Molecular Binding Kinetics of CYP P450 by Using the Infinitesimal Generator Approach

Master thesis

by

Paula Breitbach

Matrikelnr. 337500

Technical University Berlin

First Supervisor: PD Dr. Konstantin Fackeldey

Second Supervisor: PD Dr. Marcus Weber

May 18, 2018

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Die selbstständige und eigenständige Anfertigung versichert an Eides statt:

Ort, Datum	Unterschrift

Zusammenfassung

Wir betrachten das Enzym CYP 3A4, dass für den Stoffwechsel vieler Medikamente und Xenobiotika verantwortlich ist. Es gehört zur Familie der Cytochromen P450, deren Mitglieder alle einen ähnlichen Aufbau aufweisen, jedoch auf verschiedene Liganden spezifiziert sind. Die katalysierte Reaktion ist bekannt, aber der Weg, den die Liganden zum aktiven Zentrum nehmen, ist noch weitgehend ungeklärt. Das aktive Zentrum liegt im Inneren des Enzyms und ist nur über verschiedene Zugangskanäle erreichbar, die jedoch meist durch Residuen des CYPs blockiert sind.

Wir untersuchen mit mathematischen Methoden für 7 Liganden den Weg an die Bindestelle im aktiven Zentrum. Da vollständige Simulationen zu zeitintensiv sind, verfolgen wir einen Netzwerkansatz. Die Bewegungen des Liganden können als stetige Trajektorie eines stochastischen Prozesses modelliert werden. Wir betrachten den infinitesimalen Generator des Prozesses. Dann diskretisieren wir den Zustandsraum gemäß einer Voronoi Zerlegung, die durch um das CYP verteilte Ligandenpositionen definiert ist. Nach Fackeldey, Lie und Weber [9] berechnen wir Übergangsraten zwischen einzelnen Ligandenpositionen. Dafür müssen wir nur die Interaktionsenergie zwischen CYP und Ligand der jeweiligen Position ermitteln. Dies können wir mit Simulationen erreichen. Mithilfe der Übergangsraten analysieren wir für jeden Liganden die ‘besten’ Pfade von außen zur Bindestelle.

Die resultierenden Pfade für die einzelnen Liganden weisen eine große Variabilität auf. Je nach Ligand scheinen andere Zugangskanäle günstig zu sein.

Contents

Introduction **1**

Approach **2**

1 Cytochrome P450 3A4 **3**

1.1 Cytochrome P450 3A4 Structure and Function 4

1.2 Ligands 8

2 Molecular Modeling **9**

2.1 Markov Processes 10

2.2 Discretization 18

2.3 Square Root Approximation 21

2.4 Refinement 24

2.5 Gromacs Molecular Dynamics 27

3 Simulation **28**

3.1 Simulation and Methods 28

3.2 Refinement - Example 37

3.3 Results 39

Conclusion **50**

A Appendix **54**

Introduction

In this thesis, we adress a biomolecular problem by mathematical methods. We analyze a specific enzyme, CYP 3A4, that belongs to a superfamily of enzymes which is responsible for the metabolization of drugs and xenobiotics. This enzyme metabolizes a high percentage of all marketed drugs and is therefore especially important to consider in drug design. It is crucial to understand the mechanisms of drug metabolism in the design of new drugs in order to avoid toxicity and guarantee effectiveness.

In drug design, it is important to know about drug metabolism, as the effect of the drug highly depends on the concentration in the body. With a too low concentration, which could be caused by a low dosage or a high rate of metabolization, the drug might be ineffective. But on the other hand, a low rate of metabolization would lead to an accumulation of the drug in the body which could have toxic side effects. It is crucial that drugs do not accumulate in the body, but get metabolized eventually and cleared out of the body.

The catalyzation and reaction itself is quite well understood [14, 26]. It takes place at the enzyme's active site, which is located inside the enzyme. The CYP 3A4 has a complex three dimensional tertiary structure. The active site is surrounded by secondary structure elements that obstruct the access from the enzyme's surface to the active centre. The substrate specifity due to the access mechanism to the active site is not solved yet. In this thesis, we analyze the binding pathways for seven different ligands.

We can mathematically model molecular systems by Markov processes. We then discretize the state space using a Voronoi tessellation with ligand positions as center points and approximate a transition rate matrix by simulation data. Simulations are carried out by the Gromacs software package [15]. While using simulation data is usually related to time consuming trajectory computation, we choose a novel approach by Weber, Fackeldey and Lie [9], the Square Root Approximation. It is a network approach, approximating transition rates between adjacent ligand positions by a formula that only requires the potential energy values of two adjacent positions. This could be a simple approach to analyze binding kinetics with less computational cost than other simulation approaches. The transition rate matrix can be evaluated in regard to the 'best' path from outside to inside. We additionally carry out a refinement procedure following [25].

Approach

We put the ligand in different positions around the CYP molecule. The idea is to create a transition network, each node representing a ligand position in a high dimensional space. With help of the interaction energy between ligand and enzyme for each position, we compute the transition rates between the ligand positions using the Square Root Approximation. We then analyze the transition rate matrix to identify the most probable pathway from the protein's surface to the active site. One interesting question is if the pathways differ considerably when using different ligands. And is there an predominant channel for all ligands?

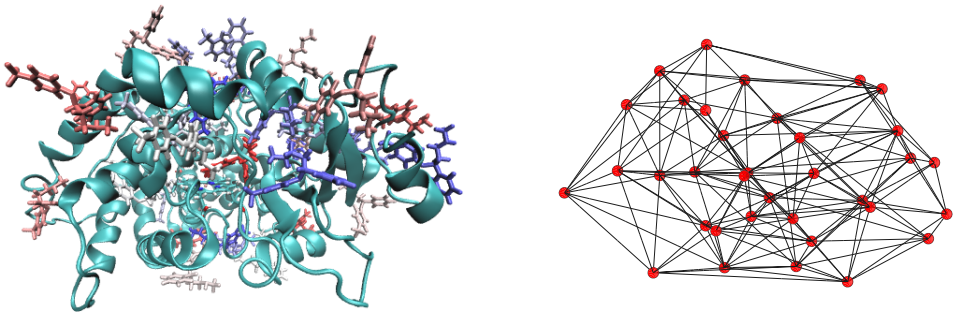

Figure 1: Example of a ligand position network. 35 fluoxetine positions and the CYP 3A4 enzyme. Right: A simple network representation of the positions

1 Cytochrome P450 3A4

The molecule we deal with is a protein, more specific, an enzyme. We will shortly introduce some biological terms which we will use to describe the structural features and mechanisms of CYP 3A4.

Proteins are built from amino acids. *Amino acids* are chemical components, consisting of an amino group, a carboxyl group and a side chain, which is specific for each amino acid. Amino acids form covalent bonds, called peptide bonds. The amino group of one amino acid reacts with the carboxyl group of a second amino acid, resulting in two connected amino acids. Two amino acids form a bipeptide. If multiple amino acids are involved, this results in amino acid chains. Short amino acid chains are called *peptides*, whereas chains of more than 100 amino acids are called *proteins*. The *primary structure* of a protein denotes its amino acid sequence. Hydrogen bonds are formed between the amino acids in the chain, yielding an energy minimizing fold. Depending on the amino acid sequence, different shapes are formed. There are *α-helices*, *β-sheets* and *random coils*. This is the *secondary structure*. Apart from that, there are bonds formed between the amino acid side chains. They cause the secondary structure elements to further fold into a complex three dimensional shape. This is called the *tertiary structure*. The function of a protein is highly affected by its fold. Proteins serve various purposes in organisms including building structures, transport proteins, signalling and catalyzing chemical reactions. One important feature is the catalyzation of reactions for other molecules. Proteins with this function are called *enzymes*. There is usually one region in the enzyme at which the substrates bind and the reaction takes place, which is called the *active site*. In some contexts the enzyme is also called *receptor* and the smaller molecule, that reacts catalyzed by the enzyme, is called *substrate* or *ligand*. The enzymatic reaction can involve a *cofactor* or *prosthetic group*. In the case of Cytochrome P450 enzymes, this is the *heme* group, which consists of a porphyrin ring and an iron ion, see Figure 2.

The substrate binds to the binding site next to the heme group and the Cytochrome P450 catalyzes an oxidation of the substrate. The product leaves the binding site and eventually after further modification it can be excreted from the body. In this case, *metabolism* describes the breakdown of the ligand molecule resulting in excretion.

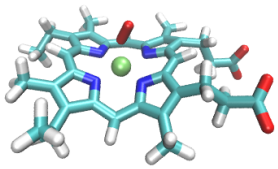

Figure 2: The heme group, iron ion in green

1.1 Cytochrome P450 3A4 Structure and Function

The biomolecule, we deal with in this thesis, is Cytochrome P450 3A4, abbreviated CYP 3A4. It belongs to a large family of enzymes with similar structural features. Hence, we shortly describe the structure and function of Cytochrome P450 enzymes in general and then concentrate on CYP 3A4.

Cytochrome P450s Cytochrome P450 is a superfamily of proteins, more precisely enzymes, responsible for the metabolism of drugs and xenobiotics, synthesis of steroids and fatty acid metabolism [28, 6]. They are found in highest concentration in the liver and gut, but they exist in almost all tissues [14]. The superfamily of Cytochrome P450 is divided in around 70 families. Members of the same family share at least 40 % sequence identity. The members of a family are further separated into subfamilies with at least 55 % sequence identity [26]. While the sequence differs considerably between different Cytochrome P450 families and subfamilies, the heme cofactor and their overall fold is conserved [29, 6, 11]. They share a similar secondary and tertiary structure, with the heme moiety bound to a helix in the core of the enzyme.

Drug Metabolism The main function of Cytochrome P450s is the drug and xenobiotic metabolism. In fact, most drugs are metabolized by a member of the cytochrome P450 family. They catalyze monooxygenation of their substrate molecules. The reaction takes place at the enzyme's active site, which is located next to the heme prosthetic group. Hydrophobic substrate molecules get oxidated, which eventually allow excretion. As described in the introduction, knowledge of the mechanism of drug metabolism is extremely important in drug design.

The chemical reaction that takes place at the active site of cytochrome is fairly well understood, see [14, 26]. However, the access routes the substrates take are less known. How do substrates enter the active site? Do they prefer different channels?

Cytochrome P450s differ in plasticity and each is specified for its own set of substrates. Substrate specificity not only depends on the active site interactions, but also on the possibility to reach the active site, which is usually enclosed by the residues lining possible access channels. It is assumed that channel gating plays a role in this [28, 6]. For more information on channel gating, see [7].

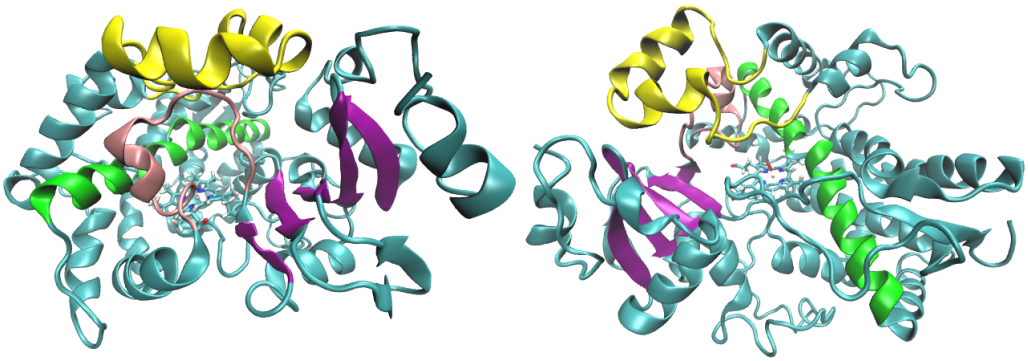

Figure 3: CYP 3A4 with colored secondary structure elements: B-C loop in pink, F-G loop in yellow, I helix in green, β1-sheet in purple; the heme group is in licorice representation

CYP 3A4 Structure In this thesis, we analyze Cytochrome P450 3A4 (short: CYP3A4), which metabolizes about 50 % of all marketed drugs [6]. CYP 3A4 is one of the most important and one of the most flexible enzymes, considering the broad range of substrates it metabolizes. Cytochrome P450 3A4 is composed of 468 amino acids and a heme prosthetic group as cofactor. The amino acid chain of CYP 3A4 arranges in α-helices, β-sheets and random coils as secondary structure elements. They form a spherical tertiary structure with the heme group buried inside the enzyme. The secondary structure elements and the overall fold are similar for each Cytochrome P450, but the amino acid sequences differ, leading to different residues lining the access channels. The heme moiety is located deep inside the enzyme, bound to the L helix and surrounded by two flexible loop regions.

We label the secondary structure elements according to [14] to adress certain regions of the CYP. Helices are labeled by letters A to L, while beta-sheets are labeled by numbers 1 to 5. Shorter helix elements are assigned to the next large helix element, labeled as 'prime'. Labels start at the N-terminus. A good schematic structure with labels can be found in [14]. Figure 3 shows two views of CYP 3A4 with some important secondary structure elements colored, including the I helix, the β1 sheet, as well as the flexible B-C and F-G loops. Two additional views with labeled structure elements are shown in the appendix.

Between helices and beta-sheets, there are random-coil-structures. These structures don't show an identifiable secondary structure. But therefore, these regions are more flexible and play an important role in substrate channel opening. The often mentioned B-C and F-G loops have a high proportion of random-coil-structures and are located close to the active site. Most of the identified pathways for substrate access or egress involve the B-C or F-G loop.

Possible substrates are only able to get from the protein surface to the active site if the cytochrome secondary structure shows a corresponding open conformation, where open means that the enzyme's residues allow for passing through to the active site by appropriate side chain and structural movements. It is suggested that ligand interactions induce the opening of channels. Certain residues of the CYP molecule act as gating residues, moving upon ligand interactions [6].

Possible Substrate Access Channels Although the process of metabolism itself is well understood, the access route that ligands take to the active site is generally unknown. Especially the range of different size substrates suggests a strong flexibility of secondary structure elements. To accomodate substrates like cyclosporine, wide channels have to open up. Possible substrate access and egress channels for Cytochrome P450s have been found and analyzed with different methods [6, 3, 28]. The results differ especially with the chosen ligand. In different studies, 7 to 11 channels for ligand access have been found in P450s [3, 6]. While some researchers put emphasis on the access channels or gating mechanisms, others were interested in the possible product egress routes or in the channel variance for different CYP family members. Methods included MD simulations, TMP, comparison of open and closed structures and Steered Molecular Dynamics.

Many channels were found by investigating open and closed crystal structures of CYP P450s. In [29], Zawaira et al. used crystal structures from the protein data bank, one complexed with a ligand and one structure in absence of a ligand. Usually, the complexed structure shows open tunnels for ligand access or egress. They analyzed differences in open and closed structures, with a special emphasis on the channel lining gating residues. They furthermore studied the side chain movement of certain residues lining the channels. These were identified as gating residues. In [6], pulling simulations were performed. The ligand of interest was put in the binding position and then pulled in the direction of already visible tunnel exits in the pdb structure (1TQN) and 18 different directions. The SMD simulations revealed channels 1, 2a, 2b, 2c, 2e, 3 and S. The tunnel lengths and radii were analyzed. From a wider opening they concluded a higher probability for ligand access or egress. Additionally the forces needed to pull the ligand were taken into account.
Cojocaru et al. suggest the additional channels 4,5,2f and 2ac [3], which were found by analyzing newly available crystal structures.

The channels are labelled 1,2,3,4,5 with channel 2 subdivided into pathway (pw) 2a to 2f as in [3]. We try to use many of the already revealed channels for the network construction. However, the exact channel position might slightly vary with the chosen CYP crystal structure. That is why we give the name of the channel and the channel entrance location as used in the simulations. We will use the names according to Figure 4 and 5 and their location

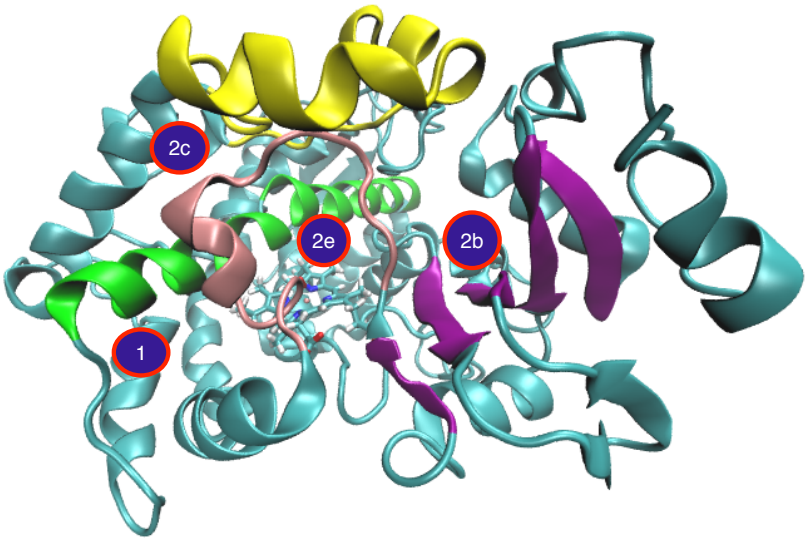

Figure 4: Frontview of CYP 3A4 with channels 1, 2b, 2c and 2e

should in general coincide with the ones found in literature. We describe the location of the channels as used in the result section to avoid confusion. The location can also be seen in Figure 4 (frontview) and 5 (backview). Two views are included to include all channels.

1 Between the C,H and I helix,

2a between the F-G loop, the B-C loop and the β1 sheet

2b Between the β1 sheet and the B-C loop

2c next to the B-C loop and the G-helix

2d next to the A helix

2e through the B-C loop

2f between the F-G loop and the β4 sheet / C-terminus loop

3 through the F-G loop, also includes channel **4** from [3], which is located closer to the F' and G' helices

S between the F-G loop, the β4 sheet and the I helix, was suggested to be the solvent channel [3]

5 close to channel **2a**, next to the C-terminus loop

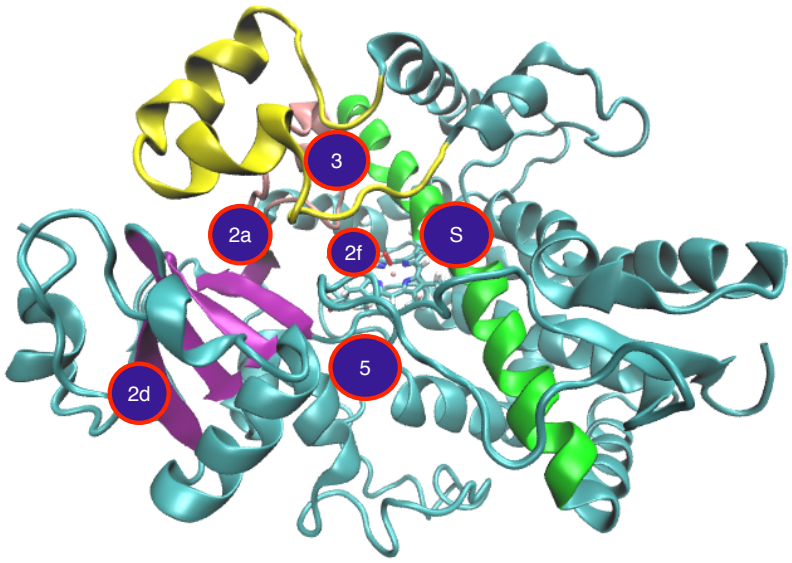

Figure 5: Backview of CYP 3A4 with channels 2a, 2d, 3 and S

1.2 Ligands

We study the molecular binding kinetics of CYP 3A4 on the basis of seven different substrates. The substrates differ by size and flexibility. We analyze their differences in the transition network and we try to find out which is their preferred pathway into the binding pocket.

The substrates are pharmaceuticals used for varying therapeutic purposes. The size ranges from 26 to 76 atoms. See more details about the ligands in the Results Section 3.3.

Ligand
Nicotine
Bromperidol
Fluoxetine
Nifedipine
Sufentanil
Terfenadine
Voriconazole

2 Molecular Modeling

In this section, we deal with the mathematical modeling of a molecular system. We consider a molecule consisting of N atoms. The N atoms are held together by covalent or ionic bonds. However, the molecule is flexible, bonds can rotate and stretch, which results in different three-dimensional structures. Especially for large proteins, the conformation is a crucial factor for their biological functionality. A conformation of a molecule can be described by a vector which contains the three dimensional coordinates for each atom, which is a vector in $\mathbb{R}^{3N}$. Additionally we can consider the momentum of each atom. That yields the state space $\mathbb{R}^{6N} = \Omega \times \Sigma$, where $\Omega = \mathbb{R}^{3N}$ is the position space and $\Sigma = \mathbb{R}^{3N}$ is the momentum space. Typically, one is mainly interested in the position coordinates. Most physical properties arise from the conformation only [21]. Consequently, we represent a molecules state by a vector in $\Omega = \mathbb{R}^{3N}$ and the molecular dynamics by a trajectory in the $3N$-dimensional state space.

In the full state space $\Omega \times \Sigma$, the evolution of a system can be modeled by the Hamiltonian $H(q, p)$, defined by Hamilton's Equations

$$\frac{\partial p}{\partial t} = -\frac{\partial H}{\partial q} \qquad \frac{\partial q}{\partial t} = \frac{\partial H}{\partial p}, \tag{1}$$

where q denotes the position vector and p the momentum vector. We remark that the Hamiltonian gives a deterministic evolution of the system. Given an initial state (q_0, p_0), the trajectory is determined by (1). $H(q, p)$ denotes the energy of the system at position (q, p) in phase space.

$$H(q, p) = U(q) + K(p), \tag{2}$$

where $U(q)$ denotes the potential energy and $K(p)$ the kinetic energy. This modeling corresponds to an energy conserving trajectory. While this is true for isolated systems, when modeling real world systems, we usually have systems that allow energy exchange with its surroundings.

In statistical physics, three different types of systems are considered. There are isolated, closed and open systems.

- In isolated systems, neither an exchange of matter nor energy with the surroundings is allowed.
- In a closed system, there is no transfer of matter (i.e. no chemical reactions or particle exchange) but a transfer of energy, usually in form of heat with the surroundings.
- In an open system, there is transfer of matter and energy.

The systems we are interested in are closed systems. There is no exchange of matter, but a transfer of energy. Imagine a closed system in contact with a heat bath. The heat bath has an infinitely large heat capacity (which in reality is usually justified by the heat bath being much larger than the system). That means it can exchange heat with the system, while keeping its temperature constant.

As a consequence, our system in contact with the heat bath can have different energy levels, because it exchanges energy in form of heat with the surroundings. To model this, a canonical ensemble (or NVT ensemble) is used. We imagine copies of the system with different initial conditions, each evolving independently from the others. They each have a constant number of particles N, a constant volume V and a constant temperature T, but they may differ in their total energy levels. When the system is in thermal equilibrium with the heat bath, the probability to find the system in a certain state is proportional to the state's energy. The system's states are distributed according to the Boltzmann distribution which is dependent on the temperature

$$\pi(x) = \frac{1}{Z} e^{-\beta V(x)}, \tag{3}$$

where β is $\frac{1}{k_B T}$, with the boltzmann constant k_B and the absolute temperature T. $V(x)$ is the energy of state x and Z is the normalization constant, i.e. $Z = \int_X \pi(x) dx$. We see that it is more likely to find states with a low energy than with a high energy. However, with increasing temperature, the states become more equally distributed.

To get a more realistic model of a biomolecule, it has to be taken into account that a molecule in the body can exchange energy with its surroundings. It is usually surrounded by an aqueous solution in the body cells at a certain temperature around 310 K. It is in contact with the solvent and can exchange heat while the body temperature remains constant.

Arriving at the notion of a canonical ensemble, we assume the process can be modeled by a stochastic process.

2.1 Markov Processes

We concentrate on Markov Processes for Molecular Modeling. It is a special class of stochastic processes, which we want to use to model the canonical ensemble of the CYP-ligand complexes. Markov Processes can be applied to many real-world problems. They are simpler to analyze because ot their defining Markov property. This Markov property says that the transition probability from the present state to the next state depends on the present

state but not on the history of the process, i.e. the states before the present state. The markov property is a reasonable assumption for a process modeling ensemble kinetics, as explained in [24]. Since we look at the process from a statistical point of view and we are not interested in one single realization but in the ensemble kinetics of the canonical ensemble, the Markov property can be assumed.

We define a transfer operator that propagates densities or membership functions on a continuous state space, which can be identified with the position space $\mathbb{R}^{3N}$ of a system with N atoms. The usual approach is then to project the transfer operator on a lower dimensional space, using the Galerkin discretization and an appropriate discretization of the state space. Subsequently, the projected transfer operator is approximated by simulation data. We will use a slightly different approach by making use of the infinitesimal generator of the process. We can discretize the generator und approximate it by the Square Root Approximation [9].

We start to define Markov Processes on finite state spaces and a discrete index set in order to introduce important properties of Markov Processes and the transition probability matrix. After that, we can easily extend the Markov process to a continuous index set, introducing the transition rate matrix. Then, we define Markov Processes on a continuous state space and in continuous time with the respective notion of the transfer operator. First, we start with some basic definitions concerning stochastic processes.

Let $(\Omega^*, \mathcal{F}, P)$ be a probability space; that consists of the sample space Ω^*, the σ-field $\mathcal{F}$, containing all possible events and the probability measure P on $\mathcal{F}$, which assigns a probability to each event. Let $(E, \mathcal{E})$ be a measurable space.

Definition 1 (Random Variable) *A $(\mathcal{F}, \mathcal{E})$-measurable function $X : \Omega^* \to E$ is called a random variable.*

The measurability assures that each element from $\mathcal{E}$ has a probability assigned that X takes that value, i.e. the preimage of A lies in $\mathcal{F}$ and then $P(X \in A) = P(X^{-1}(A))$ for $A \in \mathcal{E}$ is well-defined.

Instead of single random variables, we consider a stochastic process evolving in time. It is a random variable for each element from the index set (usually $\mathbb{N}$ or $\mathbb{R}$).

Definition 2 (stochastic Process) *A stochastic process $(X_t)_{t\in\mathcal{I}}$ with index set $\mathcal{I}$ and state space E is a collection of random variables*

$$X_t : \Omega^* \to E, \qquad \forall t \in \mathcal{I}$$

For a fixed $\omega \in \Omega^*$, $X_t(\omega) : \mathcal{I} \to E$ is a trajectory or a realization of the process.

To define the Markov property, we need to define the conditional propbability first.

Definition 3 *Given $A, B \in \mathcal{F}$ events in $\mathcal{F}$, with $P(B) > 0$, then the probability of A conditioned on B is defined as*

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

We note, that he conditional probability $P(\cdot \mid A)$ for $A \in \mathcal{F}$ with $P(A) > 0$ is again a probability measure on $(\Omega^*, \mathcal{F}, P)$.

Discrete Case: Markov Chain The simplest class of Markov Processes to analyze are those which are discrete in time and space. We consider the index set $I = \mathbb{N}_0$ and a finite set of states E, i.e. $|E| = n \in \mathbb{N}$. They are called Markov Chains.

Definition 4 *A stochastic process $(X_n)_{n \in \mathbb{N}_0}$ on a finite state space E is called a Markov process, if for all $x_{n+1}, x_n, \ldots, x_0 \in E$ with $P(X_n = x_n, \ldots, X_0 = x_0) > 0$, it holds*

$$P(X_{n+1} = x_{n+1} \mid X_n = x_n, \ldots, X_0 = x_0) = P(X_{n+1} = x_{n+1} \mid X_n = x_n).$$

This property is called the Markov property. It denotes a "memorylessness" of the process. Conditional on the present state x_n, the propagation is independent of the visited states before n. The probability for the transition to the next state is only dependent on the present state.

Furthermore, if the probability for a transition from state i to state j does not depend on n, i.e.

$$P(X_{n+1} = j \mid X_n = i) = P(X_1 = j \mid X_0 = i) = p_{ij},$$

then the process is called time homogeneous. It means, being in a state i, the one step transition probability to a state j is the same, no matter at what point of time. We concentrate on time homogeneous processes throughout this thesis.

We can define the transition probability matrix P.

$$P(i, j) := P(X_{n+1} = j \mid X_n = i) \qquad \text{for all } i, j = 1, \ldots, n \tag{4}$$

The entry $P(i,j)$ of P contains the transition probability to get from state i to state j in one timestep. This is well defined because of the assumed time homogeneity.

Definition 5 *A square matrix P is called (row-)stochastic, if*

$$P(i,j) \geq 0 \qquad \forall i,j \in E$$

$$\sum_{j \in E} P(i,j) = 1 \qquad \forall i \in E$$

The transition matrix P is a stochastic matrix, what becomes clear when we recall that $P(\cdot|A)$ is a probability measure.

An initial distribution is given by the probability vector $v^{(0)}$, with $v_i^{(0)} = P(X_0 = i)$.

Definition 6 *A probability vector is a row vector $v \in \mathbb{R}^n$ with*

$$v_i \geq 0, \qquad i = 1, \ldots, n$$

$$\sum_{i=1}^{n} v_i = 1$$

An entry v_i can be interpreted as the probability to be in state i. The transition matrix P propagates a probability distribution. If v^0 denotes the initial probability distribution for X_0, then $v^1 = Pv^0$ is the probability distribution after one timestep.

$$v^2 = Pv^1 = PPv^0 = P^2v^0$$

Vice versa, a stochastic matrix $P \in \mathbb{R}^{n \times n}$ and an initial distribution $v^0 \in \mathbb{R}^n$ define a time homogeneous Markov chain on a state space with n. A Markov Chain on a finite state space can be represented by a graph, where the nodes represent the possible states and the directed edges represent the transition probability.

We have $Pe = e$ with $e = (1, 1, \ldots, 1)^T$, hence, e is an eigenvector of P corresponding to the eigenvalue 1. Since P is a stochastic matrix, it follows that $|\lambda| \leq 1$ for all eigenvalues λ of P.

A time-continuous Markov Process or Markovian Jump Process In the last paragraph we had a discrete time Markov Process on a finite state space. Now, we will keep the finite state space, but extend the index set of the process to $T = \mathbb{R}_{\geq 0}$. That means the

process jumps between the states, but the jump could happen at any time. That is why it is also called a Markovian Jump process. Since the process is memoryless, the wait time for the next jump is exponentially distributed.

The Markov property follows straightforward from the discrete-time case.

Definition 7 *$(X_t)_{t\in T}$ is called a Markov Process, if for $i_0, \dots, i_{n+1} \in E$ and $t_0 < t_1 < \dots < t_{n+1} \in T$ with $P(X_{t_n} = i_n, X_{t_n-1} = i_{n-1}, \dots, X_{t_0} = i_0) > 0$ it holds*

$$P(X_{t_n+1} = i_{n+1} \mid X_{t_n} = i_n, X_{t_n-1} = i_{n-1}, \dots, X_{t_0} = i_0) = P(X_{t_n+1} = i_{n+1} \mid X_{t_n} = i_n)$$

Assuming time-homogenity, it holds for $s, t \in T$

$$P(X_{t+s} = i_{n+1} \mid X_s = i_n) = P(X_{t+0} = i_{n+1} \mid X_0 = i_n) = p_{i_n i_{n+1}}(t).$$

It denotes the probability to be in state i_{n+1} after a time lag of t, when being in state i_n. Hence, the transition probability does not depend on the point in time. But it depends on the time lag t. (The probability that the process $(X_t)_{t\in T}$ goes from state i to j in a second might be different from that in a minute or two seconds). When we fix the time lag, we have a transition probability matrix as in the discrete time case. But each time lag corresponds to a different one. Instead for one probability transition matrix P, we arrive at a family $(P(t))_{t\geq 0}$ to characterize the Markov Process.

$(P(t))_{t\geq 0}$, given by a time homogeneous Markov Process, satisfies the Chapman-Kolmogorov equation for all $s, t \geq 0$.

$$P(t+s) = P(t)P(s) \tag{5}$$

Due to (5), $(P(t))_{t\geq 0}$ defines a semigroup of transition matrices. A semigroup of transition matrices $(P(t))_{t\geq 0}$ and an initial distribution $v^0 \in \mathbb{R}^n$ define a continuous time Markov Process.

Instead for the full semigroup $(P(t))_{t\geq 0}$, there is a another notion to analyze continuous time Markov Processes. Assume that the limit

$$q(i,j) := \lim_{t\to 0+} \frac{P(t)(i,j)}{t} \tag{6}$$

exists for $i \neq j$. Then, $q(i,j)$ denotes the rate for a transition from i to j. Furthermore, we assume

$$\sum_{j\in E, i\neq j} q(i,j) < \infty \qquad \text{for each } i \in E \tag{7}$$

Then, we set

$$q(i,i) := -\sum_{j\in E, i\neq j} q(i,j). \tag{8}$$

We usually require two further properties of $(P(t))_{t\geq 0}$ in order to define the transition rate matrix Q.

$$(i) \quad P(0) = I \tag{9}$$

$$(ii) \quad \lim_{t\to 0+} P(t)(i,j) = \delta_{ij} \tag{10}$$

These requirements seem quite natural. They assure that there are no jumps in zero time and that the probability to jump in a small time interval is also small. The second property is called *standard*. We can define the transition rate matrix under the aforementioned conditions. The standard assumption assures differentiability.

Definition 8 (transition rate matrix) *The semigroup $(P(t))_{t\geq 0}$, satisfying the standard assumptions, defines the matrix Q by*

$$Q = \lim_{t\to 0+} \frac{P(t) - I}{t} \tag{11}$$

We interpret $Q(i,j)$ as the instantaneous rate from state i to state j ($i \neq j$). The diagonal element $Q(i,i)$ can be interpreted as the exponential distribution parameter for the holding time in state i.

Under further conditions on the matrix Q which are fulfilled due to the finite state space, we can generate the semigroup of transition matrices from it. Q is called the infinitesimal Generator of the semigroup $(P(t))_{t\geq 0}$ and it holds

$$\exp(Qt) = P(t).$$

We can calculate $P(t)$ for every t by Q and Q is not dependent on the time-step anymore. Therefore, it is often more convenient to study Q instead of P. The row sum of Q is 0 and

$$Q(i,i) = -\sum_{j\neq i, j=1}^{n} Q(i,j),$$

which follows from (11), the finite state space and $P(t)(i,i) = 1 - \sum_{j\neq i} P(t)(i,j)$.

A Markov Process on a Continuous State Space Now, we define a Markov process on a continuous state space, i.e. $E = \Omega = \mathbb{R}^{3N}$, in continuous time. We can't define it via a semigroup of transition probability matrices, since the state space is not countable anymore. The index set is $T = \mathbb{R}_{\geq 0}$.

The natural filtration $(\mathcal{F}_t)_{t\in T}$ of the process $(X_t)_{t\in T}$ on the state space $(\Omega, \mathcal{E})$, is the collection of σ-algebras defined by

$$\mathcal{F}_t = \sigma\left\{X_s^{-1}(A), \quad s \leq t, A \in \mathcal{E}\right\} \qquad \text{for } t \in T.$$

The σ-algebra $\mathcal{F}_s$ contains all the information up to time s. It holds $\mathcal{F}_s \subset \mathcal{F}_t$ for all $s, t \in T$ with $s \leq t$, thus $(\mathcal{F}_t)_{t\in\mathbb{R}}$ is a filtration.

Definition 9 (Markov property) *A stochastic process $(X_t)_{t\in T}$ adapted to the filtration $(\mathcal{F}_t)_{t\in T}$ is called Markov Process, if*

$$P(X_t \in A \mid \mathcal{F}_s) = P(X_t \in A \mid X_s)$$

for all $s, t \in T$ with $s < t$ and for all $A \in \mathcal{A}$.

To define a transfer operator, which replaces the transition probability matrix in the continuous case, we introduce the notion of a transition function.

Definition 10 *Let $(\Omega, \mathcal{A})$ and $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ two measurable spaces. A transition function is a function from $T \times \Omega \times \mathcal{A}$ to $[0, 1]$, which satisfies*

(1) $p(t, x, \cdot) : \mathcal{A} \to [0, 1]$ *is a probability measure for every $x \in \Omega$ and $t \in T$*

(2) $p(t, \cdot, A)$ *is a measurable function for every $A \in \mathcal{A}$ and $t \in T$*

(3) $p(0, x, X \setminus \{x\}) = 0$ *for every $x \in \Omega$*

(4) $p(t + s, x, A) = \int_\Omega p(t, x, dy)p(s, y, A)$ *for all $x \in \Omega$, $s, t \in T$, $A \in \mathcal{A}$*

.

It assigns a probability to each starting point x, that the process is in set $A \in \mathcal{A}$ after a time lag t. $(X_t)_{t\in T}$ is a Markov Process, if

$$p(t, x, A) = P(X_t \in A \mid X_0 = x)$$

for every $t \in T$ and every $A \in \mathcal{A}$.

μ is an invariant measure for the process $(X_t)_{t\in T}$, if

$$\mu(A) = \int p(t, x, A)\mu(dx) \qquad \text{for all } A \in \mathcal{A} \text{ and } t \in T.$$

The natural extension of the transition matrix for finite state space Markov processes is the transfer operator. Instead of acting on probability vectors, it acts on density functions or membership functions in the appropriate $L_\mu^{1,2,\infty}(\Omega)$ space [19].

We can define the transfer operator (here: forward transfer operator) via the stochastic transition function.

Definition 11 *The forward transfer operator $\mathcal{T}^t$ for a fixed timelag t is defined by*

$$\int_A \mathcal{T}^t u(y)\mu(dy) = \int_\Omega p(t, x, A)u(x)\mu(dx),$$

where $p(t, x, A)$ is a transition function.

Remark: the defined transfer operator corresponds to a forward transfer operator. There is also the notion of the backward transfer operator $\mathcal{P}^t$, which is also called transfer operator and is the adjoint operator of $\mathcal{T}^t$.

Definition 12 *The backward transfer operator $\mathcal{P}^t$ for a fixed timelag t is defined by*

$$\mathcal{P}^t u(x) = \mathbb{E}(u(X_t)|X_0 = x) = \int u(y)p(t, x, dy),$$

where $p(t, x, A)$ is a transition function.

The forward transfer operator propagates a density $u \in L_\mu^1$ by time t, hence $\mathcal{T}^t u(x)$ is the density that results by applying the dynamics of the process X_t on the density $u(x)$. The backward transfer operator acts on membership functions $v \in L_\mu^\infty$.

The transfer operator defined by the stochastic transition function in (10) has certain properties. For both defined transfer operators $\mathbb{1}$ is an eigenfunction. While $\mathcal{P}^t\mathbb{1} = \mathbb{1}$ follows directly from the definition of $\mathcal{P}^t$, we need that μ is an invariant measure for $\mathcal{T}^t\mathbb{1} = \mathbb{1}$. The Chapman-Kolmogorov equation holds for the transfer operator as defined before, due to (4) in definition 10.

The transfer operator $(\mathcal{P}^t)_{t\in T}$ defines the operator $\mathcal{Q}$ by

$$\mathcal{Q} = \lim_{t\to 0} \frac{\mathcal{P}^t - \mathcal{I}}{t}, \tag{12}$$

where $\mathcal{I}$ denotes the identity operator. Since the realizations of the molecular process $(X_t)_{t\in T}$ are continuous trajectories, the limit in (12) exists. And furthermore, since $(\mathcal{P}^t)_{t\in T}$ fulfills the Chapman-Kolmogorov equation, it is the infinitesimal generator of $(\mathcal{P}^t)_{t\in T}$ [24]. We will use an infinitesimal generator approximation to characterize the dynamics of our molecular system. First, we have to discretize the state space. We will discuss that in the next section.

2.2 Discretization

Molecular ensemble dynamics can be represented by a Markov Process on the continuous state space Ω. The operators $\mathcal{P}^t$ and $\mathcal{Q}$, that govern the dynamics, act on the infinite dimensional function spaces. The operators are usually unknown for real processes. Therefore, we aim at finding a discretization of the continuous state space into a finite number of states. The resulting projected transfer operator on the finite dimensional space can be approximated by simulation data.

Ansatz Space Instead of a infinite-dimensional function space, we consider the finite subspace $\mathcal{D} = \text{span}\{\phi_1, \ldots, \phi_n\}$ with appropriate functions $\{\phi_1, \ldots, \phi_n\}$. The simplest approach is to consider characteristic functions on sets that decompose the state space into n non-overlapping sets. But there are advantages of other methods [22].

In general, the functions $\phi_1, \ldots, \phi_n$ should be non-negative and fulfill the partition of unity condition.

$$\begin{aligned} \phi_i(x) &\geq 0, \qquad \text{for all } i = 1, \ldots, n \\ \sum_i \phi_i(x) &= 1, \qquad \text{for all } x \in \Omega. \end{aligned}$$

For characteristic functions of a full state space decompostion, this is obviously fulfilled. In the set-based approach, we partition the state space Ω into n distinct sets.

$$\Omega = \cup C_i, \qquad i = 1, \ldots, n$$

$$C_i \cap C_j = \emptyset, \qquad for\, i \neq j$$

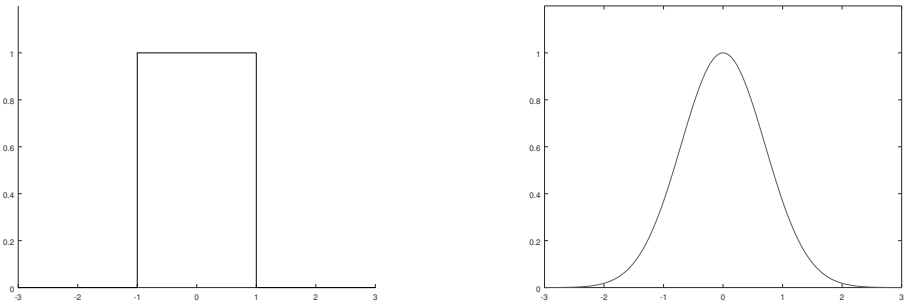

Figure 6: Left: Set based basis function. Right: Fuzzy set basis function in form of (13)

Then, the characteristic functions $\{\phi_1, \ldots, \phi_n\} = \{\mathbb{1}_{C_1}, \ldots, \mathbb{1}_{C_n}\}$ form the basis of $\mathcal{D}$. The sets could be mesh-based, for example. The problem is then the growing number of sets for high-dimensional state spaces. In [22, 21, 16], it is explained why it can be advantageous to move away from the set-based approach to a fuzzy-set approach using overlapping basis functions.

Let $\phi_1, \ldots, \phi_n$ be radial monotonic decreasing functions centered at base points $q_1, \ldots, q_n$ in the state space. Assume they fulfill the conditions of positivity and the partition of unity. There are several favorable properties these functions should have to assure good discretization, i.e. the discretization error should be small. This is explained in detail in [21]. The basis functions used are usually radial basis functions, centered at base points $q_i \in \Omega$ and monotonic decreasing with the distance to q_i. A good choice of functions would be for example

$$\phi_i(q) = \frac{e^{-\alpha_i d(q,q_i)^2}}{\sum_{j=1}^{n} e^{-\alpha_j d(q,q_j)^2}}, \tag{13}$$

with a shape parameter α_i and the distance function d. The parameters α_i determine the steepness of the function. A large α_i yields a quickly decreasing function ϕ_i. Ideally, α_i should be chosen to guarantee an appropriate amount of overlap between the different basis functions. That is why we choose α_i to be proportional to $1/d_{min}(q_i)^2$, where $d_{min}(q_i)$ is the minimal distance from q_i to the next base point q_j. It can be seen as a kind of set-based discretization but with overlapping sets and we can only assign a grade of membership to a certain set. In contrast to classic sets, the fuzzy sets are not strictly separated. A position can belong to more than one fuzzy-set, especially in the transition regions. This is more realistic, since positions in a transition region between molecular conformations can not be clearly associated with one of them. The degree of overlapping in the case of functions as defined in (13) is determined by the choice of the shape parameter α.

With either of these approaches, we can define the Galerkin projection. Therefore, we introduce the Boltzman weighted scalar product for L^1 or L^2 functions.

$$\langle u, v\rangle_\mu = \int u(x)v(x)\mu(dx),$$

with the Boltmann function μ.

Definition 13 *Let $v \in L^2(\mu)$. The Galerkin projection of v is defined as*

$$\mathcal{G}(v) = \sum_{i=1}^{n} \frac{\langle \phi_i, v\rangle_\mu}{\langle \phi_i, \mathbb{1}\rangle_\mu}\phi_i.$$

We can also project the transfer operator $\mathcal{P}^t$ to act on the subspace $\mathcal{D} = \text{span}\{\phi_1, \ldots, \phi_n\}$.

Definition 14 *The Galerkin projection of the transfer operator $\mathcal{P}^t$ is given by*

$$P_{ij} = \frac{\langle \phi_i, \mathcal{P}^t\phi_j\rangle_\mu}{\langle \phi_i, \mathbb{1}\rangle_\mu} = \frac{\langle \mathcal{T}^t\phi_i, \phi_j\rangle_\mu}{\langle \phi_i, \mathbb{1}\rangle_\mu}. \tag{14}$$

This gives us a matrix representation P on the finite subspace $\mathcal{D}$. P is a stochastic matrix. It holds $P_{ij} \geq 0$ since $\phi_i \geq 0$ and $\mathcal{P}^t\phi_i \geq 0$ and the row sum of P is 1

$$\sum_j P_{ij} = \frac{\left\langle \phi_i, \sum_j \mathcal{P}^t\phi_j \right\rangle_\mu}{\langle \phi_i, \mathbb{1}\rangle_\mu} = \frac{\langle \phi_i, \mathcal{P}^t\mathbb{1}\rangle_\mu}{\langle \phi_i, \mathbb{1}\rangle_\mu} = \frac{\langle \phi_i, \mathbb{1}\rangle_\mu}{\langle \phi_i, \mathbb{1}\rangle_\mu} = 1.$$

In case of a set-based approach, $P(i, j)$ from (14) can be interpreted as the conditional probability for the dynamics starting in set C_i to reach set C_j after time t.

$$P_{ij} = \frac{\left\langle \mathcal{T}^t\mathbb{1}_{C_i}, \mathbb{1}_{C_j}\right\rangle_\mu}{\langle \mathbb{1}_{C_i}, \mathbb{1}\rangle_\mu} = p(t, C_i, C_j)$$

To approximate the projected transfer operator in practice, one can simply count transitions from trajectories starting in set C_i to set C_j. We apply the same Galerkin discretization to the infinitesimal Generator $\mathcal{Q}$ of $(\mathcal{P}^t)_{t\in T}$

$$Q_{ij} = \frac{\langle \phi_i, Q\phi_j\rangle_\mu}{\langle \phi_i, \phi_j\rangle_\mu}.$$

The discretized version of $\mathcal{P}^t$ or $\mathcal{Q}$ can then be approximated by simulation data or other methods. We cannot directly compute it, since the transfer operator or the infinitesimal

generator are usually not explicitly known (and even if so, the integrals were hard to compute). In [25], it is shown how to approximate P_{ij} or Q_{ij} by simulation data. Sampling positions and many trajectories of length t are required. Since the system we want to approximate is high-dimensional, the calculation of as many trajectories would come with high computational costs.

Instead we approximate Q by a novel approach by Fackeldey, Weber and Lie, called the Square Root Approximation [9]. This is based on using a state's energy to estimate transition rates between ligand states. The advantage of this novel approach is the reduced computational cost. It does not rely on short- or long-term trajectory calculations, but requires simple energy calculations instead!

2.3 Square Root Approximation

To approximate the Galerkin Discretization as in [25], we would need a lot of simulation data, which is not feasible. Instead we use another discretization of $\mathcal{Q}$ based on a Voronoi tessellation for which we don't need to simulate trajectories. We start with a certain number of base points, i.e. ligand positions around the CYP 3A4 molecule, defining Voronoi regions and then we compute potential energy values for each ligand's position. For that purpose, only energy minimization is required instead of molecular dynamics simulations. We use Theorem 4 from [24] or Theorem 1 from [9].

As a discretization basis we use a Voronoi tessellation of the position space $\Omega = \cup_{i=1}^{n} \Omega_i$ with Voronoi regions Ω_i. We start with n base points $q_1, \dots, q_n$ in the position space Ω. The Voronoi regions Ω_i are defined by

$$\Omega_i := \{q \in \Omega \mid d(q, q_i) \leqslant d(q, q_j), \quad \forall i \neq j\},$$

with a distance function d. The positions that belong to the Voronoi region Ω_i are those that have a shorter (or equal) distance to q_i than to any other base point q_j.

Theorem 1 (Square Root Approximation, Theorem 1 from [9]) *Given a Voronoi tessellation of the position space ($\Omega = \cup_{i=1}^{n} \Omega_i$) and the transfer operator $\mathcal{P}^t$, such that its infintesimal Generator $\mathcal{Q}$ exists. $P(\tau)$ is the discretization of $\mathcal{P}^\tau$ based on the Voronoi tessellation, then $Q := \frac{\partial}{\partial \tau} P(\tau)(i,j)$ is given by*

$$Q(i,j) = \int_{\Omega_i \cap \Omega_j} z(q) \pi_i(q) dS(q) \tag{15}$$

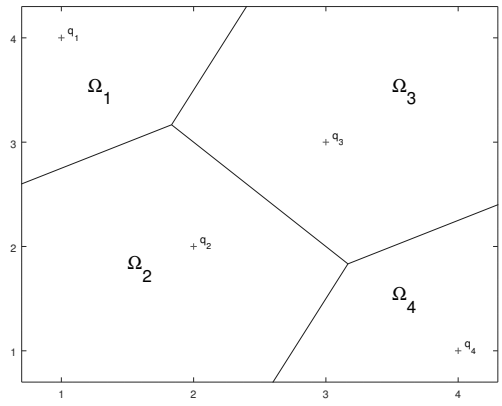

Figure 7: Voronoi tesselation example in 2 dimensions

for $i \neq j$, where $z(q)$ is the flux through q, π_i is the restricted Boltzman density to Ω_i and dS is the surface measure on $\Omega_i \cap \Omega_j$.

A proof can be found in Weber[24]. If we interpret $Q(i,j)$ as the transition rate from Voronoi region Ω_i to Ω_j, it is given by the flux from Ω_i to Ω_j through the surface $\Omega_i \cap \Omega_j$ weighted with the restricted Boltzman measure on Ω_i. We see that $Q(i,j) = 0$, if the Voronoi regions Ω_i and Ω_j are not adjacent, since then $\Omega_i \cap \Omega_j = \emptyset$. Now, we find an approximation of the transition rate matrix (15), since we can't calculate it directly. The following derivation of an approximation for $Q(i,j)$ is based on [9] and [24]. It is an approximation in two steps.

First, we rewrite (15) to $Q(i,j) = s_{ij}\langle z \rangle_{ij}/w_i$

$$Q(i,j) = \int_{\Omega_i \cap \Omega_j} z(q)\pi_i(q)dS(q) \tag{16}$$

$$= \int_{\Omega_i \cap \Omega_j} z(q)\frac{s_{ij}}{s_{ij}}\pi_i(q)dS(q) \tag{17}$$

$$= \int_{\Omega_i \cap \Omega_j} z(q)\frac{s_{ij}}{s_{ij}}\frac{\pi_q(q)}{w_i}dS(q) \tag{18}$$

$$= s_{ij}\langle z \rangle_{ij}/w_i \tag{19}$$

where s_{ij} is Boltzman weight of the surface $\Omega_i \cap \Omega_j$, w_i is the Boltzman weight of Ω_i and

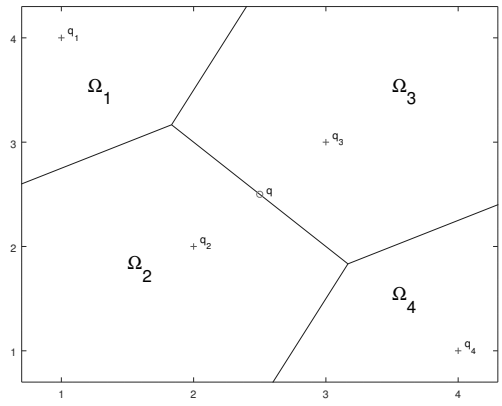

Figure 8: Voronoi tessellation, the Boltzmann weight of every q on the surface $\Omega_2 \cap \Omega_3$ is approximated by $V(q) \approx \frac{1}{2}(V(q_2) + V(q_3))$

$\langle z \rangle_{ij}$ is the flux per unit area from Ω_i to Ω_j across $\Omega_i \cap \Omega_j$.

$$s_{ij} = \int_{\Omega_i \cap \Omega_j} \pi_q(q) dS(q) \tag{20}$$

$$w_i = \int_{\Omega} 1_{\Omega_i}(q) \pi_q(q) dq \tag{21}$$

$$\langle z \rangle_{ij} = \int_{\Omega_i \cap \Omega_j} z(q) \frac{\pi_q(q)}{\int_{\Omega_i \cap \Omega_j} \pi_q(\hat{q}) dS(\hat{q})} dS(q) \tag{22}$$

If we now assume a constant flux $\hat{z}$ instead of $\langle z \rangle_{ij}$, we can simply approximate (15) by

$$Q'(i,j) = \hat{z} s_{ij} / w_i. \tag{23}$$

The transition rate from set i to set j is given by the flux per unit area times the Boltzman weight of the surface between the Voronoi regions i and j, divided by the Boltzman weight of Voronoi region i.

The second approximation we make is the approximation of the w_i and s_{ij}. The real Boltzman weights are unknown, but we can obtain the potential energy values of single points by simulation and then approximate the Boltzman weights for Voronoi regions. By construction of the Voronoi regions, we have one base point in the center of each region. We use this base point as an approximation for the Boltzman weight. For the Boltzman weight of the surface, we use a simple linear approximation. By construction, $q \in \Omega_i \cap \Omega_j$ has the same distance to base point q_i than to q_j.

We approximate the potential energy function by

$$V(q) \approx \frac{1}{2}(V(q_j) + V(q_i)). \tag{24}$$

That yields

$$s_{ij}(q) = \frac{1}{Z} e^{-\beta V(q)} \tag{25}$$

$$\approx \frac{1}{Z} e^{-\beta \frac{1}{2}(V(q_i)+V(q_j))} \tag{26}$$

$$= \frac{1}{Z} (e^{-\beta V(q_i)} e^{-\beta V(q_j)})^{\frac{1}{2}} \tag{27}$$

$$= \frac{1}{Z} \sqrt{e^{-\beta V(q_i)} e^{-\beta V(q_j)}}. \tag{28}$$

The Boltzmann weight w_i of region Ω_i with center point q_i is simply approximated by

$$w_i \approx \frac{1}{Z} e^{-\beta V(q_i)}. \tag{29}$$

Combining (23), (25) and (29), we arrive at

$$Q''(i,j) = z \sqrt{\frac{e^{-\beta V(q_j)}}{e^{-\beta V(q_i)}}} G(i,j), \tag{30}$$

for $i \neq j$, where G is the adjacency matrix with $G(i,j) = 1$, if $\Omega_i \cap \Omega_j \neq \emptyset$. For a transition rate matrix on a finite state space, it holds $Q(i,i) = -\sum_{j,j\neq i} Q(i,j)$, see section 2.1, thus we set

$$Q''(i,i) = -\sum_{j,j\neq i} Q(i,j).$$

2.4 Refinement

How do we know if we haven chosen enough ligand positions, i.e. base points q_i, to sufficiently approximate the dynamics of the ligand? The processes dynamics are governed by the transfer operator $\mathcal{P}^t$ or $\mathcal{Q}$, respectively. There are fast and slow dynamics. We are interested in the path the ligand takes from the surface to the binding site buried inside the Cytochrome P450 3A4 and the corresponding conformational changes. On a short timescale, the bondlengths and angles between the atoms of a molecule will oscillate around some value. The conformational changes and movement into the binding pocket will happen at a larger timescale.

These slow processes can be associated with the dominant eigenvalues of $\mathcal{P}^t$ or Q and

their eigenfunctions [25, 18]. Therefore, we consider the eigenvalue problem $\mathcal{P}^t u = \lambda u$ or $\mathcal{Q}v = \lambda v$ respectively. The discretization error is closely related to the discretization error of the dominant eigenfunctions [25, 17].

We can see that by considering metastable conformations. A conformation is an almost invariant subset of the state space. We can characterize that by $\mathcal{P}^t\chi \approx \chi$. It shows that this concept is related to the eigenvectors close to 1. We say χ is metastable. It means the process stays a long time in this conformation and transitions between different conformations are only rare events. Thus, the dynamics between metastable sets are slow processes. For more details about clustering the state space into metastable subsets, see [17, 24]. We will use the fact that the discretization can be considered good if the discretized eigenvectors are good approximations of the eigenfunctions corresponding to the dominant eigenvalues [17, 18].

The largest eigenvalue (in modulus) of $\mathcal{P}^t$ is 1 [25]. If the associated process is reversible, i.e.

$$\mu(x)p(t,x,y) = \mu(y)p(t,y,x) \qquad \text{for all } x, y \in \Omega$$

where μ denotes the invariant measure, then $\mathcal{P}^t$ is self-adjoint and all its eigenvalues are real-valued, such that the eigenvalues can be ordered $1 = \lambda_1 > \lambda_2 > \ldots > \lambda_n$ [25]. The dominant eigenvalues are those that are close to 1. Assume, n_c denotes the number of dominant eigenvalues. $\mathcal{P}^t$ and $\mathcal{Q}$ have the same eigenfunctions and their eigenvalues are related through

$$\lambda_{P^t} = e^{-t\lambda_Q}.$$

Thus, we can solve either of the eigenvalue problems. We just have to keep in mind that dominant eigenvalues of $\mathcal{P}^t$ are close to 1, while those of $\mathcal{Q}$ are close to 0. Since we will work with the infinitesimal generator approximation Q, we will concentrate on the respective eigenvalue problem

$$\mathcal{Q}v = \lambda v,$$

with eigenfunctions v. Because the dynamics $\mathcal{Q}$ are unknown, following [25], we project it on the basis function space

$$\mathcal{D} = \text{span}\{\phi_1, \ldots, \phi_n\}, \qquad v = \sum a_i \phi_i.$$

The eigenvalue problem for the basis function space then rewrites

$$\sum (\mathcal{Q}a_i\phi_i - \lambda a_i \phi_i) = 0.$$

Multiplying by $\phi_k / \langle \phi_k, 1 \rangle_\mu$, using the scalar product $\langle \cdot, \cdot \rangle_\mu$, this yields the generalized eigen-

value problem

$$Qv = \lambda Mv \tag{31}$$

or in matrix notation

$$QV = MV\Lambda. \tag{32}$$

With the matrices Q and M with entries

$$Q(k,j) = \frac{\langle \phi_j, \mathcal{Q}\phi_k \rangle_\mu}{\langle \phi_k, \mathbb{1} \rangle_\mu} \tag{33}$$

and

$$M(k,j) = \frac{\langle \phi_k, \phi_j \rangle_\mu}{\langle \phi_k, \mathbb{1} \rangle_\mu}. \tag{34}$$

If eigenvectors of Q are also eigenvectors of M, then they solve the generalized eigenvalue problem. Assume V is the matrix containing the dominant eigenvectors of Q in its columns. Then, it holds

$$QV = V\Lambda_Q,$$

with $\Lambda_Q = \mathrm{diag}(\lambda_1, \ldots, \lambda_{n_C})$ is the diagonal matrix of dominant eigenvalues of Q. If it holds

$$MV = V\Lambda_M,$$

i.e. the dominant eigenvectors of Q form an M-invariant subspace, then V also solves the generalized eigenvalue problem (32)

$$QV = V\Lambda_Q = V\Lambda_Q\Lambda_M\Lambda_M^{-1} = MV\Lambda_Q\Lambda_M^{-1} = MV\Lambda,$$

with $\Lambda = \Lambda_Q\Lambda_M^{-1}$. The Λ_i commute, because Λ_i are diagonal matrices.

If Λ_M exists, such that $MV = V\Lambda_M$, we can solve the generalized eigenvalue problem. It is the best possible approximation of the eigenvalue problem $\mathcal{Q}u = \lambda u$ on the subspace $\mathcal{D}$ [25]. In practice, we evaluate the angle between the subspaces spanned by the columns of MV and V, which should be small. Thereby, we see whether the dominant eigenvectors of Q are close to being eigenvectors of M.

Hence, the value of the angle can serve as an indicator for the quality of the discretization. We can furthermore identify the basis functions that need refinement by

$$r_k = \sum_{i=1}^{n_c} |\left(Mv_i - \Pi^i(Mv_i)\right)_k|, \tag{35}$$

where v_i are the eigenvectors, i.e. the i-th. column of V, and Π^i is the orthogonal projection

on v_i. A large r_k indicates that we need refinement in region Ω_k with center position q_k. Positions that have only a small overlap in M, i.e. that are relatively far away (in comparison to other positions), but have a high energy difference, resulting in a high transition rate from the high energy position to the low energy position, are identified to need refinement. Why we need more base points in transition regions, becomes clear, if we think about approximation quality of eigenvectors.

2.5 Gromacs Molecular Dynamics

We use Gromacs 5.1.4 for all simulations in this thesis. It is a software for molecular dynamics simulation. It has to be provided with an input coordinates file, which specifies the x, y and z coordinates of each atom. Furthermore a topology has to be supplied. Then the program computes the forces that arise due to the current positions of different atom types. Integrating Newtons Equations of motion, new positions and velocities for all atoms are computed.

$$F = ma$$

The chosen timestep highly affects the precision of the resulting trajectory.

In order to calculate the forces, certain parameters have to be determined. There are non-bonded interactions and bonded interactions. For example, the atoms of a molecule are connected by chemical bonds. These bonds can differ in length. Due to attracting and repulsive forces the bond length will oscillate around a certain value. That value is different depending on the atoms involved. Analoguously, the angles fluctuate, where angles are formed by three or four bonded atoms. These dynamics are modeled as an oscillation, thus a spring equation, with the parameters 'spring constant' and 'equilibrium value'. The parameters differ with the involved atom types and are specified in the force field. They are indentified by experiments. We use the AMBER 99sb force field in our simulations. The non-bonded interactions (which is the largest part of the computational cost) are theoretically calculated for every pair of atoms. To reduce computational cost, there are different methods invoked as a cut-off distance, neighbor lists or PME. The non-bonded interactions include the Coulomb potential, that arises due to varying charges, and the Lennard-Jones potential, modeling the repulsive and attractive forces between atoms (without charges). All bonded (bond lengths, bond angles, dihedral angles) and non-bonded (Coulomb- and Lennard-Jones-Potential) interactions sum up to a potential. The resulting force can be computed from the potential function.

To produce a canonical ensemble, a thermostat can be included during the simulation. These thermostats assure the correct average temperature and ideally a correct distribution

of states.

As a first step in a simulation, an energy minimization is performed. It finds the next energy minimum (using the gradient of the potential function). This is done to avoid large forces in the simulation, which will result in a system blow up. For the energy minimization, the steepest gradients or conjugate gradient algorithms are used. At least a combination of both work fine in most cases. When the system is energy minimized, the next step is equilibration. That means, the correct average temperature and pressure are set. After careful preparation, the actual MD simulation can take place.

Furthermore, non-equilibrium dynamics can be performed with Gromacs. During pulling simulations, a virtual particle is attached to a chosen pull group. A spring connects pull group and particle. Then the particle moves with a user-defined rate and force, pulling the pull group across the simulation box. We use pulling simulations to open ligand access channels of the Cytochrome P450 3A4. While pulling the ligand away from the binding pocket in different directions, the CYP atoms adjust and rearrange due to the forces that arise. The forces during this simulation are biased since we apply an external force. That is why it is important that we perform an energy minimization before we compute the potential of the pulling positions.

3 Simulation

For the Square Root Approximation (30) of the discretized generator, we need to know the potential energy of different ligand positions. We use Gromacs' energy minimization to obtain the required values.

All simulations were carried out using the MD simulation software package Gromacs 5.1.4 [15, 1, 12]. The molecular images are created with VMD [8] [20]. Crystal structures are from the PDB databank (www.rcsb.org) [2] and the DrugBank database (www.drugbank.ca) [27]. Throughout simulations, we use a tip4p water model, the AMBER99sb force field [13] and the v-rescale thermostat.

3.1 Simulation and Methods

Idea The idea is to put the ligand molecule in different 'valid' positions around the CYP 3A4 and calculate the approximated transition rates between the positions. First, we perform a MD simulation of CYP 3A4 without a ligand. The aim is to obtain the CYP in

different conformations. As described in section 1.1, there are many possible access channels for the ligands. But they are not already 'open' in the crystal structure of CYP 3A4. We hope the ligand access channels reveal during the simulation. The structures with open channels are then used for the LES (Ligand Excluded Surface) computation. For each ligand and different conformations and orientations of this ligand, we compute the valid positions on the surface of the CYP molecule[1]. We pick a number of ligand positions, set up simulation systems for them and perform an energy minimization. We will see that additional pulling simulations are required to obtain valid positions in the ligand access channels. The ligand's dynamics can be modeled by a Markov Process. We approximate the discretized Q according to section 2.3, inserting the interaction energy between ligand and CYP for every position.We interpret the $Q(i,j)$ as transition rate from ligand position i to ligand position j. Additionally, we compute the matrix M from (34) to identify positions which need refinement.

Ligand Excluded Surface The ligand excluded surface (LES) is a function defining the ligand accessible positions at the surface of a receptor molecule [10]. It is individual to each ligand. We use it to generate a picking set of 'valid' ligand positions, where 'valid' means, that they don't intersect (the atoms, molecules). Around each ligand and receptor atom, we can put a sphere representing its van der waals radius. The van der waals radius marks the distance of closest apporach of another atom. Usually the solvent excluded surface is computed, where the solvent is represented by a sphere of radius 1.4 Å. In the case of larger, flexible ligand molecules, a more exact approach has to be considered, since there might be ligand conformations fitting in cavities of the receptor molecule, while a bounding sphere might not fit. A sphere might not be a good representation of the ligands shape.

The algorithm works grid-based. It places the ligand in different conformations and rotations at each gridpoint and decides if this is a valid position. Hence, it takes a number of ligand conformations, a number of orientations and a grid spacing as input. If a state is valid can be expressed by the function

$$l(k,T,R) = \begin{cases} 1, & \text{if } ||p_i^r - (Rp_{jk}^l + T)|| \geq r_i^r + r_j^l \quad \forall i = 1,\ldots,n\, j = 1,\ldots,m \\ 0, & \text{else} \end{cases} \tag{36}$$

p_i^r are the atom positions of the receptor molecule, p_j^l are those of the ligand molecule. r_i^r and r_j^l are the atomic radii, k is the conformation of the ligand molecule, T the translation and R the rotation of the ligand molecule. A ligand's state can be described by k, T and R and is valid if $l(k,T,R) = 1$.

[1] thanks to Norbert for computing the ligand valid positions

To reduce computational costs, the algorithm works in two phases. In the first phase, the ligand is represented by a sphere containing all ligand atoms (the largest of all conformations). Where this sphere can be placed without intersecting with the receptors atoms, the ligand can be positioned in all orientations and conformations. These are ligand accessible positions. Then, another sphere is tested, which is inscribed in the ligand. Where this sphere can't be positioned, the ligand can't be placed in any conformation or orientation. These positions are not accessible by the ligand. In phase two, the remaining grid positions are tested for validity, trying each conformation and orientation, according to (36). We use these ligand positions with in the respective conformation and orientation as picking set.

Ligand Simulation To provide different input conformations of the ligand molecule for the LES algorithm in order to compute the valid positions, we perform ligand simulations. We obtain a structure file of each ligand from www.drugbank.ca [27]. The topologies are created by acpype [4]. For each of the seven ligands, we perform a 1.2 ns MD simulation. The preparation includes solvation, energy minimization, NVT and NPT equilibration. Then the simulation takes place at 500 K to sufficiently sample the state space. Recalling the Boltzmann density (3), we note that the states are more equally distributed on the state space, if the chosen temperature is high. The resulting trajectory is fitted and we pick 10 to 15 different conformations, invoking the picking algorithm described in 3.1. We use the euclidian distance in $\mathbb{R}^{3N_l}$, where N_l is the number of ligand atoms of the coresponding ligand, as distance function. The chosen conformations are written to pdb files and are used for the 'valid' ligand position calculation as described in the previous paragraph.

CYP Simulation We try to obtain different CYP conformations with channels opened up, in order to put the ligands in consecutive positions from surface to binding pocket and finally analyze the network for the path the ligand takes to the binding position. To obtain a large conformational variety, similarly to the ligand simulations, we perform the simulation of Cytochrome P450 3A4 at a high temperature. The CYP 3A4 coordinates file is 1TQN from the PDB databank (www.rcsb.org). Missing residues were added[2] as well as the topology which uses the AMBER99sb force field. In absence of a ligand, we prepare the system for CYP 3A4. We use a dodecahedral box shape, solvate, neutralize with CL ions, energy minimize and equilibrate the system with a temperature of 500 K. Thereafter, we carry out a 1.5 ns MD simulation. From the resulting trajectory we choose a few promising timeframes to be analyzed by the LES algorithm. Since nicotine is the smallest ligand molecule in the selection, the valid ligand positions are at first computed for this molecule. It turns out that the conformational variety of CYP 3A4 during the simulation is not as

[2]thanks to Vedat for providing the CYP 3A4 pdb file and topology.

large as hoped. There is no consecutive pathway from surface to binding pocket of even the smallest of our ligand molecules possible. The valid ligand position computation shows that only one of the channels, namely pw 2e, nearly connects outside and inside. Additionally pw 2a, pw 2b and pw 2f partly open up during simulation, but ligand positioning in these channels is still not possible. Since we know from literature that more channels exist [3], we decide to perform additional pulling simulations to obtain valid ligand positions in the still closed channels. Neverthelsess, we choose a CYP conformation from simulation to compute the ligand positions at the surface and the inside of the CYP molecule.

Picking Algorithm To pick different ligand conformations and pick the ligand positions for the transition network a picking algorithm is used. We try to obtain a large variety of positions or conformations respectively. Therefore, we consider the distances of points in the picking set to the already picked points. We choose the next point with the maximal distance to the closest point from the chosen points. The algorithm is from [25]. The set of all possible positions is denoted as S and the set of already picked points as $Q*$.

1. Choose a random r (or $r = 1$, when used to create the network, in order to include the binding position as first position)
2. Compute distances $d(s, q)$ for all $s \in S$ and $q \in Q*$
3. Pick the next point from S with the maximal minimum distance $\max_{s \in S} \min_{q \in Q*} d(s, q)$
4. Stop, when the number m of required positions is reached.

The results differ depending on the chosen distance function.

Pulling Simulations To open up more access channels to the active centre, we perform pulling simulations. We put the ligand in the suggested binding position in the binding pocket. Then we apply external forces in the simulation to pull the ligand slowly across the CYP molecule towards the CYPs surface. We are interested in the other way round, i.e. the path from the surface towards the binding site, but nevertheless, the simulation will induce conformational changes upon the CYP molecule, opening up the access channels. Since we want to start pulling with the ligand in its binding mode, we first have to identify the ligands binding position and orientation.

Identifying the Binding Mode The binding site is already identified. It is located next to the heme group. However, the orientation of each is ligand is individual. For each of

the seven ligands, we prepare 60 systems. We take for each ligand one input structure and compute 60 orientations of the ligand. We use the CYP 3A4 pdb file (1TQN) with known binding site coordinates. Then, we complex each of the ligands with this CYP structure. We set up the systems, prepare and simulate. Then, for each complex, we observe the potential energy throughout simulation. Taking into account the non-bonded potential energy terms, i.e. the sum of the Coulomb and the Lennard-Jones potential, between ligand and CYP and ligand and solvent, we compute the average value over simulation time. The system that results in the minimal average energy is the suggested binding mode.

Pulling Proceeding from the binding mode of each ligand, we choose directions towards the tunnel entrances according to Figure 4 and 5. We use the identified channels [3, 6] that are visible in the respective receptor-ligand-complex. Although there are a few residues blocking the access, we can still see the location of the tunnel entrances. With the Gromacs pull code, a rate of $0.03\,\mathrm{nm\,ps^{-1}}$ and a force of $1000\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$, we perform 9 pulling simulations for each ligand. In general, the Cytochrome channels open up during pulling the ligand out. While in some directions pulling is fast and easy, in other directions the Cytochrome molecule needs more time to adjust and change its conformation to allow the ligand to pass through. Some pulling simulations are repeated with slightly different directions, since the ligand did not take the requested channel in the first simulation. Figure 9 shows five positions of sufentanil being pulled out via channel 2c.

Picking and Energy Minimization - Network We collect all valid ligand positions computed by the LES algorithm and the positions from the pulling simulations. We consider the ligand position of each timeframe from the pulling simulation as one position. It is important to keep it complexed with the respective CYP structure, since it undergoes conformational changes and the ligand position is only a valid position with this CYP structure. The positions are represented by the $3N_l$ position coordinates, 3 coordinates for each atom, where N_l is the number of ligand atoms. Additionally we can compute the three dimensional center of each position, if we are only interested in spatial movement.

Let the N atom position vectors be denoted by $(x, y, z)_i$ for $i = 1, \ldots, N$. Then the center of the molecule is defined as

$$(x_c, y_c, z_c) = \frac{1}{N} \sum_{i=1}^{N} (x, y, z)_i.$$

From the set of all positions (picking set) we want to choose evenly spaced positions around the CYP molecule. We therefore calculate the three dimensional centers and use the picking

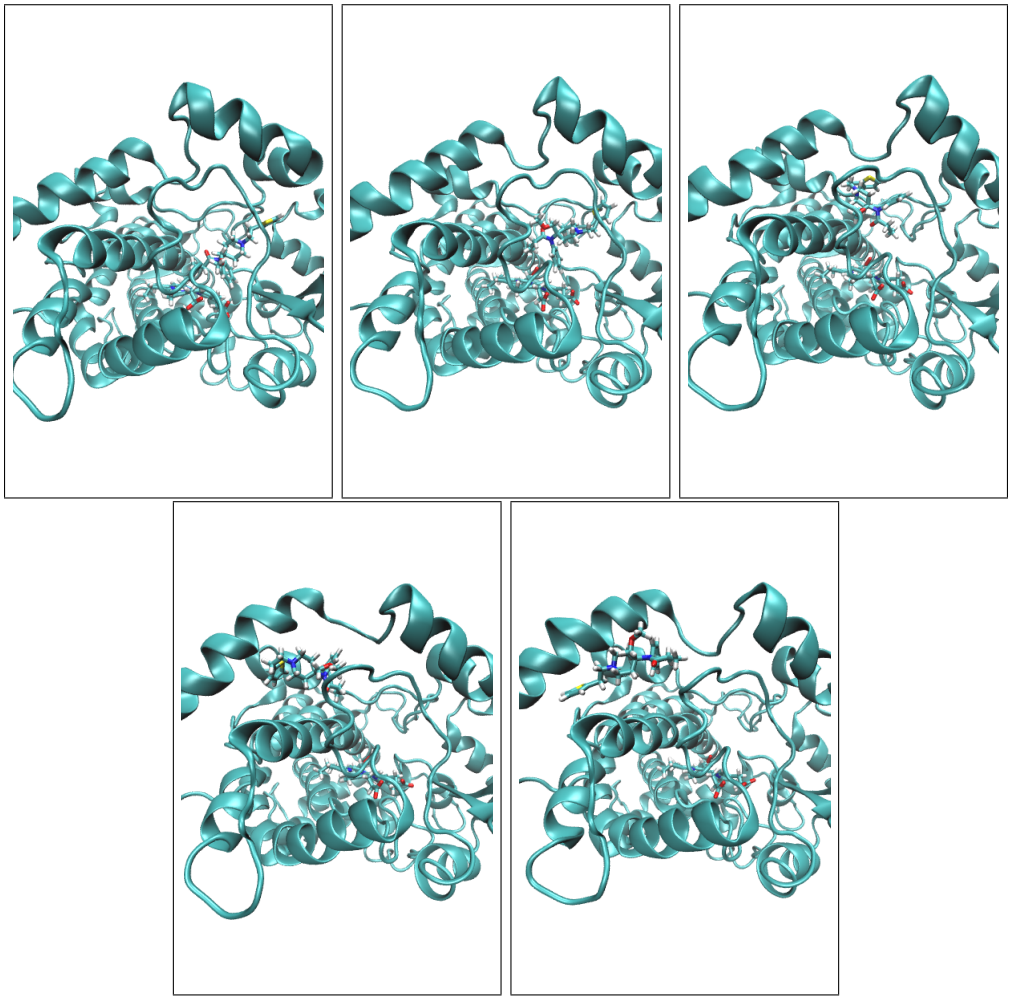

Figure 9: Series of pulling snapshots: Sufentanil in channel 2c. Note the conformational changes in the B-C and F-G loop

algorithm to pick 100 to 150 positions. Since we want to inlcude the binding positions in our network, we set $r = 1$, which corresponds to the binding mode in our picking set, instead of a random number. Each picked position is a state (or node) in the network. Now, we need the interaction energy between receptor and ligand to compute the transition rates between the positions. For each picked ligand position, we complex it with the CYP 3A4 structure that was used to calculate the valid positions or leave it complexed with CYP in case of the pulling positions. Then, we perform an energy minimization. For the positions obtained from the pulling simulation, we perform an energy minimization in vacuum first. Otherwise, the system tends to blow up due to the large forces. Often, we use a combination of the steepest descent and the conjugate gradient algorithm. After the minimization, the complexes are fitted with the heme group as reference. The energy is extracted from each postition.

Computation of Q and M With the coordinates and energies of the positions, we build the network. The transition rates between positions can be computed. We recall the approximation formula for $Q(ij)$ from section 2.3 for $i \neq j$

$$Q(i,j) \approx \sqrt{\exp(-\beta V(q_j))/\exp(-\beta V(q_i))}\hat{z}N(i,j). \tag{37}$$

Hence, the transition rate depends on the potential energy values and the adjacency relation. The rate is only positive, if two positions are adjacent. We use a delaunay triangulation or equivalently a Voronoi tessellation to define adjacency. The three dimensional centers of the ligand positions define the base points for the Voronoi tessellation. A Voronoi region Ω_i for base point q_i is defined as

$$\Omega_i = \{x \in \Omega : d(x, q_i) \leq d(x, q_j) \text{ for all } j \neq i\}, \tag{38}$$

with the euclidian distance in $\mathbb{R}^3$ as distance function. An example for 2 dimensions is shown in Figure 7.

Two base points q_i and q_j are adjacent, if their Voronoi regions Ω_i and Ω_j are adjacent, which is the case, if they share a boundary, i.e. $\Omega_i \cap \Omega_j \neq \emptyset$. We have the adjacency matrix N with entries

$$N(i,j) = \begin{cases} 1, & \text{if } q_i \text{ and } q_j \text{ are adjacent} \\ 0, & \text{else} \end{cases}.$$

$V(q)$ denotes the potential energy function of q. We only take the interaction energy between receptor and ligand into account. It consists of the Lennard-Jones potential and the Coulomb

potential.

$$V(q) = LJ(q, cyp) + Coulomb(q, cyp)$$

Gromacs energies are in $\mathrm{kJ\,mol^{-1}}$ The parameter β denotes $1/(k_B T)$, but since $V(q)$ is in $\mathrm{kJ\,mol^{-1}}$, we use $\beta = 1/(RT)$ with the gas constant R in $\mathrm{kJ\,K^{-1}\,mol^{-1}}$ and the absolute temperature T. The temperature is 310 K, body temperature. This yields a value of $\beta \approx 0.388$ We calculate the off-diagonal entries of Q according to (37) and set

$$Q(i,i) := -\sum_{j=1, j\neq i}^{m} Q(i,j) \qquad \text{for all } i = 1, \ldots, m.$$

We interpret as $Q(i,j)$ for $i \neq j$ as transition rate between adjacent positions. (It is 0 for non-adjacent positions).

Aditionally, we compute the 'overlapping' matrix M to evaluate where we may need more ligand positions to appropriately represent the ligands dynamics. We follow section 2.4 and [25] and compare the dominant eigenraum of Q with that of M. Evaluating (35), identifies the positions q_i that need refinement. More positions are required in energy transition regions, see section 2.4. The basis functions are radial monotonic decreasing functions, as suggested in [21].

$$\phi_i(x) = \exp(-\alpha_i d(x, q_i)^2)\gamma(x)^{-1}, \tag{39}$$

where $\gamma(x)^{-1}$ is the normalization constant, i.e. $\gamma(x) = \sum_{i=1}^{m} \phi_i(x)$. The α_i parameter for basis function ϕ_i centered at q_i should be proportional to $1/d_{min}(i)^2$, where $d_{min}(i)$ is the minimal distance from q_i to any other base point q_j. The matrix M contains values, representing the overlap of fuzzy sets. $M(i,j)$ is the overlap of ϕ_i and ϕ_j with radial basis functions ϕ as in (39).

Fackeldey and Weber describe in [25] how integrals of the form $\langle \phi_i, u \rangle_\mu$ can be approximated. With the Boltzmann measure μ, we see that

$$\phi_i(x)\mu(x) = e^{-\alpha_i d(x,q_i)^2 - \beta V(x)}\gamma(x)^{-1} Z^{-1}.$$

This yields a new measure μ_i with energy function $V(x) + \alpha_i/\beta d(x, q_i)^2$. We can then approximate the integral as a sum using evaluation points that are sampled according to μ_i. We take only q_i as sampling point and approximate M from (39) as follows

$$M(i,j) \approx \phi_j(q_i) = e^{-\alpha_j d(q_i, q_j)^2}\gamma(q_i)^{-1}, \tag{40}$$

with $\alpha_j = 1/d_{min}(j)^2$, d the euclidian distance in $\mathbb{R}^{3N_l}$, where N_l is the number of ligand atoms and $\gamma = \sum_{i=1}^{m} \phi_i(q_i)$. Hence, M is a stochastic matrix.

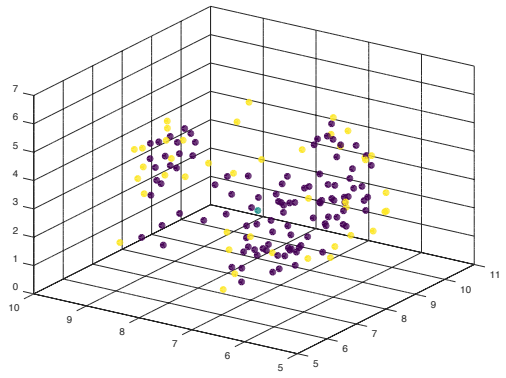

Figure 10: Fluoxetine positions with the starting positions colored in yellow and the binding position colored in green. Each position is represented by its three dimensional center.

Analysis of Q and Refinement We can add positions according to the refinement requirements. Finally, we interprete Q and try to find the 'best' path to the binding pocket of CYP 3A4. There are different algorithms to analyze Q, for example the PCCA+ [5] or the Dijkstra algorithm. The question is, which path, i.e. sequence of positions, will the ligand take to get from the surface of the CYP 3A4 to the binding pocket. We will use a simple greedy algorithm, which picks the next position according to the highest rate. That results in a sequence of positions. Due to the fact, that we don't allow to go back to an already visited position, the binding position is reached after a finite number of steps. The starting positions we are interested in are those located at the surface of the Cytochrome molecule. We take those points that form the convex hull of the position point set. In Figure 10, we see the ligand positions of fluoxetine with the starting positions colored in yellow.

For each surface position, we look at the path returned by the greedy algorithm. The last few positions before reaching the binding position are those that indicate which channel refers to the path. In most cases, we can assign one access channel to the given position sequence. Sometimes, there are jumps from one channel to another, or even jumps from the surface of the Cytochrome molecule directly to the binding position. This can happen due to the fact that we defined adjacency through Voronoi regions and the limitations of CYP residues inbetween adjacent positions are not included in the adjacency calculation. To solve that problem, the Voronoi regions could be calculated not with the euclidian distance function in $\mathbb{R}^3$ but with a surface distance function, measuring the distance to the next position along the Ligand Excluded surface. One problem arising with this function is that different CYP conformations are used with the positions. The pulling and to a small amount also the energy minimization induce conformational changes on the CYP molecule. For simplicity, we used the euclidian distance function for the adjacency calculation. Detailed results for each ligand can be found in the results section.

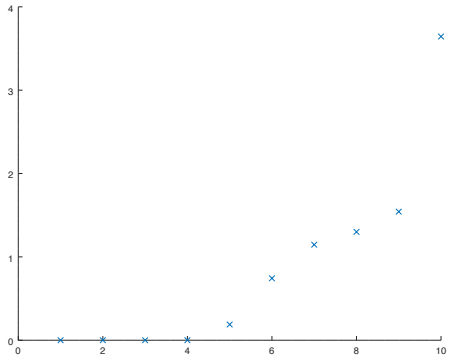

Figure 11: The 10 largest eigenvalues (in modulus) of the Q matrix for nifedipine

3.2 Refinement - Example

We look at the network for the ligand nifedipine and evaluate it according to section 2.4. Do we need more positions in certain regions to appropriately model the ligand dynamics? We picked 150 positions from our picking set. As described in the last section, we carried out the energy minimization and computation of Q and M. Then, we calculate the dominant eigenvectors of Q. Therefore, we need to identify the corresponding dominant eigenvalues. We notice 5 eigenvalues close to zero and a gap between the 5th and 6th eigenvalue, see Figure 11. According to section 2.4, we take the corresponding 5 largest eigenvectors, see also [18, 16]. Define V as the matrix containing the eigenvectors corresponding to the largest 5 eigenvalues as columns. Next, we calculate the angle between the subspaces spanned by MV and V. If it is 0, it means the dominant eigenvectors of Q are eigenvectors of M as well and we are finished. The angle is 0.39739 in radians, which corresponds to $22,77°$. The value has only limited significance, since it scales with the chosen α_i values in the computation of M. However, we will see if we can improve the value by adding new positions to the network.

We identify energy transition regions by (35). The positions with the highest values are identified to be refined. In this example, the first position to refine is position 150. The reason becomes clear if we look at the Q matrix. Row 150 of Q shows a maximum in column 38. There is a high energy difference and thus a high rate from position 150 to 38, while being not close and having only a small overlap in M. The two positions represented by their three dimensional centers are shown in Figure 12. We try to refine and add another position close to position 150 (ideally between position 150 and position 38). But the picking set does not include any position in-between, see Figure 13. There are cyp residues and no valid positions. Hence, it is not possible to refine that energy transition region. Another approach for this problem is to define another adjacency relation, since the two positions are

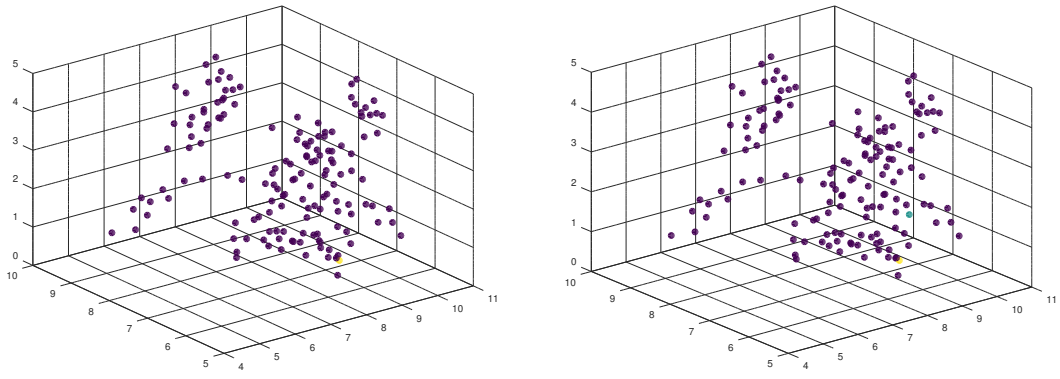

Figure 12: Positions represented by their three-dimensional centers and the position to refine in yellow. The position with a low energy, i.e. the position to which the yellow position has a high transition rate, is colored green in the right picture.

relatively far away and essentially separated by the CYP molecule. But due to the Voronoi definition with the euclidian distance, they are adjacent. It would be good to define a new adjacency relation based on the CYP's surface which excludes positions seperated by the molecule from being adjacent.

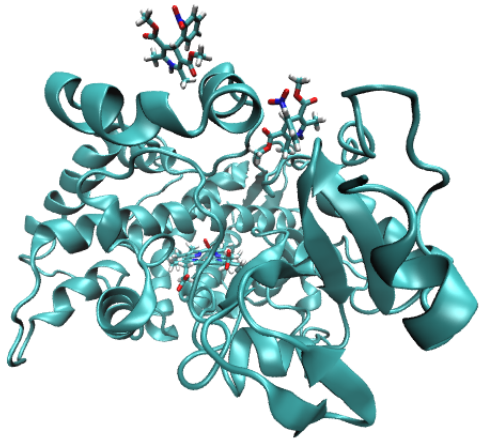

Figure 13: CYP 3A4 with the two adjacent ligand positions shown. They have a high energy difference and therefore refinement is required in the transition region. However, in-between the two positions, there is no valid ligand positioning possible due to the CYP's helix element.

3.3 Results

For each ligand, we computed a matrix Q. We interpret its entries as transition rates in a network, where the ligand positions are represented by nodes. We analyze it by a simple greedy algorithm. The starting positions are those that form the convex hull of all positions. The end position is the suggested binding position. The algorithm returns a path, i.e. a sequence of positions, from each starting position to the binding position.

We analyze these sequences by having a closer look at the positions preceding the binding position. Those are the positions that determine which pathway the ligand takes into the binding pocket. We can associate the sequence of positions with a ligand access channel. In most cases, the positions determine the channel quite clearly. However, in some of the sequences, there are jumps between channels. Then we can't associate a single channel with it. There are also cases where direct jumps from the CYP's surface to the binding pocket happen. Having identified the channels, we compare the number of starting positions that pass through the respective channel. If the sequence jumps between two channels, we indicate that by giving both channels, first the one that seems more probable due to the position sequence, and mark it with an asterisk. Channel 2f/S are not considered separately. We collect the ligand specific data in a table. Additionally, there are images included showing the prevailing pathways. They depict the sequence of the ligand positions close to the binding position corresponding to the dominant channels. The Cytochrome is in New Cartoon representation and the Heme group and ligand molecules are in Licorice representation. The ligand molecules in these images are colored in different shades to better distinguish the individual ligand positions. We included different views for better visibility of the channel entrances. Furthermore, we give the length of the position sequence. A long sequence could indicate that the transition to the binding position might not be energetically favorable.

Remark: Sometimes, there seem to be steric clashes between the CYP molecule and the ligand molecules. This is due to the fact that the CYP conformation changes during the pulling simulation (and to a small extent also in the energy minimization), hence, the ligand molecules are positioned relative to slightly different CYP conformations. In order to make the picture better recognizable, we show only one of the CYP conformations in the picture. So, some ligand molecules overlap with CYP atoms, but this is presumably not the CYP conformation belonging to that ligand molecule.

Remark 2: In the licorice representation of the ligand molecules, the colors represent the atom types as follows (cyan: carbon; red: oxygen; pink: fluorine; yellow: sulfur; blue: nitrogen; purple: bromine; white: hydrogen).

Nicotine

Nicotine is a toxic alkaloid known from tobacco products. With only 26 atoms, it is the smallest ligand that is analyzed in this thesis. The structure with two carbon-rings restrains its flexibility. Therefore, we only consider one conformation to compute the valid ligand positions.

Figure 14: Nicotine : 3D structure and skeletal formula. Skeletal formula image from the Drugbank database (www.drugbank.ca)

We picked 100 Points by invoking the picking algorithm from section 3.1. We apply the greedy algorithm to 21 starting positions. The end node is the binding position. 6 starting positions result in in a sequence corresponding to pathway 2e, which accesses through the B-C loop. The positions are shown in Figure 15. It is a channel which is already partly open in the Cytochrome P450 3A4 crystal structure. 14 times they take a path where the last jump happens from under the molecule to the binding position. The positions before could indicate a preference for channel S/2f, but it can't be determined clearly, see Figure 16. We observe one time channel 3. The sequences leading to the binding position for nicotine are rather long, hence, the access seems unfavorable.

Channel	number
pw 2e	6
pw 3	1
not defined/ jump	14

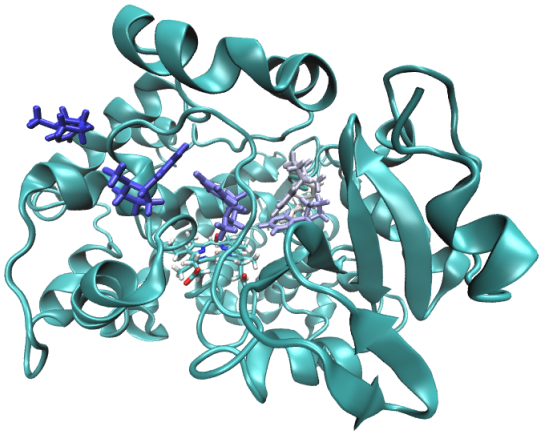

Figure 15: Nicotine's pathway into the binding pocket, corresponds to channel 2e

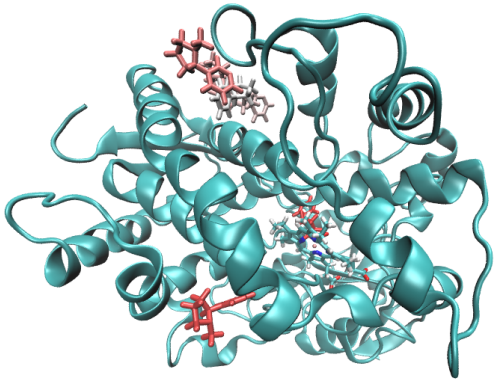

Figure 16: Nicotine: ligand positions that do not clearly indicate one channel

Bromperidol

Bromperidol is an antipsychotic. Applying the greedy algorithm, 2 dominant pathways are identified. 33 of the chosen 63 starting positions have their best pathway passing through position 124 before reaching the binding site. We relate this position to pathway S or 2f (between the β4 sheet and the F-G loop). 14 of the starting points take their way into the binding pocket through pathway 2b, which exits between the B-C loop and the β1 sheet. There are other channels identified, but each representing only a few (1-3) starting positions. Other starting positions include jumps from the surface to the inside in their position sequences, which is not realistic.

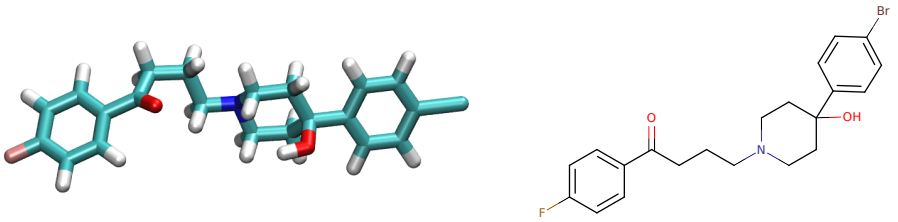

Figure 17: Bromperidol : 3D structure and skeletal formula. Skeletal formula image from the Drugbank database (www.drugbank.ca)

Hence, the two dominant pathways are 2f/S and 2b. The ligand positions corresponding to these channels are shown in Figure 18. The number of positions before reaching the binding position is between 2 and 5.

Channel	number
pw S/2f	33
pw 2b	14
pw 1	3
pw 2a	2
pw 3	3
not defined/ jump	8

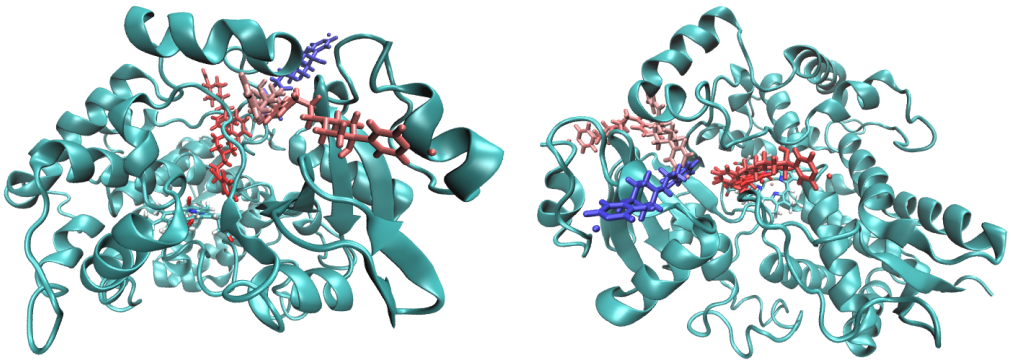

Figure 18: Two views of bromperidol's preferred channels. Channel S in dark red and channel 2b in light red. The blue ligand molecule indicates channel 2a.

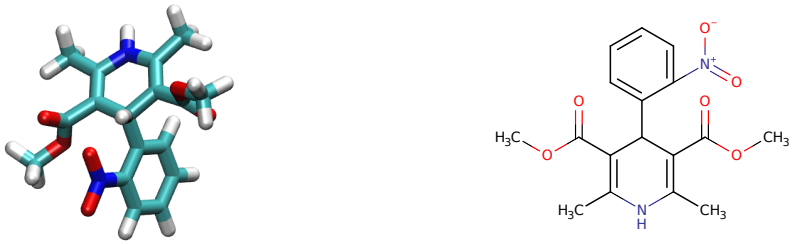

Figure 19: Nifedipine : : 3D structure and skeletal formula. Skeletal formula image from the Drugbank database (www.drugbank.ca)

Nifedipine

Nifedipine is a calcium channel blocker. It shows two prevailing pathways into the binding pocket. One can be identified with channel 1 and one corresponds to channel 5 or possibly channel 2a, see Figure 20. Nifedipine was one of the few ligands with successful pulling for channel 5 . This channel accesses next to the β4 sheet, close to channel 2a. The number of positions is between 5 and 11.

Channel	number
pw 5/ 2a*:	14
pw 1:	17
not defined/ jump:	3

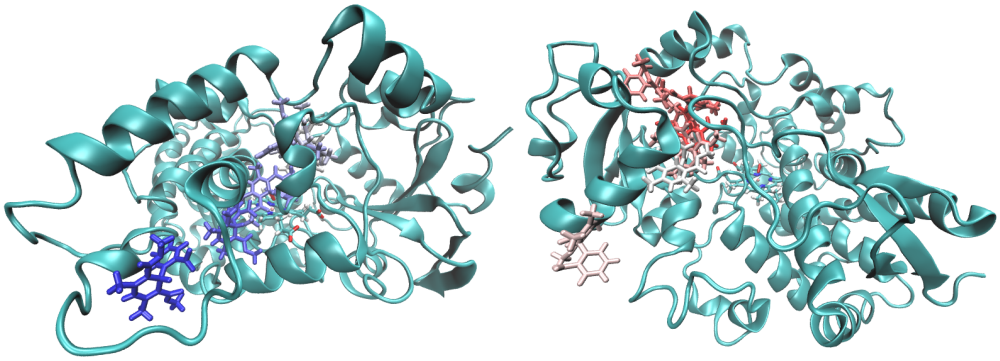

Figure 20: Nifedipine channels. Left: Channel 1. Right: Channel 5

Fluoxetine

Figure 21: Fluoxetine : 3D structure and skeletal formula. Skeletal formula image from the Drugbank database (www.drugbank.ca)

Fluoxetine is an antidepressant. There are two paths mainly used by fluoxetine. One corresponds presumably to pathway 2d, entering next to the A helix. The other path could correspond to either pathway 1 or pathway S, due to jumps between channels. Figure 22 shows the ligand positions indicating pathway 1 and pathway 2d, while Figure 23 shows the positions in pathway S. The path length ranges from 4 to 9.

Channel	number
pw 1 / S* :	20
pw 2d / 2b* :	19
not defined/ jump:	4

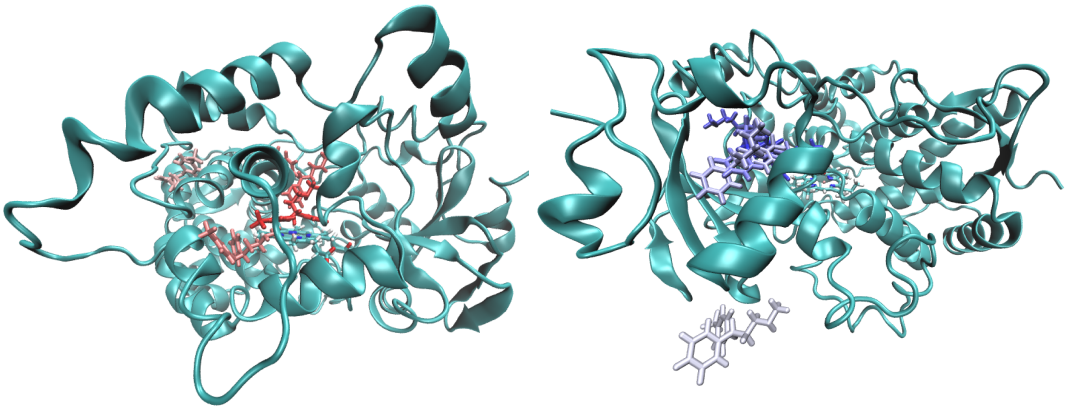

Figure 22: Fluoxetine channels. Left: Channel 1. Right: Channel 2d

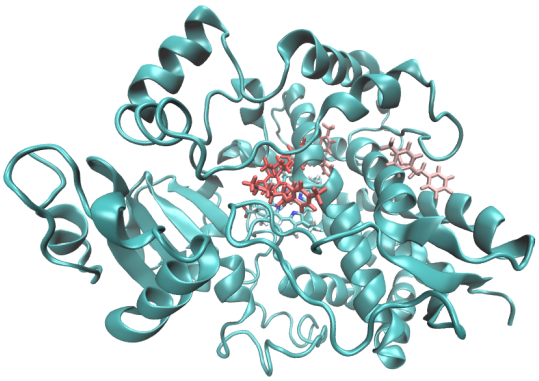

Figure 23: Fluoxetine positions in channel S

Terfenadine

Terfenadine is an antihistamine, but has been withdrawn from the market due to the risk of causing cardiac arrhythmia. It consists of 76 atoms and is the largest ligand used in this analysis.

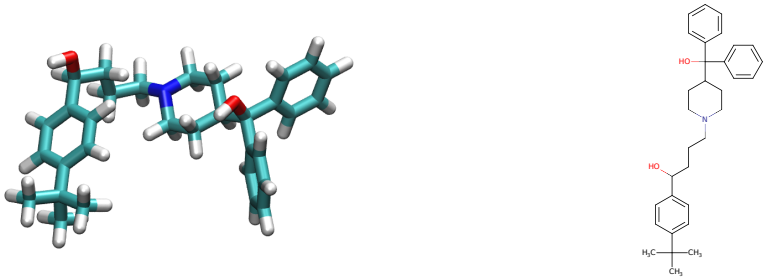

Figure 24: Terfenadine : 3D structure and skeletal formula. Skeletal formula image from the Drugbank database (www.drugbank.ca)

The greedy algorithm reveals one dominant path to the active site. It corresponds to either pathway 3 or pathway S. Due to jumps between channel 3 and the solvent channel S, we can't isolate the corresponding channel. Additionally, there is a small number of starting positions pointing to channel 2b, which can be clearly seen in Figure 25. The path length with 2 to 7 is quite short.

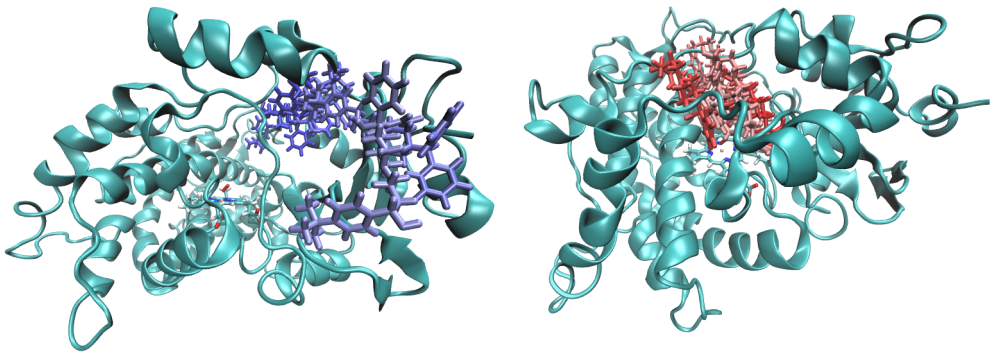

Figure 25: Terfenadine channels. Left: the 2b channel next to the B-C loop. Right: channel 3 right through the F-G loop

Channel	number
pw 3/S*:	24
pw 2b:	5
not defined/ jump:	4

*due to jumps between pw S and pw 3

Sufentanil

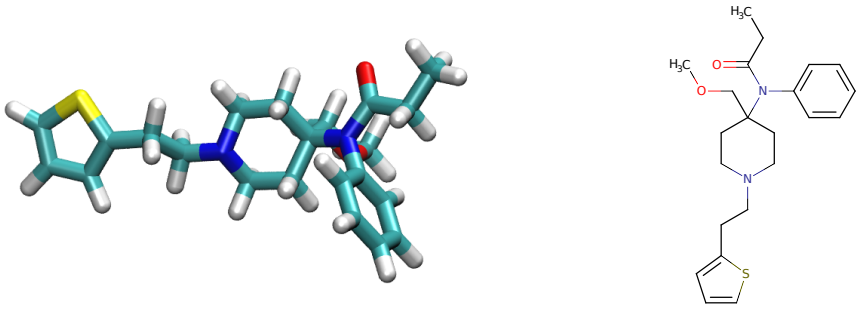

Figure 26: Sufentanil : 3D structure and skeletal formula. Skeletal formula image from the Drugbank database (www.drugbank.ca)

Sufentanil is an analgesic (painkiller). There are three channels showing a similar frequency of occurence. The first is channel 2a, with its entrance between the β1 sheet, the B-C loop and the F-G loop. Moreover, the ligand shows a preference for channel S and for channel 2c, next to the B-C loop. Figure 27 shows the corresponding ligand positions. The length is 3 to 7.

Channel	number
pw 2a	11
pw 2c:	11
pw S/2f	11
pw 2b	1
not defined/ jump:	9

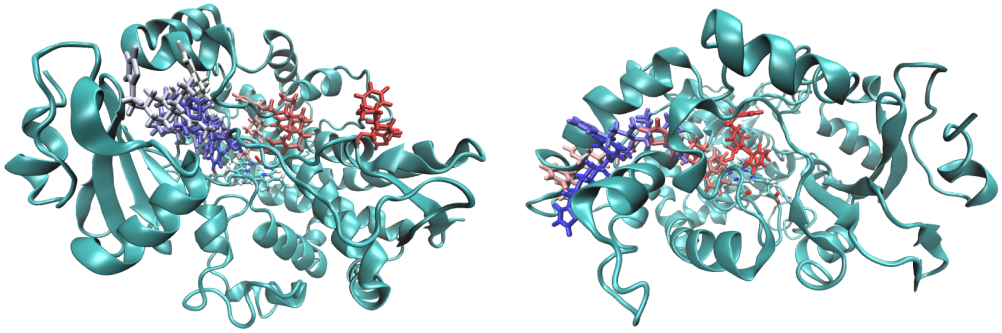

Figure 27: Pathways into the binding pocket for Sufentanil. Left: 2a (blue) and S (red); Right: 2c

Voriconazole

Voriconazole is an antifungal. The transition network reveals one dominant channel to access the active site, which is pathway 3. Ligands enter this channel through the F-G loop. Almost all voriconazole starting positions take this pathway, which can be seen in Figure 29. The path length is between 8 and 18.

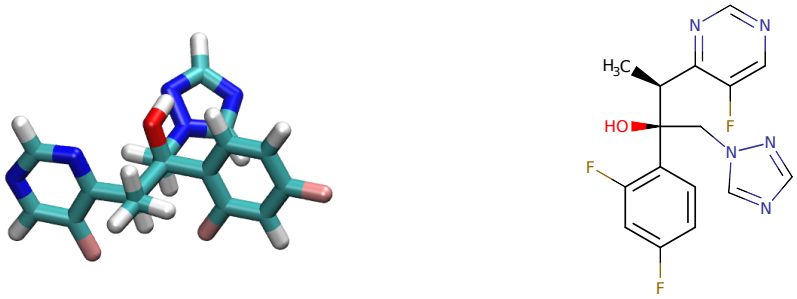

Figure 28: Voriconazole : 3D structure and skeletal formula. Skeletal formula image from the Drugbank database (www.drugbank.ca)

Channel	number
pw 3	38
not defined/ jump:	2

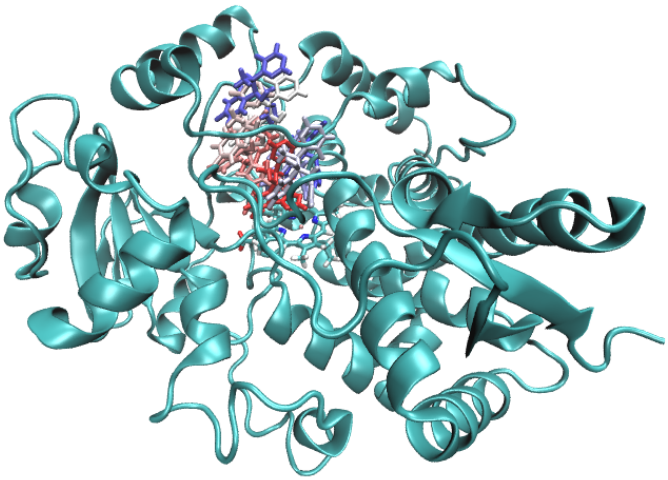

Figure 29: Voriconazole positions along pw 3 through the F-G loop

Comparison of all Ligands and Channels

There is a high variance in the preferred access channels to the active site. Table 1 lists all observed channels and the corresponding ligands. We only listed a channel for a ligand if there are 5 or more starting positions taking the sequence across the channel. We see that there is no dominant access route taken by all ligands, but instead every channel appears to be taken by at least one ligand. The preferred access channel seem to be heavily ligand specific. Surprisingly, channel 2e is only the suggested access channel of one ligand, namely nicotine, although it is already partly open in the CYP 3A4 crystal structure we start with.

Table 1: Channels and ligands with 5 or more starting positions yielding a sequence of positions that points to this channel

pw	nicotine	bromperidol	nifedipine	fluoxetine	terfenadine	sufentanil	voriconazole
1			x	x			
2a						x	
2b		x			x		
2c					x		
2d				x			
2e	x						
2f/S		x				x	
3					x		x
5			x				

Conclusion

We modeled the molecular binding kinetics of seven different ligands with Cytochrome P450 3A4 using a network approach. Assuming a Markov Process , we computed the transition rate matrix Q by the Square Root Approximation. We performed approximately 1000 energy minimizations. The good thing is that energy minimization is fast compared to trajectory computing. Since there were only a few valid positions in the access channels, we performed additional pulling simulations. The pulling simulations and preparations were rather time consuming. However, the results are a lot more meaningful with the additional positions, obtained by the pulling simulations.

There were two problems occuring due to the chosen Voronoi adjacency relation. We used the euclidian distance in $\mathbb{R}^3$ to define the adjacency of nodes. That caused problems for the refinement procedure, since there were adjacent positions with a high energy difference, that were in reality separated by the CYP structure. Hence, we could not choose valid positions in-between to refine. In reality, the direct path between the positions is obstructed, such that they should not be adjacent in the first place. Secondly, this adjacency relation was problematic for some of the resulting 'best' paths. We cannot be sure that unfavorable positions weren't skipped in the 'best' path, which might bias our results. Moreover, it allowed direct jumps from the CYP's surface to the binding position or between channels, as long as the positions were adjacent according to the Voronoi tessellation. That yielded some useless results. Nevertheless, this did not happen for the majority of the resulting paths. Both mentioned problems could probably be solved by using a different distance function. One could for example define a distance function along the Cytochrome's surface. This would change the Voronoi regions. The Ligand Excluded Surface [10] could be invoked to define the new distance function. Yet, one has to think about the CYP conformation to use for that, since we used different conformations corresponding to each ligand due to the conformational change during pulling simulations.

Despite these problems, the results seem reasonable. We have identified one or more dominant channels for each ligand and we have seen that different ligands prefer different channels, which might explain the broad range of ligands metabolized by CYP 3A4. The distance function for the adjacency relation and the algorithmic evaluation of the transition rate matrix could be adjusted for further improvement and individual requirements. We conclude that the infinitesimal generator approach by Weber, Fackeldey and Lie [9] yielded useful results. Furthermore, it assured a feasible computation time, since only energy minimization was required for the network positions. It is simple and fast if the desired positions are accessible by the ligand, what makes it a valuable approach for high dimensional systems!

References

[1] M. J. Abraham, T. Murtola, R. Schulz, Sz. Páll, J. C. Smith, B. Hess, and E. Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19 – 25, 2015.

[2] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, T N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000.

[3] V. Cojocaru, P. J. Winn, and R. C. Wade. The ins and outs of cytochrome p450s. *Biochimica et Biophysica Acta 1770*, pages 390–401, 2007.

[4] A. W. Sousa da Silva and W. F. Vranken. Acpype - antechamber python parser interface. *BMC Research Notes*, 5(1):367, Jul 2012.

[5] P. Deuflhard and M. Weber. Robust perron cluster analysis in conformation dynamics. *Linear Algebra and its Applications*, 398:161 – 184, 2005. Special Issue on Matrices and Mathematical Biology.

[6] D. Fishelovitch, S. Shaik, H. J. Wolfson, and R. Nussinov. Theoretical characterization of substrate access/exit channels in the human cytochrome p450 3a4 enzyme: Involvement of phenylalanine residues in the gating mechanism. *Journal of Physical Chemistry B*, 113(39):13018–13025, 2009.

[7] A. Gora, J. Brezovsky, and J. Damborsky. Gates of enzymes. *Chemical Reviews*, 113(8):5871–5923, 2013. PMID: 23617803.

[8] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[9] H. Ch. Lie, K. Fackeldey, and M. Weber. A square root approximation of transition rates for a markov state model. *SIAM. J. Matrix Anal. Appl.*, 34(2):738 – 756, 2013.

[10] N. Lindow, D. Baum, and H. C. Hege. Ligand excluded surface: A new type of molecular surface. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2486–2495, Dec 2014.

[11] P. C. Nair, R. A. McKinnon, and J. O. Miners. Cytochrome p450 structure-function: insights from molecular dynamics simulations. *Drug Metabolism Reviews*, 48(3), 2016.

[12] Sz. Páll, M. J. Abraham, C. Kutzner, B. Hess, and E. Lindahl. Tackling exascale software challenges in molecular dynamics simulations with gromacs. In St. Markidis and E. Laure, editors, *Solving Software Challenges for Exascale*, pages 3–27, Cham, 2015. Springer International Publishing.

[13] V. S. Pande and E. J. Sorin. Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophysical Journal*, 88(4):2472 – 2493, 2005.

[14] J. A. Peterson and S. E. Graham. A close family resemblance: the importance of structure in understanding cytochrome p450. *Structure*, 6(9):1079–1085, 1998.

[15] S. Pronk, Sz. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl. Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013.

[16] M. Sarich. *Projected Transfer Operators*. PhD thesis, Freie Universität Berlin, 2011.

[17] M. Sarich, F. Noé, and Ch. Schütte. On the approximation quality of markov state models. *Multiscale Model. Simul.*, 8(4):1154 – 1177, 2010.

[18] Ch. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A direct approach to conformational dynamics based on hybrid monte carlo. *Journal of Computational Physics*, 151(1):146 – 168, 1999.

[19] Ch. Schütte, W. Huisinga, and P. Deuflhard. Transfer operator approach to conformational dynamics in biomolecular systems. In *ERGODIC THEORY, ANALYSIS, AND ECIENT SIMULATION OF DYNAMICAL SYSTEMS*, pages 191–223. Springer, 1999.

[20] J. Stone. *An Efficient Library for Parallel Ray Tracing and Animation*. Master's thesis, Computer Science Department, University of Missouri-Rolla, April 1998.

[21] M. Weber. *Meshless Methods in Confirmation Dynamics*. PhD thesis, Freie Universität Berlin, 2006.

[22] M. Weber. Conformation-based transition state theory. Technical Report 07-18, ZIB, Takustr.7, 14195 Berlin, 2007.

[23] M. Weber. An efficient analysis of rare events in canonical ensemble dynamics. Technical Report 08-36, ZIB, Takustr.7, 14195 Berlin, 2008.

[24] M. Weber. A subspace approach to molecular markov state models via a new infinitesimal generator, 2011.

[25] M. Weber, K. Fackeldey, and Ch. Schütte. Set-free markov state model building. *Journal of Chemical Physics*, 146(12), 2017.

[26] D. Werck-Reichhart and R. Feyereisen. Cytochromes p450: a success story. *Genome Biology*, 1(6):reviews3003.1, Dec 2000.

[27] D.S. Wishart, C. Knox, A.C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, 2006.

[28] X. Yu, V. Cojocaru, and R. C. Wade. Conformational diversity and ligand tunnels of mammalian cytochrome p450s. *Biotechnology and Applied Biochemistry*, 60(1), 2013.

[29] A. Zawaira, L. Coulson, M. Gallotta, O. Karimanzira, and J. Blackburn. On the deduction and analysis of singlet and two-state gating-models from the static structures of mammalian cyp450. *Journal of Structural Biology*, 173(2), 2011.

A Appendix

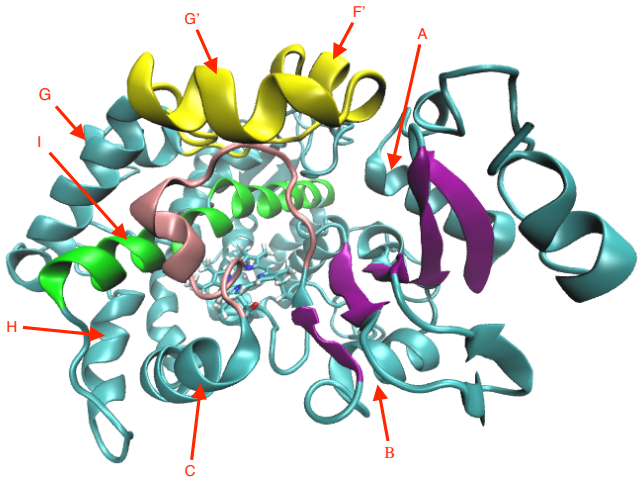

Figure 30: CYP 3A4 with labeled secondary structure elements, view 1

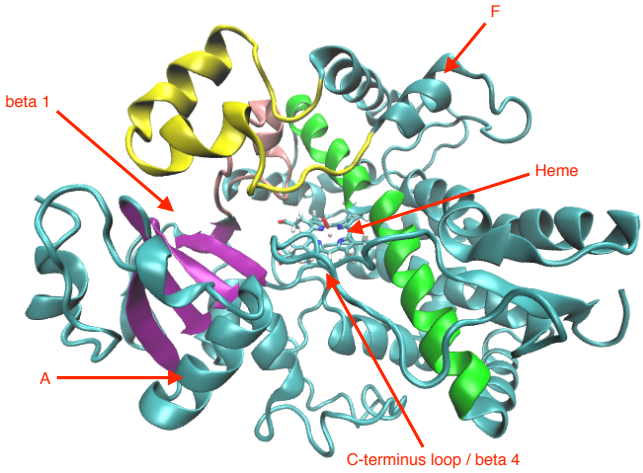

Figure 31: CYP 3A4 with labeled secondary structure elements, view 2