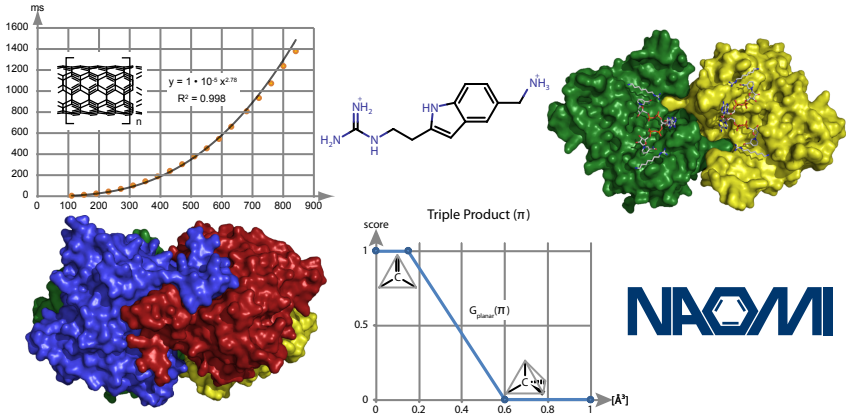
In silico modeling of small molecules and design of eIF-5A activation inhibitors



Cumulative Dissertation

with the aim of achieving the degree

Dr. rer. nat.

at the Faculty of Mathematics, Informatics and Natural Sciences

Department of Informatics

University of Hamburg

submitted by

Adrian Kolodzik

Hamburg, July 2014

Gutachter:
Prof. Dr. Matthias Rarey
Prof. Dr. Chris Meier
Prof. Dr. Johannes Kirchmair

Tag der Disputation:
31. Oktober 2014

Für meine Eltern Arnold Kolodzik und Ute Zarsen-Kolodzik, meine Brüder Matthias Zarsen und Karsten Mundhenk und meine Freundin Stefanie Lange

Kurzfassung

Die vorliegende Dissertationsschrift beschreibt meine Forschung in der Abteilung für Algorithmisches Molekulares Design im Zentrum für Bioinformatik der Universität Hamburg in der Zeit vom 15.02.2008 bis zum 30.11.2011. Die Resultate wurden in fünf Postern, zwei Konferenz-Vorträgen und sechs Publikationen in wissenschaftlichen Fachzeitschriften veröffentlicht (siehe Kapitel 7 und Anhang A). Eine weitere Publikation ist in Vorbereitung und zwei Lehrbuchkapitel befinden sich in redaktioneller Bearbeitung.

Im Rahmen der Wirkstoffentwicklung werden routinemäßig Programme eingesetzt, die kleine chemische Moleküle und ihre Interaktionen mit Proteinen modellieren. Durch virtuelles (*in silico*) Screening werden aus großen Substanz-Bibliotheken Moleküle selektiert, die mit hoher Wahrscheinlichkeit an Ziel-Proteine binden. Diese Moleküle werden dann im Labor (*in vitro*) auf ihre Aktivität untersucht und können nach weiteren Tests als Wirkstoffe in Medikamenten eingesetzt werden.

Voraussetzung für die korrekte Berechnung von molekularen Interaktionen und Eigenschaften ist die konsistente Behandlung von Molekülstrukturen aus unterschiedlichen Dateiformaten. Viele Programme erfüllen dieses Kriterium nicht und einige Eigenschaften, wie beispielsweise das kleinste Set der kleinsten Ringe einer Molekülstruktur, sind konzeptionell nicht eindeutig.

Um diese Probleme zu adressieren, habe ich die Software-Bibliothek NAOMI mitentwickelt, die Moleküle aus unterschiedlichen Dateiformaten konsistent verarbeitet. Es bildet die Grundlage für weitere Softwareentwicklungen. Weiterhin habe ich Erweiterungen von NAOMI zur Behandlung molekularer Zustände (prototrope Tautomere und Protonierungszustände) mitentworfen. Mit den *Unique Ring Families* entwickelte ich ein Konzept, das zum ersten Mal eindeutig und intuitiv molekulare Ring-Topologien beschreibt und deren effiziente Berechnung (in polynomieller Zeit) erlaubt. Durch die Anwendung von NAOMI in Kombination mit etablierter Software konnte ich zwei Inhibitoren der humanen Deoxyhypusin-Synthase (DHS) entwerfen, die erfolgreich die Vermehrung von HI-Viren in Zellkultur hemmen.

Abstract

This dissertation describes my research in the group for Computational Molecular Design of the Center for Bioinformatics at the University of Hamburg from February 15^{th}, 2008 until November 30^{th}, 2011. The results have been published in five poster presentations, two talks on scientific conferences, and six publications in scientific journals (see chapter 7 and appendix A). One additional publication is in preparation and two chapters for a text book are in the editorial process.

In the drug development process, programs are routinely used to model small molecules and their interactions with proteins. Virtual (*in silico*) screening allows the selection of compounds from large substance libraries which have a high probability of binding to a target protein. Most promising molecules can then be tested in a laboratory (*in vitro*) for their activity and can become the active substances of new drugs if additional tests are successful.

A prerequisite of the accurate calculation of molecular interactions and properties is the consistent handling of molecular structures from different file formats. Many programs do not meet this requirement and some properties such as the smallest set of smallest rings of a molecular structure are ambiguous by definition.

To address these problems, I co-developed the software library NAOMI which consistently handles molecules from different file formats. NAOMI now forms the basis for a number of additional software developments. Furthermore, I have co-developed an extension of NAOMI to model different molecular states such as prototropic tautomers and protonation states. With the *Unique Ring Families* (URFs) I developed a concept which represents the first unique and intuitive description of molecular ring topologies, that can be calculated efficiently (in polynomial time). Using NAOMI in combination with a number of virtual screening tools, I have been able to successfully design two inhibitors of human deoxyhypusine synthase (DHS), which inhibit the replication of HI viruses in cell-culture.

Contents

Chapter 1

Introduction

1.1 Overview

This thesis describes my research in the group for computational molecular design [1] at the Center for Bioinformatics (ZBH) [2] of the University of Hamburg. The group is specialized on cheminformatics [3, 4] and focuses on the computer-based (*in silico*) modeling of molecular structures and their interactions. An important part of cheminformatics is the design of small molecular inhibitors of proteins. Since proteins have a vast number of functions in biological systems, their inhibition can have strong effects on organisms. In 2008 the ZBH became a member of the consortium "Combating Drug Resistance in Chronic Myeloid Leukemia and HIV-1 Infections". The consortium was funded by the *Bundesministerium für Bildung und Forschung* (BMBF) [5]. Since the project required different areas of expertise, the consortium consisted of six different groups with experience in Chemistry, *in vitro* and *in vivo* testing of small molecules, x-ray crystallography [6], and cheminformatics. In the context of this consortium it was one goal to design inhibitors against human Deoxyhypusine Synthase (DHS) [7] *in silico*. DHS is a protein, which is involved in the replication of the human immunodeficiency virus (HIV) [8] via activation of the protein eIF-5A [9]. Inhibition of DHS therefore is a promising approach to inhibit HIV replication and represents a new strategy to treat HIV infections. Besides treating HIV infections, the project also aimed at the design of drugs to treat chronic myeloid leukemia (CML) [10, 11]. CML is a disorder, which leads to an unregulated growth of myeloid blood cells. It is caused by a chromosomal translocation between the chromosomes 9 and 22. This leads to an elongation of chromosome 9 and a truncation of chromosome 22. The truncated chromosome 22 is also called Philadelphia chromosome [12] and includes a fusion of the genes BCR and ABL. BCR-ABL positive cells contain a constitutively activated

tyrosine kinase. Due to the fact that activated eIF-5A plays a role in the proliferation of these dysregulated cells [13], DHS is also a target for the therapy of CML.

To develop inhibitors against DHS, the cheminformatics tools, which had been developed at the ZBH, were applied. These tools are designed to support the design of small molecular protein inhibitors. During the application of the available tools, their limitations became obvious. Molecular structures were not correctly processed and results varied for different input file formats. An analysis revealed that these issues were not caused by technical problems of the underlying code basis, but by the way molecular structures were modeled. The analysis of commercially and freely available tools showed similar limitations [A1]. To address these issues, a new way of modeling chemical structures was designed and implemented at the ZBH. It was the short-term goal of this approach to support the design of DHS inhibitors as well as other ongoing projects. The long-term goal was to address these cheminformatics issues in general and to provide a thoroughly validated code basis for future drug development projects.

The following sections introduce different cheminformatics aspects to provide a foundation for the following chapters. Chapter 2 describes the development of cheminformatics methods to address the observed limitations of cheminformatics tools. It includes an introduction to commonly used molecular file formats and describes how these can be processed consistently with the newly developed tool NAOMI [A1]. It also describes the extension of the initial NAOMI tool to read molecular structures from PDB [14] files and model different representations of molecules. Chapter 3 introduces a new concept for the perception of molecular ring topologies, which addresses the shortcomings of commonly used alternatives. Chapter 4 focuses on the *in silico* design of inhibitors against DHS and CNI-1493. A conclusion and an outlook are provided in chapter 5.

1.2 *In silico* Drug Design

During the last decades a number of approaches for the *in silico* design of drugs have been developed. These approaches use the increasing power of modern computer systems to identify new potential ligands of proteins, to develop hypotheses about their binding modes, and to optimize already known ligands. Depending on the input data, the strategies for *in silico* drug design can be divided into two basic categories:

Ligand-based drug design approaches solely rely on data of known active and inactive molecules. Based on the assumption that similar structures have similar properties, structure-activity relationships (SARs) [15] can be analyzed. A quantitative structure-activity relationship (QSAR) [16, 17] is an example of a ligand-based design strategy,

which predicts a compounds activity based on a linear combination of molecular properties. If binding affinities or inhibitory effects at different concentrations are known for a sufficient number of molecules, a model can be built which correlates the molecular properties with the measured affinities or inhibitory effects. This model can then be used to predict the binding affinities of similar molecules, which have not yet been tested. Despite the common use of QSAR models in cheminformatics, the underlying assumption, that similar molecules have similar properties, is not always true. In some cases even small modifications of a molecule can lead to significant changes of the molecule's binding affinity. This phenomenon is called activity cliff [18].

Structure-based drug design [19, 20] makes use of known protein structures and tries to identify binders, which have an optimal fit to the protein's ligand binding site. Molecular docking [21] belongs to these structure-based design approaches and LeadIT [22] is an example of a software which provides docking functionalities. A docking with LeadIT consists of the following steps. First a ligand is virtually placed into the binding site of a protein. The particular orientation of the ligand in the context of the protein is called a pose. For each pose potential protein ligand interactions [23] are calculated. These include, for example, electrostatic interactions and van der Waals contacts. In addition to interactions which support the binding of a ligand, protein-ligand and intra-ligand clashes are calculated and penalized. Modern scoring functions also include entropic effects by modeling the desolvation of the ligand and the protein upon binding [24]. The resulting score can be used to estimate the binding affinity and to classify ligands as predicted binders or predicted non-binders. This classification serves as a filter to select promising compounds from compound libraries. Selected compounds can then be tested *in vitro* or *in vivo* for inhibitory activity or toxic side effects.

Despite the increasing performance of modern computers and the development of enhanced methods for the *in silico* design of new drugs, the accurate prediction of binding affinities is still a challenge. A first obstacle are the different representations of input structures. These often originate from different sources or follow different conventions. Furthermore, they are sometimes provided in different file formats or lack information and are therefore under-determined. This poses a problem since ligand-based drug design approaches rely on accurate information about known active molecules. A second challenge is the correct modeling of the ligand in the context of the protein binding site. The binding site of a protein can strongly influence a ligand and lead to unusual protonation states or tautomers [25, 26], which are only rarely observed in solution. Incorrectly charged structures and tautomeric forms of a ligand can lead to an incorrect

prediction of molecular interactions and consequently to incorrectly predicted binding affinities [27].

A description of the current methods of *in silico* drug design and the validation of the different methods will be published in the chapter "Structure Based Virtual Screening" of the next edition of the "Handbook of Cheminformatics" [B1].

1.3 Molecular Descriptors

As described in the last section, a number of different tools and approaches are available to predict small molecule inhibitors of target proteins. Even if a compound is predicted to inhibit a protein, it might still be inactive or it might be unsuited to serve as a potential drug candidate. Properties which are important for a new drug are the drug's absorption, distribution, metabolism, excretion, and toxicity. These properties are summarized under the term ADMET [28, 29]. The importance of these aspects vary for individual drugs. Intra-cellular targets require, for example, that a drug passes cell membranes. Toxic effects are usually undesired, but can, to some degree, be tolerated if no alternative options for the treatment of a disease are available. To minimize the risk of spending efforts on the analysis of compounds, which are likely to fail at later stages of the drug development process, molecules are usually filtered by molecular descriptors at early stages. This is especially important since the costs for the development of a drug have been estimated to be approximately 1.8 billion dollars [30]. If liabilities of lead structures can be identified early, the associated losses can be minimized.

A common descriptor which is used in these filters is the logP [31, 32] value. The logP is the logarithm of the ratio of a molecule's concentration in its neutral form in the two solvents octanol and water. Therefore, the logP is a measure of lipophilicity and an indicator of a molecule's ability to passively pass cell membranes as well as an indicator of its potential oral availability. Another commonly used descriptor is a molecule's number of rotatable bonds. It describes a molecule's flexibility and is consequently an indicator of a molecule's loss of entropy upon binding. A set of filters which combines multiple descriptors to assess a molecule's oral bioavailability is Lipinski's *Rule-of-Five* [33]:

- molecular weight $\leq$ 500 Da
- logP ≤ 5
- ≤ 10 hydrogen bond acceptors
- ≤ 5 hydrogen bond donors

The *Rule-of-Five* has been established by analyzing the properties of known drugs. A number of variations of this rule exists. Lead structures in the drug development process are often optimized by adding or replacing small substructures to finally form a drug. This usually leads to molecules with an increased molecular weight as well as a higher number of hydrogen bond donors, hydrogen bond acceptors, and rotatable bonds. Lead structures are therefore often smaller than drugs and are described by a different set of rules [34].

Using these filters, molecules with desirable physico-chemical properties can be selected. The calculation of the individual descriptors depends on the particular molecular representations, i.e. the tautomeric forms and protonation states. Even if molecular representations are sensible and pass the drug-like filters, they can still be reactive or instable. Molecules containing reactive substructures should therefore be excluded. The identification of the molecular substructures can be achieved by SMARTS [35] pattern matching, which is described in the next section. All of these descriptors support the selection of promising drug candidates and reduce the number of molecules which have to be tested in later stages of the drug development process. For a detailed list of available descriptors see the book "Molecular Descriptors for Chemoinformatics" [36] by Todeschini and Consonni.

1.4 SMARTS Pattern Matching

A SMILES string is a 1D representation of a molecular structure. SMARTS is an acronym for **SM**iles **AR**bitrary **T**arget **S**pecification. SMARTS strings describe molecular patterns and represent an extension of the SMILES language. In contrast to SMILES strings, SMARTS allows logical operators as well as specialized symbols to further describe atoms and bonds. An example is the description of atoms which are connected to a particular number of heavy atoms or part of a certain number of rings. These patterns can be fairly easy to interpret but can also be highly complex and include recursive definitions. To visualize even complex SMARTS patterns, the tool SMARTSviewer [37] is available. An example of SMARTS patterns and their visualizations with SMARTSviewer is shown in Figure 1.1.

SMARTS patterns are a valuable tool to identify molecules which contain certain substructures. A common application is the search for molecules which include substructures that are known to be required for the binding to a target protein. Another application is the exclusion of reactive functional groups. An example is the exclusion of "Pan Assay Interference Compounds" (PAINS) [38] from high throughput screens. These include compounds with functional groups that are not suitable to be used in

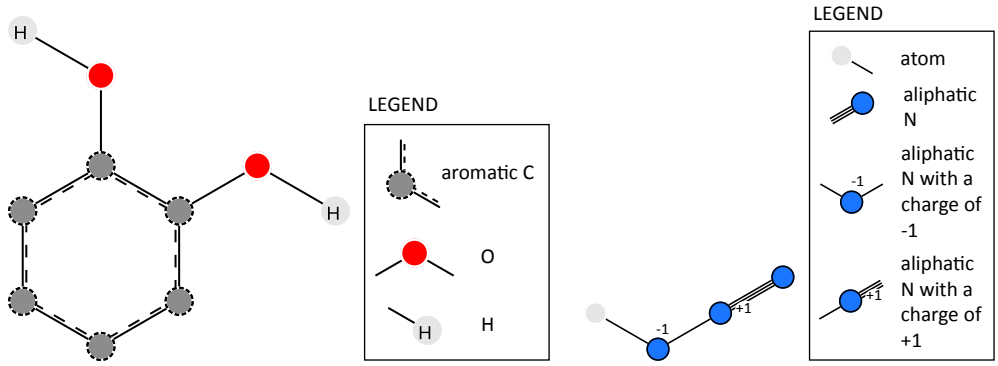


Figure 1.1: Shown are two different SMARTS patterns and their corresponding depictions as generated by SMARTSviewer. A: SMARTS pattern matching catechols. B: SMARTS pattern matching azides.

high throughput screening (HTS) campaigns. Due to these common applications, large public databases like the ZINC [39] - and the Pubchem [40] databases directly support the matching of SMARTS expressions on their web frontend.

Despite their common use, the results of matching a SMARTS expression can be undefined if the ring memberships of individual atoms are matched. The ring definition of SMARTS is based on the smallest set of smallest rings (SSSR) [41]. An SSSR is not necessarily unique for a given molecule, which can lead to ambiguous and counter-intuitive results for the calculated ring-membership of individual atoms. While this only has an impact on certain combinations of SMARTS patterns and molecular structures, it illustrates the need for a new and improved concept to describe molecular ring topologies.

Chapter 2

In silico Modeling of Molecules

2.1 Introduction to Molecular File Formats

A key element of all cheminformatics methods is the handling of data about molecular structures. Molecular structures have been represented in the form of Lewis structures for almost a hundred years [42]. Cheminformatics applications commonly interpret molecules as undirected and labelled graphs [43]. An undirected graph is a set of objects (called vertices or nodes) which are connected by bidirectional links (called edges). Atoms are represented by a graph's vertices and bonds are represented by edges connecting the vertices. Vertices usually have annotated properties such as chemical elements and coordinates. Edges are often labelled with bond orders. To share these molecular representations, a number of file formats are available. The most commonly used file formats are Accelrys SDF V2000 [44] (formerly MDL SDF), Tripos MOL2 [45], Daylight SMILES [35, 46] and PDB [14]. While almost all molecular databases provide structural data of small molecules in at least one of these formats, the underlying chemical models differ significantly.

The Accelrys SDF format may contain multiple entries. Each entry includes a connection table which describes a single localized valence bond structure of a molecule. It contains coordinates and explicit formal charges for each atom and only allows single, double, and triple bonds. Hydrogen atoms are commonly omitted. Furthermore, Accelrys SDFiles support additional data fields to annotate custom properties of a molecule. While this file format requires a large amount of disc space, it precisely describes a particular valence bond structure and allows the flexible annotation of additional properties.

SMILES is an acronym for **S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem [35]. A SMILES string is a 1D representation of a molecular structure. Charges are explicitly

annotated but hydrogen atoms are usually omitted. Hydrogen atoms can be derived from localized bond orders and formal charges. In addition to localized bond orders, SMILES strings can contain symbols for aromatic atoms and bonds. A single SMILES string can therefore describe different Kekulé forms [47]. An advantage of SMILES strings is the small amount of disc space, which is required to encode a molecule. Additionally, it represents a well-defined valence bond structure, if symbols for aromatic atoms and aromatic bonds are not used. Simple SMILES strings have the additional advantage of being human-readable. A special type of SMILES string is a USMILES [48], which can serve as a unique identifier for a molecule. A limitation of SMILES strings is their lack of atomic coordinates. SMILES strings can therefore not be used to store specific conformations of molecules.

The Tripos MOL2 format describes molecules by a connection table of bonds and atoms with annotated properties. Hydrogen atoms are frequently omitted, but each atom is assigned a Sybyl Type [49]. This includes information about the atom's element as well as its hybridization state. Nonetheless, Sybyl Types are not available for each element. Furthermore, some Sybyl Types are ambiguous and can describe multiple elements. Examples are the Sybyl Types *Het* and *Hev*, which describe hetero atoms (N, O, S, P) and heavy (non-hydrogen) atoms, respectively. The bonds in a MOL2 file are labelled as single, double, triple, or aromatic. The MOL2 file format provides coordinates for all atoms and supports the annotation of different concepts for the calculation of charges. Consequently, it is an ideal format for storing molecular data from cheminformatics applications. Examples of such applications are molecular dynamic - and molecular mechanic calculations. There are two disadvantages of using the MOL2 format. Firstly, a large amount of disc space is required to encode the structural information. Secondly, the description of molecular entities can be ambiguous. This is especially likely if hydrogen atoms and charges are omitted and aromatic bonds are used in the MOL2 file. In this case, multiple valence bond structures and protonation states can match the description of a single MOL2 entry. Furthermore, if no Sybyl type is defined or if the Sybyl type is ambiguous, the atom's annotated name has to be used to identify its element. Since there is no standardized way of doing this, the correct interpretation of MOL2 files can be challenging.

The PDB file format describes atoms by their element symbols and their atomic coordinates. Connections between atoms can be explicitly specified. Bond orders are not annotated. For atoms belonging to standard amino acids, the determination of bond orders is straightforward, since templates or residues can be annotated. For atoms of small compounds, which do not have an annotated residue or template, the assignment of bond orders based on atomic coordinates can be difficult. The PDB

format is therefore ideal to store the results of experimental methods, which are used to determine 3D coordinates of the atoms of proteins. An example of such a method is x-ray crystallography [6]. Due to the lack of annotated bond orders and templates, the PDB format is only of limited use to encode structures of small molecules.

Due to their different advantages, these chemical file formats are commonly used in the scientific community. It is therefore a basic requirement for every cheminformatics tool to handle all these file formats consistently.

2.2 Reading Different File Formats

Despite advances in cheminformatics software, the consistent modeling of molecules from different file formats remains a difficult task. To achieve this goal, the tool NAOMI [A1] was developed. It provides a solid basis for the application of virtual screening methods at the Center for Bioinformatics. The publication describing the tool in detail is included in the appendix of this dissertation in chapter A.1.

NAOMI consistently handles the molecular file formats SMILES [35, 46], MOL2 [45], SDF [44], and PDB [14]. They do not only differ with respect to their format specifications, but have different underlying chemical models. While SMILES and MOL2 files support the description of conjugated rings with aromatic bonds, SDF entries describe single valence bond structures with localized bond orders and formal charges. PDB files of small molecules usually do not even contain bonds but only three dimensional atomic coordinates of the molecules' atoms.

Because of these differences, the demands on the software also significantly differ for each of these formats. Due to the localized bond orders in SDFiles, the interpretation of the underlying molecular structures is straightforward. Since no information about electronically conjugated systems is annotated, aromaticity and delocalized bonds have to be determined by the software if they are required for an application. On the contrary, MOL2 files and SMILES strings can include information about electronically conjugated systems. To check the chemical validity of delocalized descriptions, localized valence bond structures have to be generated. The interpretation of PDB files requires the detection of bonds and bond orders based on the atomic coordinates. Since the determination of these coordinates is error prone, a software has to cope with this uncertainty. A number of geometric properties can be calculated on the basis of atomic coordinates to determine bonds and bond orders. These properties include atomic distances, bond angles, torsion angles, and measures of planarity. Due to the experimental error of the determination of the atomic coordinates, the calculation of bond orders is not straightforward. If multiple solutions are equally acceptable with respect to the

coordinates, chemical knowledge is required to select the most likely structures. The handling of PDB files is therefore complex and computationally challenging. A robust approach which uses the NAOMI framework to read PDB files has been published in the context of this dissertation [A2]. Chapter A.4 of the appendix includes a copy of the publication. For the design of DHS inhibitors, docking experiments were carried out with the software LeadIT [22] (version 1.4). A fundamental limitation was the incorrect protonation of small molecules from MOL2 and PDB files. This lead to incorrect results of initial docking experiments. These experiments had to be repeated and manually corrected. The NAOMI framework and its extensions represented a significant improvement compared to these early versions of LeadIT. Using NAOMI it was possible to first preprocess input structures and then import the structures into LeadIT without allowing further changes to the protonation.

In the following chapters, the concepts behind NAOMI will be described in more detail. NAOMI assigns localized bond orders and uses three distinct levels to describe atoms (see Figure 2.1). At the first level an element type is assigned to each atom. In NAOMI, an atom's element is described by an object which annotates information, which only relies on the atom's chemical element and isotope. This atom type therefore includes the atomic number, atomic weight and number of valence electrons. A second level of description is the valence state type, which is assigned to each atom. The assignment only depends on an atom's valence state and includes information about an atom's formal charge and the number of single bonds, double bonds, and triple bonds. Each valence state has a unique name which consists of the element symbol followed by the number of single bonds, double bonds, triple bonds, and the formal charge. The valence state types and localized bond orders of a molecule describe a localized valence bond structure. As a third level of description, each atom is linked to an atom type. An atom type describes an atom in the context of the molecule. It includes information about an atom's ideal geometry, the corresponding Sybyl type, and it annotates if the atom is part of an aromatic ring.

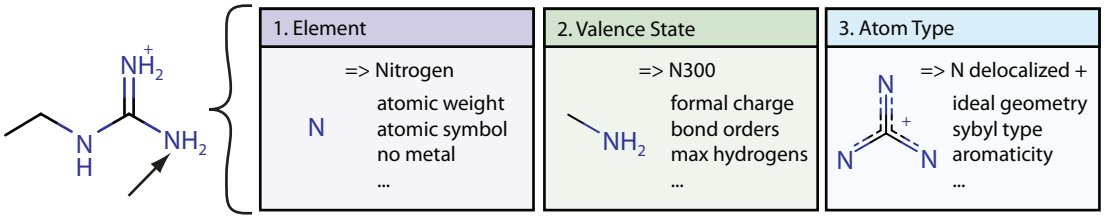


Figure 2.1: NAOMI describes each atom at three levels.

These three levels of description support different applications. Simple applications like the calculation of the molecular weight are independent of the respective valence state types or atom types. They only require the information which is stored as part of the element type. A localized valence bond structure can be checked for validity using the valence state types. If the bonds of a molecular structure are compatible with the assigned valence states, the valence bond form is valid. To identify chemically similar atoms, atom types are a suitable level of description since they include information about the atom's context. The two terminal nitrogen atoms of a guanidinium group illustrate this (see Figure 2.1). They have different valence states ("N300" and "N210+") depending on the position of the double bond. Due to the delocalized charge, the atoms can be considered equivalent. This is reflected by an identical atom type assignment ("N delocalized +") for both atoms. A second example are neutral aliphatic and aromatic carbon atoms with two single bonds and one double bond. While the valence states are identical, the atom types are different due to the delocalized electronic system of the aromatic ring.

While this illustrates the advantages of the NAOMI model from a conceptional point of view, the described approach is also computationally efficient. Element types, valence state types, and atom types can be described by a finite number of static objects which only have to be stored in memory once. Furthermore, the validity check of a molecule is straightforward, since the assignment of valence states is restricted to valid types. If a molecule contains atoms for which no valence state types exist, a simple correction is applied (e.g., excess hydrogen atoms are removed). If the correction attempt is unsuccessful, the molecule is considered invalid. In this case, the molecule is rejected. By only allowing valid valence states and localized bond orders, it can be easily checked if a valid valence bond form exists for a molecular representation. This is achieved in two steps. Firstly, valence states, which are compatible with the given representation, are assigned to each atom. Secondly, bond orders are assigned. A valid valence bond structure exists, if there is an assignment of bond orders, which is compatible with the valence state types of the atoms.

To analyze if the assignments of elements, valence states, and atom types are consistently handled by the NAOMI framework, a number of validation strategies were developed. The basis for these validation approaches is the generation of USMILES strings by the NAOMI framework. These can serve as molecular identifiers since they provide a unique 1D representation for each molecule.

To validate the file conversion with NAOMI, the structures of the directory of useful decoys (DUD) [50] were used as input. The DUD decoy and DUD ligand data sets were downloaded [51] in the SDF format as well as the MOL2 format. In a first step, the

molecular structures were initialized from both formats and converted to USMILES [48]. All corresponding entries in MOL2 and SDF format for which different USMILES were generated, were individually inspected to validate the initial data set. Afterwards, the following four aspects of format conversions were analyzed:

1. Error Correction:
   Are incorrect molecular representations corrected?
2. Conversion:
   Does the interconversion of formats result in identical structures?
3. Consistency:
   Are two consecutive conversion steps leading to consistent results?
4. Robustness:
   Can input from other tools be handled consistently? Does a tool improve the results of a consecutively used tool?

The results were compared to the commercial software CORINA [52], the commercial software MOE [53], and the freely available software Open Babel [54]. For a more detailed description of this analysis, see the publication [A1] which is attached in appendix A.1 of this thesis. As shown in the publication, NAOMI is superior to the tools CORINA, MOE, and OpenBabel in all of these aspects.

This strength of the NAOMI framework is a direct result of the concept to assign a valence bond type to each atom of a molecule and to only allow localized bond orders. If the assignment of valence state types and bond orders is carried out consistently for different file formats, the conversion between the file formats is straightforward. Nevertheless, the concept of NAOMI is limited to the representation of Lewis structures. Electron-deficient bonding is an example of a bond type, which cannot be correctly modeled by the NAOMI framework. Diborane [55] (see Figure 2.2) contains two B-H-B bonds. Each of these bonds contains two valence electrons which are distributed between the three atoms. A bond which is connected to three atoms is not available in the NAOMI framework and a hydrogen atom cannot have more than one bond. A hydrogen atom which is connected to two atoms (in this case boron) would therefore be considered invalid. This limitation is acceptable for regular cheminformatics tasks since electron deficient bonds are usually not handled in the drug design process.

The NAOMI framework now serves as a basis for further developments at the Center for Bioinformatics since almost all cheminformatics applications require data from external sources. NAOMI is not limited to simple file conversions, but can be used to further analyze and modify molecular structures. An example of an application which makes use of the NAOMI framework and which can potentially enhances molecular

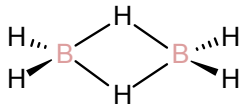


Figure 2.2: Molecular structure of diborane.

docking experiments, is the consistent handling of prototropic tautomers and protonation states. The handling of these different molecular states is described in more detail in the following section.

2.3 Prototropic Tautomers and Protonation States

Cheminformatics applications like molecular docking use input from various sources. Different sources do not only differ with respect to the provided file formats. They can also follow different standards for the representation of molecular structures. These standards commonly include the selection of standardized protonation states and prototropic tautomers. An example is the normalization of carboxylates, which are in general represented in their charged form. Six-membered conjugated rings are usually depicted with alternating single and double bonds, if possible. In the following, these different representations of molecules will be summarized under the term molecular states.

Molecular docking calculations are an example of a class of cheminformatics applications which can be sensitive to the particular positions of individual hydrogen atoms. This is the case since these applications usually involve the calculation of explicit hydrogen bonds. Structures which have been standardized in different ways can therefore lead to different results [27]. Another application for which molecular states are important is the registration of molecular structures in public databases. A solution to this issue is the generation of a canonical molecular state followed by the selection of reasonable number of representative molecular states. The number of molecular states of a compound can be high due to the combination of multiple molecular states of different functional groups. For many applications it is, however, favorable to register only one unique representative. This requires the canonical selection of a single molecular state independent of the input form.

To tackle these tasks, the valence state model of NAOMI has been used to describe molecular states and their relations. A molecular structure can be described by the valence states of its atoms and their connectivity. Molecular graphs with the same connectivity but alternative valence states will be called valence state combinations.

A valence state combination is valid if localized bond orders can be assigned in a way which is compatible to the valence states.

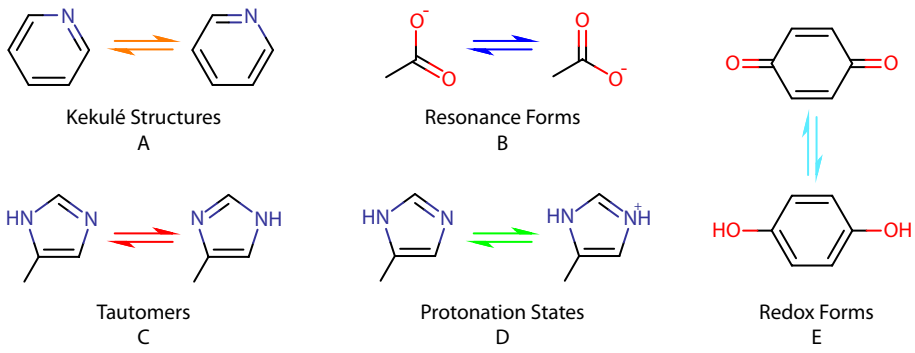


Figure 2.3: The valence states of the NAOMI model can be used to describe different molecular states. In different Kekulé structures (A) the corresponding atoms have identical valence states. In different resonance forms (B) a charge switches from one atom to another atom which is part of the same conjugated system. In tautomers (C), a hydrogen atom is transferred from one atom to another atom in the same conjugated system. Protonation states (D) involve the addition or substraction of single hydrogen atoms. Quinone (E) is the oxidized form of hydroquinone.

To identify different molecular states, substitutions of valence states are classified. Figure 2.3 shows a number of examples of distinct molecular states. In different Kekulé structures (2.3A) all atoms have identical valence states. Resonance forms (2.3B) involve the transfer of a charge between two atoms which are part of the same conjugated system. This transfer leads to a change of bond orders. In the shown example, the two oxygen atoms switch the valence states O100- and O200. Tautomeric forms (2.3C), do not have different charges but involve a transfer of a hydrogen between atoms of the same conjugated system. This results in a switch of single and double bonds for the affected atoms. The atom in a tautomer, which has a higher number of single bonds before the transition, will be called a tautomer donor in the following. After changing its valence state it has a lower number of single bonds and becomes the corresponding tautomer acceptor. Tautomers contain an identical number of tautomer acceptors and tautomer donors which undergo opposite transitions. In the depicted example, the nitrogen atoms switch the valence states N110 and N300. Tautomers and resonance forms each require an even number of atoms which change their states. Protonation states (2.3D) involve the addition or subtraction of a hydrogen which results in a change of the charge. In the shown example a nitrogen atom changes its valence state from N110 to N210+. The last example (2.3E) shows two redox forms of a molecule, which can be

transformed into each other by the addition or subtraction of two hydrogen atoms. For a more detailed description of the different molecular states and their handling by the NAOMI model, please refer to the corresponding publication [A3] which is attached to this dissertation in section A.5 of the appendix.

The description of molecular states within the NAOMI model was used to enumerate and canonize molecular structures using a branch and bound algorithm with a simple scoring scheme. The results of the generation of different protonation states and tautomers were compared to commercially and freely available software. The combination of tautomers and protonation states will be summarized under the term protomers in the following. The extension of the NAOMI framework was used to produce protomers for four different data sets including the ZINC clean leads with approximately 6 million compounds. For cases with more than one protomer, an average of 2.5 molecular states were generated. The best scored protomer according to NAOMI was identical to the initial representation in the ZINC data set in more than 83% of the cases. In about 16% of the cases the representation in the ZINC database was found in the set of molecular states with a score of at least 75% percent of the maximum score. Consequently, this extended set of protomers included the input structure in more than 99% of the cases. This is a strong indication of the quality of the produced molecular states. An average runtime of 0.45 ms was required to generate molecular states for each molecule of the data set on a standard PC*. Due to the quality of the resulting structures and the fast processing time, the presented approach is suitable to enumerate a set of meaningful molecular states. This can improve the performance of cheminformatics applications like molecular docking.

Pattern matching is commonly used in local approaches for the identification of known tautomeric transformations [56, 57]. There is always the risk of missing patterns if conjugated systems are too large to be adequately captured by small molecular patterns. Furthermore, it is a challenging task to avoid that multiple patterns match the same part of a molecule and lead to inconsistent results. Global approaches[58, 59] often enumerate a large number of possibly artificial molecular states.

NAOMI combines the advantages of these approaches. The procedures for the generation of valence bond structures can be described as a global approach, while the scoring scheme adds the chemical knowledge of local approaches. Additionally, NAOMI consistently handles different types of molecular states like resonance forms and ionization states. Despite these advantages of the presented extension of NAOMI, the concept is not suitable to calculate the energetically most favorable tautomer. It can, however, be easily adapted to serve different purposes.

*Intel Core i5-3570 CPU (4x 3.40 GHz) with 8 GB of main memory

In addition to the generation of a set of reasonable states, the generation of a canonical state was also implemented. The implementation was tested by first enumerating molecular states for the input structure followed by the generation of a canonical form for each of these states with NAOMI. The resulting canonical form was identical in all of the cases. The presented approach can therefore be used to canonize molecular states and to identify different states.

Chapter 3

Unique Ring Families

3.1 Introduction to Ring Perception Concepts

Ring perception is a key step in a number of cheminformatics application. As described in section 1.4, SMARTS expressions can be used to match atoms which are part of a specific number of rings. The number of rings refers to an arbitrary selected smallest set of smallest rings (SSSR) of the molecular graph. Besides matching SMARTS expressions, rings of molecular graphs are used to identify aromatic systems, calculate atomic 2D or 3D coordinates, and determine molecular scaffolds. A detailed introduction of ring perception concepts will be published in the chapter "Ring Perception" of the next edition of the "Handbook of Cheminformatics" [B2].

The calculation of an SSSR is only one of a number of different ring perception concepts, which are commonly used in cheminformatics. These ring perception concepts serve different purposes and have different advantages and disadvantages. In the following paragraph, a number of graph theoretical terms will be introduced [43]. Using these terms, the different ring perception concepts will be described in more detail. The terms will also be used in the results section of this thesis to describe a newly developed and advanced ring perception concept.

In computer science, a subgraph is called a cycle if each vertex has a degree of two. A connected cycle is called a ring. The size of a cycle C will be denoted as $|C|$ and is equal to its number of edges (bonds). The set of edges of a cycle will be written as $E(C)$. A basic idea of ring perception is the representation of rings by the incidence vectors of their edges. Cycles can be combined by calculating the symmetric difference ($\oplus$) of their edges' incidence vectors. This operation will be called addition of cycles. The result of the addition of two cycles is also a cycle. A set of cycles is called a cycle basis if all cycles of the graph can be calculated by the addition of a subset of its cycles.

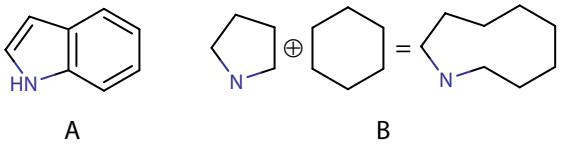


Figure 3.1: A: The molecular structure of indole can be interpreted as a molecular graph. B: The corresponding graph contains 3 cycles: a 5-cycle, a 6-cycle, and a 9-cycle. The symmetric difference of the incidence vectors of the 5-cycle and the 6-cycle results in the 9-cycle.

The size of a cycle basis is the sum of its cycles' sizes. A minimum cycle basis is a cycle basis with minimum size. Figure 3.1 shows the molecular graph of indole and its cycles. Each set of two of its cycles forms a cycle basis since the third cycle can be constructed by the symmetric difference of the incidence vectors of the other cycles. The 5-cycle and the 6-cycle form a minimum cycle basis and represent an SSSR.

An efficient algorithm to calculate an SSSR has been published [41]. The concept is sufficient for most molecules, but it can lead to inconsistent results for molecules containing multiple SSSRs. The molecular graph of cubane illustrates this problem (see Figure 3.2A). Cubane contains six different SSSRs each containing only five of the six 4-cycles. For each of the SSSRs, four atoms belong to three of its cycles while the remaining four atoms belong to only two of its cycles. Since an SSSR forms the basis of the ring property of the SMARTS language, the pattern [R3] would only match four of the eight equivalent atoms.

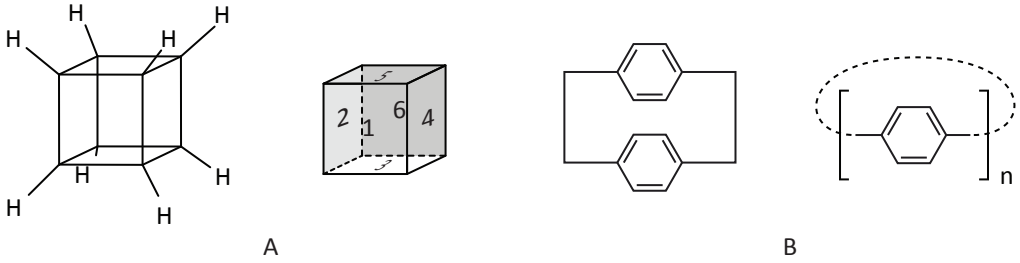


Figure 3.2: The molecular graph of cubane contains six 4-cycles (A). Molecular graphs having a structure as shown in B contain n 6-cycles and 2^n macrocycles.

A conceptually straightforward but computationally demanding concept of ring perception is the calculation of all rings (Ω). Ω can be calculated by an algorithm introduced by Hanser [60]. In contrast to the calculation of an SSSR, the results of the calculation of Ω are unique. Since the number of rings can increase exponentially with the number of atoms of a molecule, this approach is not always feasible in the context

of cheminformatics. An example of a concept which is focusing on the calculation of synthetically meaningful cycles is the essential set of essential rings (ESER) [61]. This concept lacks the complete description of a molecule's rings since it does not necessarily include a cycle basis. The relevant cycles (RCs) [62], which are also known as K rings, represent a set of cycles which circumvents these problems. RCs are defined as the union of all minimum cycle bases of a graph and completely describe a molecule's cycles. Furthermore, they are unique and their number is small for most molecules. For molecules which contain small para-bridged rings in large macrocycles, the number of RCs can grow exponentially with respect to the number of atoms. The disadvantage of this approach becomes obvious if RCs are calculated for para-bridged 6-cycles in macrocycles (see Figure 3.2B). In Vismara's algorithm [62] for the calculation of the exponential number of RCs a polynomial number of relevant cycle prototypes (RCPs) is calculated in an intermediate step. The set of RCPs is not unique, but it completely describes a molecule's rings since it represents a cycle basis.

3.2 Motivation to Develop a New Concept

As described in the last section, ring perception is crucial in a number of different tasks in cheminformatics. Therefore, the implementation of a ring perception algorithm was a key step in the development of the software NAOMI. None of the approaches, which are listed in the last section, combine the three important properties of chemical ring descriptions, namely being unique, chemically meaningful, and efficient to compute.

The calculation of an arbitrary SSSR is not a suitable approach for a tool aiming at consistently modeling molecules. An approach which circumvents this problem is the calculation of RCs as described by Vismara [62]. While it yields identical results for molecules containing only a single SSSR, the number of RCs is unique even for complex molecules. At the beginning of the implementation of the NAOMI tool, the potentially large number of RCs for a molecule was not a significant problem, since most molecules only contain ring systems of low complexity. To assess the influence of the calculation of RCs on the runtime of every-day tasks, the RCs were calculated for the 32 593 299 molecules of the PubChem dataset. 31 706 629 compounds (97% of the data set) contained at least one ring. For these molecules, the RCs could be calculated in less than one ms in 31 658 230 (99.8%) of the cases on a standard PC*. Only for 6 molecules the calculation of RCs required more than 1s with a maximum of 3s.

While this seems to be acceptable at a first glance, some procedures in the NAOMI framework perform further calculations for each ring of a molecule. Aromaticity, for

*Intel Core2 Quad Q9550 CPU (4x 2.83 GHz) with 4 GB of main memory; single thread only

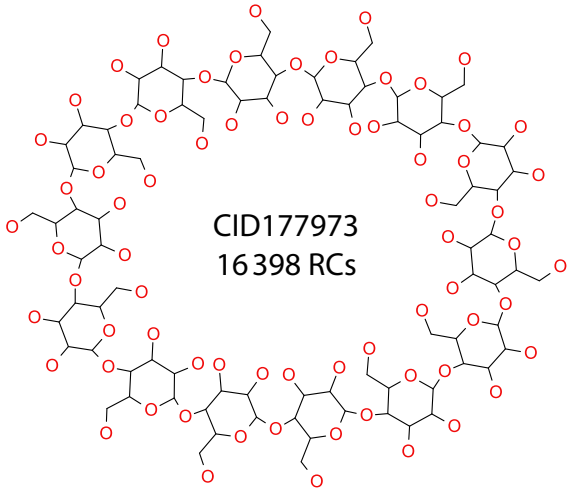


Figure 3.3: The molecular structure of CID177973 contains 16 398 RCs.

example, is determined by analyzing the valence states of each ring of a molecule. Furthermore, each ring is individually analyzed if 3D atomic coordinates have to be calculated. This can result in high runtimes for molecules containing a large number of RCs. An example is molecule CID177973 from the PubChem database, which is shown in Figure 3.3. The NAOMI framework usually requires only 0.5 milliseconds for the initialization of a molecule from an input file. The processing of CID177973 is significantly slower and can require several minutes on a standard PC. The compound has a similar structure as described in Figure 3.2B and contains $14 + 2^{14} = 16\,398$ RCs.

Due to these issues, neither the calculation of a single SSSR nor the calculation of all RCs represents a suitable approach for the calculation of molecular rings. Thus, a new model had to be developed, which shares the advantages of both approaches without sharing their disadvantages. This resulted in the development of the Unique Ring Families which will be introduced in the next paragraph.

3.3 Introduction to Unique Ring Families

Unique Ring Families (URFs) are defined on the basis of RCs. Two RCs C_1 and C_2 of a molecular graph G are URF-pair-related if the following three conditions hold:

1. $|C_1| = |C_2|$
2. $E(C_1) \cap E(C_2) \neq \emptyset$
3. It exists a set S of strictly smaller rings in G such that $C_1 \oplus (\bigoplus_{c \in S} c) = C_2$

A URF is defined as the transitive closure of the URF-pair relation. Consequently, all cycles of a URF have the same size and can be constructed by the addition of an arbitrary cycle of the same URF and a subset of smaller cycles. Since this definition of URFs is based on the RCs of a molecular graph, the URFs can be calculated in two steps. At first, all RCs are calculated as described by Vismara. Secondly, the three above mentioned properties are checked and the RCs are assigned to URFs. This approach requires the calculation of the exponential number of RCs followed by an analysis of every 2-pair of RCs. Hence, this approach is computationally demanding and the required time for the calculation of URFs would increase exponentially with the number of RCs.

To improve the calculation of URFs, an algorithm has been developed, which determines the number of URFs and the edges belonging to each URF on the basis of the polynomial number of RCPs. While the polynomial runtime complexity of the calculation of URFs is important, the most important advantage of URFs is the intuitive and unique description of a molecule's rings. Cubane, for example, contains six URFs. For cyclophane-like structures, which only contain n small para-bridged cycles (see Figure 3.2B), $n+1$ URFs are perceived. The structure shown in Figure 3.3 includes a total of 15 URFs. Out of these, 14 URFs contain 6 edges and represent 6-cycles. The remaining URF contains 112 edges and represents the large macrocycle including the smaller 6-cycles.

URFs combine three important properties for the description of molecular ring topologies: (1) They provide a unique description of a molecule's rings. (2) URFs can be efficiently calculated (in polynomial runtime). (3) The description of a molecule's rings by URFs is intuitive. Due to these advantages, URFs are suitable to become a standard concept for the description of molecular ring topologies.

Despite these advantages, there is also a disadvantage of the presented concept. The number of rings of an SSSR is equal to the cyclomatic number. Let $E(G)$ be the number of edges (bonds) of a connected molecular graph and $V(G)$ the number of vertices (atoms). The cyclomatic number r can be calculated using the following equation:

$$r = E(G) - V(G) + 1 \tag{3.1}$$

Currently, there is no such formula to calculate the number of URFs. Nevertheless, the number of URFs can be estimated since it is greater than or equal to r and smaller than or equal to the number of RCPs. The number of RCPs can be estimated according to theorem 4 of Vismara's publication [62]. Let u be the number of URFs of a connected

molecular graph. Then u can be estimated with the following equation:

$$r \leq u \leq rV(G) + E(G)^2 \tag{3.2}$$

A more detailed description of this ring perception concept can be found in the corresponding publication [A4] which is attached to this dissertation in appendix A.3.

Chapter 4

In silico Design of eIF-5A Activation Inhibitors

4.1 Deoxyhypusine Synthase (DHS)

The previous sections introduce different aspects of cheminformatics, which form the basis of *in silico* drug design. In the context of this thesis cheminformatics tools were used to design inhibitors of the human protein DHS. The DHS will be described in more detail in the following.

Human DHS is an enzyme which is involved in the activation of the human elongation initiation factor 5A (eIF-5A). Two modifications to lysine 50 (lys-50) of eIF-5A are required for its activation [9, 63] (see Figure 4.1). The first reaction is catalyzed by DHS and involves the transfer of an amino-butyl residue from spermidine to lys-50 of eIF-5A. The resulting deoxyhypusine is further modified by the Deoxyhypusine Hydroxylase (DOHH) to form the unusual amino acid hypusine. This thesis focuses on the inhibition of eIF-5A activation via inhibition of DHS. Efforts to identify inhibitors of the DOHH are not discussed in this thesis.

Activated eIF-5A is a cellular cofactor of the HIV Rev protein [64–66]. The Rev protein is required for the export of unspliced viral mRNA from the nucleoplasm to the cytoplasm [67, 68]. Consequently, DHS is a required host cell factor for the HIV replication and represents a promising target for the development of new anti-HIV therapies. Since it is a host cell factor, DHS is not directly affected by viral mutations. The development of viral resistances against DHS inhibitors is therefore unlikely. Inhibiting a host cell factor bares the risk of negative side effects on the host. In addition to its role in the replication of HIV, the hypusination of eIF-5A is essential for cellular

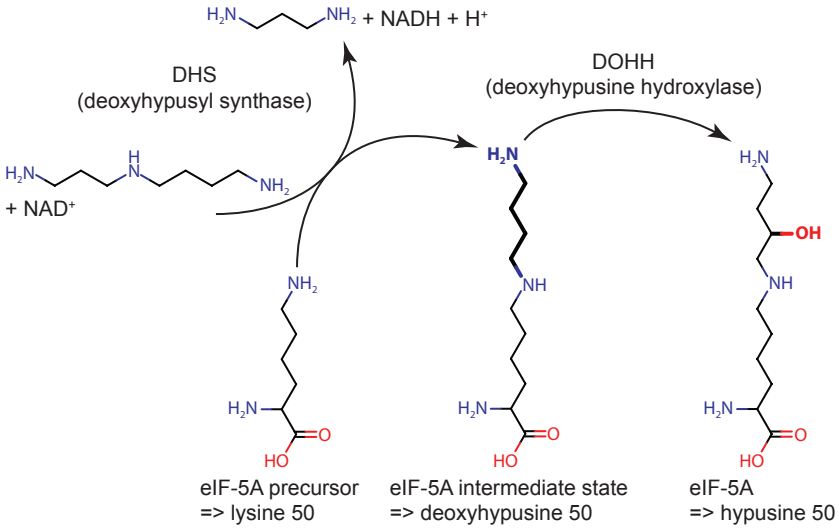


Figure 4.1: The activation of eIF-5A consists of two steps. In the first step, which is catalyzed by DHS, an aminobutyl residue is transferred from spermidine to lysine 50 of eIF-5A. In the second step, the deoxyhypusine is oxidized. This reaction is catalyzed by the DOHH.

proliferation and plays a role in the progression of CML [13]. Furthermore, it has been published that activated eIF-5A is required for the elongation step of translation [69].

Figure 4.2 shows the structure of DHS as published under PDB code 1RQD [7]. DHS is a tetrameric enzyme (see Figure 4.2A) with four ligand binding sites at the contact surfaces of the monomers (see Figures 4.2B and 4.2C). The two most potent known inhibitors of DHS are N-1-guanyl-1,7-diaminoheptane (GC_7) [70] and CNI-1493 [71] (see Figure 4.3). The compound CNI-1493 is also known as semapimod. GC_7 has structural similarity to spermidine which is the natural substrate of DHS. A number of GC_7 - and spermidine analogs [70] have already been tested for their inhibitory activity against DHS. Furthermore, GC_7 has been co-crystallized with NAD in the DHS binding site [7] (see Figure 4.2B). Analogs of spermidine like 1,8-diaminooctane and GC_7 show dose-dependent inhibition of Rev and suppression of HIV replication [72, 73]. Despite these promising experiments, GC_7 is only of limited use as a drug candidate due to possible interference with the natural metabolism of spermidine. Since GC_7 is a potent inhibitor of DHS and since its binding site is known, it can serve as a lead structure for the development of new drugs.

CNI-1493 is an inhibitor of DHS which was tested in two clinical phase 2 trials against Crohn's disease [74–76]. Due to its size and number of charges, CNI-1493 is not an optimal drug candidate. Nonetheless, it can serve as a lead structure for ligand

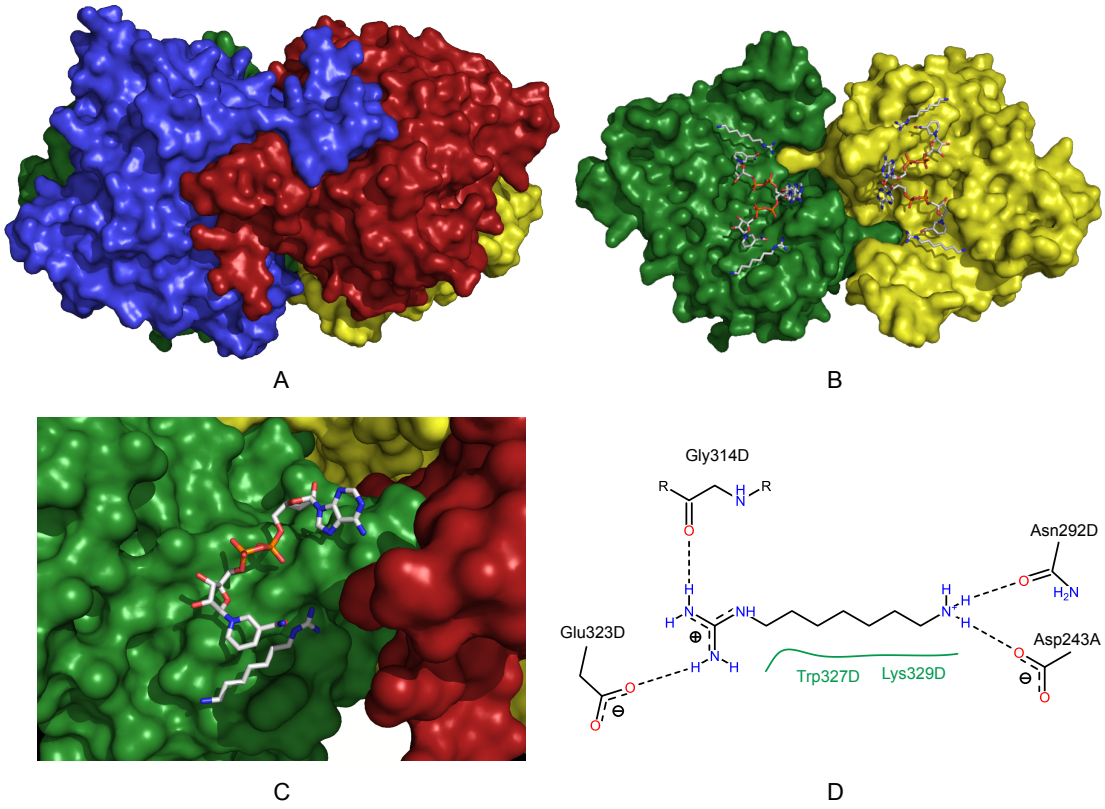


Figure 4.2: A: DHS is a tetrameric protein B: There are four binding sites at the contact surfaces of the monomers. C: GC7 and NAD bind close to each other in the binding site. D: GC_7 interacts with different amino acids of both protein chains which form each of the binding sites.

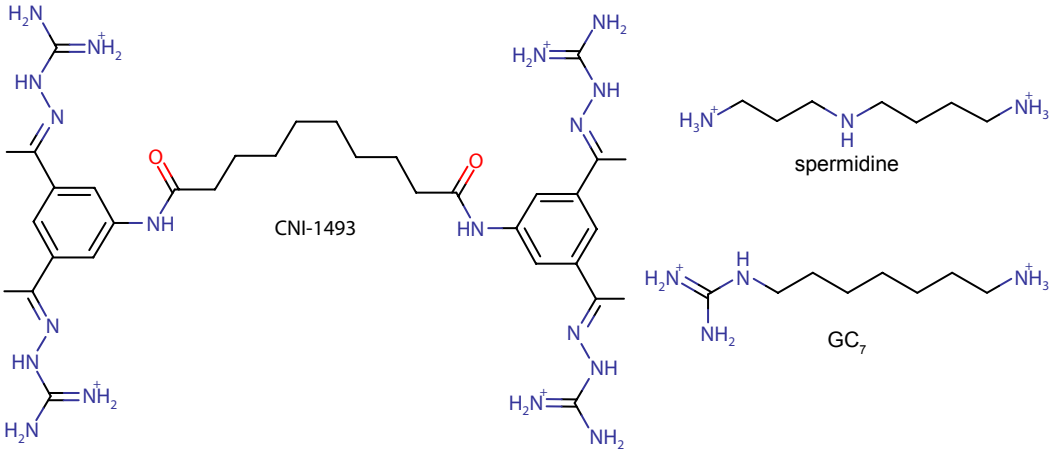


Figure 4.3: The natural substrate of DHS is spermidine. Two potent inhibitors are GC_7 and CNI-1493. GC_7 has a high similarity to spermidine and binds to DHS at the spermidine binding site. The binding site of CNI-1493 was unknown at the beginning of the work for this thesis.

optimization approaches. In contrast to GC_7 the binding site and mode of action of CNI-1493 were still unknown.

4.2 Initial Screening Efforts

The inhibition of DHS represents a promising strategy to develop novel HIV - and CML therapies as described in section 4.1. As a first step of the design of new DHS inhibitors, the clean leads subset of the ZINC data set [39] was screened *in silico*. The ZINC database contains molecular structures of compounds which are described as being commercially available. Compounds from the ZINC database can therefore, in principle, be directly ordered from compound vendors if they are identified as hit compounds in a virtual screen. This allows small lead times for following experiments. Since the database includes millions of molecular structures, the first virtual screening was performed with the TrixX [77] software, which is specifically designed for the fast screening of large data sets into multiple receptors. The disadvantage of the software is the required preprocessing of the input data. For each molecule a number of conformers have to be generated. This requires time and disc space for storing the data. Since the ZINC dataset was already preprocessed at the ZBH, this was not an issue.

Initial hits of the TrixX screening were again analyzed using LeadIT followed by the even more accurate and time consuming HYDE [24] scoring function. The compounds which were selected for further analysis are shown in Figure 4.4. In addition to the

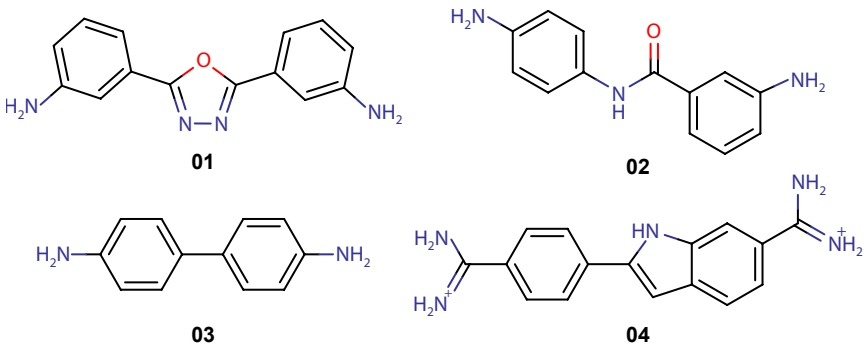


Figure 4.4: Compounds selected for *in vitro* testing according to virtual screening of the ZINC database.

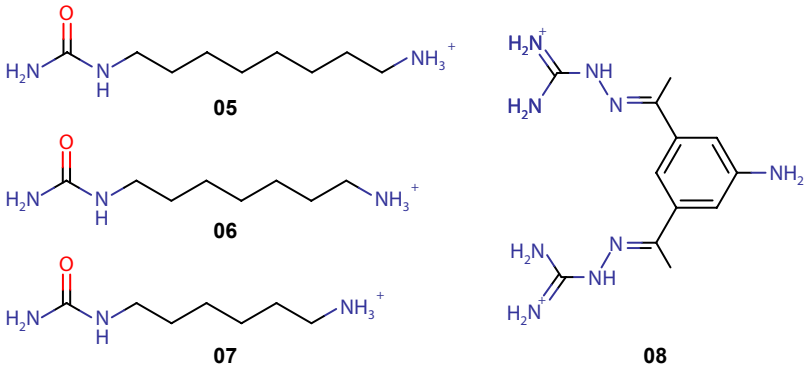


Figure 4.5: Modifications of GC_7 (compounds **05**, **06**, and **07**) and a substructure of CNI-1493 (compound **08**) were selected for further testing.

screening of the ZINC database, the known inhibitors GC_7 and CNI-1493 were modified. The guanidinium group of GC_7 was replaced by a urea group. The length of the carbon chain which connects the primary amine with the urea group was varied between 6 and 8 atoms. The corresponding compounds are shown in Figure 4.5 (compounds **05** - **08**). In contrast to the guanidinium group, the urea group is not charged which could potentially increase the ability of these molecules to pass cell membranes. Analyzing the inhibitory effects of these molecules was a first test to replace the charged guanidinium group. Compound **08** is a substructure of CNI-1493, which was selected to test the hypothesis that one part of the symmetric CNI-1493 is sufficient to inhibit DHS.

Compounds **01** to **08** were tested for anti-retroviral activity in cell culture as well as inhibition of DHS in an enzymatic assay by the group of Prof. Dr. Hauber from the Heinrich Pette Institute [78]. Furthermore, the cell toxicity was analyzed. First tests of compounds **01** to **04** showed good antiviral activities for compounds **01**, **02**, and

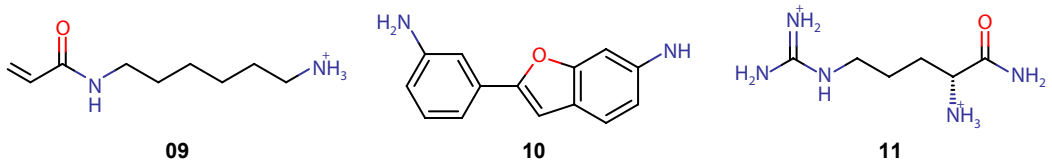


Figure 4.6: Compounds selected for further analysis after initial testing of compounds **01** - **08**.

04 while showing only low toxicity. Following these initial results, further compounds were selected for the analysis. These compounds are shown in Figure 4.6. In a second set of experiments, inhibition of DHS was only observed for compounds **04** and **09**. At a concentration of 80 μM compound **04** inhibited DHS activity by 11% and compound **09** showed 14% inhibition of DHS. Additionally, the inhibition of DHS by compound **04** was observed to be dose-dependent. The activity of DHS was reduced by 99% at a concentration of 360 μM of compound **04**. Due to these properties, **04** served as a lead structure for further experiments.

Compound **04** which is also known as DAPI (4',6-Diamidin-2-phenylindol) is commercially available as a fluorescent dye. DAPI can intercalate into DNA and it can bind to the minor groove [79]. It is commonly used for histological staining. These properties of compound **04** explain the observed cell toxicity. Since DAPI is classified as an irritant and since the observed inhibition of DHS was limited, compound **04** had to be further optimized.

4.3 Improved Analogs of Lead Structure

Due to the cell toxicity and the limited inhibitory activity of compound **04**, improved analogs had to be developed. As tested during the design of GC_7, the combination of a guanidinium group and a primary amine showed the strongest interaction with the terminal parts of the DHS binding site [70]. In addition to these substructures, amidines were also considered during the optimization of compound **04**, to allow the identification of new lead structures. Besides including these substructures, the analogs of compound **04** were also designed to have an increased flexibility. Compound **04** contains a conjugated electronic system which stabilizes a planar conformation. This enables the compound to intercalate into DNA. To limit the risk of intercalation and avoid the corresponding side effects, flexible linkers were included into the newly designed analogs of compound **04**. A further advantage of this flexibility is the ability of the corresponding compounds to adapt to the DHS binding site, which can result

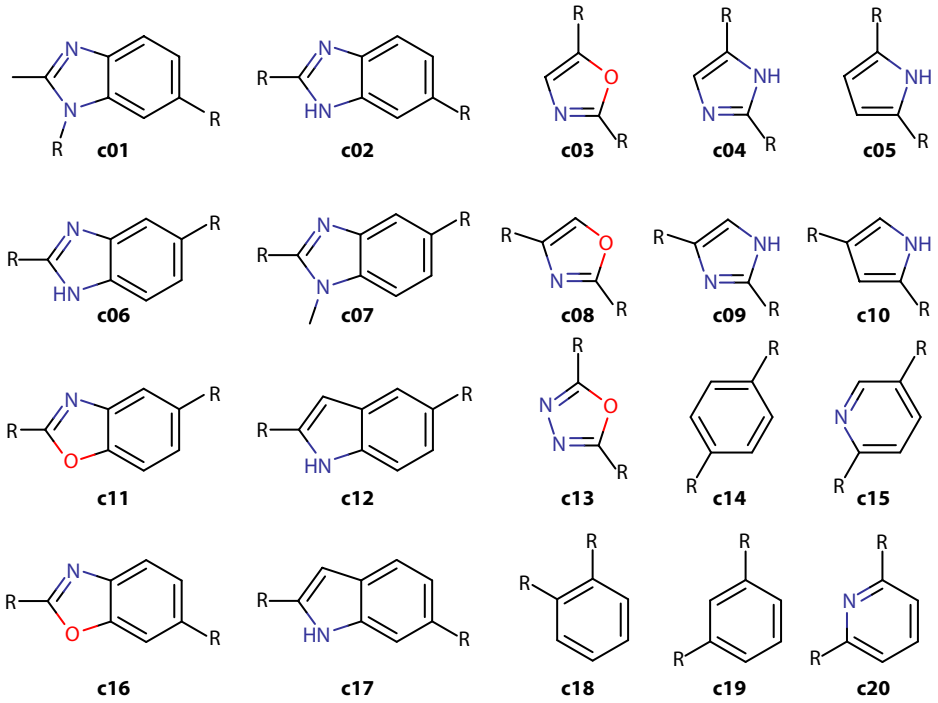


Figure 4.7: Fragments selected to interact with TRP-327 and HIS-288 of the DHS binding site.

in geometrically optimized interactions and therefore increased binding affinities. A possible disadvantage is the increased loss of entropy during protein binding due to the increased flexibility. To optimize the inhibitory effect of newly synthesized inhibitors compared to GC_7, a scaffold hopping [80] approach was pursued. Different variations of the molecular core were analyzed. Compounds were selected to allow π-π stacking interactions with tryptophane 327 of the DHS binding site.

Each molecular structure was formed by one core fragment (see Figure 4.7) with two connected anchor fragments (see Figure 4.8). This lead to a total of 875 unique molecular structures which were generated *in silico* and docked into the DHS binding site using the software LeadIT. The four structures with the best docking scores are shown in Figure 4.9. Out of these, compound **12** demonstrated dose-depend inhibition of DHS and inhibited HIV replication by 14% at a concentration of 2μM in cell culture. For further information on the design, synthesis, and testing of these inhibitors please see the corresponding publication [A5], which is attached to this thesis in chapter A.6 of the appendix. Due to the observed activity of compound **12** and the inhibition of HIV in cell culture, the compound and a number of derivatives were patented [E1]. To design compounds with improved activity against DHS and HIV further studies

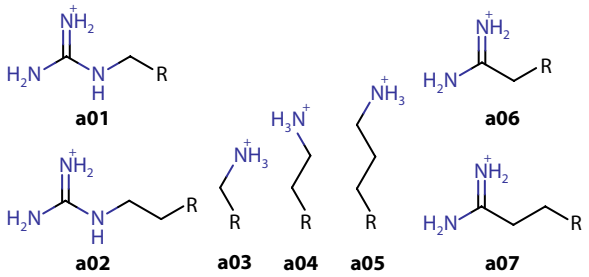


Figure 4.8: Fragments selected to interact with terminal parts of the DHS binding site.

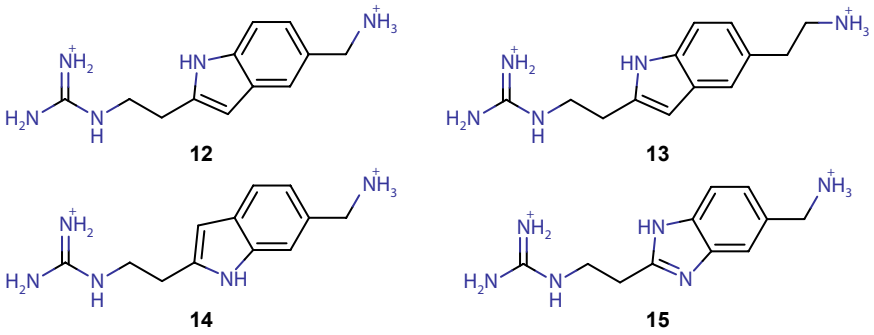


Figure 4.9: Compounds selected from fragment space according to docking results.

are required which involve additional cycles of *in silico* design, chemical synthesis, and activity determination. The synthesis of derivatives of compound **12** is currently in progress and the resulting compounds will soon be tested for their activity against DHS and their inhibition of HIV replication.

4.4 Development of CNI-1493 Analogs

In addition to the design of DHS inhibitors which are based on the known DHS inhibitor GC_7 a second approach focused on the development CNI-1493 analogs. As described in section 4.2 compound **08** is a substructure of CNI-1493 and does not inhibit DHS. A crystal structure which was produced by the group of Prof. Dr. Hilgenfeld and which has not yet been published, showed the compound CNI-1493 bound to the surface of DHS in a stacked conformation (see Figure 4.10).

The observed binding site of CNI-1493 is close to the entrance of the spermidine binding site and involves amino acids of two chains of DHS. This lead to the hypothesis that the binding of CNI-1493 blocks the entrance to the spermidine binding site and thereby inhibits the activation of eIF-5A. To analyze if the stacked conformation of

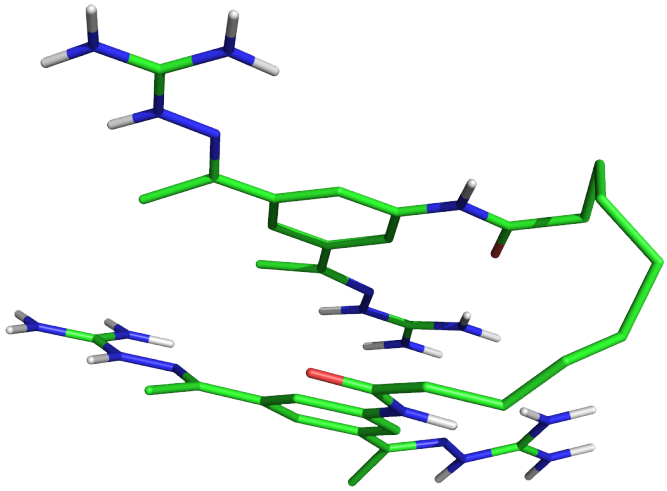


Figure 4.10: Conformation of CNI-1493 as observed by the group of Prof. Dr. Hilgenfeld.

CNI-1493 is required for its inhibitory activity and to find an optimal length of the alkyl chain, different analogs of CNI-1493 were synthesized and tested. An overview of the analogs is given in Figure 4.11.

For compound **21** a stronger inhibition of DHS compared to CNI-1493 was observed at concentrations below 4 μM. Therefore, it was selected for analysis of its effects against CML. The corresponding experiments are described in the publication [A6], which is attached to this thesis in appendix A.2. In the publication, compound **21** is referred to as DHSI-15. In summary, compound **21** demonstrated strong anti-proliferative effects on BCR-ABL positive and negative cells. Due to the effect on BCR-ABL negative cells, it is unlikely to be useful as a treatment against CML.

In addition to compound **21**, compound **26** showed a similar inhibition of DHS compared to CNI-1493. All other analogs showed a weaker inhibition. Due to the conjugated electronic system in the central part of compound **26**, a stacked conformation of compound **26** is highly unlikely. Its inhibitory activity therefore indicates that the stacked conformation of CNI-1493 is not required for the inhibition of DHS. Consequently, further experiments are needed to verify the active conformation and the binding site of CNI-1493. A publication which summarizes these results and includes synthetic protocols for all CNI analogs, which are shown in Figure 4.11, is currently in preparation for submission. The initial submission failed since the crystal structure shown in Figure 4.10 has not yet been published by Prof. Dr. Hilgenfeld and his group. As soon as the crystal structure is publicly available, the results of the testing of analogs of CNI-1493 will also be submitted for publication.

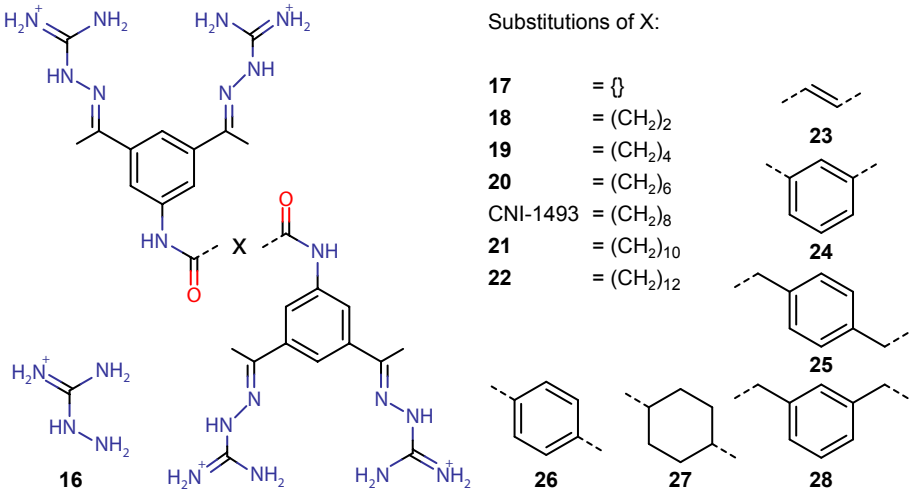


Figure 4.11: Variations of CNI-1493 which were selected for *in vitro* testing of inhibitory activity against DHS.

Chapter 5

Conclusion and Outlook

The first major contribution of this thesis to the field of chemoinformatics is the co-development of NAOMI and its extensions to process PDB files and to generate different molecular states. These developments have accelerated the research at the ZBH significantly. The NAOMI framework allows a consistent and efficient handling of valence bond structures from different file formats and forms the basis for a number of current developments in the group for "Algorithmic Molecular Design" at the ZBH. MONA [81], LOFT [82, 83], and the ChemBioNavigator [84] are examples of published tools which are based on NAOMI. In addition to the direct impact on the work at the ZBH, the publication which introduced NAOMI also describes validation procedures for cheminformatics tools in general. These can help developers to validate and improve chemoinformatics applications.

The NAOMI framework could be used as a basis for further cheminformatics applications. An example is the currently available generation of a set of sensible molecular states. This feature of the NAOMI framework could significantly enhance the results of a docking software like LeadIT if it becomes properly integrated. Force fields which currently use the Sybyl Types for parametrization could possibly be improved by instead using the chemically meaningful atom types of the NAOMI model. Since the NAOMI framework was developed according to the needs at the Center for Bioinformatics, only the most common chemical file formats are currently supported. Further improvements could therefore include the support of additional input and output formats. INCHI [85, 86] codes are an example of a currently unsupported but widely used way to represent molecular structures. In addition to the support of already existing file formats, a new file format could be developed on the basis of the NAOMI framework and specialized on the description of small molecules.

5. CONCLUSION AND OUTLOOK

The second contribution of this thesis to the area of Cheminformatics is the introduction of URFs. The concept of URFs combines three important properties for the description of molecular ring topologies. URFs are unique, can be efficiently calculated in polynomial time, and provide an intuitive description of molecular ring topologies. Thereby, URFs tackle the problems of the ambiguous SSSR rings, which are still widely used. The corresponding publication [A4] has raised the awareness of the SSSR's problems and brought alternative ring perception concepts to the attention of the scientific community [87]. Despite the fact that the NAOMI framework supports the calculation of URFs, it currently still calculates the RCs for each molecule (see appendix C) to determine aromaticity and stereo information. If URFs would be used for this purpose instead of the RCs, the runtime for the initialization of molecules with complex ring systems could be significantly reduced. Furthermore, the aromaticity detection could be changed to identify URFs as aromatic, if each bond of the URF is part of at least one ring that is aromatic. This would avoid the need to identify a path with alternating single and double bonds for large macrocycles. Another application for URFs is the calculation of 3D atomic coordinates.

While cheminformatics is an area of research by itself, it also serves the purpose of supporting the development of new drugs. In the presented work, a variety of cheminformatics tools have been applied to develop new inhibitors against human DHS. The result is a new dose-dependent inhibitor (see compound **12** in Figure 4.9) of DHS, which inhibits HIV replication in cell culture without cytotoxic side effects at active concentrations. The design of this inhibitor is the third significant contribution of the work presented in this thesis. Compound **12** is a lead structure which can serve as a starting point for further lead optimization studies. It inhibits DHS and has validated that DHS is a promising target for future HIV treatments. Since DHS is a host cell factor, the danger of upcoming resistances against drugs targeting this enzyme is minimal. Therefore, inhibitors of human DHS have the potential to support modern therapies against HIV infections. In addition to the development of compound **12**, which was based on the known inhibitor GC_7, analogs of CNI-1493 were also tested for their inhibition of DHS. Based on a crystal structure which included CNI-1493 in a stacked confirmation bound to the surface of DHS (see Figure 4.10), different analogs of CNI-1493 were designed to test if the stacked conformation is required. Compound **21** showed a stronger inhibition of DHS compared to CNI-1493 and could potentially adapt a similarly stacked conformation. It was furthermore tested against CML and showed strong anti-proliferative effects against BCR-ABL positive and BCR-ABL negative cells. The anti-proliferative effect against BCR-ABL negative cells represents a potential problem of DHS as a target and should be further investigated. The mostly

planar compound **26** also showed inhibitory activity against DHS. The stacked conformations of CNI-1493 bound to DHS as observed by Prof. Hilgenfeld seems not to be a requirement for the binding to DHS. This could be explained by multiple binding modes. Further analogs of the developed inhibitors are currently being synthesized and will soon be tested for anti-retroviral activity and inhibition of DHS *in vitro*.

In summary, the work which is presented in this thesis introduced significant improvements to a number of aspects of cheminformatics and has demonstrated the value of cheminformatics as part of the drug development process.

Chapter 6

Bibliography (other authors)

(1) Center for Bioinformatics - Computational Molecular Design., accessed June 28th, 2014, `http://www.zbh.uni-hamburg.de/en/research/computational-molecular-design.html`.

(2) Center for Bioinformatics., accessed June 28th, 2014, `http://www.zbh.uni-hamburg.de/en`.

(3) Willett, P. Chemoinformatics: a history. *WIREs Comput. Mol. Sci.* **2011**, *1*, 46–56.

(4) Chen, W. L. Chemoinformatics: Past, Present, and Future. *J. Chem. Inf. Model.* **2006**, *46*, 2230–2255.

(5) Bundesministerium für Bildung und Forschung., accessed June 28th, 2014, `http://www.bmbf.de/`.

(6) Rupp, B., *Biomolecular Crystallography: Principles, Practice and Application to Structural Biology*; Taylor and Francis Ltd.: 2009.

(7) Umland, T. C.; Wolff, E. C.; Park, M. H.; Davies, D. R. A New Crystal Structure of Deoxyhypusine Synthase Reveals the Configuration of the Active Enzyme and of an Enzyme-NAD-Inhibitor Ternary Complex. *J. Biol. Chem.* **2004**, *279*, 28697–28705.

(8) Volberding, P.; Sande, M.; Lange, J.; Greene, W.; Gallant, J., *Global HIV/AIDS Medicine*; Saunders: 2007.

(9) Park, M. H. The post-translational synthesis of a polyamine-derived amino acid, hypusine, in the eukaryiotic translation initiation factor 5A (eIF5A). *J. Biochem.* **2006**, *139*, 161–169.

6. BIBLIOGRAPHY (OTHER AUTHORS)

(10) Cortes, J. E.; Talpaz, M.; Kantarjian, H. Chronic myelogenous leukemia: A review. *Am. J. Med.* **2014**, *100*, 555–570.

(11) Fausel, C. Targeted chronic myeloid leukemia therapy: seeking a cure. *J. Manag. Care. Pharm.* **2007**, *13*, 8–11.

(12) Nowell, P. C. Discovery of the Philadelphia chromosome: a personal perspective. *J. Clin. Invest.* **2007**, *117*, 2033–2035.

(13) Balabanov, S.; Gontarewicz, S.; Ziegler, P.; Hartmann, U.; Kammer, W.; Copland, M.; Brassat, U.; Priemer, M.; Hauber, I.; Wilhelm, T.; Schwarz, G.; Kanz, L.; Bokemeyer, C.; Hauber, J.; Holyoake, T. L.; Nordheim, A.; Brümmendorf, T. H. Hypusination of eukaryotic initiation factor 5A (eIF5A): a novel therapeutic target in BCR-ABL–positive leukemias identified by a proteomics approach. *Blood* **2006**, *109*, 1701–1711.

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

(15) Johnson, M. A.; Maggiora, G. M., *Concepts and Applications of Molecular Similarity*; Taylor and Francis Ltd.: 1990.

(16) Hansch, C. Quantitative approach to biochemical structure-activity relationships. *Accounts Chem. Res.* **1969**, *2*, 232–239.

(17) Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **1964**, *7*, 395–399.

(18) Maggiora, G. M. On Outliers and Activity Cliffs – Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.

(19) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3–50.

(20) Anderson, A. C. The Process of Structure-Based Drug Design. *Chem. Biol.* **2014**, *10*, 787–797.

(21) Brooijmans, N.; Kuntz, I. D. Molecular Recognition and Docking Algorithms. *Annu. Rev. Bioph. Biom.* **2003**, *32*, 335–373.

(22) LeadIT version 2.0.2, BioSolveIT GmbH, accessed June 28th, 2014, `http://www.biosolveit.de/LeadIT`.

(23) Helms, V., *Protein-Ligand Interactions: From Molecular Recognition to Drug Design*; WILEY-VCH Verlag: 2005.

(24) Reulecke, I.; Lange, G.; Albrecht, J.; Klein, R.; Rarey, M. Towards an Integrated Description of Hydrogen Bonding and Dehydration: Decreasing False Positives in Virtual Screening with the HYDE Scoring Function. *ChemMedChem* **2008**, *3*, 885–897.

(25) IUPAC Gold Book - tautomerism., accessed June 28th, 2014, `http://goldbook.iupac.org/T06252.html`.

(26) Sayle, R. So you think you understand tautomerism? *J. Comput. Aided Mol. Des.* **2010**, *24*, 485–496.

(27) Ten Brink, T.; Exner, T. Influence of Protonation, Tautomeric, and Stereoisomeric States on Protein-Ligand Docking Results. *J. Chem. Inf. Model.* **2009**, *49*, 1535–1546.

(28) Cheng, F.; Li, W.; Liu, G.; Tang, Y. In Silico ADMET Prediction: Recent Advances, Current Challenges and Future Trends. *Curr. Top. Med. Chem* **2013**, *13*, 1273–1289.

(29) Selick, H. E.; Beresford, A. P.; Tarbit, M. H. The emerging importance of predictive ADME simulation in drug discovery. *Drug Discov. Today* **2002**, *7*, 109–116.

(30) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **2010**, *9*, 203–214.

(31) Sangster, J., *Octanol-Water Partition Coefficients: Fundamentals and Physical Chemistry*; WILEY-VCH Verlag: 1997.

(32) Kellogg, G. E.; Abraham, D. J. Hydrophobicity: is LogPo/w more than the sum of its parts? *Eur. J. Med. Chem.* **2000**, *35*, 651–661.

(33) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3–26.

(34) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comp. Sci.* **2001**, *41*, 1308–1315.

6. BIBLIOGRAPHY (OTHER AUTHORS)

(35) Daylight Theory Manual 4.9., accessed June 28th, 2014, `http://www.daylight.com/dayhtml/doc/theory/index.pdf`.

(36) Todeschini, R.; Consonni, V., *Molecular Descriptors for Chemoinformatics*; Wiley-VCH Verlag GmbH: 2009.

(37) Schomburg, K.; Ehrlich, H.-C.; Stierand, K.; Rarey, M. From Structure Diagrams to Visual Chemical Patterns. *J. Chem. Inf. Model.* **2010**, *50*, 1529–1535.

(38) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719–2740.

(39) Irwin, J. J.; Shoichet, B. K. ZINC - a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

(40) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, 623–633.

(41) Zamora, A. An Algorithm for Finding the Smallest Set of Smallest Rings. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 40–43.

(42) Lewis, G. N. The Atom and the Molecule. *J. Am. Chem. Soc.* **1916**, *38*, 762–785.

(43) Diestel, R., *Graph Theory*; Springer-Verlag: 2010.

(44) Accelrys SDF File Format., accessed June 28th, 2014, `http://accelrys.com/products/informatics/cheminformatics/ctfile-formats/no-fee.php`.

(45) Tripos Mol2 File Format., accessed June 28th, 2014, `http://tripos.com/data/support/mol2.pdf`.

(46) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comp. Sci.* **1988**, *28*, 31–36.

(47) Ruske, W. August Kekulé und die Entwicklung der chemischen Strukturtheorie. *Naturwissenschaften* **1965**, *52*, 485–489.

(48) Weininger, D.; Weininger, A.; Weininger, J. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.

(49) Babel., accessed June 28th, 2014, http://www.tripos.com/mol2/atom_types.html.

(50) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(51) DUD - A Directory of Useful Decoys., accessed June 28th, 2014, http://dud.docking.org/r2/.

(52) CORINA - Fast Generation of High-Quality 3D Molecular Models., version 3.48, accessed June 28th, 2014, http://www.molecular-networks.com/products/corina.

(53) MOE., version 2010.10, accessed June 28th, 2014, http://www.chemcomp.com/software.htm.

(54) Babel., accessed June 28th, 2014, http://www.eyesopen.com/docs/babel/current/pdf/BABEL.pdf.

(55) Mayer, I. Bond orders in three-centre bonds: an analytical investigation into the electronic structure of diborane and the three-centre four-electron bonds of hypervalent sulphur. *J. Mol. Struc.-THEOCHEM* **1989**, *186*, 43–52.

(56) Milletti, F.; Storchi, L.; Sforna, G.; Cross, S.; Cruciani, G. Tautomer enumeration and stability prediction for virtual screening on large chemical databases. *J. Chem. Inf. Model.* **2009**, *49*, 68–75.

(57) Kochev, N. T.; Paskaleva, V. H.; Jeliazkova, N. Ambit-Tautomer: An Open Source Tool for Tautomer Generation. *Mol. Inf.* **2013**, *32*, 481–504.

(58) Haranczyk, M.; Gutowski, M. Quantum Mechanical Energy-Based Screening of Combinatorially Generated Library of Tautomers. TauTGen: A Tautomer Generator Program. *J. Chem. Inf. Model.* **2007**, *47*, 686–694.

(59) Thalheim, T.; Vollmer, A.; Ebert, R.-U.; Kühne, R.; Schüürmann, G. Tautomer Identification and Tautomer Structure Generation Based on the InChI Code. *J. Chem. Inf. Model.* **2010**, *50*, 1223–1232.

(60) Hanser, T.; Jauffret, P.; Kaufmann, G. A New Algorithm for Exhaustive Ring Perception in a Molecular Graph. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1146–1152.

(61) Fujita, S. A new algorithm for selection of synthetically important rings. The essential set of essential rings for organic structures. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 78–82.

6. BIBLIOGRAPHY (OTHER AUTHORS)

(62) Vismara, P. Union of all the minimum cycle bases of a graph. *Electron. J. Comb.* **1997**, *4*, 1–15.

(63) Wolff, E. C.; Kang, K. R.; Kim, Y. S.; Park, M. H. Posttranslational synthesis of hypusine: evolutionary progression and specificity of the hypusine modification. *Amino Acids* **2007**, *33*, 341–350.

(64) Bevec, D.; Jaksche, H.; Oft, M.; Wöhl, T.; Himmelspach, M.; Pacher, A.; Schebesta, M.; Koettnitz, K.; Dobrovnik, M.; Csonga, R.; Lottspeich, F.; Hauber, J. Inhibition of HIV-1 replication in lymphocytes by mutants of the Rev cofactor eIF-5A. *Science* **1996**, *271*, 1858–1860.

(65) Ruhl, M.; Himmelspach, M; Bahr, G. M.; Hammerschmid, F.; Jaksche, H.; Wolff, B.; Aschauer, H.; Farrington, G. K.; Probst, H.; Bevec, D.; Hauber, J. Eukaryotic Initiation Factor 5A Is a Cellular Target of the Human Immunodeficiency Virus Type 1 Rev Activation Domain Mediating Trans-Activation. *J. Cell. Biol.* **1993**, *123*, 1309–1320.

(66) Strebel, K. Virus-host interactions: role of HIV proteins Vif, Tat, and Rev. *AIDS* **2003**, *17*, 25–34.

(67) Pollard, V. W.; Malim, M. H. The HIV-1 Rev protein. *Annu. Rev. Microbiol.* **1998**, *52*, 491–532.

(68) Groom, H. C.; Anderson, E. C.; Lever, A. M. Rev: beyond nuclear export. *J. Gen. Virol.* **2009**, *90*, 1303–1318.

(69) Li, C. H.; Ohn, T.; Ivanov, P.; Tisdale, S.; Anderson, P. eIF5A Promotes Translation Elongation, Polysome Disassembly and Stress Granule Assembly. *PLoS One* **2010**, *5*, 1–13.

(70) Lee, Y. B.; Park M. H. Folk, J. E. Diamine and triamine analogs and derivatives as inhibitors of deoxyhypusine synthase: synthesis and biological activity. *J. Med. Chem.* **1995**, *38*, 3053–3061.

(71) Specht, S.; Sarite, S.; Hauber, I.; Hauber, J.; Gorbig, U.; Meier, C.; Bevec, D.; Hoerauf, A.; Kaiser, A. The guanylhydrazone CNI-1493: an inhibitor with dual activity against malaria-inhibition of host cell pro-inflammatory cytokine release and parasitic deoxyhypusine synthase. *Parasitol. Res.* **2008**, *102*, 1177–1184.

(72) Hart, R. A.; Billaud, J.-N.; Choi, S. J.; Phillips, T. R. Effects of 1,8-Diaminooctane on the FIV Rev Regulatory System. *Virology* **2002**, *304*, 97–104.

(73) Hauber, I.; Bevec, D.; Heukeshoven, J.; Krätzer, F.; Horn, F.; Choidas, A.; Harrer, T.; Hauber, J. Identification of cellular deoxyhypusine synthase as a novel target for antiretroviral therapy. *J. Clin. Invest.* **2005**, *115*, 76–85.

(74) Long-term Study of Semapimod (CNI-1493) for Treatment of Crohn's Disease., accessed June 28th, 2014, `http://clinicaltrials.gov/ct2/show/NCT00740103`.

(75) CNI-1493 for Treatment of Moderate to Severe Crohn's Disease., accessed June 28th, 2014, `http://clinicaltrials.gov/ct2/show/NCT00038766`.

(76) Hommes, D.; van den Blink, B.; Plasse, T.; Bartelsman, J.; Xu, C.; Macpherson, B.; Tytgat, G.; Peppelenbosch, M.; Van Deventer, S. Inhibition of stress-activated MAP kinases induces clinical improvement in moderate to severe Crohn's disease. *Gastroenterol.* **2002**, *122*, 7–14.

(77) Schlosser, J.; Rarey, M. Beyond the Virtual Screening Paradigm: Structure-Based Searching for New Lead Compounds. *J. Chem. Inf. Model.* **2009**, *49*, 800–809.

(78) Heinrich Pette Institute., accessed June 28th, 2014, `http://www.hpi-hamburg.de/en/`.

(79) Wilson, W. D.; Tanious, F. A.; Barton, H. J.; Jones, R. L.; Fox, K.; Wydra, R. L.; Strekowski, L. DNA sequence dependent binding modes of 4',6-diamidino-2-phenylindole (DAPI). *Biochemistry* **1990**, *29*, 8452–8461.

(80) Böhm, H.-J.; Flohr, A.; Stahl, M. Scaffold hopping. *Drug Discov. Today: Technol.* **2004**, *1*, 217–224.

(81) Hilbig, M.; Urbaczek, S.; Groth, I.; Heuser, S.; Rarey, M. MONA - Interactive manipulation of molecule collections. *J. Cheminform.* **2013**, *5*, 1–10.

(82) Fischer, J. R.; Lessel, U.; Rarey, M. LoFT: Similarity-Driven Multiobjective Focused Library Design. *J. Chem. Inf. Model.* **2010**, *50*, 1–21.

(83) Lessel, U.; Wellenzohn, B.; Fischer, J. R.; Rarey, M. Design of Combinatorial Libraries for the Exploration of Virtual Hits from Fragment Space Searches with LoFT. *J. Chem. Inf. Model.* **2012**, *52*, 373–379.

(84) Stierand, K.; Harder, T.; Marek, T.; Hilbig, M.; Lemmen, C.; Rarey, M. The Internet as Scientific Knowledge Base: Navigating the Chem-Bio Space. *Mol. Inform.* **2012**, *31*, 543–546.

(85) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the worldwide chemical structure identifier standard. *J. Cheminform.* **2013**, *5*, 7.

6. BIBLIOGRAPHY (OTHER AUTHORS)

(86) InChI version 1, software version 1.04 (2011)-Technical Manual., accessed June 28th, 2014, `http://www.inchi-trust.org/fileadmin/user_upload/software/inchi-v1.04/InChI_TechMan.pdf`.

(87) May, J.; Steinbeck, C. Efficient ring perception for the Chemistry Development Kit. *J. Cheminform.* **2014**, *6*, 3.

(88) Alexandrescu, A., *Modern C++ Design: Generic Programming and Design Patterns Applied*; Addison Wesley: 2001.

(89) Zentrum für Bioinformatik - NAOMI., accessed June 28th, 2014, `http://www.zbh.uni-hamburg.de/NAOMI`.

(90) Qt., version 4.7, accessed June 28th, 2014, `http://qt-project.org/`.

(91) Tarjan, R.; Vishkin, U. An Efficient Parallel Biconnectivity Algorithm. *SIAM J. Comput.* **1985**, *14*, 862–874.

Chapter 7

Bibliography (Adrian Kolodzik)

7.1 Journal Articles

(A1) Urbaczek, S.; **Kolodzik, A.**; Fischer, R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI - On the almost trivial task of reading molecules from different file formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199–3207.

(A2) Urbaczek, S.; **Kolodzik, A.**; Groth, I.; Heuser, S.; Rarey, M. Reading PDB: Perception of Molecules from 3D Atomic Coordinates. *J. Chem. Inf. Model.* **2013**, *53*, 76–87.

(A3) Urbaczek, S.*; **Kolodzik, A.***; Rarey, M. The Valence State Combination Model: A Generic Framework for Handling Tautomers and Protonation States. *J. Chem. Inf. Model.* **2014**, *54*, 756–766.
* equal contribution

(A4) **Kolodzik, A.**; Urbaczek, S.; Rarey, M. Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies. *J. Chem. Inf. Model.* **2012**, *52*, 2013–2021.

(A5) Schröder, M.*; **Kolodzik, A.***; Pfaff, K.; Priyadarshini, P.; Krepstakies, M.; Hauber, J.; Rarey, M.; Meier, C. In silico Design, Synthesis, and Screening of Novel Deoxyhypusine Synthase Inhibitors Targeting HIV-1 Replication. *ChemMedChem* **2014**, *9*, 940–952.
* equal contribution

(A6) Ziegler, P.; Chahoud, T.; Wilhelm, T.; Pällman, N.; Braig, M.; Wiehle, V.; Ziegler, S.; Schröder, M.; Meier, C.; **Kolodzik, A**.; Rarey, M.; Panse, J.; Hauber, J.; Balabanov, S.; Brümmendorf, T. H. Evaluation of deoxyhypusine synthase inhibitors targeting BCR-ABL positive leukemias. *Invest. New Drug* **2012**, *30*, 2274–2283.

7.2 Book Chapters

(B1) **Kolodzik, A.**; Schneider, N.; Rarey, M. In *Structure Based Virtual Screening*, Engel, T., Ed., in editorial process.

(B2) **Kolodzik, A.**; Rarey, M. In *Ring Perception*, Engel, T., Ed., in editorial process.

7.3 Talks

(C1) Rarey, M.; **Kolodzik, A.**; Urbaczek, S. Let's Talk About Rings., 8th German Conference on Chemoinformatics, Goslar, 2012.

(C2) **Kolodzik, A.**; Urbaczek, S.; Rarey, M. Unique cycle families: A set of unique and chemically meaningful rings., 242nd ACS National Meeting, Denver, 2011.

(C3) **Kolodzik, A.**; Hilbig, M.; von Behren, M. M.; Heumaier, A.; Otto, T.; Urbaczek, S.; Rarey, M. MONA: A solution for handling large data sets in drug discovery., 242nd ACS National Meeting, Denver, 2011.

7.4 Posters

(D1) Urbaczek, S.; **Kolodzik, A.**; Heuser, S.; Groth, I.; Rarey, M. NAOMI - On the almost trivial task of reading molecules from different file formats., Gordon Research Converence, Boston, Massachusetts, 2011.

(D2) Urbaczek, S.; **Kolodzik, A.**; Rarey, M. NAOMI - On the almost trivial task of reading molecules from different file formats., International Conference on Chemical Structures, Noordwijkerhout, Netherlands, 2011.

(D3) Schröder, M.; **Kolodzik, A.**; Windshuegel, B.; Krepstakies, M.; Priyadarshini P. amd Hauber, J.; Rarey, M; Meier, C. Rational Drug Design - Sreening and Synthesis of Potential Deoxyhypusine Synthase Inhibitors Targeting HIV-1 Replication., 23rd International Converence on Antiviral Research, San Francisco, California, 2010.

(D4) **Kolodzik, A.**; Rarey, M. Computational methods in drug development., 6th Status Seminar Chemical Biology, Frankfurt a.M, Germany, 2009.

(D5) **Kolodzik, A.**; Rarey, M. Computational methods in drug development., Academic Drug Development in Oncology - Translating Basic Science Research into Innovative Treatments, Berlin, Germany, 2009.

7.5 Patents

(E1) Priyadarshini, P.; Schröder, M.; Rarey, M.; **Kolodzik, A.**; Hauber, J.; Krepstakies, M.; Meier, C. Deoxyhypusin-Synthese Inhibitoren. pat., DE 102012103405 A1, Oct. 24, 2013.

7.6 Individual Contributions to Journal Articles

7.6.1 Urbaczek, S.; Kolodzik, A.; Fischer, R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI - On the almost trivial task of reading molecules from different file formats. *J. Chem. Inf. Model.* 2011, *51*, 3199–3207

The NAOMI software library was developed and implemented in a joint effort by Sascha Urbaczek, Adrian Kolodzik, Dr. Robert Fischer and Dr. Tobias Lippert. Sascha Urbaczek focused on the development of concepts of the chemical model and the computational representation of molecules in the NAOMI framework. Adrian Kolodzik focused on the adaption to file formats and supported the validation of the software. Dr. Robert Fischer's focus was the handling of fragment spaces and the general software design. Dr. Tobias Lippert focused on the efficient setup of the library and a design to support the handling of protein structures. Dr. Inken Groth, Dr. Stefan Heuser, and Dr. Tanja Schulz-Gasch provided general support. Prof. Dr. Matthias Rarey supervised the work.

7.6.2 Kolodzik, A.; Urbaczek, S.; Rarey, M. Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies. *J. Chem. Inf. Model.* 2012, *52*, 2013–2021

The general idea of the URFs was developed by Sascha Urbaczek and Adrian Kolodzik in a joint effort. Adrian Kolodzik both established the theoretical foundation for the description of the URFs and developed the algorithms for their calculation in polynomial time. Furthermore, Adrian Kolodzik integrated the method into the NAOMI framework. Together, Adrian Kolodzik and Sascha Urbaczek designed the validation procedures which were carried out by Adrian Kolodzik. Prof. Dr. Matthias Rarey supervised the work and helped with the development of the final concept.

7.6.3 Urbaczek, S.; Kolodzik, A.; Groth, I.; Heuser, S.; Rarey, M. Reading PDB: Perception of Molecules from 3D Atomic Coordinates. *J. Chem. Inf. Model.* 2013, *53*, 76–87

The initial concepts and algorithmic ideas for the perception of molecules from 3D coordinates were developed by Sascha Urbaczek and Adrian Kolodzik in a joint effort. Sascha Urbaczek implemented the algorithms and methods for the individual steps of the workflow and derived parameters for the scoring of individual valence state assignments. Furthermore, Sascha Urbaczek integrated the methods in the NAOMI

framework. Together, Adrian Kolodzik and Sascha Urbaczek designed the validation procedures which were carried out by Sascha Urbaczek. Prof. Dr. Matthias Rarey supervised the work.

7.6.4 Urbaczek, S.*; Kolodzik, A.*; Rarey, M. The Valence State Combination Model: A Generic Framework for Handling Tautomers and Protonation States. *J. Chem. Inf. Model.* 2014, *54*, 756–766
*** equal contribution**

The general concepts for the generation of tautomers and protonation states were developed by Sascha Urbaczek and Adrian Kolodzik in a joint effort. Sascha Urbaczek implemented the algorithmic ideas for the selection of valence states, the generation of valid valence bonds structures and the scoring. Furthermore, Sascha Urbaczek integrated the implementation into the NAOMI library. Adrian Kolodzik developed the functionality for the partitioning of molecules into zones. Adrian Kolodzik and Sascha Urbaczek performed the evaluation in a joint effort. Prof. Dr. Matthias Rarey supervised the work.

7.6.5 Schröder, M.*; Kolodzik, A.*; Pfaff, K.; Priyadarshini, P.; Krepstakies, M.; Hauber, J.; Rarey, M.; Meier, C. In silico Design, Synthesis, and Screening of Novel Deoxyhypusine Synthase Inhibitors Targeting HIV-1 Replication. *ChemMedChem* 2014, *9*, 940–952
*** equal contribution**

Potential inhibitors were synthesized by Dr. Marcus Schröder. The chemical synthesis was supervised by Prof. Dr. Chris Meier. Adrian Kolodzik designed potential DHS inhibitors *in silico*. The *in silico* design was supervised by Prof. Dr. Matthias Rarey. The cell toxicity of the potential inhibitors and the assessment of their inhibitory activity was performed by Katharina Pfaff, Dr. Poornima Priyadarshini, and Dr. Marcel Krepstakies. These tests were supervised by Prof. Dr. Joachim Hauber.

7.6.6 Ziegler, P.; Chahoud, T.; Wilhelm, T.; Pällman, N.; Braig, M.; Wiehle, V.; Ziegler, S.; Schröder, M.; Meier, C.; Kolodzik, A., et al. Evaluation of deoxyhypusine synthase inhibitors targeting BCR-ABL positive leukemias. *Invest. New Drug* 2012, *30*, 2274–2283

Dr. Patrick Ziegler, Dr. Tuhama Chahoud, Dr. Thomas Wilhelm, Nora Pällman, Dr. Melanie Braig, Valeska Wiehle, Dr. Susanne Ziegler, Dr. Marcus Schröder, and Dr. Jens Panse synthesized and evaluated the new inhibitor of DHS in a joint effort. This work was supervised by Prof. Dr. Chris Meier, Prof. Dr. Joachim Hauber, Dr. Stefan Balabanov, and Prof. Dr. Tim Henrik Brümmendorf. Adrian Kolodzik discussed the results in the context of half-yearly meetings of the consortium "Combating drug resistance in chronic myeloid leukemia and HIV-1 infection" and helped with proof-reading of the manuscript. Prof. Dr. Matthias Rarey supervised the work as a member of the consortium.

Glossary

Ω	set of all rings
ADMET	absorption, distribution, metabolism, excretion, and toxicity
CML	chronic myeloid leukemia; a hematopoietic stem cell disease
DAPI	4',6-diamidino-2-phenylindole; a fluorescent dye which strongly binds to DNA
DHS	deoxyhypusine synthase
DOHH	deoxyhypusine hydroxylase
eIF-5A	elongation initiation factor 5A
ESER	essential set of essential rings
GC_7	N-1-guanyl-1,7-diaminoheptane
HIV	human immunodeficiency virus
HTS	high throughput screening
PAINS	pan assay interference compounds
QSAR	quantitative structure-activity relationship
RC	relevant cycle
RCP	relevant cycle prototypes
SAR	structure-activity relationship
SMARTS	smiles arbitrary target specification; a language to describe molecular patterns by strings
SMILES	simplified molecular input line entry system; a 1D representation of molecular structures
SSSR	smallest set of smallest rings

7. BIBLIOGRAPHY (ADRIAN KOLODZIK)

URF	unique ring family
USMILES	unique SMILES
ZBH	Center for Bioinformatics of the University of Hamburg
ZINC	ZINC is not commercial (recursive acronym); a curated collection of commercially available compounds

Appendix A

Publications in Scientific Journals

A.1 Urbaczek, S.; Kolodzik, A.; Fischer, R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI - On the almost trivial task of reading molecules from different file formats. *J. Chem. Inf. Model.* 2011, *51*, 3199–3207

ACS direct link:
http://pubs.acs.org/articlesonrequest/AOR-hRTTf9abf9PGggQX9ztR

JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

NAOMI: On the Almost Trivial Task of Reading Molecules from Different File formats

Sascha Urbaczek,[†] Adrian Kolodzik,[†] J. Robert Fischer,[†] Tobias Lippert,[†] Stefan Heuser,[‡,∥] Inken Groth,[‡] Tanja Schulz-Gasch,[§] and Matthias Rarey*[,†]

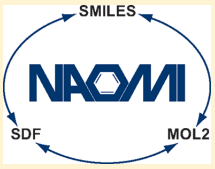
[†]Center for Bioinformatics (ZBH), Bundesstrasse 43, 20146 Hamburg, Germany

[‡]Research Active Ingredients, Beiersdorf AG, Troplowitzstrasse 15, 22529 Hamburg, Germany

[§]Pharmaceutical Division, F. Hoffmann-La Roche Ltd., CH-4070 Basel, Switzerland

Supporting Information

ABSTRACT: In most cheminformatics workflows, chemical information is stored in files which provide the necessary data for subsequent calculations. The correct interpretation of the file formats is an important prerequisite to obtain meaningful results. Consistent reading of molecules from files, however, is not an easy task. Each file format implicitly represents an underlying chemical model, which has to be taken into consideration when the input data is processed. Additionally, many data sources contain invalid molecules. These have to be identified and either corrected or discarded. We present the chemical file format converter NAOMI, which provides efficient procedures for reliable handling of molecules from the common chemical file formats SDF,[1] MOL2,[2] and SMILES.[3] These procedures are based on a consistent chemical model which has been designed for the appropriate representation of molecules relevant in the context of drug discovery. NAOMI's functionality is tested by round robin file IO exercises with public data sets, which we believe should become a standard test for every cheminformatics tool.



INTRODUCTION

Chemical file formats provide the necessary data for application programs and offer a means to share results with other scientists in a computer readable form. For small molecules, the most commonly used formats are Symyx SDF V2000 (formerly MDL SDF),[1] Tripos MOL2,[2] and Daylight SMILES.[3] Virtually all public databases provide molecular files of at least one of these types.

Unfortunately, many programs do not accept all formats as input or generate only some of them as output. Hence, file format converters are needed to exchange data between these tools. This becomes especially important if several of these tools are combined in a workflow. The consistent conversion of molecules is crucial at this point, since even minor alterations might result in errors in subsequent calculations.

The conversion process is difficult and error prone. File formats implicitly represent an underlying chemical model which has to be taken into account. Hence, the file format conversion is actually a conversion between different chemical representations. Furthermore, some programs generate files that do not follow format specifications or contain errors. Converters must thus be able to identify errors and ambiguities in input data and resolve them consistently or discard the corresponding molecule.

Since chemical file formats play such a central role in cheminformatics, every tool and software package must be able to read and write molecular files. Hence, every tool that supports more than one file format can be used as a converter. However, there are tools which have specifically been designed for file format conversion, such as the free software OpenBabel[4] and, more recently, fconv[5] or the commercial tools MOL2Mol,[6] MN. Convert,[7] and Babel.[8] Furthermore, there is a large number of programming libraries for cheminformatics, both open source and proprietary, which provide the necessary functionality to read and write molecules. Evidently, these can be used to implement converter tools. Examples of such libraries are Open Babel,[9] CDK,[10] CACTVS,[11] JOELib,[12] PerlMol,[13] OEChem,[14] and RDKit.[15] Additionally, some tools are routinely used for file format conversions, although that is not their specific purpose. Typical examples are programs for the generation of 3D coordinates, such as CORINA,[16] LigPrep,[17] and CONCORD.[18]

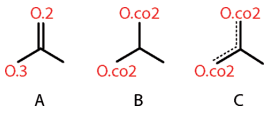


Figure 1. Different representations of carboxylates as observed in MOL2 files.

We have implemented a new tool for the consistent conversion of chemical file formats called NAOMI. This converter is based on a robust chemical model which is designed to appropriately describe organic molecules relevant in the context of drug discovery. It provides a reliable and accurate internal representation which allows for a consistent interconversion of the widely used

Received: July 14, 2011
Published: November 08, 2011

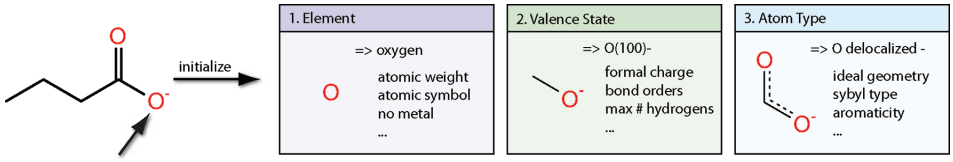


Figure 2. Annotation of the three levels of chemical information for an oxygen of a carboxylate.

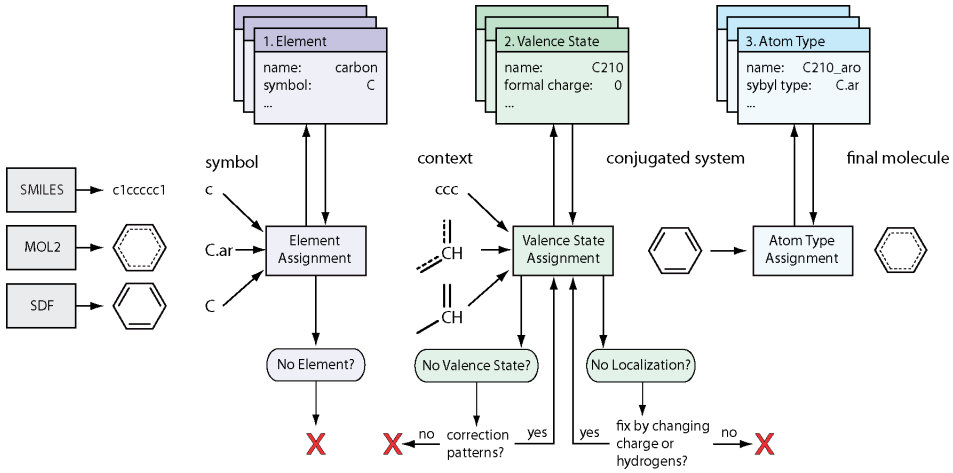


Figure 3. Schematical view of the three steps of molecule initialization.

molecular file formats SDF V2000,[1] MOL2,[2] and SMILES.[3] NAOMI also supports reading and writing SDF V3000 files but does currently not implement all associated features, e.g., self-contained sequence representation. NAOMI checks the chemical validity of molecules and calculates molecular descriptors independent of input file formats.

Although file IO is a task all cheminformatics tools have to perform, not very much is known about the methodologies applied to address the problems related to file conversion. We assume that many tools use approaches very similar to NAOMI, but unfortunately these are mostly not published. Furthermore, file IO and conversion is rarely tested and validated exhaustively. The aim of this paper is to explicitly put the focus on these tasks to demonstrate the complexity and typical pitfalls. We present a round robin test for cheminformatics tools able to read and write different file formats and advocate the use of such tests routinely.

File Format Conversion. The conversion of file formats involves two steps: First, the information provided by the input format is interpreted to build an internal representation of the molecule. Second, all relevant data for the target format is derived from this representation. Due to the different underlying chemical models of the file formats, the conversion usually involves switching from one chemical description to another. Thus, it is important to consider the requirements and limitations of these descriptions.

The Symyx SDF format[1] represents molecules by a single valence bond structure, also called Lewis structure.[19] Hydrogens are frequently omitted to save disk space, while the file format specification ensures the presence of formal charges. The valence bond description has limitations concerning kekule and resonance

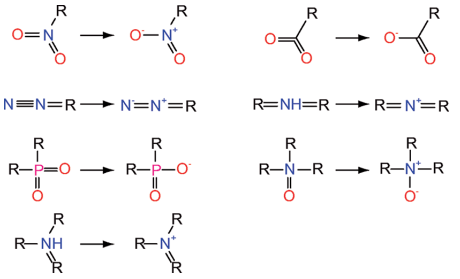


Figure 4. If no valence state can be identified for an atom, then a set of simple correction patterns is applied.

structures, since multiple equivalent valence bond forms of the same molecule may exist.

SMILES[20] can represent molecules by a single valence bond structure, whereas hydrogens are virtually always omitted. The format also implements the concept of aromatic atoms and bonds, which allows to represent aromatic systems with different equivalent kekule forms by a single delocalized description. According to the Daylight theory manual,[20] aromaticity in SMILES is however not intended to model physicochemical properties (Daylight theory manual, page 14). Nevertheless, aromatic atoms and bonds are commonly used to describe molecules which are aromatic in a chemical sense, although a single valence bond structure would be sufficient to characterize these molecules unambiguously.

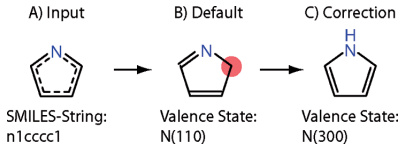


Figure 5. If an input file annotates aromatic atoms and bonds (A), default valence states are assigned in a first step (B). If this attempt is not successful, alternative valence states are considered (C) to correct the input.

The TRIPOS MOL2 format implements the concept of aromatic atoms and bonds, too. Furthermore, the format offers the possibility to describe equivalent resonance forms of common functional groups, such as carboxylates and guanidinium groups, with a delocalized representation. This is realized using specific atom types, called sybyl types, which include information about the atom's hybridization. Usually, MOL2 files do not provide formal charges, but hydrogens are specified. Unfortunately, there is no exact documentation on how the sybyl types must be assigned. This leads to considerable differences between MOL2 files written by different tools. As shown in Figure 1 there are many ways to combine sybyl types, bond orders, and charges to describe the same functional group.

METHODOLOGY

Chemical Model. A consistent chemical model is the keystone for an appropriate internal representation of molecules in cheminformatics application. It also provides the framework for the identification and correction of erroneous input molecules.

The atom-centered chemical model of NAOMI comprises three different levels of chemical information which are assigned to each atom during an initialization procedure. Each level extends the environment that is considered and provides a more detailed description of the atom.

The element is the first and most basic level of description. It provides properties which depend only on the atom's chemical element. These properties comprise the element symbol, the element name, the atomic number, the atomic weight, the van der Waals radius, the number of valence electrons, the covalent radius, and whether the element is considered a metal.

The valence state is the second level of chemical information and extends the scope of the chemical element by taking bonds and formal charges into account. Each valence state represents a valid bond pattern of an atom in a valence bond structure of the molecule. Valence states contain topological information which include formal charge, number of bonds, bond orders, number of free electrons, and whether the corresponding atom can be part of a conjugated or aromatic system.

The atom type extends the valence state to model effects, such as aromaticity and the existence of equivalent resonance forms. This is needed to compensate for the shortcomings of a localized molecular description.

To determine an atom type, the atom and all atoms in its conjugated system (if applicable) are considered. Atom types provide an ideal geometry, a corresponding sybyl type, mark atoms as conjugated or aromatic, and contain information about delocalized electrons. Additionally, an atom type marks the corresponding atom as a hydrogen-bond acceptor or as a potential hydrogen-bond donor.

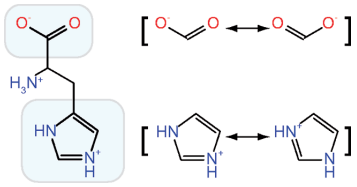


Figure 6. Molecules are partitioned into zones of conjugated atoms. The two oxygen atoms of the carboxylate group and the two nitrogen atoms of the imidazole ring have different valence states but identical atom types. Therefore, the valence states describe a localized structure with a defined formal charge, and the atom types describe a delocalized structure, with a delocalized charge.

Each atom is assigned a corresponding element, valence state and atom type (see Figure 2). Valence states ensure that each molecule has a valid valence bond structure, while atom types allow easy access to a delocalized description.

The basic assumption of the chemical model is that organic molecules which are relevant in the drug discovery context can always be represented by at least one valence bond structure. If that is not the case, then the molecule will either be corrected or discarded. Since there are no strict valence rules for metallic elements, only monatomic ions are accepted. Molecules containing covalently bound metals are currently not supported by the model.

Molecule Initialization. *Overview.* During the molecule initialization data from input files is used to build the internal representation of the molecule. This task is carried out in three separate steps (see Figure 3).

Element Assignment. First, the molecular graph is built from the connectivity data provided by the input file. During this process, the element for each atom is determined, and initial bond types are assigned. The perception of elements, bond types, and connectivity from the different file formats is implemented according to their respective specifications. All elements of the periodic table and bonds of type single, double, triple, and aromatic are supported. Molecules which have atoms or bonds with undefined types are discarded at this point, since this information is required in the subsequent steps.

The initial data are used to generate a valid valence bond form of the molecule. A valence bond form is valid if valence states can be assigned to all atoms and the aromatic bonds can be localized. If no valence bond form can be generated and no correction is possible, the molecule is discarded.

Valence State Assignment. They are selected on basis of the formal charge and bond orders of the atom. Hence, molecules with formal charges, hydrogens, and localized bond orders are the optimal input for this procedure. In this case, the assignment is straightforward and unambiguous. The omission of hydrogens or the use of aromatic bonds, which basically corresponds to the omission of bond orders, also poses no problem, since the remaining properties are still sufficient to reach an unambiguous assignment. If charges or multiple properties are missing, then additional data from the input format is necessary to resolve ambiguities.

Each file format makes use of a different molecular representation and applies certain strategies to omit redundant information. Hence, individual assignment procedures are needed for each file format.

Molecules from SDF are supplied in a valence bond form, which allows a direct comparison to valence states. If hydrogens are

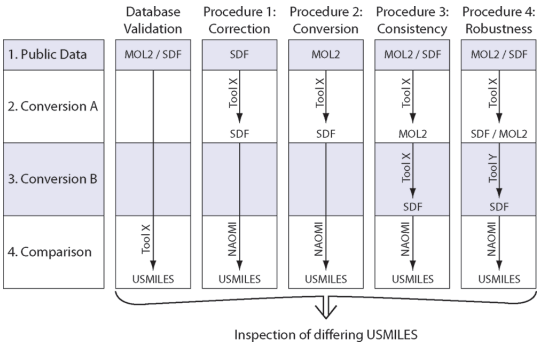


Figure 7. Various procedures test different aspects of file format conversions. During these procedures, molecules are converted by different combinations of tools. USMILES are used for the comparison of the resulting molecules.

Table 1. Options Used for Computing Time Benchmarks

tool/options	explanation
CORINA	
−d wh	write hydrogens to output file
−d no 3d	disable generation of 3D coordinates
−t n	do not write trace file
MOE	
−SVL script	(see Supporting Information)
NAOMI	
−v 0	do not print messages to shell
Open Babel	
−o can	generate USMILES (only for SMILES as output)

omitted, formal charges and multiple bonds are sufficient to unambiguously identify the correct valence states.

Molecules from SMILES may provide information on the bond orders explicitly, whereas hydrogens are virtually always omitted. If this is the case, the assignment works the same way as for SDF. Additionally, SMILES implements the concept of aromatic bonds. This means that bond orders and hydrogens can be missing, and hence ambiguities arise for certain types of atoms. The most prominent example is the pyrrole-like aromatic nitrogen (see Figure 5) which has to be provided with explicit hydrogens for an unambiguous assignment.

Molecules from MOL2 usually have all hydrogens attached but lack the specification of formal charges. If they also contain aromatic bonds, two properties are missing. These ambiguities can only be resolved by using sybyl type information. Additionally, some resonance forms of common functional groups are indicated by specific sybyl types. Their bond types and valence states are adapted accordingly in a postprocessing step.

If no valence state could be found for an atom, the atom's environment is checked by using simple patterns representing common valence errors (see Figure 4). If a pattern matches, then a valence state is assigned, and the bond orders and valence states of the environment are adapted. Otherwise the molecule is discarded.

Afterward, the bonds marked as aromatic in the input file are localized to ensure a valid valence bond form. Information about the localized bond orders for each atom is provided by its corresponding valence state. The information is used in a recursive algorithm to assign defined bond orders to all bonds.

Table 2. Validation of Input Data Sets by NAOMI

data set	no. molecules	no. rejected molecules	corrected molecules	no. diffs MOL2 ↔ SDF
DUD ligands[23]	3961	0	10	0
DUD decoys[23]	124 413	1	13	0

If the assignment of bond orders was not successful using the default valence states, all atoms of a molecule are checked for an alternative valence state assignment using rule sets specific to the respective file formats (see Figure 5). All combinations of these alternatives are enumerated, and the most probable solution is picked by a simple scoring scheme. The score is calculated as the sum of atoms which have the same valence states with respect to the initial structure. Thus, the procedure assures a minimum deviation from the default assignment. If there are multiple solutions with equal scores, a canonical solution is picked. If no solution could be found, then the molecule is discarded.

Atom Type Assignment. At this point, a valid valence bond form of the molecule is available and can be accessed during subsequent calculations. Since all necessary information can now be derived from the internal representation, the following steps are independent of the input file format.

The next step is the generation of a delocalized description for the molecule. The description allows to overcome the limitations of the valence bond representation concerning kekule and resonance structures. Although these aspects are handled by separate procedures, both need information about the molecule's rings. These are calculated using the relevant cycles algorithm as described by Vismara.[21]

Since equivalent kekule structures can only occur in cyclic systems, this information is stored directly in the molecule's rings. A ring is marked as delocalized if it has alternating single and double bonds and the number of delocalized electrons does fulfill Hueckel's rule. Bonds from rings which are already marked are considered both single and double during the check of neighboring rings. To ensure that the assignment for all rings is independent from the initial valence bond form, the assignment procedure is repeated until the total number of marked rings does not change anymore.

For the identification of equivalent resonance forms, the molecule is partitioned into zones which correspond to its conjugated systems (see Figure 6). This is done by using the information provided by the valence states in combination with the molecule's rings. Each zone is checked for pairs of atoms for which a formal charge can be exchanged. These atoms can be identified by comparison of their corresponding valence states. Then all possible resonance forms are enumerated, and all atoms with delocalized charges are marked. Finally, suitable atom types are selected from a list provided by the valence state using the information about the conjugated system and the delocalization of the atom.

After the initialization procedure, the molecule is represented by a valence bond description (valence states and bond orders) and a delocalized description (atom types and delocalization flags). Both descriptions can be used in subsequent steps.

Validation. To evaluate the quality of file format conversions, a method for comparing input and converted molecules is required. Unfortunately, there is no direct way to determine if two molecular representations are identical. This is especially true if they are stored in different file formats.

The comparison of unique SMILES (USMILES)[22] is an easy and verifiable way to identify differences between molecules. Two things have to be taken into consideration with this approach: First, USMILES generated with different tools are often not identical. This means that the method will only be reliable if the USMILES come from the same source. Second, some file format specific information will be lost during the conversion. Therefore, USMILES should be obtained from SDF files, since it provides an unambiguous valence bond structure.

The public DUD ligand and the DUD decoy[23] data sets are used in all validation procedures. To establish a reference for the comparison, both were converted from SDF and MOL2 to USMILES (see Figure 7). These USMILES serve as a basis to determine whether molecules change during conversion steps.

To investigate a tool's ability to convert file formats, four validation procedures are used as shown in Figure 7. In the first procedure, the internal error correction of the tools is analyzed by conversion of molecules from SDF to SDF. The ability to convert molecules from one format into another is investigated in the second procedure by converting molecules from MOL2 to SDF. The third procedure focuses on a tool's internal consistency by converting back and forth using the same tool twice. Finally, the robustness is checked by using different tools subsequently in a pipeline.

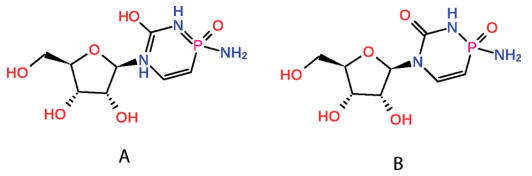


Figure 8. Molecule ZINC0153034: (A) Rejected by NAOMI in DUD decoy data set and (B) in current ZINC database.

Table 3. Data Sets Converted To USMILES by MOE and Open Babel[a]

			MOL2 ↔ SDF	
tool	data set	no. rejected molecules	no. diffs	% of data
MOE	DUD ligands	0	1598	40%
	DUD decoys	0	67 042	54%
Open Babel	DUD ligands	0	1875	47%
	DUD decoys	0	46 987	38%

[a] Shown are the differences between the generated USMILES originating from MOL2 and SDF.

Table 4. Investigation of Correction Functionality

	DUD decoys		DUD ligands	
tool	no. rejected	no. corrected	no. rejected	no. corrected
CORINA	0	0	0	0
MOE	0	13	0	8
NAOMI	1	13	0	10
Open Babel	0	0	0	0

All validation procedures are performed with CORINA,[24] MOE,[25] Open Babel,[4] and NAOMI. CORINA is commonly used for generating 3D coordinates and for molecular file format conversion and is considered the gold standard. MOE is used for a variety of applications in drug design and supports preparation of ligands for subsequent calculations. This includes the generation of protonation states and tautomers as well as filtering according to molecular descriptors. Correct and consistent reading and writing of molecules forms the basis for these applications. An open source alternative to these tools is Open Babel. Open Babel supports a variety of molecular file formats and is designed to be used as a file format converter.

Computing Time Benchmarks. Although the consistency and the quality of the converted molecules are of superior importance, computing times play a significant role due to the increasing sizes of current data sets. Hence, the runtime behavior is analyzed in order to assess their applicability in large setups.

To investigate NAOMI's performance, the ZINC-everything data set is converted from and to MOL2, SDF, and USMILES. Measured computing times are compared to the commonly used tools CORINA, Open Babel, and MOE. For an unbiased comparison, optional settings of these tools are selected to yield similar results compared to NAOMI. Therefore, generation of USMILES and writing of hydrogens are enforced, and output of additional information is minimized (see Table 1). Conversion from SMILES to MOL2 and SDF using CORINA is omitted since CORINA automatically generates 3D coordinates upon conversion. Furthermore, SMILES is not supported as an output format by CORINA. Although, NAOMI is able to conduct its calculations in parallel, this option is disabled for an easier comparison. All file format conversions are performed on a Linux PC with two Intel Xeon CPUs (2.53 GHz) and 32 GB of main memory.

RESULTS

Data Set Validation. Results of the validation of the DUD ligand and DUD decoy data sets[23] are shown in Table 2. NAOMI successfully converts all molecules except one from MOL2 and SDF to USMILES. A small number of incorrectly protonated nitrogens are corrected. One molecule (ZINC1583034) is rejected, as it contains invalid phosphorus and nitrogen atoms (see Figure 8) which cannot be corrected and localized. Since USMILES

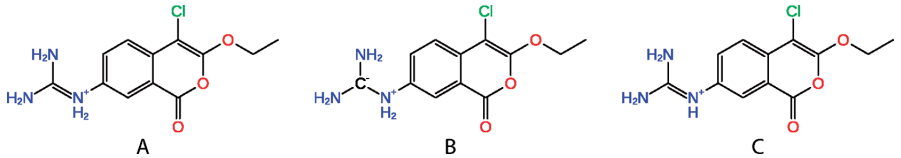


Figure 9. (A) Molecule from DUD ligand data set. (B) Corrected molecule from MOE. (C) Corrected molecule from NAOMI.

generated by NAOMI are identical for both file formats, they can serve as a reference for the following validation procedures.

Both data sets could also be successfully converted to USMILES by MOE and Open Babel. The molecule which was rejected by NAOMI is neither discarded nor corrected by both tools. USMILES originating from MOL2 and SDF, however, differ significantly (see Table 3).

Table 5. Investigation of Conversion Functionality

	DUD decoys		DUD ligands	
tool	no. diffs	% of data	no. diffs	% of data
CORINA	5522	4%	439	11%
MOE	4287	3%	181	5%
NAOMI	0	0%	0	0%
Open Babel	13 469	11%	966	24%

Table 6. Investigation of tool consistency

		DUD decoys		DUD ligands	
tool	starting file format	no. diffs	% of data	no. diffs	% of data
CORINA	MOL2	5522	4%	439	11%
	SDF	4174	3%	235	6%
MOE	MOL2	5770	5%	457	12%
	SDF	5683	5%	453	11%
NAOMI	MOL2	0	0%	0	0%
	SDF	0	0%	0	0%
Open Babel	MOL2	17 351	14%	1168	29%
	SDF	17 364	14%	1168	29%

Tool Validation 1: Correction. As mentioned above, the DUD data sets contain 24 invalid molecules in total of which one has been rejected and 23 could be corrected. CORINA and Open Babel convert those without performing any error correction (Table 4). MOE and NAOMI correct the nitrogens with invalid protonation states with differing results (see Figure 9 for an example). Additionally, NAOMI corrects invalid phosphate groups.

Tool Validation 2: Conversion. Results of the investigation of the conversion functionality (see Figure 7) are shown in Table 5. By inspection of the differing molecules, we were able to identify a small number of error classes that will be discussed for every tool:

CORINA places positive charges on carbon atoms of guanidinium- and amidinium-like groups. This error also occurs in five-membered aromatic rings containing this substructure.

MOE places positive charges on carbon atoms of guanidinium- and amidinium-like groups in five-membered aromatic rings. Depending on the substituents, the carbon atom is either charged twice or a carbon atom next to it is negatively charged.

Open Babel's most prominent class of errors is the incorrect conversion of aromatic systems containing charged nitrogen atoms. All bonds in these systems are converted to single bonds in the resulting SDF file. The second kind of error concerns protonation states. Open Babel does not consider input hydrogens to determine formal charges. Therefore, many atoms are neutralized during the conversion process. Since MOL2 entries often do not provide formal charges, this may lead to unexpected results.

Table 7. Investigation of Tool Robustness

			DUD decoys		DUD ligands	
tool X	tool Y	starting file format	no. diffs	% of data	no. diffs	% of data
CORINA	MOE	MOL2	4265	3%	176	4%
		SDF	5931	5%	449	11%
	NAOMI	MOL2	58	0%	0	0%
		SDF	4149	3%	235	6%
	Open Babel	MOL2	5522	4%	439	11%
		SDF	19 192	15%	1371	35%
MOE	CORINA	MOL2	6755	5%	504	13%
		SDF	4656	4%	245	6%
	NAOMI	MOL2	3159	3%	167	4%
		SDF	4585	4%	239	6%
	Open Babel	MOL2	4483	4%	174	4%
		SDF	19 311	16%	1374	35%
NAOMI	CORINA	MOL2	0	0%	0	0%
		SDF	643	1%	17	0%
	MOE	MOL2	176	0%	0	0%
		SDF	1217	1%	221	6%
	Open Babel	MOL2	0	0%	0	0%
		SDF	14 172	11%	1164	29%
Open Babel	CORINA	MOL2	29 896	24%	1887	48%
		SDF	10 047	8%	289	7%
	MOE	MOL2	13 693	11%	973	25%
		SDF	43 285	35%	1703	43%
	NAOMI	MOL2	13 469	11%	966	24%
		SDF	1790	1%	24	1%

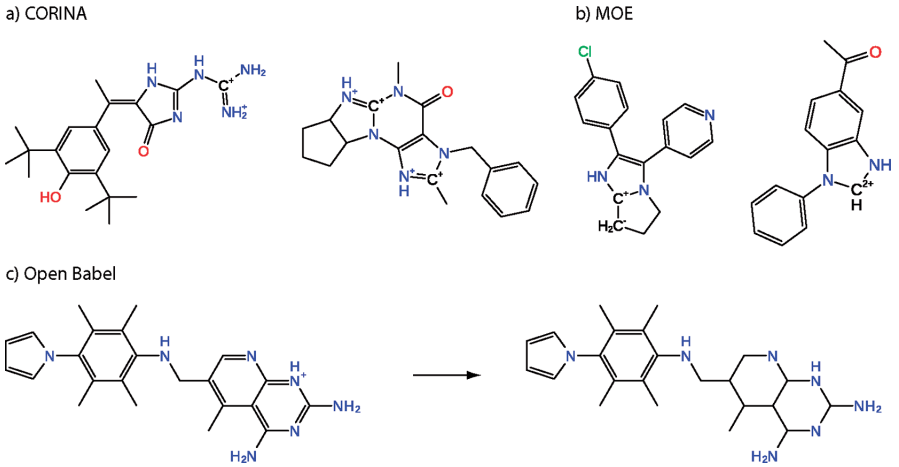


Figure 10. Examples of conversion problems with CORINA, MOE, and Open Babel.

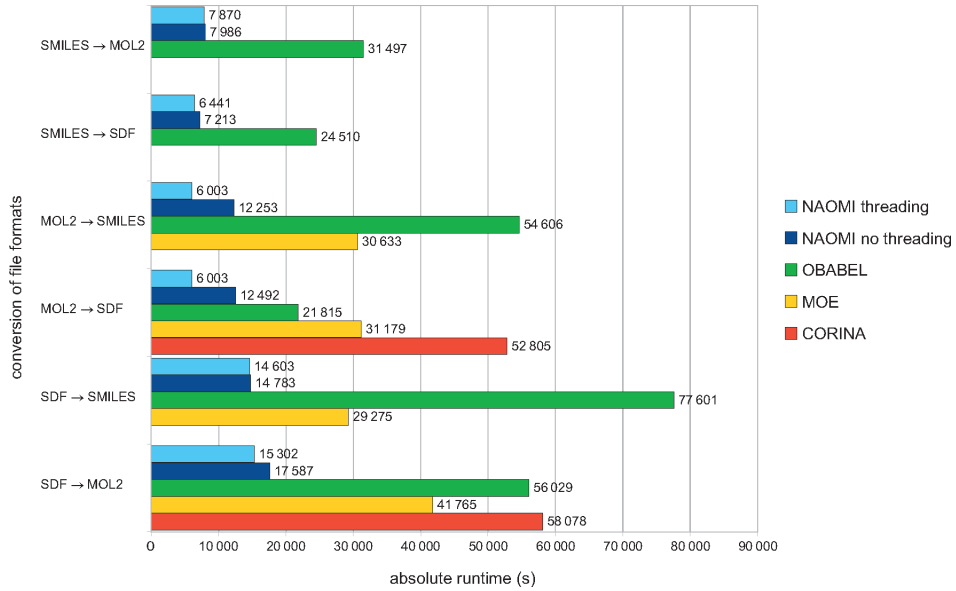


Figure 11. Computing times (wall clock time) for file format conversion of the ZINC-everything data set. For CORINA and MOE, only the computation from SDF and MOL2 are comparable, since the conversion from SMILES includes 3D coordinate generation which is not the case for NAOMI and OpenBabel. Furthermore, CORINA does support SMILES as output format.

Tool Validation 3: Consistency. Results of the investigation of consistency (see Figure 7) are shown in Table 6. Starting from MOL2, the numbers of differences should be identical to those of validation procedure 2 (see Table 5), since no additional file format conversion is performed. A higher number of errors indicates inconsistencies in reading and writing from and to MOL2. Starting from SDF, no differences at all should occur.

CORINA and NAOMI convert molecules consistently in both cases. The differences which were observed for CORINA when the first input was provided from SDF are introduced by switching from a delocalized to a localized description. Nevertheless, they only represent different valid resonance forms of the original data and are therefore not considered conversion errors. MOE and Open Babel show inconsistencies in both cases.

Tool Validation 4: Robustness. The robustness of the investigated tools is analyzed by combining two different tools in a pipeline. Since tools tend to interpret input from file formats differently, the molecules can change with each additional tool included in the workflow. Table 7 indicates that inconsistencies during file format conversion are not uncommon and depend both on the kind of tools used and on the order in which they are combined.

Furthermore, the success of the conversion strongly depends on the source of the input data. The experiment clearly shows that all tools benefit significantly from preprocessing data sources with NAOMI toward consistency and high quality (see Figure 10).

Computing Time Benchmarks. Figure 11 summarizes the computing times for conversion of the ZINC-everything data set. Since NAOMI is designed for large scale cheminformatics applications, it is not surprising that it is substantially faster than the modeling platform MOE. NAOMI supports multithreading resulting in a speed-up by another factor of 1.4. For SDF and SMILES, file IO is usually the rate-determining step. Therefore, threading does not lead to an improvement of runtimes. The MOL2 format however needs a more advanced initialization procedure, thus leading to gains in runtimes when threading is enabled.

In summary, NAOMI achieves a conversion speed of up to 2841 molecules/second on a PC with two Intel Xeon CPUs (2.53 GHz) and 32 GB of main memory.

CONCLUSION

Handling chemical structures is and remains a complex task. File formats contain chemical descriptions at different levels of detail and are therefore not easy to convert. Since the description of file formats are sometimes ambiguous when it comes to details, software tools tend to interpret them differently. This in turn causes errors in data sets and misinterpretations in tools. For the cheminformatics community, it would be a great benefit to build clear standards for file formats and to certify software with respect to these standards.

Meanwhile, it is important that software tools are at least self-consistent when reading and writing file formats. Evidently, errors in reading molecules from files usually have a substantial impact on downstream algorithms and methods. NAOMI will most certainly have flaws of its own, and in order to find them, constistency checks as those presented are needed. We urge that more of these tests should be published and that the existing ones become a standard validation procedure for all cheminformatics applications.

The command-line converter NAOMI has been implemented in C++ and can be downloaded at http://www.zbh.uni-hamburg.de/naomi. It will be available free of charge for academic use. A convenient graphical user interface for NAOMI's functionality will soon be provided by the chemical library preprocessor MONA (see http://www.zbh.uni-hamburg.de/mona).

ASSOCIATED CONTENT

Supporting Information. Original and corrected structures for both DUD data sets are provided. The corrected structures are supplied in the same file format as the respective input files (SDF or MOL2). Furthermore, a text file containing the SVL commands used for the computations with MOE is supplied. This material is available free of charge via the Internet at http://pubs.acs.org/.

AUTHOR INFORMATION

Corresponding Author
*E-mail: rarey@zbh.uni-hamburg.de.

Present Address
[||]Current address: Georg Simon Ohm University of Applied Sciences, Nuremberg, Germany.

ACKNOWLEDGMENT

The authors thank Stefan Wefing for his initial ideas concerning the chemical model, Dr. Holger Claußen for testing, Rene Kraus for IO support, and Matthias Hilbig for supplying a graphical interface.

REFERENCES

(1) *Symyx CTfile Formats*; http://www.symyx.com/downloads/public/ctfile/ctfile.jsp, (accessed January 27, 2011).
(2) *TRIPOS Mol2 File Format*; http://tripos.com/data/support/mol2.pdf, (accessed January 27, 2011).
(3) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
(4) *The Open Babel Package*, version 2.3.0; http://openbabel.org, (accessed January 18, 2011),
(5) Neudert, G.; Klebe, G. fconv: format conversion, manipulation and feature computation of molecular data. *Bioinformatics* **2011**, *27*, 1021–1022.
(6) *Mol2Mol*; http://www.gunda.hu/mol2mol/index.html, (accessed January 27, 2011).
(7) *MN.Convert*; Molecular Networks GmbH - Computerchemie: Erlangen, Germany; http://www.molecular-networks.com/products/convert, (accessed January 27, 2011)
(8) *Babel*; OpenEye Scientific Software, Inc.: Santa Fe, NM; http://www.eyesopen.com/docs/babel/current/pdf/BABEL.pdf, (accessed January 27, 2011).
(9) Guha, R.; Howard, M.; Hutchison, G.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. The Blue Obelisk-Interoperability in Chemical Informatics. *J. Chem. Inf. Model.* **2006**, *46*, 991–998.
(10) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.
(11) Ihlenfeldt, W.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Compatibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109–116.
(12) *JOELib/JOELib2*; http://sourceforge.net/projects/joelib/, (accessed January 27, 2011).
(13) *PerlMol*; http://www.perlmol.org/, (accessed January 27, 2011).
(14) *OEChem*; OpenEye Scientific Software, Inc.: Santa Fe, NM; http://www.eyesopen.com/oechem-tk, (accessed January 27, 2011).
(15) *RDKit*; http://rdkit.org/, (accessed Jan 27, 2011).
(16) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000–1008.
(17) *LigPrep*;Schrödinger, LLC: Cambridge, MA; http://www.schrodinger.com/products/14/10/, (accessed January 27, 2011).
(18) *Concord*; Tripos: St. Louis, MO; http://tripos.com/data/SYBYL/Concord_072505.pdf, (accessed January 27, 2011).
(19) Lewis, G. N. The Atom and the Molecule. *J. Am. Chem. Soc.* **1916**, *38*, 762–785.
(20) *Daylight theory manual*, Daylight version 4.9; Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA; http://www.daylight.com/dayhtml/doc/theory/index.pdf, (accessed March 8, 2011).
(21) Vismara, P. Union of all the minimum cycle bases of a graph. *Electron. J. Comb.* **1997**, *4*, 1–15.
(22) Weininger, D.; Weininger, A.; Weininger, J. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.
(23) Huang, N.; Shoichet, B.; Irwin, J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801 Data sets (SDF and Mol2) downloaded April 12, 2011.

(24) *CORINA - Fast Generation of High-Quality 3D Molecular Models*, version 3.48; Molecular Networks GmbH - Computerchemie : Erlangen, Germany; http://www.molecular-networks.com/products/corina, (accessed January 18, 2011).
(25) *MOE*, version 2010.10; Chemical Computing Group: Montreal, Quebec, Canada; http://www.chemcomp.com/software.htm, (accessed January 18, 2011).

A.2 Ziegler, P.; Chahoud, T.; Wilhelm, T.; Pällman, N.; Braig, M.; Wiehle, V.; Ziegler, S.; Schröder, M.; Meier, C.; Kolodzik, A., et al. Evaluation of deoxyhypusine synthase inhibitors targeting BCR-ABL positive leukemias. *Invest. New Drug* 2012, *30*, 2274–2283

Springer link:
http://dx.doi.org/10.1007/s10637-012-9810-1

Invest New Drugs
DOI 10.1007/s10637-012-9810-1

PRECLINICAL STUDIES

Evaluation of deoxyhypusine synthase inhibitors targeting BCR-ABL positive leukemias

Patrick Ziegler · Tuhama Chahoud · Thomas Wilhelm · Nora Pällman · Melanie Braig · Valeska Wiehle · Susanne Ziegler · Marcus Schröder · Chris Meier · Adrian Kolodzik · Matthias Rarey · Jens Panse · Joachim Hauber · Stefan Balabanov · Tim H. Brümmendorf

Received: 15 January 2012 / Accepted: 28 February 2012

Summary Effective inhibition of BCR-ABL tyrosine kinase activity with Imatinib represents a breakthrough in the treatment of patients with chronic myeloid leukemia (CML). However, more than 30 % of patients with CML in chronic phase do not respond adequately to Imatinib and the drug seems not to affect the quiescent pool of BCR-ABL positive leukemic stem and progenitor cells. Therefore, despite encouraging clinical results, Imatinib can still not be considered a curative treatment option in CML. We recently reported downregulation of eukaryotic initiation factor 5A (eIF5A) in Imatinib treated K562 cells. Furthermore, the inhibition of eIF5A by siRNA in combination with Imatinib has been shown to exert synergistic cytotoxic effects on BCR-ABL positive cell lines. Based on the structure of known deoxyhypusine synthase (DHS) inhibitors such as CNI-1493, a drug design approach was applied to develop potential compounds targeting DHS. Here we report the biological evaluation of selected novel (DHSI-15) as compared to established (CNI-1493, deoxyspergualin) DHS inhibitors. We show that upon the compounds tested, DHSI-15 and deoxyspergualin exert strongest antiproliferative effects on BCR-ABL cells including Imatinib resistant mutants. However, this effect did not seem to be restricted to BCR-ABL positive cell lines or primary cells. Both compounds are able to induce apoptosis/necrosis during long term incubation of BCR-ABL positive BA/F3 derivates.

Stefan Balabanov and Tim H. Brümmendorf equal senior authorship

P. Ziegler · V. Wiehle · S. Ziegler · J. Panse · T. H. Brümmendorf
Klinik für Onkologie, Hämatologie und Stammzelltransplantation, Universitätsklinikum der RWTH, Aachen University, Aachen, Germany

T. Chahoud · T. Wilhelm · N. Pällman · M. Braig · S. Balabanov
Department of Oncology, Haematology and Bone marrow transplantation with section Pneumology, Hubertus Wald-Tumor Zentrum (UCCH), University Hospital Eppendorf (UKE), Hamburg, Germany

T. Wilhelm
Department of Biochemistry, University Hospital Aachen (UKA) of the Rheinisch.-Westfälische Technische Hochschule Aachen, Aachen, Germany

M. Schröder · C. Meier
Department of Chemistry, Organic Chemistry, Faculty of Sciences, Hamburg, Germany

A. Kolodzik · M. Rarey
Center for Bioinformatics, University of Hamburg, Hamburg, Germany

J. Hauber
Heinrich Pette Institute - Leibniz Institute for Experimental Virology, Hamburg, Germany

T. H. Brümmendorf (✉)
Klinik für Hämatologie, Onkologie und Stammzelltransplantation (Med. Klinik IV), Pauwelsstraße 30, 52074 Aachen, Germany
e-mail: tbruemmendorf@ukaachen.de

T. Chahoud
III. medizinische Klinik und Poliklinik- Nephrologie, Rheumatologie, Sektion Endokrinologie/Diabetologie, University Hospital Eppendorf (UKE), Hamburg, Germany

Published online: 14 March 2012

Pharmacological synergism can be observed for deoxyspergualin and Imatinib, but not for DHSI-15 and Imatinib. Finally we show that deoxyspergualin is able to inhibit proliferation of CD34+ progenitor cells from CML patients. We conclude that inhibition of deoxyhypusine synthase (DHS) can be supportive for the anti-proliferative treatment of leukemia and merits further investigation including other cancers.

Keywords CML · BCR-ABL · Hypusine modification · eIF5A

Introduction

Chronic myeloid leukemia (CML) represents a hematopoietic stem cell disease characterized by accumulation of mature and immature myeloid cells in the blood, bone marrow and the spleen of affected patients [1]. CML is known to arise from a reciprocal chromosomal translocation between chromosome 9 and chromosome 22 in a hematopoietic stem and progenitor cell [2]. The resulting constitutively activated and dysregulated tyrosine kinase, BCR-ABL, represents the underlying molecular mechanism of CML pathogenesis [3]. Effective inhibition of BCR-ABL tyrosine kinase activity is therefore the key element in the treatment of patients with chronic myeloid leukemia. Imatinib has been found to stop disease progression and to reverse hematologic abnormalities in CML patients [4–6]. Imatinib exerts its effects by the following modulating events: inhibition of BCR-ABL kinase activity, autophosphorylation and substrate phosphorylation, blockade of the proliferation of BCR-ABL+leukemic cells and induction of apoptosis [7].

However, resistance to Imatinib develops at a frequency of 2 % per year in chronic phase during the course of treatment and the drug seems not to affect the quiescent pool of BCR-ABL positive leukemic stem and progenitor cells [8]. Although the clinical use of Imatinib has revolutionized the treatment of CML, the finding of new cellular targets and the identification of synergistic drugs are of pivotal importance in the quest for CML cure and long term therapy [9].

We recently reported down regulation of eukaryotic initiation factor 5A (eIF5A) in Imatinib treated K562 and primary leukemic cells [10]. Interestingly, the eIF5A protein is activated by a unique post-translational modification which generates a 2-fold positively charged amino acid termed hypusine and which is similar to the polyamine spermidine [11]. Hypusination is essential for cellular proliferation and is thought to be a highly conserved process in all eukaryotes [12] and in some archaea [13]. Active eIF5A is thought to have its main role in the elongation step of translation [14]. Particularly in mammalian cells hypusinated eIF5A has also been shown to participate in the nucleocytoplasmic transport of specific cellular mRNA [15, 16]. Moreover, it is well established that eIF5A functions in retroviral mRNA transport and binds to HIV-1 and FIV, and HTLV-I Rex transactivator proteins [17, 18]. Deoxyhypusine synthase (DHS) and deoxyhypusine hydroxylase (DOHH) are the enzymes catalyzing the modification of the eIF5A protein [11] and the inhibition of eIF5A by siRNA or DHS-inhibitors in combination with Imatinib exert synergistic cytotoxic effects on BCR-ABL positive cell lines [10].

In the current project we aimed to develop novel DHS-inhibitors potentially useful for later clinical use. Based on the structure of known DHS inhibitors such as CNI-1493, new compounds targeting DHS were developed. From 20 new substances tested, only a few were able to reliably inhibit hypusine synthesis in an *in vitro* assay (data not shown). Here we report the biological evaluation of one of these new compounds, DHSI-15, which successfully passed in vitro testing. We compared its effect with the effects of the chemical lead substance CNI-1493 and with deoxyspergualin, both of which have been shown to be DHS inhibitors *in vitro* [19–21].

Material and methods

Chemicals

Imatinib was obtained form Toronto Research Chemicals (Toronto, ON). Deoxyspergualin was obtained from Nippon Kayaku Co.,Ltd. DHSI-15 and CNI-1493 were synthesized according to modified literature known procedures [22, 23]. Diethylmalonate was *in situ* deprotonated with sodium hydride before adding 5-nitro isophthaloyl dichloride. The intermediate was not isolated but reacted in a hydrolytic decarboxylation to 5-nitro-1,3-diacetylbenzene in 29 % yield. The reduction of the nitro group to the 3,5-diacetylaniline was achieved by tin(II)chloride in 98 % yield. The DHSI-15-tetraketone and CNI-1493-tetraketone, respectively were synthesized by the reaction of 3,5-diacetylaniline with 1,12-dodecanoicdiacid chloride or sebacinic acid dichloride in dry dichloromethane/pyridine (38 % and 46 %). The tetraketones were converted to the guanoyl hydrazones DHSI-15 and CNI-1493 with aminoguanidine hydrochloride in ethanol/water 9:1 (v/v) under acid catalysis. DHSI-15 and CNI-1493 precipitated as pure compounds at −26 °C and were isolated as their corresponding hydrochlorides by filtration and washing steps in yields of 68 % and 76 %. *N*,*N*′-Bis[3,5-bis[1(aminoiminomethyl)-hydrazoethyl]phenyl]dodecanediamide-tetrahydrochloride. **DHSI-15**: ^{1}H NMR (400 MHz, DMSO-d_6): δ=11.24 (brs, 4H, guanidino-NH), 10.16 (s, 2H, amide-NH), 8.12 (d, J=1.5 Hz, 4H, H-2/-6), 8.04 (t, J=1.5 Hz, 2H, H-4), 7.81 (brs, 12H, guanidino-N_2H_3), 2.37 (s, 12H, CH_3), 2.33 (t, J=7.5 Hz, 4H, H-a), 1.62–1.57 (m, 4H, H-b), 1.28 (s, 12H, H-c/-d/-e) ppm. ^{13}C NMR

(101 MHz, DMSO-d_6): δ=171.5 (amide-C), 156.0 (guanidino-C), 151.6 (imino-C), 139.4 (C-1), 137.5 (C-3/-5), 118.8 (C-4), 118.7 (C-2/-6), 36.2 (C-a), 28.8, 28.7, 28.6 (C-c/-d/-e), 24.9 (C-b), 15.1 (CH_3) ppm. HRMS (FAB) m/z=calcd 773.4834 $[M+H]^+$, found 773.4938 $[M+H]^+$. *N*,*N*′-Bis[3,5-bis[1(aminoiminomethyl)-hydrazoethyl]phenyl]decanediamide-tetrahydrochloride **CNI-1493**: ^{1}H NMR (400 MHz, DMSO-d_6): δ=11.23 (brs, 4H, guanidino-NH), 10.17 (s, 2H, amide-NH), 8.12 (d, *J*= 1.6 Hz, 4H, H-2/-6), 8.04 (t, *J*=1.6 Hz, 2H, H-4), 7.81 (brs, 12H, guanidino-N_2H_3), 2.37 (s, 12H, CH_3), 2.33 (t, *J*=7.4 Hz, 4H, H-a), 1.62–1.59 (m, 4H, H-b), 1.33–1.30 (m, 8H, H-c/-d) ppm. ^{13}C NMR (101 MHz, DMSO-d_6): δ=171.6 (amide-C), 156.0 (guanidino-C), 151.8 (imino-C), 139.4 (C-1), 137.4 (C-3/-5), 118.9 (C-2/-4/-6), 36.4 (C-a), 28.7 (C-c/-d), 25.0 (C-b), 15.2 (CH_3) ppm. HRMS (FAB) m/z=calcd 745.4598 $[M+H]^+$, found 745.4614 $[M+H]^+$.

Stock solutions of imatinib, DHSI-15, CNI-1493 and deoxyspergualin (all 10 mg/mL; in dimethyl sulfoxide [DMSO]/H_2O [1:1]) were stored at −20 °C.

Cell culture techniques

Ba/F3, BA/F3p210, BA/F3-T315I, BA/F3-E255K, BA/F3-M351T and 32D cells were cultured in RPMI 1640 medium (Invitrogen, Paisley, UK) at 37 °C in a humified atmosphere of 5 % CO_2 as described previously [24]. The following supplements were added: fetal bovine serum (10 %, Biochrom, Berlin, Germany), natriumpyruvat (1 mM), nonessential amino acids (1 mM), Penicilin (50 U) and Streptomycin (50 μM) to all cell lines, and 1 ng/ml recombinant murine interleukin-3 (RnD systems), Ba/F3 and 32D cell line only. NIH-3 T3 cells and MEFs were cultured in DMEM (Invitrogen) supplemented with 10 % fetal bovine serum, 50 U Penicilin, 50 μM Streptomycin and in addition 4 mM L-Glutamine (Biochrom) and 25 μM ß-Mercaptoethanol (Sigma) only for culturing MEFs.

For short-term proliferation assays (5 days) cells were seeded at a density of 1×10^4/mL in 24-well plates. The inhibitors were added at the beginning of the culture period in the concentrations indicated. Control cells were incubated in the same concentration of DMSO. The concentration of DMSO in all assays including inhibitors dissolved in DMSO, was less or equal to 0.1 % and had no effect on cell growth or viability. For long-term proliferation assays (13 days) cells were seeded at a density of 5×10^6/ 10 mL in T25 flasks (BD Biosciences). Cells were grown in the presence of 4 μM of the indicated inhibitors, a dose which was found to inhibit 50 % of the cells in short-term proliferation assays. Cells were counted at days 3, 6, 8, 10 and 13 and reseeded at a density of 5×10^6/ 10 mL and fresh inhibitor was added.

RNA isolation, cDNA synthesis and quantitative PCR

RNA was isolated from cells using TriFast (Peqlab) according to the manufacturer's protocol. After DNase treatment, cDNA was prepared by reverse transcription of 1 μg total RNA using random hexamer primer and M-MuLV reverse transcriptase (Fermentas). Quantitative real time PCR was performed in thermal cycler (Stratagene) using Kapa Sybr Fast Master Mix. Quanti Tect primers for eIF5A-1, eIF5A-2, DHS, DOHH and GAPDH were purchased from Qiagen. Conditions for real time PCR reaction were as follows: 1 cycle of 95 °C for 3 min and 40 cycles of 95 °C for 15 s, 57 °C for 30 s and 72 °C for 30 s. PCR was performed in triplicates and relative expression was calculated using the $2^{-\Delta\Delta CT}$ method.

Short-term expansion of human CD34+ stem and progenitor cells

Peripheral blood from patients diagnosed with CML, either in untreated chronic phase or at blast crisis or from patients with newly diagnosed AML were obtained upon written informed consent in concordance with the local ethics committe. CML was confirmed by the presence of a BCR-ABL fusion transcript in RT-PCR and according to the WHO diagnostic criteria. After Ficoll gradient centrifugation, CD34+ cells were purified using a Midi-MACS CD34 isolation kit (Miltenyi Biotec, Bergisch-Gladbach, Germany) as described elsewhere. Purity of CD34+ cells determined by flow cytometry was above 95 %. For expansion, cells were plated in triplicates in serum free medium containing human stem-cell factor (100 ng/mL), human Flt-3 Ligand (100 ng/mL), human thrombopoietin (50 ng/mL), human interleukin-3 and −6 (IL-3 and IL-6, respectively, both 20 ng/mL), and granulocyte colony-stimulating factor (20 ng/mL) and the inhibitors as reported previously. Where needed, an additional cocktail of cytokines and inhibitor containing medium was added after 5 days [25].

Cell counting

Cells were counted using flow cytometry and Flow-Count fluorospheres (Beckmann Coulter). After washing, harvested cells were resuspended in PBS containing 10 % FCS, 2 mM EDTA and 7-aminoactinomycin D. Immediately prior to analysis, 50 μl of Flow-Count fluorospheres were added. Absolute cell counts were automatically determined using a Gallios FACS-analyzer. The system software calculated cell numbers using the following formula: cells per microliter=[(viable cells counted)/(fluorospheres counted)]×fluorospheres/microliter.

Analysis of DNA fragmentation by flow cytometry

Cells from long-term proliferation (day 6 and day 13) were harvested, washed with phosphate buffered saline (PBS) and fixed in cold 70 % ethanol for at least 20 min. After repeated washes in PBS, cells were incubated in PBS containing RNAse A (100 μg/mL) and propidium iodide (10 μg/mL) for 30 min on ice; where possible at least 10.000 cells were analyzed from each sample [26].

MTT assay

Cells were grown in the presence of the indicated inhibitors or DMSO for 6 days at a density of 5×10^6/ 10 mL and subsequently seeded as 6 replicates into 96-well flat bottomed microtiter plates (BD Biosciences, Heidelberg, Germany) at a density of 9×10^3/ 150 μl. After overnight culture increasing concentrations of Imatinib were added. Forty-eight hours later the ability of remaining viable cells to transfom 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) into formazan was assessed [27]. Fraction affected (F_a) and dose-relationship at the point of IC_{50} were analyzed using CalcuSyne Software (Biosoft).

Results

Murine cell lines and CD34+ enriched cells from CML and AML patients express mRNA for eIF5A-1 and eIF5A-2 as well as for eIF5A modifying enzymes

EIF5A is an essential protein required for cellular proliferation [28]. In mammalian cells, two different genes coding for two isoforms of eIF5A, which are 84 % identical in humans, can be detected [29]. Whereas eIF5A-1 protein and mRNA are detected in all cells, only mRNA of eIF5A-2 seems to be expressed ubiquitously, whereas eIF5A-2 protein expression varies widely [30]. EIF5A contains the unusual amino acid hypusine [N(epsilon)-(4-amino-2-hydroxybutyl)lysine] and its biosynthesis occurs within the eIF5A precursor involving the enzymes deoxyhypusine synthase (DHS) and deoxyhypusine hydroxylase (DOHH). We first evaluated the expression pattern of eIF5A-1 and eIF5A-2 as well as the modifying enzymes DHS and DOHH in different murine cell lines (Fig. 1a) and in primary CD34+ cells by RT-PCR from healthy donors as well as from CML and AML patients (Fig. 1b). In primary cells and in all cell lines tested, eIF5A-1 mRNA was expressed at low to high levels, whereas the expression of eIF5A-2 was restricted to some cell lines. Within the analyzed murine cells, embryonic stem cells (ES) showed highest eIF5A-1 expression, while in mouse embryonic fibroblasts (MEFs), NIH-3T3 and BA/F3 cells mRNA levels for eIF5A-1 were low, with the weakest expression in 32D cells (Fig. 1a). In contrast, eIF5A-2 mRNA was strongly expressed in MEFs and NIH-3T3 and weakly in ES cells. No expression could be detected in BA/F3 and 32D cells. DHS and DOHH mRNA were expressed at comparable levels in all murine cell lines tested, with NIH-3T3 cells showing highest expression for DHS and MEFs showing highest expression for DOHH.

Within human CD34+ cells (Fig. 1b), mRNA expression for eIF5A-1 was found in all primary isolates. CD34+ cells from CML patients showed significantly increased eIF5A-1 expression levels when compared with healthy donors. Expression levels of eIF5A-1 in CML seemed to be independent of disease state. Expression of DHS and DOHH was also enhanced in CML patients. There was correlation between mRNA expression of DHS, DOHH and eIF5A-1 within the same patient sample, suggesting a common regulation of eIF5A-1 and its modifying enzymes. In summary, these results show that eIF5A and its modifying enzymes are expressed at varying levels in different murine cell lines and human primary CD34+ cells with enhanced expression in CML patients, suggesting that all cells should be prone to respond to hypusination inhibitors in vitro.

Inhibition of proliferation of BA/F3 p210 cells including the imatinib resistant mutants M351T, E255K, and T315I by deoxyhypusine synthase inhibitors

In order to test proliferation of BCR-ABL positive and negative cells in the presence of hypusination inhibitors, different specimens of the murine cell line BA/F3 were used. BA/F3 WT is a BCR-ABL negative cell line, which is able to proliferate in the presence of IL-3 only [27]. At day 5 of incubation, a significant inhibition of proliferation of BA/F3 WT was obvious using DHSI-15 at a concentration of 2 μM (p= 0.0082) or using deoxyspergualin (DSG) at a dose concentration of 1 μM (p=0.02) (Fig. 2a and b). CNI-1493 had only weak effects on the proliferation of BA/F3 WT cells at all concentrations tested and diminution of cell growth in the presence of CNI-1493 never reached significance when compared to DMSO treated control cells (Fig. 2c).

The BCR-ABL positive cell line BA/F3 p210 was significantly inhibited at day 5 using DSG (2 μM; p=0.003) or DHSI-15 (4 μM; p=0.01) (Fig. 2a and b). Similar to Ba/F3 WT cells, effects on cell counts seen with CNI-1493 were marginal (Fig. 2c). These results demonstrate that the DHS inhibitors DHSI-15 and DSG have anti-proliferative effects and that these effects are independent from the presence or absence of BCR-ABL in the BA/F3 cell line, while CNI-1493 showed no growth inhibitory effect.

Ba/F3 p210 is an Imatinib sensitive cell line. In order to evaluate if Imatinib resistant cell lines would also be affected by DHS-inhibitor treatment, we used BA/F3 p210 M351T, BA/F3 p210 E255K and BA/F3 p210 T315I in

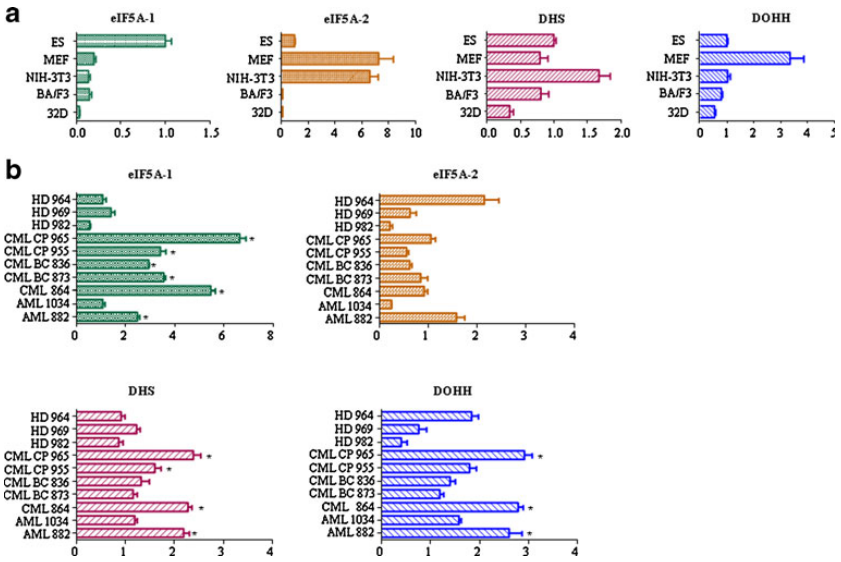


Fig. 1 eIF5A-1, DHS and DOHH but not eIF5A-2 are ubiquitously expressed in primary mouse cells and murine cell lines as well as in CD34+ cells from patients with CML and AML. **a** Quantitative expression analysis of eIF5A-1, eIF5A-2, deoxyhypusine synthase (DHS) and deoxyhypusine hydroxylase (DOHH) mRNA in mouse ES cells, MEFs, NIH-3 T3, BA/F3 WT and 32D cells. Expression levels were normalized against GAPDH and compared with the expression of ES cells (set to 1). Mean values±SEM of three experiments are shown. **b** mRNA expression of eIF5A-1, eIF5A-2, DHS and DOHH in CD34+ cells from healthy donors (HD) and patients with CML in chronic phase (CP) and blast crisis (BC), from one *BCR-ABL*-negative CML patient (864) and two AML patients. Expression levels were analyzed by quantitative RT-PCR, normalized against GAPDH and compared with the average expression of three healthy donors (HD, set to 1), (*$p < 0{,}05$). Mean values±SEM of three experiments are shown

proliferation assays. As shown in Fig. 2a and c, both the partially resistant BA/F3 p210 M351T and BA/F3 p210 E255K cell lines as well as the highly resistant BA/F3 p210 T315I cell line were significantly inhibited by DSG and DHSI-15 at concentrations similar to those effective in Ba/F3 p210, whereas CNI-1493 had only little effect on cell growth of these cell lines (Fig. 2c). These results suggested that mutations in the *bcr-abl* gene, rendering cells resistant to Imatinib, had no effect on DHS inhibition *in vitro*.

Pharmacologic effects of deoxyhypusine synthase-inhibitors and imatinib in BCR-ABL positive, imatinib sensitive cell lines

In order to test for potential pharmacological effects between DHS inhibitors and Imatinib, MTT assays were performed. Given the long half live of the eIF5A protein (~7 days) [31], cells were pre-incubated with DHS inhibitors for 6 days prior to adding Imatinib. We used the highly

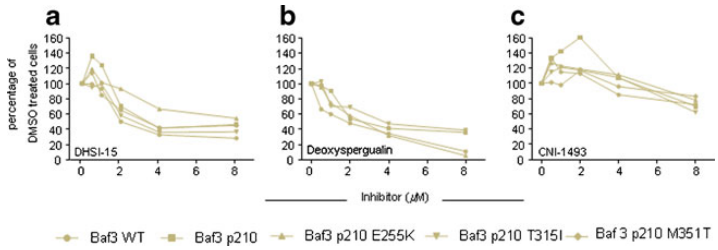


Fig. 2 Effect of direct inhibition of hypusination on proliferation of BA/F3 and BA/F3p210 cells including the mutants M351T, T315I, E255K. Cells were grown in the presence of DMSO and the indicated inhibitor concentrations or in the presence of DMSO alone for 5 days. The percentage of cell growth was calculated relative to the cell growth of DMSO treated cells, which was set to 100 %. Data shown are the mean of three independent experiments. A, B and C show inhibitors of deoxyhypusine synthase (DHSI-15, Deoxyspergualin, CNI-1493)

Imatinib sensitive Ba/F3 p210 as well as the BA/F3 p210 M351T, BA/F3 p210 E255K cell lines, the latter displaying different degrees of resistance against Imatinib. As shown in Fig. 3, pre-incubation with DSG, but not with DHSI-15, reduced the necessary concentration of Imatinib required for 50 % inhibition in all cell lines tested. In particular, the IC_{50} Imatinib for BA/F3 p210 without pre-incubation was 0.38 μM (BA/F3 p210 E255K: 7.3 μM; BA/F3 p210 M351T: 2.1 μM), while with DSG pre-incubation, it was reduced to 0.04 μM (BA/F3 p210 E255K: 5.2 μM; BA/F3 p210 M351T: 0.4 μM) (Fig. 3b, d, f), suggesting synergistic or additive interaction between Imatinib and DSG in Ba/F3 p210 and BA/F3 p210 M351T cell lines. DHSI-15 pre-incubation had no effect in lowering IC_{50} Imatinib in any of the cell lines tested (Fig. 3a, c, e). CNI-1493 was not tested due to its marginal effects on proliferation of BA/F3 cells (Fig. 2b). From these experiments, we conclude that only the DHS inhibitor DSG, but not DHSI-15, was able to sensitize the BCR-ABL positive cell lines for treatment with Imatinib.

DHSI-15 and deoxyspergualine induce apoptosis/necrosis in BA/F3 cell lines independent of the presence or absence of BCR-ABL

There is growing evidence that the unhypusinated form of the eIF5A protein is an inducer of apoptosis [32]. We therefore analyzed BA/F3 WT cells and its BCR-ABL positive derivates for DNA-fragmentation at day 6 and at day 13 during culture with DHS inhibitors. Incubation with DHSI-15 and DSG was strongly anti-proliferative and increased apoptosis or necrosis at late time points of the treatment. Apoptosis and necrosis was indicated as an increase in sub-G_1 DNA content during DHSI-15 treatment (Fig. 4a and b). Whereas at day 6 of DHSI-15 treatment cells showed mostly G_2/M growth arrest and only a few sub-G1 cells could be detected (Fig. 4a), at day 13, all cell lines tested displayed highly significant proportions of apoptotic/necrotic cells, independent of the presence or absence of BCR-ABL (Fig. 4a and b). Similar results could be obtained by incubating BA/F3 cells with DSG (Fig. 4b). At day 13 of DSG incubation, amounts of induction of apoptosis/necrosis were detectable in the following order: BA/F3 WT>BA/F3 p210>BA/F3 p210 T315I>BA/F3 p210 E255K>BA/F3 p210 M351T. We therefore conclude that the deoxyhypusine synthase inhibitors DSG and DHSI-15 have apoptotic/necrotic effects on the BA/F3 cell lines, independent of the presence or absence of BCR-ABL, most likely by increasing the cellular content of the unmodified form of eIF5A.

Inhibition of CD34+ cell growth with deoxyspergualin

Given the IC_{50}-Imatinib lowering capacity of DSG on BA/F3 cells in combined treatment protocols (Fig. 3b, d, f), we went on to test for effects of DSG on hematopoietic stem

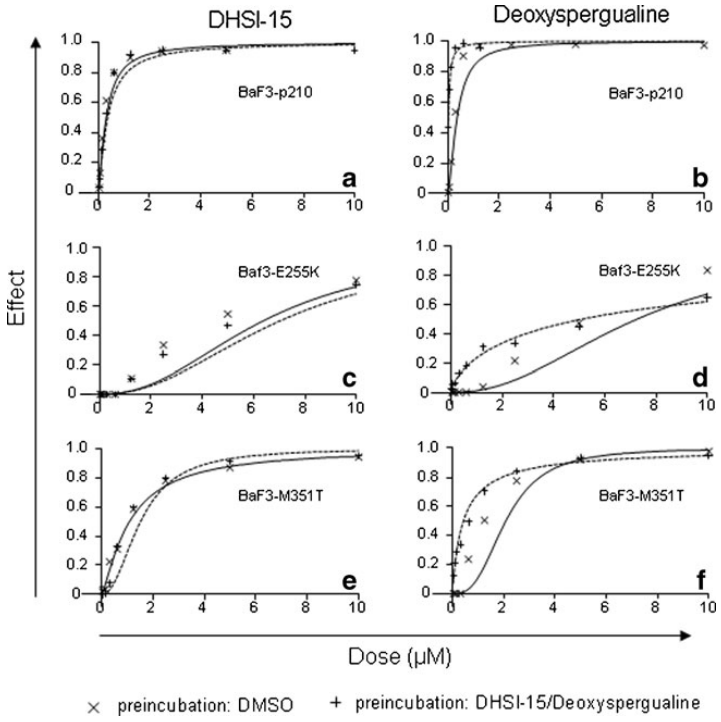


Fig. 3 Direct inhibition of hypusination does not synergize with Imatinib to induce cytotoxicity in BCR-ABL positive, imatinib sensitive BA/F3 cells. Dose-effect relationship for combined therapy with DHSI-15 or Deoxyspergualin and Imatinib in BA/F3 p210, BA/F3 T315I, BA/F3 E255K and BA/F3 M351T cells after 48 h. Cells were preincubated with hypusination inhibitors or DMSO only for 6 days prior to adding Imatinib. Effects on viability and cell growth were assessed using MTT assay. Results were calculated using Calcusyn software

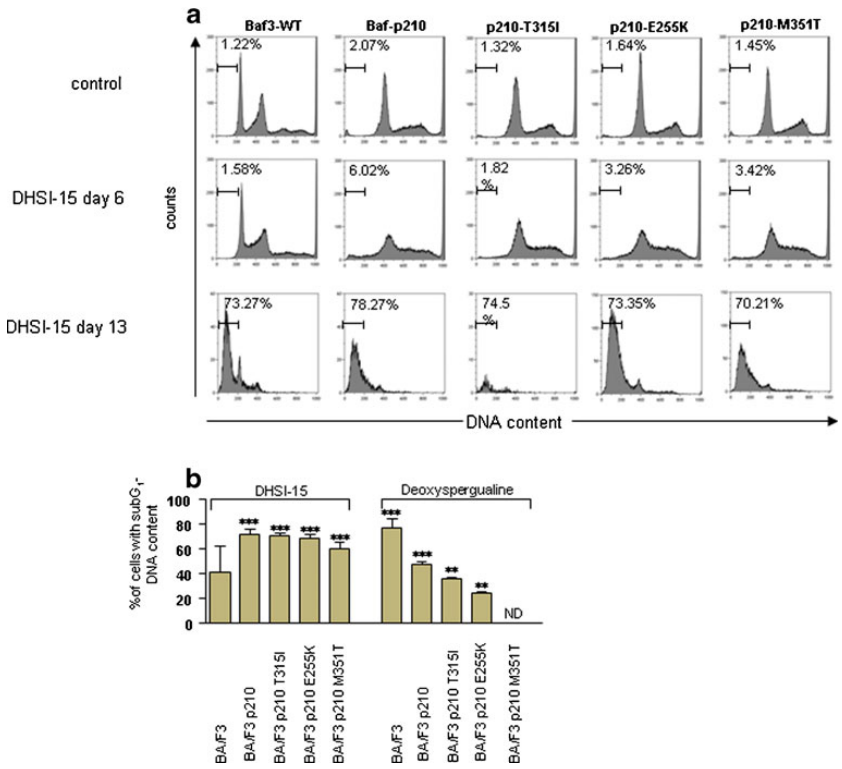


Fig. 4 Direct inhibition of hypusination induces DNA-fragmentation in BCR-ABL negative and positive BA/F3 cells in long term proliferation assays. BA/F3 WT, BA/F3 p210, BA/F3 T315I, BA/F3 E255K and BA/F3 M351T were grown in the presence of the deoxyhypusine synthase inhibitors DHSI-15 and Deoxyspergualine or in the presence of DMSO for 13 days. **a** FACS-plots show apoptotic and necrotic cell fractions and G_2/M arrest of propidium-iodide stained cells after 6 and 13 days of DHSI-15 or DMSO culture. **b** Percentages of apoptotic cells at day 13 are given as means (±SD) of three independent experiments. Statistical significant differences are indicated (***$p < 0.001$)

and progenitor cells. CD34+ cells purified from CML patients at diagnosis and healthy donors were subjected to proliferation assays. Cells were grown in the presence of increasing concentrations of DSG for 9 days and cells were counted at day 3, day 6 and day 9. As shown, DSG inhibited both cells from healthy donors as well as from patients with diagnosed CML in a time and concentration dependent manner (Fig. 5a and b). However, whereas effects on CD34+ cells from healthy donors were less pronounced at any concentration at day 3, inhibition increased at day 6 and day 9 and exceeded the effects seen on CD34+ CML cells. These data suggested that the effects of DSG on human stem and progenitor cell growth were independent on the presence of BCR-ABL. We then analysed the proliferation of CD34+ cells from the same sources under increasing concentrations of Imatinib (Fig. 5c and d). Cells were grown for 3 days and counted at day 1, day 2 and day 3. The degree of expansion as compared to untreated cells decreased for CD34+ CML cells in a dose and time dependent fashion. Similar results but less pronounced could be seen for CD34+ cells from healthy donors confirming inhibitory effects of Imatinib on normal progenitor cells in vitro [25]. By combining a fixed dose of DSG and increasing doses of Imatinib (Fig. 5e and f) additive anti-proliferative effects could be found in a very similar fashion between normal CD34+ cells and Imatinib as well as between CD34+ CML cells and Imatinib strongly arguing against a therapeutic window for this combination in CML compared with normal cells.

Discussion

Continuous treatment of chronic myeloid leukemia with Imatinib is able to induce long lasting responses in a large cohort of patients [33, 34]. To this end Imatinib is the standard therapeutic regimen for patients with newly diagnosed CML. However there are three major problems associated with long-term Imatinib treatment: 1. Primary or acquired resistance to Imatinib during therapy. 2. Limited effects of Imatinib in accelerated phase or blast crisis CML. 3. Evolution of minimal residual disease due to persistence of BCR-ABL positive hematopoietic stem and progenitor cells [35, 36]. Therefore, the development of new drugs supporting or complementing Imatinib therapy alongside with the development of novel tyrosine kinase inhibitors with an enhanced spectrum of BCR-ABL activity such as Nilotinib, Dasatinib or Bosutinib are clearly needed [37].

Previously, we showed that the inhibition of eIF5A either by siRNA or by using N^1-guanyl-1,7-diaminoheptan (GC7) has anti-proliferative effects in leukemias [10]. Since GC7 is

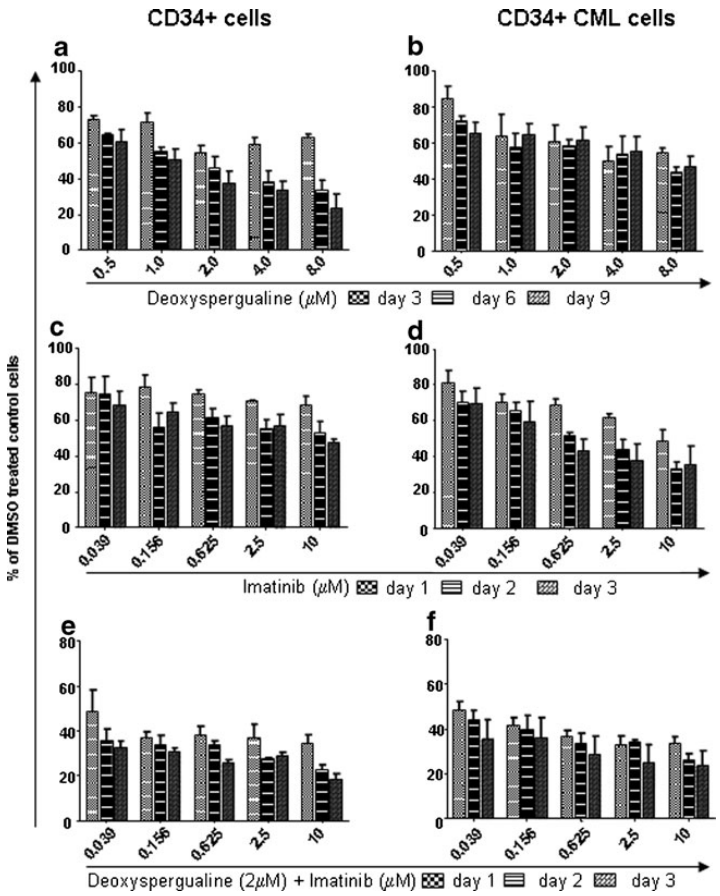


Fig. 5 Effects of Deoxyspergualine and Imatinib on progenitor cells from healthy donors and CML patients at diagnosis. CD34+ cells from healthy donors or from CML patients in the untreated chronic phase of CML were incubated with increasing doses of Deoxyspergualine (**a**, **b**) and assayed at day 3, 6 and 9. Same cells were subjected to increasing doses of Imatinib (**c**, **d**) or increasing doses of Imatinib and a fixed dose of Deoxyspergualine treatment (**e**, **f**). Cells were analysed at day 1, day 2 and day 3. Bar graphs represent the mean percentage (±SD) of cell growth as calculated relative to the cell growth of DMSO treated cells, which was set to 100 %

not suitable for the treatment of patients due to its pharmacokinetic characteristics, we screened the known DHS–inhibitors deoxyspergualin, CNI-1493 and the new DHS-inhibitor DHSI-15, for their anti-proliferative effects in leukemic cells.

Deoxyspergualin and CNI-1493 are potent compounds which are both in clinical use. Deoxyspergualin was synthesized from spergualin in 1982 [38]. Deoxyspergualin has been shown to inhibit the growth of activated naïve T-cells and its clinical use is dicussed for a variety of hyperreactive inflammatory diseases [39]. Inhibition of deoxyhypusine synthase by deoxyspergualin was first described in 2002 [21].

CNI-1493 (Semapimod) was developed as an inhibitor of arginine transport and nitric oxide production in macrophages [22]. It prevents acute inflammation and endotoxin lethality. CNI-1493 was tested in a preliminary clinical trial with patients suffering from severe Crohn's disease, where a clinical response was seen in 67 % of the patients at week 4 of treatment [40]. In 2004, CNI-1493 was identified in a screening assay for the identification of deoxyhypusine synthase inhibitors [19]. By comparing all three deoxyhypusine synthase inhibitors in parallel, we see limited effects with CNI-1493 in all murine cell lines tested. In contrast to that, the results we obtain with DHSI-15 and deoxyspergualin are promising as both compounds exert strong cytostatic and proapoptotic effects *in vitro*, independent of the presence and/or the mutational state of BCR-ABL. This is in agreement with the finding that the unhypusinated eIF5A is the pro-apoptotic form and that its accumulation during long-term DHS- inhibitor treatment is detrimental for a cell. In fact, it has been shown that mutant forms of eIF5A which are not capable of being hypusinated, induce loss of mitochondrial transmembrane potential, release of cytochrome c and caspase activation [41].

However, although pre-treatment with DSG lowered the IC_{50} of Imatinib in hematopoietic cells, we are not able to observe pharmacologic synergism of IM and DHSI-15 in murine cell lines with a subcomplete resistance to IM. This highly suggests that DHSI-15 does not rank in this respect with the previously described DHS-Inhibitor N^1-guanyl-1,7-diaminoheptan and the compared DHSI- inhibitor DSG. Additionally and in agreement with the experiments using

murine cell-lines, DSG showed efficacy in primary CML cells as well as in CD34+ cells derived from healthy donors.

We therefore hypothesize that although DSG and DHSI-15 target BCR-ABL positive as well as negative cells *in vitro*, due to an obvious lack of a therapeutic window observed in BCR-ABL-positive as opposed to normal CD34+ hematopoietic stem and progenitor cells, a clinical use of these compounds alone or in combination is unlikely to be demonstrated. This is further underscored by the deep responses which can be achieved with 1st (Imatinib), 2nd (Nilotinib, Dasatinib and Bosutinib) [42–44] and emerging 3rd generation (e.g. Ponatinib) [45] tyrosine kinase inhibitors (TKI) in both first and second line treatment and the promising *in vitro* data on combination therapy of TKI with already clinically available drugs such as JAK2-inhibitors, SMO inhibitors or autophagy inhibitors (e.g. chloroquine). Taken together, despite of their specific mode of action and pronounced anti-proliferative activity, hypusination inhibitors (HI) are unlikely to add to the treatment options in chronic phase CML. On the contrary, HI might be of potential use for the treatment of either late stage CML (accelerated phase or blast crisis) or in AML and/or BCR-ABL+/− ALL as a mere cytoreductive principle alone or in combination with chemotherapy in the future.

Acknowledgements A START Grant of the Medical Faculty of RWTH Aachen University (PZ) and a network-grant of the German Federal Ministry of Education and Research (BMBF; grant No. 01GUO715-717 to THB, JH, CM and MR) supported this work.

Conflict of interest The authors declare that they have no conflict of interest

Ethical standards The experiments comply with the current laws of Germany

References

1. Champlin RE, Golde DW (1985) Chronic myelogenous leukemia: recent advances. Blood 65(5):1039–1047
2. Lobo NA et al (2007) The biology of cancer stem cells. Annu Rev Cell Dev Biol 23:675–699
3. Sattler M, Griffin JD (2003) Molecular mechanisms of transformation by the BCR-ABL oncogene. Semin Hematol 40(2 Suppl 2):4–10
4. Sawyers CL et al (2002) Imatinib induces hematologic and cytogenetic responses in patients with chronic myelogenous leukemia in myeloid blast crisis: results of a phase II study. Blood 99 (10):3530–3539
5. Talpaz M et al (2002) Imatinib induces durable hematologic and cytogenetic responses in patients with accelerated phase chronic myeloid leukemia: results of a phase 2 study. Blood 99(6):1928–1937
6. O'Brien SG et al (2003) Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. N Engl J Med 348(11):994–1004
7. Capdeville R et al (2002) Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. Nat Rev Drug Discov 1 (7):493–502
8. Deininger M, Buchdunger E, Druker BJ (2005) The development of imatinib as a therapeutic agent for chronic myeloid leukemia. Blood 105(7):2640–2653
9. O'Hare T, Corbin AS, Druker BJ (2006) Targeted CML therapy: controlling drug resistance, seeking cure. Curr Opin Genet Dev 16 (1):92–99
10. Balabanov S et al (2007) Hypusination of eukaryotic initiation factor 5A (eIF5A): a novel therapeutic target in BCR-ABL-positive leukemias identified by a proteomics approach. Blood 109(4):1701–1711
11. Park MH, Cooper HL, Folk JE (1982) The biosynthesis of protein-bound hypusine (N epsilon -(4-amino-2-hydroxybutyl)lysine). lysine as the amino acid precursor and the intermediate role of deoxyhypusine (N epsilon -(4-aminobutyl)lysine). J Biol Chem 257(12):7217–7222
12. Gordon ED et al (1987) Eukaryotic initiation factor 4D, the hypusine-containing protein, is conserved among eukaryotes. J Biol Chem 262(34):16585–16589
13. Bartig D et al (1992) The archaebacterial hypusine-containing protein. Structural features suggest common ancestry with eukaryotic translation initiation factor 5A. Eur J Biochem 204(2):751–758
14. Li CH et al. eIF5A promotes translation elongation, polysome disassembly and stress granule assembly. PLoS One 5(4):e9942.
15. Kruse M et al (2000) Inhibition of CD83 cell surface expression during dendritic cell maturation by interference with nuclear export of CD83 mRNA. J Exp Med 191(9):1581–1590
16. Maier B et al. The unique hypusine modification of eIF5A promotes islet beta cell inflammation and dysfunction in mice. J Clin Invest 120(6): 2156–70.
17. Hart RA et al (2002) Effects of 1,8-diaminooctane on the FIV Rev regulatory system. Virology 304(1):97–104
18. Katahira J et al (1995) Effects of translation initiation factor eIF-5A on the functioning of human T-cell leukemia virus type I Rex and human immunodeficiency virus Rev inhibited trans dominantly by a Rex mutant deficient in RNA binding. J Virol 69(5):3125–3133
19. Sommer MN et al (2004) Screening assay for the identification of deoxyhypusine synthase inhibitors. J Biomol Screen 9(5):434–438
20. Hauber I et al (2005) Identification of cellular deoxyhypusine synthase as a novel target for antiretroviral therapy. J Clin Invest 115(1):76–85
21. Nishimura K et al (2002) Inhibition of cell growth through inactivation of eukaryotic translation initiation factor 5A (eIF5A) by deoxyspergualin. Biochem J 363(Pt 3):761–768
22. Bianchi M et al (1995) An inhibitor of macrophage arginine transport and nitric oxide production (CNI-1493) prevents acute inflammation and endotoxin lethality. Mol Med 1(3):254–266
23. Ulrich P, Cerami A (1984) Trypanocidal 1,3-arylene diketone bis (guanylhydrazone)s. Structure-activity relationships among substituted and heterocyclic analogues. J Med Chem 27(1):35–40
24. Balabanov S et al. Abcg2 overexpression represents a novel mechanism for acquired resistance to the multi-kinase inhibitor Danusertib in BCR-ABL-positive cells in vitro. PLoS One 6(4): e19164.
25. Bartolovic K et al (2004) Inhibitory effect of imatinib on normal progenitor cells in vitro. Blood 103(2):523–529
26. Gontarewicz A et al (2008) Simultaneous targeting of aurora kinases and Bcr-Abl kinase by the small molecule inhibitor PHA-739358 is effective against imatinib-resistant BCR-ABL mutations including T315I. Blood 111(8):4355–4364
27. Hartmann U et al (2005) Telomere length and telomerase activity in the BCR-ABL-transformed murine Pro-B cell line BaF3 is unaffected by treatment with imatinib. Exp Hematol 33(5):542–549

28. Park MH, Wolff EC, Folk JE (1993) Is hypusine essential for eukaryotic cell proliferation? Trends Biochem Sci 18(12):475–479
29. Clement PM et al (2003) Identification and characterization of eukaryotic initiation factor 5A-2. Eur J Biochem 270(21):4254–4263
30. Clement PM et al (2006) Differential expression of eIF5A-1 and eIF5A-2 in human cancer cells. FEBS J 273(6):1102–1114
31. Nishimura K et al (2005) Independent roles of eIF5A and polyamines in cell proliferation. Biochem J 385(Pt 3):779–785
32. Li AL et al (2004) A novel eIF5A complex functions as a regulator of p53 and p53-dependent apoptosis. J Biol Chem 279(47):49251–49258
33. Capdeville R et al (2008) Report of an international expanded access program of imatinib in adults with Philadelphia chromosome positive leukemias. Ann Oncol 19(7):1320–1326
34. Hochhaus A et al (2009) Six-year follow-up of patients receiving imatinib for the first-line treatment of chronic myeloid leukemia. Leukemia 23(6):1054–1061
35. Branford S, Hughes T (2006) Detection of BCR-ABL mutations and resistance to imatinib mesylate. Methods Mol Med 125:93–106
36. Zonder JA, Schiffer CA (2006) Update on practical aspects of the treatment of chronic myeloid leukemia with imatinib mesylate. Curr Hematol Malig Rep 1(3):141–151
37. O'Hare T et al. Targeting the BCR-ABL signaling pathway in therapy-resistant Philadelphia chromosome-positive leukemia. Clin Cancer Res 17(2): 212–21.
38. Iwasawa H et al (1982) Synthesis of (−)-15-deoxyspergualin and (−)-spergualin-15-phosphate. J Antibiot (Tokyo) 35(12):1665–1669
39. Valdivia LA et al (1991) Suppressor cells induced by donor-specific transfusion and deoxyspergualin in rat cardiac xenografts. Transplantation 52(4):594–599
40. Hommes D et al (2002) Inhibition of stress-activated MAP kinases induces clinical improvement in moderate to severe Crohn's disease. Gastroenterology 122(1):7–14
41. Sun Z et al. Apoptosis induction by eIF5A1 involves activation of the intrinsic mitochondrial pathway. J Cell Physiol 223(3): 798–809.
42. Saglio G et al. Nilotinib versus imatinib for newly diagnosed chronic myeloid leukemia. N Engl J Med 362(24): 2251–9.
43. Kantarjian H et al. Dasatinib versus imatinib in newly diagnosed chronic-phase chronic myeloid leukemia. N Engl J Med 362(24): 2260–70.
44. Cortes JE et al. Safety and efficacy of bosutinib (SKI-606) in chronic phase Philadelphia chromosome-positive chronic myeloid leukemia patients with resistance or intolerance to imatinib. Blood 118(17): 4567–76.
45. O'Hare T et al (2009) AP24534, a pan-BCR-ABL inhibitor for chronic myeloid leukemia, potently inhibits the T315I mutant and overcomes mutation-based resistance. Cancer Cell 16(5):401–412

A.3 Kolodzik, A.; Urbaczek, S.; Rarey, M. Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies. *J. Chem. Inf. Model.* 2012, *52*, 2013–2021

ACS direct link:
http://pubs.acs.org/articlesonrequest/AOR-WBYvgICi3nrRd4gG9DNN

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

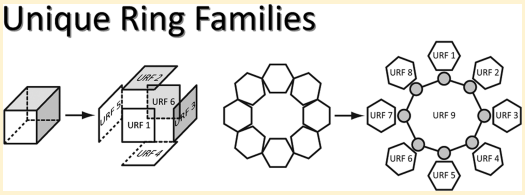
Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies

Adrian Kolodzik,[†,‡] Sascha Urbaczek,[†] and Matthias Rarey*[,†]

[†]Center for Bioinformatics (ZBH), University of Hamburg, Bundesstr. 43, 20146 Hamburg, Germany

Supporting Information

ABSTRACT: The perception of a set of rings forms the basis for a number of chemoinformatics applications, e.g. the systematic naming of compounds, the calculation of molecular descriptors, the matching of SMARTS expressions, and the generation of atomic coordinates. We introduce the concept of unique ring families (URFs) as an extension of the concept of relevant cycles (RCs).[1,2] URFs are consistent for different atom orders and represent an intuitive description of the rings of a molecular graph. Furthermore, in contrast to RCs, URFs are polynomial in number. We provide an algorithm to efficiently calculate URFs in polynomial time and demonstrate their suitability for real-time applications by providing computing time benchmarks for the PubChem Database.[3] URFs combine three important properties of chemical ring descriptions, for the first time, namely being unique, chemically meaningful, and efficient to compute. Therefore, URFs are a valuable alternative to the commonly used concept of the smallest set of smallest rings (SSSR) and would be suited to become the standard measure for ring topologies of small molecules.



INTRODUCTION

Ring perception is a crucial step in many chemoinformatics applications, including the calculation of molecular descriptors, the matching of SMARTS expressions, and the generation of two- and three-dimensional atomic coordinates. In order to obtain consistent results, a set of rings has to be unique in the sense that it depends only on the molecule's topology. Efficient algorithms and ring perception concepts that lead to a limited number of cycles provide the means for interactive applications. Chemically meaningful rings allow for an easy analysis and interpretation of the resulting set of rings. Due to their high relevance in chemistry, several computational methods for automatic ring perception have been developed over the past 35 years.[4] Each of these methods has deficiencies in being either not unique or not polynomial in number or not chemically meaningful. The paper of Berger et al.[5] impressively demonstrates this for a number of ring perception concepts including the widely used SSSR.[4]

A molecule can be interpreted as a simple, connected, unweighted and undirected graph $G = (V,E)$ where the atoms are interpreted as a set of vertices V and bonds are considered a set of edges E. A cycle is a subgraph of G such that any vertex degree is exactly two. A connected cycle is called elementary. Since elementary cycles meet our expectation of rings in a molecular graphs we will use the terms elementary cycle and ring synonymously. $E(v_1,v_2)$ is the edge connecting the vertices v_1 and v_2. For the set of vertices or edges of a cycle (or a general subgraph) C, we will write $V(C)$ and $E(C)$, respectively. A cycle C containing the edges $E(C)$ has a length of $|C|$ which is equal to its number of edges $|E(C)|$. It can be described by the incidence vector of its edges. A cycle with n edges is called n-cycle. A connected n-cycle is called n-ring. A chord is an edge e connecting two vertices of a ring C with $e \notin E(C)$. A ring is chord-less if it has no chord. Cycles can be combined to larger ones by forming the symmetric difference of their edges; this operation is considered the "addition" of cycles. In order to describe the addition of cycles, we utilize the xor operator $\oplus$ in agreement with the nomenclature used by Berger et al.[5] Thus, the addition of two cycles C_A and C_B that forms the cycle C_C will be written as $C_A \oplus C_B = C_C$. All cycles of G form the cycle space $S(G)$. A cycle base $B(G)$ is a subset of $S(G)$ that allows to construct all cycles of $S(G)$ by the addition operation. The length of $B(G)$ is equal to the sum of the lengths of its cycles. All cycles of a cycle base are elementary.

In the following, we will discuss common concepts of ring perception in order to motivate our new approach. The **set of all rings**[6] (Ω) includes all elementary rings of a molecular graph and efficient algorithms for its calculation have been developed. The number of rings and the computational runtimes, however, grow dramatically for complex ringsystems. Additionally, not all resulting rings are meaningful in a chemical context, and Ω is, thus, a unique description that is neither chemically meaningful nor polynomial in size.

The most frequently applied strategy of ring perception is the calculation of the **smallest set of smallest rings**[4] (SSSR) which is a subset of Ω. An SSSR represents a minimum cycle base (MCB). It contains a polynomial number of rings and can be calculated in polynomial time.[7] If a molecular graph contains only a single MCB, the SSSR is unique and intuitive. If this is

Received: December 31, 2011
Published: July 10, 2012

not the case, the resulting SSSR is arbitrary and depends on the specific algorithm used for its construction. Furthermore, the selected SSSR often depends on the input atom order.[8]

The problems arising from nonunique ring descriptions can be exemplified with SMARTS pattern matching. According to page 20 of the Daylight Theory Manual,[9] the SMARTS pattern [R3] describes an atom which is part of three SSSR rings. The matching of this SMARTS pattern on the highly symmetric molecule cubane (SMILES = C12C3C4C1C5C2C3C45) using the Daylight web service[10] illustrates the problems arising from the SSSR's lack of uniqueness (see Figure 1). Any combination

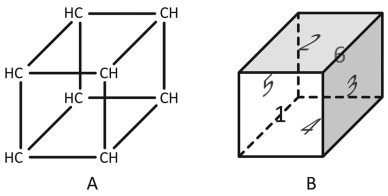


Figure 1. Cubane contains six alternative MCBs. Each combination of five of the 4-rings forms an SSSR.

of five of the shown 4-rings forms a valid SSSR. The sixth ring can be constructed by adding the rings of the SSSR.[11] Consequently, the SMARTS pattern [R3] only matches four of the eight equivalent carbon atoms depending on the selected SSSR.

The **essential set of essential rings** (ESER)[12] and the approaches published by Corey[13] and Wipke[14] try to perceive chemically meaningful rings by calculating an MCB and adding rings up to a certain size or rings including certain elements. Due to their heuristic nature, these approaches lack a mathematical foundation and are not suitable for all kinds of molecular graphs.[5]

In addition, there is a number of graph theoretical ring perception concepts which are limited to planar graphs. The **minimum planar cycle base** and the **extended set of smallest rings**[15] are examples of such concepts. Since molecular graphs are not necessarily planar, these ring perception concepts are of limited use for general applications in chemoinformatics.

The **set of β-rings**[16] is defined on a plane embedding of a molecular graph. The chord-less faces of the embedding are processed by increasing size. The set of β-rings includes all faces representing 3-rings or 4-rings. Additionally, it contains all faces which are linearly independent of three or less shorter faces already contained in the set. Berger et al.[5] suggested to use the **β*-rings** instead. These rings are calculated on all chord-less rings of a graph instead of the chord-less faces of a specific plane embedding. In contrast to the set of β-rings, the set of β*-rings is unique but contains an exponential number of rings.

An additional set of rings which is defined for general graphs is the **set of smallest cycles at edges** (SSCE).[17] The SSCE is calculated on the basis of Ω by recursively deleting all edges included in more than one ring. The SSCE does, however, not necessarily contain a cycle base. Consequently, it does not provide a complete description of the rings of a molecular graph.

Relevant cycles[1,2] (RCs) are defined as the union of all MCBs. They comprise a unique set of rings and an intuitive description of most molecular graphs. Some molecules, however, contain an exponential number of RCs. Examples are cyclophane-like structures which will be discussed in more detail in the following sections.

To tackle the exponential number of rings, Gleiss et al.[11] suggested to classify RCs into **interchangeability classes** (ICs). ICs are calculated by dividing RCs into essential and interchangeable rings. An essential ring is included in all MCBs. Rings which are not essential are called interchangeable. An IC contains either a single essential ring or all interchangeable rings which can be constructed from a subset of the IC and shorter cycles. While treating the rings of an interchangeability class as a union can be suitable for the prediction of RNA secondary structures, this concept is not generally applicable in chemoinformatics. For example, the description of the six RCs of cubane or the 6-rings of fulleren as single ICs is too coarse for most applications and, especially in the case of fullerene, it is not chemically meaningful.

Relevant cycle families (RCFs)[1] are conceptually similar to ICs. An RCF contains all RCs generated on the basis of a single relevant cycle prototype (RCP). RCPs are not unique and their number depends on the order of the molecule's atoms. Since each RCP results in an RCF, the RCFs are also not unique and their number can vary for a molecule.

None of the mentioned concepts of ring perception efficiently calculate a complete and polynomial set of unique and chemically intuitive rings for molecular graphs. We introduce the concept of **unique ring families** (URFs), which meets all of these requirements.

UNIQUE RING FAMILIES

Generation of Relevant Cycles. Since unique ring families (URFs) are defined on the basis of RCs, we provide a short outline of Vismara's RC detection algorithm.[1] The perception of RCs involves five consecutive steps which are explained below (see Figure 2):

1. Calculate all 2-connected components of the molecular graph G.
2. For each 2-connected component, calculate the shortest paths from each vertex r to each other vertex, only passing through vertices which follow r in an arbitrary but fixed order π.
3. Calculate RCPs by combining pairs of shortest paths of identical size starting from the same vertex r.
4. Eliminate RCPs which linearly depend on strictly smaller cycles with respect to cycle addition.
5. Calculate RCs by a backtracking procedure on the basis of the RCPs.

2-Connected components of the molecular graph can be calculated using the algorithm published by Tarjan.[18] The 2-connected components will be called ringsystems in the following sections. An order π of the vertices is established by sorting them according to their degree in descending order. Vertices of identical degree are ordered arbitrarily. This ordering guarantees polynomial runtime complexity for the calculation of RCPs. In the second step, a breadth-first-search is used to calculate a single shortest path $P(r,t)$ from each vertex r to each other vertex through vertices following r in the ordering π. Thus, only paths through vertices of equal or lower degree are considered. If two shortest paths $P(r,p)$ and $P(r,q)$ of identical size solely share the vertex r, and if furthermore p and q are directly connected by an edge, an uneven ring is identified. If p and q are both directly connected to a vertex z which is neither a member of $P(r,p)$ nor a member of $P(r,q)$, an

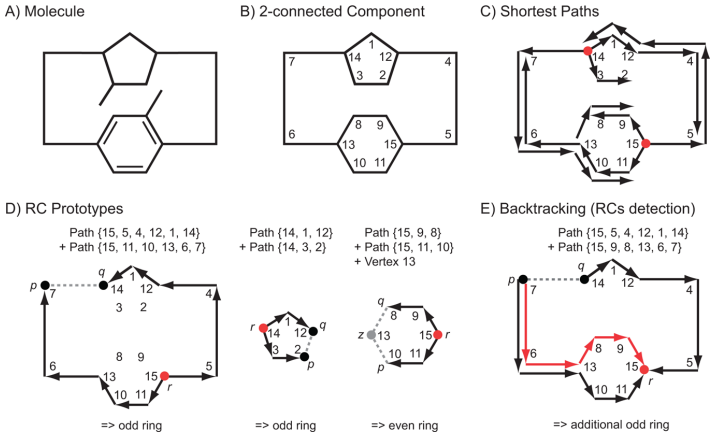


Figure 2. Process to identify the RCs of a molecular graph. (A) At first, 2-connected components are calculated (B) and vertices are ordered according to their degree. Vertices of higher degree are labeled with higher numbers than vertices of lower degree. (C) Shortest paths only passing through vertices following r in the order π are calculated from each vertex r to each other vertex of the graph (shown for vertices 14 and 15). (D) The polynomial number of RCPs are calculated on the basis of the identified shortest paths. Two shortest paths of equal lengths which only share the vertex r form an uneven RCP if their end points (p, q) are adjacent. If they share an adjacent vertex z, they form an even RCP. (E) RCs are enumerated on the basis of RCPs by combining alternative shortest paths (red arrows) connecting p or q to r.

even RCP is identified. The length of the shortest paths used to identify an RCP of size n is therefore given by the following equation:

$$|E(P(r,p))| = |E(P(r,q))| = \begin{cases} \frac{n-1}{2} & \text{if } n \text{ is odd} \\ \frac{n}{2} - 1 & \text{if } n \text{ is even} \end{cases} \quad (1)$$

As described above, only a single shortest path is considered for each pair of vertices. Multiple shortest paths connecting two vertices may exist, however. Thus, the polynomial number of RCPs represent only a subset of the exponential number of RCs. To identify all RCs on the basis of the RCPs, Vismara's algorithm uses a backtracking procedure. The set of RCs calculated during backtracking on the basis of a single RCP is defined as an RCF. This backtracking procedure includes the following steps:

First, for each RCP the set S_p of all shortest paths from p to r and the set S_q of all shortest paths from q to r are calculated. If an RCP is uneven, each combination of $P(r,p) \in S_p$ and $P(r,q) \in S_q$ forms an uneven RC with the edge $E(p,q)$ (see, for example, the 11-ring in Figure 2E). If an RCP is even, each combination of $P(r,p) \in S_p$ and $P(r,q) \in S_q$ forms an even RC with the edges $E(p,z)$ and $E(q,z)$.

Note that all RCs of an RCF have the same size. If their size is uneven, they share at least the vertices r, p, and q and the edge $E(p,q)$. Otherwise, they share at least the vertices r, p, q, and z and the edges $E(p,z)$ and $E(q,z)$. All RCFs of a molecular graph are disjoint with respect to their rings and their union forms the set of all RCs of a graph. In the following, the RCF of a ring C_x will be called RCF_x. Furthermore, we will write $E(RCF_x)$ and $V(RCF_x)$ to denote the union of the edges or vertices of all rings of an RCF_x, respectively.

Introduction of Unique Ring Families. On the basis of the RCs of a molecular graph, we define the terms URF-pair-related and URF-related as follows:

Definition 1. Let C_1 and C_2 be two RCs of a graph G, then C_1 and C_2 are *URF-pair-related* if and only if all of the following conditions hold:

1. $|C_1|=|C_2|$
2. $E(C_1) \cap E(C_2) \neq \emptyset$
3. It exists a set S of strictly smaller rings in G such that $C_1 \oplus (\oplus_{c \in S} c) = C_2$

Definition 2 . The URF-relation is defined as the transitive closure of the *URF-pair-relation*. A URF is defined as the set of URF-related RCs and hence represents an equivalence class. The length |URF| is defined as the length of each of its RCs. The number of URFs of a graph is called *URF-number*.

For an efficient calculation of molecular ring topologies in case of complex ringsystems, a description of rings should be at most polynomial in number with respect to the size of the graph. In the following, we estimate the URF-number of a molecular graph by comparing it to the polynomial number of RCFs.

Theorem 1. Any two rings of an RCF are URF-related.

Due to the construction of RCFs as described above, any two RCs of an RCF have identical lengths and share at least either an edge $E(p,q)$ or the edges $E(p,z)$ and $E(q,z)$. Thus, all rings of an RCF meet conditions 1 and 2 of definition 1. Furthermore, the RCs of an RCF differ only by alternative shortest paths replacing $P(r,p)$ or $P(r,q)$. As a consequence of eq 1, the following equation describes the length of two shortest paths used to construct an RCP of size n:

$$|P(r,p) \in S_p| = |P(r,q) \in S_q| < \frac{n}{2} \quad (2)$$

Since $P(r,p)$ contains less than half of the edges of the RCP, the symmetric difference of any two alternative shortest paths of S_p forms a set of rings which are smaller than n. Since the same is true for any two alternative paths of S_q, each two rings of an RCF can be constructed by cycle addition of each other and a set of smaller rings. Hence, all rings of an RCF meet condition 3 of definition 1. Consequently, any two RCs of an

RCF are URF-related and the URF-number is less or equal to the number of RCFs. Since the number of RCPs and RCFs is polynomial according to Theorem 4 of Vismara's paper,[1] the URF-number is at most polynomial, too.

Calculation of URFs. In the following, we provide an algorithm to calculate the polynomial number of URFs in polynomial time on the basis of the RCPs. The algorithm uses the described properties of RCPs as well as their linear dependency with respect to cycle addition in order to describe URFs by their edges sets.

Lemma 1 . Let C_A and C_B be two URF-related RCs, then C_A and C_B linearly depend on each other and a set of smaller rings with respect to cycle addition.

According to condition 3 of definition 1, two URF-pair-related RCs linearly depend on each other and a set of smaller rings with respect to cycle addition. Since a URF consists of the transitive closure of the URF-pair-relation, any two URF-related RCs linearly depend on each other and a set of smaller rings. Thus, URFs can be calculated in three steps.

1. Calculate RCPs according to Vismara's algorithm.
2. Let $B_<(G)$ be a subset of a minimum cycle basis B with $B_<(G) = \{C \in B | |C| < |C_A| = |C_B|\}$. Identify all 2-pairs of RCPs (C_A, C_B) with

$$C_A \oplus \left(\bigoplus_{c \in B_<(G)} c \right) = C_B \tag{3}$$

   Note that this operation is already performed during the calculation of RCPs. In Vismara's ring construction algorithm, a Gaussian elimination is used to eliminate rings which depend on smaller rings. Any ring C_A which depends on smaller and equal sized rings is marked as relevant. If the set of equal sized rings on which C_A depends on, only consists of a single ring C_B, C_A and C_B are marked as potentially URF-related. Furthermore, please note that any two rings of $RCF_A \cup RCF_B$ meet conditions 1 and 3 for being URF-pair-related.
3. If any two rings of RCF_A and RCF_B share an edge, these two rings are URF-pair-related. Since the URF-relation is an equivalence relation, C_A and C_B are URF-related if

$$E(\mathrm{RCF_A}) \cap E(\mathrm{RCF_B}) \neq \emptyset \tag{4}$$

In order to calculate RCPs according to Vismara's algorithm, rings which linearly depend on strictly smaller rings are eliminated. If a ring depends linearly on rings of the same size and strictly smaller rings, it is marked as relevant. All RCs of identical size which are identified in this step to be linearly dependent on each other and a set of smaller rings form pairs of possibly URF-related RCPs. For each RCP, all edges and vertices belonging to the same RCF can be identified using a simple breadth first search starting from r followed by a backtracking procedure involving the following steps:

1. Starting from r each vertex v is labeled according to its distance d_v to r using a breadth-first-search.
2. E_{cur} and V_{cur} represent the vertices and edges currently identified as belonging to E_{RCF} and V_{RCF}, respectively. V_{cur} is initialized with $V_{cur} \leftarrow \{p,q,z\}$ if C_A has even size and $V_{cur} \leftarrow \{p,q\}$ if C_A has uneven size. E_{cur} is initialized with $E_{cur} \leftarrow \{E(p,z),E(q,z)\}$ if C_A has even size and $E_{cur} \leftarrow \{E(p,q)\}$ if C_A has uneven size. A list Q of vertices is initialized with $Q \leftarrow \{p,q\}$.
3. For a vertex $v_{cur} \in Q$ identify each directly connected vertex v_{adj}. If $d_{vcur} - 1 = d_{v_{adj}}$, then
   - $E_{cur} \leftarrow E_{cur} \cup E(v_{cur}, v_{adj})$
   - $V_{cur} \leftarrow V_{cur} \cup (v_{adj})$
   - $Q \leftarrow Q \cup v_{adj}$ if $v_{adj} \notin V_{curr}$
4. $Q \leftarrow Q \backslash v_{cur}$.
5. If $Q = \emptyset$, then $E_{cur} = E(\mathrm{RCF_A})$ and $V_{cur} = V(\mathrm{RCF_A})$. Otherwise, continue with step 3.

For a connected graph containing $|E|$ edges and $|V|$ vertices, RCPs can be calculated in $O(Z|E|^3)$ with $Z = |E| - |V| + 1$ being the cyclomatic number of G.[1] A Gaussian elimination to identify RCPs of identical size, which depend on each other and strictly smaller rings, can be performed in $O(|E|R^2)$ operations with R being the number of RCPs. The sets of edges belonging to each RCF are calculated in $O(|E|R)$. Finally, the edge set intersections of all 2-pairs of RCFs can be calculated in $O(|E|R^2)$. According to Visamara[1] the number of RCPs (R) is limited by the following relation:

$$R \leq 2|E|^2 + Z|V| \Rightarrow R \leq 2|E|^2 + |E||V| \tag{5}$$

Consequently, the Gaussian elimination and the calculation of the edge intersection of 2-pairs of RCPs are the speed-limiting steps and URFs can be perceived in $O(|E|^5+|V|^2)$. Thus, URFs represent a polynomial description of the ring topologies of a molecular graph and can be calculated in polynomial time.

Interpretation of URFs. From a chemical perspective, URFs can be best understood by calculating the union of the edges of all URF-related rings. Since a URF can contain smaller URFs, it can be illustrated by merging these smaller URFs to single nodes. This illustration represents a quotient graph of the partition of smaller URFs. Examples of molecular graphs and their corresponding RCs, RCPs and URFs are shown in Figures 3 and 4.

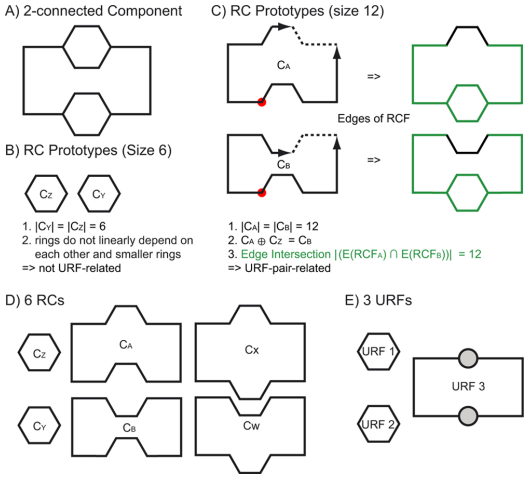


Figure 3. Ring system (A) containing 2 RCPs of size 6 (B) and 2 RCPs of size 12 (C). The two small rings form individual URFs (E). The two 12-rings belong to the same URF since they have the same size, share edges, and are linearly dependent on each other and one of the 6-rings. The molecular graph contains a total of six RCs (D) and three URFs (E). The URFs are illustrated as a quotient graph with the smaller URFs merged to individual nodes.

Compared to common strategies of ring perception, URFs have the major advantages that they are unique, intuitive, polynomial in number and provide a complete description of

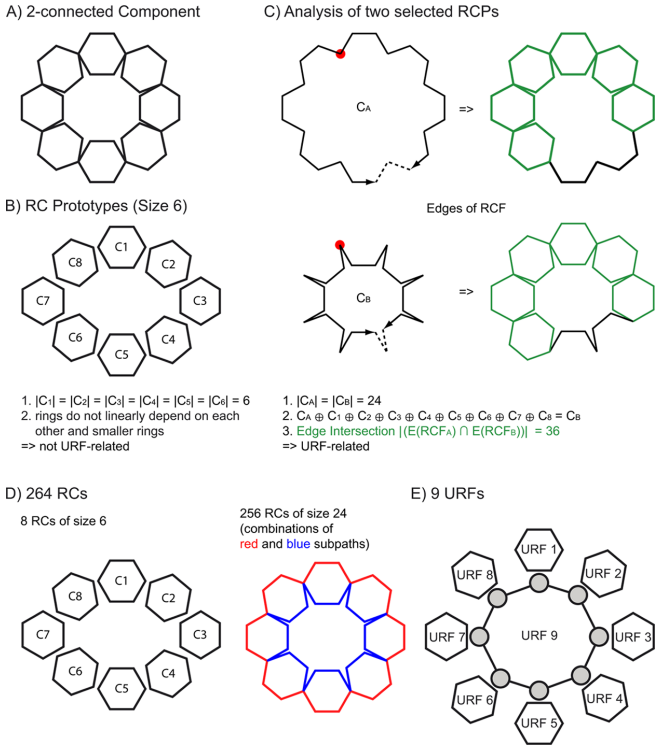
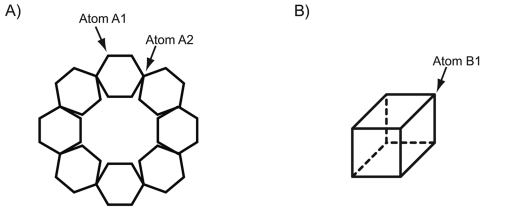


Figure 4. Ring system (A) consisting of 8 RCPs of size 6 (B) and 4 RCPs of size 24 of which 2 are illustrated (C). While the large RCPs have the same size and are linearly dependent according to condition 3 of definition 1, they do not share any edge. Their RCFs, however, share 36 edges. Note that this demonstrates, that two URF-related rings are not necessarily URF-pair-related. (D) The molecular graph contains 8 RCs of size 6 and 256 RCs of size 24. The set of all 264 RCs can be represented by 9 URFs. Eight URFs each contain a single 6-ring. One URF represents a macrocycle including the small URFs. This URF is illustrated as a quotient graph of the partition of smaller URFs. Note that the number of RCs increases exponentially with the number of para-bridged 6-rings, while the number of URFs increases linearly and stays intuitive.

the ring topology of a molecular graph. Macrocycles with para-substituated rings are a well-known problem (see Figures 3 and 4). The molecular structure shown in Figure 4 contains 264 RCs and 256 different possible SSSR cycle bases. The 256 large RCs belong to the same URF, resulting in 9 URFs. Thereby, URFs model the intuitive description of the molecule as a macrocycle containing eight smaller rings.

A frequently found specification in chemical patterns is the number of rings an atom is involved in. In the pattern language SMARTS, this is modeled with the R-feature. As discussed in the introduction, the R-feature is based on an SSSR which causes problems due to nonuniqueness. So far, no alternative approach resulting in a unique and polynomial number of ring representatives was available. Describing atoms by the number of URFs they are involved in represents an easy to implement solution to this problem.

Figure 5A shows the number of rings that contain the atoms A1 and A2. Using SSSRs, the result depends on the selected cycle base. In contrast, the number of RCs is large and chemically nonintuitive. Similar problems occur for symmetric cyclic structures like cubane (see Figure 5B). The calculation of URFs results in a consistent and chemically meaningful value for each atom. Furthermore, if an application requires the construction of an MCB, this can be easily achieved by selecting a



Ring Concept	Molecule A	Atom A1	Atom A2
URFs	9	2	3
SSSR	9	1 or 2	3
RCs	264	129	258

Ring Concept	Molecule B	Atom B1
URFs	6	3
SSSR	5	2 or 3
RCs	6	3

Figure 5. Two complex ring systems with their number of SSSR-rings, relevant cycles, and unique ring families. Additionally, ring memberships for the marked atoms are listed.

single arbitrary RCP of each URF followed by a Gaussian elimination of the resulting set of rings. Since the number of URFs is greater than or equal to the number of cycles of an MCB and smaller than or equal to the number of RCPs, the URF-number can be estimated by the following equation:

$$(E - V + 1) \leq \text{URF} - \text{number} \leq (2E^2 + EV) \qquad (6)$$

dx.doi.org/10.1021/ci200629w | J. Chem. Inf. Model. 2012, 52, 2013−2021

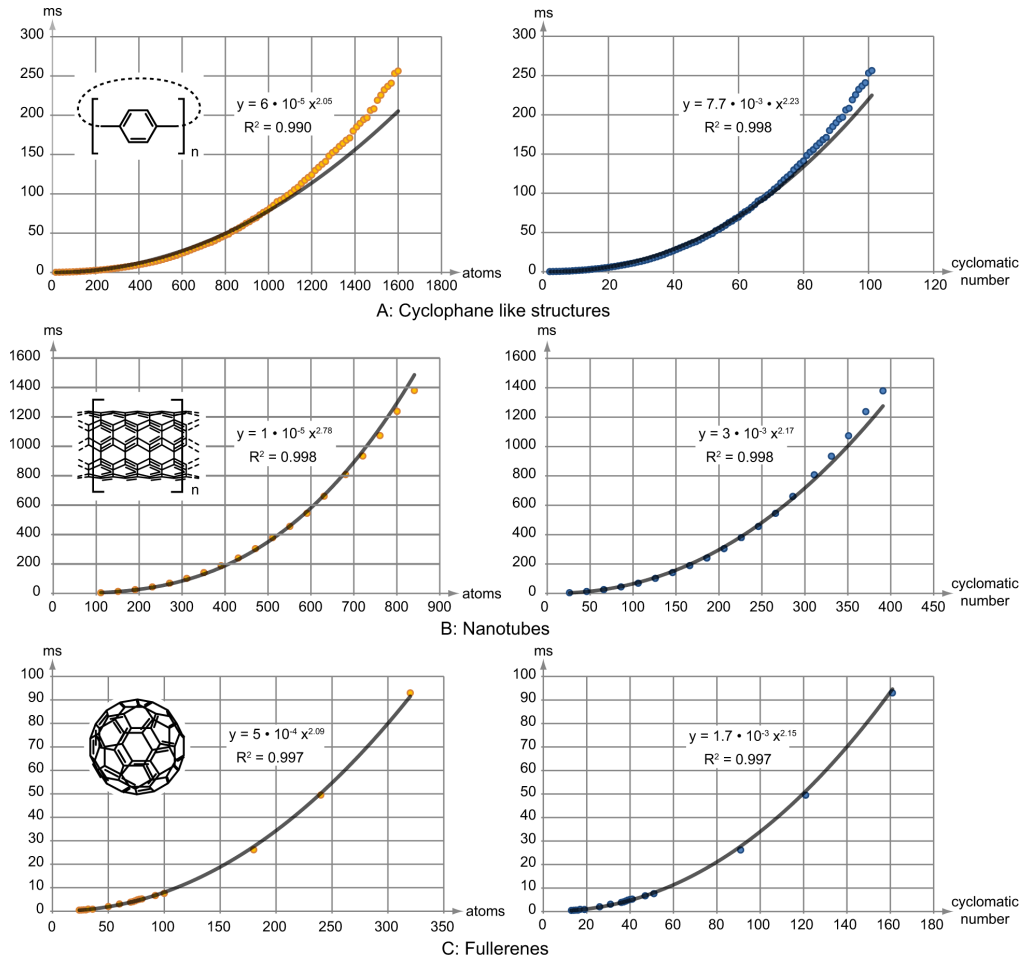


Figure 6. Required runtimes for the calculation of URFs depending on the number of atoms (left) and the cyclomatic number (right) for cyclophane-like structures (A), nanotubes (B), and fullerenes (C).

■ COMPUTING TIME BENCHMARKS

Ring perception is an important step in almost all chemoinformatics tasks. Applications which process large data sets thus require a fast method to identify the rings of molecular graphs. To check the large-scale applicability of the described method to calculate URFs, we measured the runtimes for the perception of URFs for a number of test sets. Time measurements were performed in a single thread on a PC with an Intel Core2 Quad Q9550 CPU (2.83 GHz) and 4 GB of main memory. For each molecule of the data set, the runtime for 100 iterations of ring perception was measured and on the basis of this measurement, the average runtime for a single ring perception was calculated. For file-IO we used the NAOMI framework.[19] Measured runtimes shown in Figure 7 do not include file I/O and molecular preprocessing. The data structures of the NAOMI framework are not specifically optimized for the detection of URFs but focus on the correct chemical modeling of small molecules. The listed runtimes thus provide an estimate of URF detection in the context of a common cheminformatic application.

To investigate the maximum runtime for the perception of URFs, we generated a number of molecules containing highly complex ring systems. First, we generated cyclophane-like structures that contain a large macrocycle with n para-bridged 6-rings. The generated molecules have a cyclomatic number Z_n of $Z_n = n + 1$, contain $n^2 + n$ RCs and $n + 1$ URFs. The runtime for the calculation of the URFs of these molecules is shown in Figure 6A. The required runtime for molecules containing $|V|$ atoms and a cyclomatic number of Z increases approximately with $|V|^2$ and Z^2.

As a second type of molecules that contain complex rings, single walled nanotubes were generated using ConTub.[20] While the parameters i and k were set to 5 nm, the length of the nanotube was increased in steps of 5 nm starting with a length of 10 nm up to a maximum of 100 nm. Both V and Z increase linearly with the length of the nanotube. As shown in Figure 6B,

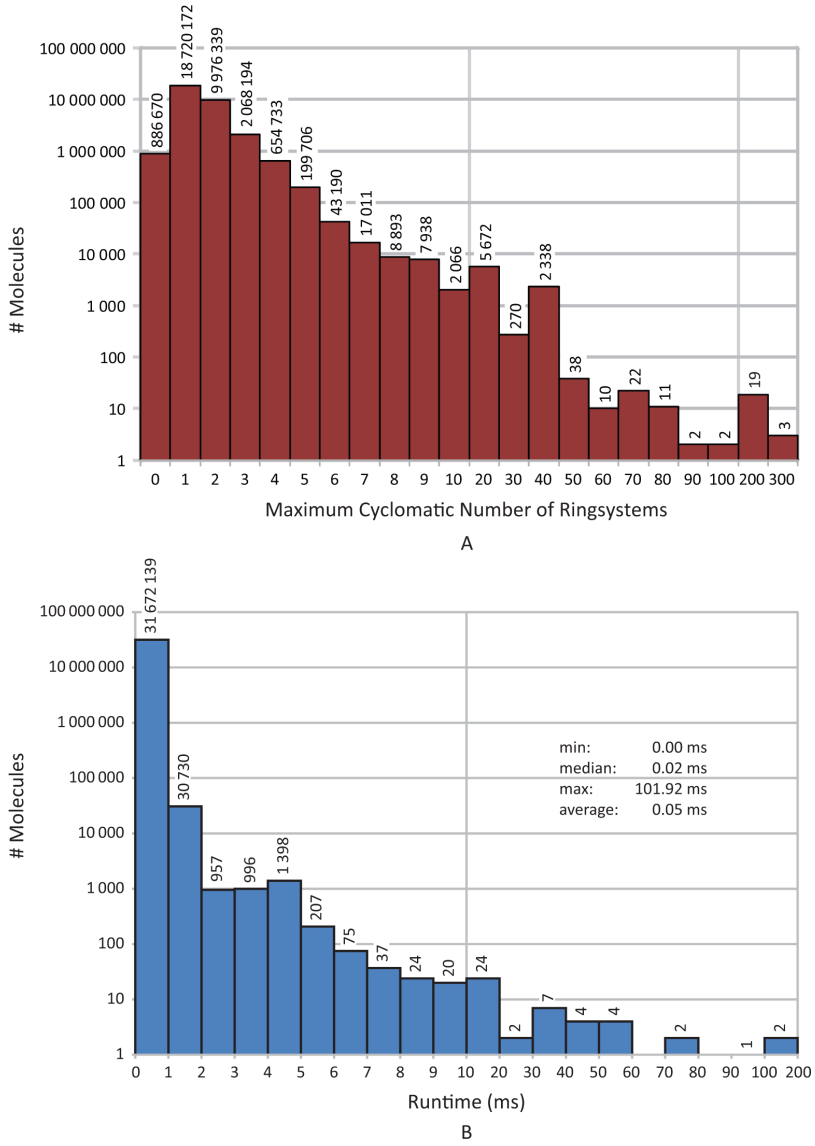


Figure 7. (A) Maximum cyclomatic number of the ringsystems of the molecules of the Pubchem-2D data set. (B) Benchmarks for URF perception for those molecules of the PubChem-2D data set having a cyclomatic number of at least one.

the runtime for the calculation of URFs increases slower than V^3 or Z^3.

As a third set of complex molecules, a number of fullerenes ranging from C24 to C320 were generated. Coordinates of these molecules were taken from a Fortran program specialized in the generation of fullerenes.[21] The runtime requirement again increased approximately with V^2 as well as with Z^2 (see Figure 6C).

Finally, to investigate the runtime which is required to perceive rings of commonly used molecules, we perceived URFs for the PubChem Compound 2D data set.[3] The data set was downloaded on March 27th, 2011 from ftp://ftp.ncbi.nlm.nih.gov/pubchem/Compound/CURRENT-Full/ and contains 32 593 299 molecular structures. These include a number of molecules of high complexity not present in the respective 3D data set. Figure 7A illustrates the complexity of the data set by showing the maximum cyclomatic number for the ringsystems of each molecule.

Shown runtimes represent the required runtime for 100 iterations of ring perception. Nevertheless, these runtimes are close to zero for most common molecules. The median for the percerption of URFs for a molecule of the Pubchem Data set is

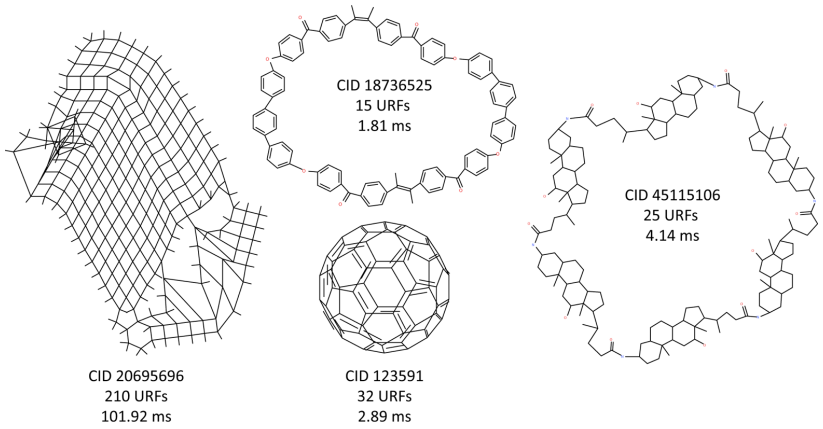


Figure 8. Runtimes and compound IDs for a selection of molecules of the Pubchem-2D data set.

0.02 ms, the average runtime is 0.05 ms, and the maximum runtime is 102 ms. This demonstrates that URFs can be calculated on the fly even for interactive applications and large databases. Only 34 490 molecules (0.11% of the database) show runtimes of more than 1 ms for the calculation of URFs. A list of those 100 molecules which require the highest runtimes for the calculation of URFs is added to this paper as Supporting Information. Some representative examples are shown in Figure 8.

A common molecular file format conversion, tested with Open Babel for the ZINC-everything data set, requires approximately 2 ms.[19] Due to the low runtime for calculating URFs of about 0.02 ms for commonly used molecules, the perception of URFs is suitable for high throughput chemoinformatics applications. Even for an artificially complex cylophane-like structure containing $100 + 2^{100}$ RCs, the URFs can be calculated in less than 2 s.

CONCLUSION

We have introduced the concept of unique ring families (URFs). In contrast to common ring perception approaches, URFs are polynomial in number, unique, and provide a complete description of the rings of a molecular graph. Furthermore, we have described an efficient method to calculate URFs in polynomial time. We demonstrated its applicability on large scale by showing computing time benchmarks for the Pubchem 2D data set. For these reasons, URFs represent a valuable alternative to common ring perception concepts and are worthwhile to be considered as a standard description for ring topologies in molecular graphs.

ASSOCIATED CONTENT

Supporting Information

100 molecular structures of the PubChem Database which require the highest runtimes for the perception of URFs. This material is available free of charge via the Internet at http://pubs.acs.org/.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

Present Address

‡Evotec AG, Essener Bogen 7, 22419 Hamburg. Phone: 0049 40 56081 230. Email: Adrian.Kolodzik@evotec.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to thank Christian Ehrlich and J. Robert Fischer for helpful comments and proofreading of the manuscript. Furthermore, the authors thank J. Robert Fischer and Tobias Lippert for their work on the NAOMI framework, which was used for reading the molecules of the Pubchem data set.

REFERENCES

(1) Vismara, P. Union of all the minimum cycle bases of a graph. *Electron. J. Comb.* **1997**, *4*, 1−15.
(2) Plotkin, M. Mathematical Basis of Ring-Finding Algorithms in CIDS. *J. Chem. Doc.* **1971**, *11*, 60−63.
(3) Wang, Y.; Xiao, J.; Suzek, T.; Zhang, J.; Wang, J.; Bryant, S. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, 623−33.
(4) Zamora, A. An Algorithm for Finding the Smallest Set of Smallest Rings. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 40−43.
(5) Berger, F.; Flamm, C.; Gleiss, P.; Leydold, J.; Stadler, P. Counterexamples in Chemical Ring Perception. *J. Chem. Inf. Model.* **2004**, *44*, 323−331.
(6) Hanser, T.; Jauffret, P.; Kaufmann, G. A New Algorithm for Exhaustive Ring Perception in a Molecular Graph. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1146−1152.
(7) Balducci, R.; Pearlman, R. S. Efficient exact solution of the ring perception problem. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 822−831.
(8) Carta, G.; Onnis, V.; Knox, A.; Fayne, D.; Lloyd, D. Permuting input for more effective sampling of 3D conformer space. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 179−190.
(9) *Daylight Theory Manual 4.9*. http://www.daylight.com/dayhtml/doc/theory/index.pdf (accessed June 9th, 2012).
(10) *Daylight Depictmatch*. http://www.daylight.com/daycgi_tutorials/depictmatch.cgi (accessed June 9th, 2012).
(11) Petra M. Gleiss, J. L.; Stadler, P. F. Interchangeability of Relevant Cycles in Graphs. *Electron. J. Comb.* **2000**, 1−16.
(12) Fujita, S. A new algorithm for selection of synthetically important rings. The essential set of essential rings for organic structures. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 78−82.

(13) Corey, E.; Perersson, G. Algorithm for machine perception of synthetically significant rings in complex cyclic organic structures. *J. Am. Chem. Soc.* **1972**, *94*, 460−465.
(14) Wipke, W.; Dyott, T. Use of Ring Assemblies in Ring Perception Algorithm. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 140−147.
(15) Downs, G.; Gillet, V.; Holliday, J.; Lynch, M. Theoretical aspects of ring perception and development of the extended set of smallest rings concept. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 187−206.
(16) Nickelsen, H. Ringbegriffe in der Chemie-Dokumentation. *Nachr. Dok.* **1971**, *3*, 121−123.
(17) Dury, L.; Latour, T.; Leherte, L.; Barberis, F.; Vercauteren, D. A new graph descriptor for molecules containing cycles. Application as screening criterion for searchingmolecular structures within large databases of organic compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1437−1445.
(18) Tarjan, R.; Vishkin, U. An Efficient Parallel Biconnectivity Algorithm. *SIAM J. Comput.* **1985**, *14*, 862−874.
(19) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI - On the almost trivial task of reading molecules from different file formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199−3207.
(20) Melchor, S.; Martin-Martinez, F. J.; Dobado, J. A. CoNTub v2.0 - Algorithms for Constructing C3-Symmetric Models of Three-Nanotube Junctions. *J. Chem. Inf. Model.* **2011**, *51*, 1492−1505.
(21) Schwerdtfeger, P. *Topological Analysis of Fullerenes - A Fortran Program*. http://ctcp.massey.ac.nz/index.php?group=page=fullerenes menu=fulleren (accessed March 11th, 2012).

A.4 Urbaczek, S.; Kolodzik, A.; Groth, I.; Heuser, S.; Rarey, M. Reading PDB: Perception of Molecules from 3D Atomic Coordinates. *J. Chem. Inf. Model.* 2013, *53*, 76–87

ACS direct link:
http://pubs.acs.org/articlesonrequest/AOR-RKZNmYgpW6rVhdXcQ8Gr

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Reading PDB: Perception of Molecules from 3D Atomic Coordinates

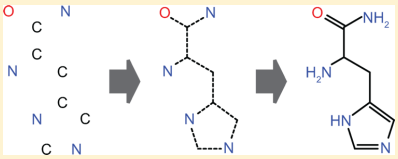
Sascha Urbaczek,[†] Adrian Kolodzik,[†,∥] Inken Groth,[‡] Stefan Heuser,[‡,§] and Matthias Rarey*[,†]

[†]Center for Bioinformatics (ZBH), University of Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany

[‡]Research Active Ingredients, Beiersdorf AG, Troplowitzstrasse 15, 22529 Hamburg, Germany

Supporting Information

ABSTRACT: The analysis of small molecule crystal structures is a common way to gather valuable information for drug development. The necessary structural data is usually provided in specific file formats containing only element identities and three-dimensional atomic coordinates as reliable chemical information. Consequently, the automated perception of molecular structures from atomic coordinates has become a standard task in cheminformatics. The molecules generated by such methods must be both chemically valid and reasonable to provide a reliable basis for subsequent calculations. This can be a difficult task since the provided coordinates may deviate from ideal molecular geometries due to experimental uncertainties or low resolution. Additionally, the quality of the input data often differs significantly thus making it difficult to distinguish between actual structural features and mere geometric distortions. We present a method for the generation of molecular structures from atomic coordinates based on the recently published NAOMI model. By making use of this consistent chemical description, our method is able to generate reliable results even with input data of low quality. Molecules from 363 Protein Data Bank (PDB) entries could be perceived with a success rate of 98%, a result which could not be achieved with previously described methods. The robustness of our approach has been assessed by processing all small molecules from the PDB and comparing them to reference structures. The complete data set can be processed in less than 3 min, thus showing that our approach is suitable for large scale applications.



INTRODUCTION

Crystal structures of protein−ligand complexes provide valuable insights into the interactions between proteins and small molecules. The statistical analysis of these structures has become an important tool in many different areas of research in the life sciences. Because of the large number of entries, the Protein Data Bank (PDB)[1] is the most important resource for experimentally determined structures of protein−ligand complexes. The structural data in the PDB is made available via different chemical file formats (PDB, mmCIF, PDBML/XML),[2] of which the PDB format[3] is the most common. PDB files contain element identities, three-dimensional coordinates, and connectivities for all atoms. However, unlike many other chemical file formats, this format does neither provide information about bond orders, formal charges, and aromaticity nor any kind of atom typing. Many cheminformatics methods and tools, however, depend on those and similar properties. Hence, when PDB files are supported as input, those properties have to be derived from the information provided by the file format. Although many current software packages include functionality to perceive molecular structures from three-dimensional coordinates, only a small number of these approaches has been published.[4−10]

The initial steps of all methods are similar to a certain extent. First, covalent bonds between atoms are identified by either using distance criteria or by simply relying on the connectivity data (CONECT entries) provided by the PDB format. In some approaches, this step is followed by a valence check during which spurious bonds arising from distorted geometries are removed. Subsequently, possible hybridizations for atoms are determined by analyzing bond lengths and bond angles. In the next step bond orders and atom types are assigned. Depending on the way these assignments are handled, the methods can be divided into two classes. Approaches from the first class determine bond orders independently of hybridization states, either by using the bond lengths directly or by matching of functional group patterns. This is often followed by an additional step during which inconsistencies in the assignments are handled. In methods from the second class, bond orders are derived directly from previously determined hybridization states using different bond localization routines.

We present a new method for the perception of molecular structures from three-dimensional atomic coordinates, which is based on the recently published NAOMI model.[11] Using its robust chemical description, the molecules are constructed in a hierarchical scoring approach. The first steps are based on the local geometry of each individual atom, whereas later steps include larger parts of the atom's environment to generate a correct chemical representation. This bottom-up approach has the advantage that it does not rely on definite assignments at early stages, for example, by assigning bond orders by torsion angles, or by matching of functional group patterns. In contrast to previously published methods, the final solution is selected from a list of potential candidate structures which are ranked using both confidence values for the atoms' geometry and

Received: July 30, 2012
Published: November 25, 2012

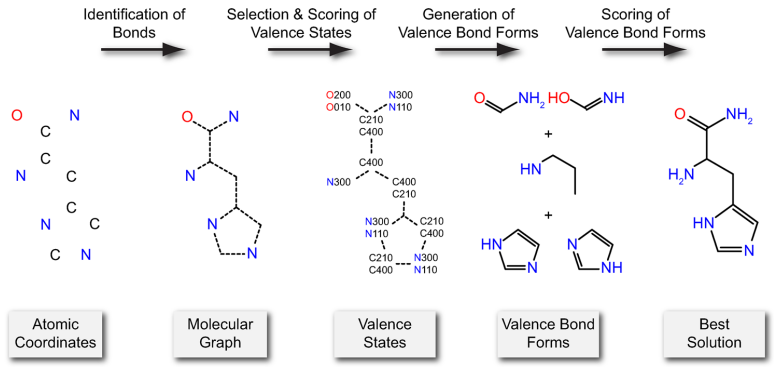


Figure 1. Schematic view of the workflow for the generation of molecules from three-dimensional coordinates.

chemical knowledge. This combination is the key to circumvent the shortcomings of other approaches, which either put too much focus on the provided coordinates or simply ignore them by using pattern-matching. The method's robustness and reliability are validated in different procedures by comparing reference molecules to the generated molecular structures. Furthermore, benchmark studies show its suitability for large scale applications.

METHODOLOGY

Overview. The aim of the presented method is the generation of both chemically valid and reasonable molecular structures from element identities and three-dimensional atomic coordinates. A molecule is considered chemically valid if a valence bond structure (Lewis structure) can be found, in which the valences of the atoms' elements are not violated. Not every possible valid valence bond form, however, provides a reasonable description of the molecule. On the one hand, geometric features, for example, interatomic distances and planar groups, must be reflected in the assigned bond orders. On the other hand, common standards for the representation of particular functional groups and resonance forms should be met. The last point is especially important since resonance forms and, depending on the quality of the provided coordinates, even tautomeric forms can not be deduced from geometry alone. For this purpose, we make use of the NAOMI model,[11] which has been successfully applied to the consistent conversion of chemical file formats. In this model, atoms are represented by three chemical descriptors, namely element, valence state, and atom type, which are assigned in three consecutive steps. Valence states represent valid bond order distributions for atoms in valence bond structures. They are defined by an element identity, the number of associated single, double, and triple bonds and a formal charge (e.g., N400+ for quaternary nitrogen atoms). As will be explained below, valence states can be used to generate valence bond forms if the atoms' connectivities are known. Atom types are derived from valence states and are thus independent of the input file format.

The perception of molecular structures from atomic coordinates is performed in four steps (see Figure 1). At first, covalent bonds are identified on the basis of interatomic distances. The second step comprises identification of possible valence states for each atom and scoring according to the atom's local environment. In the third step valence bond forms of the molecule are generated by enumerating valid combinations of valence states and their associated bond orders. These combinations are scored in the final step to determine the most appropriate valence bond representation of the molecule. The strategy adopted in our method is based on the opinion that the best possible compatibility between the perceived molecules and the provided coordinates should be sought. We believe, that the best way to do so is to build the molecular structure based on the atom's local geometries and use chemical knowledge only when either inconsistencies are encountered or ambiguities need to be resolved.

Identification of Bonds. To determine if a covalent bond exists between two atoms, the distance criterion originally proposed by Meng[4] is applied. A bond is created if

$$\delta_{\text{bond}} = r_{ij} - (R_i + R_j) < 0.4\ \text{Å} \quad (1)$$

where r_{ij} is the distance between the atoms i and j and R_i and R_j denote the covalent radii[12] of the atoms' corresponding elements. The high tolerance value of 0.4 Å in eq 1 ensures that no potential covalent bond is missed during the identification process. The softness of the criterion can, however, lead to an erroneous bond perception in case of distorted geometries. The resulting superfluous bonds give rise to two different types of chemical errors, which can readily occur at the same time. On the one hand, the atom's number of bonds may exceed the maximum valence of its associated element. On the other hand, distorted geometries may lead to the formation of incorrect cyclic structures (usually rings of size three or four). To deal with these errors, the bond perception is performed in several consecutive steps: (1) identification of bonds between non-hydrogen atoms, (2) valence check for all atoms and removal of superfluous bonds, (3) perception of the molecule's rings, (4) length check for all ring bonds and removal of superfluous bonds, and (5) identification of hydrogen bonds.

After the perception of all non-hydrogen bonds, each atom is checked for violations of its valence. This is done by comparing the number of identified bonds to the number of allowed bonds for its element. If a violation is encountered, long bonds ($\delta_{\text{bond}} > 0.1$ Å) are removed in order of their lengths until either the valence is restored or all long bonds are eliminated. In case of short non-hydrogen bonds ($r_{ij} < 0.5 \cdot (R_i + R_j)$), the coordinates are considered incorrect and the molecule cannot be constructed. After the ring perception each ring is checked for long bonds ($\delta_{\text{bond}} > 0.1$ Å). If such a ring is encountered, its longest bond is removed and the molecule's rings are

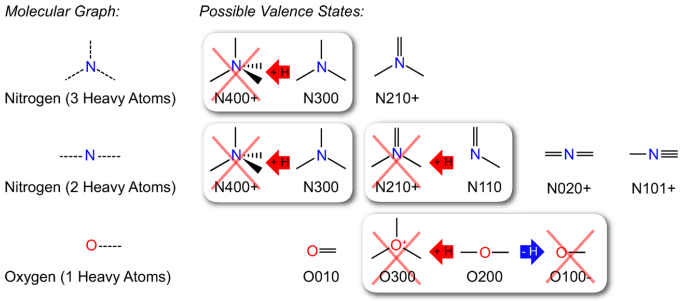


Figure 2. Examples for the selection of valence states. The crossed-out states are not selected since they can be deduced from the corresponding neutral states shown in the same box.

Table 1. Most Common Candidate Valence States for Typical Elements in Organic Molecules[a]

element	valence states						
hydrogen	H100						
carbon	C400	C210	C101	C020			
oxygen	O200	O010	O110+	O300+	O001+		
nitrogen	N300	N110	N210+	N400+	N020+	N101+	N001
phospohrous	P310	P300	P400+				
sulfur	S220	S210	S300+	S200	S110+	S010	S001+

[a]Valence states are represented as element symbol followed by the number of single, double, and triple bonds and the formal charge.

recalculated. This process is repeated until all long bonds in rings are eliminated. In contrast to non-hydrogen atoms, hydrogens are only allowed to have one bond and only the closest heavy atom needs to be identified. The hydrogen bond is created if the resulting bond is not short ($\Rightarrow r_{ij} \geq 0.5 \cdot (R_i + R_j)$) and the heavy atom's valence is not violated. Otherwise the hydrogen atom is discarded.

Selection of Valence States. In the next step, suitable valence states are selected from a list of allowed states for the respective element for each atom. Since bond orders have not been assigned at this point and formal charges are usually not provided, the number of bonds from the previous step is the only criterion for this selection. Valence states are selected in two cases. First, if the valence state and the atom have an identical number of bonds. Second, if the atom's bond count is smaller, but the missing bonds can be saturated by hydrogens. Charged valence states are only considered if no corresponding neutral state exists or a formal charge has been specified for the atom. Examples for this identification procedure are shown in Figure 2.

In many cases, this results in an ambiguous assignment since multiple valence states may be compatible with a particular number of bonds. To deal with this ambiguity, all selected valence states are scored to determine the most appropriate choice as explained below. This score reflects the state's compatibility with the atom's local environment, which is characterized by the spatial distribution of the atom's neighbors and their respective element identities. The use of a predefined list of valid valence states is an important aspect of ensuring a molecule's chemical validity. Atoms with an invalid number of bonds can be easily identified by the fact that no candidate valence state has been found. This evidently applies to all cases, where the number of bonds exceeds the maximum allowed number for the respective element. In addition to that, it is also possible to identify atoms with unusual bond counts in case of higher row elements such as sulfur or phosphorus. A typical example would be a phosphate group that is missing two of its terminal oxygen atoms thus leaving the central phosphorus with only two covalent bonds. This constellation is rather unlikely in organic molecules and simply saturating the atom's valences by addition of hydrogens seems questionable in a chemical sense. If no candidate valence state for an atom can be found, the molecule is considered incorrect and cannot be constructed. The most common candidate valence states for typical elements in organic molecules are shown in Table 1.

Evaluation of Geometrical Parameters. The compatibility of valence states is mainly assessed on the basis of the atom's local geometry. For that purpose, several geometrical parameters g are evaluated and used to derive scores $G_p(g)$ for different chemical properties p, for example, bond orders. These scores are calculated according to the following scheme. For each property, a minimum and a maximum value are defined, which correspond to the scores of 0.0 and 1.0, respectively (see Table 2). Between the minimum and the maximum values a linear function is used.

The absolute value of the scalar **triple product** π of the normalized bond vectors connecting an atom and its neighbors

Table 2. Parameters for the Calculation of Scores $G_p(g)$ for Different Properties p[a]

property	parameter	minimum (0.0)	maximum (1.0)
$G_{planar}(\pi)$	π	≥0.6	≤0.15
$G_{linear}(\alpha)$	α[°]	≤150	≥170
$G_{sp^2}(\alpha)$	α[°]	≤114	≥118
$G_{single}(\delta)$	δ[Å]	≤−0.1	≥−0.04
$G_{double}(\delta)$	δ[Å]	≥−0.04	≤−0.1
$G_{triple}(\delta)$	δ[Å]	≥−0.15	≤−0.25
$G_{planar}(\tau)$	τ[°]	≥40	≤10

[a]Between the minimum and the maximum values a linear function is used.

dx.doi.org/10.1021/ci300358c | J. Chem. Inf. Model. 2013, 53, 76−87

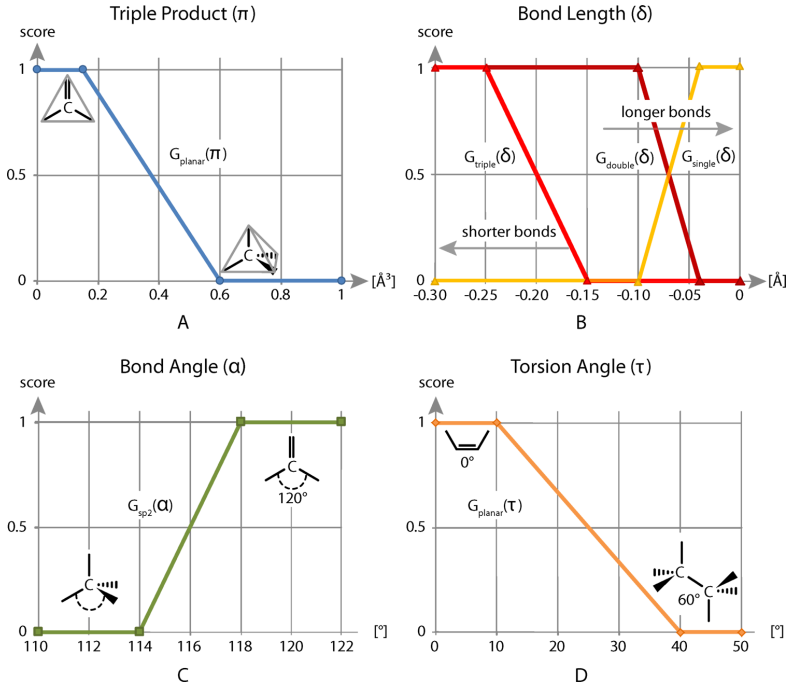


Figure 3. A: Score for the planarity of an atom using the triple product. B: Score for bond orders using the bond length. C: Score for an sp^2 hybridization on the basis of an atom's bond angle. D: Score for planarity using the largest torsion angle.

is a direct measure for its planarity ($G_{planar}(\pi)$) and can thus be used to distinguish sp^2 from sp^3 hybridizations. A triple product smaller than 0.15 indicates planarity, whereas a value larger than 0.6 (the triple product of an ideal tetrahedron is approximately 0.7) indicates the opposite. **Bond angles** α are used to determine the hybridization of an atom. They are especially important for the identification of linear geometries ($G_{linear}(\alpha)$), for example, in the presence of triple bonds. Because of the large difference to the bond angles of other hybridizations, sp hybridization can be easily distinguished. The smaller difference between the angles associated with sp^2 and sp^3 hybridizations makes the distinction between these cases rather difficult ($G_{sp^2}(\alpha)$). Scores for particular bond orders ($G_{single}(\delta)$, $G_{double}(\delta)$, $G_{triple}(\delta)$) are determined using the **bond length** δ which is calculated as described in eq 1. In the case of double bonds, the largest **torsion angle** τ at the respective bond is taken into consideration ($G_{planar}(\tau)$). Torsion angles can be used to check if the atoms surrounding the bond partners are coplanar, which is a precondition for double bonds. By taking torsion angles into account, invalid double bond assignments due to shortened interatomic distances can be avoided. Single bonds joining an aromatic ring with either an alkyl substituent or another aromatic ring are typical examples for this case. Although the bond length might be shortened, the torsion angle often clearly contradicts the double bond order. The torsion bond probability $G_{double}(\delta,\tau)$ is the product of $G_{double}(\delta)$ and $G_{planar}(\tau)$. For atoms in rings, **torsion angles** τ can be used to determine the planarity of the ring. In this case, only bonds in the same ring are included during the calculation of the largest torsion angle.

Probabilities of Hybridization States. The scores $G_p(g)$ are the basis for the calculation of probabilities for different hybridization states P_{hyb}. Since the number and kind of parameters used strongly depends on the atom's topology, each case is discussed separately.

For atoms with one bond, the bond length is the only available geometrical parameter.

$$P_{sp} = G_{triple}(\delta) \quad (2)$$

$$P_{sp^2} = G_{double}(\delta) - G_{triple}(\delta) \quad (3)$$

$$P_{sp^3} = G_{single}(\delta) \quad (4)$$

In this case the probabilities for the hybridization states correspond to the scores for the respective bond orders as described in eqs 2−4. Since sp hybridization is always associated with a linear geometry, the number of bonds at the atom's neighbor is checked. If the neighbor has more than two bonds this condition cannot be fulfilled and the value of P_{sp} is added to P_{sp^2} and then set to 0.0.

For atoms with two bonds, one bond angle and two bond lengths are available. The score for the presence of a double bond at the atom A_{double} is calculated as the sum of the torsion bond scores $G_{double}(\delta, \tau)$ of the atom's bonds, whereas its maximum value is limited to 1.0.

$$A_{double} = \min(1.0, \sum G_{double}(\delta, \tau)) \quad (5)$$

The sum in eq 5 is used to account for the limitations of valence bond structures. In delocalized systems, for example, aromatic rings, bonds can have lengths between the expected

dx.doi.org/10.1021/ci300358c | J. Chem. Inf. Model. 2013, 53, 76−87

values of single and double bonds. In this case, the score for the presence of a double bond might be underestimated if only the larger of both values is considered. Because of the geometric restraints in small rings, we then distinguish two cases. For atoms in an acyclic environment or in large rings (at least eight atoms), the following probability scheme is used:

$$P_{sp} = 2/3 \cdot (G_{linear}(\alpha) + 0.5 \cdot A_{double}) \quad (6)$$

$$P_{sp^2} = \begin{cases} 1/2 \cdot (1.0 - P_{sp}) & \text{if} \quad P_{sp} > 0.0 \\ 2/3 \cdot (A_{double} + 0.5 \cdot G_{sp^2}(\alpha)) & \text{else} \end{cases} \quad (7)$$

$$P_{sp^3} = 1.0 - (P_{sp^2} + P_{sp}) \quad (8)$$

Since only the sp hybridization is compatible with a linear geometry, the bond angle has a higher weighting factor in the calculation of the associated probability in eq 6. For the probability of an sp^2 hybridization in eq 7 it is considered less reliable due to the small difference to the ideal value of the sp^3 hybridization. If the atom is part of a small ring (less than eight atoms), ring torsion angles can be used as an additional parameter to assess the planarity of the respective ring. Furthermore, a linear arrangement is extremely unlikely in these cases, so that only sp^2 and sp^3 hybridizations need to be considered. The probabilities and scores are adapted in the following way:

$$P_{sp^2} = 2/5 \cdot (A_{double} + A_{planar} + 0.5 \cdot G_{sp^2}(\alpha)) \quad (9)$$

$$P_{sp^3} = 1.0 - P_{sp^2} \quad (10)$$

Since bond angles in rings with a size smaller than six are strongly influenced by the strain of the cyclic arrangement, they are not a reliable measure for the atom's hybridization. In this case, the score is automatically set to 0.5 to indicate that no decision can be made. The planarity score A_{planar} in eq 9 for an atom is the minimum of the $G_{planar}(\tau)$ (see Figure 3D) scores of each bond.

For atoms with three bonds, three bond angles, three bond lengths, and one triple product can be calculated. Since sp hybridization is not possible in this case, a decision between sp^2 and a sp^3 hybridization has to be made. For the calculation of the atom's angle score $A_{sp^2}(\alpha)$ the mean bond angle $\overline{\alpha}$ is used.

$$P_{sp^2} = 1/6 \cdot (3 \cdot G_{planar}(\pi) + 2 \cdot A_{double} + A_{sp^2}(\alpha)) \quad (11)$$

$$P_{sp^3} = 1.0 - P_{sp^2} \quad (12)$$

Again, the geometrical parameters are not considered equally reliable which is reflected in the different weighting factors in eq 11. The scoring of valence states for atoms with four or more bonds is solely based on scores for bond orders, and no probabilities for hybridizations need to be calculated for these cases.

Scoring of Valence States. The probabilities P_{hyb} from the previous step are used to calculate integer-based scores for all selected valence states of each atom. This score reflects the compatibility between the atom's local environment and the respective valence state and is used to identify the best suited state for an individual atom. Additionally, the absolute value of the score also provides a measure of confidence, which can be used to compare possible valence state assignments for different atoms. The scoring procedure makes use of the fact that valence states are not compatible with all hybridization states. In case of compatibility, the score S_{VS} is calculated using the probability P_{hyb} according to the following scheme:

$$S_{VS} = \begin{cases} 1 & \text{if} \quad P_{hyb} < 0.6 \\ \lfloor P_{hyb} \cdot c + 0.5 \rfloor & \text{else} \end{cases} \quad (13)$$

The confidence factor c in eq 13 determines the maximum value of the score and depends on the topology of the respective atom (see Table 3). The values are based on the number of geometrical parameters available for the calculation of the probabilities P_{hyb}. A single bond length, for example, is not well suited to reliably distinguish between hybridizations, since even small geometrical distortions may cause the bond order perception to fail. This lack of reliability is reflected in a small confidence factor of 2.0 for atoms with one bond. The integer-based scheme ensures that only those valence states which are clearly favored by the atom's local geometry receive scores larger than one. This prevents the elimination of valence states based on small geometrical differences.

Table 3. Confidence Values for Different Topologies

topology	confidence c
1 bond	2.0
2 bonds(acyclic)	3.0
2 bonds(cyclic)	4.0
≥3 bonds	5.0

If the compatibility between the selected valence states and their associated hybridization states is mutually exclusive, the scoring procedure is straightforward. Because of the limitation of valence bond structures, this is, however, not always the case (see Figure 4 for examples). On the one hand, there are atoms which are represented by the same valence state but have different hybridizations, such as nitrogens in amines and amides. These cases are handled by assigning the largest score obtained for all compatible hybridizations to the respective valence state. On the other hand, some atoms are not sufficiently represented by a single valence state such as oxygens in a carboxylate group. In this case both compatible valence states receive identical scores. Examples for the scoring procedure are shown in Figure 5.

The calculation of scores for atoms with four or more bonds can in most cases be avoided due to the fact that there is only one suitable valence state. If this is not the case, the multiple bond score A_{double} introduced in eq 5 is used in place of P_{hyb} to calculate the score for all selected valence states. This is always sufficient to distinguish between the alternatives.

In some cases, it is beneficial to remove valence states from the list of candidates if their associated hybridization is not compatible with the atom's local geometry ($P_{hyb} = 0.0$). These valence states will not be considered during the generation of valence state forms, which in turn reduces the complexity of the next steps. Since distorted geometries could easily lead to the premature exclusion of relevant valence states, this is only done in two rather unambiguous cases. First, if the corresponding valence state is only compatible with an sp hybridization and second if the atom has three bonds.

Distorted geometries can also result in incorrect scores which will eventually lead to undesired valence bond structures. This is especially true if atoms with only one bond are involved since the resulting assignment cannot be corrected by the valence states of the surrounding atoms. To avoid these errors, valence

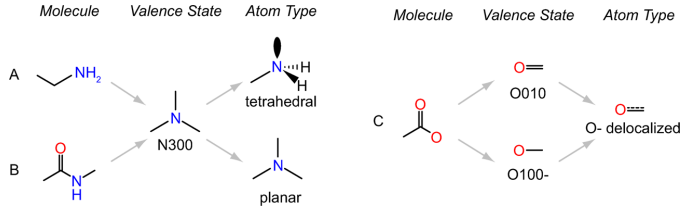


Figure 4. Limitations of valence bond structures: (a) Nitrogens in amides and amines have the same valence states but different geometries. (b) Oxygens in carboxylates have different valence states but have the same bond length.

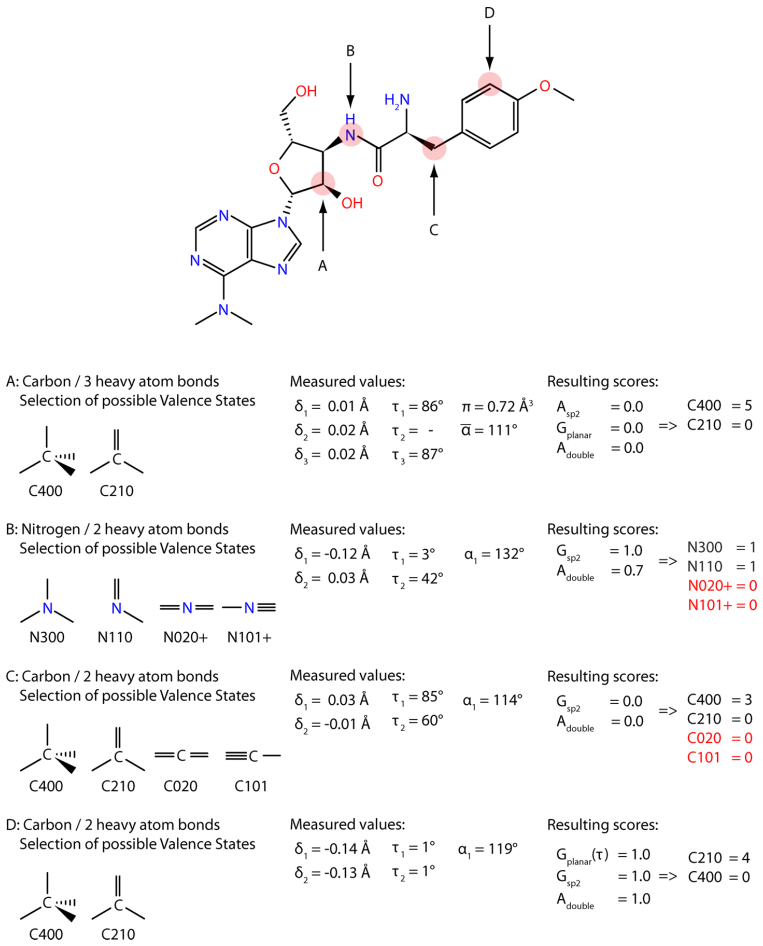


Figure 5. Examples for the valence state scoring procedure. Relevant geometrical parameters are triple products, bond lengths, bond angles and torsion angles. If bond angles and bond lengths do not indicate a linear geometry, valence states associated with a linear geometry are excluded (marked in red). For the atoms A, C, and D the geometrical parameters clearly support one of the valence states. In the case of atom B, there is no clear preference and two valence states are equally probable.

state scores of atoms that are part of specific substructures are increased by +2 (see Figure 6). These resulting scores are, however, not high enough to change the assignment in case of a perfect geometric compatibility.

The purpose of the described procedure is to provide reliable scores which can be used to identify the best valence state representation of the molecule. Due to the differing quality of available input data, this must also apply if the provided coordinates are of poor quality. As mentioned above, the scores are not only used to find the best choice for an individual atom but also to compare assignments between atoms. This means that valence states with higher scores have a stronger influence

dx.doi.org/10.1021/ci300358c | J. Chem. Inf. Model. 2013, 53, 76−87

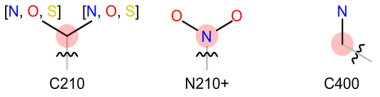


Figure 6. Additional scores of +2 are assigned to valence states for atoms (marked with red spheres) in specific substructures. The number of bonds at the atoms corresponds to the number of bonds identified during the bond perception.

on the resulting valence state form. The score does not only depend on the number of available parameters at the respective atom, but also on their consistency. This is assessed by the individual evaluation of the geometrical parameters during the calculation of the probabilities P_{hyb}. A value of 1.0 is only possible if all geometrical parameters are consistent, which in turn results in low scores for atoms with inconsistent local geometries.

Generation of Valence Bond Forms. In the next step, chemically valid valence bond representations of the molecule are generated by assigning valence states to all atoms and bond orders to all bonds. A combination of valence states is valid if a bond order distribution can be generated which is in accordance with the valence states of the atoms. The score of such a combination is calculated as the sum of the scores of the valence states from the previous step. The best combination can be identified by enumerating all valid combinations with a maximum score. For the enumeration a branch and bound algorithm with a depth-first search strategy is used. Prior to the enumeration, the list of valence states for each atom is checked for cases where only one valence state is remaining. This state is assigned directly and the orders of the adjacent bonds are adapted accordingly. Afterward, the molecule is partitioned into zones containing atoms connected by bonds with unassigned bond orders. The individual processing of each zone further decreases the number of possible combinations. If a single best scored combination for a zone exists, it is selected. Otherwise, combinations with equal scores are ranked using additional geometrical and chemical criteria as described below.

Scoring of Valence Bond Forms. Each combination of valence states generated in the last step is a valid valence bond form (in the sense that no valences are violated) and is also compatible with the local geometry of the atoms. This does, however, not necessarily imply that each form provides a reasonable description of the molecule. On the one hand, discrepancies between the assigned bond orders and the actual bond lengths might exist, which could not be resolved during the atom based valence state scoring procedure. On the other hand, the combination might contain unusual representations of functional groups or conjugated systems, which could not be excluded using geometrical parameters alone. Hence, an additional scoring scheme, which makes explicit use of the assigned bond orders, is applied to distinguish reasonable from undesired valence bond forms. In contrast to the previous steps, where geometrical parameters had a high priority, this step focuses mainly on chemical aspects.

Prior to the scoring procedure, valence states and bond orders are assigned if they are identical in all generated valence bond forms. Afterward, the molecule is again partitioned into zones containing atoms connected by bonds with unassigned bond orders. Then, substructures (see Figure 7) including at least one of the unassigned atoms are identified in each of the remaining valence bond forms. These substructures correspond to preferred representations of functional groups and for each

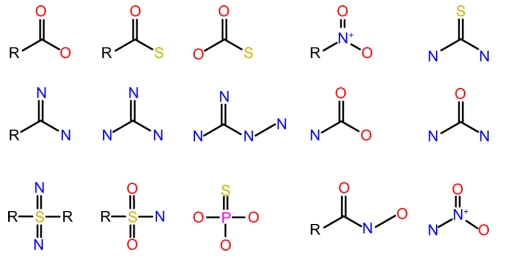


Figure 7. Substructures representing favored representations of particular functional groups in valence bond structures. The R represents both carbon and hydrogen.

match a score of +1 is assigned to the respective form. If an unassigned atom is part of a ring with a size smaller than eight, Hueckel's rule is applied to assess its aromaticity. Valence bond forms receive a score of +1 for each ring, where the rule is fulfilled. It must be stressed that our approach does not favor particular functional groups or aromatic rings in general but only if the geometrical parameters were not sufficient to resolve the structure.

If a bond with an unassigned bond order is part of a substructure or ring which has been scored in the previous step, no further scoring is performed. Otherwise, $G_{order}(\delta)$ from Table 2 is used to determine if the current bond order is compatible with the calculated bond length. In the case of double bonds, $G_{double}(\delta,\tau)$ is used. If the respective value exceeds a threshold of 0.7 a score of +1 is assigned to the valence bond form. Hence, solutions in which bond lengths do not correspond to the assigned bond orders receive lower scores. If the bond is also part of a ring with less than eight atoms, $G_{planar}(\tau)$ is used as an additional parameter to assess the bond's planarity. A score of +1 is assigned if either $G_{planar}(\tau)$ is smaller than 0.3 (planar geometry) for a double bond or $G_{planar}(\tau)$ is larger than 0.7 (ring is not planar) for a single bond.

Again, only the solutions with the largest scores are kept. If there are still multiple solutions left, they are considered equivalent and a canonization scheme is used to choose a unique form for each zone. Since a detailed explanation of the canonization algorithm extends the scope of this publication, only a brief description of the general idea will be given. The atoms of the respective zone are ordered in a procedure similar to the CANON algorithm[13] used for the generation of USMILES. The zone is then processed atom by atom according to this newly generated order. At each step the respective solutions are sorted by the valence states (using ids as sorting criterion) of the particular atom and all solutions with lower ranks are eliminated. This process is repeated until only one solution remains. Obviously, it is also possible to omit the canonization and use the solutions for each zone to enumerate all equivalent valence bond forms of the molecule.

RESULTS AND DISCUSSION

Validation with Curated Structures. In a first validation procedure we tested if our method was able to generate the expected valence bond structures for small molecules from different PDB entries. The success was verified by comparison of the resulting molecules to manually curated reference structures provided as USMILES.[13] Small molecules were extracted from PDB entries used in the studies of Hendlich[6] and Labute.[7] Because of its importance in the field of

cheminformatics, we also included the ligands from the PDB entries of the Astex Diverse Set.[14] The complete validation set consists of 563 molecules from 363 PDB entries. Both PDB entries and SMILES files for the respective compounds are provided as Supporting Information. Table 4 lists the PDB entries and the component names of the ligands for which our method failed to generate the expected structure. Five of these examples are shown together with the reference structures taken from the respective publications in Figure 8.

Table 4. PDB IDs and Component Names of All Molecules for Which Our Method Did Not Generate the Expected Structure

Labute[7]	Hendlich[6]	Astex[14]
2R04 (W71)	1MIO (CFM)	1G9V (HEM)
3FX2 (FMN)	1PMP (OLA)	1Q4G (HEM)
5TLN (BAN)	6RSA (UVC)	
8XIA (XLS)		

The dihydro-oxazol ring of ligand W71 from 2R04 (see Figure 8A) is perceived as oxazol. Because of a short bond of C4A to the nitrogen atom and the planarity of the five-membered ring, the valence state C210 (which is compatible with an sp^2 hybridization) receives a higher score. This eventually leads to a structure including an aromatic ring. One of the hydroxy groups of the flavin mononucleotide ligand FMN from 3FX2 (see Figure 8B) is interpreted as a carbonyl group. In this case the valence state C210 is favored due to the trigonal planar geometry of C2′. The same also applies to the α carbon CA2 in BAN from 5TLN (see Figure 8E). The carbonyl group of the molecule XLS from 8XIA (see Figure 8C) is interpreted as a hydroxy group because of the tetrahedral geometry at C2. The double bonds of the olefinic moiety of OLA (1PMP) (see Figure 8D) and of one of the vinylic groups in HEM (1G9V, 1Q4G) are perceived as single bonds due to the bond lengths and associated bond angles.

Our method was able to generate the correct structure in 98% of the cases. All observed differences were caused by strong deviations from the expected molecular geometries. The valence bond forms generated by our method are, however, equally reasonable in a chemical sense and also in agreement with the supplied atomic coordinates. Only in the case of BAN the generated structure does not correspond to the tautomeric form which would be expected for the isolated compound with respect to the hydroxamic acid group. The molecular geometry may, however, be influenced by the interactions with a metal atom in the protein−ligand complex. The PDB entry CFM contains an Fe−Mo−S cluster, for which our method does not produce a valence bond form but isolated atoms. Since valence bond forms are not well suited to describe metal clusters, we do not consider this a perception error, but think it should be mentioned at this point. The same is true for the vanadate in 6RSA in which no bonds between the oxygens and the vanadium atom are formed. The uridine molecule, however, is perceived correctly.

Comparison with Other Methods. To compare our results with those of other existing methods, we used the tools I-interpret,[15] fconv[16] and MOE[17] to generate molecules for the above-mentioned 363 PDB entries. This was done by first converting the entries from PDB to SDF (since fconv does not support sdf as output format, mol2 was chosen in this case) and then using the converted file as input for the comparison to the reference structures. The results are summarized in Table 5. Since our method will be part of the NAOMI converter, it is referred to as NAOMI in the table. The comparison to the reference structures was done using the NAOMI framework. Since all files (PDB input, SDF/MOL2 files from different tools, SMILES for comparison) are supplied as Supporting

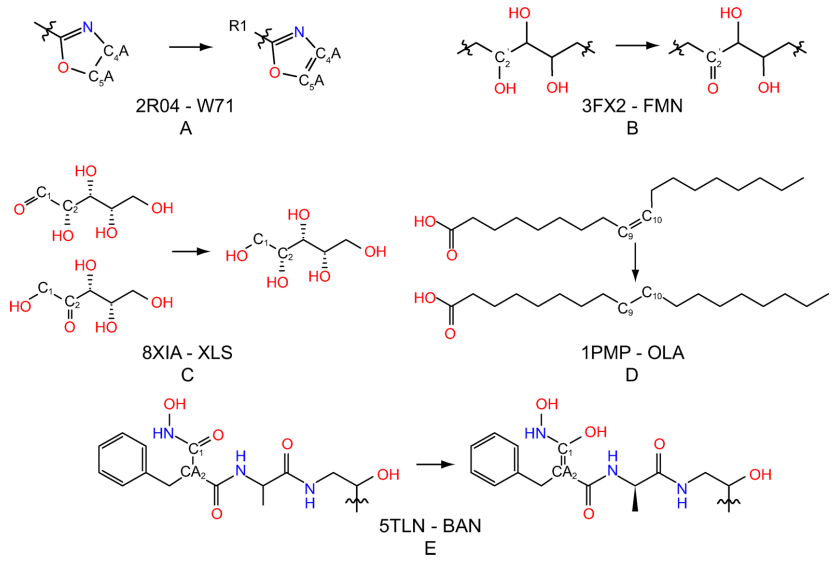
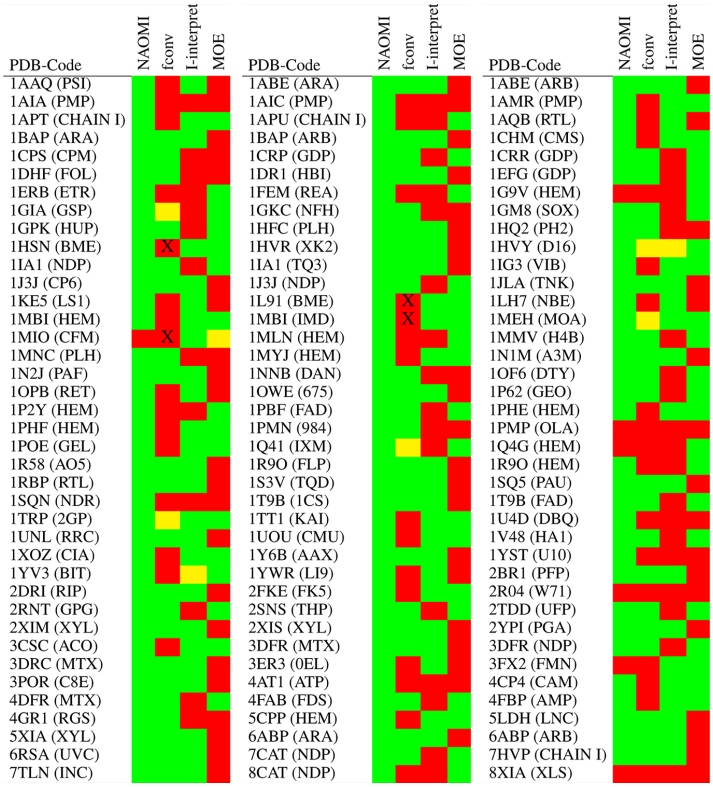


Figure 8. Five of the nine molecules for which our method did not generate the expected structure. The expected results are shown on the left side of the arrow, the results of our method on the right. The names from the PDB files are listed for all atoms for which incorrect valence states were identified.

dx.doi.org/10.1021/ci300358c | J. Chem. Inf. Model. 2013, 53, 76−87

Table 5. Results of the Generation of Molecules from the 363 PDB Entries Using Different Tools[a]



[a]The colors represent the quality of the resulting structures. Green cells: Correct structure. Yellow cells: Suboptimal structure. Red cells: Structure substantially differing from reference. X: No structure generated.

Information, the comparison can be carried out using other tools with the same functionality. The differences between the generated molecules and the references can be divided into two categories. First, there are molecules for which hybridization states or bond orders have been differently assigned. All of these differences are caused by deviations from the expected geometries and are thus directly linked to the quality of the respective coordinates. Second, there are molecules with unusual or even chemically unreasonable resonance or tautomeric forms. Although these differences are not wrong considering the molecule's geometry, they deviate from conventions concerning the representation of particular substructures. Depending on the gravity of these deviations, the solutions are either considered invalid or simply not optimal. Examples for both cases are shown in Figure 9.

Table 5 shows that many differences appearing with other tools are avoided by our method. Incorrect perceptions because of geometrical distortions are often prevented by considering all aspects of an atom's environment. The confidence values for valence states are derived from multiple geometrical parameters so that the assignment has a certain stability against small geometrical distortions. This is a considerable advantage over methods which rely on definite assignments based on particular geometrical parameters. By considering the confidence values of surrounding atoms during the generation of valence bond

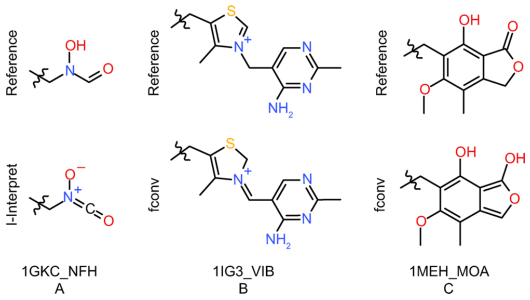


Figure 9. Comparison of reference structures and perceived structures generated by other tools. The structures A and B are classified as errors, whereas structure C is classified as not optimal.

structures even strong distortions can be compensated in some cases. The explicit inclusion of chemical knowledge in the last step of the workflow helps to reliably resolve the remaining ambiguities. Errors concerning the representation of molecules typically occur with methods that put too much emphasis on the evaluation of the geometrical parameters during the generation of valence bond forms. One has to keep in mind that localized bond orders are only an approximation and do

dx.doi.org/10.1021/ci300358c | J. Chem. Inf. Model. 2013, 53, 76−87

not have to strictly adhere to molecular geometries. By scoring multiple alternative structures using a combination of chemical and geometrical criteria, our method is able to generate molecules that are both in agreement with the atomic coordinates and chemically reasonable.

Validation with Complete Ligand Expo Data Set. The main purpose of our method is the automatic generation of reasonable molecular representations for large data sets. To show that our method is both, efficient and robust, we applied it to all entries of the Ligand Expo data set[18] in PDB format and analyzed the results in terms of runtime and quality. The generated structures were compared to the respective molecules in the SDF format, which are also provided on the Ligand Expo Web site.[19] Again, USMILES served as a basis for the comparison. Since the NAOMI model does not support covalently bound metal atoms, all metal bonds were ignored and only the largest resulting component was used. Additionally, monatomic entries were skipped, since the ionization state of single atoms can not be deduced without knowledge of the environment. Empty entries and entries with multiple disconnected components were also ignored, since this usually indicates missing atoms. Some entries were rejected due to unusually small distances between atoms (coordinate errors). The results of this procedure are summarized in Table 6.

Table 6. Results of the Analysis of the 602704 Entries in the Ligand Expo Data Set for Both SDF and PDB

	SDF	PDB
no. total	602704	602704
mo. format errors	0	3015
no. empty entries	7688	7678
no. monatomic entries	241002	239452
no. disconnected entries	10254	10193
no. coordinate errors	499	939
no. converted entries	343261	341427
no. compared entries	334121	

Both data sets initially contained 602704 entries, of which 334121 (55.4%) were eventually used for comparison. To avoid inconsistencies concerning ionization states, all molecules were neutralized in advance (see Figure 10). In 91.7% (306341) of the cases identical valence bond structures were found. The reasons for the observed 27780 differences are quite diverse, as shown shown in Table 7.

In 10012 (36.0%) of the cases, a different tautomeric form of the molecule was generated. Tautomeric forms can often not be distinguished on the basis of the provided coordinates and multiple solutions are equally acceptable. As described above these cases are handled by a canonization procedure, so that different tautomeric forms do not indicate perception errors but rather different default representations. Typical examples for substructures with equivalent tautomeric states are substituted imidazoles, pyrimidones, and guanidinium groups. 810 (2.9%)

Table 7. Analysis of the Reasons for Different Valence Bond Structures for the 334121 Compared Entries of the Ligand Expo Data Set[a]

	entries	% of data set	% of differences
no. different valence bond form	27780	8.3	100
no. different tautomeric form	10012	3.0	36.0
no. different oxidation state	810	0.2	2.9
no. different bond order	10349	3.1	37.3
no. different terminal bond order	6063	1.8	21.8
no. small molecule	3523	1.1	12.7

[a]Molecules are considered small if they have less than 8 heavy atoms.

of the differences were due to different oxidation states of particular heterocyclic compounds such as NAD/NADP. As with tautomers, these states can not be reliably distinguished on the basis of atomic coordinates, especially in entries with low resolution. Therefore, these cases are also not considered perception errors meaning that 94.9% of the results are essentially identical.

The remaining 16958 entries were further investigated in order to determine the reason for the incorrect perception. These entries correspond to 2341 different components, of which the 20 with the highest counts are shown in Table 8. Evidently, 22.8% (3864) of the differences are caused by only 1% of the components. These entries will be used for the discussion of specific problems encountered with the LigandExpo data set.

The errors associated with HEM are almost exclusively caused by the vinylic double bonds. As discussed above, the number of available geometrical parameters for the determination of bond orders for terminal bonds is small and makes the perception less stable with respect to deviations from ideal geometries. PGV, BCR, PEK, PEV, and OLC are molecules with long aliphatic chains and a specific number of double bonds. In many entries there is a considerable disagreement between our method and the LigandExpo references concerning both the presence and position of these double bonds. We have encountered numerous examples where we did not even find a single shortened bond length in the molecule although a double bond was present in the LigandExpo structure. Many of the incorrect perceptions concerning FAD, NAD, and UMP are caused by strong geometrical distortions of the respective aromatic rings. In some cases torsion angles that reach up to 40° are encountered in these usually completely planar structures. In case of CYC, BLA, and MDO exocyclic carbon−carbon double bonds at five-membered aromatic heterocyclic are interpreted as single bonds. These assignments were in all cases a result of an unambiguous single bond length at the respective bond. The difference from the entries ACB, MLE, and MYR are caused by the specific way covalently bound compounds are handled in the PDB format. If a molecule is bound to a residue of a protein or nucleic acid, the atom involved in this bond is usually assigned to the residue.

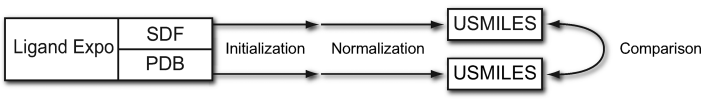


Figure 10. Scheme for the comparison of molecules from the Ligand Expo data set. Generated molecules from the PDB format are compared to the respective structures from the SDF format.

Table 8. PDB Component Names and Numbers of Errors for Those Molecules for Which the Most Errors Occured

name	no. errors	name	no. errors	name	no. errors	name	no. errors	name	no. errors
HEM	1794	PGV	194	CYC	187	BCR	182	LLP	164
ACB	124	FAD	124	PEK	120	PEV	102	MLE	84
PSO	124	BLA	83	1MA	81	MYR	80	7MG	80
OLC	79	MDO	77	NAD	77	PDU	76	UMP	73

This means that the compound in the entry does not represent an isolated molecule and that necessary information is missing. These errors can often be avoided when the complete PDB entry including the protein environment is used. The reasons for the differences encountered for PSO are quite similar. The psoralen is also covalently bound to a nucleotide but in this case no atoms from the initial component are missing. This connection is, nevertheless, reflected in the coordinates by a change of hybridization geometries for the carbon atoms in the five-membered ring. Since the molecule contained in the LigandExpo data set is an isolated psoralen, the different perception is not surprising. In case of LLP, PDU, and 7MG the structures provided by the LigandExpo data set seem to be wrong (see Figure 11).

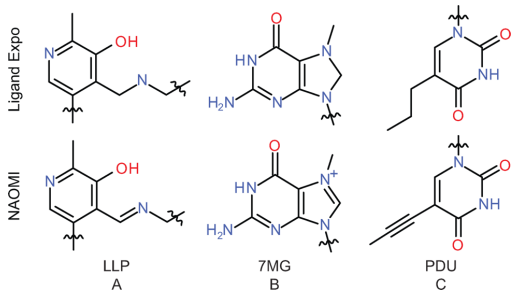


Figure 11. Comparison of inconsistent structures from the LigandExpo data set to those generated by NAOMI.

The compound LLP represents a lysine residue covalently linked to a pyridoxal phosphate via an imine group. This double bond is not present in any of the structures from LigandExpo although it is reflected by a short bond length in the coordinates. The name for the compound PDU on the LigandExpo Web site is 5(1-propynyl)-2′-deoxyuridine-5-monophosphate which indicates the presence of a triple bond. This is also confirmed by an analysis of the molecule's geometry. This triple bond is, however, not present in the reference structure. 7MG is supposed to be 7*N*-methyl-guanosine-5′-monophosphate, a molecule with a charged five-membered heterocycle which is generated by our method. The structure found in the LigandExpo data set, however, has a carbon atom with an sp^3 hybridization in the five-membered ring.

We think that these examples are sufficient to provide a general overview of the reasons for the observed differences. A special case worth mentioning are molecules with fewer than eight heavy atoms, such as solvents and auxiliary agents. Because of the extreme deviations from ideal geometries, these entries can often not be handled on the basis of atomic coordinates alone. We believe that in some cases these molecules were of minor interest to the researchers and less care was taken during the structure determination process.

When interpreting the results of the comparison one has to keep in mind that our method solely relies on the atomic coordinates provided by the file format. The reference molecules in the Ligand Expo data set are, however, derived from various inputs. In particular, this includes information about the components provided by the crystallographers. This means that the provided coordinates are not necessarily in perfect agreement with the structures present in the data set. In the end 10349 (61.0%) of the 16958 remaining entries differ by only one bond order and the respective bond is terminal in 6063 (35.8%) of these cases. This shows that the generated structures, even if they are not identical, are generally in good agreement for the larger part of the molecules.

Runtimes. The runtimes for the conversion from both the PDB and the SDF format to USMILES are shown in Table 9.

Table 9. Runtimes for the Conversion of the Ligand Expo Data Set from PDB and SDF to USMILES

data set	entries	runtime (s)
PDB (all)	602704	147
SDF (all)		79
PDB (>7 atoms)	204797	110
SDF (>7 atoms)		64

The conversion from SDF provides a point of reference for the performance of our method, since the steps after the generation of the valence bond structure are identical for both formats. Due to the numerous monatomic and small molecules (e.g., solvent molecules) in the data set, we also used a subset where all entries with less than eight atoms have been excluded. This data set provides a more realistic picture of the average runtimes per molecule. The molecule entries in the PDB format were only supplied as single files in a tar archive, which can cause large IO overhead. To avoid this, we concatenated all files into one large file which is a common procedure for other formats such as SDF.

Time measurements were performed on a PC with an Intel Core2 Quad Q9550 CPU (2.83 GHz) and 4 GB of main memory. The average runtime for the conversion of a single molecule from the PDB format is approximately 1 ms. The comparison to the value obtained for the SDF format (0.4 ms/molecule) shows, that the runtimes lie well in the range of conventional file format conversions. Our method can hence be used even in large scale applications.

CONCLUSION

We have presented a novel method for the perception of molecular structures from atomic coordinates. This method is based on the recently published NAOMI model,[11] which has been developed for the appropriate representation of organic molecules. The robustness of our approach has been assessed by processing the Ligand Expo data set in PDB format and comparing the resulting molecules to the structures from the corresponding SDF files. The results are correct in more than

dx.doi.org/10.1021/ci300358c | J. Chem. Inf. Model. 2013, 53, 76−87

95% of the cases showing that our method is able to produce reasonable results even when working with coordinates of varying quality. The method's accuracy has been demonstrated by comparison to manually curated molecules from previously published benchmarking sets. Our method was successful in 98% of the cases and was able to generate reasonable molecular representations even from structures with distorted geometries. A direct comparison to the tools fconv, I-interpret, and MOE shows that the combination of geometrical and chemical criteria used in our method is the key to avoid many assignment problems. Due to the average runtime of less than 1 ms per molecule the method is perfectly suitable for large scale applications.

Since the method is based on the NAOMI model, it is currently limited to organic molecules which can be represented by valence bond structures. This limitation does, however, only exclude a small number of molecules in the PDB and is thus considered acceptable. Because of missing hydrogen atoms and low resolution of most PDB entries the appropriate tautomeric form can usually not be deduced from the atomic coordinates alone. This would require a more advanced analysis of the ligand's energy or the explicit consideration of the molecule's environment, for example, the binding pocket of the protein, neither of which are in the scope of our method. The method is included in the current version of the NAOMI-converter which can be downloaded at http://www.zbh.uni-hamburg.de/naomi. It is available free of charge for academic use.

■ ASSOCIATED CONTENT

Supporting Information

PDB files of the 563 molecules used in the validation studies, the corresponding USMILES of the reference structures, and the converted molecules for which the perception was considered incorrect are provided. This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

Present Addresses

[§]Georg Simon Ohm University of Applied Sciences, Kesslerplatz 12, 90121 Nuremberg, Germany.

[∥]Evotec AG, Essener Bogen 7, 22419 Hamburg.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(2) PDB File Formats. http://www.pdb.org/pdb/static.do?p=file_formats/index.jsp (accessed Oct 19, 2011).

(3) PDB Format, version 3.3. http://www.wwpdb.org/documentation/format33/v3.3.html (accessed Oct 19, 2011).

(4) Meng, E.; Lewis, R. Determination of Molecular Topology and Atomic Hybridization States from Heavy Atom Coordinates. *J. Comput. Chem.* **1991**, *12*, 891−898.

(5) Baber, J.; Hodgkin, E. Automatic Assignment of Chemical Connectivity to Organic Molecules in the Cambridge Structural Database. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 401−406.

(6) Hendlich, M.; Rippmann, F.; Barnickel, G. BALI: Automatic Assignment of Bond and Atom Types for Protein Ligands in the Brookhaven Protein Databank. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 774−778.

(7) Labute, P. On the Perception of Molecules from 3D Atomic Coordinates. *J. Chem. Inf. Model.* **2005**, *45*, 215−221.

(8) Froeyen, M.; Herdewijn, P. Correct Bond Order Assignment in a Molecular Framework Using Integer Linear Programming with Application to Molecules Where Only Non-Hydrogen Atom Coordinates Are Available. *J. Chem. Inf. Model.* **2005**, *45*, 1267−1274.

(9) Zhao, Y.; Cheng, T.; Wang, R. Automatic Perception of Organic Molecules Based on Essential Structural Information. *J. Chem. Inf. Model.* **2007**, *47*, 1379−1385.

(10) Sayle, R. PDB: Cruft to Content (Perception of Molecular Connectivity from 3D Coordinates). Daylight Chemical Information Systems Inc. MUG'01 Presentation, 2001. http://www.daylight.com/meetings/mug01/Sayle/m4xbondage.html (accessed Oct 18, 2011).

(11) Urbaczek, S.; Kolodzik, A.; Fischer, R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI - On the Almost Trivial Task of Reading Molecules from Different File Formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199−3207.

(12) Cordero, B.; Gomez, V.; Platero-Prats, A. E.; Reves, M.; Echeverria, J.; Cremades, E.; Barragan, F.; Alvarez, S. Covalent radii revisited. *Dalton Trans.* **2008**, 2832−2838.

(13) Weininger, D.; Weininger, A.; Weininger, J. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.

(14) Hartshorn, M.; Verdonk, M.; Chessari, G.; Brewerton, S.; Mooij, W.; Mortenson, P.; Murray, C. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726−741.

(15) I-Interpret, version 1.0, Shanghai Institute of Organic Chemistry. http://www.sioc-ccbg.ac.cn/?p=42 software=i-interpret (accessed Oct 4, 2012).

(16) fconv—A tool not only for file conversion, version 1.24, Gerd Neudert, University of Marburg. http://pc1664.pharmazie.uni-marburg.de/drugscore/fconv_download.php (accessed Oct 4, 2012).

(17) Molecular Operating Environment (MOE), version 2011.10, Chemical Computing Group Inc. http://www.chemcomp.com/software.htm, (accessed Oct 4, 2012).

(18) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H.; Westbrook, J. Ligand Depot: A Data Warehouse for Ligands Bound to Macromolecules. *Bioinformatics* **2004**, *20*, 2153−2155.

(19) Ligand Expo, RCSB PDB. http://ligand-expo.rcsb.org/ (SDF and PDB dataset downloaded Jul 10, 2012).

A.5 Urbaczek, S.*; Kolodzik, A.*; Rarey, M. The Valence State Combination Model: A Generic Framework for Handling Tautomers and Protonation States. *J. Chem. Inf. Model.* 2014, *54*, 756–766
* equal contribution

ACS direct link:
http://pubs.acs.org/articlesonrequest/AOR-aNbSPU9D97tS5tadunet

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

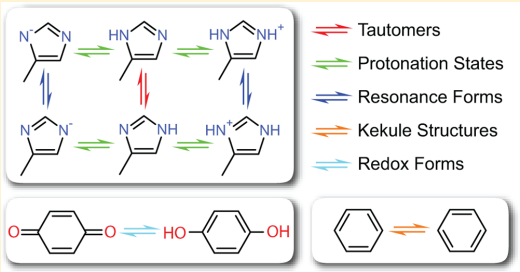
The Valence State Combination Model: A Generic Framework for Handling Tautomers and Protonation States

Sascha Urbaczek,[†] Adrian Kolodzik,[†] and Matthias Rarey*

University of Hamburg, Center for Bioinformatics (ZBH), Bundesstrasse 43, 20146 Hamburg, Germany

Supporting Information

ABSTRACT: The consistent handling of molecules is probably the most basic and important requirement in the field of cheminformatics. Reliable results can only be obtained if the underlying calculations are independent of the specific way molecules are represented in the input data. However, ensuring consistency is a complex task with many pitfalls, an important one being the fact that the same molecule can be represented by different valence bond structures. In order to achieve reliability, a cheminformatics system needs to solve two fundamental problems. First, different choices of valence bond structures must be identified as the same molecule. Second, for each molecule all valence bond structures relevant to the context must be taken into consideration. The latter is especially important with regard to tautomers and protonation states, as these have considerable influence on physicochemical properties of molecules. We present a comprehensive method for the rapid and consistent generation of reasonable tautomers and protonation states for molecules relevant in the context of drug design. This method is based on a generic scheme, the Valence State Combination Model, which has been designed for the enumeration and scoring of valence bond structures in large data sets. In order to ensure our method's consistency, we have developed procedures which can serve as a general validation scheme for similar approaches. The analysis of both the average number of generated structures and the associated runtimes shows that our method is perfectly suited for typical cheminformatics applications. By comparison with frequently used and curated public data sets, we can demonstrate that the tautomers and protonation state produced by our method are chemically reasonable.



INTRODUCTION

One of the most fundamental requirements in cheminformatics is the consistent handling of molecules from different sources. There is always the implicit assumption that the results of cheminformatics software applications are only dependent on the actual compounds and not on the way these are provided in the input data. Yet, apart from problems arising from the interpretation of data from chemical file formats, there are certain ambiguities in the way molecules are represented which considerably complicate this task. Virtually all modern cheminformatics systems are based on a description of molecules by valence bond structures (Lewis structures). The inherent limitations of this molecular representation and their implications on tautomer generation have been recently discussed in detail by Sayle.[1] In the following, we will largely follow the nomenclature used in his publication and refer back to particular aspects mentioned therein.

The main problem with respect to consistency is the fact that different valence bond structures can represent the same molecule. Some of these correspond to distinct chemical entities, e.g., tautomers and protonation states, whereas others are artifacts of valence theory, i.e., resonance forms and Kekule structures. In some contexts even oxidation states may be interpreted as alternative forms of the same molecule (see Figure 1 for examples).

From a formal point of view, each of these valence bond structures could be chosen as a representation for a particular compound. In practice, not all members of this set of alternatives are equally likely to be encountered due to automated normalization procedures and manual curation. However, despite all these efforts, a certain degree of ambiguity cannot be entirely avoided. The resulting implications for cheminformatics systems in general[2] and large compound databases in particular[3] have been thoroughly investigated in the literature. In his publication, Sayle[1] has identified five specific tasks associated with the ambiguities of molecular representations. With respect to consistency; these are comparison (#1) and, more importantly, canonicalization (#2). A cheminformatics system must be able to reliably identify and treat alternative valence bond structures as the same molecule. This is usually done by conversion to a canonical form which serves as input for subsequent methods.

Received: December 6, 2013
Published: February 18, 2014

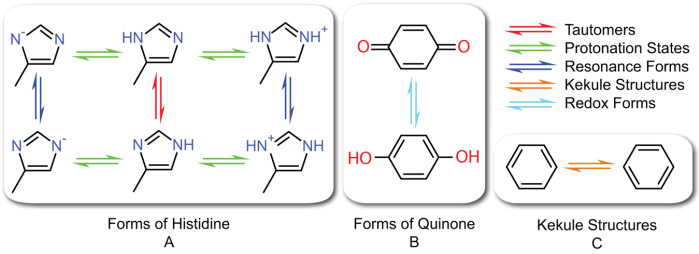


Figure 1. (A) Different valence bond structures of the imidazole ring of histidine including prototropic tautomers, protonation states, and resonance forms. (B) Two oxidation forms (quinone and hydroquinone) may in some context be considered as the same molecule. (C) Kekule structures are valence bond structures of aromatic rings with alternating single and double bonds.

The generation of unique identifiers, e.g., InChI,[4,5] is a typical application scenario for canonicalization procedures.

Another quite opposite problem arises with regard to the general reliability of cheminformatics calculations. In many cases, it is necessary to consider multiple valence bond structures to sufficiently represent a molecule. The most prominent examples are certainly tautomers and protonation states, which will be summarized under the term protomers in the following. Since these correspond to actual physical entities, their respective ratios can have significant influence on a compound's observed physicochemical properties.[6−11] The problem is, however, not exclusive to this scenario, as different resonance forms also play a role during the calculation of partial charges.[12] The respective tasks identified by Sayle[1] are (complete) enumeration (#3) and selection (#4). Both refer to the generation of valence bond structures, the difference being that selection (#4) restricts the results to a subset containing only relevant, e.g., energetically stable, solutions. Virtual screening techniques such as molecular docking are applications in which selection (#4) plays an important role. Relying on only one valence bond structure can lead to false-negative results as particular protomers may interact differently with target proteins. On the other hand, a large number of (possibly energetically unfavorable) alternatives can result in an increased false-positive rate and unnecessarily high runtimes. The general implications on structure-based and ligand-based screening methods have been investigated in several publications.[13−15] The final task mentioned by Sayle[1] is prediction (#5), which extends selection by additionally ranking the relevant solutions by their respective energy.

The basic problem associated with the interconversion of valence bond structures is to transform groups of atoms according to specific rules with respect to bond orders and atomic properties (formal charges, bound hydrogens). As has been proposed by Sayle,[1] the methods developed for that purpose can be roughly divided into two categories: (1) Local approaches rely on pattern matching to identify relevant groups of atoms. These patterns are associated with rules describing the respective changes in the molecule. Pattern-based methods thus only use transformations that were anticipated in advance, thereby reducing the risk of generating unexpected and probably unwanted results. On the other hand, there is always the possibility of omitting relevant structures due to missing patterns. This can occur even if rules of a similar type are already included in the pattern library. Transformations covering long bond paths are a typical example for that problem. There are multiple publications describing local methodologies in the literature.[13,16−18] (2) Global approaches predefine substructures in a molecule, identify atoms with variable states within, and subsequently enumerate valid valence bond structures. This is usually done in a more generic manner than matching specific patterns, so that the results can easily contain completely artificial, i.e. chemically unreasonable, results. These either have to be omitted directly during or removed after the enumeration procedure. The omission of transformations in more complex structures, however, is generally not a problem. Global approaches have also been described in the literature[19−21] and other sources.[22] It must be noted that the previous differentiation between the two types of methods has been introduced mainly for classification purposes. Local approaches, for instance, often include a number of long-range patterns which, in combination with the underlying transformation engine, makes them suitable for the handling of the vast majority of molecules relevant in the field of drug design.

Here, we present the valence state combination model, a new concept for the description and classification of valence bond structures based on the NAOMI[23] framework. Using this model, we have developed, based on similar ideas as the ones presented by Sayle et al.,[22] an extended and significantly improved method for the generation of valence bond structures which falls into the general category of global approaches. By application of a generic scoring scheme, this method combines the inherent consistency of the global strategy with the high reliability generally attained by local approaches. In contrast to previously published global methods, our approach consistently deals with all aspects relevant for the generation of protomers, including resonance forms and ionization states. Our method has been used to solve three common cheminformatics tasks, namely the generation of a canonical form (canonicalization), the generation of a preferential representation (normalization), and the generation of a set of reasonable protomers (generation). We have tested each application with respect to consistency using a general and comprehensible validation scheme. Furthermore, we have assessed the general suitability of our approach for common cheminformatics applications on the basis of these three operations. The criteria for the evaluation comprise runtime, the average number of generated structures, and the quality of the resulting protomers.

METHODOLOGY

Valence State Combination Model. Valence bond structures of molecules are generally represented as graphs in which nodes correspond to atoms and edges correspond to bonds. Each atom is associated with an element and a formal charge and each bond with a localized bond order (single,

dx.doi.org/10.1021/ci400724v | J. Chem. Inf. Model. 2014, 54, 756−766

double, or triple). In the NAOMI model,[23] this description is extended by an atom-based valence state descriptor. A valence state is a chemically valid combination of bond orders and formal charge for a particular element (see Figure 2). This

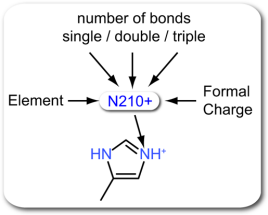


Figure 2. Example of a valence state descriptor for a nitrogen atom. The descriptor comprises the atom's element, bond order distribution, and formal charge.

additional descriptor is used to ensure the chemical validity of a molecule. A valence bond structure is valid if a valence state with the given bond orders and formal charge exists for each atom. Furthermore, valence states provide the means to systematically classify and generate different valence bond structures of molecules as explained below.

A set of valence states for all atoms of the molecule is called a valence state combination (VSC). A VSC is valid if a distribution of bond orders compatible with these valence states exists. Valid VSCs thus correspond to valence bond structures associated with a particular heavy atom skeleton. Note that bond orders are not part of the VSC representation; they are used for validation purposes only. Relations between valence bond structures can be determined by comparison of their corresponding VSCs (see Figure 3).

The description of these relations is based on atoms with different valence states, considering both their number and their types. Depending on the changed properties, substitutions of valence states for atoms are classified as protonation type, tautomer type, and resonance type as shown in Table 1. The involved states are called donors (higher number of single bonds) or acceptors (lower number of single bonds). The respective numbers of substitutions in VSCs are denoted as $\Delta_{type}(D \rightarrow A)$ and $\Delta_{type}(A \rightarrow D)$.

Table 2 lists the six basic relation types together with their conditions. Distinct valence bond structures with identical VSCs correspond to Kekule forms. They differ only in their respective bond order distribution. If all substitutions between two VSCs are of the protonation type, two cases need to be distinguished. When changing a donor to an acceptor or vice versa, the formal charge of the respective atom changes due to the addition or removal of hydrogen atoms. If the number of substitutions of donors and acceptors is not equal, the total charge of the molecule is altered, resulting in a different ionization state. Otherwise, the net charge of the molecule is identical, meaning that protons are merely occupying different locations. Tautomers and mesomers contain only changes of the tautomer-type and the resonance-type, respectively. Additionally, the number of donors and acceptor substitutions must be equal. Otherwise, the VSCs represent different redox forms of the molecule.

Substitution types can also occur in mixed constellations, and the resulting relations are best described as combinations of the just presented basic types. The 1-hydroxy-2-pyridone mentioned by Sayle[1] is an interesting example. The valence bond structures shown in Figure 4 can be best characterized as different resonance forms, a zwitterionic and a neutral one, with different proton positions.

The algorithms presented in the following chapters are based on the VSC representation of molecules. One of its major advantages is the fact that all of the potentially relevant molecule states can be consistently generated by considering different types of valence state substitutions. By explicitly handling all the different cases described in this section, a high degree of generality can be achieved.

Overview. The complete workflow for the generation of valence bond structures is shown in Figure 5. In the first step, the molecule is subdivided into multiple nonoverlapping substructures which are then treated independently. This partitioning reduces the computational costs for both the generation and the subsequent scoring of VSCs. A partition is considered valid if the independent enumeration of VSCs of each part and a subsequent combination of these lead to the same VSCs as if the enumeration would have been performed on the whole molecule. A partition is optimal if it is valid and has the smallest possible substructures. In the following sections, two partitioning schemes (generic and heuristic) are presented. Both are applied for the solution of different cheminformatics tasks described in later sections.

After partitioning, the atoms of each substructure are checked for alternative valence state assignments. Which valence states are included strongly depends on the context and will be explained in more detail later. As well as partitioning, valence state selection has a strong influence on the computational costs of the subsequent steps. The more alternatives are selected, the more VSCs must be generated and potentially scored. An optimal selection scheme thus only selects valence states for atoms that actually need to be modified. Again, two selection schemes (generic and heuristic) for different applications will be presented.

In the next step, VSCs are generated for each substructure using the alternative states selected in the previous step. Each of these VSCs is checked for validity by attempting to calculate a bond order distribution. VSCs for which this is not possible are invalid and therefore rejected. During the calculation, additional boundary conditions, e.g., the oxidation state of the initial molecule, are preserved.

The resulting VSCs are all chemically valid but may still contain undesired valence bond structures. These include unstable tautomers, unlikely protonation states, unreasonable resonance forms, or unusual representations of functional groups. In order to identify and eventually remove these VSCs, a pattern-based scoring scheme is applied. The resulting score expresses how well a particular substructure of the molecule is represented by the respective VSC. It must be stressed that the scoring scheme has not been designed to accurately predict the ratios between different molecular species. Its two main purposes are the elimination of completely artificial representations, i.e., energetically inaccessible states, and the coarse categorization of the remaining VSCs into stability classes. After eliminating all undesired VSCs, the final valence bond structures are completely enumerated by combining the VSC of the different substructures.

Partitioning of Molecules. The partitioning algorithm is based on the exclusion of atoms and bonds from the molecular graph and the subsequent identification of the remaining connected components. These will be referred to as Multi State

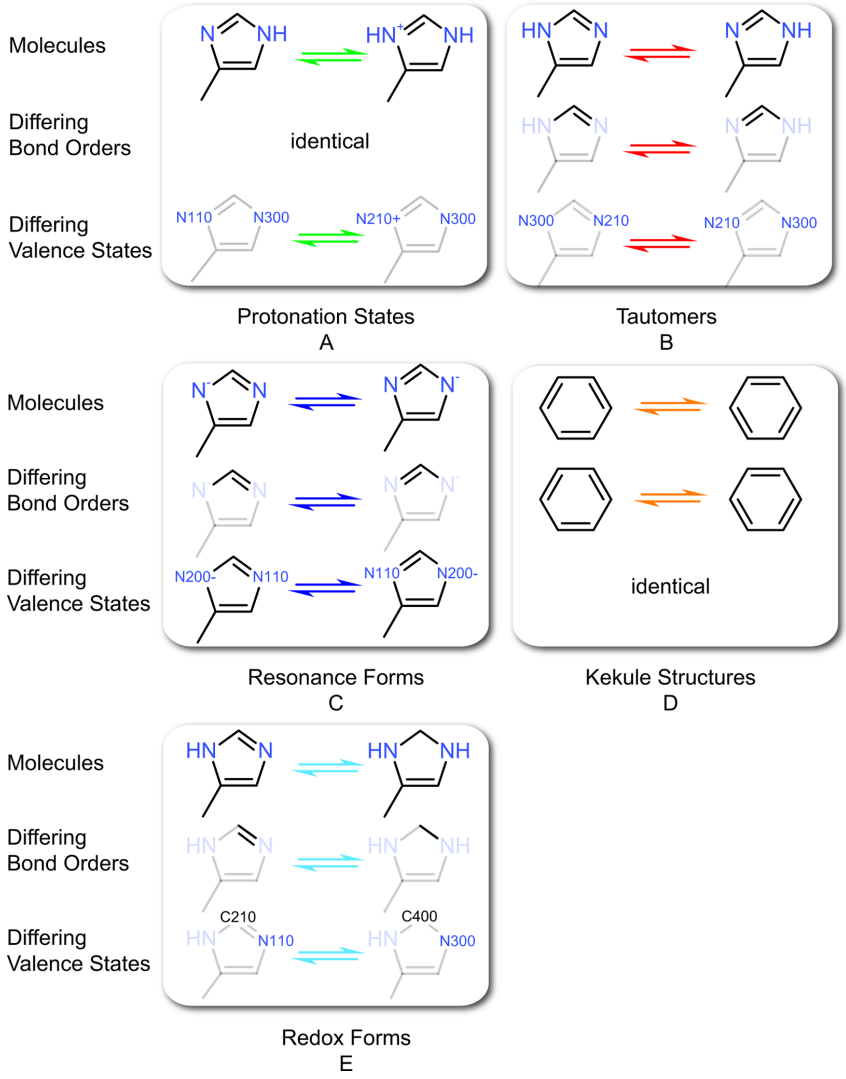


Figure 3. Differences between protonation states (A), tautomers (B), resonance forms (C), Kekule structures (D), and redox forms (E).

Table 1. Substitution Types for Valence States Including the Affected Properties[a]

				examples	
type	double bonds	# bonds	charge	donor	acceptor
protonation	0	±	±	O200	O100-
resonance	±	0	±	O100-	O010
tautomer	±	±	0	O200	O010

[a]Changed properties are marked with a ± and unchanged properties with 0. The pairs of valence states on the right side of the table represent common substitutions for oxygen atoms.

Table 2. Relations between Valence Bond Structures on the Basis of Valence State Substitution

relation	substitution type	condition
kekule	none	
ionization	protonation	$\Delta(D \rightarrow A) \neq \Delta(A \rightarrow D)$
protonation	protonation	$\Delta(D \rightarrow A) = \Delta(A \rightarrow D)$
mesomer	resonance	$\Delta(D \rightarrow A) = \Delta(A \rightarrow D)$
tautomer	tautomer	$\Delta(D \rightarrow A) = \Delta(A \rightarrow D)$
redox	resonance	$\Delta(D \rightarrow A) \neq \Delta(A \rightarrow D)$
	tautomer	$\Delta(D \rightarrow A) \neq \Delta(A \rightarrow D)$

Partitions (MSP) in the following discussion. The **generic** partitioning scheme only involves the exclusion of sp3-hybridized carbon atoms (corresponds to valence state C400). There are only two particular cases in which atoms with valence state C400 are included in MSPs: first, if the atom is bound to an atom with valence state C210, which in turn has

dx.doi.org/10.1021/ci400724v | J. Chem. Inf. Model. 2014, 54, 756−766

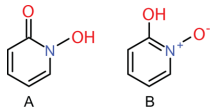


Figure 4. Example for the combination of valence state substitutions. The relation between the pyridone form (A) and the pyridine form (B) cannot be described by one of the basic types from Table 2.

at least one neighbor with the element nitrogen, oxygen or sulfur, and second, if the atom is part of a ring and is the only atom with valence state C400 in this ring. Bonds are excluded if one of the connected atoms is excluded.

The MSPs resulting from the generic partitioning scheme are usually large, and it is often possible to further reduce their size. This is achieved by removing bonds within the MSPs with the goal to effectively split them into smaller substructures. The exclusion of a bond is only valid if its bond order in the current structure is identical in all relevant VSCs. Since the final bond orders are not known at this point, the decision that a bond will keep its current type must be in accordance with the subsequent scoring procedure. This means that VSCs with a different bond order would be rejected in the follwing steps in any case.

The **heuristic** partitioning scheme builds on the results from the generic scheme and uses a set of rules to identify additional bonds for exclusion. These rules are based on the classification of each MSP into conjugated rings, conjugated chains, and functional groups. Rings are considered conjugated if all of their atoms are part of the respective MSP. Conjugated chains consist only of carbon atoms which have a multiple bond and are bound only to other carbon atoms. The remaining connected components represent functional groups. In a first step, bonds connecting functional groups with conjugated rings or conjugated chains are investigated. A bond is excluded if it is a single bond and the atom from the functional groups does not fulfill one of the following two criteria: (1) It has a valence state of type N300. (2) It has a valence state of type O200 or S200 and only one non-hydrogen bond. In these cases, a change in bond order is not unlikely, as is shown for two examples in Figure 6.

Since conjugated chains consist of only carbon atoms, they are merely bridges between the other two types of

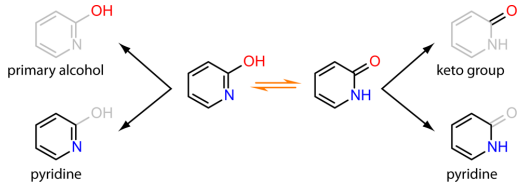


Figure 6. Two examples for which functional groups and conjugated rings have to be treated as a union to avoid missing VSCs.

substructures. Therefore, if a conjugated chain has only one bond to another structure (ring or functional group), this bond can be safely excluded. This is also done if the chain has multiple bonds which were previously excluded by the functional group rule. The complete partitioning of the NAD + molecule is shown as an example in Figure 7.

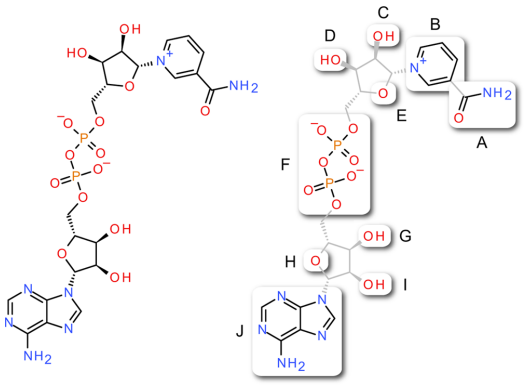


Figure 7. Partitioning of NAD+ into functional groups and conjugated rings. The amide group and the pyridine ring have been separated, whereas the amino group remains connected to the purine.

Selection of Valence States. The selection of valence states is based on the substitution types introduced above (see Table 1). Each substitution corresponds to a pair of valence states which are known in advance and can be retrieved starting

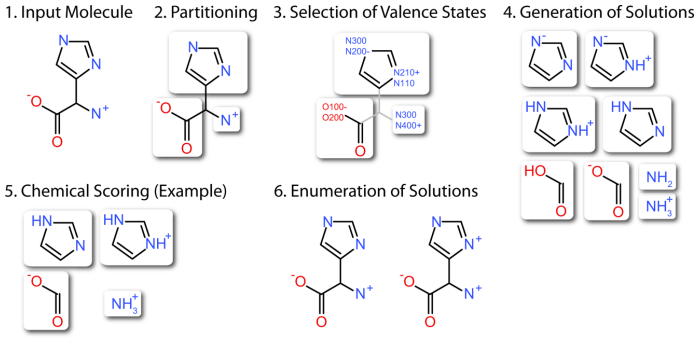


Figure 5. Overview of the generation of protonation states for an input molecule (1). In a first step, the molecule is partitioned into substructures (2) which are handled separately. In the next step, alternative valence states are selected (3). Afterward, valid VSCs are generated (4). These are scored (5), and only the best solutions for each zone are retained. The final list of valence bond structures results from the combination of all remaining VSCs.

from any valence state. A list of alternatives for an atom can thus be easily obtained by consecutively and uniquely adding the respective members of the pairs for each of the relevant substitution types. In order to select an alternative assignment, the compatibility with the atom's topology must be ensured. This means that the number of bonds of the valence state must be larger than or equal to the atom's number of non-hydrogen bonds. Otherwise, the assignment would correspond to the removal of non-hydrogen bonds. Although this may be interesting with respect to transformations such as ring−chain tautomerism, it will not be further considered here.

For the sake of generality, the **generic** selection procedure includes all possible valence states for each atom in a MSP. This usually results in potentially many more alternatives than are actually needed. The **heuristic** selection scheme aims at reducing this number by explicitely excluding valence states for particular atoms. The problem at this point is similar to the one discussed in the previous section. The final VSCs are not yet known, and the decisions must be in accordance with the subsequent scoring procedure in order to avoid missing VSCs.

The exclusion of particular valence states in the **heuristic** selection scheme is based solely on an analysis of the atom's environment. For atoms in functional groups, this includes their direct neighbors from the same functional group. These are transformed into a SMILES-like identifier which reflects the valence bond structure of the input molecule. This identifier is looked up in a list of predefined structures. If the identifier is present, information concerning the exclusion of particular substitution types is retrieved. In this way, groups that already have a preferred representation in the initial valence bond structure need not be modified. The information provided from the patterns is described in Figure 8.

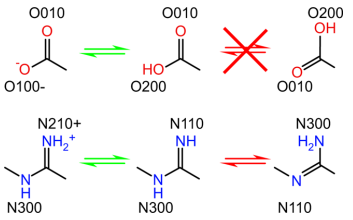


Figure 8. Selection of valence states for carboxylic acid and amidine groups. Due to the group's symmetry, different tautomers of carboxylic acids are not considered.

For the generation of tautomers, carboxylic acids are irrelevant. Due to the symmetry of the group the transfer of the hydrogen from one oxygen to the other would only result in a different rotamer. In this case both oxygen atoms are excluded from tautomer substitution. With respect to protonation, both the charged and the neutral form need to be included. This means that both oxygens are not excluded from protonation substitution. The same procedure is applied to atoms in conjugated rings with the ring constituting the atom's environment. If the identifier is not included, the **generic** scheme is used to identify alternative states for the atom.

Generation of Valid VSCs. Prior to the generation of VSCs, each MSP is analyzed to ensure that the generation of additional states is at all possible. MSPs can be ignored if no atom with alternative valence states could be found. For tautomers and mesomers, i.e. if new bond order distributions are to be generated, MSPs can also be omitted if only either donors or acceptors are present. In this case, no substitution of valence states is possible (see Figure 9 for examples). Changing the number of donors and acceptors corresponds to changing the oxidation state of the molecule, which is not desired in most contexts.

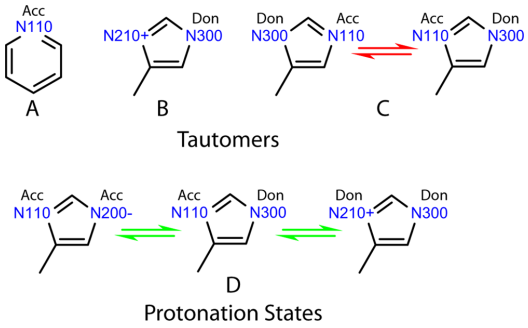


Figure 9. Criteria for the generation of additional states. The generation of tautomers requires at least one tautomer acceptor and one tautomer donor in a zone. Pyridine (A) has only a single tautomer acceptor, and the imidazolium ion (B) only has one tautomer donor. No tautomers can be generated in such cases. Imidazole (C) contains a tautomer acceptor as well as a tautomer donor and can tautomerize. Protonation states (D) can also be generated if a molecule only contains either protonation-type donors or acceptors.

The algorithm for the generation of valid VSCs is based on a backtracking procedure with pruning. The atoms of the MSP are processed in a specific order which is established prior to the actual assignment procedure. The algorithm starts with terminal atoms, i.e. atoms with only one bond in the MSP, followed by internal atoms with at least one terminal neighbor. The remaining atoms are processed last. The order of the atoms inside the three classes is arbitrary and does not affect the result. As a combinatorial problem, the procedure can be represented by a tree, where each node corresponds to the assignment of a valence state to an atom. Inner nodes thus represent partial VSCs while the tree's leaves correspond to complete VSCs. For each node, the chemical validity of the corresponding VSC is verified. In most cases, this can be performed without actually generating bond orders for the bonds of the MSP. The checks are based on the compatibility between valence states of different atoms with respect to the expected bond types as well as their oxidation states: (1) For atoms with only one bond in the MSP, the assignment of a valence state is equivalent to the assignment of a bond order to the corresponding bond. The compatibility with the atom's neighbors can be easily checked by ensuring that the count of this particular bond type is not exceeded. This check is always performed when an atom with terminal neighbors is encountered. (2) When reaching a leaf, the valence states with an uneven number of multiple bonds are counted. If this number is uneven, no valid bond order distribution exists, and the VSC can be further ignored. (3) The number of donors in the initial valence bond structures is counted in order to retain the molecule's oxidation state. VSCs differing in the number of donors compared to the initial valence bond structures can be discarded. Note that since information about being a donor or

acceptor is also stored in the valence states, VSCs not fulfilling this boundary condition can be easily identified. (4) Eventually, for each VSC passing all previous checks, a recursive bond localization routine is used which assigns bond orders to all bonds in the MSP. If this routine is successful, the solution represents a valid valence bond structure and is stored.

Scoring of VSCs. Scores for each VSC are calculated under consideration of the bond order distribution generated in the previous step. The scoring procedure is mainly based on the recognition of predefined structural fragments contained within particular substructures, i.e., conjugated rings and functional groups, of the molecule. The final score of the VSC (S_{VSC}) is calculated as the sum of the individual scores obtained for each of these substructures (see eq 1). Please note that due to changes in bond orders and valence states, the scores have to be recalculated for each VSC.

$$S_{VSC} = \sum S_{ring} + \sum S_{group} \quad (1)$$

$$S_{ring} = \sum \text{cycle} + \sum S_{sub} \quad (2)$$

$$S_{group} = \sum S_{subgroup} \quad (3)$$

The structural fragments in the substructures are identified using canonical SMILES-like identifiers. These are generated on the basis of the bond types and valence states of the respective VSC. The predefined data are stored in multiple databases which can be queried with the identifiers in order to retrieve the score associated with a fragment.

In case of conjugated rings, the score S_{ring} comprises two types of contributions, one from the ring itself, S_{cycle}, and one from its substituents, S_{sub} (see eq 2). The reference point for S_{cycle} is the isolated aromatic system without exocyclic double bonds, e.g., pyrrole for a five-membered ring with one nitrogen atom. In case there are multiple structures fulfilling this requirement, e.g., the 1H and 2H tautomers of 1,2,3-triazole, one is arbitrarily selected. The score of the reference system is set to an arbitrary value of 100. If a ring with an identical heavy atom connectivity does contain a structural deviation from the reference, e.g., an sp^3 hybridized carbon atom, the associated fragment has an individual score. This can be higher or lower depending on the stability assigned to this particular arrangement. The substructures representing ring substituents comprise the ring atom, the exocyclic atom, and the exocyclic atom's direct neighbors. The associated scores have fixed values and are independent from the concrete ring system they are connected to. Again, one particular representation of the substituent, the one with an exocyclic single bond and without charges, receives an arbitrary reference score of 100. Functional groups are first treated as a whole; i.e., an identifier for the complete group is generated. If the pattern was present, the associated score is directly set as the score of the substructure. Otherwise, the group is partitioned into smaller pieces which serve as starting points for further queries. In this case, the score for the group is composed of the scores of the smaller fragments (see eq 3). The reference system for a subgroup is preferably neutral and corresponds to the most stable tautomeric form where possible.

If no predefined data are available in any of the three cases, a generic score is calculated according to eq 4:

$$S_{generic} = \max(0, 80 - \sum P) \quad (4)$$

This is done by subtracting various penalties (P) which are summarized together with the respective conditions in Table 3.

Table 3. Classification and Conditions for the Penalties used during the Calculation of Generic Scores

substructure	type	penalty	condition
ring	aromaticity	20	nonaromatic ring (Hueckel's rule)
ring	charge	20	single charge in ring
ring	charge	80	multiple charges in ring and substituents
ring	stability	80	three consecutive donors[a] in the ring
substituent	bond order	20	substituent has exocyclic double bond
substituent	charge	20	single charge in substituent
substituent	charge	80	multiple charges in substituent
group	charge	80	multiple positive charges in group

[a]Donors are atoms with the following valence states: O200, N300, S200.

Since $S_{generic}$ is used only as a fallback, the respective maximal score is deliberately set lower than that of the reference system. If the sum of the penalties (P) exceeds 80, the score of the substructure is set to zero.

The relative differences between the scores of rings, substituents, and functional groups have been derived from multiple pairs of tautomers and ionization states for which the major form was known from either experiments or theoretical calculations.[24] The databases currently contain 252 entries in total (113 in cycles, 121 in subgroups, 18 in substituents). Examples for ring and functional groups patterns are shown in Figure 10.

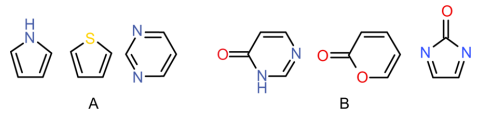


Figure 10. Examples for ring and functional groups patterns. (A) A reference score of 100 is assigned to isolated aromatic rings without exocyclic double bonds. (B) The score for rings with exocyclic double bonds comprises one contribution from the ring and another from the carbonyl substituent.

VSC MODEL APPLICATIONS

In the following applications, we will consider resonance forms, prototropic tautomers, and protonation states as instances of the same molecule, whereas oxidation forms are interpreted as distinct chemical species. The method is, however, not restricted to this assumption in general and can be easily modified so that different types of valence bond structures are perceived as identical.

Canonicalization. The generation of a canonical representation is the first workflow in which our method is applied. Canonical representations are mainly used to determine whether two valence bond structures represent the same molecule. In this context, it does not matter if the result corresponds to the most stable form or even a chemically reasonable one.

The workflow starts with the partitioning of the molecule into MSPs and the selection of alternative valence states as described above. The atoms of each MSP are then sorted in a

dx.doi.org/10.1021/ci400724v | J. Chem. Inf. Model. 2014, 54, 756−766

canonical way using a variant of the Morgan extended sums algorithm.[25] The backtracking algorithm in the generation step processes the atoms in this exact order until the first valid VSC has been found. This VSC serves as the canonical form of the respective substructure. Since no additional scoring is needed, the canonical VSCs of each substructure can be directly combined to yield the canonical representation for the complete molecule.

For the canonicalization to work correctly, the results must be identical for each possible valence bond structure of the molecule provided as input. This can only be achieved if the substructures generated in the partitioning step and the lists of valence states identified in the selection step are both identical in each case. The heuristic algorithms for partitioning and selection are therefore inappropriate, and the generic variants are applied. Since only a single valid VSC must be generated in the end, the size of the substructures and the number of alternative states are of less importance for the resulting compute time. Nevertheless, in order to further accelerate the process, all charged valence states are transformed into their neutral states where possible (considering the number of hydrogens) using the protonation-type substitution. Consequently, only tautomer-type and resonance-type substitutions need to be considered in the next steps.

The canonicalization procedure applied to the atoms of each substructure differs only in one aspect from the CANON algorithm used for the generation of USMILES.[26] In the CANON algorithm, the atomic invariants correspond to the atom's valence state in combination with the number of attached hydrogens. This means that the initial ranks of atoms can normally be deduced by comparing valence states and hydrogens. In case of a yet unknown valence bond structure, the final valence state of an atom is, however, not defined. Instead, a list of valence states is used to describe the topology of each atom and provide the initial ranks. Furthermore, the number of non-hydrogen bonds serves as a replacement for the number of hydrogens.

Normalization. The aim of normalization is the generation of a canonical valence bond structure which additionally adheres to common conventions for the representation of molecules. This task seems, at least at first glance, quite similar to the previously described canonicalization. The main difference results from the necessity of a scoring step in order to determine the best suited choice for the molecule. This implies that multiple VSCs have to be generated and compared with each other. Here, we have chosen a neutral form as normalized representation, meaning that all atoms are neutralized when possible (considering bound hydrogens). The only exception to this rule is functional groups which are represented in a zwitterionic form by convention, e.g., nitro groups and n-oxides. The method is, however, not restricted to this preference and can be easily modified so that, for instance, the preferred ionization state is generated.

Again, the workflow starts with the partitioning of the molecule into substructures and the selection of alternative valence states. Due to the enumeration of VSCs in the later steps, the size of the substructures and the number of states are relevant factors. Therefore, the heuristic strategies for both partitioning and state selection are used. In contrast to canonicalization, the initial substructures and alternative valence states do not have to be identical for each starting structure. The additional scoring step ensures that the results are consistent.

In the next two steps, valid VSCs are generated and scores are assigned as explained in the sections above. For each substructure, only those solutions with the highest score are retained. If there is only one VSC left for a substructure, it can be directly assigned, and no further steps are necessary. Otherwise, a canonical solution has to be picked from the VSCs with the highest score. This is done using the canonicalization method described in the previous section. However, since this method only works correctly in case of identical MSPs and lists of valence states, a preprocessing step is required. The respective MSP is repartitioned by exclusion of bonds having the same bond type in all VSCs. Additionally, all valence states which could not be found in one of the remaining VSCs are removed from the lists of alternatives. This eventually creates the necessary conditions for the canonicalization procedure.

Generation. The last application of our method is the generation of a set of reasonable tautomers and protonation states of a molecule. The resulting molecules can be used as input for methods that rely on the positions of hydrogen atoms such as docking. They can also serve as a starting point for the determination of the energetically most stable form of a molecule under consideration of the molecule's local environment, e.g., bound to a protein. The inclusion of multiple resonance structures, although possible with our method, is not considered useful in this context.

The initial steps of the workflow are identical to those described for normalization. But instead of canonically selecting one of the remaining VSCs of each zone, the combinations are enumerated in order to generate a set of molecules. One major difference from the previously presented approaches is the possibility of generating duplicates due to molecular symmetry. This is avoided by removing VSCs from each zone that would lead to identical valence bond structures in the resulting molecules. For this purpose, automorphism classes for atoms are calculated using the Morgan algorithm, which is also used for the canonicalization. In combination with the respective valence state of an atom in a VSC, these classes can be used to generate a string representation of each VSC in a zone, which are used to identify and remove duplicates.

For molecules containing more than one ionizable group, it is usually not desirable to enumerate all combinations of VSCs from the respective zones. To avoid chemically unreasonable species with a high number of charges, the maximum number of charges in the complete molecule is restricted by three simple rules: (1) The number of charged groups must be smaller than four, (2) the number of pairs of oppositely charged groups is smaller than two, and (3) the maximum number of positive charges in a ring system is restricted to one.

RESULTS AND DISCUSSION

The three applications presented in the previous sections are the basis for the evaluation of our method in terms of consistency, quality, and performance. Throughout these studies, the following commonly used public data sets served as input: (1) ZINC clean leads[27,28] (ZINC-CL), (2) LigandExpo component dictionary[29,30] (LEXPO-CD), (3) Drugbank[31,32] (DRUGBANK), and (4) ChEMBL.[33,34] All calculations and runtime measurements were performed on a PC with an Intel Core i5−3570 CPU (3.40 GHz) and 8 GB of main memory.

Consistency. Independence from the initial valence bond structure of a molecule is a fundamental requirement of the presented method and has been thoroughly investigated for

dx.doi.org/10.1021/ci400724v | J. Chem. Inf. Model. 2014, 54, 756−766

Table 4. Runtimes for the Three Workflows with Different Data Sets

	ZINC-CL	LEXPO-CD	DRUGBANK	ChEMBL
# total molecules	5735035	17310	6583	1318187
runtime canonicalization [ms/cmpd]	0.28	0.41	0.45	0.71
runtime normalization [ms/cmpd]	0.31	0.50	0.6	0.73
runtime generation [ms/cmpd]	0.45	0.75	0.75	1.37

each of the three applications. Consistency can be verified by a simple and straightforward procedure. The starting point is a set containing different representations of the same molecule, e.g., as different molecule entries in a file. After applying the respective workflow to each representation, the resulting molecules are converted to USMILES for comparison. If the method is consistent, all resulting USMILES are identical. In case of enumeration, lists of USMILES must be compared.

The best way to ensure consistency would be to test all possible valence bond structures of the molecule with the procedure described above. This is, however, not feasible in many cases due to the prohibitively large number of resulting molecular states. We therefore decided to reduce the set by exclusion of protonation and ionization states (see Table 2 for our definition), since the main complexity of the task results from valence bond structures with different bond order distributions.

The input structures needed for the assessment of our method's consistency were generated using a workflow corresponding to the one described for the canonicalization of molecules. But instead of selecting a canonical form, we generated valid VSCs without any scoring step and enumerated all possible combinations. Identical results could be achieved for all three workflows, canonicalization, normalization, and generation, with all four data sets mentioned above.

Runtimes. Table 4 lists the runtimes for the three workflows with the above-mentioned data sets. The results for canonicalization and normalization are comparable in both cases, whereas the time needed for the generation of a set of states is higher. This is not surprising since the workflow involves the enumeration of multiple molecule states and the built-in elimination of duplicates based on automorphism classes. In all cases, an average runtime lower than 1.5 ms per molecule is measured, thus showing that our method is suitable for processing large data sets. The similarity of results for canonicalization and normalization are most probably a consequence of the normalization procedures used during the curation of the used data sets. As has been explained above, the runtimes for normalization are highly dependent on the input form of the molecule, and the process is accelerated by reasonable initial representations.

Normalization. The main purpose of normalization is to transform different input forms of the same molecule into an identical and at the same time chemically reasonable representation. We have already shown that our normalization workflow is consistent for the four data sets used in this study. Here, we will focus on the second aspect. We believe that the best way to investigate if results are chemically reasonable is to compare the resulting valence bond structure with those found in frequently used and curated public data sets.

The procedure applied for this purpose is again based on the comparison of USMILES. Directly using the input molecule and the normalized version is, unfortunately, not suitable in many cases. As has been explained above, a canonicalization step at the end of the workflow is used to arbitrarily select one of multiple equally acceptable solutions. This makes the comparison to a reference structure, which has most likely been normalized by a different procedure, pointless. We therefore decided to enumerate all combinations of VSCs with the highest score and to check if the input structure is contained within the obtained set (best). A negative result does, however, not necessarily mean that our method generated an unreasonable result. The representation in the data set could simply correspond to a VSC which received a lower score based on our scoring scheme. For that reason, we additionally enumerated all VSCs with a score of at least 75% of the best score and also searched in this extended set (extended). The results of this validation procedure are summarized in Table 5.

Table 5. Classification of the Input Structures from Three Data Sets into Mutually Exclusive Categories for the Generation of Tautomers

	LEXPO-CD	DRUGBANK	ChEMBL
# total molecules	17310	6583	1318187
# molecules (best)	16837	6431	1252408
# molecules (extended)	364	118	52491
# molecules (not found)	135	48	11433

The differences encountered during the process can be subdivided into two classes. First, there are input structures which are not found in the best set, but in the extended set. These correspond in many cases to keto and enol tautomers of aromatic heterocycles, which are ranked differently by our scoring method (see 138 in Figure 11). Second, there are input structures which are not present in either of both sets. After visual inspection, we think that the results generated by our method are in general at least equally acceptable and in some cases even better than the representations found in the data set. The latter especially applies to charged structures for which a reasonable neutral form can be formulated (see 3MC in Figure 11). The normalized molecules generated by our method are provided as Supporting Information for all entries of LEXPO-CD and DRUGBANK which were not included in either of the two sets.

Finding the input structure in a set of equally scored alternatives is, however, only one aspect of the method's performance. Additionally, one has to make sure that the success is not simply based on the enumeration of an unreasonably large number of representations. For that reason, the sizes of the respective sets are also an important performance indicator and are shown in Table 6.

The average number of generated states is considerably lower than the result of an exhaustive enumeration. Only for a small percentage of molecules (less than 0.5%) does the number of equivalent structures actually exceed a size of five. This is in all cases caused by the combination of states from independent zones, e.g., molecules having multiple imidazole rings.

Generation. The aim of our generation workflow is to generate a set of chemically reasonable protomers of a molecule

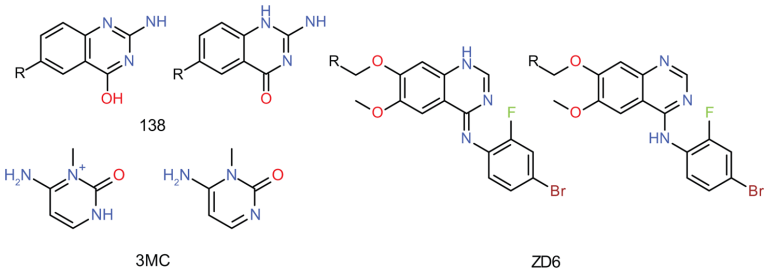


Figure 11. Examples of differences between the normalized forms generated by our method (right side) and those found in the Ligand Expo data set (left side).

Table 6. Number of Molecules with More than One and More than Five Tautomers in the Best Set[a]

	ZINC-CL	LEXPO-CD	DRUGBANK	ChEMBL
# total molecules	5735035	17310	6583	1318187
# tautomers >1	207207	1483	520	93430
# tautomers >5	671	5	5	4699
# average	2.27	2.21	2.37	8.3

[a]The provided average refers only to cases with more than one tautomer.

for typical cheminformatics applications, e.g., docking calculations. Considering this context, the resulting set should only contain states which are realistically expected to be stable in a protein−ligand complex. In order to assess the quality of our results, we used ZINC-CL as a reference set since it was generated for the exact same scenario. The procedure is identical to the one described for the evaluation of the normalization workflow. The input structure is searched in two sets, one containing the states with the highest score (best) and one containing states with a score of at least 75% of the highest score (extended). The results of the procedure are summarized in Table 7.

Table 7. Classification of the Input Structures from the ZINC-CL Data Set into Mutually Exclusive Categories for the Generation of Protomers

	ZINC-CL
# total molecules	5735035
# molecules (best)	4764463
# molecules (not best)	914921
# molecules (not found)	55651

As has already been discussed above, one important parameter for the evaluation of the method's performance certainly is the number of generated states. The results for all four data sets are summarized in Table 8.

CONCLUSION

The simple fact that the same molecule can be represented by different valence bond structures constitutes a complex challenge for cheminformatics applications. It complicates the determination of molecular identity and makes the results of cheminformatics calculations prone to inconsistencies. Furthermore, it imposes the task of selecting the best suited structure or structures for the respective context of application. The identification, description, and consistent handling of these different molecular representations is thus a fundamental requirement in the field of cheminformatics.

To cope with these problems, we have introduced a formalism which describes different valence bond structures of a molecule on the basis of the recently published NAOMI model. Using this description, we developed a general method for their fast and consistent enumeration and presented three exemplary applications. In our validation, we have shown that the devised methodology can be successfully applied to relevant tasks in cheminformatics in a consistent manner. We have also demonstrated the low runtime of our approach which makes it suitable for processing large data sets.

Table 8. Number of Molecules with More than One and More than Five Protomers in the Best Set[a]

	ZINC-CL	LEXPO-CD	DRUGBANK	ChEMBL
# total molecules	5735035	17310	6583	1318187
# protomers >1	1007976	2221	770	159663
# protomers >5	9240	183	78	13231
# average	2.54	3.14	3.20	4.40

[a]The provided average refers only to cases with more than one protomer.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information includes the normalized structures of all entries from the Ligand Expo Component Dictionary and Drugbank whose input form was not included in the results of our method. These are provided as separate SMILES files. This material is available free of charge via the Internet at http://pubs.acs.org/.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

Author Contributions

†Equal contribution.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank one of the reviewers for bringing the existence of the freely available source code of the method developed by Sayle and Delaney[35] to their attention.

REFERENCES

(1) Sayle, R. So you think you understand tautomerism? *J. Comput.-Aided Mol. Des.* **2010**, *24*, 485−496.
(2) Warr, W. A. Tautomerism in chemical information management systems. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 497−520.
(3) Sitzmann, M.; Ihlenfeldt, W.-D.; Nicklaus, M. Tautomerism in large databases. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 521−551.
(4) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the worldwide chemical structure identifier standard. *J. Cheminform.* **2013**, *5*, 7.
(5) InChI version 1, software version 1.04 (2011)-Technical Manual. http://www.inchi-trust.org/fileadmin/user_upload/software/inchi-v1.04/INChI_TechMan.pdf (last accessed Dec 06, 2013).
(6) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and Original pK_a Prediction Method Using Grid Molecular Interaction Fields. *J. Chem. Inf. Model.* **2007**, *47*, 2172−2181.
(7) Shelley, J.; Cholleti, A.; Frye, L.; Greenwood, J.; Timlin, M.; Uchimaya, M. Epik: a software program for pK_a prediction and protonation state generation for drug-like molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 681−691.
(8) Martin, Y. Let's not forget tautomers. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 693−704.
(9) Clark, T. Tautomers and reference 3D-structures: the orphans of in silico drug design. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 605−611.
(10) Cramer, R. Tautomers and topomers: challenging the uncertainties of direct physicochemical modeling. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 617−620.
(11) Greenwood, J.; Calkins, D.; Sullivan, A.; Shelley, J. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 591−604.
(12) Gilson, M.; Gilson, H.; Potter, M. Fast assignment of accurate partial atomic charges: an electronegativity equalization method that accounts for alternate resonance forms. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1982−1997.
(13) Oellien, F.; Cramer, J.; Beyer, C.; Ihlenfeldt, W.-D.; Selzer, P. M. The Impact of Tautomer Forms on Pharmacophore-Based Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 2342−2354.
(14) ten Brink, T.; Exner, T. Influence of Protonation, Tautomeric, and Stereoisomeric States on Protein-Ligand Docking Results. *J. Chem. Inf. Model.* **2009**, *49*, 1535−1546.
(15) Kalliokoski, T.; Salo, H.; Lahtela-Kakkonen, M.; Poso, A. The effect of ligand-based tautomer and protomer prediction on structure-based virtual screening. *J. Chem. Inf. Model.* **2009**, *49*, 2742−2748.
(16) Kenny, P.; Sadowski, J. In *Chemoinformatics in Drug Discovery*; Oprea, T., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2005; pp 271−285.
(17) Milletti, F.; Storchi, L.; Sforna, G.; Cross, S.; Cruciani, G. Tautomer enumeration and stability prediction for virtual screening on large chemical databases. *J. Chem. Inf. Model.* **2009**, *49*, 68−75.
(18) Kochev, N. T.; Paskaleva, V. H.; Jeliazkova, N. Ambit-Tautomer: An Open Source Tool for Tautomer Generation. *Mol. Inf.* **2013**, *32*, 481−504.
(19) Haranczyk, M.; Gutowski, M. Quantum Mechanical Energy-Based Screening of Combinatorially Generated Library of Tautomers. TauTGen: A Tautomer Generator Program. *J. Chem. Inf. Model.* **2007**, *47*, 686−694.
(20) Thalheim, T.; Vollmer, A.; Ebert, R.-U.; Kühne, R.; Schüürmann, G. Tautomer Identification and Tautomer Structure Generation Based on the InChI Code. *J. Chem. Inf. Model.* **2010**, *50*, 1223−1232.
(21) Will, T.; Hutter, M. C.; Jauch, J.; Helms, V. Batch tautomer generation with MolTPC. *J. Comput. Chem.* **2013**, *34*, 2485−2492.
(22) Sayle, R.; Delany, J. In *Innovative Computational Applications: the Interface of Library Design, Bioinformatics, Structure Based Drug Design and Virtual Screening*; IIRG publishers: San Franciso, CA, 1999. http://www.daylight.com/meetings/emug99/Delany/taut_html/sld001.htm (accessed Jan 30, 2014).
(23) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI - On the almost trivial task of reading molecules from different file formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199−3207.
(24) Raczynska, E.; Kosinska, W.; Osmialowski, B.; Gawinecki, R. Tautomeric Equilibria in Relation to Pi-Electron Delocalization. *Chem. Rev.* **2005**, *105*, 3561−3612.
(25) Morgan, H. L. The generation of a unique machine description for chemical structures - a technique developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.
(26) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.
(27) Irwin, J.; Shoichet, B. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(28) ZINC Database - Version 12. https://zinc.docking.org/ (Clean Leads Reference as SMILES downloaded Dec 03, 2013).
(29) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H.; Westbrook, J. Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* **2004**, *20*, 2153−2155.
(30) Ligand Expo; RCSB PDB. http://ligand-expo.rcsb.org/ (chemical component dictionary as SMILES (OpenEye with stereo) downloaded Jul 10, 2012).
(31) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A.; Wishart, D. DrugBank 3.0: a comprehensive resource for "Omics" research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035−D1041.
(32) DrugBank 3.0. http://www.drugbank.ca/ (all drugs as SDF downloaded Dec 03, 2013).
(33) Bellis, L. J.; et al. Collation and data-mining of literature bioactivity data for drug discovery. *Biochem. Soc. Trans.* **2011**, *39*, 1365−1370.
(34) ChEMBLdb - Version 17. https://www.ebi.ac.uk/ (ChEMBLdb including an SDF file downloaded Dec 03, 2013).
(35) Source code of the tautomer generation method by Sayle and Delany. http://www.daylight.com/meetings/emug99/Delany/tautomers/ (accessed Jan 30, 2014).

A.6 Schröder, M.*; Kolodzik, A.*; Pfaff, K.; Priyadarshini, P.; Krepstakies, M.; Hauber, J.; Rarey, M.; Meier, C. In silico Design, Synthesis, and Screening of Novel Deoxyhypusine Synthase Inhibitors Targeting HIV-1 Replication. *ChemMedChem* 2014, *9*, 940–952
*** equal contribution**

Wiley link:
http://dx.doi.org/10.1002/cmdc.201300481

CHEMMEDCHEM
FULL PAPERS

ChemPubSoc Europe

DOI: 10.1002/cmdc.201300481

In silico Design, Synthesis, and Screening of Novel Deoxyhypusine Synthase Inhibitors Targeting HIV-1 Replication

Marcus Schroeder,[a] Adrian Kolodzik,[b] Katharina Pfaff,[a] Poornima Priyadarshini,[c] Marcel Krepstakies,[c] Joachim Hauber,[c] Matthias Rarey,[b] and Chris Meier*[a]

The human enzyme deoxyhypusine synthase (DHS) is an important host cell factor that participates in the post-translational hypusine modification of eukaryotic initiation factor 5A (eIF-5A). Hypusine-modified eIF-5A plays a role in a number of diseases, including HIV infection/AIDS. Thus, DHS represents a novel and attractive drug target. So far, four crystal structures are available, and various substances have been tested for inhibition of human DHS. Among these inhibitors, *N*-1-guanyl-1,7-diaminoheptane (GC7) has been co-crystallized in the active site of DHS. However, despite its potency, GC7 is not selective enough to be used in drug applications. Therefore, new compounds that target DHS are needed. Herein we report the in silico design, chemical synthesis, and biological evaluation of new DHS inhibitors. One of these inhibitors showed dose-dependent inhibition of DHS in vitro, as well as suppression of HIV replication in cell cultures. Furthermore, the compound exhibited no cytotoxic effects at active concentrations. Thus, this designed compound demonstrated proof of principle and represents a promising starting point for the development of new drug candidates to specifically interfere with DHS activity.

Introduction

A key step in HIV chemotherapy was the introduction of combination anti-retroviral therapy (cART) in the mid-1990s, which significantly prolonged patients' expectancy of life.[1] However, to decrease cART-related toxicities[2,3] and the development of potential (multi)drug resistance during long-term cART,[4] it is important to identify new targets for therapy and to develop novel anti-retroviral drugs (that is, low-molecular-weight inhibitors).[5] A general limitation to the development of anti-retroviral drugs that, like most cART regimens, commonly target virus-encoded enzymes is the high mutation rate of retroviruses;[1] these mutations frequently lead to the occurrence of drug-resistant strains.[6] However, a possibility to circumvent these problems is to address host cell components that are essential for virus replication and, because they are of cellular origin, are not subject to virus mutation. Within the HIV replication cycle, various host cell factors play an important role.[7,8] For example, eukaryotic initiation factor 5A (eIF-5A), a cellular protein that primarily promotes the elongation step of translation,[9] particularly during the biosynthesis of polyproline motifs,[10] has also been shown to act as a cellular cofactor of the HIV Rev protein.[11,12] Rev is an essential viral regulator that primarily mediates the nucleocytoplasmic transport and translation of incompletely spliced and unspliced viral transcripts.[13,14]

Activation of eIF-5A involves a unique spermidine-dependent post-translational modification of a specific lysine residue into the unusual amino acid hypusine (Figure 1). This modification is catalyzed by the sequential action of human deoxyhypusine synthase (DHS) and deoxyhypusine hydroxylase (DOHH).[15] Previously, derivatives (that is, polyamine analogues) of the natural DHS substrate spermidine were tested regarding their effect on DHS activity, which showed that *N*-1-guanyl-1,7-diaminoheptane (GC7) is a potent inhibitor.[16,17] However, the potential application of GC7 in vivo is limited, due to its unselective binding properties and high structural similarity to spermidine. This may result in potential undesired side effects, for example, in spermidine biosynthesis and metabolism.[18]

Importantly, other DHS inhibitors, including the spermidine analogue 1,8-diaminooctane, have been shown to significantly suppress virus replication by inhibiting Rev activity in a dose-dependent manner.[19,20] Thus, the targeting of DHS, for example, through the synthesis of improved GC7 derivatives obtained by rational drug design, may be a promising strategy to efficiently block HIV replication, including the replication of viruses that are otherwise resistant to current cART.

Based on the X-ray crystal data for DHS in the Protein Data Bank (PDB)[21] and the known inhibitor GC7, structure- and ligand-based drug design approaches were applied in order to discover novel DHS inhibitors. We report herein the in silico

[a] *Dr. M. Schroeder,[+] K. Pfaff, Prof. Dr. C. Meier*
Organic Chemistry, Department of Chemistry, Faculty of Sciences
University of Hamburg
Martin-Luther-King-Platz 6, 20146 Hamburg (Germany)
E-mail: chris.meier@chemie.uni-hamburg.de

[b] *A. Kolodzik,[+] Prof. Dr. M. Rarey*
Center for Bioinformatics, University of Hamburg
Bundesstraße 43, 20146 Hamburg (Germany)

[c] *Dr. P. Priyadarshini, Dr. M. Krepstakies, Prof. Dr. J. Hauber*
Heinrich Pette Institute, Leibniz Institute for Experimental Virology
Martinistraße 52, 20251 Hamburg (Germany)

[+] *These authors contributed equally to this work.*

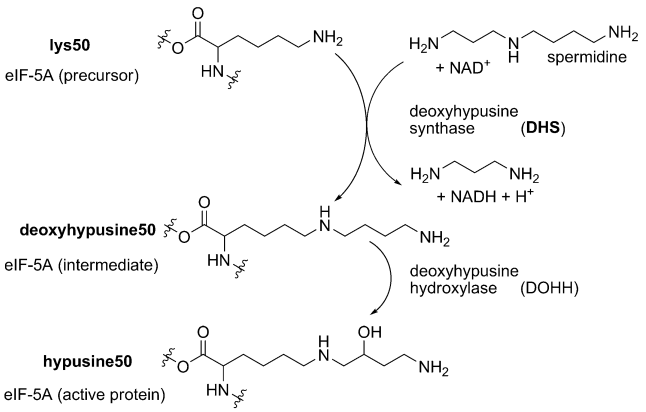


Figure 1. In vivo post-translational activation of eIF-5A catalyzed by DHS and DOHH.

design, synthesis, and biological evaluation of several selected compounds, which were tested for inhibition of DHS in an enzyme assay, for inhibition of HIV-1 replication in vitro, and for potential cytotoxic effects.

Results and Discussion

Virtual screening

To identify new binders to DHS, we combined ligand-based and structure-based drug design approaches. These include large-scale virtual screening (TrixX B-MI),[22] scaffold hopping and ligand decoration (ReCore),[23] combinatorial library design (Loft),[24] virtual screening (LeadIT),[25] and rescoring (HYDE).[26]

The two crystal structures 1RQD and 1ROZ represent the active form of DHS,[27] so these were selected for subsequent docking experiments. However, in these crystal structures, only protein dimers were found, whereas the biologically active enzyme complex is a tetramer.[28] Therefore, the tetrameric protein was built in accordance with the crystal structure parameters by using the Molecular Operating Environment (MOE) software.[29] The two additional amino acid chains were named "C" and "D", respectively (Figure 2). The binding site at the interface of chain A and

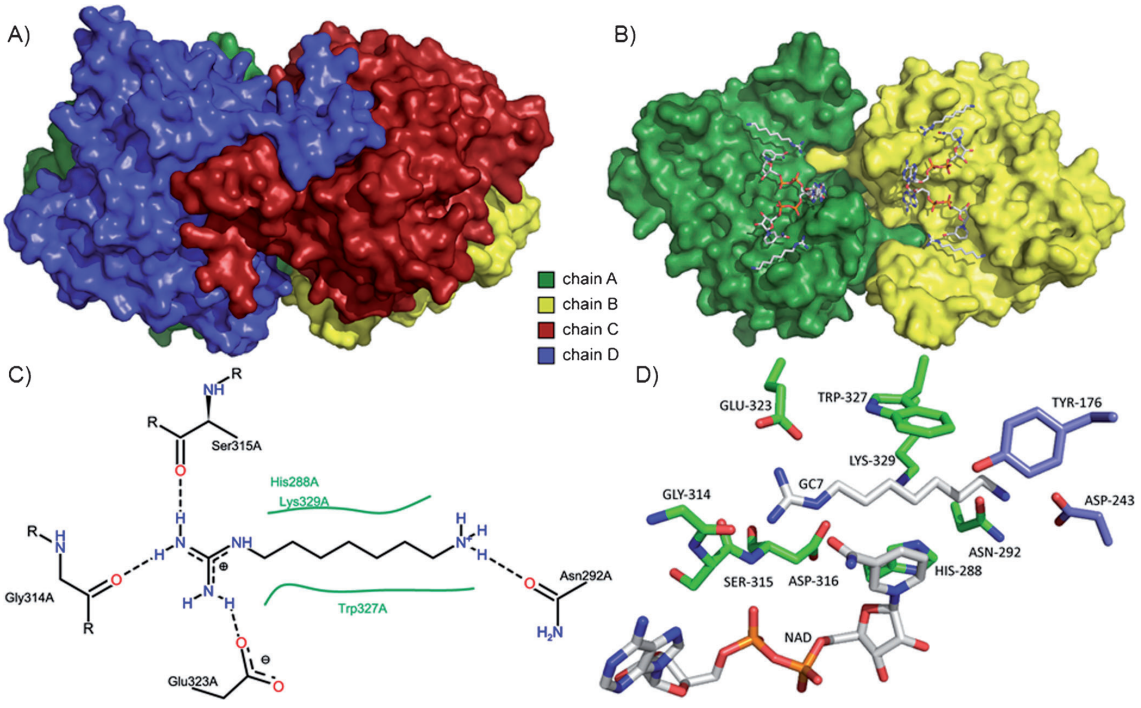


Figure 2. A) Tetrameric form of DHS generated from PDB crystal structure 1RQD. B) Chains A and B with cofactor NAD and inhibitor GC7 in all four binding sites. C) Interactions of GC7 with the DHS binding site. D) Three-dimensional view of the binding site.

chain D involving Asp243 of chain A and Glu323 of chain D was analyzed.

To define the binding site for subsequent docking experiments, the crystal structures (1RQD and 1ROZ) were aligned and superposed according to their α carbon (CA) atoms (root mean square deviation, RMSD = 0.179 Å). The coordinates of the co-crystallized ligand (GC7) and cofactor (nicotinamide adenine dinucleotide, NAD) of crystal structure 1RQD were then transferred into 1ROZ. Two alternative binding sites were defined. The first variant only included amino acids with at least one atom within a distance of less than 6.5 Å from GC7 or NAD. Both the cofactor NAD and the ligand GC7 were removed before docking. The second variant was defined accordingly around GC7 with the cofactor NAD present.

Consequently, the second variant of the binding site is relatively small compared with the first variant. Both variants of the binding site are used with their standard protonation and additional variants of His288. His288 is considered as being either protonated at the NE2 nitrogen (τ nitrogen) atom or rotated by 180° and protonated at the ND1 nitrogen (π nitrogen) atom, which possibly leads to an interaction with the carboxylate group of Asp316.

All ligands that were used for docking were first processed with the CORINA[30] software and subsequently with the NAOMI[31] program to ensure reasonable coordinates and valid valence bond forms. An overview of the development workflow is shown in Figure 3. In a first step, the clean-leads subset of the ZINC database[32] was screened for potential inhibitors of DHS. This dataset contains only lead-like compounds without reactive groups (for example, epoxides). In order to employ the TrixX BMI virtual screening approach, rotamers were generated by using the TrixX conformer generator.[33] The resulting rotamers were subsequently docked into the binding site of 1RQD. Solutions were only considered for further analysis if they had a predicted interaction to the OD2 oxygen atom of Asp243 in chain A and at least one additional interaction to the OE1 or OE2 oxygen atoms of Glu323 in chain D. Thereby, a similar interaction pattern to that of GC7 is ensured. Furthermore, proposed binders with a score above −10 (weak binders) were discarded to filter for medium-to-strong binders.

Based on experimental results, a focused fragment space was designed to match the characteristics of the DHS binding site. Molecules of this fragment space allow systematic testing of combinations of different anchor groups, cores, and linker lengths. Core fragments **c1**–**c20** (Figure 4) were chosen to build the center of generated molecules and to form an additional π–π stacking interaction to Trp327 in chain D of DHS, which cannot be established by the aliphatic compound GC7.

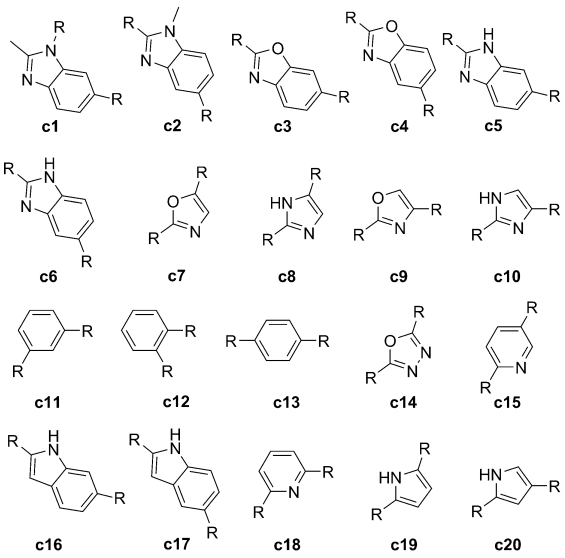


Figure 4. Core fragments designed to interact with Trp327 in chain D of DHS. The linkers (R) are virtually replaced by anchor fragments. Each combination of a single core fragment and two anchor fragments forms a molecule used for subsequent molecular docking calculations.

Anchor fragments **a1**–**a7** (Figure 5) were designed to interact with the hydrogen-bond acceptors of the binding site. All possible combinations of these fragments were enumerated. The resulting molecules were filtered for lead likeness.

Scaffold hopping and ligand decoration were performed with ReCore, by using the standard set of fragments supplied with the software. This set consists of fragments generated by fragmenting the drug-like compounds of the ZINC database according to the BRICS[34] shredding rules. This approach generates further potential lead structures with possibly higher binding affinities and additional interactions with the protein binding site.

In summary, we have followed three strategies. We used large-scale virtual screening of the ZINC database to identify commercially available compounds that can be purchased and tested for inhibition of DHS. This approach does not use any knowledge about already identified binders, so it has the potential to yield completely different inhibitors of DHS. Scaffold hopping was used to directly improve the binding affinity of lead structures and known binders. Only parts of the already known inhibitors are replaced by this

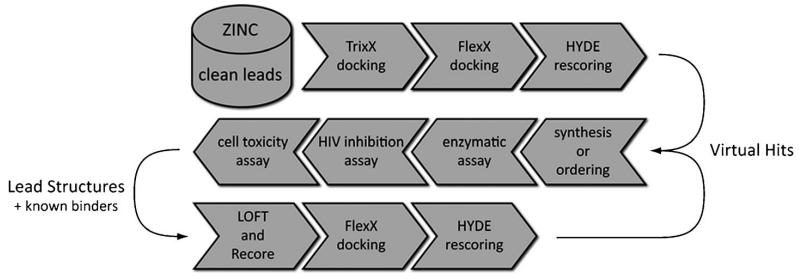


Figure 3. Workflow for the development of new potential DHS inhibitors.

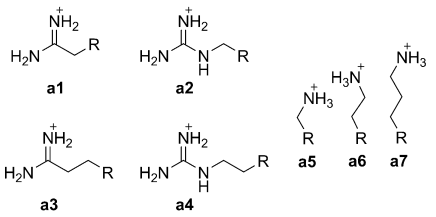


Figure 5. Anchor fragments designed to interact with hydrogen acceptors of the DHS binding site. The linkers (R) are virtually replaced by core fragments. Each combination of a single core fragment and two anchor fragments forms a molecule used for subsequent molecular docking calculations.

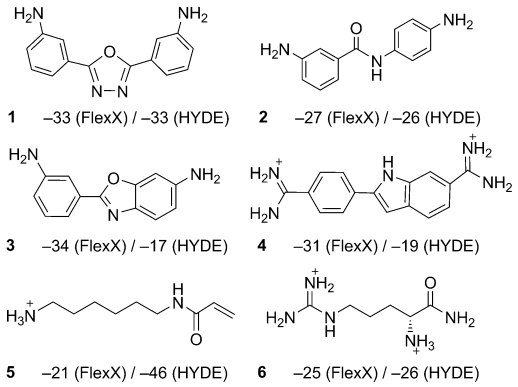


Figure 6. Compounds **1**–**6** were selected as potential DHS inhibitors on the basis of a TrixX BMI screening of the ZINC database.

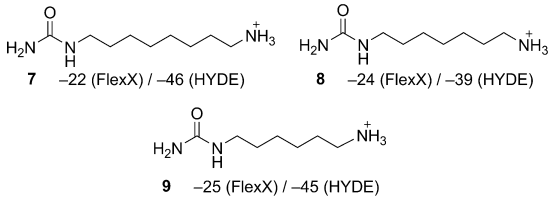


Figure 7. Suggested inhibitors designed on the basis of GC7 with the guanidinium group replaced by a urea group. The shown FlexX and HYDE scores represent the best score of all performed docking experiments for each compound.

approach, so the resulting molecules have a high probability of showing inhibitory activity against DHS. However, molecules designed by this approach, which are not commercially available, have to be individually synthesized. A fragment space was used to systematically investigate a high number of similar molecular structures. Thus, only the most promising representative of a class of similar compounds has to be synthesized and tested in vitro.

All of the molecules selected by the above-mentioned approaches were docked into the binding site of DHS by using the LeadIT docking software (version 2.0.2) based on the FlexX docking approach.[35] The maximum overlap volume of ligand and protein was set to 2 Å^3 and the clash factor was set to 0.7 to limit false-positive docking solutions. All other parameters were kept at the default settings. Furthermore, the same interaction filter was applied as that used for the TrixX BMI virtual screening.

Rescoring was performed with the HYDE scoring function (version 3.25).[26] This scoring function models desolvation effects, so it has a higher accuracy in predicting binding affinities than the standard scoring functions implemented within TrixX or LeadIT, respectively. For every ligand docked with LeadIT, the ten best poses were stochastically optimized and rescored with the HYDE software. All poses with a negative HYDE score were manually inspected for incorrect conformations or wrong protonation states.

Docking studies

Pooled molecules from the initial TrixX BMI screening and structures from scaffold hopping and ligand decoration approaches were docked into the binding site of DHS. The most promising compounds as judged by the FlexX and HYDE scores were further analyzed for synthetic accessibility or commercial availability. Compounds **4** and **6** were commercially available, whereas compounds **1**–**3** and **5** were selected for synthesis (Figure 6).

In addition, substitution of the charged guanidinium group by a urea group should improve the abilities of the molecules to pass through cell membranes relative to the ability of GC7 and resulted in high (negative) FlexX and HYDE scores (Figure 7).

Out of the compounds that are shown in Figures 6 and 7, only compound **4** inhibited DHS in an enzymatic assay and the replication of HIV in cell cultures (data not shown). Compound **4** therefore served as a basis for the following inhibitor design. To systematically enhance the inhibitory activity and to decrease the observed cell toxicity of compound **4**, a fragment space was designed as described above. The resulting molecules address the hydrogen-bond acceptors of the terminal parts of the binding site, as well as the aromatic ring of Trp327 in the center of the binding site. Furthermore, the generated molecules were sufficiently flexible to adapt to the binding site of DHS and to avoid intercalation into DNA. They shared predicted interactions with Trp327, Glu323, and Asp243 of DHS. Out of these molecules, compound **10** was selected as a target compound because it had the highest predicted binding affinity (Figure 8). The predicted interactions of compound **10** with

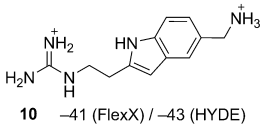


Figure 8. Selected target molecule from a second set of predicted binders generated from focused fragment space with the corresponding FlexX and HYDE scores.

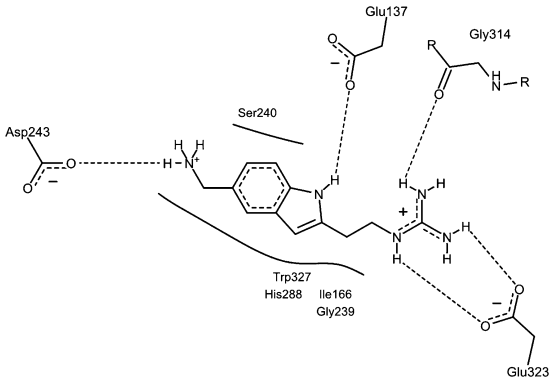


Figure 9. Predicted binding mode of compound **10** to DHS as calculated by the LeadIT 2.0.2 software.

the active site of DHS are shown in Figure 9. Compound **10** has not been described before in the literature.

Chemical synthesis of the in silico designed potential DHS binders

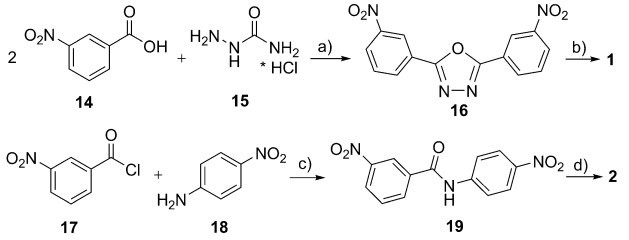
The compounds can be subdivided into three groups: the more rigid and more GC7-unlike compounds **1**–**6**, the flexible and more GC7-like analogues **7**–**9**, and compound **10**, which was derived from the focused fragment space (Figures 6, 7, and 8). It was possible to prepare the proposed compounds from the first two groups in a maximum of two steps by starting from cheap reagents and commercially available starting materials. For the synthesis of the more rigid aromatic compounds **1**–**3**, the synthesis routes are given in Schemes 1 and 2. To avoid additional protection and deprotection steps, the nitro group was chosen as a masked amino group precursor in all cases. The corresponding nitro compounds **16** (for **1**), **19** (for **2**), and **21** (for **3**) were synthesized in yields of 16–91%, despite the fact that the electron-withdrawing nitro group lowers the reactivity of the aromatic system. The substituted 1,3,4-oxadiazole compound **16** was synthesized by cyclization of 3-nitrobenzoic acid with aminourea hydrochloride in polyphosphoric acid in 26% yield.[36] Heating of 3-nitrobenzoyl chloride and 4-nitroaniline in pyridine to reflux, as reported by Hu et al., led to 3-nitro-*N*-(4-nitrophenyl)-benzamide (**19**; in 91% yield).[37]

6-Nitro-2-(3-nitrophenyl)benzoxazole (**21**) was prepared first by applying a copper-catalyzed method reported by Ueda and Nagasawa that led from the benzanilide **19** by way of an oxidative ring closure to the benzoxazole **21** in 27% yield.[38] As shown in Scheme 2, the alternative route described by Hausner et al. allowed the synthesis to be shortened by one step.[39] In this case, benzoxazole **21** was obtained from 2-amino-4-nitrophenol (**20**) and 3-nitrobenzoic acid (**14**) in a nonoptimized yield of 16%.

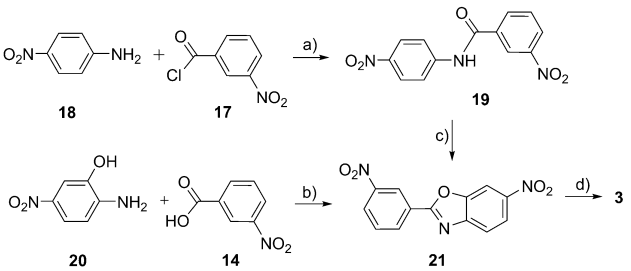
For the last step, different reduction methods were tried to convert the nitro group into the amino group. Among the various procedures, the reduction with tin(II) chloride (Scheme 1) gave the best results for **16** and **19**. However, in case of **21**, the reduction with tin(II) chloride led to a ring opening of the oxazole moiety. Therefore, in this case, hydrogen under pressure and palladium on charcoal were used (Scheme 2).

Beside these rigid aromatic systems, three flexible aliphatic derivatives of GC7 were synthesized (Figure 7 and Scheme 3), in which the guanidine moiety was replaced by a urea function and the alkyl chain length varied from six to eight carbon atoms. Compound **5**, in which the guanidine moiety was substituted with an acryloyl moiety and the alkyl chain length was shortened to six carbon atoms, was also synthesized.

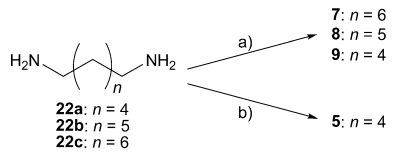
A method reported by Miyagawa et al. led to target compound **5** in 35% yield by adding acryloyl chloride dropwise to a diluted solution of 1,6-diaminohexane in dry methanol.[40] In solvents like pyridine, CH_2Cl_2, or THF with triethylamine as a base, the only isolated product was the disubstituted diamine.[41–43] Compounds **7**–**9** were prepared according to the method of Boden et al. in yields varying from 11 to 39%.[44] The synthesis included washing steps for purification, so parts of the product were also dissolved, which resulted in the poor yields for **7**–**9**.



Scheme 1. Synthesis of the more rigid aromatic compounds **1** and **2** based on the virtual screening results. *Reagents and conditions*: a) P_2O_5, phosphoric acid, reflux, 2 h, 26%; b) $SnCl_2$, conc. HCl, 50 °C, 2 h, 84%; c) pyridine, reflux, 18 h, [N_2], 91%; d) $SnCl_2$, conc. HCl, 50 °C, 2 h, 99%.



Scheme 2. Synthesis of 6-amino-2-(3-aminophenyl)benzoxazole (**3**). *Reagents and conditions*: a) Pyridine, reflux, 18 h, [N_2], 91%; b) $B(OH)_3$, xylene, reflux, 12 h, 16%; c) *o*-dichlorobenzene, copper triflate, reflux, 22 h, 27%; d) Pd/C, H_2, EtOH, RT, 68 h, 25%.

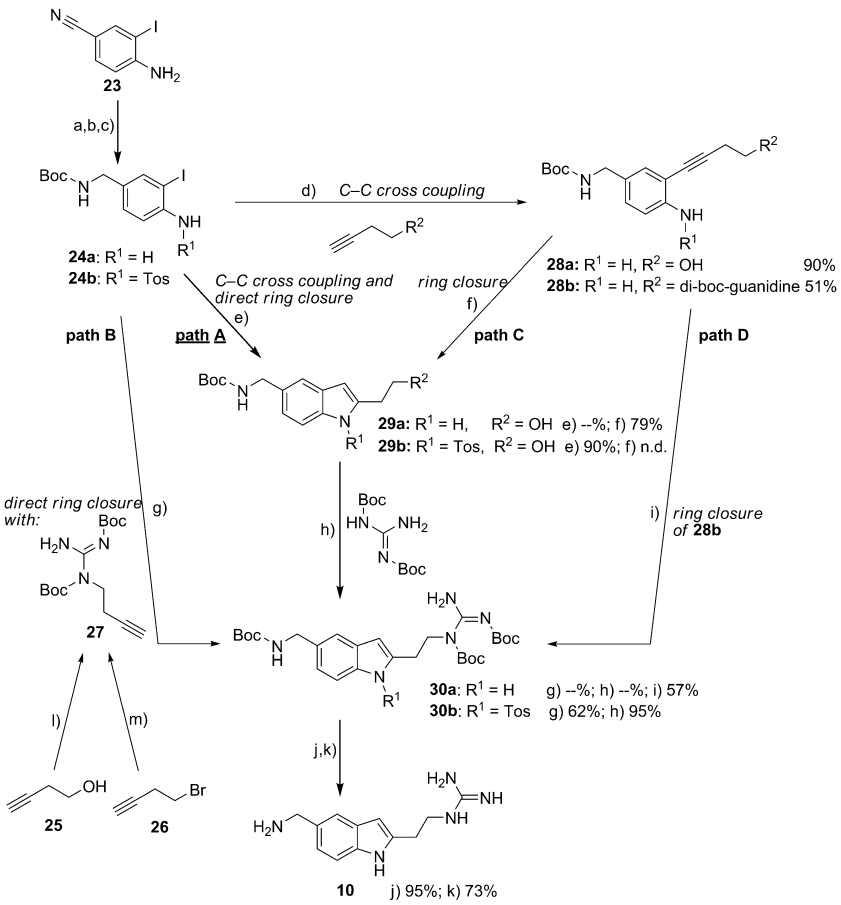


Scheme 3. Overview of the synthesized GC7 analogues **5** and **7**–**9**. *Reagents and conditions*: a) Trimethylsilylisocyanate (0.35 equiv), THF, RT, 4 h, [N_2], 11–39%; b) acryloyl chloride (0.4 equiv), CH_3OH, 0 °C→RT, 16 h, [N_2], 35%.

For compound **10**, the substituted indole moiety is the key structure element. Due to the particular substitution pattern, no suitable indole precursor was commercially available. Thus, the synthesis was started with 4-amino-3-iodobenzonitrile (**23**), which bears a nitrile function as a precursor for the benzyl amine group and an iodine atom in the *ortho* position to the amine (Scheme 4). Different C–C cross-coupling and ring-closure methods with 4-substituted but-1-ynes were tested to yield the indole (paths A–D). Initially, the nitrile group was reduced with BH_3·THF, and this was followed by acidic workup according to the method of Li et al.[45] The following Boc protection was carried out with di-*tert*-butyldicarbonate in 72–80% yield. The in situ coupling and cyclization of **24a** with but-3-yn-1-ol (**25**) was only successful by using the reaction conditions described by Larock et al. (Scheme 4, path A).[46] However, even then, only an inseparable mixture of the 2- and 3-substituted indoles was isolated in a yield of 35%.

Thus, the reactivity of the aromatic amine moiety of **24a** was enhanced by *N*-tosylation with tosyl chloride to give precursor **24b** in 92%.[47,48] A copper iodide/$Pd(PPh_3)_2Cl_2$ mediated reaction with three equivalents of but-3-yn-1-ol yielded the indole **29b** in 90% (Scheme 4, path A).[44] In analogy to known procedures, the hydroxy group of the indole **29b** was converted into the *N*,*N*′-bis-*tert*-butoxycarbonylguanidino group by using the Mitsunobu reaction in 95% yield.[49] The cleavage of the tosyl group gave the best results by using 5 M NaOH in methanol and the product was isolated in 95% yield. The final deprotection step was the acidic cleavage of all of the Boc groups of **30a** by treatment with 2 M HCl in CH_3CN. Purification on reverse-phase (RP) silica gel with H_2O as the eluent and freeze drying gave the pure compound **10** as the hydrochloride salt in 73% yield.

It is noteworthy that a protecting group on the hydroxy function of but-3-yn-1-ol (**25**) was not needed for the cyclization. It actually caused a negative effect, because no product



Scheme 4. Overview of the synthesis of the in silico designed DHS inhibitor **10**. *Reagents and conditions*: a) BH_3·THF (3 equiv), THF, reflux, 4 h, [N_2], then 1 M HCl (3 equiv), reflux, 1 h; b) 1. Et_3N (4 equiv), DMAP (0.2 equiv), 2. di-*tert*-butyldicarbonate (1–7.3 equiv), 0 °C→RT, 5 h, [N_2], 72–80% (over two steps); c) pyridine (3 equiv), TosCl (1.2 equiv), CH_2Cl_2, RT, 42 h, [N_2], 92%; d) DBU (1.5 equiv), Pd EnCat TPP30 (3.5 mol%), alkyne (1.5 equiv), reflux, 17 h, [N_2]; e) alkyne (3 equiv), CuI (0.2 equiv), $Pd(PPh_3)_2Cl_2$ (10 mol%), NEt_3, DMF, 85 °C, 17 h, [N_2]; f) KH (1.2 equiv), $Cu(OAc)_2$ (1.1 equiv), 1,2-dichloroethane, 70 °C, 3–4 d, [N_2], 79%; g) see (e); h) *N*,*N*′-bis-*tert*-butoxycarbonylguanidine (1.5 equiv), PPh_3 (1.5 equiv), DIAD (1.5 equiv), 0 °C→reflux, 3 h, [N_2]; i) see (f); j) 5 M NaOH/CH_3OH, RT, 40 h, [N_2]; k) 2 M HCl/CH_3CN, RT, 28 h; l) see (h), 87%; m) *N*,*N*′-bis-*tert*-butoxycarbonylguanidine (1 equiv), NaH (1 equiv), DMF, [N_2], 10%. Boc: *tert*-butoxycarbonyl; Tos: toluene-4-sulfonyl; n.d.: not determined; DMAP: 4-dimethylaminopyridine; DBU: 1,8-diazabicyclo[5.4.0]undec-7-ene; DMF: *N*,*N*-dimethylformamide; DIAD: diisopropylazodicarboxylate.

was isolated with the methodology of Larock et al. Moreover, for 4-amino-3-(4-*O*-*tert*-butyldimethylsilyl-but-1-yn-1-yl)benzonitrile under the cyclization conditions of Hiroya et al., the 5-cyano-2-vinyl-1*H*-indole was isolated as the main product in 90%.[50] To avoid additional protection steps, the approach of the nitrile reduction after indole synthesis was tested, too. Unfortunately, this was not successful and led to reduction of the indole as a side reaction.[51–53]

The synthesis of *N*,*N*′-bis-*tert*-butoxycarbonylguanidinobut-3-yne (**27**) prior to cross-coupling and cyclization with **24a**/**b** would save a step in the linear synthesis route (Scheme 4, path B), so alkyne **27** was synthesized from but-3-yn-1-ol (**25**) under Mitsunobu reaction conditions[49] (87%) and from 4-bromo-1-butyne (**26**) by nucleophilic substitution[54] (10%). A ring-closure reaction of 4-amino-3-(4-bromobut-1-yn-1-yl)benzonitrile led only to the elimination product 5-cyano-2-vinyl-1*H*-indole in poor yields (14%). Due to the lower yields obtained with *N*,*N*′-bis-*tert*-butoxycarbonylguanidinobut-3-yne (**27**) in the transition-metal-catalyzed cyclizations and the high efficiency of the Mitsunobu reaction conditions, it is more convenient to convert the hydroxy group after the indole synthesis. However, the Mitsunobu reaction with indole **29a** was not successful, possibly due to side reactions of the free indole nitrogen atom (Scheme 4).

To circumvent the tosylation/detosylation but still use the modified "one-pot" indole synthesis from Adachi et al.,[47] we aimed to convert compound **23** directly to the *O*-*tert*-butyl-4-*N*-(*O*-*tert*-butoxycarbonyl)-3-iodobenzylcarbamate. However, even if a large excess of di-*tert*-butyldicarbonate (7.3 equiv) was used, the *bis*-Boc protection was not achieved.

Interestingly, by using a combination of Sonogashira cross-coupling reaction conditions[55,56] and subsequent cyclization according to the method of Hiroya et al., indole **30a** was obtained without the need for a protecting group on the arylamine nitrogen atom (Scheme 4, path D).[50] Although this route was shorter, the yields were significantly lower than those of the first route (path A).

In summary, the most efficient synthetic route yielded, through reduction, Boc protection, tosylation, "one-pot" cyclization with but-3-yn-1-ol, guanidine introduction under Mitsunobu reaction conditions, detosylation, and cleavage of the Boc protecting groups, the target compound **10** in an overall yield of 44% over seven steps (Scheme 4, path A). One advantage of this route is the variability of the starting materials because different hydroxy-substituted alkynes similar to **25** may be suitable and the aromatic precursor (such as **24b**) can also be varied, for example, with different electron-donating substituents and various substitution patterns. Thus, the established synthetic route should enable the preparation of derivatives of compound **10** with, for example, varied alkyl chain lengths.

Inhibition of DHS activity in vitro

In order to investigate the ability of the new compounds **1**–**10** to inhibit DHS, all of the compounds were tested in an enzymatic in vitro DHS assay. The DHS reaction was simulated by employing recombinantly expressed eIF-5A and DHS, together with the substrates NAD and ^{3}H-labeled spermidine. The reaction was performed with the indicated concentrations of derivatives **1**–**10**, GC7 as a reference control, or only dimethylsulfoxide (DMSO) as a negative control, respectively. As a measure for DHS activity, the relative amounts of the tritium-labeled aminobutyl moiety transferred from spermidine to eIF-5A were detected.

Unfortunately, there was no significant inhibition of DHS detected for the in silico designed compounds **1**–**9** of the first two groups (Figures 6 and 7). Compounds **1**–**3** and **6**–**9** showed no inhibition, whereas compounds **4** and **5** showed weak inhibition (11% and 14%, respectively). GC7 showed a 45% inhibition under the assay conditions (80 μM GC7 or the corresponding compound; negative control subtracted from the absolute inhibition). In contrast, the novel designed potential binder **10** (Figure 8) showed significant dose-dependent inhibition of DHS ($IC_{50} \approx 12$ μM; Figure 10a,b).

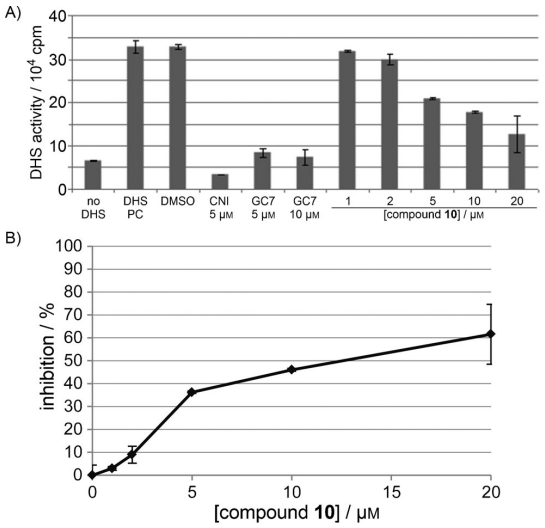


Figure 10. A) DHS inhibition assay with compound **10**. B) Dose-dependent inhibition of DHS by the novel inhibitor **10**.

Inhibition of HIV-1 replication in vitro

It has been previously demonstrated that the eIF-5A-modifying enzyme DHS may serve as a novel target for anti-retroviral therapy.[19,20] Therefore, the potential inhibitory effect on the replication of HIV-1 in tissue cultures was analyzed. For this purpose, PM1 lymphocytes were incubated for seven days, either in the presence of compound **10**, the established DHS inhibitor GC7 (positive control), or culture medium alone (negative control). Subsequently, the respective cultures were infected with HIV-1_{BaL} and culturing was continued for another week, at which time the p24 antigen levels in the culture supernatants were determined. Both compounds moderately inhibited the formation of HIV-1 progeny in a dose-dependent

manner (Figure 11). Clearly, GC7 was about twice as active as the in silico designed compound **10**. Compound **10** at a concentration of 2 μM achieved a moderate HIV-1 inhibition rate of 14%. Cell viability testing (by alamarBlue Assay) in the respective cultures failed to detect drug-induced cytotoxicity (data not shown).

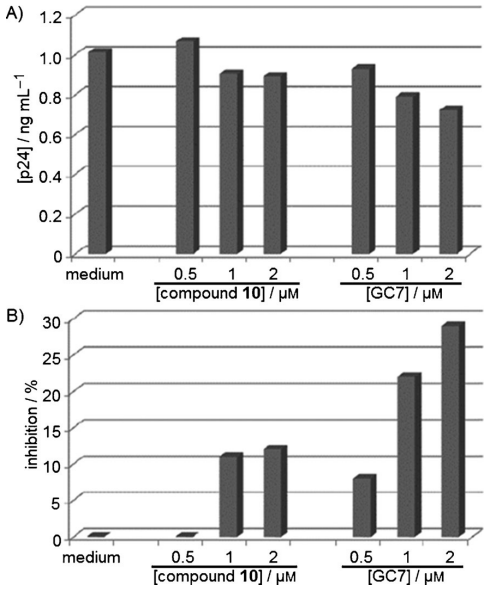


Figure 11. Antiviral effects of DHS inhibitors. PM1 cell cultures were incubated in the presence of the indicated concentrations of compound **10**, GC7, or in medium alone for seven days, infected with the CCR5-tropic HIV-1 isolate BaL, and further cultivated for another week. A) Release of viral particles was determined by an HIV-1 p24 antigen-specific enzyme-linked immunosorbent assay at day seven post-infection. B) The percentage of inhibition of virus replication relative to the replication in the control culture without drugs (medium), which was arbitrarily set at 100%.

Conclusions

In the presented approach, we have described a way of designing a new small-molecule inhibitor of deoxyhypusine synthase. A compound was successfully designed, synthesized, and validated to show dose-dependent inhibition of DHS in vitro and suppression of HIV replication in cell cultures. Thereby, we have demonstrated that the in silico design of DHS inhibitors may serve as a starting point to develop new drugs against diseases, such as HIV infection, in which the hypusine-containing protein eIF-5A plays a critical role. Although the observed antiviral activity for compound **10** is not very high, the data found here can be used now for a further development of the concept. The synthesis pathways described for compound **10** are flexible enough to prepare more structurally diverse molecules. Variations of compound **10** are currently being synthesized and will soon be analyzed with respect to their inhibition of DHS activity and HIV replication. Moreover, compound **10** is being used as a hit structure for further virtual screening optimization and also co-crystallization with DHS.

Experimental Section

Chemistry

All air- or water-sensitive reactions were performed in flame-dried glassware under a nitrogen atmosphere. Compounds **4** and **6** were obtained from Sigma–Aldrich. Commercially available solvents and reagents were used without further purification with the following exceptions: dichloromethane (CH_2Cl_2) was distilled from calcium hydride and stored over activated molecular sieves. THF was dried over potassium/benzophenone, distilled under nitrogen, and stored over activated molecular sieves. Petroleum ether 50–70, EtOAc, CH_2Cl_2, and CH_3OH employed in chromatography were distilled before use. For column chromatography, silica gel 60 (230–400 mesh) was used. For thin layer chromatography (TLC), precoated aluminum 60 F_{254} plates with a 0.2 mm layer of silica gel containing a fluorescence indicator were used. NMR spectra were recorded on 400 or 500 MHz spectrometers (Bruker AMX 400, Bruker AV 400, or Bruker DRX 500). All ^1H and ^{13}C NMR chemical shifts (δ) are quoted in parts per million (ppm) and were calibrated against solvent signals. High-resolution (HR) mass spectra were obtained with a VG Analytical VG/70–250F spectrometer (FAB; matrix was *m*-nitrobenzyl alcohol). HR ESI mass spectra were obtained with an Agilent Technologies ESI-TOF 6224 spectrometer. IR spectra were acquired by using a Bruker Alpha-P IR spectrometer within the 400–4000 cm^{-1} range in the attenuated total reflection (ATR) mode. Analytical HPLC was carried out on a VWR-Hitachi LaChromElite HPLC system, which consisted of a VWR-Hitachi L-2130 pump, an L-2200 autosampler, and an L-2455diode array detector. The column used was a Nucleodur C18 Gravity, 5 μm (Macherey–Nagel). Elution was performed with a 100 mM ammonium formiate buffer (pH 3.0)/acetonitrile (Sigma–Aldrich, HPLC grade) eluent, 5–44% CH_3CN (0–12 min), a flow rate of 1.3 mL min^{-1}, and UV detection at 264 nm.

General procedure 1: **Mitsunobu introduction of the guanidine**: The reaction was carried out under a nitrogen atmosphere. A solution of the corresponding alcohol, PPh_3 (1.5 equiv), and *N*,*N*′-bis-*tert*-butoxycarbonylguanidine (1.5 equiv) in dry THF (15 mL $mmol^{-1}$ alcohol) was cooled down to 0 °C. This was followed by dropwise addition of DIAD (1.5 equiv). After that, the reaction mixture was heated at reflux for 3 h. The solvent was removed in vacuo and the crude product was purified by column chromatography on silica gel (petroleum ether 50–70/EtOAc, 4:1→1:1).

General procedure 2: **Synthesis of the indole**: The reaction was carried out under a nitrogen atmosphere. Dry triethylamine (30 equiv) was added to a suspension of *O*-*tert*-butyl-3-iodo-4-(4-methylphenylsulfonylamido)benzylcarbamate (**24b**), CuI (0.2 equiv), and $Pd(PPh_3)_2Cl_2$ (0.1 equiv) in dry DMF (10–15 mL $mmol^{-1}$). This was followed by dropwise addition of the substituted alkyne (3 equiv). After being stirred for 17 h at 85 °C, the suspension was diluted with ethyl acetate and washed with H_2O (3×). The organic layer was dried over sodium sulfate and the solvent was removed under reduced pressure. The crude product was then purified by column chromatography on silica gel.

2,5-Di-(3-nitrophenyl)-1,3,4-oxadiazole (16): For the polyphosphoric acid solution, P_2O_5 (15 g) was added to phosphoric acid (10 mL). 3-Nitrobenzoic acid (**14**; 1.00 g, 5.99 mmol) and aminourea hydrochloride (**15**; 0.67 g, 5.99 mmol) were added to this solution. After being stirred at 170 °C for 2 h, the reaction mixture was

poured into ice/H_2O. The precipitated product was filtered and washed with H_2O to yield a colorless solid (482 mg, 1.55 mmol, 26%). ^{1}H NMR (400 MHz, $CDCl_3$): δ = 9.00 (dd, J = 1.7, 1.7 Hz, 2H; H2), 8.64 (dt, J = 7.9, 1.4 Hz, 2H; H6), 8.51 (ddd, J = 8.2, 1.0, 1.0 Hz, 2H; H4), 7.96 ppm (dd, J = 8.0, 8.0 Hz, 2H; H5); ^{13}C NMR (101 MHz, $CDCl_3$): δ = 163.6 (oxadiazole C), 148.9 (aryl CNO_2), 132.8 (aryl C6), 130.8 (aryl C5), 126.8 (aryl C4), 125.2 (C1), 122.2 ppm (aryl C2); TLC: R_f = 0.82 (CH_2Cl_2/CH_3OH, 9:1); IR (neat): $\tilde{\nu}$ = 3093, 1517, 1350, 1062, 912, 713, 667 cm^{-1}; HRMS (FAB): m/z calcd: 313.0572 $[M+H]^+$; found: 313.0576.

3-Nitro-*N*-(4-nitrophenyl)benzamide (19): The reaction was carried out under a nitrogen atmosphere. 4-Nitroaniline (**18**; 1.00 g, 7.24 mmol) was dissolved in dry pyridine (24 mL) and then 3-nitrobenzoyl chloride (**17**; 1.61 g, 8.69 mmol, 1.2 equiv) was added. After that, the reaction mixture was heated at reflux for 18 h and then poured into ice/H_2O. The precipitated product was filtered and washed with H_2O to yield a yellow solid (1.90 g, 6.61 mmol, 91%). ^{1}H NMR (400 MHz, [D$_6$]DMSO): δ = 11.10 (s, 1H; amide NH), 8.82 (dd, J = 1.8, 1.8 Hz, 1H; H3), 8.48 (ddd, J = 8.2, 1.5, 1.5 Hz, 1H; H7), 8.43 (ddd, J = 8.0, 1.2, 1.2 Hz, 1H; H5), 8.30 (d, J = 9.1 Hz, 2H; H10), 8.07 (d, J = 9.2 Hz, 2H; H9), 7.88 ppm (dd, J = 8.1, 8.1 Hz, 1H; H6); ^{13}C NMR (101 MHz, [D$_6$]DMSO): δ = 164.1 (carbonyl C), 147.6 (C4), 145.0 (C8), 142.6 (C11), 135.6 (C2), 134.4 (C5), 130.3 (C6), 126.6 (C7), 124.8 (C10), 122.6 (C3), 120.2 ppm (C9); TLC: R_f = 0.85 (CH_2Cl_2/CH_3OH, 9:1); IR (neat): $\tilde{\nu}$ = 3367, 3082, 1686, 1525, 1347, 1299, 1109, 848, 704 cm^{-1}; HRMS (FAB): m/z calcd: 288.0620 $[M+H]^+$; found: 288.0610.

6-Nitro-2-(3-nitrophenyl)benzoxazole (21): Method a: 2-Amino-5-nitrophenol (**20**; 771 mg, 5.00 mmol) and 3-nitrobenzoic acid (**14**; 836 mg, 5.00 mmol) were suspended in xylene (15 mL). After addition of boronic acid (340 mg, 5.50 mmol, 1.1 equiv), the reaction mixture was heated at reflux for 8 h and then the solvent was removed in vacuo. The crude product was purified by column chromatography on silica gel (CH_2Cl_2/CH_3OH, 99:1) to yield a pale-rose solid (229 mg, 0.803 mmol, 16%). ^{1}H NMR (400 MHz, $CDCl_3$): δ = 9.15 (dd, J = 1.8, 1.8 Hz, 1H; H9), 8.63 (ddd, J = 7.7, 1.6, 1.1 Hz, 1H; H13), 8.57 (dd, 1H, J = 2.3, 0.4 Hz; H6), 8.48 (ddd, J = 8.3, 2.3, 1.1 Hz, 1H; H11), 8.39 (dd, J = 8.8, 2.2 Hz, 1H; H4), 7.93 (dd, J = 8.8, 0.3 Hz, 1H; H3), 7.81 ppm (t, J = 8.0 Hz, 1H; H12); ^{13}C NMR (101 MHz, $CDCl_3$): δ = 149.7 (C7), 148.5 (C2), 146.4 (C1), 145.5 (C10), 133.5 (C13), 130.5 (C12), 127.0 (C11), 123.2 (C8), 122.7 (C3), 122.2 (C9), 121.5 (C5), 121.2 (C4), 107.6 ppm (C6); TLC: R_f = 0.41 (petroleum ether 50–70/EtOAc, 4:1); IR (neat): $\tilde{\nu}$ = 3104, 2922, 2852, 1521, 1345, 1061, 815, 733, 707 cm^{-1}; HRMS (FAB): m/z calcd: 286.0386 $[M+H]^+$; found: 286.0463.

Method b: In a round-bottomed flask, 3-nitro-*N*-(4-nitrophenyl)benzamide (**19**; 104 mg, 362 μmol) was dissolved in *o*-dichlorobenzene (1.2 mL). This was followed by addition of copper triflate (30 mg, 83 μmol, 0.2 equiv). After that, the reaction mixture was heated at reflux for 22 h and then the solvent was removed in vacuo. The crude product was purified by column chromatography on silica gel (petroleum ether 50–70/EtOAc, 6:1→1:1) to yield a pale-rose solid (28 mg, 99 μmol, 27%). The analytical data were identical to those reported above.

6-Amino-2-(3-aminophenyl)benzoxazole (3): The reaction was carried out under a hydrogen atmosphere. 6-Nitro-2-(3-nitrophenyl)benzoxazole (**21**; 75 mg, 33 μmol) and Pd/C (10 mg) were suspended in dry ethanol (8 mL). This was followed by activation with ultrasound for 30 s. The reaction mixture was then stirred at room temperature under hydrogen pressure for 68 h, before being filtered and extracted with methanol. The crude product was purified by column chromatography on silica gel (CH_2Cl_2/CH_3OH, 97:3) to yield a colorless solid (18 mg, 80 μmol, 25%). ^{1}H NMR (400 MHz, $CDCl_3$): δ = 7.37 (d, J = 8.2 Hz, 1H; H3), 7.31 (s, 1H; H9), 7.24–7.14 (m, 2H; H12, H13), 6.79 (s, 1H; H6), 6.70 (d, J = 7.4 Hz, 1H; H11), 6.62 (d, J = 8.4 Hz, 1H; H4), 5.38 ppm (2×s, 4H; 2×NH_2); ^{13}C NMR (101 MHz, $CDCl_3$): δ = 159.8 (C1), 151.7 (C2), 149.2 (C8), 147.8 (C7), 131.9 (C5), 129.5 (C12), 127.6 (C10), 119.5 (C3), 116.2 (C11), 113.9 (C13), 112.3 (C4), 111.3 (C9), 94.2 ppm (C6); TLC: R_f = 0.47 (CH_2Cl_2/CH_3OH, 9:1); IR (neat): $\tilde{\nu}$ = 3413, 3314, 3205, 1628, 1489, 1354, 1146, 811, 717, 622 cm^{-1}; HRMS (ESI$^+$): m/z calcd: 226.0902 $[M+H]^+$; found: 226.0974.

2,5-Di-(3-aminophenyl)-1,3,4-oxadiazole (1): In a 25 mL round-bottomed flask, tin(II) chloride (608 mg, 3.20 mmol, 10 equiv) was dissolved in concentrated HCl (8 mL) and warmed up to 50 °C. This was followed by slow addition of 2,5-di-(3-nitrophenyl)-1,3,4-oxadiazole (**16**; 100 mg, 0.320 mmol). Afterward, the solution was stirred for a further 1.5 h and then poured, under gas formation, into a mixture of potassium carbonate (10 g) and ice/H_2O (100 mL). The product was extracted with EtOAc (3×) and the combined organic layers were dried over sodium sulfate. Finally, the solvent was removed in vacuo to yield a pale-yellow solid (68 mg, 0.27 mmol, 84%). ^{1}H NMR (400 MHz, [D$_6$]DMSO): δ = 7.29 (dd, J = 1.8, 1.8 Hz, 2H; H2), 7.26–7.21 (m, 2H; H5), 7.21–7.18 (m, 2H; H6), 6.79 (ddd, J = 7.8, 2.3, 1.3 Hz, 2H; H4), 5.52 ppm (s, 4H; NH_2); ^{13}C NMR (101 MHz, [D$_6$]DMSO): δ = 164.2 (oxadiazole C), 149.4 (C1), 129.8 (C5), 123.5 (C3), 117.2 (C4), 113.7 (C6), 111.0 ppm (C2); TLC: R_f = 0.61 (CH_2Cl_2/CH_3OH, 9:1); IR (neat): $\tilde{\nu}$ = 3207, 1592, 1468, 1317, 781, 678 cm^{-1}; HRMS (FAB): m/z calcd: 253.1083 $[M+H]^+$; found: 253.1089.

3-Amino-*N*-(4-aminophenyl)benzamide (2): In a 25 mL round-bottomed flask, tin(II) chloride (727 mg, 3.83 mmol, 11 equiv) was dissolved in concentrated HCl (7 mL) and warmed up to 50 °C. This was followed by slow addition of 3-nitro-*N*-(4-nitrophenyl)benzamide (**19**; 100 mg, 0.348 mmol). Next, the solution was stirred for a further 2 h and then poured, under gas formation, into a mixture of potassium carbonate (8.25 g) and ice/H_2O (100 mL). The product was extracted with EtOAc (3×) and the combined organic layers were dried over sodium sulfate. Finally, the solvent was removed in vacuo to yield a colorless solid (78 mg, 0.345 mmol, 99%). ^{1}H NMR (400 MHz, [D$_6$]DMSO): δ = 9.64 (s, 1H; amide NH), 7.35 (dd, J = 6.8, 1.9 Hz, 2H; H9), 7.12–7.00 (m, 3H; H2, H5, H6), 6.70 (ddd, J = 7.9, 2.3, 1.0 Hz, 1H; H4), 6.52 (dd, J = 6.7, 2.1 Hz, 2H; H10), 5.23 (s, 2H; NH_2), 4.87 ppm (s, 2H; NH_2); ^{13}C NMR (101 MHz, [D$_6$]DMSO): δ = 165.3 (amide C), 148.4 (C3), 144.6 (C11), 136.1 (C1), 128.4 (C5), 128.2 (C8), 122.0 (C9), 116.1 (C4), 114.3 (C6), 113.4 (C10), 112.7 ppm (C2); TLC: R_f = 0.52 (CH_2Cl_2/CH_3OH, 9:1); IR (neat): $\tilde{\nu}$ = 3325, 3218, 1597, 1583, 1511, 1320, 1246, 817, 505 cm^{-1}; HRMS (FAB): m/z calcd: 228.1136 $[M+H]^+$; found: 228.1132.

(6-Aminohexyl)urea (9): The reaction was carried out under a nitrogen atmosphere. A solution of trimethylsilylisocyanate (0.20 mL, 1.5 mmol, 0.35 equiv) in dry THF (45 mL) was added dropwise to a solution of 1,6-diaminohexane (**22a**; 500 mg, 4.30 mmol) in dry THF (15 mL) over a period of 3 h. The reaction mixture was then stirred for 2 h at room temperature. This was followed by addition of H_2O (6 mL) and additional stirring for 2 h. Next, the solvent was removed in vacuo and the residue was suspended in hot EtOAc and filtered. The remaining solid was washed with H_2O (10 mL). The product was isolated by freeze drying of the aqueous phase to yield a colorless foam (68 mg, 0.43 mmol, 28%). ^{1}H NMR (400 MHz, [D$_6$]DMSO): δ = 5.92 (t, J = 6.3 Hz, 1H; urea NH), 5.34 (s, 2H; urea NH_2), 3.44 (br s, 2H; NH_2), 2.93 (q, J = 6.6 Hz, 2H; H1), 2.53 (t, J = 6.7 Hz, 2H; H6), 1.39–1.20 ppm (m, 8H; alkyl H); ^{13}C NMR (101 MHz,

[D_6]DMSO): δ = 158.6 (urea C), 41.1 (C6), 38.9 (C1), 30.0, 26.2, 26.1, 26.0 ppm (alkyl C); IR (neat): $\tilde{\nu}$ = 3301, 2932, 2854, 1643, 1556, 1356, 1228, 587 cm^{-1}; HRMS (FAB): *m/z* calcd: 160.1449 $[M+H]^+$; found: 160.1451.

(7-Aminoheptyl)urea (8): The reaction was carried out under a nitrogen atmosphere. A solution of trimethylsilylisocyanate (0.18 mL, 1.4 mmol, 0.35 equiv) in dry THF (39 mL) was added dropwise to a solution of 1,7-diaminoheptane (**22b**; 500 mg, 3.85 mmol) in dry THF (13 mL) over a period of 3 h. The reaction mixture was stirred for 2 h at room temperature. This was followed by addition of H_2O (6 mL) and additional stirring for 2 h. The solvent was then removed in vacuo and the residue was suspended in hot EtOAc and filtered. The remaining solid was washed with H_2O (10 mL). The product was isolated by freeze drying of the aqueous phase to yield a colorless foam (24 mg, 0.14 mmol, 11%). ^{1}H NMR (400 MHz, [D_6]DMSO): δ = 5.92 (t, J = 6.4 Hz, 1H; urea NH), 5.34 (s, 2H; urea NH_2), 3.78 (br s, 2H; NH_2), 2.93 (q, J = 6.6 Hz, 2H; H1), 2.59 (t, J = 7.1 Hz, 2H; H7), 1.42–1.21 ppm (m, 10H; alkyl H); ^{13}C NMR (101 MHz, [D_6]DMSO): δ = 158.6 (urea C), 40.2 (C7), 38.7 (C1), 31.0, 29.9, 28.5, 26.3, 26.2 ppm (alkyl C); IR (neat): $\tilde{\nu}$ = 3349, 2931, 2852, 1651, 1596, 1553, 1262, 1225 cm^{-1}; HRMS (FAB): *m/z* calcd: 174.1606 $[M+H]^+$; found: 174.1614.

(8-Aminooctyl)urea (7): The reaction was carried out under a nitrogen atmosphere. A solution of trimethylsilylisocyanate (0.17 mL, 1.2 mmol, 0.35 equiv) in dry THF (37 mL) was added dropwise to a solution of 1,8-diaminooctane (**22c**; 500 mg, 3.5 mmol) in dry THF (13 mL) over a period of 2 h. The reaction mixture was then stirred for 2 h at room temperature. This was followed by addition of H_2O (6 mL) and additional stirring for 2 h. Next, the solvent was removed in vacuo and the residue was suspended in hot EtOAc and filtered. The remaining solid was washed with H_2O (10 mL). The product was isolated by freeze drying of the aqueous phase to yield a colorless foam (90 mg, 0.48 mmol, 39%). ^{1}H NMR (400 MHz, [D_6]DMSO): δ = 5.93 (t, J = 6.0 Hz, 1H; urea NH), 5.35 (s, 2H; urea NH_2), 3.60 (br s, 2H; NH_2), 2.92 (q, J = 6.6 Hz, 2H; H1), 2.56 (t, J = 7.2 Hz, 2H; H8), 1.40–1.30 (m, 4H; H2, H7), 1.27–1.22 ppm (m, 8H; alkyl H); ^{13}C NMR (101 MHz, [D_6]DMSO): δ = 158.7 (urea C), 40.8 (C8), 38.9 (C1), 31.6, 30.0, 28.9, 28.7, 26.4, 26.2 ppm (alkyl C); IR (neat): $\tilde{\nu}$ = 3393, 1928, 2851, 1655, 1594, 1558, 1404, 655, 572 cm^{-1}; HRMS (FAB): *m/z* calcd: 188.1762 $[M+H]^+$; found: 188.1763.

***N*-(6-Aminohexyl)prop-2-enamide (5)**: The reaction was carried out under a nitrogen atmosphere. A solution of 1,6-diaminohexane (**22a**; 500 mg, 4.30 mmol) in dry methanol (20 mL) was cooled to 0 °C. This was followed by dropwise addition of acryloyl chloride (0.35 mL, 4.3 mmol) over 1 h. The reaction mixture was then allowed to warm up to room temperature and stirred for a further 30 min. The solvent was removed in vacuo and the crude product was purified by column chromatography on silica gel (CH_2Cl_2/CH_3OH/Et_3N, 80:18:2) to yield a colorless solid (253 mg, 1.49 mmol, 35%). ^{1}H NMR (400 MHz, [D_6]DMSO): δ = 8.05 (s, 1H; amide NH), 6.20 (dd, J = 17.1, 10.3 Hz, 1H; H2), 6.05 (dd, J = 17.1, 2.3 Hz, 1H; H1a), 5.55 (dd, J = 10.0, 2.3 Hz, 1H; H1b), 3.25 (br s, 2H; NH_2), 3.10 (q, J = 6.8 Hz, 2H; H4), 2.53–2.51 (m, 2H; H9), 1.44–1.22 ppm (m, 8H; alkyl H); ^{13}C NMR (101 MHz, [D_6]DMSO): δ = 164.4 (amide C), 131.9 (C2), 124.6 (C3), 41.4 (C9), 38.5 (C4), 33.0, 29.1, 26.4, 26.0 ppm (alkylC); IR (neat): $\tilde{\nu}$ = 3296, 3057, 2930, 2856, 1653, 1542, 1239, 951 cm^{-1}; HRMS (FAB): *m/z* calcd: 171.1497 $[M+H]^+$; found: 171.1499.

4-Amino-3-iodo-*N*-benzylamine hydrochloride (31): The reaction was carried out under a nitrogen atmosphere. 1 M BH_3·THF solution (60 mL, 60 mmol, 3 equiv) was added dropwise to a solution of 4-amino-3-iodo-benzonitrile (**23**; 4.86 g, 20.0 mmol) in dry THF (60 mL). The reaction mixture was then heated at reflux for 4 h. After the reaction mixture had cooled down to room temperature, 2 M HCl (30 mL, 60 mmol, 3 equiv) was added. This was followed by further heating at reflux for 1 h. The solvent was removed in vacuo to yield a yellowish solid, which was used without further purification.

***O-tert*-Butyl-4-amino-3-iodobenzylcarbamate (24a)**: The reaction was carried out under a nitrogen atmosphere. A suspension of dry triethylamine (11 mL, 80 mmol, 4 equiv) and **31** (5.68 g, 20.0 mmol) in dry CH_2Cl_2 (90 mL) was cooled to 0 °C. After that, DMAP (487 mg, 3.99 mmol, 0.2 equiv) and di-*tert*-butyldicarbonate (9.20 g, 42.1 mmol, 2.1 equiv) were added. The reaction mixture was stirred at room temperature for 5 h. The solution was washed with H_2O (3×) and the aqueous phase was re-extracted with H_2O. The combined organic layers were dried over sodium sulfate and the solvent was removed under reduced pressure. The crude product was purified by column chromatography on silica gel (petroleum ether 50–70/EtOAc, 2:1) to yield a pale-orange oil (4.99 g, 14.4 mmol, 72%). ^{1}H NMR (400 MHz, [D_6]DMSO): δ = 7.41 (s, 1H; H2), 7.24–7.21 (t, J = 5.8 Hz 1H; NH), 6.95 (dd, J = 8.2, 1.9 Hz, 1H; H6), 6.68 (d, J = 8.2 Hz, 1H; H5), 5.09 (s, 2H; NH_2), 3.91 (d, J = 6.4 Hz, 2H; $BnCH_2$), 1.37 ppm (s, 9H; 3×BocCH_3); ^{13}C NMR (101 MHz, [D_6]DMSO): δ = 155.6 (Boc C=O), 147.2 (C1), 137.1 (C2), 129.8 (C3), 128.2 (C6), 114.1 (C5), 82.8 (C4), 77.6 (Boc $C(CH_3)_3$), 42.2 ($BnCH_2$), 28.2 ppm (Boc $C(CH_3)_3$); TLC: R_f = 0.55 (petroleum ether 50–70/EtOAc, 2:1); IR (neat): $\tilde{\nu}$ = 3342, 2975, 1686, 1614, 1496, 1365, 1247, 1154, 1027, 783 cm^{-1}; HRMS (ESI$^+$): *m/z* calcd: 349.0413 $[M+H]^+$; found: 349.0406.

***O-tert*-Butyl-3-iodo-4-(4-methylphenylsulfonylamido)benzylcarbamate (24b)**: The reaction was carried out under a nitrogen atmosphere. Dry pyridine (2.0 mL, 26 mmol, 3 equiv) and *p*-tosyl chloride (1.97 g, 10.3 mmol, 1.2 equiv) were added to a solution of **24a** (3.00 g, 8.61 mmol) in dry CH_2Cl_2 (36 mL). After being stirred at room temperature for 42 h, the reaction mixture was diluted with CH_2Cl_2 and extracted with H_2O (3×). The combined organic layers were dried over sodium sulfate and the solvent was removed under reduced pressure. The crude product was purified by column chromatography on silica gel (petroleum ether 50–70/EtOAc, 2:1) to yield an orange oil (3.99 g, 7.93 mmol, 92%). ^{1}H NMR (400 MHz, [D_6]DMSO): δ = 9.61 (s, 1H; Tos NH), 7.67 (s, 1H; H2), 7.58 (d, J = 7.8 Hz, 2H; H2′, H6′), 7.36 (d, J = 7.8 Hz, 3H; H3′, H5′, Boc NH), 7.13 (d, J = 8.4 Hz, 1H; H5), 6.90 (d, J = 8.4 Hz, 1H; H6), 4.07–3.99 (m, 2H; Bn CH_2), 2.38 (s, 3H; Tos CH_3), 1.38 ppm (s, 9H; 3×Boc CH_3); ^{13}C NMR (101 MHz, [D_6]DMSO): δ = 156.8 (Boc C=O), 143.5 (C4′), 141.4 (C1), 138.2 (C3), 138.0 (C2), 129.8 (C3′, C5′), 128.5 (C1′), 127.7 (C5), 127.3 (C6), 127.1 (C2′, C6′), 99.8 (C4), 78.6 (Boc $C(CH_3)_3$), 42.6 (Bn CH_2), 28.4 (Boc $C(CH_3)_3$), 21.4 ppm (Tos CH_3); TLC: R_f = 0.43 (petroleum ether 50–70/EtOAc, 2:1); IR (neat): $\tilde{\nu}$ = 3325, 2976, 2930, 2217, 1688, 1487, 1332, 1246, 1157, 664, 548 cm^{-1}; HRMS (ESI$^+$): *m/z* calcd: 525.0321 $[M+Na]^+$; found: 525.0322.

***O-tert*-Butyl-4-amino-3-(4-*N*,*N*′-bis-*tert*-butoxycarbonylguanidino-but-1-yne-1-yl)benzylcarbamate (28b)**: The reaction was carried out under a nitrogen atmosphere. DBU (0.21 mL, 1.37 mmol, 1.5 equiv) was added to a suspension of **24a** (306 mg, 0.878 mmol) and Pd EnCat TPP30 (77 mg, 3.5 mol%; 0.4 mmol Pd per 1 g) in dry CH_3CN (8 mL), followed by addition of **27** (410 mg, 1.37 mmol, 1.5 equiv). The reaction mixture was heated at reflux for 17 h and, after cooling down to room temperature, was filtered. The residue was washed with CH_2Cl_2 and CH_3OH. The filtrate was concentrated in vacuo and the crude product was purified by column chroma-

tography on silica gel (CH_2Cl_2/CH_3OH, 99:1→19:1) to yield a colorless solid (237 mg, 0.446 mmol, 51%). ^1H NMR (400 MHz, $CDCl_3$): $\delta = 9.08$ (s, 2H; guanidine NH_2), 7.17 (t, $J = 5.9$ Hz, 1H; NH), 6.96 (s, 1H; H2), 6.89 (d, $J = 8.2$ Hz, 1H; H6), 6.60 (d, $J = 8.2$ Hz, 1H; H5), 5.13 (s, 2H; NH_2), 4.01 (t, $J = 7.7$ Hz, 2H; H10), 3.90 (d, $J = 5.9$ Hz, 2H; Bn CH_2), 2.74 (t, $J = 7.0$ Hz, 2H; H9), 1.49 (s, 9H; 3×Boc CH_3), 1.41 (s, 9H; 3×Boc CH_3), 1.37 ppm (s, 9H; 3×Boc CH_3); ^{13}C NMR (101 MHz, $CDCl_3$): $\delta = 162.8$ (C11), 159.4 (Boc C=O), 155.6 (Boc C=O), 153.9 (Boc C=O), 148.1 (C4), 130.4 (C2), 128.3 (C6), 127.3 (C1), 113.6 (C5), 106.0 (C3), 91.4 (C8), 83.7 (Boc $C(CH_3)$), 79.0 (C7), 77.8 (Boc $C(CH_3)$), 77.6 (Boc $C(CH_3)$), 43.2 (C10), 42.7 (Bn CH_2), 28.3 (Boc $C(CH_3)$), 28.0 (Boc $C(CH_3)$), 27.5 (Boc $C(CH_3)$), 19.2 ppm (C9); TLC: $R_f = 0.55$ (CH_2Cl_2/CH_3OH, 9:1); IR (neat): $\tilde{\nu} = 3373$, 2976, 1709, 1608, 1502, 1365, 1245, 1141, 1004, 744 cm^{-1}; HRMS (ESI^+): *m/z* calcd: 532.3135 $[M+H]^+$; found: 532.3136.

5-(*O*-*tert*-Butoxycarbonyl)aminomethyl-2-(2-hydroxyethyl)-1-*N*-(4-methylphenylsulfonyl)indole (29b): The reaction was performed according to general procedure 2 by using **24b** (500 mg, 0.995 mmol), **25** (3 equiv), and dry DMF (10 mL). A CH_2Cl_2/CH_3OH gradient from 99:1 to 19:1 was used for the column chromatography to yield a pale-yellow oil (399 mg, 0.898 mmol, 90%). ^1H NMR (400 MHz, $CDCl_3$): $\delta = 7.94$ (d, $J = 8.2$ Hz, 1H; H7), 7.66 (d, $J = 8.2$ Hz, 2H; H2′, H6′), 7.34 (d, $J = 8.1$ Hz, 3H; H3′, H5′, NH), 7.30 (s, 1H; H4), 7.14 (d, $J = 8.7$ Hz, 1H; H6), 6.59 (s, 1H; H3), 4.77 (t, $J = 5.5$ Hz, 1H; OH), 4.15 (d, $J = 5.9$ Hz, 2H; Bn CH_2), 3.76 (dt, $J = 6.4$, 9.5 Hz, 2H; H9), 3.14 (t, $J = 6.6$ Hz, 2H; H8), 2.30 (s, 3H; Tos CH_3), 1.38 ppm (s, 9H; 3×Boc CH_3); ^{13}C NMR (101 MHz, $CDCl_3$): $\delta = 156.1$ (Boc C=O), 144.9 (C4′), 138.9 (C2), 134.4 (C7a), 130.4 (C3′, C5′), 129.4 (C3a), 126.0 (C2′, C6′), 122.8 (C5), 119.2 (C4), 110.1 (C3), 75.2 (Boc $C(CH_3)_3$), 59.5 (C9), 43.0 (Bn CH_2), 32.2 (C8), 28.4 (Boc $C(CH_3)_3$), 21.3 ppm (Tos CH_3); TLC: $R_f = 0.37$ (CH_2Cl_2/CH_3OH, 19:1); IR (neat): $\tilde{\nu} = 3400$, 3054, 2977, 2930, 1510, 1469, 1364, 1090, 961, 670 cm^{-1}; HRMS (ESI^+): *m/z* calcd: 467.1617 $[M+Na]^+$; found: 467.1615.

***N*,*N*′-Bis-*tert*-Butoxycarbonylguanidinobut-3-yne (27)**: The reaction was performed according to general procedure 1 by using **25** (0.38 mL, 5.0 mmol) to yield a colorless solid (1.36 g, 4.37 mmol, 87%). ^1H NMR (400 MHz, $CDCl_3$): $\delta = 9.06$ (s, 2H; NH_2), 3.90 (t, $J = 8.3$ Hz, 2H; H1), 2.89 (t, $J = 2.5$ Hz, 1H; H4), 2.44 (dt, $J = 8.1$, 2.9 Hz, 2H; H2), 1.50 (s, 9H; 3×Boc CH_3), 1.41 ppm (s, 9H; 3×Boc CH_3); ^{13}C NMR (101 MHz, $CDCl_3$): $\delta = 160.9$ (C5), 154.8 (2×Boc C=O), 83.0 (C3), 80.0 (2×Boc $C(CH_3)_3$), 73.3 (C4), 43.4 (C1), 28.4 (2×Boc $C(CH_3)_3$), 18.3 ppm (C2); TLC: $R_f = 0.42$ (petroleum ether 50–70/EtOAc, 2:1); IR (neat): $\tilde{\nu} = 3376$, 3197, 2984, 2965, 1637, 1512, 1450, 1308, 1120, 1003, 597 cm^{-1}; MS (FAB): *m/z* calcd: 312.2 $[M+H]^+$; found: 312.3.

5-(*O*-*tert*-Butoxycarbonyl)aminomethyl-2-(2-*N*,*N*′-bis-*tert*-butoxycarbonylguanidinoethyl)-1-*N*-(4-methylphenylsulfonyl)indole (30b): Method a: The reaction was performed according to general procedure 2 by using **24b** (254 mg, 0.506 mmol), **27** (3 equiv), and dry DMF (7 mL). A petroleum ether 50–70/EtOAc gradient from 4:1 to 1:1 was used for the column chromatography to yield a pale-yellow oil (212 mg, 0.309 mmol, 62%). ^1H NMR (400 MHz, $CDCl_3$): $\delta = 9.12$ (s, 2H; NH_2), 8.01 (d, $J = 8.6$ Hz, 1H; H7), 7.71 (d, $J = 8.2$ Hz, 2H; H2′, H6′), 7.37–7.30 (m, 3H; H3′, H5′, NH), 7.29 (s, 1H; H4), 7.18 (dd, $J = 8.8$, 1.2 Hz, 1H; H6), 6.50 (s, 1H; H3), 4.20 (t, $J = 6.3$ Hz, 2H; H9), 4.14 (d, $J = 5.9$ Hz, 2H; Bn CH_2), 3.23 (t, $J = 5.8$ Hz, 2H; H8), 2.30 (s, 3H; Tos CH_3), 1.41 (s, 9H; 3×Boc CH_3), 1.38 (s, 9H; 3×Boc CH_3), 1.19 ppm (s, 9H; 3×Boc CH_3); ^{13}C NMR (101 MHz, $CDCl_3$): $\delta = 160.0$ (Boc C=O), 156.1 (Boc C=O), 154.5 (C10), 145.6 (C4′), 138.8 (C2), 135.5 (C3a), 135.5 (C1′), 130.5 (C3′, C5′), 130.0 (C7a), 129.9 (C5), 126.3 (C2′, C6′), 123.9 (C6), 119.0 (C4), 114.3 (C7), 110.8 (C3), 84.7 (Boc $C(CH_3)_3$), 83.9 (Boc $C(CH_3)_3$), 78.3 (Boc $C(CH_3)_3$), 43.6 (Bn CH_2). 44.2 (C9), 28.5 (Boc $C(CH_3)_3$), 28.1 (C8), 27.5 (Boc $C(CH_3)_3$), 21.4 ppm (Tos CH_3); TLC: $R_f = 0.42$ (petroleum ether 50–70/EtOAc, 2:1); IR (neat): $\tilde{\nu} = 3378$, 2931, 1709, 1607, 1504, 1471, 1443, 1365, 1282, 1162, 811, 545 cm^{-1}; HRMS (FAB): *m/z* calcd: 686.3 $[M+H]^+$; found: 686.5.

Method b: The reaction was performed according to general procedure 1 by using **29b** (399 mg, 0.897 mmol) to yield a pale-yellow oil (585 mg, 0.853 mmol, 95%). The analytical data were identical to those reported above.

5-(*O*-*tert*-Butoxycarbonyl)aminomethyl-2-(2-*N*,*N*′-bis-*tert*-butoxycarbonylguanidinoethyl)-1*H*-indole (30a): Method a: The reaction was carried out under a nitrogen atmosphere. A solution of **30b** (59 mg, 86 μmol) in dry methanol (3 mL) was mixed with NaOH (600 mg, 15 mmol). The viscose reaction mixture was stirred at room temperature for 40 h. After dilution with EtOAc, the solution was washed with H_2O (3×). The aqueous phase was extracted with EtOAc and the combined organic layers were dried over sodium sulfate. After removal of the solvent, the crude product was purified by column chromatography on silica gel (CH_2Cl_2/CH_3OH, 19:1→9:1) to yield a colorless solid (43 mg, 82 μmol, 95%). ^1H NMR (400 MHz, $CDCl_3$): $\delta = 10.94$ (s, 1H; NH1), 7.27 (br s, 2H; H4, Boc NH), 7.22 (d, $J = 8.2$ Hz, 1H; H7), 6.92 (d, $J = 8.2$ Hz, 1H; H6), 6.17 (s, 1H; H3), 4.15 (d, $J = 6.3$ Hz, 1H; Bn CH_2), 3.52 (t, $J = 7.9$ Hz, 2H; H9), 2.90 (br s, 2H; H8), 1.42 (s, 9H; 3×Boc CH_3), 1.38 (s, 9H; 3×Boc CH_3), 1.29 ppm (s, 9H; 3×Boc CH_3); ^{13}C NMR (101 MHz, $CDCl_3$): $\delta = 160.9$ (C10), 155.7 (2×Boc C=O), 136.9 (C2), 135.1 (C7a), 130.4 (C5), 128.1 (C3a), 120.2 (C6), 117.7 (C4), 110.4 (C7), 99.0 (C3), 77.5 (C9), 44.0 (Bn CH_2), 28.3 (2×Boc $C(CH_3)$), 28.1 ppm (C8); TLC: $R_f = 0.28$ (CH_2Cl_2/CH_3OH, 9:1); IR (neat): $\tilde{\nu} = 3281$, 2930, 1683, 1632, 1486, 1390, 1146, 966, 776 cm^{-1}; HRMS (FAB): *m/z* calcd: 532.3135 $[M+H]^+$; found: 532.3122.

Method b: The reaction was carried out under a nitrogen atmosphere. KH (19 mg, 0.48 mmol, 1.2 equiv) was added to a suspension of **28b** (214 mg, 0.402 mmol) and $Cu(OAc)_2$ (80 mg, 0.44 mmol, 1.1 equiv) in dry 1,2-dichloroethane (11 mL). After that, the reaction mixture was stirred at 70 °C for 4 d. The crude product was purified by column chromatography on silica gel (CH_2Cl_2/CH_3OH, 99:1→9:1) to yield a pale-yellow solid (123 mg, 0.230 mmol, 57%). The analytical data were identical to those reported above.

2-(2-Guanidinoethyl)-5-aminomethyl-1*H*-indole (10): In a 25 mL round-bottomed flask, **30a** (55 mg, 0.10 mmol) was dissolved in CH_3CN/2 M HCl (1:1; 8 mL). After the reaction mixture had been stirred at room temperature for 28 h, the solvent was removed under reduced pressure. The crude product was purified by column chromatography on RP silica gel with H_2O as the eluent. Freeze drying of the product fractions yielded a pale-rose solid of **10** as the hydrochloride salt (22 mg, 73 μmol, 73%). ^1H NMR (400 MHz, $CDCl_3$): $\delta = 11.37$ (s, 1H; NH1), 8.33 (s, 3H; Bn NH_2, guanidine NH), 7.86 (t, $J = 5.8$ Hz, 1H; guanidine NH), 7.54 (s, 1H; H4), 7.32 (d, $J = 8.7$ Hz, 1H; H7), 7.13 (dd, $J = 8.3$, 1.7 Hz, 1H; H6), 6.27 (s, 1H; H3), 4.01 (dt, $J = 5.8$, 5.8 Hz, 2H; Bn CH_2), 3.52 (dt, $J = 6.9$, 6.3 Hz, 2H; H9), 2.94 ppm (t, $J = 7.3$ Hz, 2H; H8); ^{13}C NMR (101 MHz, $CDCl_3$): $\delta = 156.9$ (C10), 136.9 (C2), 135.4 (C7a), 127.9 (C3a), 123.8 (C5), 120.8 (C6), 120.4 (C4), 110.7 (C7), 99.1 (C3), 58.7 (C9), 42.6 (Bn CH_2), 28.8 ppm (C8); IR (neat): $\tilde{\nu} = 3326$, 3232, 3052, 2923, 1683, 1486, 1309, 1164, 967, 804, 641 cm^{-1}; HRMS (FAB): *m/z* calcd: 232.1557 $[M+H]^+$; found: 232.1552.

Biochemistry

DHS activity assay: The deoxyhypusine synthase reaction was conducted, in principle, as described previously.[57] Briefly, the reaction mixture contained eIF5A precursor protein (5 μg), DHS (3 μg), and [^{3}H]spermidine (2 μCi, 32.4 Ci mmol^{-1}) in a total volume of 200 μL of reaction buffer (300 mM glycine–NaOH (pH 9.0) containing 1 mM NAD^+, 1 mM 1,4-dithiothreitol (DTT), and 50 μg mL^{-1} bovine serum albumin (BSA)). To test for potential influence of the new inhibitors on DHS activity, the enzymatic reaction was performed in the presence of the indicated concentrations of the indicated compounds **1**–**10**, or with GC7 as a positive control, or without any compounds as a solvent or negative control, respectively. After 1 h at 37 °C, the reaction was stopped by adding 20 mM spermidine (200 μL) in PBS and was transferred onto a Millipore GSWPO2500 nitrocellulose membrane, which was previously blocked with 20 mM spermidine/PBS for 1 h. The reaction mixture was vacuum filtered and then the membrane was washed with PBS (5 mL). Finally, the membranes were air dried and the tritium signal was measured in a liquid scintillation counter.

Overexpression and purification of DHS and eIF-5A: *Escherichia coli* strain BL21 (DE3) (Novagen) was employed for protein expression.

His-DHS purification: The pTricHis-DHS clone was used for the preparation for recombinant DHS.[55] A 1 L culture of Luria–Bertani medium containing 50 μg mL^{-1} ampicillin was inoculated with 50 mL volume of an overnight culture of *E. coli* BL21 (DE3) transformed with pTricHis-DHS (preinoculum) and grown at 37 °C until the optical density value at 600 nm reached 0.6–0.8 (~3 h). The Tre recombinase gene expression was induced by addition of 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG) and the culture was grown for a further period of 4 h at 37 °C. After incubation, the bacteria were harvested by centrifugation (10 min, 5000 *g*, 4 °C) and stored at −80 °C until further use.

The cell pellet was resuspended in lysis buffer (30 mL; 50 mM tris(hydroxymethyl)aminomethane (Tris; pH 8) containing 300 mM NaCl and 30 mM imidazole) and lysed by sonication. The soluble fraction was separated by centrifugation at 12000 rpm for 15 min. A bed volume of 1 mL of buffer-equilibrated nickel–nitriloacetate (Ni-NTA) beads was added to the supernatant and incubated with gentle rocking in an end-to-end rotor at 4 °C for 1 h. The beads were then packed in a column and washed with a wash buffer (50 mM Tris (pH 8) containing 300 mM NaCl and 30 mM imidazole), followed by a second wash buffer (50 mM Tris (pH 8) containing 300 mM NaCl and 50 mM imidazole). The protein was eluted with 50 mM Tris (pH 8) containing 300 mM NaCl and 300 mM imidazole. The purified His-DHS protein was analyzed by SDS-PAGE, dialyzed against 300 mM glycine–NaOH (pH 9) and 10% glycerol, and stored at −20 °C.

eIF-5A expression and purification: The pGEX-eIF5A clone was used to prepare eIF-5A protein, as described previously but with minor modifications.[58] GST-eIF5A overexpression was performed as described for DHS.

The IPTG-induced cell pellet was resuspended in lysis buffer (40 mL; 10 mM PBS (pH 7.4) containing 1 mM DTT, 5 mM ethylenediaminetetraacetate (EDTA), 2.6 mM $MnCl_2$, and 26 mM $MgCl_2$) with 0.5 μg mL^{-1} DNAse and protease inhibitors (0.1 mM phenylmethanesulfonyl fluoride (PMSF), 2 μg mL^{-1} leupeptin, and 2 μg mL^{-1} aprotinin) and lysed by sonication. 1% Triton X-100 was added to the total lysate (Triton X-100 was diluted in 1× PBS and then used) and the mixture was incubated at 4 °C for 1 h on an end-to-end rotor. The soluble fraction was separated by centrifugation at 12000 rpm for 15 min.

A bed volume of 1 mL of buffer-equilibrated glutathione (GSH) beads were added to the supernatant and the mixture was incubated with gentle rocking in an end-to-end rotor at 4 °C for 1–2 h. The lysate was passed through a column so that the beads were packed in a column. The beads were then washed with a wash buffer (100 mL; 50 mL PBS (pH 7.4) containing 1 mM DTT and 5 mM EDTA) and with a second buffer (150 mL; 50 mM Tris (pH 8) containing 200 mM NaCl and 5 mM EDTA). The protein was eluted with 50 mM Tris-HCl containing 200 mM NaCl and 20 mM reduced glutathione (pH 8.0). The purified protein fractions were dialyzed against the factor Xa cleavage buffer (50 mM Tris-HCl containing 100 mM NaCl and 1 mM $CaCl_2$). Factor Xa cleavage and final purification of eIF-5A was performed exactly as described previously.[58] eIF-5A was dialyzed in assay buffer (300 mM glycine/NaOH, pH 9) and stored at −20 °C.

HIV-1 infection experiments: PM1 cells were pre-incubated for 7 d in the presence of compound **10**, the DHS inhibitor GC7, or in medium alone. Subsequently, the cells were infected with the CCR5-tropic HIV-1 strain BaL.[59] For infection, roughly 3×10^6 cells were resuspended in culture medium (500 μL) without drugs and incubated at 37 °C for 3 h with HIV-1 viral stock (3 ng). After infection, cells were washed twice with PBS and further cultivated for another week in the presence of the compounds or in medium alone. At day 7 after infection, the p24 levels in the supernatant were determined with an enzyme-linked immunosorbent assay (Innogenetics NV). PM1 cells were maintained in RPMI medium (Invitrogen) containing 10% fetal calf serum (FCS; Biochrom AG), 100 U mL^{-1} penicillin, and 100 mg mL^{-1} streptomycin (Invitrogen).

Acknowledgements

The authors are grateful to the Bundesministerium für Bildung und Forschung (BMBF) for financial support under grant 01GU0715-0718 "Combating Drug Resistance in CML and HIV-1 Infection". We also thank the other research groups collaborating in this project: the groups of Prof. Dr. R. Hilgenfeld (University of Lübeck), Prof. Dr. T. Brümmendorf (University Hospital Aachen), and Dr. S. Balabanov (University Hospital Zürich). We thank Ilona Hauber (Heinrich Pette Institute) for technical assistance with HIV experiments.

Keywords: antiviral agents · drug design · enzymes · HIV · inhibitors

[1] M. A. Thompson, J. A. Aberg, P. Cahn, J. S. Montaner, G. Rizzardini, A. Telenti, J. M. Gatell, H. F. Günthard, S. M. Hammer, M. S. Hirsch, D. M. Jacobsen, P. Reiss, D. D. Richman, P. A. Volberding, P. Yeni, R. T. Schooley, *JAMA J. Am. Med. Assoc.* **2010**, *304*, 321–333.
[2] S. G. Deeks, A. N. Phillips, *BMJ* **2009**, *338*, a3172.
[3] A. Calmy, B. Hirschel, D. A. Cooper, A. Carr, *Antiviral Ther.* **2009**, *14*, 165–179.
[4] D. D. Richman, *Antiviral Res.* **2006**, *71*, 117–121.
[5] C. Flexner, *Nat. Rev. Drug Discovery* **2007**, *6*, 959–966.
[6] B. Larder, *AIDS* **2001**, *15*, S27–S34.
[7] S. P. Goff, *Nat. Rev. Microbiol.* **2007**, *5*, 253–263.
[8] M. Suhasini, T. M. Reddy, *Curr. HIV Res.* **2009**, *7*, 91–100.
[9] P. Saini, D. E. Eyler, R. Green, T. E. Dever, *Nature* **2009**, *459*, 118–121.
[10] E. Gutierrez, P.-S. Shin, C. J. Woolstenhulme, J. R. Kim, P. Saini, A. R. Buskirik, T. E. Dever, *Molecular Cell* **2013**, *51*, 35–45.

[11] M. Ruhl, M. Himmelspach, G. M. Bahr, F. Hammerschmid, H. Jaksche, B. Wolff, H. Aschauer, G. K. Farrington, H. Probst, D. Bevec, *J. Cell Biol.* **1993**, *123*, 1309 – 1320.
[12] D. Bevec, H. Jaksche, M. Oft, T. Woehl, M. Himmelspach, A. Pacher, M. Schebesta, K. Koettnitz, M. Dobrovnik, *Science* **1996**, *271*, 1858 – 1860.
[13] V. W. Pollard, M. H. Malim, *Annu. Rev. Microbiol.* **1998**, *52*, 491 – 532.
[14] H. C. Groom, E. C. Anderson, A. M. Lever, *J. Gen. Virol.* **2009**, *90*, 1303 – 1318.
[15] M. H. Park, *J. Biochem.* **2006**, *139*, 161 – 169.
[16] Y. B. Lee, M. H. Park, J. E. Folk, *J. Med. Chem.* **1995**, *38*, 3053 – 3061.
[17] J. Jakus, E. C. Wolff, M. H. Park, J. E. Folk, *J. Biol. Chem.* **1993**, *268*, 13151 – 13159.
[18] A. Kaiser, A. Gottwald, C. Wiersch, B. Lindenthal, W. Maier, H. M. Seitz, *Parasitol. Res.* **2001**, *87*, 963 – 972.
[19] R. A. Hart, J. N. Billaud, S. J. Choi, T. R. Phillips, *Virology* **2002**, *304*, 97 – 104.
[20] I. Hauber, D. Bevec, J. Heukeshoven, F. Krätzer, F. Horn, A. Choidas, T. Harrer, J. Hauber, *J. Clin. Invest.* **2005**, *115*, 76 – 85.
[21] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235 – 242.
[22] J. Schlosser, M. Rarey, *J. Chem. Inf. Model.* **2009**, *49*, 800 – 809.
[23] P. Maass, T. Schulz-Gasch, M. Stahl, M. Rarey, *J. Chem. Inf. Model.* **2007**, *47*, 390 – 399.
[24] J. R. Fischer, U. Lessel, M. Rarey, *J. Chem. Inf. Model.* **2010**, *50*, 1 – 21.
[25] LeadIT, version 2.0.2, BioSolveIT GmbH, Sankt Augustin (Germany), http://www.biosolveit.de/LeadIT (accessed November 12, **2013**).
[26] I. Reulecke, G. Lange, J. Albrecht, R. Klein, M. Rarey, *ChemMedChem* **2008**, *3*, 885 – 897.
[27] T. C. Umland, E. C. Wolff, M. H. Park, D. R. Davies, *J. Biol. Chem.* **2004**, *279*, 28697 – 28705.
[28] D. I. Liao, E. C. Wolff, M. H. Park, D. R. Davies, *Structure* **1998**, *6*, 23 – 32.
[29] MOE: Molecular Operating Environment, version 2012.10, Chemical Computing Group, Montreal, QC (Canada), http://www.chemcomp.com (accessed February 26, **2014**).
[30] CORINA: Fast Generation of High-Quality 3D Molecular Models, version 3.48, Molecular Networks, Erlangen (Germany), http://www.molecular-networks.com/products/corina (accessed November 12, **2013**).
[31] S. Urbaczek, A. Kolodzik, J. R. Fischer, T. Lippert, S. Heuser, I. Groth, T. Schulz-Gasch, M. Rarey, *J. Chem. Inf. Model.* **2011**, *51*, 3199 – 3207.
[32] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model.* **2005**, *45*, 177 – 182.
[33] A. Griewel, O. Kayser, J. Schlosser, M. Rarey, *J. Chem. Inf. Model.* **2009**, *49*, 2303 – 2311.
[34] J. Degen, C. Wegscheid-Gerlach, A. Zaliani, M. Rarey, *ChemMedChem* **2008**, *3*, 1503 – 1507.
[35] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, *J. Mol. Biol.* **1996**, *261*, 470 – 489.
[36] S. Wang, Z. Li, W. Hua, *Synth. Commun.* **2002**, *32*, 3339 – 3345.
[37] W. P. Hu, H. S. Yu, Y. R. Chen, Y. M. Tsai, Y. K. Chen, C. C. Liao, L. S. Chang, J. Wang, *J. Bioorg. Med. Chem.* **2008**, *16*, 5295 – 5302.
[38] S. Ueda, H. Nagasawa, *J. Org. Chem.* **2009**, *74*, 4272 – 4277.
[39] S. H. Hausner, D. Alagille, A. O. Koren, L. Amici, J. K. Staley, K. P. Cosgrove, R. M. Baldwin, G. D. Tamagnan, *Bioorg. Med. Chem. Lett.* **2009**, *19*, 543 – 545.
[40] A. Miyagawa, M. C. Z. Kasuya, K. Hatanaka, *Bull. Chem. Soc. Jpn.* **2006**, *79*, 348 – 356.
[41] G. L. Stahl, R. Walter, C. W. Smith, *J. Org. Chem.* **1978**, *43*, 2285 – 2296.
[42] A. Akinc, A. Zumbuehl, M. Goldberg, E. S. Leshchiner, V. Busini, N. Hossain, S. A. Bacallado, D. N. Nguyen, J. Fuller, R. Alvarez, A. Borodovsky, T. Borland, R. Constien, A. de Fougerolles, J. R. Dorkin, K. N. Jayaprakash, M. Jayaraman, M. John, V. Koteliansky, M. Manoharan, L. Nechev, J. Qin, T. Racie, D. Raitcheva, K. G. Rajeev, D. W. Y. Sah, J. Soutschek, I. Toudjarska, H. P. Vornlocher, T. S. Zimmermann, R. Langer, D. G. Anderson, *Nat. Biotechnol.* **2008**, *26*, 561 – 569.
[43] R. Manchanda, S. K. Agarwal, P. Kumar, A. K. Sharma, K. C. Gupta, R. Chandra, *Helv. Chim. Acta* **2002**, *85*, 2754 – 2762.
[44] P. Boden, J. M. Eden, J. Hodgson, D. C. Horwell, J. Hughes, A. T. McKnight, R. A. Lewthwaite, M. C. Pritchard, J. Raphy, *J. Med. Chem.* **1996**, *39*, 1664 – 1675.
[45] F.-N. Li, N.-J. Kim, D.-J. Chang, J. Jang, H. Jang, J.-W. Jung, K.-H. Min, Y.-S. Jeong, S.-Y. Kim, Y.-H. Park, H.-D. Kim, H.-G. Park, Y.-G. Suh, *Bioorg. Med. Chem.* **2009**, *17*, 8149 – 8160.
[46] R. C. Larock, E. K. Yum, M. D. Refvik, *J. Org. Chem.* **1998**, *63*, 7652 – 7662.
[47] H. Adachi, K. K. Palaniappan, A. A. Ivanov, N. Bergman, Z.-G. Gao, K. A. Jacobson, *J. Med. Chem.* **2007**, *50*, 1810 – 1827.
[48] M. J. Robins, Y. Peng, V. L. Damaraju, D. Mowles, G. Barron, T. Tackaberry, J. D. Young, C. E. Cass, *J. Med. Chem.* **2010**, *53*, 6040 – 6053.
[49] D. S. Dodd, A. P. Kozikowski, *Tetrahedron Lett.* **1994**, *35*, 977 – 980.
[50] K. Hiroya, S. Itoh, T. Sakamoto, *J. Org. Chem.* **2004**, *69*, 1126 – 1136.
[51] S. A. Monti, R. R. Schmidt, *Tetrahedron* **1971**, *27*, 3331 – 3339.
[52] A. J. Elliott, H. Guzik, *Tetrahedron Lett.* **1982**, *23*, 1983 – 1984.
[53] B. Robinson, *Chem. Rev.* **1969**, *69*, 785 – 797.
[54] G. Vaidyanathan, M. R. Zalutsky, *J. Org. Chem.* **1997**, *62*, 4867 – 4869.
[55] R. Chinchilla, C. Nájera, *Chem. Rev.* **2007**, *107*, 874 – 922.
[56] J. Sedelmeier, S. V. Ley, H. Lange, I. R. Baxendale, *Eur. J. Org. Chem.* **2009**, 4412 – 4420.
[57] D. Bevec, B. Kappel, H. Jaksche, R. Csonga, J. Hauber, H. Klier, A. Steinkasserer, *FEBS Lett.* **1996**, *378*, 195 – 198.
[58] M. N. Sommer, D. Bevec, B. Klebl, B. Flicke, K. Hoelscher, T. Freudenreich, I. Hauber, J. Hauber, H. Mett, *J. Biomol. Screening* **2004**, *9*, 434 – 438.
[59] S. Gartner, P. Markovits, D. M. Markovitz, M. H. Kaplan, R. C. Gallo, M. Popovic, *Science* **1986**, *233*, 215 – 219.

Received: November 25, 2013
Revised: January 29, 2014
Published online on March 11, 2014

Appendix B

NAOMI

The NAOMI [A1] framework and its extensions are implemented in C++. Tests for each class are using the framework Qt [90]. A basic class of the NAOMI framework is the class Molecule (see Figure B.1). The class Molecule describes a connected molecular graph. It includes pointers to the corresponding atoms, bonds, and ring systems. An atom object points to the corresponding objects of the classes Element, ValenceState, and AtomType. A bond object contains information about its order in a localized valence bond structure and pointers to both atoms it connects. Each object of the class Ringsystem represents a biconnected component of the molecular graph. It includes pointers to the contained atoms and rings. A ring object represents a connected cycle of a molecular graph and points to its atom objects and the containing ring system object.

Objects of the classes Molecule, Atom, Bond, Ringsystem, and Ring are dynamically created for each molecular structure. The objects of the classes Element, ValenceState, and AtomType are only generated once at the start of the program via a singleton pattern [88]. The NAOMI tool can be downloaded free of charge [89].

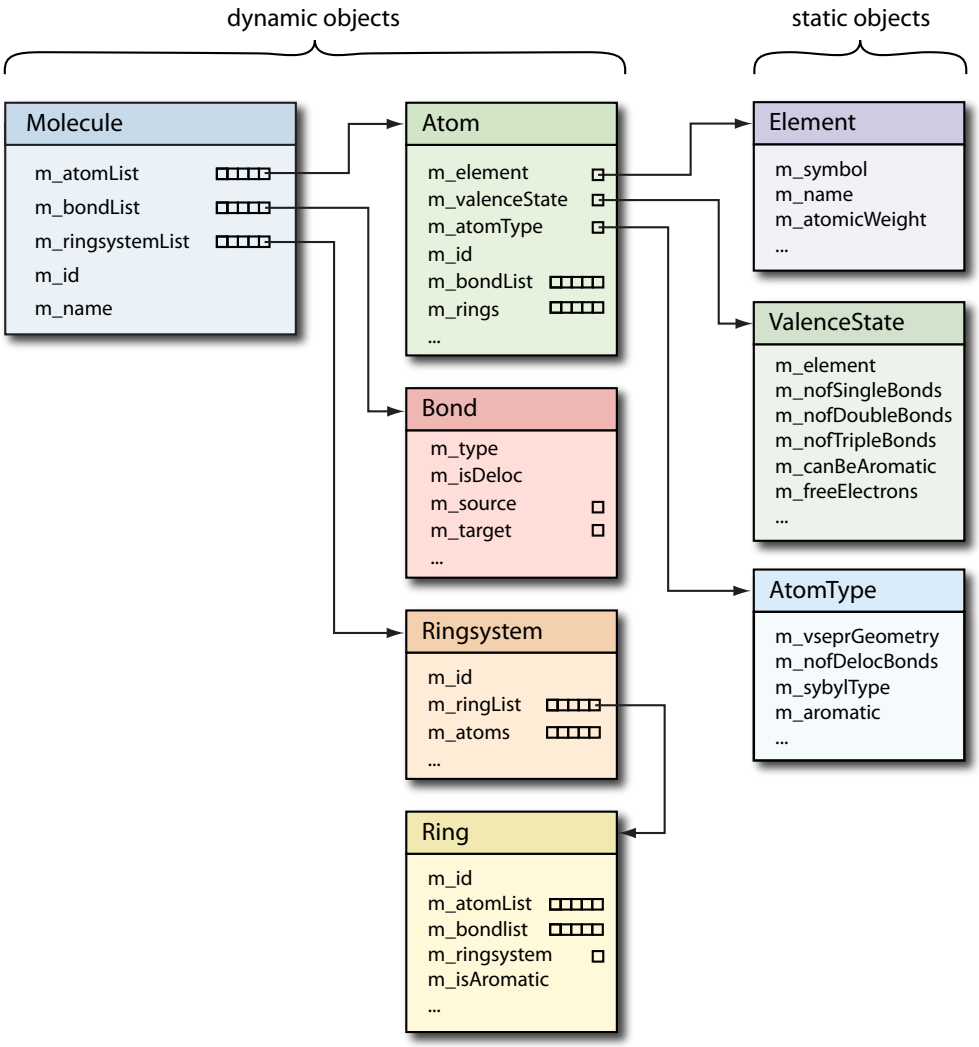


Figure B.1: Schematic view of the structure of NAOMI's central class Molecule. Only a subset of the connections from the class Molecule to the objects of the other classes are shown with arrows.

Appendix C

URF Perception

The perception of URFs is integrated into the NAOMI framework. Based on a molecular graph, the biconnected components (ring systems) are calculated as described by Tarjan [91]. For each biconnected component the RCPs are determined as described by Vismara [62].

The last step of the calculation of RCPs involves a Gaussian elimination. Cycles are represented by incidence vectors of their edges. The term cycle addition will be used as described in section 3.1. The elimination starts with a set of RCP candidates. During the elimination, cycles are processed in the order of their size (smallest first) and can fall into three categories. (1) If a cycle can be generated by cycle addition of a set of strictly smaller cycles, it is eliminated. (2) A cycle which does not fall into category one, but which can be generated by cycle addition of smaller - and equal-sized cycles, is an RC and part of a minimum cycle basis. In this case, the corresponding cycles of identical size are marked as potentially URF-related. (3) The remaining cycles are independent of smaller and equal sized cycles and therefore part of all minimum cycle bases. These cycles also belong to the RCs.

Each 2-pair of the potentially URF-related cycles (identified in step 2) are further analyzed. If the two cycles share at least one bond, they are URF-related. The transitive closure of this relation forms the URF. Some parts of the NAOMI framework require the calculation of all RCs. Therefore, the RCs are calculated on the basis of the RCPs as described by Vismara. All RCs which are generated from an RCP are URF-related and therefore belong to the same URF.

URFs can, however, be generated without the enumeration of the exponential number of RCs (see Figure C.1). Based on the set of RCPs, it is possible to identify all bonds which belong to a URF. This is done by a backtracking procedure, which is similar to the generation of RCs as described by Vismara. In Vismara's algorithm, RCs

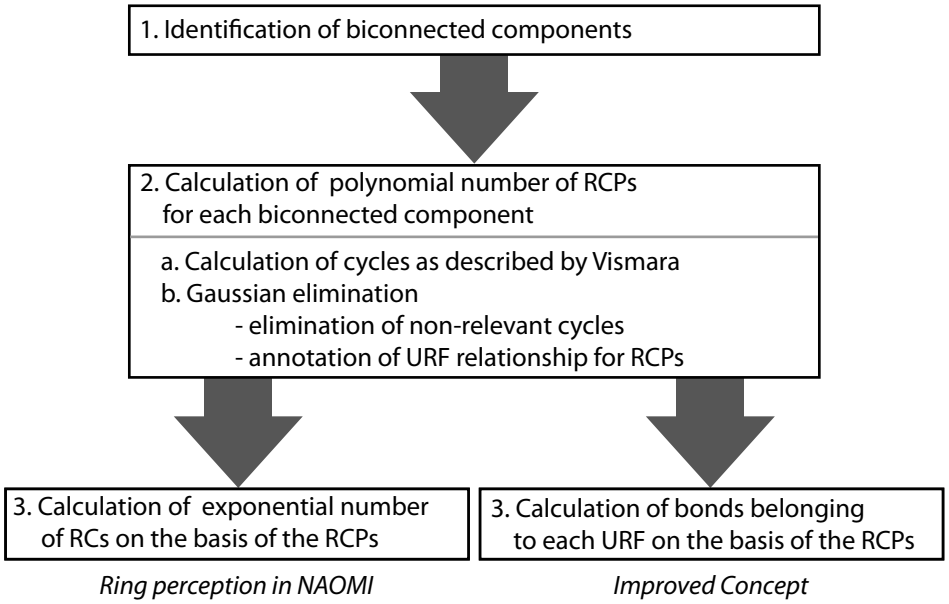


Figure C.1: URFs can be perceived without the calculation of the exponential number of RCs. The NAOMI framework still relies on the set of RCs.

are generated by identfying alternative paths of equal length for each half of the cycle and enumerating all combinations. To identify bonds of an URF, these combination do not have to be enumerated. Instead, each bond in the alternative paths is directly assigned to the URF. This process represents a significant runtime improvement. It only requires polynomial instead of exponential runtime. Nevertheless, it can only be integrated into the NAOMI framework as soon as the other parts of the framework have been modified to be independent of the calculation of RCs. The separation of the different aspects of the algorithm into cohesive function allows an easy change of the ring perception code for this purpose.

Acknowledgements

I would like to thank the "Bundesministerium für Bildung und Forschung" (BMBF) for financial support for the design of DHS inhibitors under grant 01GU0715-0718 "Combating Drug Resistance in CML and HIV-1 Infection". Furthermore, I would like to thank the co-authors of the papers, which resulted from this very productive and interesting project. I would like to especially thank Dr. Marcus Schröder for the synthesis of the *in silico* designed inhibitors and for his persistent efforts to get the corresponding papers published.

I would like to express my sincere gratitude to my former colleagues at the Center of Bioinformatics (ZBH), many of which have become good friends over the years. I really enjoyed the very open discussions with Dr. Christian Ehrlich who shared an office with me at the ZBH. I want to especially thank Sascha Urbaczek, Robert Fischer, Matthias Hilbig, and Tobias Lippert for their contributions to the NAOMI library and for their support and help during the work for this thesis. I am sincerely grateful that I had the chance to be a part of the "NAOMI team". A special thanks goes to Robert Fischer and Matthias Hilbig for proof-reading this thesis and for challenging nights with board - and card games. A big thanks also goes to my colleague Dr. Alexander Böcker-Felbek for proofreading one of the final drafts of this thesis. I thank Prof. Dr. Matthias Rarey for supervising my research at the ZBH and for his scientific guidance.

Last but not least I would like to thank my family - Arnold Kolodzik, Ute Zarsen-Kolodzik, Matthias Zarsen, and Karsten Mundhenk - as well as my girlfriend Stefanie Lange for their continuous support.

Eidesstattliche Erklärung

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

Hamburg, den 05. Juli 2014

_____

Adrian Kolodzik