ZVN AO A 303, 1ZWH RDE A 1001, 2ZX2 RAM B 198, 3ZXH E41 B 401, 2ZY1 830 A 808, 2ZYV PPS X 1501, 1ZZS DP9 A 799, 2ZZ5 6CN B 302, 4ZZX FSU A 1584, 4ZZZ FSU B 2015

B.0.1.2 ProtFlex18$_{id}$ Data Set

3A22 ARA B 771, 4A6W 5C1 A 1265, 2ALW MNM A 4001, 5AOK GOH B 1294, 3BD7 CKB A 998, 2BU9 HFV A 1333, 2CBV CGB A 1447, 4CMI M4V C 1270, 4CYP A62 A 1000, 5D05 PHE D 406, 6D6L FY4 B 300, 4DGM AGI A 406, 5DJ9 PXG B 508, 5DWR 5H7 A 401, 3E08 TRP H 403, 5E28 BC5 A 302, 2EG7 OTD A 410, 4EGO 1F1 B 502, 5EGH PC A 512, 6ELN P4A A 301, 6EN6 BJ2 C 709, 3F2P S3B A 3000, 2F7Q AOL A 5009, 3FK0 S3P A 428, 4FU9 675 A 313, 5FYR INS D 301, 5G09 6DF D 1476, 6G37 FTU A 801, 6G9U ETK C 302, 3GIQ G01 B 481, 4GQN INI C 301, 5GUD 2IT A 501, 5GWE GWM D 502, 3HKY IX6 A 579, 3HLJ V21 A 262, 2I5X UA5 B 702, 5I7S E9P A 302, 3IMG BZ2 B 302, 5IOQ DUR B 303, 5IVC 6E7 A 601, 4J5J 478 B 401, 5JE1 TOF A 302, 3JT4 JM8 A 800, 4JZX IPE B 401, 3K8D KDO C 1244, 5LOM SNW A 301, 5LPD 71U H 307, 5LRF KS3 A 901, 4LVG 20O B 601, 3M0M AOS B 3002, 3M1Z BMP B 229, 5M4J GLY B 503, 1MMK TIH A 428, 4MOL 2FG C 703, 4MWQ G39 A 513, 5MXF MFU A 401, 5N25 8HK A 302, 4N7C AEF A 202, 3NHW ZXZ A 234, 4NH7 E0G A 301, 2NMX M25 B 312, 3NQM BMP I 2, 4NR0 TCL A 302, 4NR4 2LK B 1201, 3NVZ I3A L 1, 5NZN G39 B 503, 3OF3 DIH L 500, 5OHT 9VH A 702, 2P15 EZT A 600, 3P93 KDG H 407, 4PNT IQD B 1202, 2POW I7C A 1000, 4PSR FUL B 609, 2PVW G88 A 1768, 4Q6W 3HB A 501, 4QXC OGA A 600, 2RDN 1PL A 280, 3RLQ 3RQ A 901, 4RPO T6C D 402, 4RT2 N6T A 406, 3S2J L3A A 401, 3SMR NP7 C 1000, 5TJZ PDC A 301, 5TY8 7Q1 A 302, 1U1W 3HA B 701, 3U5K 08J A 1, 4U73 Q02 A 404, 4UAU XBP B 301, 3UKJ ENO A 401, 3UU1 14B C 404, 3UV7 0CN A 318, 3UXM 0DN D 803, 2V30 U5P B 1479, 5VD3 H8H A 401, 3VHV LD1 A 1, 2VO3 M04 A 1352, 2VVN NHT B 1716, 3VXI ASC A 502, 3WH8 IFM A 502, 5WQK 7NC B 301, 4X8D AVI B 502, 2XFP XCG B 602, 2XFS J01 B 600, 2XGT NSS A 1550, 4Y3Q F02 A

401, 2Y7K SAL B 1304, 2YC3 MW5 A 1301, 4YPY 4F9 A 301, 4YRW URC B 4006, 4YXO 4JC A 305, 2ZGB 21U H 1801, 4ZLU 4PW A 602, 3ZOS 0LI B 1000, 4ZUL UN1 F 602

B.0.1.3 ProtFlex18$_{od}$ Data Set

18GS GDN A 210, 5A06 SOR F 1342, 2A2C NG1 A 459, 3A3G DLZ A 191, 1A8I GLS A 998, 2AE2 PTO A 262, 3AFH GSU A 2001, 3AG6 PAJ A 501, 5ALU HD2 A 1548, 3ANN SYE A 800, 5AQZ SGV A 1389, 3B1N MZR A 401, 3B3C PLU A 500, 6B3H CN4 B 101, 4B4V L34 B 2001, 3B6H MXD A 551, 3B7U KEL X 707, 4BB9 F1P A 702, 2BL9 CP6 A 1240, 6BLD DXJ A 502, 4BQH 9VU A 1539, 4BRK UNP A 1395, 3BXO UPP A 239, 4BZB DGT D 900, 5C1R 51N B 403, 2C29 DQH F 1332, 3C3U C2U A 351, 4C5A DS0 A 311, 4C7G NGO A 1495, 6CA3 MIG A 701, 3CKL STL B 501, 5CMM SYM A 301, 3CP6 RSX A 401, 5CPO XEN B 401, 6CSP FBM B 805, 5CXX FER C 301, 5CY3 55Y A 701, 4CZH F90 A 1335, 4DBS 0HV A 403, 3DDQ RRC C 299, 3DDU 552 A 901, 5DF1 58X B 901, 4DI9 0GY A 401, 4DK4 DUN B 301, 5DKY NOJ A 1000, 4DTZ LDP A 501, 5DY2 DIN A 402, 2E2R 2OH A 1401, 5EDB 5M8 A 201, 4EE0 GSF B 202, 4EGO 1F1 A 504, 3EI6 PL4 A 434, 5EKD 5BX A 401, 6EK3 OUL B 901, 5EOB 5QQ A 1401, 1EQC CTS A 401, 3ESS 18N A 1, 4EZ9 D3T D 901, 5F0X DAT B 504, 4F2W TDI A 301, 5F3Z 5V5 A 1003, 2F4J VX6 A 514, 1F9H PH2 A 181, 3FAZ NOS A 301, 5FBN 5WE C 702, 6FC1 MGP A 301, 2FDU D1G A 501, 4FEP 6AP B 101, 5FH8 5XK C 801, 3FJZ GPF A 429, 5FJK EM6 A 1350, 6FYR EAQ A 501, 3GER 6GU A 91, 2HOX P1T A 6001, 4HO4 THM A 303, 1I1D 16G A 905, 1K97 CIR A 502, 5KGJ X6X A 402, 3LXV 4NC M 1, 4MMM BP7 C 201, 2O06 MTA A 501, 3OEM OEM A 287, 2OFI ADK A 301, 2QIM ZEA A 160, 4QWB YYY A 401, 4RXT ARA A 401, 3TD9 PHE A 400, 5TPU TYD C 201, 2V7J TRP A 1360, 5VJF 13P B 404, 3VMK IPM A 401, 1WLJ U5P A 300, 5WP4 PC A 702, 3WU2 BCR b 622, 4WUT FCB A 404, 2YA7 ZMR C 1776, 1YRD CAM A 420, 4ZBO ETE B 303, 2ZFZ ARG D 300, 4ZJP RIP A 301, 2ZZD TLA L 4004

B.0.1.4 The Other ProtFlex18 Pockets

1A05 IPM A 401, 2A1N CAM A 1422, 2A1O CAM A 1422, 2A1O CAM B 2422, 3A22 ARA A 701, 3A22 ARA A 711, 3A22 ARA A 721, 5A4U I3A A 1213, 5A4U I3A B 1213, 5A6X MFU A 201, 5A6X MFU B 201, 1A96 XAN B 304, 2AAC FCB A 1, 2AAC FCB B 179, 2AAY GPJ A 702, 4AG9 16G A 1168, 5AGN 4JK A 800, 4AIA ADK A 400, 4AIZ 88Q D 1109, 4AJH 88S A 1334, 4AJL 88W A 1333, 5AL3 TGW A

3168, 6ALO ARG A 402, 6ALQ ARG A 403, 4AM8 PAO B 402, 4AM8 PAO C 402, 5AM9 GLU B 911, 5AMB ILE P 41, 3AN1 URC A 1333, 3AN1 URC B 1332, 5AOK GOH A 1292, 6AO8 ARG A 601, 6APS SV2 A 301, 2AQJ TRP A 650, 2ATJ BHO A 353, 5AUW QUE A 400, 3AVS OGA A 1501, 4AXX 3PG A 1421, 3AYI HCI A 907, 3AYJ PHE A 904, 3AYJ PHE B 1904, 4AYR IFL A 503, 4AYU N8P A 499, 4AYU N8P B 499, 4AZJ SEP A 500, 3B0Y DGT A 576, 6B04 C6J A 401, 6B04 C6J B 401, 2B13 TLA B 500, 3B1E 0JO B 401, 3B1E 0JO C 401, 3B1Q NOS E 401, 5B2E MQG A 302, 5B2E MQG B 302, 3B3M JI1 A 800, 3B3N JI2 A 800, 5B5E BCR b 618, 5B5E BCR b 619, 5B5E BCR B 619, 5B5E BCR b 620, 5B5E BCR B 620, 6B5E TYD B 304, 5B66 BCR B 619, 5B66 BCR b 620, 5B66 BCR b 621, 3B7E ZMR A 1001, 4B7R G39 A 801, 4B7R G39 C 801, 4B7U BCN A 1331, 2B8T THM A 4970, 4BC5 5FX A 1531, 3BEX PAU A 248, 3BEX PAU B 248, 3BEX PAU C 248, 4BG1 OGA A 900, 4BG4 ARG A 403, 4BG4 ARG B 403, 4BHG OGA A 400, 2BIB PC A 1541, 3BLB SWA A 1048, 6BL2 ICT A 502, 2BOI MFU B 700, 5BQF TLA A 404, 6BQ5 MTA B 402, 4BR2 UNP A 1501, 1BTO SSB A 378, 2BUU 4NC B 1542, 2BUZ 4NC B 1542, 4BVO TLA A 1394, 3BWL I3A A 601, 3BWY DNC A 302, 4BWL MN9 C 1297, 5BWH DHY C 403, 5BWH DHY D 403, 3BXE 13P A 401, 4BZ5 TLA A 700, 4BZ5 TLA B 700, 4BZ5 TLA C 700, 4BZB DGT B 800, 3C0V ZEA A 156, 2C1L TLA A 1363, 6C2Z P1T A 501, 3C39 3PG A 417, 4C5B DAL A 311, 4C5B DAL B 311, 4C5C DAL A 311, 4C5C DAL B 311, 2C6Z CIR A 1281, 4C6X TLM A 1419, 4C6Z TLE A 1420, 1C9S TRP C 81, 1C9S TRP F 81, 1C9S TRP R 81, 1C9S TRP U 81, 4C9L CAM A 1419, 4C9L CAM B 1419, 4C9O CAM A 423, 4C9O CAM B 423, 4C9P CAM A 423, 4C9P CAM B 423, 6C9X VOG A 701, 5CBS E42 A 301, 5CDG PFB A 404, 5CDH TLA A 401, 5CDH TLA B 401, 5CDH TLA D 401, 5CDH TLA E 401, 5CDH TLA F 401, 5CDS PFB A 404, 2CHN NGT A 1718, 3CHC ZRG A 439, 4CHS GS8 A 1217, 5CI5 T6T A 501, 2CJF RP4 B 1351, 2CJF RP4 K 3151, 6CJA F0G A 401, 6CJA F0G B 401, 6CJA F0G C 400, 2CL5 BIE A 1218, 4CLR FDB A 1270, 4CLR FDB B 1270, 5CLD ADK B 101, 4CM4 4NR A 1270, 4CM6 AOB B 1270, 4CMI M4V A 1270, 4CMK FQW A 1270, 1CQ1 PQQ A 504, 1CQ1 PQQ B 504, 3CTP XLF A 401, 4CXM MTA B 540, 2CYB TYR B 401, 2CZE U5P B 402, 2CZL TLA A 401, 1D0C INE A 760, 1D0C INE B 761, 1D0O INE A 760, 5D04 PHE C 407, 5D05 PHE A 406, 5D05 PHE B 407, 5D05 PHE C 404, 2D1G ETE A 1001, 4D1J DGJ A 600, 4D1J DGJ B 600, 4D1J DGJ C 600, 4D1J DGJ D 600, 4D1J DGJ E 600, 4D1J DGJ F 600, 4D1J DGJ G 600, 4D1O ARG B 700, 5D3U TRP A 502, 3D46 TLA A 502, 3D46 TLA B 502, 3D46 TLA C 502, 3D46 TLA D 502, 3D46 TLA E 502, 3D46 TLA F 502, 3D46 TLA G 502, 3D46 TLA H 502, 5D85 P1T A 402, 5D9Y OGA A 2001, 3DCW EZL A 301, 3DDS CFF A 904, 3DDS NBG A 901, 3DDS NBG B

901, 3DDW CFF A 903, 4DE0 0JB A 300, 4DE3 DN8 A 301, 5DEQ ARA B 301, 4DF1 BMP A 301, 6DGM 1GP A 902, 6DGM 1GP B 902, 3DHF NMN A 503, 3DJE FSA A 501, 4DJU TLA B 502, 4DJV TLA B 502, 4DJX TLA B 502, 5DJ9 PXG A 505, 5DKV T6T A 401, 5DKV T6T B 401, 5DKV T6T C 401, 4DNJ ANN A 502, 4DO1 ANN B 502, 4DO1 ANN C 502, 4DO4 DJN A 510, 4DO5 DGJ A 509, 1DRK RIP A 272, 2DRI RIP A 272, 1DUV PSQ H 402, 1DUV PSQ I 403, 3DUR KDO A 303, 4DUB LDP A 501, 3DYO IPT B 2001, 1DZ4 CAM A 502, 1DZ4 CAM B 502, 1DZ6 CAM A 502, 1DZ8 CAM A 503, 3E08 TRP A 403, 3E08 TRP B 403, 3E08 TRP C 403, 3E08 TRP D 403, 3E08 TRP F 403, 5E0I 5J6 B 500, 4E1O PLP A 1000, 4E1O PLP D 1000, 4E1O PLP F 1000, 3E2T TRP A 3, 4E30 TYD A 503, 5E3K 5JV B 501, 2E40 LGC A 1001, 3E5U 3C4 A 501, 3E5U 3C4 B 503, 3E5U 3C4 C 504, 3E5U 3C4 D 502, 3E5X 3C4 A 501, 3E5X 3C4 C 502, 3E5X 3C4 D 503, 2E68 DOR A 1353, 3E7M AT2 A 906, 4E70 N7I A 403, 3EA2 INS A 801, 4EAR IM5 A 301, 1EC8 GLR B 500, 1ED5 NRG A 1705, 5EEU TRP A 101, 5EEU TRP D 101, 5EEU TRP K 101, 5EEU TRP O 101, 5EEU TRP Q 101, 5EEU TRP R 101, 5EEU TRP S 101, 5EEV TRP A 101, 5EEV TRP D 101, 5EEV TRP K 101, 5EEV TRP N 101, 5EEV TRP O 101, 5EEV TRP Q 101, 5EEV TRP R 101, 5EEW TRP A 101, 5EEW TRP D 101, 5EEW TRP N 101, 5EEW TRP O 101, 5EEW TRP Q 101, 5EEW TRP R 101, 5EEX TRP D 101, 5EEX TRP K 101, 5EEX TRP N 101, 5EEX TRP O 101, 5EEX TRP Q 101, 5EEY TRP D 101, 5EEY TRP N 101, 5EEY TRP O 101, 5EEY TRP Q 101, 5EEZ TRP N 101, 5EEZ TRP O 101, 5EF1 TRP O 101, 4EGO 1F1 A 502, 4EGO 1F1 C 502, 4EGO 1F1 D 503, 5EGH PC B 510, 3EHW DUP A 777, 3EHW DUP B 777, 5EH5 XCZ A 303, 5EHM OEM A 301, 5EHM OEM B 301, 1EIR BPY A 301, 2EI0 BP7 A 402, 5EJ9 TPP A 601, 5EJ9 TPP B 602, 5EJ9 TPP C 601, 5EJ9 TPP D 602, 5EJ9 TPP E 601, 5EJ9 TPP F 602, 5EJ9 TPP G 601, 5EJA TD6 B 601, 4EK1 CAM A 502, 4EK1 CAM B 502, 5ELO KRS A 602, 6EN5 BJ2 C 702, 6EN6 BJ2 A 711, 6EN6 BJ2 B 710, 4EO6 0S2 A 600, 2EPN NGT A 1650, 3EPW JMQ A 1002, 1EQJ 2PG A 801, 4EQL SAL A 602, 3EVM YYY B 201, 3EVO TYD B 161, 3EWZ CNU B 481, 1EXA 394 A 450, 2EXS TRP B 2100, 3EX1 6CN B 481, 3EX2 6CN A 481, 3EXE TPP A 1005, 3EXE TPP C 1002, 3EXE TPP E 1011, 3EXE TPP G 1008, 5EXK 5AD A 403, 5EXK 5AD C 403, 5EXK 5AD E 403, 4F0S NOS A 501, 5F27 5TT A 301, 6F22 C9B A 200, 5F5N 5VD A 302, 6F6A CU5 A 301, 5F76 MTA A 301, 5F76 MTA B 301, 5F8Y X6X A 201, 5F8Y X6X A 202, 2F90 3PG A 408, 2F90 3PG B 409, 4FCK GPA A 401, 2FEU CAM A 1420, 2FEU CAM B 1421, 4FEO 6AP B 101, 5FHR DNC A 301, 5FHR DNC B 301, 5FII PHE A 901, 5FIU TLA B 1300, 5FIU Y3J A 1299, 3FJ7 PEQ A 301, 3FJX S3P A 428, 3FJZ S3P A 430, 4FJ7 DGT A 1001, 5FKY 2J4 A 1717, 6FNQ AVJ A 401, 6FNS DY8 A 401, 3FO4 6GU A 91, 2FPZ 270 A 1000, 2FPZ 270 B 1001, 3FPF

MTA A 301, 1FSG 9DG A 304, 3FTV 11X A 710, 6FT2 ARG A 306, 3FUW MTA A 272, 3FUX MTA A 272, 3FVG MS8 A 901, 3FVK 8DX A 1, 4FVY 3KJ A 804, 3FWF CAM A 420, 3FWF CAM B 420, 3FWG CAM A 420, 3FWG CAM B 420, 3FWJ CAM A 420, 6FWH 5LD A 203, 6FWH 5LD B 203, 6FWH 5LD B 204, 6FWH 5LD D 202, 6FWH 5LD E 202, 6FWH 5LD F 203, 4FXR BMP A 301, 5FYR INS A 301, 5FYR INS B 301, 5FYR INS C 301, 1G0R THM B 2531, 3G1A UP6 A 229, 3G1A UP6 B 229, 3G1V 5FU A 501, 3G24 UP6 A 229, 5G2T UAP C 511, 6G2N O84 A 302, 3G35 F13 A 1, 4G41 MTA B 300, 1G6S GPJ A 701, 1G6S S3P A 601, 1G6T S3P A 601, 2G6H ARG A 770, 2G6H ARG B 771, 2G6I ARG A 770, 2G6I ARG B 771, 2G6K ARG B 771, 2G6M ARG A 770, 2G6M ARG B 771, 2G6N ARG B 771, 4G7A AZM A 302, 4G88 API A 401, 4G88 API D 401, 3GAO XAN A 90, 3GBE NOJ A 8000, 4GC4 BMP A 301, 3GDN MXN A 531, 2GG6 S3P A 601, 2GGA GPJ A 701, 2GGA S3P A 601, 4GHG DHY C 403, 3GIQ G01 A 481, 4GJY OGA A 502, 5GLP ARA A 403, 6GL9 PHU B 1202, 3GN0 DMO A 551, 3GQY TLA A 542, 3GQY TLA B 542, 3GQY TLA C 542, 4GQN INI A 301, 4GQN INI A 302, 3GR4 TLA A 542, 3GR4 TLA B 542, 3GR4 TLA C 542, 1GTF TRP P 81, 3GUZ PAF A 177, 4GVX FUL A 303, 4GVX FUL B 304, 4GVX FUL C 303, 3GWC UFP A 260, 3GWC UFP B 260, 3GWC UFP E 260, 5GWE GWM A 502, 5GWE GWM C 502, 2GX6 RIP A 301, 2GZM DGL A 501, 2GZM DGL D 504, 4H3J TLA B 502, 2H4X 3PG A 1408, 4H4D 10E A 402, 4H4D 10E B 402, 5H41 IFM A 1203, 4H5F ARG A 317, 4H5F ARG C 312, 4H5F ARG D 305, 3H78 BE2 A 350, 4HCH TLA A 401, 4HCH TLA B 401, 4HIH RAM D 301, 3HSN HAR A 770, 4HT2 V50 C 302, 3HW8 D3T A 581, 1HXK DMJ A 1103, 1I0L 7HP A 800, 1I13 7HP A 800, 5I2C ARG A 401, 4I3B BLR B 201, 2I5X UA5 A 701, 5I6D AU6 A 402, 5I6D AU6 B 402, 5I6D AU6 C 402, 4I7N 1DJ A 202, 5I7I 3HB A 401, 5I8X ZDC A 201, 3IAR 3D1 A 501, 4IAV CXA A 402, 5IBD GGJ A 407, 5IE0 SRT A 1001, 5IE0 SRT A 1002, 4IIC IFM A 944, 4IIE CGB A 943, 3IJI ALA B 384, 4IJI 99T A 501, 4IJI 99T D 501, 5IJW DGL A 301, 5IJW DGL B 301, 3IK3 0LI B 2, 3INJ BXB A 1001, 3INL BXB A 1001, 3INL BXB B 1001, 3INL BXB C 1001, 3INL BXB D 1001, 3INL BXB E 1001, 3INL BXB F 1001, 3INL BXB G 1001, 2IOY RIP A 401, 2IOY RIP B 402, 4IO7 PHE B 301, 5IOQ DUR B 302, 5IOY 6C5 A 301, 5IOY 6C5 A 302, 3IT1 TLA A 402, 3IT1 TLA B 402, 3IT3 3AM A 343, 3ITL LRH A 601, 3ITL LRH B 602, 3ITL LRH D 604, 3ITU IBM B 999, 3ITU IBM C 999, 3ITV PSJ A 601, 3ITV PSJ B 602, 3ITV PSJ D 604, 5ITP 6DB A 301, 3IU8 T03 A 289, 4IUO QIC A 301, 3IVD URI A 603, 4IV9 TSR A 602, 2IX9 CXS B 1261, 3IX8 TX3 A 174, 3IX8 TX3 C 174, 1J1U TYR A 401, 4J25 OGA D 402, 4J25 OGA E 402, 5J42 TLA A 402, 2J78 GOX A 1451, 5J71 TLA A 501, 1JDF GLR C 2512, 1JDF GLR D 2513, 2JDM MFU B 1117, 2JDM MFU C 1117, 5JDV EVF B 302,

5JDY TOF A 302, 5JE0 AZ8 A 302, 5JE7 EVF B 302, 5JEI FWD A 301, 5JEP EVF B 302, 2JFY DGL A 1256, 2JFY DGL B 1256, 2JFZ 003 A 1256, 2JFZ DGL A 1257, 2JFZ DGL B 1257, 5JG5 EVF B 301, 5JIB OIA C 400, 2JKP CTS B 1727, 5JNA TOR A 302, 5JNA TOR B 302, 5JNA TOR C 302, 5JOY 6LW A 602, 5JOY 6LW B 602, 4JPX PHE A 301, 3JQA DX4 A 270, 4JQA TLA B 403, 4JR5 1LS B 601, 3JT4 JM8 B 800, 1JYX IPT B 2001, 1JYX IPT C 2001, 1JZ5 149 A 2001, 1JZ5 149 B 2001, 1JZ5 149 C 2001, 4JZB IPE A 404, 4JZB P2H A 405, 5K0C 6OZ B 302, 4K1K G39 B 510, 4K1W CS2 A 502, 4K1W CS2 B 502, 4K3N 1OT A 1004, 4K3N 1OT C 1004, 3K5E K5E A 369, 4K8K 1PJ A 402, 3KCZ 3AB A 1, 4KFN 1QR A 601, 4KFO 1QS A 601, 3KGF PHE A 9003, 5KIT 6TA A 501, 4KKY CAM X 503, 4KPL CS2 F 501, 4KPL KDG G 501, 4KPL KDG H 501, 3KSM BDR A 1, 3KVJ DOR A 399, 1KW8 BPY B 401, 1KW9 BPY B 401, 1KWW MFU B 601, 1KWW MFU C 701, 4KWD JF2 A 405, 4KZB NZ2 A 401, 3L0N S5P A 257, 3L0V 724 A 485, 3L2H CXS B 163, 4L4E CAM A 503, 4L4G CAM A 503, 3L63 CAM A 440, 4L6D VNL B 402, 4L6D VNL C 402, 4L6D VNL D 402, 4L6D VNL E 402, 4L6D VNL F 402, 4L6G CNL A 502, 5L9V OGA A 502, 5LBZ 6T3 A 302, 4LFG IPE A 302, 4LFG IPE B 301, 3LHU BMQ A 229, 3LHU BMQ B 229, 3LHV BMQ A 229, 3LHV BMQ B 229, 3LHV BMQ C 229, 3LHW BMQ A 229, 3LHY BMQ A 229, 3LHY BMQ B 229, 3LHZ BMQ A 229, 3LI1 BMQ A 229, 3LLD UP6 A 229, 5LL4 6YH A 306, 3LNK TLA B 4, 3LPJ TLA A 455, 3LPJ TLA B 455, 1LS6 NPO A 3001, 4LS3 HIS A 601, 5LSA DNC A 304, 3LTP BMP A 229, 3LTP BMP B 229, 4LU3 AZM A 302, 4LUJ BMP B 301, 1LVW TYD A 3002, 3LV5 BMP B 229, 3LV6 BMP B 229, 4LVB 20N A 601, 4LVD 1EB A 603, 4LVF 20P A 601, 4LVG 20O A 601, 4LW7 BMP B 301, 4LWW LWW A 601, 5LWM PHU A 1202, 4LXQ TYD A 302, 3M0H RNS A 2001, 3M0H RNS B 2002, 3M0H RNS D 2004, 3M0L PSJ A 601, 3M0L PSJ B 602, 3M0V RNS A 2001, 3M0X PSJ A 601, 3M0X PSJ C 603, 3M1Z BMP A 229, 1M2W MTL A 5600, 1M2W MTL B 6600, 3M4F CXS A 207, 5M4J GLY A 503, 1M5E AM1 A 1700, 1M5E AM1 B 1701, 1M5F AM1 A 1200, 1M5F AM1 B 1201, 4M5R MSR A 304, 4M6P 20R A 601, 5M67 3D1 C 503, 4M81 GLF A 501, 3MBH PXL A 400, 3MBH PXL B 400, 3MBH PXL C 400, 3MBH PXL D 400, 3MBH PXL E 400, 3MFW B3U A 600, 3MFW B3U B 601, 5MFQ 2J9 A 901, 1MI4 S3P A 1001, 3MI2 PFU A 1, 3MJM DOR A 1410, 1MMW VIO A 1780, 4MOG G3F A 802, 4MOL 2FG A 802, 4MOL 2FG B 802, 4MOR 2H5 D 802, 4MSS 2CZ A 401, 3MUZ IPT 2 2001, 4MUY 2E5 A 402, 5MUX TLA A 501, 5MUX TLA B 501, 5MUX TLA E 501, 3MVX BHZ A 504, 4MV0 2E6 A 402, 5MXC MFU A 401, 4MYD 164 A 301, 3N3M NUP A 2000, 3N86 RJP A 147, 3N86 RJP O 147, 4N8G DAL A 402, 4N8G DAL B 402, 4N8G DAL C 402, 1NEY 13P A 5001, 3NG7 HNL X 433, 3NHW ZXZ A 233, 4NJH 2K8 A 303, 4NJH 2K8 B 303,

4NJM 3PG A 401, 2NMX M25 A 311, 1NQX RLP A 1201, 1NQX RLP B 2201, 1NQX RLP C 3201, 2NQ7 HM5 A 410, 3NQ6 UP6 I 1, 3NQ7 BMP B 229, 3NQA BMP I 1, 3NQA BMP I 2, 3NQC BMP I 1, 3NQD BMP I 1, 3NQE BMP A 229, 3NQF BMP A 229, 3NQG BMP I 2, 3NQM BMP I 1, 4NQ8 PAF A 401, 3NRS TLA A 1001, 4NR4 2LK A 1201, 5NS8 NOJ A 506, 5NS8 NOJ B 510, 3NVS GPJ A 429, 3NVW GUN C 503, 5NWE G39 A 503, 1NXJ TLA B 392, 5NYA FB2 A 302, 5NZ4 G39 A 503, 5NZE G39 A 503, 5NZF G39 A 503, 5NZF G39 B 503, 5NZF G39 C 503, 5NZN G39 A 503, 4O10 2QF A 601, 4O13 2P1 A 601, 4O13 2P1 B 601, 5O48 9K2 A 502, 5O4V 9K2 A 502, 2O73 2AL B 1001, 2O73 2AL C 4001, 2O73 2AL D 5001, 2O73 2AL E 2001, 2O74 GUN B 2001, 2O74 GUN C 3001, 2O74 GUN D 4001, 2O74 GUN E 6001, 2O74 GUN F 5001, 2O78 TCA A 701, 2O78 TCA B 701, 2O78 TCA C 701, 2O78 TCA D 701, 2O78 TCA E 701, 2O7D DHC A 701, 2O7D DHC B 701, 2O7D DHC C 701, 2O7D DHC E 701, 2O7D DHC F 701, 1O98 2PG A 801, 3OCC DIH A 500, 3OCC DIH C 500, 3OCC DIH D 500, 3OCC DIH E 500, 3OCU NMN A 2003, 5OCM 9RH A 302, 5OCM 9RH B 302, 5OCM 9RH C 302, 5OCM 9RH D 302, 5OCM 9RH E 302, 5OCM 9RH F 302, 1OFZ FUL A 1313, 3OF3 DIH A 500, 3OF3 DIH C 500, 3OF3 DIH D 500, 3OF3 DIH E 500, 3OF3 DIH H 500, 4OGI R78 A 202, 5OHY 9VH A 704, 5OHY 9VH B 707, 5OHY 9VH C 707, 3OID TCL C 604, 3OID TCL D 602, 1OM4 ARG B 770, 5OMR V55 A 502, 3OOG YTP A 2001, 5OOA URI A 201, 3OTI TYD A 377, 2OU3 I3A A 163, 3OUT DGL A 266, 3OUT DGL B 266, 3OUT DGL C 266, 4OVT LFC A 403, 4OWO 6F0 A 404, 2OZL TPP A 2330, 3P0N BPU A 1163, 1P1U AMQ A 302, 1P1W AMQ A 428, 3P10 CTN A 165, 3P1F 3PF A 1198, 3P5Y BMP A 229, 3P5Y BMP B 229, 3P5Z BMP A 229, 3P5Z BMP B 229, 4P56 SMN B 401, 3P60 BMP B 229, 3P61 BMP B 229, 3P93 KDG D 407, 3P93 KDG G 407, 5P9R 7JJ A 303, 3PB5 F63 A 1001, 3PC4 KOU A 702, 1PG4 PRX A 999, 2PGA ANU B 7016, 2PGN TPP B 615, 4PGN 3IO A 401, 4PGN 3IO B 401, 4PGN 3IO C 401, 4PH9 IBP B 601, 4PIN AVI A 401, 4PIO AVI A 402, 3PKD Y10 A 288, 3PL8 G3F A 903, 4PML 3AB B 1202, 4PNR G18 B 1202, 4PNT IQD A 1202, 4PSR FUL A 622, 2PWD NOJ A 8000, 2PWD NOJ B 8001, 1PX4 IPT A 2001, 1PX4 IPT C 2001, 1PX4 IPT D 2001, 4PZ0 PAV A 401, 3Q12 PAF A 501, 2Q2A ARG A 904, 2Q2A ARG B 903, 3Q23 G2P B 1108, 1Q6Q LXP B 9301, 4Q7F 3D1 A 603, 2Q95 A05 A 400, 3QAX ARG A 600, 2QBL CAM A 517, 2QBM CAM A 517, 3QEZ BMP A 229, 3QEZ BMP B 229, 2QFQ S3P A 701, 2QFS S3P A 701, 2QFT GPJ A 801, 2QFT S3P A 701, 3QF0 BMP B 229, 4QFP VAL A 601, 2QJN KDG D 2004, 2QJW TLA B 179, 1QMG DMV A 620, 1QMG DMV B 620, 1QMG DMV C 620, 1QMG DMV D 620, 3QMS BMP A 229, 3QMT BMP A 229, 4QOJ STL A 302, 2QRL OGA A 500, 3QRY DMJ A 430, 2QSZ NMN A 201, 3QS8 17D A 600,

1QV6 24B A 378, 2QVH OSB A 5550, 2QVH OSB B 5551, 1QXZ M3C A 2001, 2QX0 PH2 A 181, 4QXB OGA A 600, 4QXC OGA C 600, 4R33 TRP A 503, 4R33 TRP B 503, 4R34 TRP A 505, 4R6W PC A 301, 4R93 TLA B 501, 1RDJ MFB 1 1, 3REM SAL A 301, 1RF6 GPJ A 1428, 1RF6 GPJ B 1528, 3RIE JFQ B 501, 3RIE MTA A 401, 4RK1 RIB A 401, 4RK1 RIB B 401, 4RK1 RIB C 401, 3RLR 3RR A 901, 3RLU BMP A 229, 3RLV BMP A 229, 3RLV BMP B 229, 4RLF 4MA A 1001, 3RM8 RM8 A 417, 3RME RME B 418, 4RN0 L6G B 501, 4RPO T6C A 401, 4RPO T6C C 401, 1RS7 MTL B 871, 4RU1 INS A 401, 4RU1 INS B 401, 4RU1 INS C 402, 4RU1 INS D 401, 4RU1 INS E 402, 4RU1 INS F 401, 4RU1 INS G 401, 4RU1 INS H 401, 4RU1 INS I 401, 4RU1 INS J 401, 4RU1 INS K 401, 4RV3 INS A 402, 4RXM INS A 401, 4RY0 RIP A 401, 4RY8 SR1 A 401, 3S2Z DHC B 257, 4S25 IRN A 702, 4S26 IRN A 702, 4S27 AIR A 702, 3S9Y FNU A 324, 3S9Y FNU B 324, 3SBF D8T A 404, 3SBF D8T B 404, 3SCO GLF A 477, 3SCO GLF B 1002, 3SCS GLF A 477, 1SD3 SYM A 998, 3SIZ BMP A 229, 3SLH GPJ A 444, 3SLH GPJ B 442, 3SLH GPJ C 442, 1SMO TLA B 726, 3SMR NP7 B 1000, 3SUR NGT A 2000, 1SW1 PBE A 301, 5SYI NIZ A 805, 1SZO CAX A 5001, 1SZO CAX B 5002, 1SZO CAX C 5003, 1SZO CAX D 5004, 1SZO CAX E 5005, 1SZO CAX F 5006, 1SZO CAX G 5007, 1SZO CAX H 5008, 1SZO CAX K 5011, 1SZO CAX L 5012, 3SZU H6P A 998, 3T09 149 A 2001, 3T09 149 B 2001, 3T09 149 C 2001, 3T09 149 D 2001, 3T0D 149 B 2001, 3T0D 149 C 2001, 3T0D 149 D 2001, 3T44 BE2 A 273, 1T88 CAM A 1422, 1T88 CAM B 2422, 3T9V CNI A 400, 1TC2 7HP A 800, 1TC2 PRP A 801, 3TCF UNK I 1, 3TCF UNK K 1, 3TCF UNK L 1, 3TCF UNK M 1, 3TCF UNK N 1, 5TE2 7B9 A 501, 3TG2 ISC A 501, 3TI6 G39 A 801, 3TL1 JRO A 160, 3TR9 PT1 A 1001, 4TSN PC A 202, 3TWP SAL B 404, 3TX6 ENO A 386, 4TXJ THM A 301, 4TXJ THM B 301, 4TXJ THM C 301, 5TXY 7Q1 A 303, 1TZC PA5 A 600, 5TZA 7OG B 1001, 5TZA 7OG C 1001, 5TZZ 7OJ A 1001, 5TZZ 7OJ B 1001, 5TZZ 7OJ C 1001, 5U00 7OV B 1001, 5U00 7OV C 1001, 1U1W 3HA A 700, 3U15 03M C 1, 4U1O KAI A 301, 4U21 FWD A 401, 4U22 FWD A 401, 5U2M 7T7 A 901, 1U3U BNF A 1378, 5U3F 7TS A 400, 4U4X 3C2 A 801, 5U62 7WD A 504, 5U98 1KX A 301, 5U9P TLA A 302, 5U9P TLA B 302, 5U9P TLA B 303, 5U9P TLA C 302, 5U9P TLA C 303, 5U9P TLA D 302, 5U9P TLA D 303, 4UAT XBP A 301, 4UAT XBP B 301, 4UB6 BCR B 620, 3UCC 2PG A 601, 3UES DFU A 501, 1UF5 CDT A 998, 1UHH CZP A 1001, 1UHK CZN B 2001, 3UHF DGL A 260, 3UHF DGL B 260, 3UJ9 PC A 301, 3UJE 2PG A 503, 3UJR 2PG A 503, 3UJS XSP A 602, 3UK0 ENO A 501, 1UMD TDP A 1402, 3UNA SAL A 1340, 3UNA SAL B 1340, 3UNC SAL A 1338, 4UOV AZM A 299, 3UU1 14B B 404, 4UVZ 5NN A 2165, 1UWT GTL A 1490, 1UWU GOX A 1490, 3UWQ U5P B 232, 2UZ1 TPP A 1556, 2UZ1 TPP B 1557, 2UZ1 TPP C 1557, 5UZ0 AMZ A 601,

5UZ0 AMZ B 603, 5UZ0 AMZ C 601, 3V1P BMP A 229, 4V24 GYR A 1450, 5V2C BCR a 411, 5V2C BCR b 618, 5V2C BCR B 618, 5V2Y ARG A 402, 1V3E ZMR A 1200, 2V30 U5P A 1480, 2V3D NBV A 1503, 2V5Z SAG A 1503, 3VC1 GST A 303, 3VC1 GST B 303, 3VC1 GST C 303, 3VC1 GST D 303, 3VC1 GST E 303, 3VC1 GST G 303, 3VC1 GST H 303, 3VC1 GST J 303, 3VC1 GST K 303, 3VC3 C6P A 501, 3VC3 C6P B 501, 3VC3 C6P D 501, 3VC3 C6P E 501, 3VD4 IPT B 2001, 2VFG 3PG B 1249, 3VIF LGC A 507, 2VJX IFL A 1865, 2VL4 MNM A 1865, 2VTF B3P A 1618, 2VVN NHT A 1718, 2VVT DGL A 1270, 2VVT DGL B 1270, 3VVF ARG A 301, 3VYG TLA C 302, 3VYG TLA F 302, 3VYG TLA L 302, 1W1T CHQ A 1513, 1W1Y TYP A 1509, 5W16 DGL A 301, 5W16 DGL B 301, 5W16 DGL C 301, 5W16 DGL D 301, 5W1Q DGL A 301, 5W1Q DGL B 301, 2W4I DGL A 1256, 2W4I DGL B 1255, 2W4I DGL E 1255, 2W4I DGL F 1255, 2W5R GP9 A 1644, 3W51 AJ2 A 1201, 3W5N RAM A 1202, 3W6O GCP A 801, 1W8S FBP F 270, 1W8S FBP G 270, 1W8S FBP I 270, 5WCZ NOJ A 601, 2WEJ FB2 A 1263, 2WGG TLM A 1417, 5WG7 AUD A 301, 5WG8 LB1 A 503, 5WGD EST A 601, 5WI1 AOY A 901, 2WJ6 SRT A 1286, 2WK9 PLG B 600, 3WLV AZA A 401, 3WLV AZA B 401, 3WLV AZA C 401, 3WLV AZA D 401, 2WOG ZZD A 1365, 3WQE PLP B 401, 5WQK 7NC A 301, 3WRH CAM A 503, 3WRH CAM E 503, 3WRJ CAM A 503, 3WRJ CAM E 503, 3WRL CAM A 503, 3WRL CAM E 503, 3WRM CAM A 503, 3WRM CAM F 503, 3WU2 BCR A 411, 3WU2 BCR b 621, 2WVT FHN A 1473, 2WVT FHN B 1474, 4WW8 VD9 A 305, 4WW8 VD9 C 304, 2WYW TCL B 1260, 2WYW TCL C 1260, 2WZI 5FN A 1719, 4WZZ RAM A 401, 2X0V X0V B 1290, 2X14 3PG A 1419, 5X1M DHB A 501, 5X1N DHB A 502, 3X44 PUS A 400, 5X49 01B A 604, 4X5S AZM A 302, 1X8D RNS A 1105, 4X8D 3GC A 501, 4X8D AVI A 502, 4X8E AVJ A 501, 5XBI 81U A 500, 2XCG XCG A 602, 5XDG 83U B 503, 5XDG 83U C 503, 1XES 3IO A 2000, 1XES 3IO B 3000, 1XES 3IO C 4000, 2XE6 3PG A 1417, 4XEQ PAF A 401, 4XEQ PAF B 401, 2XF3 J01 A 600, 2XFS J01 A 500, 2XFS J01 A 600, 2XFS J01 B 500, 4XFP AZA A 401, 4XFP AZA B 401, 4XFP AZA C 401, 4XFP AZA D 402, 2XH9 J01 A 1437, 2XII TA9 A 1002, 4XKN HIS A 503, 1XNZ FCD A 268, 1XON PIL A 501, 1XS6 DUT A 1194, 5XXM LGC A 802, 2XZJ KFN A 503, 5Y2P AZA A 401, 5Y2P AZA B 401, 5Y52 AZA A 401, 5Y52 AZA B 401, 5Y52 AZA C 401, 2Y7I ARG A 1245, 2Y7K SAL A 1302, 4Y9T PA1 A 401, 4YB6 HIS A 302, 4YB6 HIS B 302, 4YB6 HIS C 302, 4YB6 HIS D 302, 4YB6 HIS E 302, 5YE8 8U3 A 501, 4YMA 4E5 B 304, 4YMX ARG A 301, 1YNH SUO A 1001, 1YNH SUO B 1002, 1YNH SUO C 1003, 1YNH SUO D 1004, 4YO7 INS A 405, 1YRC CAM A 420, 2YR6 BE2 A 906, 4YRW URC A 3006, 4YTR TGK C 403, 4YTT PUF A 403, 4YW8 1WD A 704, 4YW9 1WD A 706, 2YZB URC A 2304, 2YZB URC D 2301, 2Z1Y LEU A 400,

2Z25 DOR A 1410, 2Z26 DOR A 1410, 2Z27 NCD B 2410, 4Z4R FUL B 605, 1Z5O MTA A 5233, 4Z6Q DHY C 403, 4Z6Q DHY D 403, 4Z6S 4SX C 402, 2Z9X ALA A 1502, 4ZBB GDN A 300, 4ZCW 4NG B 501, 2ZI7 GNG A 502, 2ZUX RAM A 639, 2ZUX RAM A 641, 2ZUX RAM B 639, 4ZUL UN1 B 602, 2ZVP NPO X 1201, 2ZYT PPS X 501, 2ZYU PPS X 501, 2ZZ5 6CN A 301, 2ZZD TLA C 4001, 2ZZD TLA F 4002, 2ZZD TLA I 4003

Appendix C

Scientific Contributions

My scientific contributions in the time of my PhD thesis are listed in the following.

C.1 Publications in Scientific Journals

Flachsenberg, F.; Meyder, A.; Sommer, K.; Penner, P.; Rarey, M. (2020) *A Consistent Scheme for Gradient-Based Optimization of Protein-Ligand Poses*, in preparation.
My contribution in this work is the Continuous Torsion Score, the initial design of the `NumOptimization` library and overall discussion about validation.

Schöning-Stierand, K.; Diedrich, K.; Fährrolfes, R.; Flachsenberg, F.; Agnes Meyder, A.; Nittinger, E.; Steinegger, R., Rarey, M. (2020) *ProteinsPlus: Interactive Analysis of Protein-Ligand Binding Interfaces* Nucleic Acids Research, in submission.
Second paper about the group's web server now with my tools `StructureProfiler` and `EDIAScorer` on `https://proteins.plus`. I co-wrote the section about the `StructureProfiler`.

Friedrich, N.-O.; Flachsenberg, F.; Meyder, A.; Sommer, K.; Kirchmair, J.; Rarey, M. (2019) *Conformator: A Novel Method for the Generation of Conformer Ensembles.* Journal of Chemical Information and Modeling, 59(2): 731-742.
The Conformater uses the Continuous Torsion Score (CTS, see Chapter 4) developed by me as part of a force field. I also co-wrote the section about the CTS.

Meyder, A.; Kampen, S.; Sieg, J.; Fährrolfes, R.; Friedrich, N.-O.; Flachsenberg, F.; Rarey, M. (2019) *StructureProfiler: An all-in-one Tool for 3D Protein Structure Profiling.* Bioinformatics, 35(5): 874–876.
The `Structureprofiler` is introduced in Chapter 3 and was published online in

2018. I conceptualized the project, wrote the final version of the tool and rerun al tests based on S. Kampen's and J. Sieg's preliminary work and wrote the paper.

Meyder, A.; Nittinger, E.; Lange, G.; Klein, R.; Rarey, M. (2017) *Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-ray Structures.* Journal of Chemical Information and Modeling, 57(10):2437–2447.
EDIA (Chapter 2) is published in this publication. I conceptualized the project based on previous work of E. Nittinger, implemented, tested, validated the tool and the metric and wrote the paper.

Bietz, S.; Inhester, T.; Lauck, F.; Sommer, K.; von Behren, M.; Fährrolfes, R.; Flachsenberg, F.; Meyder, A.; Nittinger, E.; Otto, T.; Hilbig, M.; Schomburg, K.; Volkamer, A.; Rarey, M. (2017) *From cheminformatics to structure-based design: Web services and desktop applications based on the NAOMI library.* Journal of Biotechnology, 261:207-214.
`EDIAScorer` joined the group's AMD ChemBio suite for standalone tools in 2017. I co-wrote the section of the tool.

Fährrolfes, R.; Bietz, S.; Flachsenberg, F.; Meyder, A.; Nittinger, E.; Otto, T.; Volkamer, A.; Rarey, M. (2017) *ProteinsPlus: a web portal for structure analysis of macromolecules.* Nucleic Acids Research, 45:W337-W343.
The first publication about the group's web server `https://proteins.plus` in 2017 with EDIA integrated into the `EDIAScorer` available for e.g. visually inspecting the structure. I co-wrote the section of the tool.

Nittinger, E.; Inhester, T.; Bietz, S.; Meyder, A.; Schomburg, K.T.; Lange, G.; Klein, R.; Rarey, M. (2017) *A Large-Scale Analysis of Hydrogen Bond Interaction Patterns in Protein-Ligand Interfaces.* Journal of Medicinal Chemistry, 60:4245-4257.
I supplied EDIA for the study.

Schomburg, K.T.; Nittinger, E.; Meyder, A.; Bietz, S.; Schneider, N.; Lange, G.; Klein, R.; Rarey, M. (2017) *Prediction of protein mutation effects based on dehydration and hydrogen bonding - A large-scale study.* Proteins, 85(8):1550-1566.
I contributed the preliminary version of GeoHYDE (see Chapter 5) for the analysis and co-wrote its section.

Friedrich, N.-O.; Meyder, A.; Sommer, K.; Flachsenberg, F.; de Bruyn Kops, C.; Rarey, M.; Kirchmair, J. (2017) *High-quality dataset of protein-bound ligand conformations and its application to benchmarking conformer ensemble generators.* Journal of Chemical Information and Modeling, 57(3): 529-539.
Platinum uses a preliminary version of my version of the EDIA for analyzing ligands (Chapter 3) and I contributed in the area of the overall quality factor discussion.

Guba, W.; Meyder, A.; Rarey, M.; Hert, J. (2016) *Torsion Library Reloaded: A New Version of Expert-Derived SMARTS rules for Assessing Conformations of Small Molecules.* Journal of Chemical Information and Modeling, 56(1): 1-5.
The TorLib16 was published through this Application Note (see Chapter 4). My contribution for the TorLib16 was the automatic error analysis such as double peak detection.

C.2 Talks

Meyder, A.; Schmidt, R.; Rarey, M. (2018) *Automatic SMARTS Hierarchy Analysis for the Updated Torsion Library and Its use in Scoring Torsion Angles* as Flash Oral Presentation at the EuroQSAR 2018 in Thessaloniki, Greece.

Meyder, A.; Nittinger, E.; Lange, G.; Klein, R.; Rarey, M. *EDIA: Estimating Electron Density Support for Individual Atoms in X-ray Structures* at the Sheffield Conference on Chemoinformatics 2016 in Sheffield, United Kingdoms.

C.3 Posters

Meyder, A.; Kampen, S.; Sieg, J.; Flachsenberg, F.; Fährrolfes, R.; Ehmki, E.; Nittinger, E.; Rarey, M. (2017) *StructureChecker: An all-in-one tool for high quality 3D structure data set assemblage* at the 13. German Conference on Chemoinformatics in Mainz, Germany.

Meyder, A.; Nittinger, E.; Lange, G.; Klein, R.; Rarey, M. (2017) *Extending Rescoring Validation with the Electron Density Score of Individual Atoms (EDIA)* at

the Gordon Research Conference for Computer-Aided Medicine Design in Mount Snow, USA.

Meyder, A.; Fährrolfes, R.; Nittinger, E.; Lange, G.; Klein, R.; Rarey, M. (2016) *A Novel Web Service To Estimate the Electron Density Support For Individual Atoms in X-ray Structures* at the 12. German Conference on Chemoinformatics in Fulda, Germany.

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.