Yunhan Chu

Development of Methods for *de novo* Design of Functional Drugs and Catalyst Compounds

Thesis for the degree of Philosophiae Doctor

Trondheim, June 2011

Norwegian University of Science and Technology
Faculty of Natural Sciences and Technology
Department of Chemistry

NTNU – Trondheim
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Natural Sciences and Technology
Department of Chemistry

ISBN 978-82-471-2876-3 (printed ver.)
ISBN 978-82-471-2877-0 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2011:165

Printed by NTNU-trykk

Acknowledgments

First, a special thanks to Prof. Bjørn K. Alsberg for introducing me to the fields of chemometrics and cheminformatics, and all his help and support as my supervisor. I also deeply acknowledge Prof. Vidar R. Jensen for his excellent instruction on my work in catalyst design. Wouter Heyndrickx is thanked for the wonderful collaboration time. Dr. Giovanni Occhipinti is appreciated for all his valuable advice and comments.

Sincere thanks are extended to colleagues at physical chemistry for their nice friendship. In particular, Terje Bruvoll is thanked for nice coffee breaks as well as helping me getting computational resource. Einar Ryeng is thanked for technical help and interesting discussions. Tjasa Kumelj is thanked for the help of using biological software.

Ingvild Eide-Haugmo, Jingzhe Jin and Fen Qin are thanked for their friendship and special concern.

I am grateful for the financial support from the Department of Chemistry. The NOTUR and NTNU local supercomputing programmes are acknowledged for provision of plenty of cpu time.

Furthermore, I am thankful for the help from Lena Frostad, Bjørn F. Syvertsen, Inger. M. Froseth and other administrative colleagues.

Our families deserve thanks for their persistent love and support throughout my life. And of course, my dear Xuezhong, you are the main source of my inspiration and the main reason why I go ahead. This thesis is dedicated to you, but I know it is nothing compared to what you have done for me.

Summary

De novo design produces novel molecular structures with desired properties by taking chemical space as a source. Open source software has had a significant impact on several areas of computer-aided molecular design, such as cheminformatics, docking studies and bioinformatics but has been slow to impact *de novo* design field in a similar way. This triggered our development of the open source software GeneGear. GeneGear is a Java-platform that is built with the chemistry development kit (CDK) and several other Java packages, such as Jmol, WEKA, JavaStat and JFreeChart. Though our development is still at a primary stage, it supports assembly of new molecules through either a systematic combinatorial library routine or a stochastic evolutionary algorithm routine. Various *in silico* methods, such as docking, molecular similarity and QSAR are allowed to be adopted to direct the molecular design process in a structure-based way or a ligand-based way. The individual quality evaluations can be parallely implemented, thereby enabling large-scale optimizations. In addition to the main implementations, some complementary approaches, such as design of a fragment library, graphical visualization of a building block or a product set, and selection of an optimal structure subset are optionally provided. In contrast to many known *de novo* tools which are highlighted with their particular way of use, GeneGear intends to assist chemists in *de novo* design with multiple methods. In Chapter 3, an overview of the software and its functionality are illustrated.

Among the available *in silico* methods discussed with GeneGear, evolutionary algorithms (EAs) are a class of powerful optimization methods, which have high advantage in the achievement of *de novo* creation of novel chemical structures. However, without explicit constraints, an EA is hampered by sampling structures which are chemically undesired for different reasons. By applying data analytical methods from the fields of machine learning, chemometrics and multivariate statistics, a knowledge-based approach is proposed in Chapter 4, which allows a user to define his own filter (called the bias filter (BF)) using a set of available positive/negative molecules to constrain the EA output of molecules towards the desired positive structure space. The BF approach requires no explicit formulation of structure constraining rules and allows the possibility of building a filter where the user does not know the underlying rules for what constitute an "acceptable" structure, which makes itself much intuitive and user friendly.

Despite the wide spread of *de novo* design methods in medicinal chemistry,

automation and computer-aided synthesis have been comparably little appreciated in organometallic and coordination chemistry. Many of the available methods for drug design are not adapted to the structural variations of such type of compounds due to their poor construction rules in addressing the knowledge about the coordination center and the neighboring ligands. Chapter 5 describes a fragment-based EA method, which is developed with GeneGear specifically for *de novo* optimization of coordination compounds. The algorithm represents a 2D coordination structure as a graph with a "core-trial-free" part concept where three kinds of pattern-sensitive operations (growing, crossover, mutation) are used to sample candidate structures. Also, it permits a high flexibility in the fitness definition and allows parallel implementation of the individual fitness-generating calculations, potentially making it possible for large-scale optimizations. The capabilities of the EA method are illustrated by a series of representative searches for optimal ruthenium-based catalysts for olefin metathesis, where the fitness of the generated structures is assessed by a QSAR obtained at the semi-empirical PM6 level. The results show that our EA approach is likely to prove a powerful tool in searching for transition metal catalysts and other functional coordination compounds with optimal properties.

Contents

Chapter 1

Introduction

1.1 Background

The design of new and improved molecules with desired physical, chemical, and biological properties is an important but challenging task in the world of chemistry, biology and pharmaceutics; the designer has to identify a small number of suitable candidates from an essentially huge chemical space. The past two decades have witnessed the wide adoption of high-throughput screening (HTS) in the field of drug discovery, which enables large libraries of available compounds to be biologically screened at rapid rates for their ability to bind or modulate targets of interest [1–3]. With the use of highly automated, robotic techniques, although HTS in principle allows every available compound against every biological assay to be tested, it is still associated with a number of practical problems, for instance, the overall expense caused by acquisition of the sheer number of synthesized samples and the difficulty in assuring of the quality of screening libraries and HTS assays. In contrast to the expensive *in vitro* process of HTS, virtual screening (VS, also known as vHTS) performs a more cost-effective *in silico* task. By using a computational scoring scheme, the latter searches large compound databases to discover a limited number of candidate molecules that most likely possess activity against the biological targets of interest [4]. Basically, the VS methods can be divided into the two categories: structure-based screening (namely docking) and ligand-based screening. The former takes advantage of the three dimensional structure of the target macromolecule and models its key interactions involved in ligand binding to detect small molecules which have good steric and energetic fit with it [5–9]. The latter uses two- or three- dimensional similarity-based methods

[10–13] or pharmacophore models [14] to identify molecules that share common structural features with some known active ligand(s). VS based on database searching has great attraction in discovery of new lead compounds as the hits can be tested immediately and the molecules are usually synthetically feasible. Nevertheless, what the screening techniques provide are chemically or structurally available molecules which only represents a small fraction of the total possible drug-like molecules. The latter was estimated to be at least 10^{60} in the chemical space [15, 16].

Clearly, it may require to find new structures which possess activities but are different from the available molecules. The process of designing new molecules is called *de novo* design. In rational drug design, *de novo* design is regarded as one of the most important approaches, as well as virtual screening. While the latter 'finds' ligands fitting the binding pocket of the receptor from a large number of available compounds, the former 'builds' ligands by modification of available structures or assembly of pre-defined atomic or fragmental building blocks. To achieve structure manipulation, a computer-based molecular representation is needed. *De novo* design programs use various ways to virtually represent a 2D molecular structure, such as SMILES [17–19], connectivity matrix [20], linear [21, 22] or topological [23] tree, or graph [24–27]. Despite the advantages of other types of molecular encoding approaches, the graph-based representation methods are highlighted by their efficient chemical resemblance and flexible operability. The structure manipulation can then be guided by the two types of design strategies: structure- (or receptor-) based design and ligand-based design.

A classical structure-based computational design is to construct molecules directly within the binding site of the target protein. The quality of the designed molecules is typically evaluated with a force-field based, an empirical or a knowledge-based scoring function [8]. The molecular building process is usually concerned with 'linking' [28–37] functional groups pre-placed at favorable interaction sites in a molecule or incrementally 'growing' [28–30, 35–45] a structure from a 'seed' fragment under the shape constraint of the binding site, though it may also be driven by other concepts, such as lattice sampling [46–48], Monte Carlo simulation [43, 44, 49–54], and combinatorial scheme [45, 55–58]. A drawback of these building strategies is that the resulting conformations of molecules will almost always be at high energies which is not common under natural circumstances [59]. Furthermore, the orientations of the building blocks locally determined by the building procedure does not necessarily yield the optimal fit for the whole molecule. To overcome these problems, some relatively new

structure-based methods, such as SYNOPSIS [59], ADAPT [27], LEA3D [22] and AutoGrow [60], use a separate structure generator to build new molecules and feed the molecules into an docking software – e.g., FlexX [61], DOCK [62] or AutoDock [63] – for fitness determination. A weakness of these methods is that they have a higher tendency to produce false-positive ligands with reliable models of bound ligands. While on the other hand, they have an added benefit that the programs are easily adapted to different kinds of evaluation functions, so some two- or three-dimensional similarity-based methods and pharmacophore models can be combined with docking to reduce the number of false-positive candidates for fairly complex scoring calculations.

In contrast to a structure-based design, a ligand-based design does not rely on a three-dimensional structure of a particular biological target, instead, it starts from the structure(s) and knowledge of one or more known ligands and encodes them into a pharmacophore model [64], a similarity measure [24, 25, 65] or a QSAR model [19, 23, 66], to determine the fitness of the new generated molecules. Both a QSAR model and molecular similarity can not only be based on three-dimensional structures but also on topological structures of molecules depending on the applied descriptors. Moreover, properties other than biological activity, such as diversity, synthetic feasibility, and absorption, distribution, metabolic, excretion and toxicity (i.e., the so-called ADMET) properties, can be taken into account as well.

The second type of structure-based design represented by SYNOPSIS etc. shares similar building strategies with ligand-based design, where separate structure generators are used to achieve structure assembly; in fact, many of the generators are driven by an evolutionary algorithm [19, 23–25, 65–69]. Evolutionary algorithms, including genetic algorithms (GA) [70–72], genetic programming (GP) [73, 74], evolution strategy (ES) [75, 76], and evolutionary programming(EP) [77, 78], represent a powerful and general class of global optimization methods. Applied to molecular structure optimization, these methods use tailored genetic operators for new structure generation and perform population-based stochastic searches for optimal solutions to a given problem, providing a practical tool for investigating the potentially huge chemical space; for key reviews, see refs [79–82]. It is thus not surprising that a number of powerful methods for evolutionary *de novo* design have been developed [22–25, 27, 31, 59, 60, 65, 67, 83, 84]. EA does not make any assumptions about the underlying fitness landscape, this generality makes it suited for both structure-based design (e.g., LigBuilder [31], ADAPT [27], SYNOPSIS [59], LEA3D [22] and AutoGrow [60]) and ligand-based design (e.g., Chemical Gen-

esis [67], TOPAS [65], JavaGenes [24], the developments of Nachbar [23] and Brown et al. [25]).

Of course, both structure- and ligand-based design complement each other depending on the situation of the application cases. It would also be desirable to be able to easily switch between different "construction" and "evaluation" strategies, since the *de novo* molecular design is a dynamic activity where the perspectives and purposes of the investigators vary according to application cases, so the design strategies may change frequently. For instance, from a structure-based design to a ligand-based design, or from a stochastic search to a systematic construction. Unfortunately, most *de novo* softwares available for computer-aided molecular design (CAMD) do not support such a flexibility and they are usually concerned with one particular way of use rather than multiple approaches to a wide range of problems. Once the user changes his design strategy, it becomes difficult to rely on only one software tool. Of course, some tools are designed in a modularized manner and in theory do support a low-cost code modification, however, most of them are proprietary softwares which do not freely provide source code that can be modified by the user, though there do exist a few exceptions such as AutoGrow [60] and JavaGenes [24].

Another fact which we must pay attention to is that so far much computer-aided efforts has been spent designing new active compounds for the pharmaceutical industry. In contrast, computer-aided synthesis is much less widespread in non-medicinal chemistry. In particular, in organometallic and coordination chemistry field, contemporary catalyst development researchers still to a large extent rely on their general chemical knowledge and intuition to provide *ad hoc* qualified guesses for lead structures. The fact that automation and computer-aided synthesis have been comparably little appreciated in coordination and transition metal chemistry is, no doubt, the result of the many obstacles that this particular chemistry poses. Many of the traditional methods for drug design are not adapted to the structural variations due to the central atom in such compounds, in particular in the case that this atom is a transition element. For example, the coordination number and geometry may vary depending on the metal and its oxidation state as well as on the ligands, thereby complicating both the construction of rules for automatic structure generation and the optimization of parameters (e.g., parametrization in force field and semi-empirical methods) for calculation of molecular descriptors and properties. Moreover, the reactivity of coordination compounds may be influenced, or even to a large extent governed, by solvent and entropy, the effects of which may be hard and costly to describe computationally. So far, to the best of our

knowledge, no automatic molecular builders capable of handling coordination compounds are available.

To improve on the situation and make a contribution to the community, we have developed "GeneGear" - an open source software that integrates multiple strategies and supports further extension for *de novo* design of various compounds. GeneGear is a Java-developed platform which was built with the popular chemistry development kit (CDK) [85], and several other Java packages, such as Jmol [86], WEKA [87], JavaStat [88] and JFreeChart [89]. Virtual combinatorial library design and evolutionary *de novo* design represent its two main application implementations, which provide users respectively a systematic and a stochastic strategy for finding novel chemical structures. Whereas the former biases to a family-based design, the latter suits a more diverse chemical space exploration. In the present thesis, we have investigated both of these approaches though more focus will be put on the latter. Structures generated from both approaches are allowed to be quality (fitness) evaluated based on various systems, such as, a receptor–ligand binding energy predicted using a docking software, e.g., AutoDock [63] and AutoDock Vina [90], a molecular similarity by the scale of a set of molecular descriptors, or a QSAR/QSPR. The individual fitness evaluations can be distributed in parallel over multiple nodes on a cluster-type architecture, thereby enabling large-scale optimizations. In addition to the main application implementations, some complementary utility methods, such as design of a fragment library, graphical visualization of a building block or product set, and selection of an optimal structure subset are also provided.

1.2 Scientific objective

It is thus helpful to present scientists the convenient tool that is of multiple use to assist their *in silico* processes of molecular design. Further exploration of the usefulness of GeneGear in the designs of drug-like or catalyst molecules with various strategies is valuable.

As "molecular evolution" has been an important and popular approach in recent decades of *de novo* design, there is an interest to investigate its capability with different application cases.

Despite its advantage in creation of novel structures, EA is also apt to produce chemically undesirable structures. Penalty by fitness function or restriction of construction by explicit rules to eliminate the undesirable structures can sometimes be unfavorable when take into account the time and programming

cost. It is of interest to investigate the possibility of using some knowledge-based approach to constrain the search space of a random EA within some chemically meaningful space.

The development of *de novo* optimization methods so far has mainly been focused on designing new active compounds for the pharmaceutical industry. In contrast, computer-aided synthesis is much less widespread in non-medicinal chemistry at large and in organometallic and coordination chemistry in particular. A development toward more automation and less *ad hoc* guess-work in this field thus has a great potential and is likely to trigger important technical developments.

1.3 Outline of the thesis

The rest of this thesis is organized as follows:

Chapter two gives the reader a basic introduction about the main concepts of *de novo* design and how *de novo* design techniques are artistically applied in exploring chemical space for novel molecules of domain interest.

Chapter three introduces the general architecture and main functions of our new developed *de novo* design software GeneGear, the usefulness and capabilities of which in *de novo* design of functional drugs and catalysts are investigated through several well-studied application examples.

Chapter four focuses on investigation of a knowledge-based approach, which is built with the data analytical methods from the fields of machine learning, chemometrics and multivariate statistics and is used to constrain the structure space generated by an random EA within a chemically meaningful area.

Chapter five concentrates on a fragment-based evolutionary algorithm, which is facilitated with pattern-sensitive structural operations, quantum chemistry knowledge, QSAR analysis, and parallel scheme, and used to optimize functional coordination compounds; each compound in our concept is characterized as a molecular graph with "core", "trial" and "free" parts.

Final concluding remarks and further perspectives are given in Chapter five.

Chapter 2

The Art of *De novo* Design in Exploring Chemical Space

2.1 Chemical Space

"Space", as Douglas Adams famously described, "is big. You just won't believe how vastly, hugely, mind-bogglingly big it is". His remark gets similar resonance when applies to chemical space, which in the general sense involves the ensemble of all possible molecules like universe populated with stars. Despite its vastness, much of chemical space contains blank, lightless galaxies, which are no of interest. In a narrow sense, 'chemical space' can be related to a region defined by a particular choice of chemical descriptors and the limits placed on them [16]. In this case, it can be examined in a manner similar to the Mercator convention in geography, where rules are equivalent to dimensions (e.g., longitude and latitude), and structures are equivalent to objects (e.g., cities and countries) [91]. Anyway, 'chemical space' is a term for practical uses, infinite and limited only by chemists' imagination. In the context of this study, we refer 'chemical space' to a total descriptor space that encompasses all small carbon-based molecules that could possibly be created. Even with such a limited scope, the space is still vast; the total number of drug-like molecules has estimated to exceed 10^{60}, which only represents a small fraction of the total possible number of small carbon-based compounds with molecular masses in the same range in the space.

2.2 Screening Techniques vs. *de novo* Design

Given the enormous size of chemical space, the challenge for scientists is to identify small molecules that are of primary interest. So far, many efforts have been paid to the drug discovery field. The standard process of drug discovery is considered to be linear (see Figure 2.1 [92]). Study of a particular disease derives human knowledge about it. After that, a biological target associated with it is identified. High-throughput screens are then designed and subsequently compounds from drug-like chemical libraries are tested in the screens for their ability to modulate the target. Selected 'hits' (compounds that shows levels of activity beyond a certain threshold level) are further optimized through the testing of other screens (often lower throughput) to give leads that have required pharmacokinetic properties. In the following, *in vivo* tests start, where leads with required efficacy are further optimized and developed into candidate drugs, which are subjected to test by human clinical trials.

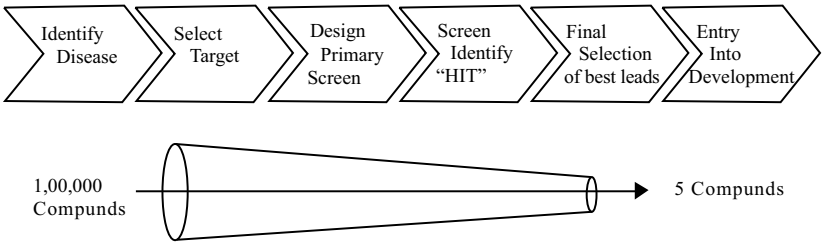


Figure 2.1: A drug discovery funnel [92].

The drug discovery field contains the most number of strategies and methods for exploring chemical space. Some of them are seen having counterparts in other fields, some are not. The investigation of this study thus by default directs to the issue of exploring the drug-like and biological relevant chemical space. And the motivation is to give the reader basic knowledge how strategies/methods from different categories and from *de novo* design in particular are being used to exploring the chemical space of interest.

Medicinal chemistry suggests that compounds that are considered to be functional with regard to specific categories of targets are clustered in discrete regions of chemical space [93]. Given this, what are the best strategies to navigate through the chemical space and direct our exploratory efforts towards the regions that are most likely to consist of molecules with useful activities/properties?

The current primary strategy used by pharmaceutical industry for identifying

small active molecules that might be starting points for potential drugs is the use of high-throughput screening (HTS). This is owing to its compatibility with the production of orally administered drugs. Through HTS, large collections of tens to hundreds of thousands of compounds are experimentally assayed for their ability to bind or modulate a biological target. However, the project itself is very costly – the compounds screened nowadays are often purely synthetic products from combinatorial chemistry as opposed to natural products used in the early era of pharmacology. Despite great advances in parallel synthesis and high-throughput screening technology, the number of compounds that can be created and tested in a reliable manner is only a tiny fraction of all the molecules of potential pharmaceutical interest. Moreover, the process is fundamentally based on "trial-and-error", which is prohibited by the vastness of chemical space.

An alternative approach, known as virtual screening (VS) is to computationally screen large libraries of molecules for compounds that complement binding sites of known targets [5, 6]. In this case, VS is a docking-based method. VS faces several fundamental challenges, including sampling various conformations of flexible molecules and calculating absolute binding energies with solvent effects. As a result, it is plagued by false-positive and false-negative predictions. Even with limitations, the field has experienced important successes, where new ligands are predicted along with their receptor-bound structures with hit rates significantly greater than with high-throughput screening [94]. Also, by the strategy of using libraries of accessible, often purchasable compounds, it saves synthetic effort required by HTS, and thus in principle allows even larger numbers of compounds to be processed against receptors for which structures are available at little cost. Thus, for those who can tolerate its false-positive and false-negative predictions, virtual screening offers a practical route to finding some interesting ligands from the known chemical space. However, by restricting itself to available compounds, VS avoids broad searches of chemical space, and thus cannot produce structurally novel molecules from the unexplored chemical space.

Molecular *de novo* design [1] produces novel molecular structures with desired properties from scratch [82, 96]. In this approach, what a medicinal chemist – and equally a *de novo* molecule-design software – faces is a virtually infinite search space, including at least 10^{60} drug-like molecules, from which the most

[1] *De novo* design is the design of bioactive compounds by incremental construction of a ligand model within a model of the receptor or enzyme active site, the structure of which is known from X-ray or nuclear magnetic resonance (NMR) data [95].

promising candidate structures have to be discovered. Such a large space prohibits an exhaustive search by any program because of the combinatorial problem. Instead of systematic searching chemical space, the *de novo* design process relies on the principle of local optimization: a *de novo* search program may randomly start from one or more discrete points in chemical space to navigate through their neighborhoods with the pressure of human knowledge, and finally converge on some local or practical optima. It may also employ some human-instructed systematic heuristics to determine what are the local best solutions.

2.3 Focuses of *de novo* Design

Applying computer programs to design of molecules seems an obvious choice since the computer can create virtual molecules much faster than human can. However, prior to any exploring step, a *de novo* design program must address precisely the following three issues: how to assemble the candidate structures; how to estimate their potential quality; and how to sample the search space effectively and efficiently [82]. The chemical space is full of molecules, each of which is a non-linear structure implicitly following rules of chemistry, e.g. every oxygen atom has two bonds, and every carbon atom four. Of course, there can be further constitutional constraints imposed by chemists for different aims. How to represent the structures in an appropriate way and generate chemically desirable rather than problematic structures is the main concern of the first question. Second, the outcome of a *de novo* design program based on any sampling algorithm should be computationally assessed of their quality. The quality evaluation function (often referred to as objective function) directs the sampling program towards more competitive structures. Although, most of the time, the definition of quality evaluation function is really depending on the investigators, it is important to use methods that predict the molecular properties reliably and time-efficiently. However, it is impossible to search systematically the set of all possible molecules. One reason is that the number of possible structural variations rises very fast with the size of the molecule. If one considers a possible variation a search step in chemical space, the number of dimensions orthogonal to the step axis increases with the number of atoms in the molecule. So a 'normal' drug molecule, which may contain e.g. 20 non-hydrogen atoms, and each atom of which has 10 possible variations, will lead to a search space of 20 dimensions with 10^{20} sampling points. This makes a systematic examination of all possible molecules be quite difficult. Thus it is crucial to make some meaningful reduction of the search space.

2.4 Principal Constraints

Chemical space is vast but most of it is not relevant to the problem being studied. The limits of bioactive chemical space are concerned with specific interactions between small molecules and the three dimensional molecular recognition patterns on particular biological targets [93], which form 'primary target constraints' for *de novo* searched candidate structures. Such constraints can be derived either from a three-dimensional target structure or some known ligands of the particular target. When the former is consulted, it is receptor-based design, otherwise, the design strategy is ligand-based. Receptor-based design starts with the binding site, the potential complementarity of which in molecular shape and physicochemical properties are substantially important for specific binding. Thus the binding site is used to derive shape constraints for ligands, as well as specific non-covalent ligand–receptor interactions. The latter in the form of interaction sites [2] (see Figure 2.2 [82] for an example) define strong and explicit requirements for successful receptor–ligand binding and have a major role in the effort to reduce the vast number of possible structures. They can be further subdivided into hydrogen bonds, electrostatic and hydrophobic interactions. hydrogen-bonding interactions are of special interest owing to the stable chemical match between hydrogen-bond acceptors and donors, and often form key interaction sites.

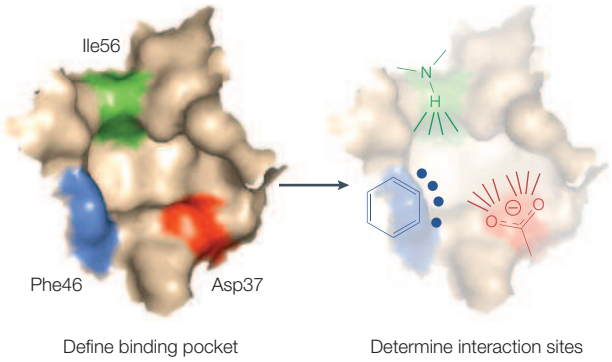


Figure 2.2: On the basis of an X-ray model of the binding pocket (PDB-identifier: 1FKF) of FK506-binding protein (FKBP-12 [97]) interaction sites were identified. Selected interaction centres are indicated by blue dots (lipophilic), green (acceptor) and red (donor) lines [82].

There are different ways to derive interaction sites from the three-dimensional

[2] An interaction site is a position in space that is not occupied by the receptor and in which a ligand atom favorably interacts with the receptor [82].

structure of the binding pocket. HSITE [98] is the first rule-based method, which considers hydrogen-bond acceptors and donors and outputs the map of hydrogen-bonding regions of the receptor. The regions are centred at ideal hydrogen-bond geometry which is empirically determined from crystal structures of available small molecules. Some later developed rule-based methods [28–30, 57, 58] made extension to this by the addition of such as the lipophilic interaction sites [28–30, 57], and interaction sites of covalent bonds and bonds to metal ions [58]. Grid-based approaches tackle this problem in another way. Typically the approaches generate a grid of points in the binding site and place different probe atoms or fragments with respect to, such as hydrogen-bonding capabilities or lipophilic properties, at each grid point for the determination of favorable interaction site and computation of interaction energies. Several *de novo* design programs [28, 29, 41] perform the software GRID [99] to achieve the above process, while others, such as LigBuilder [31], have their own implementation of this algorithm. The resolution of the grid decides the performance of grid-based methods. Although higher resolution leads to more grid points and more accurately calculated regions of the interaction sites, it also requires greater computational cost, so a compromise is necessary. In addition to indicating favorable positions of specific functional groups in the binding site as what rule- and grid- based methods have done, the Multiple Copy Simultaneous Search (MCSS) [100] method yields a set of pre-docked fragments in energetically favorable orientations at these positions. The basic scheme is as follow: first, multiple copies of fragments are randomly placed inside the binding pocket; all groups are then minimized in energy using a force field and the forces among them are neglected; next, groups that have interaction energy with protein above a certain threshold are discarded. An MCSS run leads to multiple outcomes that can be further inspected for the most promising ones [32, 50]. Similar effect can be achieved by the use of a docking software, which constructs a component of some *de novo* design programs [22, 27, 60].

In addition to optimizing properties associated with binding affinity/biological activity, properties such as absorption, distribution, metabolism, excretion, (the so-called ADME) that are important for the efficacy of a drug also need to be carefully considered. The concept of 'drug-likeness' is related to the characteristics (such as oral absorption, aqueous solubility and permeability) of compounds that are more likely to yield safe, orally bioavailable medicines. It was introduced by Lipinski's seminal analysis of the Derwent World Drug Index. The analysis shows that orally administered drugs are more likely to reside in areas of chemical space defined by a limited range of molecular properties [93]. These properties are described in Lipinski's 'rule of five' [101]

which says that, in general, an orally drug-like molecule meets the following criteria: 1) no more than five hydrogen-bond donors, 2) no more than ten hydrogen-bond acceptors, 3) a molecular mass no greater than 500 daltons, and 4) a log P value (a measure of lipophilicity) of no greater than five [3]. Since 'rule of five', more methods to predict drug-likeness have been proposed. They can be used as simple rules of thumb for restricting *de novo* search space to regions possibly enriched in 'drug-like' and 'lead-like' molecules.

Another factor which is of major importance is the ease of synthesis. So far, most *de novo* designs guarantee the validity of their searched structures by strictly complying with the basic chemical valence rules. However, this does not ensure the generation of chemically reasonable and stable structures. The issue of unstable chemical groups still can be addressed with a set of construction [20, 26, 84] or filter [31] rules. But what are more difficult to handle is the synthetic feasibility. There is no guarantee that a stable molecule retrieved from domain relevant are synthetically feasible to produce. Even though this problem was not newly recognized in the history of *de novo* design, it got intended address only recently. The common idea of solutions toward this problem by a small number of *de novo* design programs is to assemble the building blocks in agreement with some virtual organic reaction schemes. For example, the building blocks used for assembly of candidate structures by TOPAS [65] are resulted from retrosynthetic combinatorial analysis [102] of a set of common organic reactions. LigBuilder [31] uses basic chemical fragments, all of which can be classified into chemical groups and rings. PRO_SELECT [57] adopts idea from both structure-based drug design and combinatorial chemistry and selects potential substituents for a synthetically accessible template positioned in the active site of target of interest. In the approach of SYNOPSIS [59], building blocks are molecules that are selected from a starting database, and a set of chemical reactions is used as templates for synthesis of candidate structures. The reactivity of a reaction associated with a particular building block is estimated on the basis of a series of additional rules for acceptance of such a reaction, e.g.: an NH2 moiety can be oxidized to an NO2 moiety, but not when it is part of an N-NH2 moiety; an H-N-C=O moiety can be reduced to an H-N-C moiety only in absence of C=S moiety; an aliphatic halogen atom is preferred more than an NH2 moiety to an aromatic one. Alternatively, some external softwares, such as CAESA [58] and SEEDS [103], can be applied for assessing the synthetic accessibility of candidate compounds.

Nevertheless, the process of designing pharmaceutically effective molecules re-

[3] All numbers are multiples of five, which is the origin of the rule's name.

quires the generation of complex structures with many constraints. Constraints other than the binding affinity are usually referred to as secondary target constraints. The final score of a *de novo* proposed structure is sometimes calculated as a weighted sum of estimated binding affinity and other objective terms. The latter directly or indirectly contribute to the secondary constraints.

2.5 Scoring Function

The application of a *de novo* design program leads to a number of newly generated structures which are supposed to be scored with their quality, so the most promising ones can be suggested. The scoring function gives fitness value to the sampled space and guides the design process efficiently through the search space. Basically, there are receptor-based and ligand-based scoring functions as described below.

2.5.1 Receptor-based Scoring

Receptor-based scoring functions are concerned with evaluation of the potential complementarity of a candidate structure to the binding site of a target and the receptor–ligand binding affinity. The very early *de novo* design programs [33, 38, 39, 46–48, 104] only applied steric limits to guide the search, but more approaches to estimating binding free energy [105] emerged in the following years. The latter can be roughly divided into three families [8, 106]: force-field based, empirical, and knowledge-based scoring functions [4]. Force-field based methods are most computationally costly because of slow molecular dynamics (MD) or Monte Carlo (MC) simulations. The force field-based energy functions usually consists of receptor–ligand (inter) and internal ligand (intra) interaction energy, which basically includes van der Waals and electrostatic terms. The van der Waals potential energy is for general estimation of non-bonded interactions, and often modelled by a Lennard–Jones 12–6 function, as described in Equation 2.1:

$$E_{\mathrm{vdW}}(r) = \sum_{j=1}^{N} \sum_{i=1}^{N} 4\varepsilon \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.1}$$

where ε is the well depth of the potential and σ is the collision diameter of the respective atoms i and j.

[4] Empirical functions and knowledge-based potentials can be understood as very general 3D-QSARs. They differ from the usual QSAR in two points: First, the training set contains many different ligand types that bind at different receptors. Second, the structure of the receptor itself is also used in the derivation of the model [106].

The electrostatic potential energy is depicted as a pair-wise summation of Coulombic interactions, as shown in Equation 2.2:

$$E_{\text{coul}}(r) = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \frac{\sigma_{q_i q_j}}{4\pi\varepsilon_0 r_{ij}} \tag{2.2}$$

where N is the number of atoms in molecules A and B, respectively, and q the charge on each atom.
Additional terms such as bond, angle and torsion may be included as well. Standard force-field scoring functions have limitations in capturing solvation and entropic effects. They are further complicated by the requirement of cut-off distances for the treatment of non-bonded interactions, which more or less affects the accurate treatment of long-range effects involved in binding. LEGEND [45] was the first program that implemented a force-field based scoring function. And quite a few others [40, 41, 45, 52, 84, 107, 108] follows over the years.

Empirical scoring functions are a weighted sum of individual ligand–receptor interaction terms, such as hydrogen bonds, electrostatic and hydrophobic interactions, commonly supplemented by penalty terms, such as the number of rotatable ligand bonds. The weights correspond to free-energy contribution of interaction of each type. They are obtained from a regression analysis of experimentally determined binding energies and, potentially, X-ray structural information of a list of receptor–ligand complexes. The first *de novo* program that implements empirical scoring function was LUDI [28, 29]. Several other *de novo* programs [28–31, 42, 49, 57, 109] followed the same concept. Empirical functions contain many individual contributing terms which have counterparts in force-field molecular mechanics terms, however, here they are more efficient to evaluate, which makes the methods attractive. A disadvantage of these methods is their dependence on available datasets that are limited in size and feature similar ligands and receptors. This can cause their bias towards specific structural motifs, and prevent them from more general use.

Knowledge-based scoring functions are designed to reproduce structural information rather than binding energies, for which a number of atom-type interactions are defined depending on their molecular environment. Protein–ligand complexes are modelled using pair-wise atom potentials, so the binding effects can be implicitly captured. Implementations of such functions can be found in Potential of Mean Force (PMF) [110, 111] and DrugScore [112]. The latter also includes solvent-accessibility corrections to pair-wise potentials. The *de*

novo design program SMoG [43, 44] has its own implementation of this type of scoring function. It performs a statistical analysis on a set of ligand–receptor complex structures to determine the frequencies of each possible pair of atoms in contact with each other. Interactions found to occur more frequently than would be randomly expected are considered favorable, otherwise they are considered unfavorable. Binding free energy is represented as a sum of free energies (or, equivalently, potentials of mean force) of interatomic contacts that are calculated from their frequencies. Only structural information is needed to derive these frequencies, so a greater number of structures, not limited to those with known binding affinities, can be included in the analysis. The major advantage of knowledge-based functions also lies in their computational simplicity. The disadvantage is that their derivation is based on the implicit information that are extracted from limited sets of protein–ligand complex structures.

2.5.2 Ligand-based Scoring

The three-dimensional structure of a biological target can be facilitated to examine the steric and energetic fit of new candidate structures to the binding site. However, if such a three-dimensional structure is not available while one or more binding molecules are known, ligand-based design can be used. In contrast to receptor-based strategy which is inevitably confronted with the problem of conformational complexity, a ligand-based strategy can focus on either the three-dimensional or the topological structure of one or more known ligands.

A way to use information inherent to known active compounds is derivation of a three-dimensional pharmacophore model [5]. Once established, it can either be used to obtain a pseudo-receptor model or act as a direct 3D similarity template [64]. Whereas the former guides the design of structures towards those that are complementary to the constraint, the latter directs to structures that are similar to the constraint. To build a three-dimensional pharmacophore model, all applied ligands should be aligned in advance in a common binding mode. The generality of the model improves with an increase of structural diversity in the applied set of known ligands. Alternatively, a set of known ligands can be applied for development of a quantitative structure–activity relationship (QSAR) [6] model style scoring function [19, 54, 66]. Another approach

[5] A pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response [95].

[6] A quantitative structure-activity relationships is a mathematical relationship linking

is the computation of molecular similarity between a known active compound and the candidate compounds [24, 25, 65]. This is based on the calibration of a set of molecular descriptors. Both a QSAR model and similarity measure can be based on either three-dimensional structures or topological structures, depending on the choice of the descriptors.

2.5.3 Multiobjective Scoring

As noted already, a truly effectual drug molecule is more than a bioactive structure. Physicochemical properties of molecules that determine effects such as ADME, are also important, as well as other factors such as cost and synthetic feasibility. Consequently, the focus in *de novo* design can be shifted toward the generation of more complex structures within the constraint of multiple objectives. Even though these objective may be suboptimal in the single objective sense, they might be competitive and conflicting with each other when they meet together. For example, a molecular population designed on diversity alone has a tendency to contain molecules existing in non-druglike regions of chemical space, e.g., molecules with high molecular weights. Thus, there is a need for the search for solutions that offer acceptable performance in all objectives. Typically, optimization techniques formulate multiobjective problems as a weighted-sum scoring function, such as follow: $f(p) = w_1p_1 + w_2p_2 + ... + w_np_n$, where p_n is the nth property and w_n the nth weighting coefficient. Combining multiple objectives via a weighted-sum fitness function produces a single compromise. However, such an approach is often not desirable as it is not always easy to decide the appropriate weights. Furthermore, the fitness function determines the regions of the search space that are explored, and combining objectives via weights can result in that some regions not being explored. In the absence of additional information, a multiobjective optimization takes all applied objectives as equivalent. In such a case, there is a hypersurface existing in the search space that represents a continuum of alternative solutions, each of which corresponds to a compromise or tradeoff between the various objectives. Such a hypersurface is termed a Pareto frontier or a tradeoff surface and the solutions that are part of it are termed non-dominated or Pareto solutions. Multiobjective optimization seeks a set of non-dominated solutions rather than a single solution. A solution is non-dominated when there is no another one which is either equivalent or better in all the objectives and, better in at least one objective than it. In other words, one solution dominates

chemical structure and pharmacological activity in a quantitative manner for a series of compounds. Methods which can be used in QSAR include various regression and pattern recognition techniques [95].

another if it is either equivalent or, better, in all the objectives and, strictly, it is better in at least one objective [68].

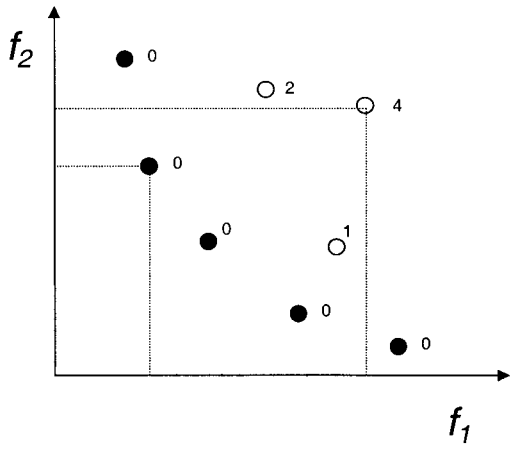


Figure 2.3: Potential solutions in a two-objective (f1 and f2) problem. The circles represent pairs of values reflecting the two objectives. The solid circles are non-dominated solutions and fall on the Pareto frontier. Unfilled circles are dominated solutions. Individuals are ranked by the number of times they are dominated; thus the non-dominated solutions are ranked zero and the dominated solutions are ranked as denoted [68].

The Multiobjective Genetic Algorithm (MOGA) [113] attempts to map out the hypersurface in the search space where all the solutions are treated as equivalent without the need for normalization. In MOGA, the ranking of the population is based on the number of times each solution is dominated (also known as Pareto ranking), as illustrated in Figure 2.3 [68] Pareto ranking allows the Pareto frontier to be mapped out by the population by evolving multiple non-dominated solutions simultaneously. The MOGA algorithm has been successfully applied to a number of problems in virtual combinatorial library design [68, 114, 115] and *de novo* design [25, 69]. Optimal solutions are the ones that are on the final Pareto frontier, which represent the most appropriate compromises of the individual objectives for the task.

2.6 Structure Assembly

2.6.1 Molecular Representation

To sample chemical space, what must be first considered is how to represent the chemical structures appropriately. A molecular structure can be represented

in one (e.g. a SMILES [116, 117] string), two (e.g. a topological tree [23] or undirected graph [24–26] or three dimensions (e.g. a file of atom coordinates and bond connectivities). Line notations like SMILES, are very compact for storing, retrieving, and communicating information, but the manipulation thereof by possible structure operations may gives rise to chemically invalid structures violating basic valence rules. Topological representations, such as a tree or graph, have good resemblance of constitutional digram of a chemical structure, however, it is well known that the properties of a chemical structure are highly dependent on the three-dimensional structure, so at some stage a realistic 3D structure is often needed to be generated. Conformational representations do the best in description of the detailed information of chemical structures, while this may also mean that they are more complicated to modify. In a word, each class of representation has its own advantage and disadvantage, which one to use is according to the actual condition of the application.

2.6.2 Building Blocks

Both atoms and fragments can used as basic building blocks for the assembly of candidate structures. Atom-based approaches are superior to fragment-based methods in generation of a diverse structure space. But at the same time, they are confronted with a much larger search space which contains more structures with chemically unacceptable constitutions. This hinder the speed of atom-based methods in finding suitable candidate compounds. Fragment-based strategies, on the other hand, reduces the combinatorial problem largely. This reduction is usually preferred, especially fragments are used that have common occurrence in drug molecules. Moreover, the definition of fragment is not that rigid: it can vary from an atom to a polycyclic ring system. As the problem coupled with pure atoms as building blocks became more and more distinct, today, it is more general that fragments with more than one atom are used as building blocks which of course can be padded with a few single-atom fragments.

2.6.3 Structural Operation

There are a list of concepts of structural operations with respect to assembly of an ensemble of new structures, such as linking, growing, random sampling, lattice-based sampling, sampling driven by molecular dynamics (MD), crossover and mutation.

The growing program [28–30, 35–45] starts with a seed fragment that is prepositioned at one of the key interaction sites of the receptor (Figure 2.4) [82].

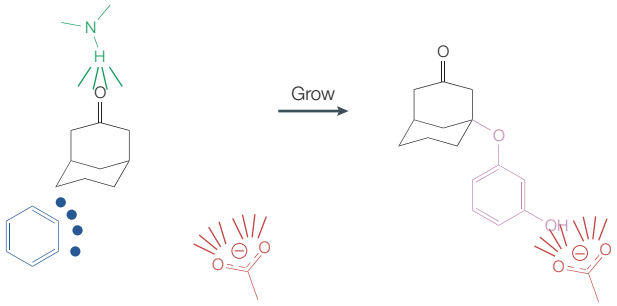


Figure 2.4: Illustration of the 'growing' concept based on the example of FK506-binding protein (FKBP-12, Figure 2.2) [82].

The seed structure gets incrementally grown within the binding site, with each suggestion being evaluated with its steric and/or energetic fit with the site. The growing strategy can run into difficulties if the active site contains more than one distinct (sub)pockets and/or if the seed is too small compared to the binding pocket. Due to the combinatorial nature of the search space, which is also associated with a variety of conformations for each single topological solution, it is not easy to suggest a molecule that fits to each part of the binding pocket. The strategy is more suitable for lead optimization. In this case, one can take the framework of a known lead compound as the seed structure and let the program build the remaining residues. Since the framework has occupied the principal part of the binding pocket, the problem of insufficient sampling will become less severe. A good example is PRO_SELECT [57] which introduced the idea of combinatorial chemistry to the growing method for the consideration of synthetic feasibility.

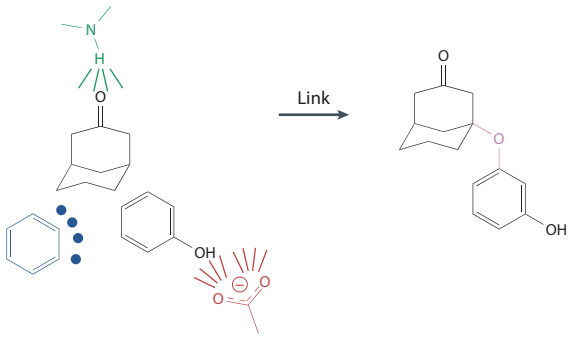


Figure 2.5: Illustration of the 'linking' concept based on the example of FK506-binding protein (FKBP-12, Figure 2.2) [82].

The linking procedure [28–37] commences with several fragments that are pre-placed independently and favorably at some key interaction sites of the receptor (Figure 2.5 [82]). The pre-placed building blocks are joined by automatically constructed or searched linkers to derive a complete molecule satisfying all key interaction sites. The linking strategy allows maximum ligand–protein interactions at the very beginning with the placement of suitable chemical fragments at optimal positions. However, linking different fragments together is not easy – any slight misplacement of fragments or fragments with no desirable spatial orientation can lead to ambiguous structures.

The lattice-based strategy concerns placement of an atomic lattice in the binding site. The lattice can be constituted by randomly and evenly distributed atoms [48, 51], regularly assigned sp^3 carbon atoms (diamond lattice) [46, 50], or pre-docked fragments [47]. Atoms in the vicinity of different interaction sites are joined by the shortest path among them and connected by newly formed bonds to yield new molecules [46–48]. Alternatively, atoms are joined to each other and to functional groups [50] by Monte Carlo simulation [50, 51] to form new molecules with the guidance of stereochemistry and potential energy minimization.

In molecular dynamics (MD) simulation, a user-specified set of molecular fragments (or atoms) are allowed to move independently about a fixed target active site. This allows the fragments to sample various low-energy orientations. When the geometries between pairs of fragments are appropriate, bonds can be formed between them to result in larger molecules. The formed bonds can in subsequence be broken in order to form energetically more favorable connections to different fragments. The connectivity of fragments gets iteratively improved over the course of the simulation, leading to a number of energetically favorable molecules. CONCEPTS [49] was the first such program, CONCERTS [52] and DycoBlock [107, 108, 118] followed.

Different from the above structure-based design strategies which construct ligands directly in the binding site of a target, ligand-based *de novo* design is not directly provided with the constraints of interaction sites. The majority of ligand-based approaches feature an evolutionary algorithm to optimize the topological molecular graphs. EAs have several optimal properties which make them attractive in *de novo* design, the most important perhaps being the ability to perform well in a search space which is large and incompletely understood. The choice of EA for chemical space sampling implicitly makes genetic operators (e.g., crossover and mutation) be responsible for the rendition of molecular diagrams. A crossover operator does an 'inter-breeding' (see

Figure 2.6 [26]), wherein 'genetic materials' (substructures) of two 'parent' individuals are exchanged and combined to generate new 'child' individuals. A mutation operator while on the other hand tends to introduce in an external variation (see Figure 2.7 [26]), wherein some small and local changes, such as changing an atom element type or bond order, adding or removing a fragment, are made to a present individual to create a new individual.

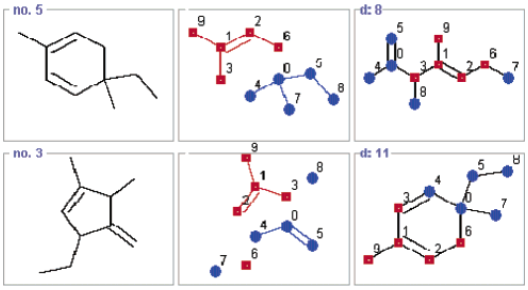


Figure 2.6: Crossover operation on a pair of structures with molecular formula $C_{10}H_{16}$ [26].

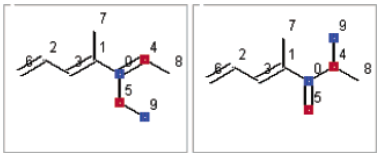


Figure 2.7: Mutation operation [119] on a structure with molecular formula $C_{10}H_{16}$, with the parent structure (left) and the offspring structure (right) [26].

Among the existing EA methods, JavaGenes [24] implements a multiple-point crossover operator which allows operations on edges of rings. TOPAS [65] uses a mutation operator that substitutes whole fragments, obeying the rules of virtual chemical reaction schemes. Nachbar [23] and Brown et al. [25] developed both crossover and mutation operators in their implementations. Chemical Genesis [67] applies crossover and mutation operators to the three-dimensional molecular structures and adds novel mutations, e.g., translation and rotation of molecule, and rotation about a bond. Ligand-based methods generally do not make use the three-dimensional structure of a target to evaluate the overall quality of generated ligands. But it is possible to use an EA structure generator like TOPAS [65] in combination with a heuristic 3D conformer builders like CORINA [120, 121], CONCORD [122, 123] and feed the designs into fast docking programs – e.g., FlexX [61], DOCK [62], or AutoDock [63] – for a

structure-based quality (fitness) evaluation. The program SYNOPSIS [59], ADAPT [27], LEA3D [22] and AutoGrow [60] are representative versions following this approach.

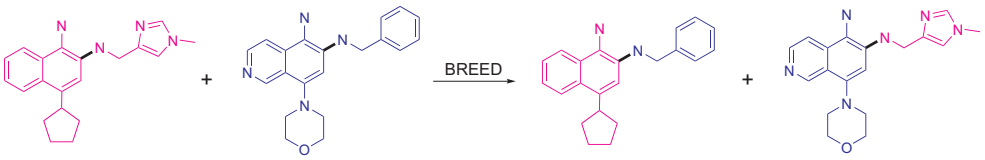


Figure 2.8: Illustration of the steps carried out by BREED to generate new ligands. The bonds that are distinctly colored in black are matching bonds.

BREED [124], a novel *de novo* design program, utilizes a special 'crossover' (recombination) idea. In this approach, a set of known ligands for a particular target in their three-dimensional active conformation is superimposed in a common reference frame by the overlay (alignment) of their target crystal structures. The overlapping bonds in all pairs of ligands are found, and the fragments on sides of each matching bond are swapped to generate novel molecules, as depicted in Figure 2.8. The method can be carried out in a recursive manner, so that the resulting offspring ligands are added into the pool of known actives and join in subsequent cycles of recombination. BREED [124] can only be applied in cases where several ligands are already known and structurally characterized. As the known ligands are pre-superimposed in their active conformation, functional groups that are known to bind the target will be present in the novel ligands in precise position and orientation.

2.7 Space Search

The vastness of chemical space is essentially caused by the combinatorial problem. Now, consider the chemical space of a normal hexane (n-hexane), each hydrogen atom of which could be substituted for another element type or chemical group. The use of a list of only 150 substituents in consideration of mono- to 14-substituted hexanes will lead to more than 10^{29} possible derivatives of n-hexane [125]. It is thus not difficult to imagine the infiniteness of the entire chemical space. To sample the search space effectively, *de novo* design has to tackle the problem of combinatorial explosion.

2.7.1 Combinatorial Search Strategies

The ideal algorithm for combinatorial search is expected to be able to solve the problem with provably few search steps and with provably optimal solution

quality. However, this is usually not realistic due to a potentially huge search space. Combinatorial search strategies offer a practical solution by lowering their expectations on one or both of the two aims. They solves instances of combinatorial problems by reduction of search space and possible use of efficient heuristics. The strategies are basically divided into breadth-first search (BFS) and depth-first search (DFS), though there can be some alternatives such as breadth-bounded search [126]. While a breadth-first search systematically goes through every solution along with every level of a search tree (or graph) (see Figure 2.9(a)), a depth-first search explores the search tree (or graph) as far as possible along each branch before backtracking (see Figure 2.9(b)).

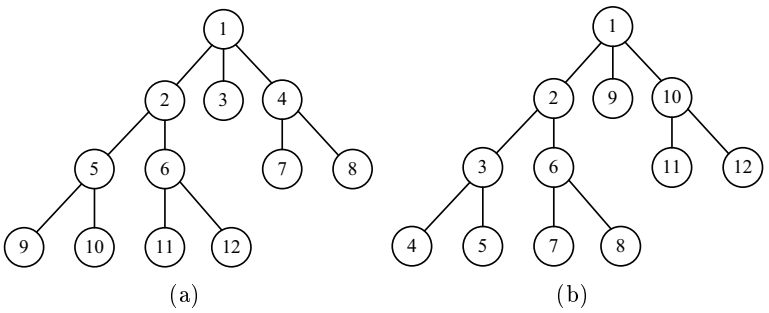


Figure 2.9: Order in which the nodes are visited in (a) breadth-first search and (b) depth-first search, respectively.

A few of *de novo* design programs implement either a breadth-first strategy [33, 35–37, 40, 47, 48] or a depth-first strategy [30, 35–39, 41, 46, 47]. A breadth-first strategy conserves all possible partial solutions at each level of the search tree and explores, sequentially, other levels until the end. Since breadth-first search performs a systematic examination, it ensures identification of the optimal solution given enough search steps, but it typically has large memory requirements. A depth-first strategy however retains only one of the possible partial solutions at each level of the search tree. Depth-first search does not guarantee to find the overall best solution, even if the best partial solution is selected each time, but it is able to reduce the search space significantly.

To make a combinatorial search feasible in a *de novo* design, the level of compromise between the requirements of 'optimality guarantee' and 'space economy' should be determined. Most of the programs that implement the breadth-first strategy use a linking method for structure assembly. Given that

the key interaction sites are already occupied with favorable fragments, only a smaller problem space (e.g., small set of fragments) is expected to be explored, which makes it possible to run an exhaustive search. RASSE [40], however implements a breadth-first search with an atom-based growing approach, whose search space is limited with the only 100 best partial solutions being selected at each growing step. SPROUT [35–37] and its predecessor [38] use depth-first search in combination with the A* search algorithm – a particular type of best-first search which estimates the cost of reaching a specific partial solution and suggests where to end up. Another possible change of depth-first search can be random selection of a partial solution at each tree level among all partial solutions [109] or among highest-scoring partial solutions [39, 41]. An incorporation of a breadth-first and a depth-first search is observed in the refined version of SPROUT [58].

2.7.2 Monte Carlo and the Metropolis criterion

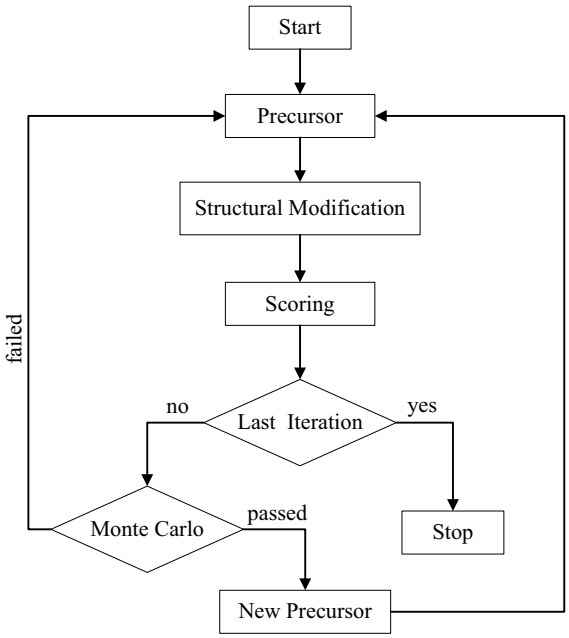


Figure 2.10: Flowchart of the main steps of Monte Carlo simulation in *de novo* design.

Many computer algorithms are said to use a 'Monte Carlo' simulation, meaning that some kind of random sampling is employed. Monte Carlo simulation generates configurations of a new molecule by making random changes

to the structure (e.g., topology, conformation, charge, hybridization) of a current molecule. Monte Carlo search alone is implemented with the program LEGEND [45]. In most instances of *de novo* design, a Monte Carlo search is combined with a Metropolis criterion, as is the simplified flowchart shown in Figure 2.10. A new molecule is generated after structure-modification of a precursor molecule. The new molecule is then scored and decided whether it is accepted or rejected. If it scores better than the precursor, it is immediately accepted. If it scores worse, it can still be accepted if it passes a test form such as the one below [54]:

$$\exp(-\beta\Delta) \geq \rho \tag{2.3}$$

Where β is a user-defined constant (≥ 0) whose value determines the strictness of the Monte Carlo procedure, Δ is the difference between the score of the new molecule and the precursor and ρ is a random number between 0 and 1. Once the new molecule get accepted, it becomes the new precursor.
The first *de novo* design program that made use of a Monte Carlo search together with the Metropolis criterion is CONCEPTS [49]; several others followed [42–44, 50–54, 127].

2.7.3 Evolutionary algorithms

Mining desirable structures from chemical space concerns a 'navigation' problem – one has to take a route through areas where less desirable compounds reside. The chemical space is very large, making it complex to navigate through. Thus, there is much interest in the development of effective, heuristic algorithms for search and optimization. Evolutionary algorithms (EAs), which can be further subdivided into genetic algorithm (GA) [70–72], genetic programming (GP) [73, 74], evolution strategy (ES) [75, 76], and evolutionary programming(EP) [77, 78], are a class of stochastic search algorithms which are proving able to provide optimal, or near-optimal, solutions to a wide range of challenging problems in a variety of disciplines. They are based on the concepts of Darwinian evolution [128]. Natural evolution produces organisms, whereas EAs seek to mimic natural evolution's ability to produce functional objects (e.g., structures, parameters and programs) by the use of analogous mechanisms: reproduction, mutation, recombination (crossover), and selection.

A typical EA starts with an initial population of proposed solutions. Each solution is evaluated by a fitness function (scoring system) and assigned with a score value indicating how well it solves a given problem. In *de novo* design case, the population is a set of small molecules (also in abstract referred to

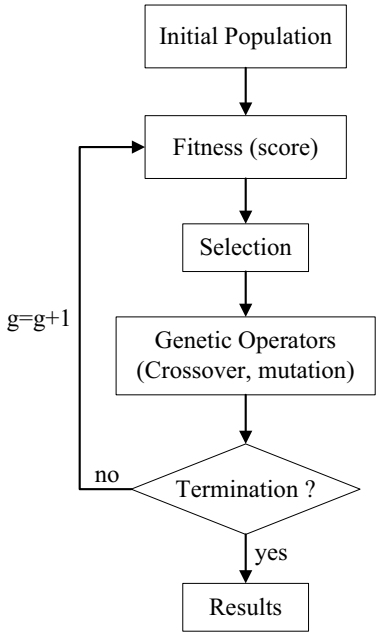


Figure 2.11: General flowchart for an EA [22].

as chromosomes or genotypes in GA) and the problem is often desired activities or properties defined by a fitness function. The population evolves over generations by the application of reconstruction operators like crossover and mutation, on selected solutions. The selection of solutions for reconstruction can be based on various mechanisms such as ranking, roulette wheel and tournament [129]. The selection of solutions, which are allowed to survive from one generation to the next, is inclined to those of better fitness. The algorithm terminates when predefined conditions are met, e.g.: sufficiently good solutions have been found or solutions are no more improved. A general flowchart of EA is shown in Figure 2.11. EAs have been widely used for investigating the potentially huge chemical space in *de novo* design area; for key reviews, see refs [79–82]. By default, an EA makes no assumptions about the fitness landscape, this generality makes it suited for both structure-based design (e.g., LigBuilder [31], ADAPT [27], SYNOPSIS [59], LEA3D [22] and AutoGrow [60]) and ligand-based design (e.g., Chemical Genesis [67], TOPAS [65], JavaGenes [24], the developments of Nachbar [23] and Brown et al. [25]). Most of these applications involve the chemical structure search in a constitutional space except few examples in which EAs are adopted to explore the conformational space [60, 67].

2.8 Application Examples

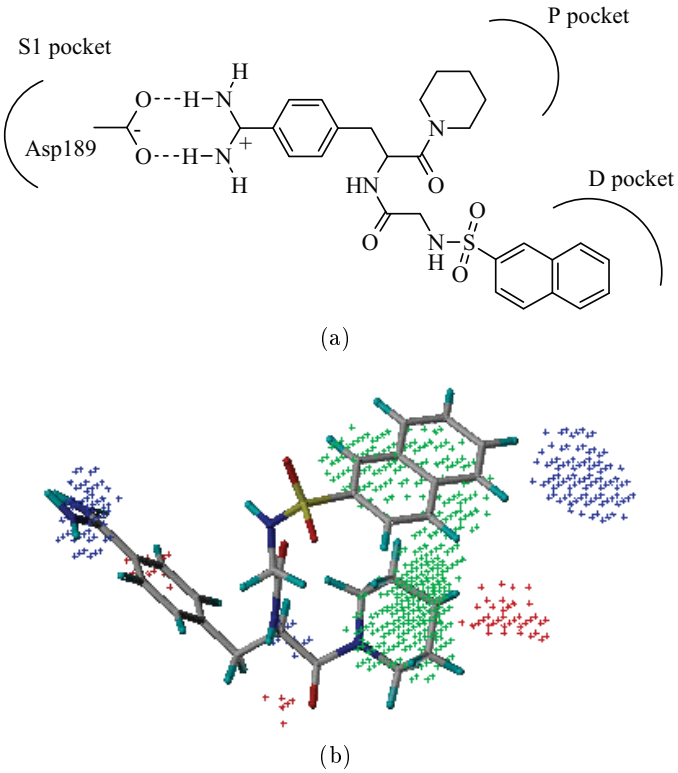
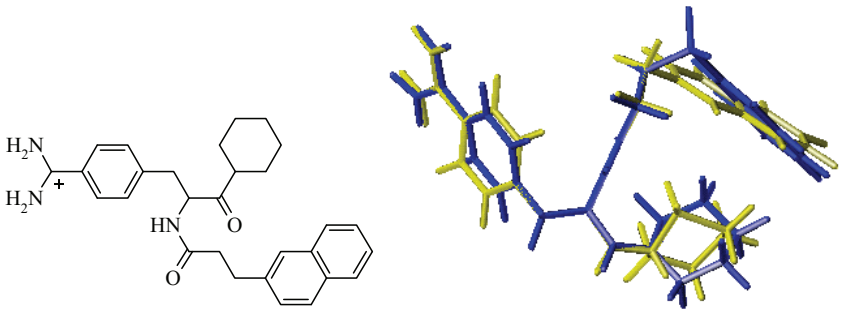


Figure 2.12: Key interaction sites of thrombin (PDB entry 1DWD): (a) NAPAP and the known interaction sites, and (b) the analysis on the key interaction sites by the POCKET program of LigBuilder (hydrogen bond donor grids in blue, hydrogen bond acceptor grids in red, and hydrophobic grids in green) [31].

Thrombin. Thrombin is a serine protease that has a central role in the cascade of blood clot formation. It constitutes a good target for the development of antithrombotic drugs. A Schematic diagram showing the binding pocket of thrombin in complex with the ligand molecule, NAPAP, is given in Figure 2.12(a) [31]. The binding pocket of thrombin contains three major interaction sites, which are denoted as S1, P, and D, respectively. NAPAP is an archetypal active site-directed inhibitor of thrombin [130]. It binds reversibly to the thrombin active site by filling the D pocket with its naphthyl group and the P pocket with the piperidine ring and by placing its basic benzamidine moiety into S1 to form a salt bridge with Asp189. In the complex of PDB

structure 1DWD, the ligand molecule, NAPAP, fits these interaction sites well and exhibits a high binding affinity to thrombin (K_i=10^{-8}M). The crystal structure of this complex was thus used in testing the ability of LigBuilder [31] in reproducing known ligand molecules with the use of a library of organic fragments.

To design ligands for thrombin, the POCKET program of LigBuilder was first used to analyze the binding pocket. The thrombin complex structure was used as the input for POCKET and the key interaction sites was reproduced faithfully, as it is shown in Figure 2.12(b) [31]: the S1 site (the blue aggregation on the left) overlaps the amidine group of NAPAP, and the D and P site (the green aggregations in the middle) overlap the hydrophobic rings of NAPAP.

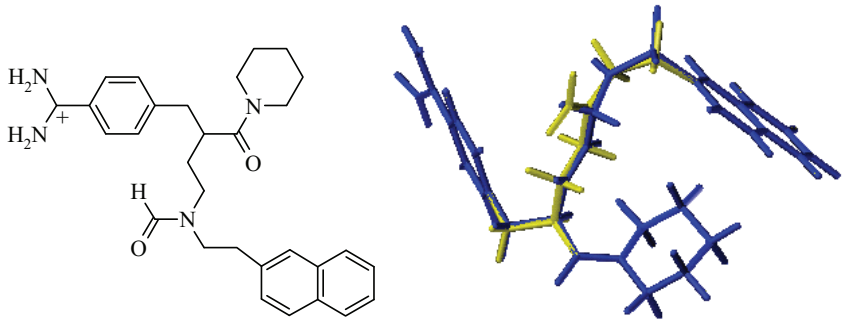


(a) (b)

Figure 2.13: (a) The most similar molecule with NAPAP given by the GROW program of LigBuilder, and (b) superimposition between the molecule (in yellow) and NAPAP (in blue) [31].

As the next step, GROW was run which used a central part (framework) of NAPAP with three growing sites as the seed structure. Since NAPAP is a good inhibitor for thrombin, the program was expected to yield some molecules similar to NAPAP. A molecule very similar to NAPAP was indeed found (see Figure 2.13(a) [31]), which has exactly the same functional group of NAPAP for the S1 site, a more hydrophobic cyclohexane ring instead of the piperidinium ring of NAPAP for the P site, and a more hydrophobic group than the sulfonamide counterpart of NAPAP for the D site. As a whole, this molecule simulates NAPAP well both in structure and conformation (see Figure 2.13(b) [31]). The K_i value of this molecule is predicted to be 10^{-9}M, which is slightly better than the one of NAPAP.

The LINK program on thrombin was also tested. In this case, the seed struc-



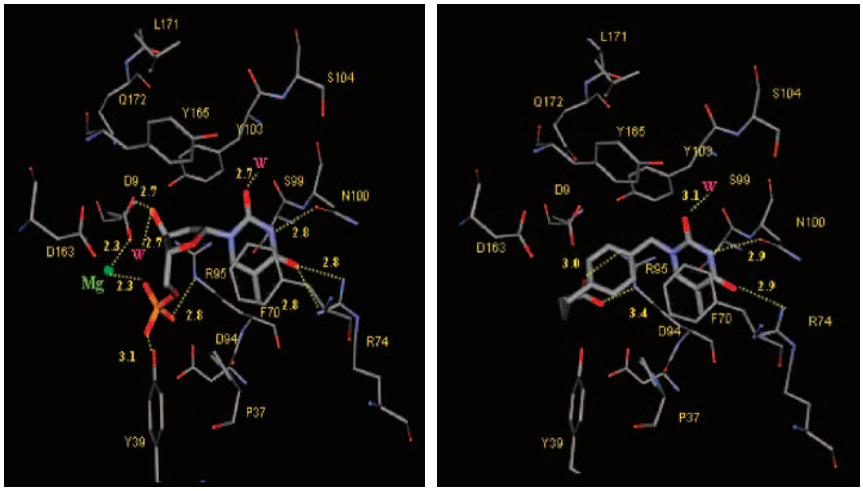
(a) (b)

Figure 2.14: (a) The most similar molecule with NAPAP given by the LINK program of LigBuilder, and (b) superimposition between the molecule (in yellow) and NAPAP (in blue) [31].

ture was extracted from the end parts of NAPAP, including three separated pieces. The purpose of this work is to test whether LINK can link the three pieces in a reasonable way within the constraints of the binding pocket. Also here a molecule very similar to NAPAP was found (see Figure 2.14(a) [31]). Although the framework of this molecule is a bit different from NAPAP, they are very close in style (see Figure 2.14(b) [31]).

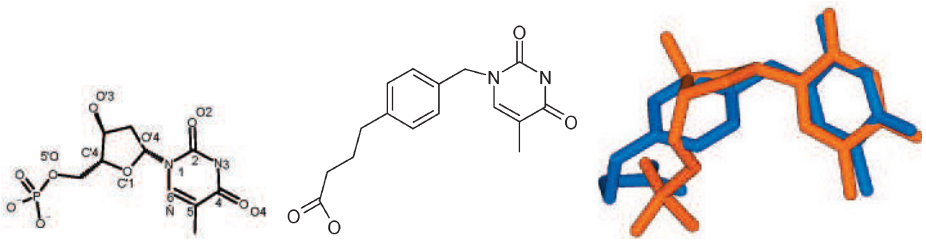
TMPKmt. TMPK belongs to the NMPK family. It is a promising target for developing new antituberculosis drugs [131, 132]. Using ATP as its preferred phosphoryl donor, TMPK reversibly phosphorylates deoxythymidine-5'-monophosphate (dTMP) to deoxythymidine-5'-diphosphate (dTDP). The dTMP binding target TMPKmt (PDB structure 1G3U, see Figure 2.15(a) [22]) was used in a ligand design by LEA3D [22] for novel thymine analogues. LEA3D used a fragment-based genetic algorithm to assemble candidate molecules, each of which was subjected to the evaluation of protein-ligand binding interaction by FlexX docking program. Prior to the *de novo* design process, FlexX had been tested to be able to reproduce the binding mode observed in the crystallographic complex with a root-mean-square of 0.8 Å.

The search of thymine analogues focused on the replacement of the sugar-phosphate moiety of substrate dTMP (Figure 2.16(a) [22]), while the thymine base moiety of it was ordered to be contained in each molecule created by LEA3D to conserve key interactions with the target. Also, the thymine building block was the main base fragment to be taken care by FlexX. It was iden-



(a) (b)

Figure 2.15: The substrate dTMP (a) and the active molecule given by LEA3D/FlexX (b) in their respective key interactions with TMPKmt [22].



(a) (b) (c)

Figure 2.16: (a) Structure of the substrate dTMP, (b) structure of the active molecule given by LEA3D, and (c) superposition between the substrate dTMP (in orange) and the given molecule (in blue) [22].

tified among the results that the sugar part of the dTMP could be replaced by a substituted benzyl group. A molecule (Figure 2.16(b) [22]) with a K_i of 16.5 μM, better than the one of dTMP (27 μM), manages good interactions with the TMPK target in Figure 2.15(b) [22]. Compared to the substrate dTMP, the given molecule forms five rather than four hydrogen-bonds with the target. It possesses a significant π-stack with Phe70 benzene ring and a good balance with Arg 95 positive charge as well as the substrate dTMP does. But only

one of the two hydrogen-bond interactions between the thymine base of dTMP and Arg74 is seen with it [22]. A superposition between the substrate dTMP and the given molecule is shown in Figure 2.16(c) [22].

HIV-1 Protease. HIV-1 protease is a retroviral aspartyl protease that performs an essential step in the life cycle of HIV, the retrovirus that causes AIDS [133]. It cleaves polyproteins newly synthesized by HIV at appropriate places and at critical time to create mature protein components of an infectious HIV virion. Without the sensitive and essential function of HIV-1 protease, HIV virions remain uninfectious. Due to this reason, inhibition of HIV-1's activity to disrupt HIV's ability in replicating and infecting additional cells remains an active subject in pharmaceutical research.

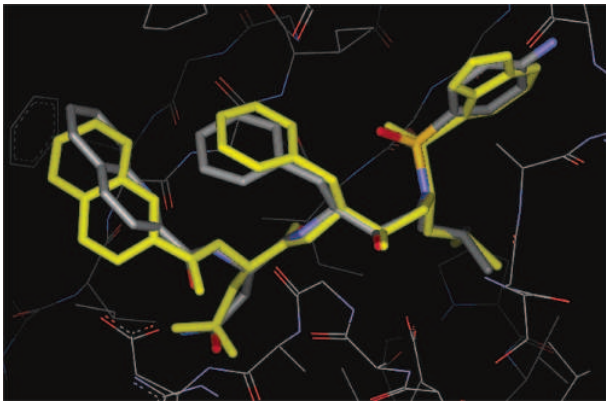


Figure 2.17: A compound generated by BREED and a close analogue of it (in yellow) which has its crystal structure available. Superimposition of them by overlaying the protein structures, yielding the RMS deviation among shared ligand atoms which is only 0.8 Å [124].

HIV-1 protease has large number and wide variety of potent inhibitors, and numerous publicly available crystal structures. It was chosen to demonstrate the capacity of BREED [124] in making use of experimentally determined structures of known ligands to produce new ligands. A brief description of BREED routine is given in Section 2.6.3. In the study, ligands with regard to the HIV-1 protease system came from the PDB crystal structures of 1HPV, 1HSG, 1HPX, 1HXB, 1B6J, 1B6K, 1HII, 1IIQ, 1OHR and 4PHV. Each compound chosen by the set was ensured to share at least one matching bond with another molecule in the set.

A first pass of the first four HIV-1 protease inhibitors of the given set through

BREED led to 20 novel compounds, and a second round of such processing involving the total 10 inhibitors added an extra 81 compounds. Deletion of structures with undesirable characteristics (no key hydrogen bonding hydroxyl group or chemically unstable functionality) from the total set of 101 compounds resulted in a list of 62 novel, chemically viable compounds. Because these compounds were designed by combining known target-bound inhibitors that are superimposed in their respective active conformation, the new structures implicitly inherited the appropriate conformation and orientation for binding to their targets. Figure 2.17 [124] shows a compound generated by BREED and a close analogue of it (in yellow) for which a crystal structure is available. Superimposition of them by overlaying the protein structures, yielding the RMS deviation among shared ligand atoms which is only 0.8 Å.

Chapter 3

GeneGear: An Open Source Software for Computer-Based *de novo* Design

Yunhan Chu and Bjørn K. Alsberg

Department of Chemistry
Norwegian University of Science and Technology
N-7491 Trondheim, Norway

This chapter is about to be submitted.

Abstract

An open source Java-developed software, called GeneGear, is introduced for *de novo* molecular design with multiple methods. It is built upon the chemistry development kit (CDK) together with the Java package Jmol, WEKA, JavaStat and JFreeChart, and interfaced with the 3D structure builder Balloon, Open Babel and ChemAxon tools together with the docking software AutoDock and Vina. Typically, a *de novo* application can be driven by either a systematic virtual combinatorial library or a stochastic evolutionary algorithm, while the former is biased to a design of compounds of same family, the latter is suited for a more diverse chemical space exploration. The quality (fitness) of the sampled structures can be estimated with either a structure-based docking prediction (usually for drug design), or a ligand-based property calculation, e.g., a similarity measure with a template structure or a QSAR regressed from a set of known compounds. Some complementary utility implementations such as designing a fragment library design, graphical visualization of a building block or a product set, and selecting an optimal structure subset are optionally supported. We tested the usefulness of our software with several different case studies, and the obtained results show high promise of using our software as a complementary computer-aided tool in design of drug and catalyst compounds.

3.1 Introduction

In chemistry, the term *de novo* design has traditionally been referred to as a structure-based design [134, 135] process which incrementally constructs bioactive compounds (drugs) within the constraint of the binding site of a target protein or enzyme [95]. Decades of theory and algorithm development has made much extension to the content of this concept, in which the design objects are not only related to druglike compounds [82, 96], but also to molecules such as catalysts [136–138], enzymes [139, 140], and proteins [141–144]. In addition to optimizing properties traditionally associated with binding affinity/biological activity, properties with respect to absorption, distribution, metabolism, excretion, toxicity (the so-called ADMET), synthetic accessibility, and chemically stability, etc. are also increasingly being used. Moreover, when the three-dimensional structure of a biological target for a structure-based design is not available, structures of known ligands of the particular targets can be used as complementary measures in an alternative ligand-based design.

In order to generate novel molecular structures, various approaches can be used. Stochastic strategies [20, 23–26] find novel structures from random sampling of the chemical space, while deterministic strategies (typically a combinatorial library design [45, 55–58]) assemble frameworks and determined side

chains at decided substitution positions to generate candidate structures. Some fragment-based design approaches fall somewhere between the two, which take prepared fragment-based building blocks but manipulate them through stochastic operations like growing [28, 31, 59, 60], linking [29, 31, 84], crossover [22, 27, 60] and mutation [22, 27, 31, 60].

The strategies that are for evaluation of the quality (fitness) of the generated structures can also be diverse. Receptor-based approaches starts with the three-dimensional structure of the target (known from X-ray crystallographic or NMR data) and examine the steric or energetic fit of the ligand binding to the receptor, where the energy can be calculated based on a force field [27, 59, 60], an empirical [22, 29, 31, 52, 57, 107], or a knowledge-based [43] approach. In contrast, ligand-based approaches make use of a three-dimensional pharmacophore model [64], a quantitative structure–activity/property relationship (QSAR/QSPR) [19, 54, 66], or a descriptor-based similarity measure to a known active compound [24, 25, 65] to estimate the fitness of the candidate structures. Also there can be a combination of the above [67, 84] or the application of multiobjective constraints [68, 69, 114, 115] that have multiple factor evaluations.

Despite the diverse existing *de novo* tools, many of them are dominated by a particular design routine which do not allows much freedom in how the system is used, modified and extended. Customers may expect inexpensive modular softwares that do one or several tasks well and can be integrated in a flexible way by in-house staff to meet their own specific needs. An open source strategy provides maximum flexibility in this respect which allows programmers to share and advance ideas and produce new softwares without having to start from scratch. Unfortunately, most of the available *de novo* softwares are proprietary softwares, though there do exist a few exceptions such as AutoGrow [60] and JavaGenes [24].

To improve on the situation and make a contribution to the community, we have developed GeneGear, an open-source Java software for *de novo* design with multiple methods. GeneGear is built on the chemistry development kit (CDK) [85] and several other chemical, statistical and graphic Java packages [86–89]. Virtual combinatorial library design and evolutionary *de novo* design represent its two main application functions, which respectively provide users a systematic and a stochastic way to sample novel candidate structures. The structures generated from both routines can be flexibly measured by various scoring systems, e.g., a receptor–ligand binding energy predicted by a docking software, a molecular similarity calculated with a set of molecular descriptors, or a self-defined QSAR/QSPR. The fitness calculation process can be

executed in parallel over multiple computational nodes in large-scale problems. In addition to the main applications, some complementary utility implementations such as design of a fragment library, graphical visualization of a building block or product set, and selection of an optimal subset are optionally supported. In the following sections, the architecture of GeneGear and its major functional modules will be explained in detail. Also, a demonstration of its functionality will be given through several well-studied application examples.

3.2 General Map of GeneGear

A general map of GeneGear is shown in Figure 3.1. Virtual combinatorial library design and evolutionary *de novo* design are two main application programs which respectively provide users with a systematic and a stochastic way to sample novel candidate structures. The central data processed by GeneGear are chemical structures and their relevant property attributes. The latter include data types of numeric, logical and string, and can be easily retrieved with vectors and matrices. The former consists of more complicated structural objects which require an appropriate representation for effective processing. There are quite a number of ways to represent a 2D molecular structure, e.g., SMILES [17–19], connectivity matrices [20], linear [21, 22] or topological trees [23], and graphs [24–27]. After a careful comparison, we decided on a graph-based representation [20, 24–26], which has more advantages than other representations. The chemistry development kit (CDK) [85] provides a proper approach to parse 2D molecular structures as graphs in a computer, where the molecular diagram is coded as a set of atom and bond objects with connectivity information stored in a data structure called an AtomContainer [26, 85]. Using the graph-based representation, chemical structures can be implemented in a flexible way. We have developed a suite of graph- and fragment-based operations - Crossover, Mutation, Grow, and Link - for identification of novel molecular structures. While the first three operations are manipulated by an evolutionary algorithm (EA) where relatively random variations occur on the basis of an existing set of structures or fragments, the last operation (Link) is used in the construction of a virtual combinatorial library where different groups of side chain fragments are connected at corresponding variation sites of a main framework (scaffold) according to some fixed virtual synthesis routine.

Of course, both an EA or a library design need a scoring system to help select high-quality structures. GeneGear has been coded to interface with the renowned and permissive docking softwares - AutoDock [63] and AutoDock Vina [90] (henceforth referred to as Vina), through which the candidate lig-

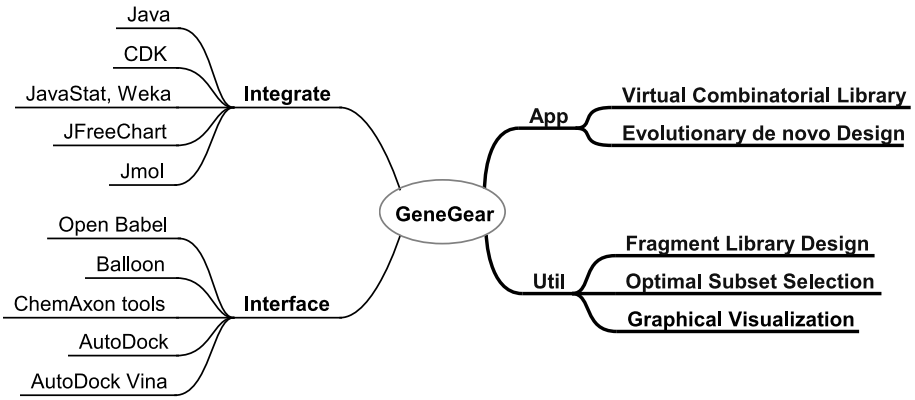


Figure 3.1: General development structure of GeneGear.

ands can be docked into the target protein for a structure-based prediction of binding affinity. While AutoDock 4 combines an semiempirical free energy force field [145] scoring function with a Lamarckian genetic algorithm (LGA) search method [63], Vina uses a gradient optimization method directed by a scoring function combining advantages of knowledge-based potentials and empirical functions [90]. Both of the softwares can take into account full ligand flexibility and limited receptor (residue) flexibility, and return bound conformations with predicted free energies of binding. AutoDock allows to determine the binding energy of a given ligand by referring it to the conformer that has lowest energy either among the population of all searched conformers or among the most-populated cluster decided by root-mean-square-deviation (rmsd) of atomic coordinates, whereas Vina only supports the first type of determination. Alternatively, a set of molecular descriptors imported from CDK QSAR package [146] or some user-defined descriptors can be applied to define a ligand-based similarity measure related to a template structure or build a PLS [147, 148]-based quantitative structure–activity/property relationship (QSAR/QSPR). These various fitness pressures can be used in combination to form an appropriate multiobjective fitness function.

Prior to any possible fitness pressure computation, an acceptable initial 3D structure for a molecule is usually required. GeneGear by default maintains interfaces to the programs of Balloon [149], Open Babel/Obconformer [150, 151], and ChemAxon [152] MolConverter/Cxcalc [153] for different ways of 3D conformational search. Balloon [149] uses distance geometry to generate an initial conformer, which is then subjected to geometric modifications by a genetic algorithm which employs a MMFF94-like molecular force field for estimation of

the energies of each new candidate conformer. Open Babel [150, 151] is often used to translate molecules between different formats. From version 2.2.3, it also provides a rule-based from-scratch 3D structure building method based on the energy minimization of a MMFF94 force field [154]. Obconformer [151] using the same force field with Monte Carlo search can be used to locate a better conformer. MolConverter [153] generates coordinates from a Minkowski-like space [155] followed by geometry optimization to a local energy minimum in 3D using the Dreiding force field [156]. A deep conformational search using the same force field level by the Cxcalc [153] program follows. It must be noted that here GeneGear only provides interfaces to these external softwares where some of them are proprietary, and it is the responsibility of the user to have valid license for these programs.

For the reason of efficiency, GeneGear supports parallel implementation of conformational search and fitness calculation of both an EA and a library design on a cluster-type architecture through the Message Passing Interface (MPI) [157] standard. The basic scheme is as follows: the user allocates n nodes for the overall job, each of which contributes m processor cores. Core one on the master node generates $(n \times m - 1)$ structures and transfers them simultaneously to the $(n \times m - 1)$ slave cores. While the slave cores are busy computing the fitness of the current batch of molecules, core one prepares the next batch of new $(n \times m - 1)$ molecules and then accepts the previous batch of calculated molecules to refresh the current population.

Complementary utility implementations, such as building a fragment library and selecting optimal subset of structures, are optionally supported by GeneGear. Furthermore, it provides graphic user interface (GUI) for visualizing and browsing a set of chemical structures; a set of molecular descriptors imported from CDK QSAR package [146] covering electronic, geometrical, topological and hybrid categories is available for capturing the chemical features of involved structures. Principle component analysis (PCA) is included in GeneGear to be used for investigating the chemical space specified by chosen structure descriptors. Several useful Java libraries are integrated for various practical needs, such as, Jmol [86] for 3D molecular visualization and rendition, JavaStat [88] for PCA analysis, WEKA [87] for PLS modelling, and JFreeChart [89] for data plotting.

3.3 Fragment Library Design

Fragments from existing drugs and compounds with known activities and properties make it likely to produce new compounds with reasonable molecular

structures. In addition, they are also more likely to be "druglike" in a general sense (such as, satisfying the "rule of five" [101] and not containing reactive functional groups) than random structures. This makes the fragment-based drug design a popular concept. However, the number of available drugs is vast, so usually only a small fraction of them is accessible. GeneGear provides a procedure to build a fragment library from fragmentation of available molecular structures. In general, the procedure consists of two steps: splitting available molecules into fragments and screening obtained fragments with a set of rules.

Splitting. Structures are hydrogen depleted and split into fragments at rotatable and non-terminal bonds (i.e., single bonds which are not a part of a ring and do not include atoms connected to only one other atom), or at variation sites of a common skeleton among a series of compounds of same family. Prior to that, some user-specified rules, such as "rule of five" [101], can be implemented to prevent undesirable structures from being processed. When the splitting operator starts on a structure, the bonds that connect to rings have higher priority to be handled than the ones that connect to general atoms. The resulting fragments are saved in MDL sdf file format with the positions where the substitution points (R-groups) and the sources from where they originate are indicated, as the example shown below:

```
   153.sdf

   3   2   0   0   0   0   0   0   0   0999 V2000
     -0.0178    1.4608    0.0101 C   0  0  0  0  0  0  0  0  0  0  0  0
      0.0021   -0.0041    0.0020 N   0  0  0  0  0  0  0  0  0  0  0  0
      1.3566    1.9679   -0.0003 N   0  0  0  0  0  0  0  0  0  0  0  0
   1   2   1   0   0   0   0
   1   3   1   0   0   0   0
M  END
>  <R-groups@>
1 2 3

>  <From>
357650.mol;362868.mol;375086.mol;612799.mol

$$$$
```

Screening. All derived fragments are subjected to certain filter rules, so duplicate and unfavored entries are removed. The "unfavored" structures are somewhat up to the specific definition of the user. Since GeneGear is distributed as open source, a user can easily adopt his own code to exclude structures and fragments which are not suitable. A library is then established which contains preferable building blocks (scaffolds and side chains). A link file is then created which contains the file paths and occurrences in known molecules of the fragments, and may act as an input to the main *de novo* design programs.

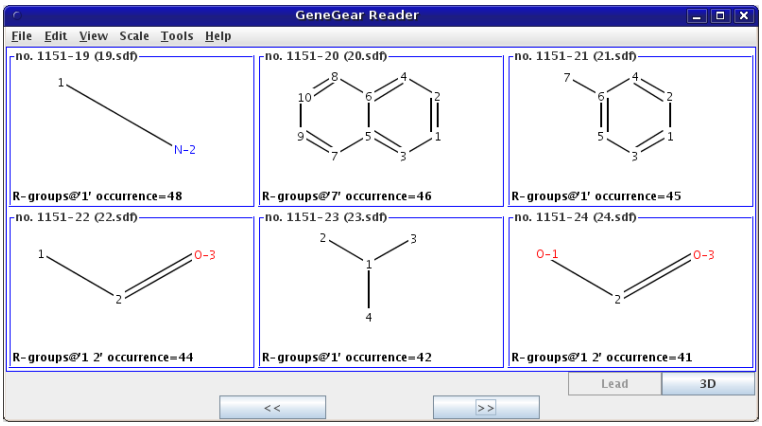


Figure 3.2: Graphical view of a fragment library built by GeneGear.

Visualization. To further study a fragment library, it is useful to present all the structures in a united graphic interface so the user can check them easily. A graphical view of the 1151 fragments built by GeneGear on the basis of 1990 compounds of the National Cancer Institute (NCI) diversity dataset [158] is shown in Figure 3.2. To build this fragment set, only the structures that possess atoms within C, N, O, S, P, F, Cl, Br and I got processed. Fragments that are charged, or possess more than 16 atoms, more than three fused rings or at least one ring with more than 7 atoms got excluded according to self-defined rules. Also, for our later case study, we exclude the fragments that possess atoms other than C, N, O, S. Finally, we get a set of 1151 fragments including 599 side chains with one R-group, and 552 scaffolds where 462 have two R-groups, 75 have three R-groups, 14 have four R-groups and 1 has five R-groups. Further investigation of this fragment set is seen in Case Study I in Section 3.6.

3.4 Optimal Subset Selection

Statistical selection of a subset of molecules that spans an important structural or physiochemical space of an original pool of structures is a technique frequently needed in various areas of chemistry, such as combinatorial synthesis [159, 160], selection of molecules for screening [161, 162], and QSAR/QSPR analysis [163–165], where diversity, coverage, and descriptive ability are usually the main goals to achieve. The process starts from mapping the molecules into a multi-dimensional structural or property space. Descriptors imported from the CDK QSAR package [146] or some self-defined descriptors can be

applied for constructing such a space, where the relative positions of the mapped molecules and the inter-distances between them can be measured. Various selection strategies, such as D-optimal design [166], space-filling (SF) design [167–170], cell-based design [171], cluster-based design [172, 173] and onion design [164, 165, 174] depending on the applications, can be applied to operate on the descriptor-based or a transformed (e.g., PCA score) space to select an appropriate molecular set. In GeneGear, the following selection algorithms are provided besides a simple random selection:

Dissimilarity Selection. A type of commonly used strategy for diversity-based selection [168, 169]. In our algorithm, the compound which is closest to all the other compounds in the same pool is selected first. In the following, the selection is processed in a loop such that the next compound to be selected is always as distant as possible from the already selected compounds. As the algorithm attempts to generate the most diverse set, compounds at the edge of the property space can be taken.

Sphere Exclusion (SE). Based on the work of Hudson et al. [170] and Wootton et al. [175], it starts with the compound that has the largest sum distance from all the other compounds, and the compounds that are closer to the selected one within the sphere of defined radius are deleted. A new compounds is then selected which is most distant from the one previously selected. The whole process is repeated until no more compounds remain. By default the distance averaged between all molecules is taken as the initial radius. As the number of selected compounds depends on the exclusion radius, it may have a number of iterations of adjusting exclusion radius steps to satisfy a predefined number of points.

D-Optimal Design (DOD). A set of compounds supporting a regression model is selected such that its model matrix X maximizes the determinant of the covariance matrix X^TX [166]. Four different models: linear, pure quadratic, full quadratic and pure cubic are by default provided by GeneGear (see Table 3.1). Which model to use is dependent on the design (sample) size. A genetic algorithm (GA) is used to search the optimal compound set with maximum X^TX volume. The D-optimality criterion ensures the outer peripheral parts of a compound space be adequately sampled, while for the inner part it is not necessarily the same. It is thus not well suited for large design sizes (many compounds) relative to the number of factors, otherwise there may be oversampling of the outer peripheral parts of the molecular space.

Table 3.1: The term and smallest sample size for each D-optimal model in GeneGear, where K denotes the number of significant factors.

Model	Term	Sample size
Linear	$1+\sum_{i=1}^{K} x_i$	$1+K+2$
Pure quadratic	$1+\sum_{i=1}^{K} x_i+\sum_{i=1}^{K} x_i^2$	$1+2K+2$
Full quadratic	$1+\sum_{i=1}^{K} x_i+\sum_{i=1}^{K} x_i^2+\sum_{i,j=1}^{K} x_i x_{j\neq i}$	$1+2K+K(K-1)/2+2$
Pure cubic	$1+\sum_{i=1}^{K} x_i+\sum_{i=1}^{K} x_i^2+\sum_{i,j=1}^{K} x_i x_{j\neq i}+\sum_{i=1}^{K} x_i^3$	$1+3K+K(K-1)/2+2$

D-Optimal Onion Design (DOOD). To overcome some of the shortcomings of D-optimal and space-filling designs, D-optimal onion design (DOOD) was developed [165, 174] and used [164]. In our code, samples are sorted according to their Euclidean distances to the centre point (the one closest to the calculated centroid of the experimental domain) and are subject to be divided into several layers (shells). The number of significant factors and sample sizes of the available models (see Table 3.1) help to decide which models to use and how they divide the applied sample space. The final selection are the collection of the compounds selected from different layers by D-optimal design with respective model.

Most Descriptive Compound (MDC). This is a method first proposed by Hudson et al. [170] It selects a subset of compounds which most effectively represents the compounds in the original population. The information of a compound is quantitatively described by a vector of reciprocal values of distance ranks of all compounds to it, (e.g., 1 for itself, 1/2 for the closest neighbor, 1/3 for the second closest, etc.) and the information of the whole data set is a vector V summed by all the compounds' information vectors. The procedure then proceeds as follow: 1) the compound that has the largest sum value namely smallest overall distance to all the other compounds is selected as the most descriptive compound (MDC), and 2) a new vector V containing the information for the remaining compounds is decided by the multiplication of the values of the current V with the values of 1 that are subtracted in advance by the corresponding values of the MDC vector. The whole procedure is repeated until all values in V are less than 1 (no more information to extract) or until the required number of compounds have been selected.

3.5 Virtual Combinatorial Library Design

In combinatorial chemistry, a combinatorial library is a collection of different but structurally related chemical compounds that are subjected to synthesis and screening to identify potential novel molecules with desired properties. To construct such a library, a process of *in silico* "virtual library" generation constructs an essential part: molecules are represented by Markush structure - a common scaffold with several variation sites labeled as R-group, each of which is associated with a list of alternative reagents [176, 177]. This can quickly lead to a large number of structures. For example, a scaffold with three R-groups, each corresponding to 20 reagents, can generate 8000 possible structures. However, the number of compounds that are synthesizable is alway lower than theoretical yield. Selecting subsets from the full virtual library for actual synthesis using various criteria is often used to decrease the cost. A virtual combinatorial library can be designed following one of the two directions [177]: (1) a directed library which is biased against a specific target, or a known pharmacophore model, and (2) an exploratory library which is target-independent and is supposed to span a wide range of structural and physiochemical characteristics.

GeneGear supports construction of both types of libraries. A virtual combinatorial library of compounds provides a common skeleton with several variation sites (R-groups) and the same number of reagent groups. A virtual compound library is constructed by enumerating and linking the different groups of reagents at the corresponding substitution sites of the common skeleton. Once the full library is established, a directed or diverse subset of compounds can be selected from it for different purposes. The directed selection strategy collects a set of product molecules that have high binding affinity to a given biological target, for which molecular docking softwares such as AutoDock [63] and Vina [90] can be used to predict the binding free energy. In contrast, the exploratory selection strategy tends to result in an informative, representative and diverse structure set. The exploratory selection can be based on either the reagent space [178–180] or the product space [162], where various selection strategies as discussed in Section 3.4 can be used for the task.

Case Study. A set of 88 serine protease inhibitors provided with experimentally determined biological activities (pKi) towards trypsin was extracted from the work of Böhm et al. [181] As shown in Figure 3.3, there is a common skeleton with two variation sites shared among the given inhibitors. We split the compounds at the variation sites of the common skeleton using the fragment library tools above, resulting in 15 R1-group related reagents and 54

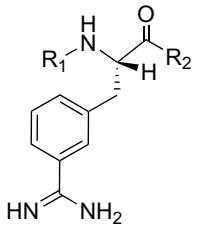


Figure 3.3: A common skeleton with two variation sites shared among structures compiled by Böhm et al. [181] Structural variations are allowed at the positions R_1 and R_2.

Table 3.2: Descriptors for scaling the structural space spanned by the 810 product compounds derived from the Böhm set.

category	molecular descriptor
3D	Charged partial surface area (CPSA) [182]
	Gravitational index [183, 184]
	Molecular length to breadth ratio [185, 186]
	Molecular distance edge (MDE) [187]
	Moment of inertia [188]
	Geometrical shape coefficients of radius-diameter diagram [189, 190]
	Weighted holistic invariant molecular (WHIM) descriptors [191]
2D	Topological polar surface area (TPSA) [192]
	Topological shape coefficients of radius-diameter diagram [189]
	XLogP [193, 194]
	Polarizability differences between all bonded atoms [195]
	numbers of hydrogen bond acceptors
	numbers of hydrogen bond donors
	numbers of atoms
	numbers of bonds

R2-group related reagents. A full product library of 810 compounds was constructed through the enumeration of the two groups of reagents at the exclusive substitution points of the common skeleton by GeneGear, where the Balloon program was called with a setting of 50 conformers and 100 generations for its GA-based search for finding a low-energy 3D conformer for each structure.

A set of 82 descriptors covering seven 3D categories and eight 2D categories,

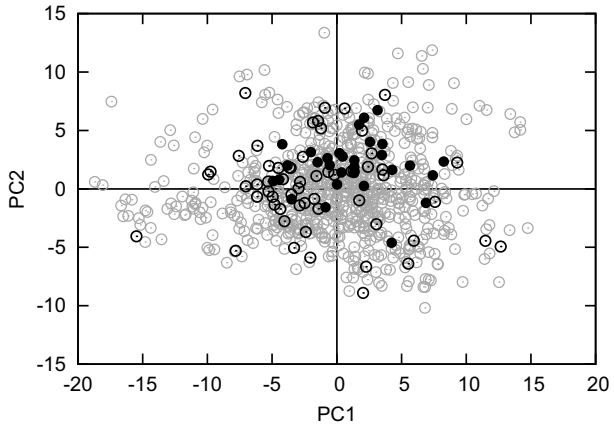


Figure 3.4: A PCA investigation of the structural space of the full set of 810 compounds built from the Böhm set, 82 descriptors covering seven 3D categories and eight 2D categories are used to scale the objects. The scores of the first two principal components (PC1 - PC2) that describe 67% of the variance in the descriptor space are shown, where the black circles represent the individuals that are covered by the Böhm set and the filled ones mark the first 30 highest active inhibitors against trypsin, while the gray circles represent the individuals that are out of the Böhm set.

as is listed in Table 3.2, was used to scale the structural space spanned by the 810 product compounds. A PCA score plot of the structural space using the first two principal components that explain 67% of the variance, is shown in Figure 3.4, where the black circles represent the individuals from the Böhm set and the filled ones mark the first 30 highest active inhibitors for trypsin, while the gray circles represent the compounds that are out of the Böhm set.

A number of virtual library designs were performed, each of which involves a selection of 50 compounds from the library of 810 compounds, see Table 3.3. The selection strategies were either directed by the binding free energy predicted by the docking software AutoDock or Vina against the crystal structure of trypsin (PDB structure 1PPH [196–198]), or based on the exploratory statistical selection algorithms described in Section 3.4. Whereas the former was only applied to the product compounds, the latter was applied to both product- and reagent-based structure space. Since the results of the docking softwares and some of the statistical selection procedures such as the D-optimal and the onion design rely on random number generators, we repeat the two types of experiments 6 and 100 times respectively using different random numbers.

The derived compound libraries were investigated with their coverage of the 88 compounds and the 30 trypsin active compounds within the Böhm set. The ones which covers the trypsin active compounds from the Böhm set more were locally referred to be better than the ones which cover less, providing a practical standard to compare performances of the different design procedures.

Table 3.3: The various virtual combinatorial library designs with their respective library coverage of 88 known compounds and 30 trypsin active compounds from the Böhm set[d].

Library[a](50[b])	Design	Coverage[c](via.Reagent[d])		Design	Coverage[c](via.Product[d])	
		Böhm(88)[e]	Try.Act.(30)[f]		Böhm(88)[e]	Try.Act.(30)[f]
Exploratory	RDM[1,†]	6	2	RDM[1,†]	5	1
	MDC[2]	13	4	MDC[2]	5	2
	SE[3]	2	0	SE[3]	5	0
	Dissim[4]	4	0	Dissim[4]	2	0
	DOD[5,†]	5	1	DOD[5,†]	5	1
	DOOD[6,†]	5	1	DOOD[6,†]	5	1
Directed	-	-	-	AutoDock[7,‡]	12	3
	-	-	-	Vina[8,‡]	12	6

[a]The type of the designed library. [b]The sample size. [c]The number of overlapped compounds. [d]The category of the design with respect to the studied objects, product-based or reagent-based. [e]The dataset compiled by Böhm et al., which includes 88 serine protease inhibitors. [f]The first 30 highest active inhibitors for trypsin within the Böhm set[d]. [1]Random selection. [2]Most descriptive compound. [3]Sphere exclusion. [4]Dissimilarity selection. [5]D-optimal design. [6]D-optimal onion design. [7]AutoDock 4.2 was used. [8]AutoDock Vina. [†]Average over 100 separate experiments. [‡]Average over 6 separate experiments.

The results of the different categories of selections in average are listed in Table 3.3. Under the reagent-based design category, it is observed that MDC shows better performance than the other exploratory selection algorithms, where 13 out of 50 selected compounds are covered by Böhm set and 4 of them are among the 30 trypsin active inhibitors. When it comes to the category of product-based design, the screening directed designs outperform the exploratory statistical designs, and the docking software Vina shows better prediction than AutoDock which is in agreement with previous studies [90, 199]. The reason for why the exploratory selections show low performance on product space could be that the structure descriptors do not properly represent the structural variation. In fact, exploratory designs are more ready for a diversity-based selection than an activity-based selection, which can be another important reason. Further investigation is outside the scope of the present paper.

3.6 Evolutionary *de novo* Design

Virtual combinatorial library design provides a systematic routine to sample a large array of novel compounds from sets of reagent building blocks, which makes it more suited for a design of compounds of same family where the researcher has precise knowledge about how specific building blocks (skeleton and reagents) are combined into specific structures. By using common skeleton with fixed substitution points for specific reagents, the search is restricted to a fraction of compounds from a dense part of a potentially huge chemical space. However, there can also be the case that one has no clear idea about the constitution of the framework of the designed structures or that a more diverse chemical space is expected to be explored. In such cases the investigator is confronted with is a much larger combinatorial problem for which exhaustive search is impractical. Evolutionary algorithms (EAs), including genetic algorithm (GA) [70–72], genetic programming (GP) [73, 74], evolution strategy (ES) [75, 76] and evolutionary programming (EP) [77, 78], take stochastic routines for global optimizations, providing a practical tool for investigation of such problems.

With the application of tailored genetic operators and a well-defined fitness function, an EA tends to find optimal solutions to a given problem through a population-based optimization. In the field of drug design, EAs have been widely used in *de novo* suggestion of novel structures [15, 81, 82]. The available *de novo* EA implementations, such as Chemical Genesis [67], JavaGenes [24], LEA [19], TOPAS [65] and CoG [25] perform ligand-based designs where a descriptor-based similarity measure, [24, 25, 65] a combined constraint of properties [67] or a QSAR/QSPR model [19, 54, 66] is used as a scoring system. On the other hand, LigBuilder [31], ADAPT [27], LEA3D [22], SYNOPSIS [59], GANDI [84] and AutoGrow[60] perform structure- (or receptor-) based designs where the three-dimensional structure of a given protein is facilitated to examine the steric or energetic fit of the designed molecules. GeneGear supports both ligand-based and structure-based designs in this respect. In the former case, a molecular similarity measure, a single property constraint, or a QSAR/QSPR (driven by such as a PLSR model) utilized with a set of available or self-developed molecular descriptors may be used as a fitness function. In the latter case, molecular docking softwares such as AutoDock [63] and Vina [90] can be prompted for a virtual screening of the candidate ligands according to their predicted binding free energies with respect to the target. Sampling new molecular structures on the basis of a pool of available structures [200] or a set of fragment libraries are both supported with some genetic operators, e.g.,

crossover and mutation. The operators can be informed to keep some common part of the implemented structures intact and only let the remaining parts be varied. In fragment-library based design cases, a number of lead frameworks can be included in a run for competition of a restricted portion of designed structure which may also have other remaining parts that are allowed to be varied more freely.

The basic scheme of our EA is configured as follow: a seed population sourced from an available pool of structures or grown from a set of fragment libraries is constructed. All the structures are by default saturated with hydrogens and subject to a conformational search performed by programs like Balloon [149], Babel [151], or MolConverter & Cxcalc [152, 153]. The one of lowest energy within a predefined number of searched conformers for each structure is saved and scored by a structure-based or ligand-based fitness function. Next, an optimization cycle consisting of five main steps starts: 1) selection of competitive parent structures (only for genetic crossover and mutation operations), 2) breeding new offspring structures by structural operations, 3) conformational search of offspring structures 4) fitness calculation of offspring structures, and 5) updating current population. Here, more competitive parent structures are found by a tournament procedure which compares pair of randomly individuals picked from the population and takes the best. New offspring structures are generated from genetic crossover or mutation, or from fragment growing. Whereas the crossover and mutation operators swap or mutate fragments either based on building blocks from a fragment library or based on some random species of the operated structures (see ref.[200] for the latter case), the growing operator is only applied to the fragment-library based case. The generated structures go through conformational search and fitness evaluation, and are used to replace, gradually, the least competitive structures in the current population. The optimization cycle continues until a predefined number of offspring structures have been produced. The offspring, together with the survivors of the current population, establish a new generation. The population evolves over generations until a predefined termination criterion is satisfied (e.g., maximum of generations or a minimum number of satisfying solutions) or exhaustion sets in (e.g., no fit solution is found within a limited number of continuous searches).

Case Study I. The following application example originates from the 1990 compound containing National Cancer Institute (NCI) diversity dataset [158]. Selected from around $140,000$ compounds, the dataset consists of a broad range of chemotypes [158]. It has been widely used in docking-based virtual

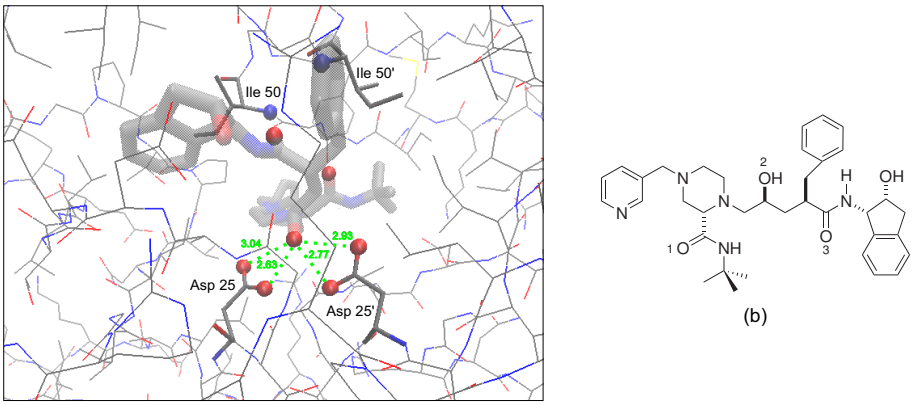


Figure 3.5: (a) Active site of HIV-1 protease in complex with its inhibitor indinavir (PDB structure 1HSG). Specific interactions between the enzyme and the inhibitor include the hydroxyl group (O2 in (b)) hydrogen bonding to the carboxyl groups of the essential Asp 25/25' enzymic residues (hydrogen bonding distances are shown in angstroms), and the amide oxygens (O1 and O3 in (b)) of the inhibitor hydrogen bonding to the backbone amide nitrogen of Ile 50/50' via a potential intervening water molecule. (b) Structure of the indinavir with its numbering scheme of its oxygen atoms.

screening studies [201–203], and specifically, it has also been used as a benchmark dataset for comparing AutoDock and Vina in the application [199] of screening for inhibitors that are actively against human immunodeficiency virus (HIV-1) protease [133], an enzyme vital to the replication of the AIDS virus. A stereoview of HIV-1 protease active site in complex with indinavir [204, 205] – one of its potent and orally bioavailable inhibitors – at the resolution of 2.0Å (PDB structure 1HSG [196, 197, 206]), is shown in Figure 3.5 (a). One important interaction between the enzyme and the inhibitor is a critical hydroxyl group (referred to as O2 in Figure 3.5 (b)) that hydrogen bonds to the carboxyl groups of the catalytically active aspartic acids (Asp 25/25'). Incorporation of structural isosteres as replacements of the hydroxyl group may lead to compounds that are potent and selective to HIV-1 protease.

An application of our fragment library tools of GeneGear on the NCI diversity set has led to a library of 1151 fragments including 552 scaffolds and 599 side chains, as it has been described in Section 3.3. We repeated the same fragmentation routine to the indinavir structure, from which 8 fragments (fr.1-fr.8) were generated including 4 scaffolds and 4 side chains as they are respectively

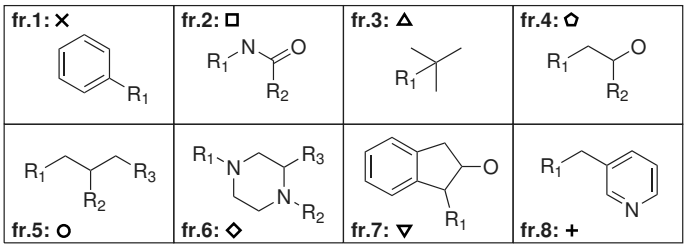


Figure 3.6: Fragments split from the structure of indinavir using the fragment library tools of GeneGear, which resulted in 4 side chains (fr.1, fr.3, fr.7 and fr.8) that are all one R-group containing, and 4 scaffolds among which fr.2 and fr.4 are both two R-group containing while fr.5 and fr.6 are three R-group containing.

shown in Figure 3.6. It was detected that the NCI diversity set covers the fragments from indinavir with 284 times for fr.1, 89 times for fr.2, 42 times for fr.3, 22 times for fr.4, and 7 times for fr.5.

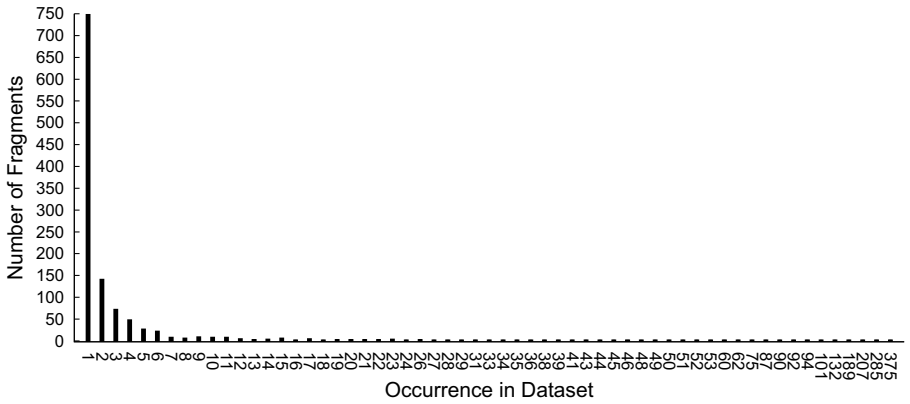
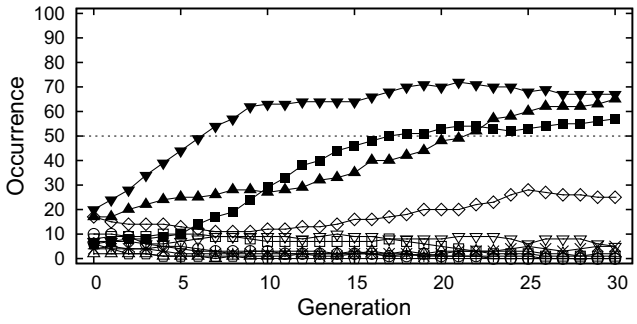


Figure 3.7: Histogram plotting on the possible occurrences of the fragments in the NCI diversity set plus the indinavir (x) vs. the associated numbers of fragments (y).

The fragments derived from the two sources were merged, which produced a fragment library with 1154 unique entries. Figure 3.7 reflects a numerical counting of fragments (y) corresponding to possible occurrences of the fragments in the NCI diversity set plus the indinavir. A selection of fragments occurring no less than 8 times in the library plus the three singly occurring indinavir fragments resulted in 98 fragments including 60 side chains with one

R-group, and 38 scaffolds where 34 have two R-groups and 4 have three R-groups. The selected set of 98 fragments was applied in both a receptor- and a ligand- based evolutionary *de novo* design of novel active inhibitors for the HIV-1 protease. In the receptor-based design, the 1HSG HIV-1 protease structure obtained from the Protein Data Bank (PDB) [196, 197] was used as the receptor which had the complexed ligand and all water removed, the binding free energy predicted by docking based on AutoDock 4.2 was used to estimated the fitness of the candidate structures. In the ligand-based design, similarity to the indinavir structure based on the description of the same seven 3D and eight 2D classes of molecular descriptors used in the virtual combinatorial library case study was applied to define a fitness function. Both the receptor- and ligand- based designs used a population size of 100 individuals and maximum 30 generations, and each was repeated 6 times with different random numbers. The occurrences of each indinavir related fragment got averaged among the respective 6 experiments along with the generations, see the plots for the two types of designs in Figure 3.8 (a) and Figure 3.9 respectively.



(a)

(b)

Figure 3.8: (a) Occurrences of the 8 indinavir related fragments averaged by the 6 receptor-based EA experiments along with the generations, as well as the fragments that are selected over 50 times by at least one averaged generation, where the legends correspond to the fragments with the same marks in Figure 3.6 and (b) respectively. (b) Fragments that were selected over 50 times by at least one average generation, including two scaffolds associated with two R-groups and one scaffold associated with three R-groups.

In the receptor-based design, see Figure 3.8 (a), all of the 8 indinavir related fragments (see Figure 3.6) got chosen by the EA initial generation, but none of them comes out on top in the end despite the relatively high persistence of fr.6. In contrast, significant selection goes to the three fragments derived from the NCI diversity set (see Figure 3.8 (b)) which have over 50 times in average been used by the final generations of the evolution runs.

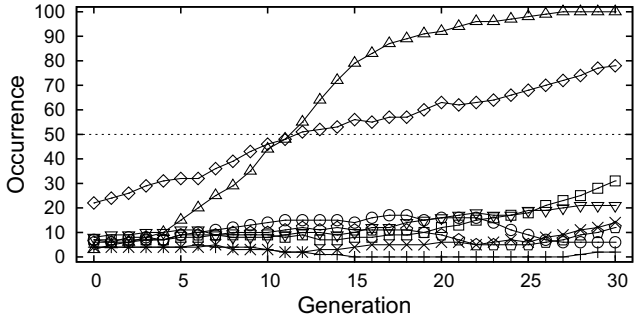


Figure 3.9: Occurrences of the 8 indinavir related fragments averaged by the 6 ligand-based EA experiments along with the generations (no other fragments selected over 50 times by the averaged generations), where the legend correspond to the fragments with the same marks in Figure 3.6.

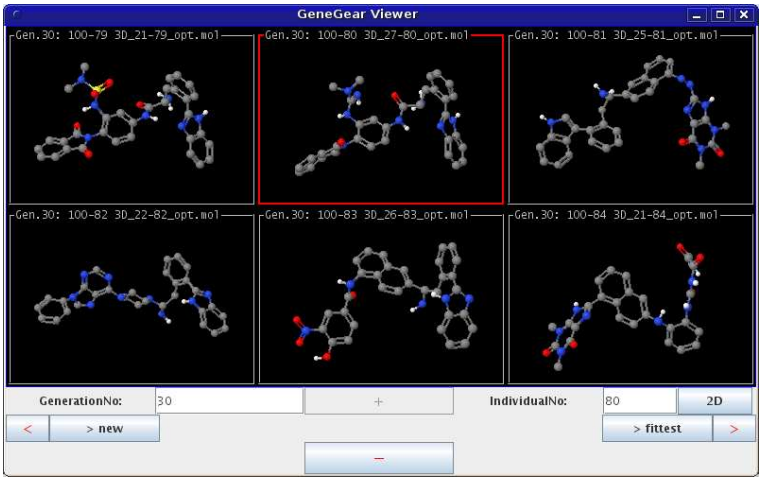


Figure 3.10: Graphical view of outcome of an EA run of GeneGear.

In the ligand-based design, see Figure 3.9, all of the 8 indinavir related fragments (see Figure 3.6) got chosen by the EA initial generation, and the ones which were in average most used by the final generations of the EA experiments

were all from the indinavir fragments (fr.3 and fr.6) which is to be expected due to the nature of the similarity pressure. Besides the evident increase of fr.3 and fr.6 in occurrence, other fragments also got a moderate or slight increase in growth except fr.5 and fr.8. The slight or even negative growth of the 6 fragments could potentially mean that the descriptors used by the similarity measure are not sufficient to capture the variation of the structures, or there are some good substitutes coming from the NCI diversity set for the fragments.

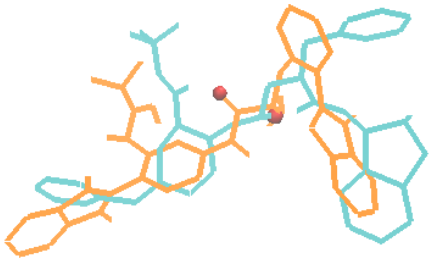


Figure 3.11: Superposition of the indinavir structure (in cyan) cut from the 1HSG complex with the AutoDock 4.2 predicted binding mode of molecule *no.*80 (in orange, marked with red frame in Figure 3.10).

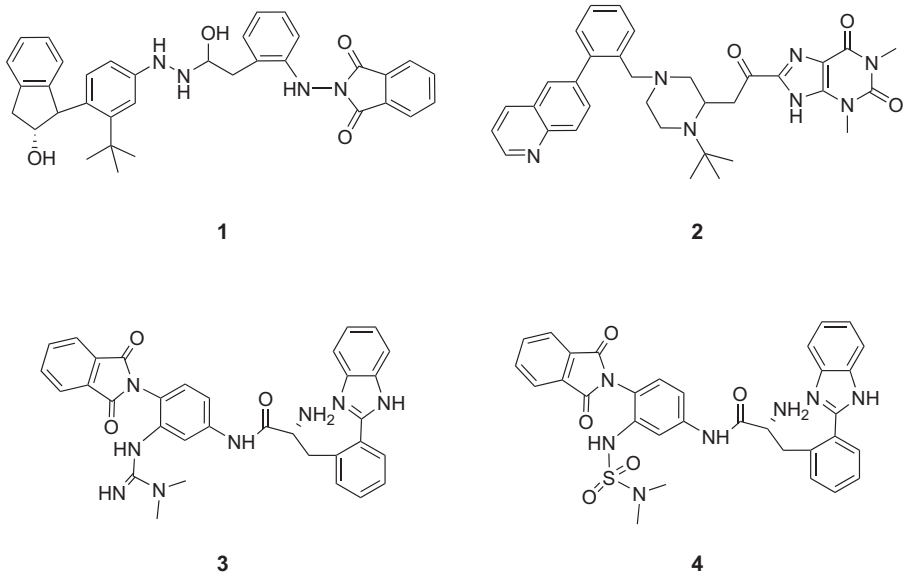


Figure 3.12: Some interesting ligands for HIV-1 protease given by GeneGear from the EA runs with the multiobjective function combining half-to-half the receptor- and ligand-based scoring strategy.

In the third type of design, we combined the receptor-based and ligand-based scoring strategy together to defined a half-to-half weighted multiobjective function. A graphical view of the outcome of one of the evolution experiments is shown in Figure 3.10, where individual *no.*80 of the population appeared in the last generation shows the highest fitness of both binding affinity to HIV-1 protease and similarity to indinavir. A superposition between the indinavir inhibitor cut from the 1HSG complex and the AutoDock 4.2 predicted binding mode of *no.*80 molecule (also referred to molecule *no.*4 in Figure 3.12) is given in Figure 3.11. As a whole, this molecule simulates indinavir well both in structure and conformation, and a carbonyl oxygen is attempting to perform similar function of the hydroxyl oxygen of indinavir though they are still quite distant in conformation. More interesting ligands for HIV-1 protease given by GeneGear output from the relevant experiments are shown in Figure 3.12.

Case Study II. In the past decades, *de novo* design has been mainly used in the design of biologically active molecules (usually drugs). However, one fact one cannot neglect is that novel compounds and materials with desired function and propertied are needed in many areas. One particularly interesting domain is transition metal catalysts. The following is a summary of using our fragment-based EA in the optimization of ruthenium catalyst for olefin metathesis [207–213]. More detailed information will be discussed elsewhere [214]. The available ruthenium olefin metathesis catalysts can be distinguished in two main classes: the first [207,208] and the second [209] generation Grubbs catalysts. The difference between these two classes lies in the kind of dative ligand L remaining in the active catalyst complex. Whereas the first generation Grubbs catalyst contains a phosphine-based ligand, the second generation contains an N-heterocyclic carbene (NHC)-based ligand (generally based on either imidazol-2-ylidene or dihydroimidazol-2-ylidene ring), and the second generation is observed to be more catalytic active than the first generation. The nature of this dative ligand significantly influences the catalytic activity. Most of the efforts at improving the performance of the ruthenium-based catalysts are, in fact, aimed at this dative ligand, by either modifying the substituents (branches) of existing phosphine or NHC structures, or changing the chemical nature of the ligand by more radical modification [211,212].

In our study, we allowed both a phosphine and an imidazol-2-ylidene based ruthenium catalyst frameworks (see Figure 3.13) to be involved in same evolution experiments, where their substitution sites are allowed to be randomly varied with structures constructed from a predefined fragment library of 1083 scaffolds and 1155 side chains. A PLSR-based QSAR model in which electronic

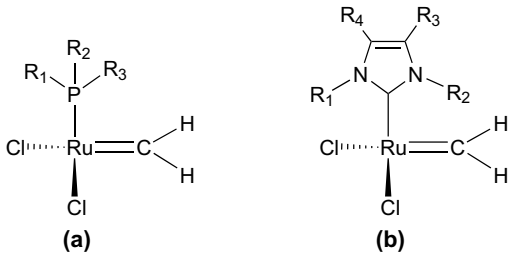


Figure 3.13: Two lead frameworks that are associated with a same ruthenium-based coordination center and different ligand skeletons (a phosphine **(a)** vs. a imidazol-2-ylidene based NHC **(b)**, which respectively dominate the chemical nature of the first and the second generation Grubbs catalysts) are used in same evolution experiments for competition.

and geometric descriptors, obtained at the semi-empirical PM6 level of theory, are correlated with catalytic activity (Q^2=0.85, RMSECV=1.46) was used as a fitness function to estimate the fitness of the newly generated structures.

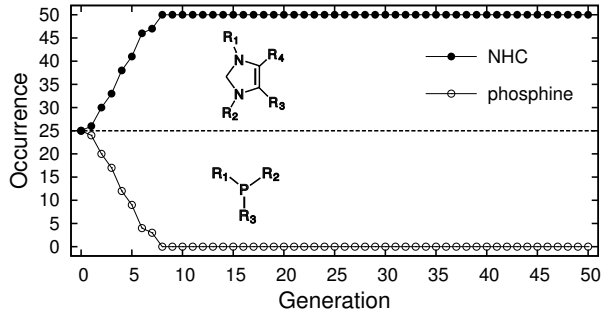


Figure 3.14: Occurrences of particular lead frameworks in conjunction with generation numbers. Phosphine- **(a)** and NHC- **(b)** based lead frameworks seen in Figure 3.13 are involved in the same evolution experiment.

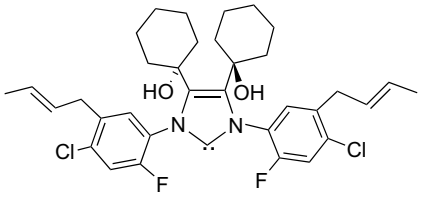


Figure 3.15: NHC-based LCl_2-Ru=CH_2 selected from the last generation of an EA, predicted to be highly active as a 14-electron complex.

In general, the EA optimizations are seen to favor catalysts known to be more active over less active ones: it successfully retraced the transition from the so-called phosphine-based first generation to the NHC-based second generation Grubbs catalysts. A trend plot of the occurrences of particular lead frameworks in conjunction with the generation numbers is shown in Figure 3.14. Finally, the optimizations also resulted in a number of new structures with high predicted activities, one of the example is shown in Figure 3.15. This work illustrated the potential of GeneGear for *in silico* development of transition metal catalysts and other functional coordination compounds.

3.7 Conclusion

We have described an open source software called GeneGear, which assists chemists in *de novo* design with multiple *in silico* methods. Depending on the application cases, the candidate structures can be sampled either by a systematic combinatorial library routine or by a stochastic evolutionary algorithm, while the former is biased to a design of compounds of same family, the latter is suited for a more diverse chemical space exploration. The fitness of the sampled structures can be estimated with either a structure-based docking prediction (usually for drug design), or a ligand-based property calculation, e.g., a similarity measure with a template structure or a QSAR regressed from a set of known compounds. Some complementary utility implementations such as designing a fragment library design, graphical visualization of a building block or a product set, and selecting an optimal subset are optionally supported in our software. The functionality and flexibility of GeneGear has been illustrated by several case studies related to the designs of functional drugs or catalysts. It shows that our software is quite effective in handling relevant problems and hopefully this open source tool will be used and further developed by scientists in the field.

3.8 Acknowledgement

The Department of Chemistry of NTNU is acknowledged for funding of this research. The NOTUR and NTNU local supercomputing programmes are thanked for provision of plenty of cpu time.

Chapter 4

A Knowledge-Based Approach for Screening Chemical Structures within *de novo* Molecular Evolution

Yunhan Chu and Bjørn K. Alsberg

Department of Chemistry
Norwegian University of Science and Technology
N-7491 Trondheim, Norway

This chapter has been published in the Journal of Chemometrics,
Volume 24 (2010), pages 399-407.

Is not included due to copyright

Chapter 5

De novo Optimization of Functional Coordination Compounds Using a Fragment-Based Evolutionary Algorithm

Yunhan Chu[†], Wouter Heyndrickx[‡], Giovanni Occhipinti[‡], Vidar R. Jensen[‡], and Bjørn K. Alsberg[†]

[†]Department of Chemistry, Norwegian University of Science and Technology, N-7491 Trondheim, [‡]Department of Chemistry, University of Bergen, Allégaten 41, N-5007 Bergen, Norway

This chapter is about to be submitted.

Is not included due to copyright

Chapter 6

Conclusions and Outlook

De novo design plays an important role in exploring chemical space, which allows a variety of computational knowledge, methods and tools to be implemented for producing novel chemical structures with desired properties. It constitutes a good complement to screening techniques.

An open source *de novo* software called GeneGear has been presented. In contrast to many known *de novo* tools which are highlighted with their a particular way of use, GeneGear assists chemists in multiple ways of designs. At current stage, both a systematic combinatorial and a stochastic evolutionary sampling routine can be implemented by GeneGear to generate novel molecular structures. Whereas the former biases to a family-based design, the latter suits a more diverse chemical space exploration. Structures generated from both are allowed to be scored of their quality (fitness) based on either separate structure-based and ligand-based evaluations or a combined function of them. The individual quality evaluations can be spread over multiple nodes in parallel on a cluster-type architecture, thereby enabling large-scale optimizations. In addition, some complementary methods, such as, design of a fragment library, graphical visualization of a building block or product set, and selection of an optimal structure subset are also supported. We demonstrated the functionality of GeneGear through several well-studied application examples where functional drugs and catalysts were required. The results shows that our software is quite effective at handling relevant problems.

Of course, the development of the software is still at the primary stage, there can be many cases which are outside the current processing ability of GeneGear.

Moreover, the performances of some important functional modules, such as 3D building and docking, are really depending on the explicit third-party softwares. Also, the software is a derivative from personal thesis work, most of the programs are written in script for Linux, while the user-friendliness has not been sufficiently addressed in the current work. These limitations should be addressed more carefully in the future development of GeneGear.

In GeneGear, evolutionary algorithm has an important role in the achievement of *de novo* creation of novel chemical structures. However, without explicit constraints, an EA tends sample chemically undesirable structures, which makes it necessary to constrain the structure space generated from *de novo* evolution. By applying data analytical methods from the fields of machine learning, chemometrics and multivariate statistics, we developed a knowledge-based approach which allows a user to use a set of predefined positive/negative molecules to create a bias filter to constrain all EA generated structures within the defined positive space. The BF approach requires no explicit formulation of structure constraining rules and allows the possibility of building a filter where the user does not know the underlying rules for what constitute an "acceptable" structure, which makes itself much intuitive and user friendly.

However, whether the approach is usable or not is really depending on whether the user can construct a sufficient predictive multivariate classification model. Moreover, there is always a risk that structures generated in the evolutionary process are outside the validity of the bias filter model. When this happen, the BF model must get updated or replaced. Some future work can be done to help the user refresh his BF model. For instance, to allow the EA to pause at specified steps for inclusion of new objects which are better sampled in the current region of the structure space and repetition of the model building.

In contrast to drug design in medicinal chemistry, automation and computer-aided synthesis have been comparably little appreciated in organometallic and coordination chemistry. Many of the available methods for drug design are not adapted to the structural variations of such type of compounds due to their ordinary construction rules in which knowledge about the coordination center and the neighboring ligands are not well addressed. We have constructed an EA method for *de novo* optimization of coordination compounds. By representing a 2D coordination structure in a graph with three kinds of fragment parts that have different construction requirements, i.e., a "core" part of coordination center environment, one or several "trial" parts consisting of meaningful ligand skeletons, and one or more "free" parts grown in diversity, we properly relate the structure variation with human knowledge. We use

three kinds of pattern-sensitive operations (growing, crossover, mutation) to sample candidate structures. High flexibility is permitted in definition of the fitness function. Except for special and well-parameterized cases, these are tasks requiring quantum chemical methods. The individual, fitness-generating calculations may be run in parallel, thereby enabling the EA method for large-scale optimizations. The capabilities of the EA method are illustrated by a series of representative searches for optimal ruthenium-based catalysts for olefin metathesis, where the fitness of the generated structures is assessed by a QSAR obtained at the semi-empirical PM6 level. The results demonstrate the high potential of our method in *in silico* development of transition metal catalysts and other functional coordination compounds.

As we have mentioned in the thesis, a fair fraction of the structures generated by the method are currently synthetically unobtainable. This can be due to a number of underlying causes, such as, single-property defined fitness, arbitrary building blocks, and/or insensible assembly schemes. However, "synthetic accessibility" is an important parameter in evaluating the quality of *de novo* generated molecules, which subsumes the availability of starting materials and the synthetic feasibility of the final product. Further efforts should be made in consideration how to address this issue well.

Bibliography

[1] J. W. Armstrong. A review of high-throughput screening approaches for drug discovery. *Am. Biotechnol. Lab.*, 17:26–28, 1999.

[2] P. Hecht. High-throughput screening: beating the odds with informatics-driven chemistry. *Curr. Drug Discov.*, pages 21–24, 2002.

[3] K. P. Mishra, L. Ganju, M. Sairam, P. K. Banerjee, and R. C. Sawhney. A review of high throughput technology for the screening of natural products. *Biomed. Pharmacother.*, 62:94–98, 2008.

[4] H. Xu and D. K. Agrafiotis. Retrospect and prospect of virtual screening in drug discovery. *Curr. Top. Med. Chem.*, 2:1305–1320, 2002.

[5] B. K. Shoichet. Virtual screening of chemical libraries. *Nature*, 432:862–865, 2004.

[6] T. Hou and X. Xu. Recent development and application of virtual screening in drug discovery: An overview. *Curr. Pharm. Des.*, 10:1011–1033, 2004.

[7] S. Ghosh, A. Nie, J. An, and Z. Huang. Structure-based virtual screening of chemical libraries for drug discovery. *Curr. Opin. Chem. Biol.*, 10:194–202, 2006.

[8] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.*, 3:935–949, 2004.

[9] G. Schneider and H. Böhm. Virtual screening and fast automated docking methods. *Comb. Chem.*, 7:64–70, 2002.

[10] P. Willett. Similarity-based approaches to virtual screening. *Biochem. Soc. Trans.*, 31:603–606, 2003.

[11] P. Willett, J. M. Barnard, and G. M. Downs. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.*, 38(6):983–996, 1998.

[12] J. Bajorath. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.*, 41:233–245, 2001.

[13] N. Nikolova and J. Jaworska. Approaches to measure chemical similarity - a review. *QSAR & Combi Sci*, 22:1006–1026, 2003.

[14] G. S. Chen, C.-S. Chang, W. M. Kan, C.-L. Chang, K. C. Wang, and J.-W. Chern. Novel lead generation through hypothetical pharmacophore three-dimensional database searching: Discovery of isoflavonoids as nonsteroidal inhibitors of rat 5 alpha-reductase. *J. Med. Chem.*, 44:3759–3763, 2001.

[15] V. J. Gillet. *De novo* molecular design. In D. E. Clark, editor, *Evolutionary algorithms in molecular design*, pages 49–69. Wiley-VCH, Weinheim, 2000.

[16] C. M. Dobson. Chemical space and biology. *Nature*, 432:824–828, 2004.

[17] L. Weber, S. Wallbaum, C. Broger, and K. Gubernator. Optimization of the biological activity of combinatorial compound libraries by a genetic algorithm. *Angew. Chem. Int. Ed. Engl.*, 34:2280–2282, 1995.

[18] K. Illgen, T. Enderle, C. Broger, and L. Weber. Simulated molecular evolution in a full combinatorial library. *Chem. Biol.*, 7:433–441, 2000.

[19] D. Douguet, E. Thoreau, and G. A. Grassy. Genetic algorithm for the automated generation of small organic molecules: Drug design using an evolutionary algorithm. *J. Comput. Aided Mol. Des.*, 14:449–466, 2000.

[20] J. Meiler and M. Will. Automated structure elucidation of organic molecules from 13C NMR spectra using genetic algorithms and neural networks. *J. Chem. Inf. Comput. Sci.*, 41:1535–1546, 2001.

[21] V. Venkatasubramanian, K. Chan, and J. M. Caruthers. Evolutionary design of molecules with desired properties using the genetic algorithm. *J. Chem. Inf. Comput. Sci.*, 35:188–195, 1995.

[22] D. Douguet, H. Munier-Lehmann, G. Labesse, and S. Pochet. LEA3D: A computer-aided ligand design for structure-based drug design. *J. Med. Chem.*, 48:2457–2468, 2005.

[23] R. B. Nachbar. Molecular evolution: Automated manipulation of hierarchical chemical topology and its application to average molecular structures. *Genet. Program. Evol. Mach.*, 1:57–94, 2000.

[24] A. Globus, J. Lawton, and T. Wipke. Automatic molecular design using evolutionary techniques. *Nanotechnology*, 10:290–299, 1999.

[25] N. Brown, B. McKay, F. Gilardoni, and J. Gasteiger. A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *J. Chem. Inf. Comput. Sci.*, 44:1079–1087, 2004.

[26] Y. Han and C. Steinbeck. Evolutionary-algorithm-based strategy for computer-assisted structure elucidation. *J. Chem. Inf. Comput. Sci.*, 44:489–498, 2004.

[27] S. C. Pegg, J. J. Haresco, and I. D. Kuntz. A genetic algorithm for structure-based *de novo* design. *J. Comput. Aided Mol. Des.*, 15:911–933, 2001.

[28] H.-J. Böhm. The computer program LUDI: A new method for the *de novo* design of enzyme inhibitors. *J. Comput. Aided Mol. Des.*, 6:67–78, 1992.

[29] H.-J. Böhm. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J. Comput. Aided Mol. Des.*, 6:593–606, 1992.

[30] D. E. Clark, D. Frenkel, S. A. Levy, J. Li, C. W. Murray, B. Robson, B. Waszkowycz, and D. R. Westhead. PRO_LIGAND: An approach to *de novo* molecular design. 1. application to the design of organic molecules. *J. Comput. Aided Mol. Des.*, 9:13–32, 1995.

[31] R. Wang, Y. Gao, and L. Lai. LigBuilder: A multi-purpose program for structure-based drug design. *J. Mol. Model.*, 6:498–516, 2000.

[32] M. B. Eisen, D. C. Wiley, M. Karplus, and R. E. Hubbard. HOOK: a program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins: Struct. Funct. Genet.*, 19:199–221, 1994.

[33] V. Tschinke and N. C. Cohen. The NEWLEAD program: a new method for the design of candidate structures from pharmacophoric hypotheses. *J. Med. Chem.*, 36:3863–3870, 1993.

[34] C. M. W. Ho and G. R. Marshall. SPLICE: A program to assemble partial query solutions from three-dimensional database searches into novel ligands. *J. Comput. Aided Mol. Des.*, 7:623–647, 1993.

[35] V. Gillet, A. P. Johnson, P. Mata, S. Sike, and P. Williams. SPROUT: A program for structure generation. *J. Comput. Aided Mol. Des.*, 7:127–153, 1993.

[36] V. J. Gillet, W. Newell, P. Mata, G. Myatt, S. Sike, Z. Zsoldos, and A. P. Johnson. SPROUT: Recent developments in the *de novo* design of molecules. *J. Chem. Inf. Comput. Sci.*, 34:297–217, 1994.

[37] P. Mata, V. J. Gillet, A. P. Johnson, J. Lampreia, G. J. Myatt, S. Sike, and A. L. Stebbings. SPROUT: 3D structure generation using templates. *J. Chem. Inf. Comput. Sci.*, 35:479–493, 1995.

[38] V. J. Gillet, A. P. Johnson, P. Mata, and S. Sike. Automated structure design in 3D. *Tetrahedron Comput. Method.*, 3:681–696, 1990.

[39] S. H. Rotstein and M. A. Murcko. GenStar: A method for *de novo* drug design. *J. Comput. Aided Mol. Des.*, 7:23–43, 1993.

[40] Z. Luo, R. Wang, and L. Lai. RASSE: A new method for structure-based drug design. *J. Chem. Inf. Comput. Sci.*, 36:1187–1194, 1996.

[41] S. H. Rotstein and M. A. Murcko. GroupBuild: a fragment-based method for *de novo* drug design. *J. Med. Chem.*, 36:1700–1710, 1993.

[42] R. S. Bohacek and C. McMartin. Multiple highly diverse structures complementary to enzyme binding sites: Results of extensive application of a *de novo* design method incorporating combinatorial growth. *J. Am. Chem. Soc.*, 116:5560–5571, 1994.

[43] R. S. DeWitte and E. I. Shakhnovich. SMoG: *de novo* design method based on simple, fast, and accurate free energy estimates. 1. methodology and supporting evidence. *J. Am. Chem. Soc.*, 118:11733–11744, 1996.

[44] A. V. Ishchenko and E. I. Shakhnovich. Small molecule growth 2001 (SMoG2001): An improved knowledge-based scoring function for protein-ligand interactions. *J. Med. Chem.*, 45:2770–2780, 2002.

[45] Y. Nishibata and A. Itai. Automatic creation of drug candidate structures based on receptor structure. starting point for artificial lead generation. *Tetrahedron*, 47:8985–8990, 1991.

[46] R. A. Lewis. Automated site-directed drug design: Approaches to the formation of 3D molecular graphs. *J. Comput. Aided Mol. Des.*, 4:205–210, 1990.

[47] R. A. Lewis, D. C. Roe, C. Huang, T. E. Ferrin, R. Langridge, and I. D. Kuntz. Automated site-directed drug design using molecular lattices. *J. Mol. Graphics*, 10:66–78, 1992.

[48] D. C. Roe and I. D. Kuntz. BUILDER v.2: improving the chemistry of a *de novo* design strategy. *J. Comput. Aided Mol. Des.*, 9:269–282, 1995.

[49] D. A. Pearlman and M. A. Murcko. CONCEPTS: New dynamic algorithm for *de novo* drug suggestion. *J. Comp. Chem.*, 14:1184–1193, 1993.

[50] A. Miranker and M. Karplus. An automated method for dynamic ligand design. *Proteins: Struct. Funct. Genet.*, 23:472–490, 1995.

[51] D. K. Gehlhaar, K. E. Moerder, D. Zichi, and Christopher. *De novo* design of enzyme inhibitors by monte carlo ligand generation. *J. Med. Chem.*, 38:466–472, 1995.

[52] D. A. Pearlman and M. A. Murcko. CONCERTS: Dynamic connection of fragments as an approach to *de novo* ligand design. *J. Med. Chem.*, 39:1651–1663, 1996.

[53] N. Todorov and P. Dean. Evaluation of a method for controlling molecular scaffold diversity in *de novo* ligand design. *J. Comput. Aided Mol. Des.*, 11:175–192, 1997.

[54] E. Pellegrini and M. J. Field. Development and testing of a *de novo* drug-design algorithm. *J. Comput. Aided Mol. Des.*, 17:621–641, 2003.

[55] I. Huc and J.-M. Lehn. Virtual combinatorial libraries: Dynamic generation of molecular and supramolecular diversity by self-assembly. *Proc. Natl. Acad. Sci.*, 94:2106–2110, 1997.

[56] E. K. Kick, D. C. Roe, A. G. Skillman, G. Liu, T. J. Ewing, Y. Sun, I. D. Kuntz, and J. A. Ellman. Structure-based design and combinatorial nanomolar inhibitors of cathepsin d. *Chem. Biol.*, 4:297–307, 1997.

[57] C. W. Murray, D. E. Clark, T. R. Auton, M. A. Firth, J. Li, R. A. Sykes, B. Waszkowycz, D. R. Westhead, and S. C. Young. PRO_SELECT: Combining structure-based drug design and combinatorial chemistry for rapid lead discovery. 1. technology. *J. Comput. Aided Mol. Des.*, 11:193–207, 1997.

[58] V. J. Gillet, G. Myatt, Z. Zsoldos, and A. P. Johnson. SPROUT, HIPPO and CAESA: Tools for *de novo* structure generation and estimation of synthetic accessibility. *Perspect. Drug Discov. Des.*, 3:34–50, 1995.

[59] H. M. Vinkers, M. R. de Jonge, F. F. D. Daeyaert, J. Heeres, L. M. H. Koymans, J. H. van Lenthe, P. J. Lewi, H. Timmerman, K. Van Aken, and P. A. J. Janssen. SYNOPSIS: Synthesize and optimize system in silico. *J. Med. Chem.*, 46:2765–2773, 2003.

[60] J. D. Durrant, R. E. Amaro, and J. A. McCammon. AutoGrow: A novel algorithm for protein inhibitor design. *Chem. Biol. Drug Des.*, 73:168–178, 2009.

[61] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, 261:470–489, 1996.

[62] T. J. A. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.*, 15:411–428, 2001.

[63] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J. Comp. Chem.*, 19:1639–1662, 1998.

[64] B. Waszkowycz, D. E. Clark, D. Frenkel, J. Li, C. W. Murray, B. Robson, and D. R. Westhead. PRO_LIGAND: An approach to *de novo* molecular design. 2. design of novel molecules from molecular field analysis (mfa) models and pharmacophores. *J. Med. Chem.*, 37:3994–4002, 1994.

[65] G. Schneider, M.-L. Lee, M. Stahl, and P. Schneider. *De novo* design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comput. Aided Mol. Des.*, 14:487–494, 2000.

[66] J. Devillers. Designing molecules with specific properties from intercommunicating hybrid systems. *J. Chem. Inf. Comput. Sci.*, 36:1061–1066, 1996.

[67] R. C. Glen and A. W. R. Payne. A genetic algorithm for the automated generation of molecules within constraints. *J. Comput. Aided Mol. Des.*, 9:181–202, 1995.

[68] V. J. Gillet, W. Khatib, P. Willett, P. J. Fleming, and D. V. S. Green. Combinatorial library design using a multiobjective genetic algorithm. *J. Chem. Inf. Comput. Sci.*, 42:375–385, 2002.

[69] C. A. Nicolaou, J. Apostolakis, and C. S. Pattichis. *De novo* drug design using multiobjective evolutionary graphs. *J. Chem. Inf. Model.*, 49:295–307, 2009.

[70] J. Holland. *Adaptation in Natural and Artificial Systems.* University of Michigan Press, 1975.

[71] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, Boston, 1989.

[72] L. Davis. *Handbook of genetic algorithms.* Van Nostrand Reinhold, New York, 1991.

[73] J. R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* MIT Press, Cambridge, 1992.

[74] W. Banzhaf, P. Nordin, R. E. Keller, , and F. D. Francone. *Genetic Programming - An Introduction: On the Automatic Evolution of Computer Programs and Its Applications.* Morgan Kaufmann, California, 1998.

[75] H.-P. Schwefel. *Evolution and Optimum Seeking.* John Wiley & Sons, New York, 1995.

[76] H.-G. Beyer. *The Theory of Evolution Strategies.* Springer-Verlag, Heidelberg, 2001.

[77] L. J. Fogel, A. J. Owens, and M. J. Walsh. *Artificial Intelligence through Simulated Evolution.* John Wiley & Sons, New York, 1966.

[78] L. J. Fogel. *Intelligence through Simulated Evolution: Forty Years of Evolutionary Programming.* John Wiley & Sons, New York, 1999.

[79] D. E. Clark and D. R. Westhead. Evolutionary algorithms in computer-aided molecular design. *J. Comput. Aided Mol. Des.*, 10:337–358, 1996.

[80] D. E. Clark. *Evolutionary algorithms in molecular design.* Wiley-VCH, Weinheim, 2000.

[81] E. Lameijer, T. Bäck, J. Kok, and A. P. Ijzerman. Evolutionary algorithms in drug design. *Nat. Comput.*, 4:177–243, 2005.

[82] G. Schneider and U. Fechner. Computer-based *de novo* design of drug-like molecules. *Nat. Rev. Drug Discov.*, 4:649–663, 2005.

[83] U. Fechner and G. Schneider. Flux (1): A virtual synthesis scheme for fragment-based *de novo* design. *J. Chem. Inf. Model.*, 46:699–707, 2006.

[84] F. Dey and A. Caflisch. Fragment-based *de novo* ligand design by multiobjective evolutionary optimization. *J. Chem. Inf. Model.*, 48:679–690, 2008.

[85] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The chemistry development kit (CDK): An open-source java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.*, 43:493–500, 2003.

[86] Jmol: an open-source Java viewer for chemical structures in 3D. http://jmol.sourceforge.net.

[87] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor.*, 11:10–18, 2009.

[88] W. H. Wei and G. J. Chen. JavaStatSoft: design patterns and features. *Comput. Stat.*, 23:235–251, 2008.

[89] JFreeChart: A free java chart library. http://www.jfree.org/jfreechart.

[90] O. Trott and A. J. Olson. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comp. Chem.*, 31:455–461, 2010.

[91] T. I. Oprea and J. Gottfries. Chemography: The art of navigating in chemical space. *J. Comb. Chem.*, 3:157–166, 2001.

[92] N. Prakash and D. A. Gareja. Cheminformatics. *J. Proteomics. Bioinform.*, 3:249–252, 2010.

[93] C. Lipinski and A. Hopkins. Navigating chemical space for biology and medicine. *Nature*, 432:855–861, 2004.

[94] T. N. Doman, S. L. McGovern, B. J. Witherbee, T. P. Kasten, R. Kurumbail, W. C. Stallings, D. T. Connolly, and B. K. Shoiche. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1b. *J. Med. Chem.*, 45:2213–2221, 2002.

[95] C. G. Wermuth, C. R. Ganellin, P. Lindberg, and A. Mitscher, L. Glossary of terms used in medicinal chemistry. *Pure Appl. Chem.*, 70:1129–1143, 1998.

[96] T. Honma. Recent advances in *de novo* design strategy for practical lead identification. *Med. Res. Rev.*, 23:606–632, 2003.

[97] R. E. Babine, T. M. Bleckman, C. R. Kissinger, R. Showalter, L. A. Pelletier, C. Lewis, K. Tucker, E. Moomaw, H. E. Parge, and J. E. Villafranca. Design, synthesis and X-ray crystallographic studies of novel FKBB-12 ligands. *Bioorg. Med. Chem. Lett.*, 5:1719–1724, 1995.

[98] D. J. Danziger and P. M. Dean. Automated site-directed drug design: A general algorithm for knowledge acquisition about hydrogen-bonding regions at protein surfaces. *Proc. R. Soc. Lond. B*, 236:101–113, 1989.

[99] P. J. Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, 28:849–857, 1985.

[100] A. Miranker and M. Karplus. Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins: Struct. Funct. Genet.*, 11:29–34, 1991.

[101] C. Lipinski, F. Lombardo, B. Dominy, and P. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, 23:3–25, 1997.

[102] X. Q. Lewell, D. B. Judd, S. P. Watson, and M. M. Hann. RECAP – retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.*, 38:511–522, 1998.

[103] T. Honma, K. Hayashi, T. Aoyama, N. Hashimoto, T. Machida, K. Fukasawa, T. Iwama, C. Ikeura, M. Ikuta, I. Suzuki-Takahashi, Y. Iwasawa, T. Hayama, S. Nishimura, and H. Morishima. Structure-based generation of a new class of potent CDK4 inhibitors: New *de novo* design strategy and library design. *J. Med. Chem.*, 44:4615–4627, 2001.

[104] R. A. Lewis and P. M. Dean. Automated site-directed drug design: The formation of molecular templates in primary structure generation. *Proc. R. Soc. Lond. B*, 236:141–162, 1989.

[105] B. O. Brandsdal, F. Österberg, M. Almlöf, and I. Feierberg. Free energy calculations and ligand binding. *Adv. Protein Chem.*, 66:123–158, 2003.

[106] J. Apostolakis and A. Caflisch. Computational ligand design. *Comb. Chem. High Throughput Screen.*, 2:91–104, 1999.

[107] H. Liu, Z. Duan, Q. Luo, and Y. Shi. Structure-based ligand design by dynamically assembling molecular building blocks at binding site. *Proteins: Struct. Funct. Genet.*, 36:462–470, 1999.

[108] J. Zhu, H. Fan, H. Liu, and Y. Shi. Structure-based ligand design for flexible proteins: Application of new F-DycoBlock. *J. Comput. Aided Mol. Des.*, 15:979–996, 2001.

[109] M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini, and R. P. Mee. Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.*, 11:425–445, 1997.

[110] I. Muegge and Y. C. Martin. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.*, 42:791–804, 1999.

[111] I. Muegge. A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. *Perspect. Drug Discov. Des.*, 20:99–114, 2000.

[112] H. Gohlke, M. Hendlich, and G. Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.*, 295:337–356, 2000.

[113] C. M. Fonseca and P. J. Fleming. Genetic algorithms for multiobjective optimization: Formulation, discussion and generalization. In S. Forrest, editor, *Genetic Algorithms: Proceedings of the Fifth International Conference*, pages 416–423. Morgan Kaufmann, San Mateo, CA, 1993.

[114] D. K. Agrafiotis. Multiobjective optimization of combinatorial libraries. *IBM J. RES. & DEV. VOL.*, 45:545–566, 2001.

[115] T. Wright, V. J. Gillet, D. V. S. Green, and S. D. Pickett. Optimizing the size and configuration of combinatorial libraries. *J. Chem. Inf. Comput. Sci.*, 43:381–390, 2003.

[116] D. Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36, 1988.

[117] D. Weininger, A. Weininger, Weininger, and J. L. SMILES 2. algorithm for generation of unique smiles notation. *J. Chem. Inf. Comput. Sci.*, 29:97–101, 1989.

[118] J. Zhu, H. Yu, H. Fan, H. Liu, and Y. Shi. Design of new selective inhibitors of cyclooxygenase-2 by dynamic assembly of molecular building blocks. *J. Comput. Aided Mol. Des.*, 15:447–463, 2001.

[119] J. L. Faulon. Stochastic generator of chemical structure. 2. Using simulated annealing to search the space of constitutional isomers. *J. Chem. Inf. Comput. Sci.*, 36(4):731–740, 1996.

[120] J. Gasteiger, C. Rudolph, and J. Sadowski. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput. Method.*, 3:537–547, 1990.

[121] J. Sadowski, C. H. Schwab, and J. Gasteiger. 3D structure generation and conformational searching. In W. Langenaeker, H. D. Winter, P. Bultinck, and J. P. Tollenaere, editors, *Computational Medicinal Chemistry for Drug Discovery*, pages 151–212. Marcel Dekker, New York, 2003.

[122] R. S. Pearlman. Rapid generation of high quality approximate 3D molecular structures. *Chem. Design Auto. News*, 2:1–6, 1987.

[123] R. S. Pearlman. 3D molecular structures: Generation and use in 3D searching. In H. Kubinyi, editor, *3D QSAR in Drug Design: Theory, Methods and Applications*, pages 41–79. ESCOM Science, Leiden, 1993.

[124] A. C. Pierce, G. Rao, and G. W. Bemis. BREED: generating novel inhibitors through hybridization of known ligands application to CDK2, P38, and HIV protease. *J. Med. Chem.*, 47:2768–2775, 2004.

[125] D. Weininger. Combinatoris of organic molecular structures. In P. von Ragué Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. Kollman, H. F. S. III, and P. R. Schreiner, editors, *Encyclopedia of Computational Chemistry*, pages 425–530. Wiley, New York, 1998.

[126] J. M. Spivey. Algebras for combinatorial search. *J. Funct. Program.*, 19:469–487, 2009.

[127] N. Todorov and P. Dean. A branch-and-bound method for optimal atom-type assignment in *de novo* ligand design. *J. Comput. Aided Mol. Des.*, 12:335–349, 1998.

[128] C. Darwin. *On the Origin of Species* Facsimile of the First Edition. Harvard Univ. Press, Cambridge, MA, [1859]1975.

[129] S. N. Sivanandam and S. N. Deepa. *Introduction to genetic algorithms.* Springer-Verlag, Berlin Heidelberg, 2008.

[130] A. Caflisch, R. Wälchli, and C. Ehrhardt. Computer-aided design of thrombin inhibitors. *News Physiol. Sci.*, 13:182–189, 1998.

[131] L. de la Sierra I, H. Munier-Lehmann, A. M. Gilles, O. Bârzu, and M. Delarue. X-ray structure of TMP kinase from mycobacterium tuberculosis complexed with tmp at 1.95 a resolution. *J. Mol. Biol.*, 311:87–100, 2001.

[132] S. Pochet, L. Dugue, D. Douguet, G. Labesse, and H. Munier-Lehmann. Nucleoside analogues as inhibitors of thymidylate kinases: possible therapeutic applications. *ChemBioChem.*, 3:108–110, 2002.

[133] A. Brik and C.-H. Wong. HIV-1 protease: mechanism and drug discovery. *Org. Biomol. Chem.*, 1:5–14, 2003.

[134] D. Joseph-McCarthy. Computational approaches to structure-based ligand design. *Pharmacology & Therapeutics*, 84:179–191, 1999.

[135] G. Klebe. Recent developments in structure-based drug design. *J. Mol. Med.*, 78:269–281, 2000.

[136] M. Havranek, A. Singh, and D. Sames. Evolution and study of polymer-supported metal catalysts for oxygen atom transfer: Oxidation of alkanes and alkenes by diamide manganese complexes. *J. Am. Chem. Soc.*, 121:8965–8966, 1999.

[137] J. M. Thomas. Design, synthesis, and in situ characterization of new solid catalysts. *Angew. Chem., Int. Ed.*, 38:3588–3628, 1999.

[138] D. Farrusseng. High-throughput heterogeneous catalysis. *Surf. Sci. Reports*, 63:487–513, 2008.

[139] D. N. Bolon, C. A. Voigt, and S. L. Mayo. *De novo* design of biocatalysts. *Curr. Opin. Chem. Biol.*, 6:125–129, 2002.

[140] R. Sterner and F. X. Schmid. *De novo* design of an enzyme. *Science*, 304:1916–1917, 2004.

[141] B. I. Dahiyat and S. L. Mayo. *De novo* protein design: Fully automated sequence selection. *Science*, 278:82–87, 1997.

[142] W. F. DeGrado and C. M. Summa. *De novo* design and structural characterization of proteins and metalloproteins. *Annu. Rev. Biochem.*, 68:779–819, 1999.

[143] C. Floudas, H. Fung, S. McAllister, M. Mönnigmann, and R. Rajgaria. Advances in protein structure prediction and *de novo* protein design: A review. *Chem. Eng. Sci.*, 61:966–988, 2006.

[144] H. K. Fung, W. J. Welsh, and C. A. Floudas. Computational *de novo* peptide and protein design: Rigid templates versus flexible templates. *Ind. Eng. Chem. Res.*, 47:993–1001, 2008.

[145] R. Huey, G. M. Morris, A. J. Olson, and D. S. Goodsell. A semiempirical free energy force field with charge-based desolvation. *J. Comp. Chem.*, 28:1145–1152, 2007.

[146] R. Guha. Methods to improve the reliability, validity and interpretability of QSAR models. Ph.D. Thesis, The Pennsylvania State University, 2005.

[147] S. Wold, A. Ruhe, H. Wold, and W. J. D. III. The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses. *SIAM J. Sci. and Stat. Comput.*, 5:735–743, 1984.

[148] A. Höskuldsson. PLS regression methods. *J. Chemometrics.*, 2:211–228, 1988.

[149] M. J. Vainio and M. S. Johnson. Generating conformer ensembles using a multiobjective genetic algorithm. *J. Chem. Inf. Model.*, 47:2462–2474, 2007.

[150] R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. K. Wegner, and E. Willighagen. The Blue Obelisk - interoperability in chemical informatics. *J. Chem. Inf. Model.*, 46:991–998, 2006.

[151] OpenBabel. http://openbabel.sourceforge.net.

[152] ChemAxon Ltd., Budapest, HU, http://www.chemaxon.com.

[153] G. Imre, A. Kalászi, I. Jákli, and O. . Farkas. Advanced automatic generation of 3D molecular structures. *European Chemistry Congress*, 1:27–31, 2006.

[154] T. A. Halgren. Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *J. Comp. Chem.*, 17:490–519, 1996.

[155] G. Imre, G. Veress, A. Volford, and O. . Farkas. Molecules from the minkowski space: an approach to building 3D molecular structures. *J. Mol. Struct. (Theochem)*, 666-667:51–59, 2003.

[156] S. L. Mayo, B. D. Olafson, and W. A. I. Goddard. Dreiding: A generic force field for molecular simulations. *J. Phys. Chem.*, 94:8897–8909, 1990.

[157] M. Snir, S. Otto, S. Huss-Lederman, D. Walker, and J. Dongarra. *MPI: The Complete Reference*. MIT Press, Cambridge, 1995.

[158] S. L. Holbeck. Update on NCI in vitro drug screen utilities. *Eur. J. Cancer*, 40:785–793, 2004.

[159] N. K. Terrett, M. Gardner, D. W. Gordon, R. J. Kobylecki, and J. Steele. Combinatorial synthesis - the design of compound libraries and their application to drug discovery. *Tetrahedron*, 51:8135–8137, 1995.

[160] S. Rose. Statistical design and application to combinatorial chemistry. *Comb. Chem.*, 7:133–138, 2002.

[161] R. E. Higgs, K. G. Bemis, I. A. Watson, and J. H. Wikel. Experimental designs for selecting molecules from large chemical databases. *J. Chem. Inf. Comput. Sci.*, 37:861–870, 1997.

[162] P. M. Andersson, M. Sjöström, S. Wold, and T. Lundstedt. Strategies for subset selection of parts of an in-house chemical library. *J. Chemom.*, 15:353–369, 2001.

[163] P. M. Andersson and T. Lundstedt. Hierarchical experimental design exemplied by QSAR evaluation of a chemical library directed towards the melanocortin 4 receptor. *J. Chemom.*, 16:490–496, 2002.

[164] L. Eriksson, T. Arnhold, B. Beck, T. Fox, E. Johansson, and J. M. Kriegl. Onion design and its application to a pharmaceutical QSAR problem. *J. Chemom.*, 18:188–202, 2004.

[165] I.-M. Olsson, J. Gottfries, and S. Wold. Controlling coverage of D-optimal onion designs and selections. *J. Chemom.*, 18:548–557, 2004.

[166] P. F. de Aguiar, B. Bourguignon, M. S. Khots, D. L. Massart, and R. Phan-Than-Luu. D-optimal designs. *Chemom. Intell, Lab. Syst.*, 30:199–210, 1995.

[167] E. Marengo and R. Todeschini. A new algorithm for optimal, distance-based experimental design. *Chemom. Intell, Lab. Syst.*, 16:37–44, 1992.

[168] M. S. Lajiness. Dissimilarity-based compound selection techniques. *Perspect. Drug Discov. Des.*, 7:65–84, 1997.

[169] J. D. Holliday, S. S. Ranade, and P. Willett. A fast algorithm for selecting sets of dissimilar molecules from large chemical databases. *Quant. Struct. Act. Relat.*, 14:501–506, 2002.

[170] B. D. Hudson, R. M. Hyde, E. Rahr, and J. Wood. Parameter based methods for compound selection from chemical databases. *Quant. Struct. Act. Relat.*, 15:285–289, 1996.

[171] R. L. H. Lam, W. J. Welch, and S. S. Young. Uniform coverage designs for molecule selection. *Technometrics*, 44:99–109, 2002.

[172] R. D. Brown and Y. C. Martin. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.*, 36:572–584, 1996.

[173] A. Linusson, S. Wold, and B. Nordén. Fuzzy clustering of 627 alcohols, guided by a strategy for cluster analysis of chemical compounds for combinatorial chemistry. *Chemom. Intell, Lab. Syst.*, 44:213–227, 1998.

[174] I.-M. Olsson, J. Gottfries, and S. Wold. D-optimal onion designs in statistical molecular design. *Chemom. Intell, Lab. Syst.*, 73:37–46, 2004.

[175] R. Wootton, R. Cranfield, G. C. Sheppey, and P. J. Goodford. Physicochemical-activity relations in practice. 2. rational selection of benzenoid substituents. *J. Med. Chem.*, 18:607–613, 1975.

[176] B. A. Leland, B. D. Christie, J. G. Nourse, D. L. Grier, R. E. Carhart, T. Maffett, S. M. Welford, and D. H. Smith. Managing the combinatorial explosion. *J. Chem. Inf. Comput. Sci.*, 37:62–70, 1997.

[177] V. S. Lobanov and D. K. Agrafiotis. Scalable methods for the construction and analysis of virtual combinatorial libraries. *Comb. Chem. High Throughput Screen.*, 5:167–178, 2002.

[178] A. Linusson, S. Wold, and B. Nordén. Statistical molecular design of peptoid libraries. *Mol. Divers.*, 4:103–114, 1999.

[179] A. Linusson, J. Gottfries, F. Lindgren, and S. Wold. Statistical molecular design of building blocks for combinatorial chemistry. *J. Med. Chem.*, 43:1320–1328, 2000.

[180] A. Linusson, J. Gottfries, T. Olsson, E. Örnskov, S. Folestad, B. Nordén, and S. Wold. Statistical molecular design, parallel synthesis, and biological evaluation of a library of thrombin inhibitors. *J. Med. Chem.*, 44:3424–3439, 2001.

[181] M. Böhm, J. Stürzebecher, and G. Klebe. Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor xa. *J. Med. Chem.*, 42:458–477, 1999.

[182] D. T. Stanton and P. C. Jurs. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. *Analytical Chemistry*, 62:2323–2329, 1990.

[183] A. R. Katritzky, L. Mu, V. S. Lobanov, and M. Karelson. Correlation of boiling points with molecular structure. 1. a training set of 298 diverse organics and a test set of 9 simple inorganics. *J. Phys. Chem.*, 100:10400–10407, 1996.

[184] M. D. Wessel, P. C. Jurs, J. W. Tolan, and S. M. Muskal. Prediction of human intestinal absorption of drug compounds from molecular structure. *J. Chem. Inf. Comput. Sci.*, 38:726–735, 1998.

[185] P. C. Jurs, J. T. Chou, and M. Yuan. Studies of chemical structure biological activity relations using pattern recognition. In R. E. C. Edward C. Olson, editor, *Computer Assisted Drug Design*, pages 103–129. American Chemical Society, Washington D.C., 1979.

[186] J. Stuper, W. E. Brügger, and P. C. Jurs. *Computer Assisted Studies of Chemical Structure and Biological Function.* John Wiley & Sons, New York, 1979.

[187] S. Liu, C. Cao, and Z. Li. Approach to estimation and prediction for normal boiling point (NBP) of alkanes based on a novel molecular distance-edge (MDE) vector, lambda. *J. Chem. Inf. Comput. Sci.*, 38:387–394, 1998.

[188] H. Goldstein. *Classical mechanics.* Addison-Wesley, Cambridge, MA, 1950.

[189] M. Petitjean. Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *J. Chem. Inf. Comput. Sci.*, 32:331–337, 1992.

[190] P. Bath, A. R. Poirrette, P. Willett, and F. H. Allen. The extent of the relationship between the graph-theoretical and the geometrical shape coefficients of chemical compounds. *J. Chem. Inf. Comput. Sci.*, 35:714–716, 1995.

[191] R. Todeschini and P. Gramatica. New 3D molecular descriptors: the WHIM theory and QSAR applications. *Perspect. Drug Discov. Des.*, 2:355–380, 1998.

[192] P. Ertl, B. Rohde, and P. Selzer. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.*, 43:3714–3717, 2000.

[193] R. Wang, Y. Fu, and L. Lai. A new atom-additive method for calculating partition coefficients. *J. Chem. Inf. Comput. Sci.*, 37(3):615–621, 1999.

[194] R. Wang, Y. Gao, and L. Lai. Calculating partition coefficient by atom-additive method. *Perspect. Drug Discov. Des.*, 19(1):47–66, 2000.

[195] Alternative Periodic Table of the Elements. http://www.sunysccc.edu/academic/mst/ptable/p-table2.htm.

[196] F. C. Bernstein, T. F. Koetzle, G. J. Williams, J. Meyer, E. E., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542, 1977.

[197] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res.*, 28:235–242, 2000.

[198] D. Turk, J. Sturzebecher, and W. Bode. Geometry of binding of the n alpha-tosylated piperidides of m-amidino-, p-amidino- and p-guanidino phenylalanine to thrombin and trypsin. X-ray crystal structures of their trypsin complexes and modeling of their thrombin complexes. *FEBS Lett.*, 287:133–138, 1991.

[199] M. W. Chang, C. Ayeni, S. Breuer, and B. E. Torbett. Virtual screening for HIV protease inhibitors: A comparison of AutoDock 4 and Vina. *PLoS ONE*, 5:e11955, 2010.

[200] C. Chu and B. K. Alsberg. A knowledge-based approach for screening chemical structures within *de novo* molecular evolution. *J. Chemom.*, 24:399–407, 2010.

[201] C. Li, L. Xu, D. W. Wolan, I. A. Wilson, and A. J. Olson. Virtual screening of human 5-aminoimidazole-4-carboxamide ribonucleotide transformylase against the NCI diversity set by use of AutoDock to identify novel nonfolate inhibitors. *J. Med. Chem.*, 47:6681–6690, 2004.

[202] K. Choowongkomon, O. Sawatdichaikul, N. Songtawee, and J. Limtrakul. Receptor-based virtual screening of EGFR kinase inhibitors from the NCI diversity database. *Molecules*, 15:4041–4054, 2010.

[203] A. Spannhoff, R. Heinke, I. Bauer, P. Trojer, E. Metzger, R. Gust, R. Schule, G. Brosch, W. Sippl, and M. Jung. Target-based approach to inhibitors of histone arginine methyltransferases. *J. Med. Chem.*, 50:2319–2325, 2007.

[204] A. Wlodawer and J. Vondrasek. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. *Annu. Rev. Biophys. Biomol. Struct.*, 27:249–284, 1998.

[205] A. Wlodawer. Rational approach to aids drug design through structural biology. *Annu. Rev. Med.*, 53:595–614, 2002.

[206] Z. Chen, Y. Li, E. Chen, D. L. Hall, P. L. Darke, C. Culberson, J. A. Shafer, and L. C. Kuo. Crystal structure at 1.9-å resolution of human immunodeficiency virus (HIV) II protease complexed with L-735,524, an orally bioavailable inhibitor of the HIV proteases. *J. Biol. Chem.*, 269:26344–26348, 1994.

[207] S. T. Nguyen, L. K. Johnson, R. H. Grubbs, and Z. J. W. Ring-opening metathesis polymerization (romp) of norbornene by a group viii carbene complex in protic media. *J. Am. Chem. Soc.*, 114:3974–3975, 1992.

[208] P. Schwab, M. B. France, J. W. Ziller, and R. H. Grubbs. A series of well-defined metathesis catalysts-synthesis of $[RuCl_2(CHR')(PR_3)_2]$ and its reactions. *Angew. Chem., Int. Ed.*, 34:2039, 1995.

[209] M. Scholl, S. Ding, C. W. Lee, and R. H. Grubbs. Synthesis and activity of a new generation of ruthenium-based olefin metathesis catalysts coordinated with 1,3-dimesityl-4,5-dihydroimidazol-2-ylidene ligands. *Org. Lett.*, 1:953–956, 1999.

[210] T. M. Trnka and R. H. Grubbs. The development of L2X2Ru=CHR olefin metathesis catalysts: An organometallic success story. *Acc. Chem. Res.*, 34:18–29, 2001.

[211] T. M. Trnka, J. P. Morgan, M. S. Sanford, T. E. Wilhelm, M. Scholl, T. L. Choi, S. Ding, M. W. Day, and R. H. Grubbs. Synthesis and activity of ruthenium alkylidene complexes coordinated with phosphine and n-heterocyclic carbene ligands. *J. Am. Chem. Soc.*, 125:2546–2558, 2003.

[212] G. C. Vougioukalakis and R. H. Grubbs. Ruthenium-based heterocyclic carbene-coordinated olefin metathesis catalysts. *Chem. Rev.*, 110:1746–1787, 2010.

[213] G. Occhipinti, B. H. R., and V. Jensen. Quantitative structure-activity relationships of ruthenium catalysts for olefin metathesis. *J. Am. Chem. Soc.*, 128:6952–6964, 2006.

[214] Y. Chu, W. Heyndrickx, G. Occhipinti, V. R. Jensen, and B. K. Alsberg. *De novo* optimization of functional coordination compounds using a fragment-based evolutionary algorithm. manuscript in preparation.

[215] C. W. Murray, D. E. Clark, T. R. Auton, M. A. Firth, J. Li, R. A. Sykes, B. Waszkowycz, D. R. Westhead, and S. C. Young. PRO_SELECT: Combining structure-based drug design and combinatorial chemistry for rapid lead discovery. 1. technology. *J. Comput. Aided Mol. Des.*, 11:193–207, 1997.

[216] N. Brown, B. Mckey, and G. J. Fingal: A novel approach to geometric fingerprinting and a comparative study of its application to 3D-QSAR modeling. *QSAR. Comb. Sci.*, 24:480–484, 2005.

[217] R. Todeschini and M. Lasagni. New molecular descriptors for 2D and 3D structures. Theory. *J. Chemom*, 8(4):263–272, 1994.

[218] V. Consonni, R. Todeschini, and M. Pavan. Structure/response correlations and similarity/diversity analysis by getaway descriptors. 1. Theory of the novel 3D molecular descriptors. *J. Chem. Inf. Comput. Sci.*, 42(3):682–692, 2002.

[219] J. Gasteiger, editor. *Handbook of Chemoinformatics: From Data to Knowledge (Representation of Molecular Structures).* Wiley-VCH, Weinheim, 2003.

[220] J. H. Schuur, P. Selzer, and J. Gasteiger. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.*, 36(2):334–344, 1996.

[221] E. Estrada and E. Uriarte. Recent advances on the role of topological indices in drug discovery research. *Current Medicinal Chemistry*, 8(13):1573–1588, 2001.

[222] S. Basak, A. T. Balaban, G. D. Grunwald, and B. D. Gute. Topological indices: Their nature and mutual relatedness. *J. Chem. Inf. Comput. Sci.*, 40(4):891–898, 2000.

[223] A. R. Katritzky and E. V. Gordeeva. Traditional topological indices vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. *J. Chem. Inf. Comput. Sci.*, 33(6):835–857, 1993.

[224] M. I. Skvortsova, I. I. Baskin, O. L. Slovokhotova, V. A. Palyulin, and N. S. Zefirov. Inverse problem in QSAR/QSPR studies for the case of topological indexes characterizing molecular shape (Kier indexes). *J. Chem. Inf. Comput. Sci.*, 33(4):630–634, 1993.

[225] M. Barker and W. Rayens. Partial least squares for discrimination. *Journal of Chemometrics*, 17(3):166–173, 2003.

[226] H. Martens and T. Naes. *Multivariate Calibration.* John Wiley & Sons, New York, 1989.

[227] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag, New York, 2007.

[228] B. R. Kowalski and C. F. Bender. K-nearest neighbor classification rule (pattern-recognition) applied to nuclear magnetic-resonance spectral interpretation. *Anal. Chem.*, 44(8):1405–1411, 1972.

[229] E. A. Patrick and F. P. Fischer. A generalized k-nearest neighbor rule. *Information and Control*, 16(2):128–152, 1970.

[230] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford Univ Press, USA, 1995.

[231] M. L. Bisani, S. Clementi, and S. Wold. Chemometrics. 2. the SIMCA method. *La Chim. e L'ind.*, 64(11):727–741, 1982.

[232] S. Wold. Pattern-recognition by means of disjoint principal components models. *Pattern Recognition*, 8(3):127–139, 1976.

[233] S. K. Murthy, S. Kasif, and S. Salzberg. A system for induction of oblique decision trees. *J. Art. Intell. Res.*, 2:1–32, 1994.

[234] B. K. Alsberg, R. Goodacre, J. J. Rowland, and D. B. Kell. Classification of pyrolysis mass spectra by fuzzy multivariate rule induction-comparison with regression, K-nearest neighbour, neural and decision-tree methods. *Analytica Chimica Acta*, 348(1-3):389–407, 1997.

[235] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[236] D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. D. De Jong, P. J. Lewi, and J. Smeyers-Verbeke. *Handbook of Chemometrics and Qualimetrics: Part A and B.* Elsevier Science, New York, 1997.

[237] F. Fontaine, M. Pastor, I. Zamora, and F. Sanz. Anchor-grind: Filling the gap between standard 3D QSAR and the grid-independent descriptors. *J. Med. Chem.*, 48(7):2687–2694, 2005.

[238] B. Buttingsrud, E. Ryeng, R. D. King, and B. K. Alsberg. Representation of molecular structure using quantum topology with inductive logic programming in structure-activity relationships. *J. Comput. Aided Mol. Des.*, 20:361–373, 2006.

[239] T. I. Netzeva, A. Worth, T. Aldenberg, R. Benigni, M. T. Cronin, P. Gramatica, J. S. Jaworska, S. Kahn, G. Klopman, C. A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G. Y. Patlewicz, R. Perkins, D. Roberts, T. Schultz, D. W. Stanton, J. J. van de Sandt, W. Tong, G. Veith, and C. Yang. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. *The report and recommendations of ECVAM Workshop 52. Altern. Lab. Anim.*, 33:155–173, 2005.

[240] B. Buttingsrud, R. D. King, and B. K. Alsberg. An alignment-free methodology for modelling field-based 3D-structure activity relationships using inductive logic programming. *J. Chemometrics*, 21(12):509–519, 2007.

[241] I. V. Tetko, I. Sushko, A. K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Öberg, R. Todeschini, D. Fourches, and A. Varnek. Critical assessment of QSAR models of environmental toxicity against tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.*, 48:1733–1746, 2008.

[242] C. J. Manly, S. Louise-May, and J. D. Hammer. The impact of informatics and computational chemistry on synthesis and screening. *Drug. Discov. Today*, 6:1101–1110, 2001.

[243] J. Tomasi. Thirty years of continuum solvation chemistry: a review, and prospects for the near future. *Theor. Chem. Acc.*, 112:184–203, 2004.

[244] W. Thiel and A. A. Voityuk. Extension of the MNDO formalism to d orbitals: Integral approximations and preliminary numerical results. *Theor. Chim. Acta*, 81:391–404, 1992.

[245] W. Thiel and A. A. Voityuk. Extension of the MNDO formalism to d orbitals: Integral approximations and preliminary numerical results. *Theor. Chim. Acta*, 93:315–315, 1996.

[246] W. Thiel and A. A. Voityuk. Extension of MNDO to d orbitals: Parameters and results for the second-row elements and for the zinc group. *J. Phys. Chem.*, 100:616–626, 1996.

[247] J. J. P. Stewart. Optimization of parameters for semiempirical methods i. method. *J. Comp. Chem.*, 10:209–220, 1989.

[248] J. J. P. Stewart. Optimization of parameters for semiempirical methods ii. applications. *J. Comp. Chem.*, 10:221–264, 1989.

[249] T. R. Cundari and J. Deng. PM3(tm) analysis of transition-metal complexes. *J. Chem. Inf. Comput. Sci.*, 39:376–381, 1999.

[250] K. J. Børve, V. R. Jensen, T. Karlsen, J. A. Støvneng, and O. Swang. Evaluation of PM3(tm) as a geometry generator in theoretical studies of transition-metal-based catalysts for polymerizing olefins. *J. Mol. Model.*, 3:193–202, 1997.

[251] J. J. P. Stewart. Optimization of parameters for semiempirical methods v: Modification of nddo approximations and application to 70 elements. *J. Mol. Model.*, 13:1173–1213, 2007.

[252] U. Kaldor and S. Wilson. *Theoretical chemistry and physics of heavy and superheavy elements*. Kluwer Academic Publishers, 2003.

[253] R. J. Deeth. The ligand field molecular mechanics model and the stereoelectronic effects of d and s electrons. *Coord. Chem. Rev.*, 212:11–34, 2001.

[254] A. C. T. van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard. Reaxff: A reactive force field for hydrocarbons. *J. Phys. Chem. A*, 105:9396–9409, 2001.

[255] J. J. P. Stewart. MOPAC2009. Stewart Computational Chemistry, Colorado Springs, CO, USA, http://openmopac.net (2008).

[256] von Grotthuss M., K. G., P. J., and R. L. Wyrwicz L. S. Ligand.info small-molecule meta-database. *Comb. Chem. High Throughput Screen.*, 7:757–61, 2005.

[257] C. M. Cragg, D. J. Newman, and K. M. Snader. For an outstanding analysis of the role of natural products in pharmaceuticals. *J. Nat. Prod.*, 60:52–60, 1997.

[258] R. Breinbauer, I. R. Vetter, and H. Waldmann. From protein domains to drug candidates-natural products as guiding principles in the design and synthesis of compound libraries. *Angew. Chem. Int. Ed.*, 41:2878–2890, 2002.

[259] T. Vorfalt, S. Leuthaußer, and H. Plenio. An [(nhc)(nhc$_{EWG}$)rucl$_2$(chph)] complex for the efficient formation of sterically hindered olefins by ring-closing metathesis. *Angew. Chem., Int. Ed.*, 121:5293–5296, 2009.

[260] T. Ritter, A. Hejl, A. G. Wenzel, T. W. Funk, and R. H. Grubbs. A standard system of characterization for olefin metathesis catalysts. *Organometallics*, 25:5740–5745, 2006.

[261] H. Clavier, A. Correa, E. C. Escudero-Adan, J. Benet-Buchholz, L. Cavallo, and S. P. Nolan. Chemodivergent metathesis of dienynes catalyzed by ruthenium-indenylidene complexes: An experimental and computational study. *Chem. Eur. J.*, 15:10244–10254, 2009.

[262] C. Adlhart and P. Chen. Mechanism and activity of ruthenium olefin metathesis catalysts: The role of ligands and substrates from a theoretical perspective. *J. Am. Chem. Soc.*, 126:3496–3510, 2004.

[263] G. Occhipinti, V. R. Jensen, K. W. Törnroos, N. A. Frøystein, and H. R. Bjørsvik. Synthesis of a new bidentate nhc-ag(i) complex and its unanticipated reaction with the hoveyda-grubbs first generation catalyst. *Tetrahedron*, 65:7186–7194, 2009.

[264] S. Tiede, A. Berger, D. Schlesiger, D. Rost, A. Luehl, and S. Blechert. Highly active chiral ruthenium-based metathesis catalysts through a monosubstitution in the n-heterocyclic carbene. *Angew. Chem., Int. Ed.*, 49:3972–3975, 2010.

[265] R. M. Thomas and R. H. Grubbs. Synthesis of telechelic polyisoprene via ring-opening metathesis polymerization in the presence of chain transfer agent. *Macromolecules*, 43:3705–3709, 2010.

[266] B. Stenne, J. Timperio, J. Savoie, T. Dudding, and S. K. Collins. Desymmetrizations forming tetrasubstituted olefins using enantioselective olefin metathesis. *Org. Lett.*, 12:2032–2035, 2010.

[267] K. M. Kuhn, T. M. Champagne, S. H. Hong, W.-H. Wei, A. Nickel, C. W. Lee, S. C. Virgil, R. H. Grubbs, and R. L. Pederson. Low catalyst loadings in olefin metathesis: Synthesis of nitrogen heterocycles by ring-closing metathesis. *Org. Lett.*, 12:984–987, 2010.

[268] R. H. Grubbs, C. K. Chung, J.-B. Bourg, and K. Kuhn. PCT Int. Appl., 2009.

[269] J. Savoie, B. Stenne, and S. K. Collins. Improved chiral olefin metathesis catalysts: Increasing the thermal and solution stability via modification of a c_1-symmetrical n-heterocyclic carbene ligand. *Adv. Syn. Cat.*, 351:1826–1832, 2009.

[270] K. M. Kuhn, J.-B. Bourg, C. K. Chung, S. C. Virgil, and R. H. Grubbs. Effects of nhc-backbone substitution on efficiency in ruthenium-based olefin metathesis. *J. Am. Chem. Soc.*, 131:5313–5320, 2009.

[271] A. Grandbois and S. K. Collins. Enantioselective synthesis of [7]helicene: Dramatic effects of olefin additives and aromatic solvents in asymmetric olefin metathesis. *Chem. Eur. J.*, 14:9323–9329, 2008.

[272] P.-A. Fournier, J. Savoie, B. Stenne, M. Bédard, A. Grandbois, and S. K. Collins. Mechanistically inspired catalysts for enantioselective desymmetrizations by olefin metathesis. *Chem. Eur. J.*, 14:8690–8695, 2008.

[273] C. K. Chung and R. H. Grubbs. Olefin metathesis catalyst: Stabilization effect of backbone substitutions of n-heterocyclic carbene. *Org. Lett.*, 10:2693–2696, 2008.

[274] Y. Schrodi, R. L. Pederson, H. Kaido, and M. J. Tupy. PCT Int. Appl., 2008.

[275] A. V. Nizovtsev, E. V. Shutko, V. V. Afanasev, T. M. Dolgina, and N. B. Bespalova. PCT Int. Appl., 2007.

[276] P.-A. Fournier and S. K. Collins. A highly active chiral ruthenium-based catalyst for enantioselective olefin metathesis. *Organometallics*, 26:2945–2949, 2007.

[277] A. D. Becke. Density-functional thermochemistry. iii. the role of exact exchange. *J. Chem. Phys.*, 98:5648–5652, 1993.

[278] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.*, 98:11623–11627, 1994.

[279] B. M. Wise, N. B. Gallagher, R. Bro, J. M. Shaver, W. Windig, and R. S. Koch. *Matlab PLS toolbox 3.5 manual*. Eigenvector Research, Manson, USA, 2004.

[280] B. F. Straub. Origin of the high activity of second-generation grubbs catalysts. *Angew. Chem., Int. Ed.*, 44:5974–5978, 2005.

[281] C. G. Broyden. The convergence of a class of double-rank minimization algorithms 2, the new algorithm. *J. of the Inst. for Math. and Applications*, 6:222–231, 1970.

[282] R. Fletcher. A new approach to variable-metric algorithms. *Computer Journal*, 13:317–322, 1970.

[283] A. Goldfarb. A family of variable-metric algorithms derived by variational means. *Mathematics of Computation*, 24:23–26, 1970.

[284] D. F. Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation*, 24:647–656, 1970.

[285] Marvin 5.3.4, ChemAxon Ltd., http://www.chemaxon.com.

[286] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, N. Scalmani, G.and Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, , and J. A. Pople. Gaussian03. Revision C.02, Gaussian, Inc., Wallingford CT, 2004.

[287] N. C. Handy and A. J. Cohen. Left-right correlation energy. *Mol. Phys.*, 99:403–412, 2001.

[288] C. Lee, W. Yang, and R. G. Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron-density. *Phys. Rev. B*, 37:785–789, 1988.

[289] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic-behavior. *Phys. Rev. A*, 38:3098–3100, 1988.

[290] P. J. Hay and W. R. Wadt. Ab initio effective core potentials for molecular calculations. potentials for main group elements Na to Bi. *J. Chem. Phys.*, 82:284–298, 1985.

[291] P. J. Hay and W. R. Wadt. Ab initio effective core potentials for molecular calculations. potentials for K to Au including the outermost core orbitals. *J. Chem. Phys.*, 82:299–310, 1985.

[292] T. H. J. Dunning and P. J. Hay. In I. Schaefer, H. F., editor, *Methods of Electronic Structure Theory*, pages 1–28. Plenum, New York, 1977.

Appendix A

Supporting Information for *De novo* Optimization of Functional Coordination Compounds

Yunhan Chu[†], Wouter Heyndrickx[‡], Giovanni Occhipinti[‡], Vidar R. Jensen[‡], and Bjørn K. Alsberg[†]

[†]Department of Chemistry, Norwegian University of Science and Technology, N-7491 Trondheim, [‡]Department of Chemistry, University of Bergen, Allégaten 41, N-5007 Bergen, Norway

Is not included due to copyright