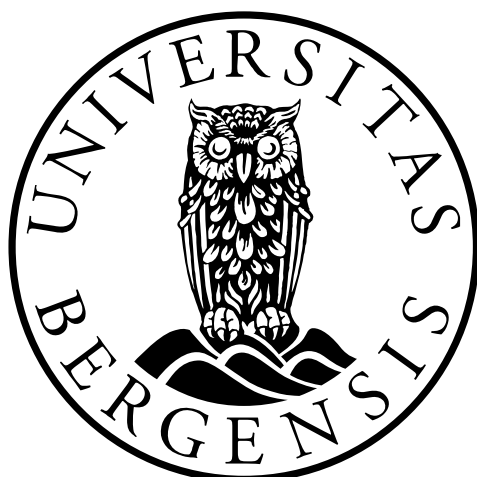


A method for automated *de novo* design of functional transition-metal compounds

Marco Foscatto



Dissertation for the degree of philosophiae doctor (PhD)
at the University of Bergen

2015

© Copyright Marco Foscatto

The material in this publication is protected by copyright law.

Year: 2015

Title: “A method for automated *de novo* design of functional transition-metal compounds”

Author: Marco Foscatto

Print: AIT OSLO AS / University of Bergen

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisor, Prof. Dr. Vidar R. Jensen, for giving me the opportunity to join this project and for providing continuous support, wise advices, and inspiring thoughts that, all in all, have made this thesis possible. I want to thank the two co-supervisors Prof. Dr. Bjørn K. Alsberg and Dr. Giovanni Occhipinti, and also Dr. Vishwesh Venkatraman for more than three years of fruitful brainstorming sessions on strategies, methods, applications and code development. I have learnt a lot from all the work done in the joint Bergen-Trondheim team, namely Vidar and Giovanni in Bergen and Bjørn, Vish, and lately also Sailesh Abburu in Trondheim, and I thank you all for this. Prof. Dr. Robert J. Deeth is thanked for the hospitality on my visit to his group in Warwick (UK) and the following collaboration on LFMM. Such project has lately involved also Benjamin J. Houghton who, as well, is tanked for discussions. Back at the University of Bergen, I'm grateful to the nanomodeling and theoretical chemistry group, in particular Prof. Dr. Knut Børve, Prof. Dr. Leif J. Sæthre, and my colleagues Sondre H. Hopen Eliasson and Vitali Koudriavtsev, for discussing projects and sharing experiences, and also the students that had to tolerate me, Mauricio Ayala Ortega and Kjell Nedreliid. Dr. Veronika Šoltészová is thanked for the development of a very useful visualization tool.

The Research Council of Norway (RCN) is acknowledges for financial support via the eVITA (grant number 205273/V30) program, the NOTUR (NN2506K) and NORSTORE (NS2506K) supercomputing programs for CPU and storage resources, and the University of Bergen for computational resources.

A special thanks goes to my family; my parents, Laura and Marcello, and my brother Andrea for all the support over the years. Finally, a huge thanks goes to my soon to be wife, Giulia, for supporting me with love and cakes, and for understanding and being patient for such a long time.

Abstract

Systematic application of automated design methods in transition metal and organometallic chemistry is hampered by the limitations of computational tools traditionally developed according to organic chemistry formalisms. Such tools are often inadequate to deal with peculiar chemical entities such as organometallic catalysts. Therefore, the present study aims to develop methods for generating and handling special chemical entities in the context of automated molecular modeling and *de novo* design. We introduced a set of formalized and versatile rules to guide automated mining of molecular fragments to be used for construction of new and synthetically realistic chemical entities with punctual control over metal coordination environments. The preparation of initial three-dimensional (3D) molecular models was made independent from tedious force field parametrization by exploiting geometrical information taken from crystallographic structures or computed models and stored in 3D fragments. The method demonstrated superior performance in the preparation of 3D models for tree-like structures of peculiar compounds and also capability of handling chemical entities that are beyond the scope of regular tools. Design of multicyclic system from acyclic 3D building blocks was achieved by including methods for the identification of closable chains of fragments and ring-closing conformational adaptation. The overall machinery, which was integrated into a previously developed evolutionary algorithm for *de novo* design, was coupled with the computationally inexpensive ligand field molecular mechanics (LFMM) method, which was implemented in Tinker as part of this work, and applied in the design of new Fe(II) spin crossover compounds with multidentate amine ligands. New potential spin crossover compounds with unexpected, yet realistic ligands were identified and proposed for further investigation. In conclusion, while refinement of the implementation is proposed to improve efficiency, the overall *de novo* molecular design method developed in this study contributes to empower the application of automated design strategies in transition metal and organometallic chemistry.

List of publications

Paper I

Automated Design of Realistic Organometallic Molecules from Fragments.
Foscato, M.; Occhipinti, G.; Venkatraman, V.; Alsberg, B. K.; Jensen, V. R.
J. Chem. Inf. Model. **2014**, *54*, 767–780.

Paper II

Automated Building of Organometallic Complexes from 3D Fragments.
Foscato, M.; Venkatraman, V.; Occhipinti, G.; Alsberg, B. K.; Jensen, V. R.
J. Chem. Inf. Model. **2014**, *54*, 1919–1931.

Paper III

Integration of Ligand Field Molecular Mechanics in Tinker.
Foscato, M.; Deeth, R. J.; Jensen, V. R.
J. Chem. Inf. Model. **2015**, Just Accepted, DOI:10.1021/acs.jcim.5b00098

Paper IV

I'll Give You a Ring: Ring Closure to Form Metal Chelates in 3D Fragment-Based *de Novo* Design.
Foscato, M.; Houghton, B. J.; Occhipinti, G.; Deeth, R. J.; Jensen, V. R.
To be submitted

Additional paper not included in this thesis:

Evolutionary *de Novo* Design of Phenothiazine Derivatives for Dye-Sensitized Solar Cells.
Venkatraman, V.; Foscato, M.; Jensen, V. R.; Alsberg, B. K.
J. Mater. Chem. A **2015**, *3* (18), 9851–9860.

The published papers are reprinted with permission from American Chemical Society. All rights reserved.

Contents

ACKNOWLEDGEMENTS	3
ABSTRACT	4
LIST OF PUBLICATIONS	5
CONTENTS	6
1. INTRODUCTION	8
1.1 AIM OF THIS WORK	11
1.2 WHAT THIS WORK IS NOT ABOUT	12
1.3 OUTLINE	12
2. COMPUTATIONAL METHODS	13
2.1 EVOLUTIONARY ALGORITHM	13
2.2 COMPUTATIONAL CHEMISTRY METHODS	14
2.2.1 <i>Quantum Mechanics</i>	14
2.2.2 <i>Molecular Mechanics</i>	26
3. GENERATING CHEMICAL ENTITIES FROM SCRATCH	32
3.1 INTRODUCTION.....	32
3.1.1 <i>Building Strategies</i>	32
3.1.2 <i>Chemical Representations</i>	35
3.2 THE GENERAL PURPOSE FRAGMENT-BASED DESIGN MACHINERY	38
3.2.1 <i>Graph Representation and Fragments</i>	38
3.2.2 <i>Generation of Fragments and Cutting Rules</i>	40
3.2.3 <i>Generation of Graphs</i>	42
3.2.4 <i>Connection rules and Compatibility Matrix</i>	44
3.2.5 <i>From Graph to 3D Molecular Model</i>	45
3.2.6 <i>Ring-Closing Potential</i>	47
4. DESIGN OF REALISTIC ORGANOMETALLIC COMPOUNDS	50
4.1 RESULTS AND DISCUSSION	50
4.2 CONCLUSION	54
5. BUILDING OF ACCURATE 3D MOLECULAR MODELS	55
5.1 INTRODUCTION.....	55
5.2 RESULTS AND DISCUSSION	56
5.2.1 <i>Case Study 1</i>	56

5.2.2	Case Study 2.....	64
5.2.3	Case Study 3.....	67
5.3	CONCLUSION.....	68
6.	FAST FITNESS FROM LIGAND FIELD MOLECULAR MECHANICS.....	69
6.1	INTRODUCTION.....	69
6.2	LIGAND FIELD MOLECULAR MECHANICS (LFMM).....	70
6.3	INTEGRATION OF LFMM IN TINKER.....	71
7.	MULTIDENTATE LIGANDS FOR FE(II) SPIN-CROSSOVER COMPOUNDS: A TEST CASE FOR RING-CLOSING DESIGN.....	73
7.1	INTRODUCTION.....	73
7.2	BRIEF COMPUTATIONAL DETAILS.....	74
7.3	RESULTS AND DISCUSSION.....	75
7.4	CONCLUSION.....	81
8.	CONCLUDING REMARKS.....	82
9.	DIRECTIONS FOR FUTURE WORK.....	84
10.	REFERENCES.....	85

1. Introduction

The growing quest for chemicals with specific properties is what molecular design is all about, and is also the underlying reason for developing automated molecular design methods: the central topic of this thesis.

From the introduction of Dalton's atomic model, the desire to understand and exploit the relations between chemical structure and functional properties has been the cornerstone of molecular design.¹ If we assume the existence of such structure-property relationship, then, in order to obtain a property of interest, we need to define a chemical system in terms of composition, structure, and stereochemistry, that intrinsically exhibits that property under certain conditions.² Unfortunately, this is not as simple as it sounds. The underlying problems are that the exact structure-property relationship is usually not known a priori, and that the target compounds represent an infinitesimal fraction of the comprehensive ensemble of all possible compounds, which is often referred as to the *chemical space*,^{3,4} thus exhaustive screening is excluded.

Building on the concept of structure-property relationship, there are two ways to tackle the design of molecules displaying a desired properties: the *forward* or *direct* strategy (i.e., from structure to property), and the *inverse* strategy (i.e., from property to structure).⁵ The first is an empirical approach, in the sense that it is based on gaining experience. Direct design involves evaluation of selected candidates, either *in vitro* or *in silico*, data analysis, and informed decisions that propose new candidates and start new iterations until satisfactory results are achieved. On the contrary, inverse design aims to deduce the identity of the target compounds from the very nature of the tailored property by inverting the mathematical formulation of structure-to-property relationships.⁵ Although inverse problems are ubiquitous in science and engineering, and the pioneering applications to molecular design have shown promising results, they are also typically challenging to solve (ill-conditioned and with multiple and unrealistic solutions).^{6,7} Therefore, the direct strategy has so far

dominated the scene of molecular design, with an undisputed role in the field of drug design. The direct design strategy is assumed also throughout the present dissertation.

With the increase of computational power and the development of efficient modeling tools, the contribution of computational chemistry has become widespread in molecular design. Unfortunately, calculated values simply cannot replace experimental data, and it is clear that neither experimental nor computational methods can possibly afford handling of the overwhelming vastness of the chemical space.⁸ Nevertheless, computational prediction of molecular properties can provide useful insights to refine the understanding of the structure-property relationship² and indicate which experiments should be performed *in vitro*, thus improving the efficiency of the design work.⁹

Computational strategies deployed in direct design include virtual screening and *de novo* design.^{10,11} While the first relies on the existence of a library of compounds to be screened, which might also be virtually synthesized on purpose,^{12,13} the *de novo* approach includes generation of candidates from scratch during the course of the computational design experiment. In fact, most of the compounds evaluated in a *de novo* design experiments are not known when the experiment begins. By building candidates from scratch, the *de novo* approach has the capability of suggesting innovative chemical features and following the trend of well performing hypothesis in an automated fashion. *De novo* algorithms are based on an iterative adaptive process where each iteration uses memory from previous experience, i.e., previously evaluated molecules, to take informed decisions aimed at modifying the molecular structures of visited candidates and creating new ones, which are then evaluated starting a new iteration.¹ In general, the three pillars characterizing this approach are (i) the generation of new candidates (problem of construction), (ii) the evaluation of each candidate (problem of scoring or fitness), and (iii) the navigation of the search space towards the optimal solution (problem of optimization).^{11,14} The combination of these three components leads to methods that explore the fitness landscape and, by automatically collecting information on the fly, attempt to converge towards the objective defined by a given fitness function.

Since the first applications reported in the literature, which date back to 1989,^{15,16} the development of automated *de novo* molecular design methods has mainly occurred within the context of organic drug design.¹⁷ Applications in other fields have been reported mainly for the design of novel enzymes,^{18,19} peptides,²⁰ nucleic acid sequences,^{21,22} and organic templates for microporous materials.^{23–25} More recently, automated *de novo* design has been applied also for the design of metal-organic frameworks,²⁶ porous organic polymers,²⁷ and functional organic compounds for light emitting diodes,²⁸ and solar cells.²⁹

These achievements were made possible by the legacy resulting from decades of development in the field of organic drug design, but the general application of *de novo* methods in transition metal and organometallic chemistry turned out challenging.³⁰ First, the design of such compounds is often characterized by a different underlying philosophy. While the central issue in drug design is the complementarity of the candidate ligand with a given biological target, both in terms of shape and intermolecular interactions, reactivity and electronic properties have dominant roles in the design of functional transition metal compounds, in particular, in the field of catalysts. Therefore, while a number of reliable empirical methods can model molecular shape and intermolecular interaction for organic systems with sufficient accuracy and at a very low computational cost, the partially filled d-orbitals that characterize transition metal centers often require specific empirical solutions, such as dedicated molecular mechanics methods,³¹ or quantum mechanical models to model both molecular and electronic structure with sufficient accuracy. Moreover, the rich chemistry of transition metal compounds introduces peculiar bond types as well as molecular and electronic structures that are often beyond the capabilities of tools and formalisms traditionally made for organic, drug-like and natural product-like molecules.^{32,33} In addition, dealing with functional transition metal compounds can

imply handling of peculiar chemical entities,ⁱ such as reaction intermediates, that are not properly handled by standard tools.

Overall, the intrinsic nature of transition metal species determines a methodological gap that challenges the application of existing automated *de novo* design methods in transition metal and organometallic chemistry.

1.1 Aim of this work

The present work aims to develop methods for *de novo* design of functional transition metal and organometallic species. In particular, this thesis focuses on improving the automated generation and handling of peculiar candidates thus overcoming the limitations that hamper broad-spectrum application of automated design tools. The following goals are set:

- Provide the capability of handling peculiar chemical entities in an automated fashion. Transition metal compounds dominate the field of catalysis where active species, intermediates, or even transition state models have to be generated and properly managed.
- Control the search space by projecting the chemists' knowledge, intuition, and intent into the automated machinery. *De novo* molecular factories are known to generate molecule that suffer from synthetic accessibility issues and easily tend to produce unrealistic candidates. Therefore, means to exert chemical control on the automated machinery are required and shall support both organic and organometallic chemistry in addition to other, possibly non-conventional, molecular assembling rules.
- Generate accurate and educated initial guesses for three-dimensional molecular models. The peculiar geometrical features of transition metal compounds are

ⁱ Chemical entities are physical entities of interest in chemistry including molecular entities, parts thereof, and chemical substances.³⁴

often pivotal in determining the molecular properties and cannot be overlooked without losing significant information (i.e., stereochemistry) and compromising the design process.

- Integrate fast fitness. Promptly available, and possibly inexpensive, methods for modeling and evaluation of the chemical properties reduce the computational price of *de novo* design.

1.2 What This Work is Not About

One of the three pillars of *de novo* design methods is the optimization algorithm. Although there exist alternatives that should be tested and compared, optimization algorithms are not explored as part of this work. Nevertheless, the methods presented here have been integrated in an evolutionary algorithm that has been described in ref. 30. The algorithm is introduced in Section 2.1, and has been applied in the design of iron complexes as discussed in Chapter 7.

1.3 Outline

After a preliminary description of the computational tools deployed in this thesis (Chapter 2), the method for the generation and modification of candidates is presented highlighting the innovations introduced in this work (Chapter 3). Focused discussion on the control strategy for the automated generation of candidates is given in Chapter 4. Next, the issue of the accurate preparation of three-dimensional models is addressed (Chapter 5). Finally, after the introduction of the fast molecular modeling tool based on the ligand field molecular mechanics (Chapter 6), the *de novo* design of Fe(II) spin crossover compounds is presented as an application of the combined *de novo* design machinery and particular emphasis is given to the management of rings (Chapter 7).

2. Computational Methods

2.1 Evolutionary Algorithm

Since the fitness landscape is characterized by a multitude of local minima and maxima, global optimization techniques are required to identify the fittest compounds.¹ Evolutionary algorithms are among the most diffused global optimization method for molecular *de novo* design.^{35,36} These algorithms are metaheuristic optimization techniques: problem-independent and approximate (i.e., not exact nor deterministic) algorithms deployed to solve general classes of problems for which specific and deterministic solving algorithms are not available.³⁷ The lack of a general and exact solution, which would be an inverse property-to-structure law, characterized molecular design problems, hence the usefulness of metaheuristic algorithms that do not use gradient nor Hessian matrix of the objective function. Moreover, these algorithms are problem-independent; the same algorithm can be applied to many different molecular design problems. To this end, an objective, or fitness, function is defined as the mean to translate properties of the candidate under evaluation into a numerical score. It should be noted that many molecular design problems have multi-objective nature, meaning that the overall value of a candidate depends on conflicting properties, such as efficacy, selectivity, synthetic accessibility, toxicity, solubility, and price, thus a multidimensional approach may be required.³⁸⁻⁴²

Evolutionary algorithms are population-based optimization techniques that involve selection of high fitness candidates and genetic operators to alter (mutation operator) and exchange (crossover operator) the constitutional features among the population members. Different implementations of the general philosophy have lead to four main groups of algorithms: genetic algorithm, genetic programming, evolutionary programming and evolutionary strategies (for a review see ref. 35). Genetic algorithms are exploited in the present work (Chapter 7).

A looping program that operates on an instantaneous population of molecules characterizes the workflow of genetic algorithms. Each iteration, which is often

referred as to a *generation*, involves (i) fitness-biased selection of high fitness individuals for producing offspring, (ii) generation of new candidates by means of mutation and crossover operators, (iii) evaluation of the new candidates, (iv) update of the population and beginning of a new iteration. This evolution loop is eventually terminated according to user-defined termination criteria. The evolutionary algorithms deployed in this work obeys this general workflow and has been described in detail in ref. 30.

2.2 Computational Chemistry Methods

The following sections introduce the computational methods used to model chemical entities and are mostly based on common textbooks on quantum chemistry and molecular modeling.^{43–47,31}

2.2.1 Quantum Mechanics

Basic Principles

The wave function $\Psi(\mathbf{x}_1, \dots, \mathbf{x}_n, t)$ characterizes the quantum mechanical description of the state of a system of n particles as a function of the combined spatial and spin coordinates (\mathbf{x}_i) of each particle i and the time (t). Although there is no clear interpretation of the wave function itself, its squared modulus $|\Psi(\mathbf{x}_1, \dots, \mathbf{x}_n, t)|^2$ is interpreted as a probability density and $|\Psi(\mathbf{x}_1, \dots, \mathbf{x}_n, t)|^2 d\mathbf{x}_1 \dots d\mathbf{x}_n$ as the probability of finding simultaneously each particle i in the corresponding infinitesimal of space $d\mathbf{x}_i$ with given spin (i.e., Born's statistical interpretation). Thus the integral over the whole space correspond to the unit, that is, $\Psi(\mathbf{x}_1, \dots, \mathbf{x}_n, t)$ is normalized. A crucial property of the wave function is that it implicitly contains all the information that can possibly be known on the system.

Such precious function is obtained as solution of the Schrödinger equation, which, for a single particle with mass m can be written in the non-relativistic form as

$$i\hbar \frac{\partial \Psi(\mathbf{x}, t)}{\partial t} = -\frac{\hbar^2}{2m} \nabla^2 \Psi(\mathbf{x}, t) + V(\mathbf{x}, t) \Psi(\mathbf{x}, t) , \quad (1)$$

where i is the imaginary unit, \hbar is Planck's constant divided by 2π , $\Psi(\mathbf{x}, t)$ the single particle wave function, $V(\mathbf{x}, t)$ is an external potential (i.e., the electrostatic potential due to the nuclei in a molecule), and ∇^2 is the Laplacian operator that is

$$\nabla^2 = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) . \quad (2)$$

Equation 1 is usually written using the Hamiltonian operator \hat{H} , which includes both the kinetic energy operator \hat{T} and the potential energy operator \hat{V} , and thus represents the total energy of the system:

$$\hat{H} = \hat{T} + \hat{V} , \quad (3)$$

$$\text{where } \hat{T} = -\frac{\hbar^2}{2m} \nabla^2 \text{ and } \hat{V} = V(\mathbf{x}, t) . \quad (4)$$

This allows the time-dependent Schrödinger equation to be written as

$$i\hbar \frac{\partial \Psi(\mathbf{x}, t)}{\partial t} = \hat{H} \Psi(\mathbf{x}, t) . \quad (5)$$

Notably, if the potential does not depend of the time t , that is $V(\mathbf{x}, t) \equiv V(\mathbf{x})$, the wave function can take the form of the product between two components: $\psi(\mathbf{x})$ only dependent of the spatial and spin coordinates, and $\tau(t)$ only dependent on time:

$$\Psi(\mathbf{x}, t) = \psi(\mathbf{x}) \tau(t) . \quad (6)$$

As a consequence of this separation of variables, the time-independent, non-relativistic Schrödinger equation is written as

$$\hat{H} \psi(\mathbf{x}) = E \psi(\mathbf{x}) , \quad (7)$$

which is an eigenvalue equation where the eigenvalue E is the total energy of the system for the state described by the stationary (i.e., $|\Psi(\mathbf{x},t)|^2 \equiv |\psi(\mathbf{x})|^2$) state $\Psi(\mathbf{x},t)$.

For a typical quantum chemical problem, where n electrons and m nuclei are to be described, the corresponding time-independent, non-relativistic Schrödinger equation, in absence of an external potentials, rises from the contributions of kinetic energy of nuclei, \hat{T}_N , that of the electrons, \hat{T}_E , the electrostatic interaction between nuclei and electrons, \hat{V}_{NE} , and the repulsions between nuclei, \hat{V}_{NN} , and between electrons, \hat{V}_{EE} .

$$\hat{H} = \hat{T}_N + \hat{T}_E + \hat{V}_{NE} + \hat{V}_{NN} + \hat{V}_{EE} \text{ , or} \quad (8)$$

$$\hat{H} = -\frac{1}{2} \sum_{A=1}^m \frac{\nabla_A^2}{M_A} - \frac{1}{2} \sum_{i=1}^n \nabla_i^2 + \sum_{A=1}^m \sum_{i=1}^n \frac{Z_A}{r_{iA}} + \sum_{A=1}^m \sum_{B>A}^m \frac{Z_A Z_B}{r_{AB}} + \sum_{i=1}^n \sum_{j>i}^n \frac{1}{r_{ij}} \text{ ,} \quad (9)$$

where M_A is the mass of nucleus A and Z_A its atomic number, r_{pq} stands for the distance between the pair p and q , and all quantities are reported in atomic units to simplify the notation.

Given that nuclei are far heavier than electrons, their motion usually takes place on a different time scale with respect to that of electrons, which move much faster. Thus the Born-Oppenheimer approximation decouples the motion of electrons and nuclei considering the motion of electrons in a system where the nuclei are static objects that have no kinetic energy ($\hat{T}_N = 0$) and experience a constant nuclear repulsion. This approximation simplifies the definition of the Hamiltonian into

$$\hat{H}_{elec} = -\frac{1}{2} \sum_{i=1}^n \nabla_i^2 + \sum_{A=1}^m \sum_{i=1}^n \frac{Z_A}{r_{iA}} + \sum_{i=1}^m \sum_{j>i}^m \frac{1}{r_{ij}} \text{ ,} \quad (10)$$

where the electronic Hamiltonian \hat{H}_{elec} considers the nuclear coordinates, $\mathbf{r}_1, \dots, \mathbf{r}_m$, as parameters rather than variables, and the same applies to the wave function, which for clarity can be written as

$$\psi_{elec} \equiv \psi(\mathbf{x}_1, \dots, \mathbf{x}_n, \{\mathbf{r}_1, \dots, \mathbf{r}_m\}) \text{ .} \quad (11)$$

This allows calculation of the electronic energy E_{elec} as

$$\hat{H}_{elec} \psi_{elec} = E_{elec} \psi_{elec} . \quad (12)$$

from which the total energy is obtained by addition of the inter-nuclear repulsion:

$$E = E_{elec} + \sum_{A=1}^m \sum_{B>A}^m \frac{Z_A Z_B}{r_{AB}} . \quad (13)$$

Unfortunately, for molecular systems larger than H_2^+ , an exact solution of the Schrödinger equation cannot be found with current techniques. Therefore, for systems of practical interest, only approximated solution are accessible. Fortunately it can be demonstrated that the expectation value of \hat{H}_{elec} calculated from any ψ_{elec}^{try} , that is the electronic energy E_{elec}^{try} , is always larger than E_{elec}^0 , which is the expectation value of the ground state ψ_{elec}^0 , for all ψ_{elec}^{try} but the actual ground state ψ_{elec}^0 . This is known as the variational principle. Finding ψ_{elec}^0 is therefore an optimization problem for which optimization techniques are applied to search the space of the acceptable solutions. Unfortunately, due to the immensity of such search space, systematic testing of all the solutions remains unachievable and further approximations are needed.

Hartree-Fock Approximation

One of the approximations introduced to facilitate the identification of ψ_{elec} is that of representing the n -electrons wave function as a Slater determinant Φ_{SD} :

$$\psi_{elec} \approx \Phi_{SD} = \frac{1}{\sqrt{n!}} \begin{vmatrix} \chi_1(\mathbf{x}_1) & \cdots & \chi_n(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \chi_1(\mathbf{x}_n) & \cdots & \chi_n(\mathbf{x}_n) \end{vmatrix} , \quad (14)$$

$$\text{where } \chi(\mathbf{x}) = \phi(\mathbf{r})\sigma(s) . \quad (15)$$

The spin orbitals $\chi(\mathbf{x})$ are orthonormal one-electron wave functions resulting from the product of a spatial orbital $\phi(\mathbf{r})$ and a spin component ($\sigma(s)$ is $\alpha(s)$ or $\beta(s)$). The determinant posses the antisymmetric property required for fermions (i.e., the wave

function change sign upon exchange of the spin orbital of two electrons), and includes all the $n!$ permutations of electrons, thus accounting for the fact that electrons are indistinguishable.

Each spin orbital is then constructed as the linear combination of a set of l basis functions $\{\eta\}_l$,

$$\chi_i = \sum_{\mu=1}^l c_{\mu i} \eta_{\mu} , \quad (16)$$

where $c_{\mu i}$ are coefficients and each η_{μ} is, in most of the cases, either a single or a combination of Slater type orbitals or Gaussian type orbitals. The firsts have the proper shape but require more computations than the seconds when it comes to calculate two electron integrals, and the seconds do not have the proper shape and thus more Gaussian functions are combined to form one basis function (i.e., contraction).

Due to the variational principle (see above), the lowest energy, i.e., the expectation values of \hat{H}_{elec} , for Φ_{SD} can be obtained optimizing the expansion coefficients ($c_{\mu i}$ in equation 16) of the linear combination of basis set functions defining each spin orbital χ_i . In order to identify the best spin orbitals that still respect the orthonormal requirements, each χ_i is obtained as solution of one-electron Hartree-Fock equations

$$\hat{f}_i \chi_i = \varepsilon_i \chi_i , \quad (17)$$

$$\text{where } \hat{f}_i = -\frac{1}{2} \nabla_i^2 - \sum_A^m \frac{Z_A}{r_{iA}} + V_{HF}(i) . \quad (18)$$

The three terms in the Fock operator \hat{f}_i take into account the kinetic energy of the electron, its potential energy due to the attraction with each of the nuclei, and the so called Hartree-Fock potential $V_{HF}(i)$ that corresponds to the repulsive potential of an electron due to the remaining $n-1$ electrons. More precisely

$$V_{HF}(\mathbf{x}_1) = \sum_j^n \left(\hat{J}_j(\mathbf{x}_1) - \hat{K}_j(\mathbf{x}_1) \right), \quad (19)$$

$$\text{where } \hat{J}_j(\mathbf{x}_1) = \int |\chi_j(\mathbf{x}_2)|^2 \frac{1}{r_{12}} d\mathbf{x}_2 \quad (20)$$

is the Coulomb operator that represents the Coulomb repulsion between an electron in \mathbf{x}_1 and the average charge of an electron in the spin orbital χ_j , and

$$\hat{K}_j(\mathbf{x}_1)\chi_i(\mathbf{x}_1) = \int \chi_j^*(\mathbf{x}_2) \frac{1}{r_{12}} \chi_i(\mathbf{x}_2) d\mathbf{x}_2 \chi_j(\mathbf{x}_1) \quad (21)$$

is the exchange operator that results from the antisymmetry of the Slater determinant. There is no classical counterpart for this operator that considers only the contribution from electrons of like spin.

Notably the action of both \hat{J} and \hat{K} depend on the spin orbitals χ that are the solution of the Hartree-Fock equation. Therefore, the solution of the problem, i.e., Φ_{SD} , is found by an iterative process where a first guess for the set of spin orbitals is provided as input and iteratively updated by solving the n Hartree-Fock equations until convergence, i.e., the input set of χ is sufficiently close to the solutions of the Hartree-Fock equations: scenario that is referred as to the self consistent field (SCF). Since spin orbitals are defined by linear combinations (equation 16), each of the Hartree-Fock equations can be rewritten as follows (expanding χ_i , multiplying for an arbitrary η_v , and integrating over space):

$$\sum_{\mu=1}^l c_{\mu i} \int \eta_v^*(\mathbf{r}) \hat{f}_i(\mathbf{r}) \eta_{\mu}(\mathbf{r}) d\mathbf{r} = \varepsilon_i \sum_{\mu=1}^l c_{\mu i} \int \eta_v^*(\mathbf{r}) \eta_{\mu}(\mathbf{r}) d\mathbf{r} \quad \text{for } 1 \leq v \leq l \quad (22)$$

The resulting system of l equations is consistently written in a matrix form (Roothaan-Hall approach),

$$\mathbf{FC} = \mathbf{SCe} \quad (23)$$

$$F_{\nu\mu} = \int \eta_\nu^*(\mathbf{r}) \hat{f}_i(\mathbf{r}) \eta_\mu(\mathbf{r}) d\mathbf{r} \quad (24)$$

$$S_{\nu\mu} = \int \eta_\nu^*(\mathbf{r}) \eta_\mu(\mathbf{r}) d\mathbf{r} \quad (25)$$

where \mathbf{F} is the Fock matrix, \mathbf{S} is the overlap matrix, \mathbf{C} the matrix of the linear expansion coefficients, and \mathbf{e} the diagonal matrix of the orbital energies. This formulation allows translation of the non-linear optimization problem of finding the wave function, into a linear algebra problem that can efficiently be handled by computational programs.

The approximation introduced in the Hartree-Fock approach has two main consequences. First, since the electron interaction is calculated against an average charge density, that is, instantaneous interaction is not accounted for, electrons tend to be too close to each other resulting in an augmented Coulomb repulsion (a.k.a. dynamic correlation). Second, the single Slater determinant approximation in the Hartree-Fock method is not suitable for nearly degenerate configurations where a better approximation would require the combination of multiple Slater determinants (a.k.a. static correlation).

As a result the variational method can only lead to a best estimate of the energy, E_{bestHF} , that is higher (less negative) of the exact ground state calculated with Born-Oppenheimer approximation and neglecting relativistic effects E_{exact} .

$$E_{corr} = E_{exact} - E_{bestHF} \quad (26)$$

The difference between these two is called correlation energy, E_{corr} .

Density Functional Theory

The number of variables involved in the electronic wave function is four times the number of electrons (three spatial coordinates and one spin variable for each electron). Nevertheless, the electron density ($\rho(\mathbf{r})$) described by the wave function depends only on the three spatial variables (the vector \mathbf{r}) regardless to the number of nuclei and electrons in the system. $\rho(\mathbf{r})$ is defined as the probability density of

finding any electron in $d\mathbf{r}_1$ with an arbitrary spin while all other electrons have arbitrary spin and positions:

$$\rho(\mathbf{r}_1) = n \int \dots \int |\psi(\mathbf{x}_1, \dots, \mathbf{x}_n)|^2 ds_1 d\mathbf{x}_2 \dots d\mathbf{x}_n \quad (27)$$

The $\rho(\mathbf{r})$ contains information such as the total number of electrons n ,

$$n = \int \rho(\mathbf{r}) d\mathbf{r} , \quad (28)$$

the position of all the nuclei, i.e., at the cusps of $\rho(\mathbf{r})$, and the nuclear charge of each nucleus from the limit of the derivate of $\rho(\mathbf{r})$ at each cusp. In addition, while the wave function has no experimental counterpart, the electron density can be observed experimentally, e.g., X-ray diffraction.

Building on Thomas and Fermi first attempts to use the electron density to describe the system (from 1927), Hohenberg and Kohn (1964) demonstrated that

$$\rho_0 \Rightarrow \{n, R_A, Z_A\} \Rightarrow \hat{H} \Rightarrow \psi_0 \Rightarrow E_0 , \quad (29)$$

since the electron density of the ground state ρ_0 fixes the number of electrons (n), the position (R_A), and charge (Z_A) of the nuclei, which determine the Hamiltonian and thus the ground state wave function ψ_0 and the energy E_0 , then the energy is a functional of the electron density (first Hohenberg-Kohn theorem):

$$E_0[\rho_0] = T[\rho_0] + E_{ee}[\rho_0] + E_{Ne}[\rho_0] , \quad (30)$$

where $T[\rho_0]$ is the kinetic energy, $E_{ee}[\rho_0]$ the potential energy due to electron-electron interaction, and $E_{Ne}[\rho_0]$ the potential energy due to nuclei-electron interaction, which is the only term that actually depend on n , R_A and Z_A , and also the only term currently known. In fact, although exact by derivation, the explicit forms of both $T[\rho_0]$ and $E_{ee}[\rho_0]$ are not known. These functionals, which are usually represented as the Hohenberg-Kohn functional

$$F_{HK}[\rho] = T[\rho] + E_{ee}[\rho] , \quad (31)$$

are not dependent on the system, i.e., independent from n , R_A and Z_A , and, if known, would be applied to any kind of system. Nevertheless, the lack of an explicit form of $F_{HK}[\rho]$ introduces the need for approximation. Moreover, the electron density is not known a priori and has to be found iteratively since, as from the second Hohenberg-Kohn theorem, only the electron density corresponding to the ground state, ρ_0 , returns E_0 , while any other approximated ρ_{trial} leads to $E_{trial} > E_0$, which is the DFT version of the variational principle.

As a means to reduce the approximation introduced by the lack of an exact form of $F_{HK}[\rho]$, the Kohn-Sham approach provided an orbital-based treatment resembling the Hartree-Fock method and that allows recovering most of the kinetic energy of $T[\rho_0]$. In fact, a single Slater determinant (Θ_s) is used as the solution to the Schrödinger equation where the Hamiltonian is built with a fictitious potential ($V_s(\mathbf{r}_i)$) that represents an ideal system of n non-interacting electrons (i.e., electrons behaving as fermions moving in the average charge field).

$$\Theta_s = \frac{1}{\sqrt{n!}} \begin{vmatrix} \vartheta_1(\mathbf{x}_1) & \cdots & \vartheta_n(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \vartheta_1(\mathbf{x}_n) & \cdots & \vartheta_n(\mathbf{x}_n) \end{vmatrix} \quad (32)$$

$$\hat{H}_s = -\frac{1}{2} \sum_{i=1}^n \nabla^2 + \sum_{i=1}^n V_s(\mathbf{r}_i) \quad (33)$$

The requirement for such a reference system is for the electron density to be equal to that of the real, fully interacting ground state of the system at hand. The so-called Kohn-Sham orbitals ϑ_i are determined as solutions of the n Kohn-Sham equations,

$$\hat{f}\vartheta_i = \varepsilon_i\vartheta_i , \quad (34)$$

$$\text{where } \hat{f} = -\frac{1}{2}\nabla^2 + V_s(\mathbf{r}) \quad (35)$$

is the Kohn-Sham operator. The kinetic energy for this reference system

$$T_S = -\frac{1}{2} \sum_{i=1}^n \langle \vartheta_i | \nabla^2 | \vartheta_i \rangle \quad (36)$$

is used as a fair approximation of that of the real, fully interacting system, $T[\rho_0]$, leaving apart only a smaller unknown correction to the kinetic energy, $T_C[\rho_0]$. Moreover, also the explicit form of the classical portion of the electron-electron Coulomb interaction $J[\rho]$ is known and is used to cover a portion of $E_{ee}[\rho_0]$, leaving apart the unknown non-classical portion of the electron-electron interaction $E_{ncee}[\rho]$. With this approach the Hohenberg-Kohn functional can be written as

$$F_{HK}[\rho] = T_S[\rho] + T_C[\rho] + J[\rho] + E_{ncee}[\rho] \quad (37)$$

$$\text{or } F_{HK}[\rho] = T_S[\rho] + J[\rho] + E_{XC}[\rho], \quad (38)$$

where $E_{XC}[\rho]$ is called the exchange-correlation functional and collects all the terms for which an explicit form is not known ($T_C[\rho_0]$ and $E_{ncee}[\rho]$). $E_{XC}[\rho]$ thus represents the connection between the reference system and the real, fully interacting one. The related exchange-correlation potential is defined as the functional derivative

$$V_{XC} \equiv \frac{\partial E_{XC}}{\partial \rho} \quad (39)$$

which allows writing $V_S(\mathbf{r})$, the fictitious potential of the Kohn-Sham operator (see equation 35), more explicitly,

$$V_S(\mathbf{r}_1) = \int \frac{\rho(\mathbf{r}_2)}{r_{12}} d\mathbf{r}_2 - \sum_A^m \frac{Z_A}{r_{1A}} + V_{XC}(\mathbf{r}_1) \quad (40)$$

where all terms with unknown explicit forms are encapsulated into the exchange-correlation potential V_{XC} . As for the Hartree-Fock method, the potential $V_S(\mathbf{r})$ depends from the solutions of the equation, i.e., the electron density by the Kohn-Sham orbitals ϑ_i , therefore the solution of the one-electron Kohn-Sham equations

(34) is found iteratively optimizing the expansion coefficients of the linear combination of basis set functions used to construct each ϑ_i , until the self consistency condition is satisfied. To this end an approximate (and to some extent empirical) form of the exchange-correlation functional $E_{xc}[\rho]$ has to be provided and, since this is the key point of the DFT method, defining a single, combined functional or two, separate functional for the exchange and the correlation contribution is an active research field.

The simplest approximation for the $E_{xc}[\rho]$ is the so-called local density approximation (LDA), which calculates the exchange and correlation contributions based on a model characterized by a constant electron density called the “uniform electron gas”. In practice, the local value $\rho(\mathbf{r})$ is assumed to have zero gradient even though $\rho(\mathbf{r})$ changes per each \mathbf{r} . Since the electron density in a molecular system is everything but constant, the LDA is a somewhat brutal approximation. Therefore, a more realistic generation of functionals considers, in addition to the local density in \mathbf{r} , also its gradient and a set of constrains to retain physical correctness of the model. This family of gradient-corrected functionals is called gradient generalized approximation (GGA) and is also said to be ‘non-local’ due to the gradient delivering information about the surroundings of the local point \mathbf{r} , and thus accounting for the non-homogeneity of the electron density. Extending this strategy even further has led to the introduction of the second derivative of the electron density and the so-called meta-GGA family of functionals. A different strategy, instead, has been that of introducing a certain amount of Hartree-Fock exact exchange into the functionals, which are then called hybrid functionals. The latter have somehow become standard for many common chemical systems, for instance, the B3LYP functional for modeling of organic molecules.

Until recent years, the DFT method has been affected by a significant shortcoming: the lack of dispersion interactions. London forces, also known as dispersion forces, are the consequence of instantaneously induced dipoles resulting by the interaction of electron densities that otherwise are not directly interacting, i.e., between non bonded

molecules or not closely bonded portions of the same molecule. As dispersion derives from the electron density of two separate points in the space, functionals considering only the local value of $\rho(\mathbf{r})$, and even its derivatives, cannot account for any dispersion. Though these non-bonded interactions are weak, if compared to those between permanent charges and multipoles, neglecting dispersion may seriously hamper accuracy. Although strategies to introduce dispersion have been developed during the last decade, and are mainly based in the inclusion of dispersion in the parametrization of the functional⁴⁸ and the use of empirical corrections (i.e., DFT-D by Grimme),⁴⁹ all the DFT calculations performed in the work presented in this thesis did not include dispersion (see Chapter 5). This conscious approximation is simply motivated by the need of comparing, and partially reproducing, previous calculations reported in the literature (see Paper II).

Semi-empirical Method

Approximate quantum mechanical methods have been developed in response to the high computational cost of *ab initio* calculations. In fact, in order to solve the Hartree-Fock equations (17), or the analogous Kohn-Sham equations (34) in DFT, the Roothaan-Hall approach (equation 23) requires the calculation of the matrix elements $F_{\nu\mu}$ which contain two-electrons integrals, for instance, resulting from the application of the Coulomb and exchange operators (\hat{J} and \hat{K}), that can involve up to four different basis functions.

$$\int \int \phi_i^*(\mathbf{r}_1)\phi_j(\mathbf{r}_1)\frac{1}{r_{12}}\phi_k^*(\mathbf{r}_2)\phi_l(\mathbf{r}_2)d\mathbf{r}_1d\mathbf{r}_2 \quad (41)$$

Since the calculation of these integrals is the more costly component of Hartree-Fock method, the semi-empirical approach aims to reduce the computational requirements by neglecting many of these integrals and providing empirical parameters to be used in lieu of other integrals. In addition, only valence electrons are treated explicitly while core electrons are included in the corresponding nucleus and their shielding taken into account in the definition of the valence electrons-nuclei interactions. Moreover, to further simplify the problem a minimal basis set where one Slater type

function per orbitals is used and off-diagonal elements of the overlap matrix (\mathbf{S} in equation 23) are often set to zero. Criteria for selecting the integrals to be neglected or approximated by empirical parameters are mainly based on the location of the basis functions, i.e., the atom on which ϕ_i , ϕ_j , ϕ_k , and ϕ_l are centred. For instance, the class of NDDO (neglect of diatomic differential overlap) methods, set

$$\int \phi_i^*(\mathbf{r}_1)\phi_j(\mathbf{r}_1)d\mathbf{r}_1 = 0 \quad (42)$$

where ϕ_i and ϕ_j belong to two different atoms, i.e., the diatomic differential overlap, but retains two-electrons, two-center integrals where the pairs ϕ_i , ϕ_j , and ϕ_k , ϕ_l are located on two different atoms, e.g., atoms A and B, but ϕ_i and ϕ_j are on A, while ϕ_k and ϕ_l on B. Empirical parameters include the Slater orbital exponents, and parameters to be used in lieu of the core-integrals, which calculate the kinetic energy of an electron moving in the field of nuclei shielded by the core electrons plus the potential energy of attraction towards such shielded nuclei. Again, different parameters are used for the core-integrals based on the location of the basis function involved and also on the type of such functions (i.e., orbital type s, p, or d).

A modern version still based on the NDDO approach include the parametric method number 6 (PM6), which was applied in Paper II (Chapter 5), that was extensively re-parametrized aiming for wider scope both in terms of training set (about 9000 species) and elements included (70 elements). The latest method (PM7) has also been provided with dispersion and H-bond terms that improve the treatment of intermolecular interactions.

2.2.2 Molecular Mechanics

Molecular mechanic (MM) methods are characterized by (i) approximate description of the molecular system based on atoms with implicit rather than explicit electrons, and (ii) mathematical models based on classical rather than quantum mechanics. The combination of these two characteristics makes MM methods orders of magnitude faster than any quantum mechanical and semi-empirical methods.

The lack of explicit electrons means that the fundamental particles in MM models are atoms, or even larger units (i.e., united-atoms and coarse grained approaches),^{50–52} each of which can be considered as if holding electrons in a fixed configuration. This approach finds justification in the Born-Oppenheimer approximation (see above). Neglecting the explicit treatment of electrons allows MM methods to deal with lower number of particles with respect to the quantum methods, but the obvious downside is that electronic properties and fine electronic effects are often neglected.

The energy of the system is calculated as a function of the atomic positions by means of empirical laws that describe the interatomic interactions following the principles of classical mechanics. As it is known that classical physics cannot accurately describe all phenomena occurring at atomic and molecular scale, the MM approach is based on the quantification of the energetic penalty, and the resulting force, associated with the deviation of the geometry from a reference condition that is defined a priori by means of empiric parameters. For this reason, the MM potential energy, which is not quantized, is more properly referred as to the “strain energy” and represents a relative quantity that is bound to the definition of the system at hand, i.e., number and type of atoms, bonds and geometrical features. A general definition of the strain energy is as follows:

$$E_{tot} = \sum_{ij}^{bonds} E_{ij}^{str} + \sum_{ijk}^{angles} E_{ijk}^{bend} + \sum_{ijkl}^{torsions} E_{ijkl}^{tors} + \sum_{i>j}^n E_{ij}^{nb} + \sum E^{other} , \quad (43)$$

where bond stretching, E^{str} , angle bending, E^{bend} , bond torsion, E^{tor} , non-bonded interactions among the n atoms, E^{nb} , are the four most common energy contributions, and E^{other} underlines that many other contributions can possibly be included in the model for a more refined description of the potential energy surface. Overall each independent term gives a contribution of “strain”, but it is the minimization of E_{tot} by variation of the atoms coordinated that allows identifying the compromise between all terms. Therefore, for each contribution a functional form, which aim to return the potential energy from the set of atom coordinates, has to be

provided together with a set of empirical parameters, thus *de facto* defining the force field under which each atom moves.

The stretching of bonds is usually approximated with a harmonic potential centred on a given equilibrium distance r_0 and weighted by the force constant k_b

$$E_{ij}^{str} = \frac{k_b}{2} (r_{ij} - r_0)^2, \quad (44)$$

where r_{ij} is the interatomic distance. Better accuracy, in particular with respect to anharmonicity, is obtained including higher order terms with the corresponding force constants (k_b' and k_b'' that may be related by a constant),

$$E_{ij}^{str} = k_b' (r_{ij} - r_0)^2 + k_b'' (r_{ij} - r_0)^3. \quad (45)$$

Although desirable, the more accurate Morse function is rarely used due to the need for three parameters, i.e., well depth (D), curvature (α), and reference distance (r_0), and the lower computational efficiency,

$$E_{ij}^{str} = D \left[1 - e^{-\alpha(r_{ij} - r_0)} \right]^2. \quad (46)$$

Also the angle bending term is most often represented by a harmonic potential, though a more general formulation includes, as for the bond stretching, higher order terms:

$$E_{ijk}^{bend} = k_a' (\theta_{ijk} - \theta_0)^2 + k_a'' (\theta_{ijk} - \theta_0)^3 + k_a''' (\theta_{ijk} - \theta_0)^4 + \dots, \quad (47)$$

where θ_0 is the reference value, and k_a' , k_a'' , and k_a''' are the weighted force constants.

The terms accounting for torsions about bonds need to reflect the periodicity of the motion, hence trigonometric functions are often used to define torsional potentials:

$$E_{ijkl}^{tor} = \sum_{m=0}^M \frac{k_m}{2} \left(1 + \cos(m\varpi_{ijkl} - \gamma) \right), \quad (48)$$

where ϖ_{ijkl} is the torsion angle, γ a phase factor, k_m are the contribution to the rotational barrier of the bond, and m is the multiplicity. While the use of torsion potentials is common practice, it should be noted that 1-4 interactions between the atoms i and l in the sequence $i-j-k-l$ might provide a discrete approximation of the energy profile. This approximation was exploited extensively in Papers II and IV (see sections 3.2.5 and 3.2.6).

The terms describing non-bonded interactions involve two major contributions: electrostatics, and van der Waals. The first, is usually formulated as an ensemble of Coulomb interactions (equation 49), possibly implementing a distance-dependent dielectric ϵ_0 (equation 50), that are calculated between partial and punctual charges (q_i and q_j) that may or may not be centred on atoms so to permit the description of multipoles.

$$E_{ij}^{chg} = \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}} \quad (49)$$

$$E_{ij}^{chgDD} = \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}^2} \quad (50)$$

Unfortunately, the definition of the partial charges is far from trivial due to the lack of explicit electrons in the model. Therefore, methods have been developed to assign fixed partial charges as empirical parameter, or calculate a balanced distribution of partial charges on the fly. Nevertheless, since partial charges are not experimentally observable, there is no unambiguous method to provide these values. Moreover, empirical assignation of partial charges may not be sufficiently general. Therefore, when modeling transition metal species, electrostatics is often absorbed in other functions rather than explicitly accounted for. For instance, parameters of other functions, such as bonding and van der Waals terms, are optimized as to reproduce experimental or *ab initio* results that do include electrostatic effects thus encapsulating part of the electrostatic contribution. Nevertheless, methods exist even to explicitly account for polarization.^{53,54}

The van der Waals term collect both attractive and repulsive contributions. The firsts represent the ensemble of favourable interactions resulting from the instantaneous dipoles due to fluctuations of the electron clouds, and are referred as to the London forces. These attractive forces are usually modelled with a r^{-6} term. The second contributions account for the strong repulsion occurring when non-bonded atoms are too close and Pauli's principle prohibits the overlap of their electron clouds. A common potential representing both these contributions is that known as the Lennard-Jones potential, which is generally written as

$$E_{ij}^{\text{vdw}} = k\varepsilon_{\text{vdw}} \left[\left(\frac{\sigma}{r_{ij}} \right)^n - \left(\frac{\sigma}{r_{ij}} \right)^m \right], \quad (51)$$

$$\text{where } k = \frac{n}{n-m} \left(\frac{n}{m} \right)^{m/(n-m)} \quad (52)$$

and ε_{vdw} defined the well depth, σ is the collision parameter, which defines the interatomic distance at which repulsion equals attraction, and the exponents are often $n = 12$ and $m = 6$.

A number of other terms are often included in the definition of the force field. These include improper torsion or out-of-plane terms, which account for the out-of-planarity displacement of atom l with respect to ijk plane in the tripod $i-j(-k)-l$, cross-terms, which account for the coupling between modes (i.e., stretching of two adjacent bonds, stretch-bend, bending of adjacent angles, bending-torsion), and H-bond. The actual list of terms constitute the fingerprint of a force field as it defines which kind of interatomic interactions and geometrical effects the force is capable of reproducing.

The need for defining empirical parameters for all these terms introduces to another fundamental characteristic of most MM methods. That is, the use of atom types to specify, and *de facto* impose, the properties of every atom in the system. Atom types allow assignation of the ensemble of empiric parameters adopted, or calculated on the fly with empiric rules, for each MM term. This is based on the empirical

consideration that atoms of a given element, in a similar chemical environment are likely to behave similarly. Therefore, the functional forms and parameters capable of reproducing experimental and *ab initio* geometries for some molecules, can be used to model other, yet similar molecules. This property is known as transferability of the force field, and is crucial for making predictions for new molecules based on parameters obtained from proper parameterization process.

As a consequence of the need for transferability, many force fields have developed a plethora of parameter sets for the most common elements found in organic and biological chemistry resulting in the use of different atom types for the same element depending on the chemical context surrounding an atom with given atomic number. Unfortunately, the picture is more complicated when transition metals are taken into account. In fact, these atoms display variability of oxidation and spin states, coordination numbers and geometries that make transition metal species particularly challenging for MM. Nevertheless, the prospect of modeling transition metal species with fast empirical MM approach has stimulated the development of method capable of accounting for the peculiarities of transition metals. Among others, the ligand field molecular mechanics (LFMM) method introduced the effect of the ligand field stabilization energy in standard, organic chemistry force fields. The LFMM method was integrated in a widely available MM tool as part of the work in this thesis, i.e., Paper III (Chapter 6), with the aim to increase the availability of the method and its use in *de novo* design. In fact, the resulting integrated tool was deployed extensively in Paper IV for the design of iron compounds (Chapter 7).

Finally, as MM is an empirical approach there is no truly “correct” combination of functional forms and parameters. Instead, the MM methods conform to the need of reproducing results that are defined either experimentally or with more accurate computational modeling methods. This property of the empiric approach was exploited in Paper IV where a special force field was developed to obtain multiple molecular ring closures from acyclic chains of atoms (see Section 3.2.6).

3. Generating Chemical Entities from Scratch

The automated generation of new candidates is the characteristic feature of *de novo* design methods that allows identification of candidates beyond traditional trends and expectation biases. In this chapter the background about automated generation and modification of molecules is introduced (Section 3.1) highlighting the limitations that led to the development of a new method, which is described in Section 3.2. The new method is suited not only for simple molecules but also for chemical entities with peculiar geometries and arrangements of atoms, reaction intermediates, and even transition state models.

3.1 Introduction

3.1.1 Building Strategies

Building Blocks: Atoms vs Fragments

Strategies for automated generation of molecular objects can be divided according to the type of building blocks used, which can be either atoms or molecular fragments.¹⁷ Methods implementing atom-based building strategies combine atoms of various elements according to valence rules and average elemental composition. This strategy can potentially build every possible chemical object,¹³ and atom-based building strategies have been implemented in a number of *de novo* methods exploiting such extreme generality.^{55–61,38,62} However, the major drawback of atom-based builders is that most of the combinations of atoms do not represent stable molecules or are not synthetically accessible, thus strategies to control the assembling of atoms have to be included.⁶³ The problem is even worse for transition metal compounds, due to the high coordination number and extended list of bond types.^{30,64} In general, automated molecular builders need to balance novelty and confined chemical space.⁶⁵ This is a well-known issue in *de novo* drug design, and strategies to reduce the amount of unrealistic candidates generated from atom-based builders are mostly based on filters that reject unacceptable chemical features^{13,66} or evaluate the synthetic accessibility on the fly.^{67–71}

Alternatively, molecules are built by assembling molecular fragment rather than single atoms.^{72–83} While the use of multi-atom building blocks implies a low resolution in the exploration of the chemical space,⁸⁴ it also allows to control the type of functionalities generated and to avoid most of the unrealistic candidates. The combination of selected molecular fragment and connection rules efficiently confines the chemical space according to the specific needs and defines a subspace referred as to the fragment space.^{65,85} Connection rules are needed to avoid formation of undesired chemical features that would otherwise require the application of filters as for the atom-based approach.⁶⁶ Fragments are typically generated by retrosynthetic fragmentation of existing molecules⁸⁶ and reconnected according to synthetic principles thus introducing the chemical reasoning directly into the automated builder.^{72,75,87,88}

A radically different approach for automated builders is that of virtual synthesis.^{89–92} This method generate molecules by performing virtual reactions between molecules that are taken from databases of commercially available reagents, thus mimicking the experimental workflow. However, automation of this approach requires robust and standardized classification of synthetic reactions and functional group reactivity,^{93–95} none of which is currently available for transition metal chemistry.

Builders for Transition Metal Compounds

All methods so far reported in the literature for the automated generation of transition metal species implement fragment-based building strategies.^{30,96,97} The software developed by Hay and co-workers (named HostDesigner)⁹⁸ is meant to build libraries of receptors (hosts) that show affinity for a given target (guest). The algorithm combines “complex fragments”, which encapsulate each host-guest bonding interaction, and linker fragments, which are meant to connect pairs of complex fragments. Such user-defined complex fragments specify the metal-coordinating moieties, thus preventing the formation of any undesired bond involving the metal atom, but all connections with the linkers are formed replacing hydrogen atoms without consideration of synthetic accessibility. HostDesigner has been applied to the design of receptors for organic molecules^{99,100} and also for multidentate ligands with

high affinity for transition metal ions.^{96,101,102} Instead, Rothenberg and co-workers^{97,103} deployed an automated building process to create combinatorial libraries of catalysts with bidentate ligands. The strategy was based on classifying fragments according to their role in the final transition metal complex (i.e., metal, ligand groups, bridges, and decorating groups). Contrarily to HostDesigner, metal–ligand bonds were generated by connection of specific classes of fragments (i.e., the ligand groups and the metal atom). A similar classification of the building blocks was applied by some of the authors involved in the present project with the aim to overcome the limits of the chemical representation deployed (see section 3.1.2).³⁰ In particular, the transition metal complex was divided in three layers: a metal-containing core fragment, which encapsulated also some of the ligands, a specific class of metal-ligating fragments (the “trial parts”), and a further class of fragments acting as decorations for both core and trial parts (the “free parts”). Again, the use of classes of fragment was exploited to impose the chemical surrounding of the metal atom, but no control on the rest of the connections was possible.

Connection Rules

A common feature in the three methods introduced in the previous section, is that fragments are organized in classes and the building process assembles compounds following a layered scheme that, as a blueprint, defines which fragment class to use in which position of the final molecule.^{30,97} These approaches are therefore based on a fragment class compatibility scheme. On the contrary, the retrosynthetic strategy diffused in organic drug design focuses on the chemical features of the attachment points on each connected fragment rather than on the fragment as a whole; that is, an attachment point compatibility scheme is applied.^{72,75,87,88} The approach based on fragment class compatibility is significantly more limited than that based on attachment points classes. In particular, since the class of the fragment is a property of the whole building block, the fragment class does not provide information on the single attachment points. Moreover, without a precise definition of the chemistry generated by the connection of two fragments of different class, the method is prone to faulty generation of unrealistic or unacceptable functionalities, unless all possible

connection are evaluated manually prior to deployment. Finally, the class of a fragment is an arbitrary, system dependent property resulting from manual work, rather than data mining, and limits the possibility of reusing libraries of fragments.

3.1.2 Chemical Representations

Implementation of an automated builder requires the use of a virtual representation that allows definition and modification of chemical entities and related information. The choice of the chemical representation for an automated builder can seriously affect the capabilities of the builder to handle peculiar species.

The Panorama of Chemical Representations

In all branches of computer-aided chemical research the use of machine-readable representations of chemical compounds is mostly established.¹⁰⁴ Nevertheless, continuous developments demonstrate that as the cheminformatic tasks evolves so does the need for suitable representation of chemical entities, concepts, and information.¹⁰⁵ In fact, the recent developments tend to integrate various types of information in a machine readable fashion as to facilitate the execution of automated tasks.¹⁰⁶

The coordinate-less, string-like representations are often utilized as a general purpose chemical nomenclature that is readable both by humans and machines, and allows definition of the composition, connectivity, and stereochemical descriptors.¹⁰⁷ Typically, the simplified molecular-input line-entry system (SMILES)^{108,109} serves this purpose, but also the modular chemical descriptor language (MCDL)^{110,111} and the InChI^{112,113} molecular identifiers are used. Connection tables are alternative graph-based representations that have led to some of the most diffuse file formats for sharing chemical data,¹¹⁴ and may include bi- or three-dimensional (3D) spatial coordinates. String-like and graph-based representations are often simplified by the removal of non-chiral hydrogen. This simplification, which leads to a significant reduction of the number of atoms, is based on the assumption that the presence of such hydrogen atoms can be inferred from the hydrogen-depleted structure. Further simplification leads to the so-called reduced representations,⁵¹ where the smallest

building units are groups of atoms, or even groups of small molecules, thus providing a coarse-grained description of the system.^{50,51} Instead, an all-atom 3D representation that provides atomic coordinates, either in the form of Cartesian or internal coordinates,¹¹⁵ constitutes the current standard for input/output in the majority of molecular mechanics and quantum chemistry tools. Notably, while most of molecular mechanic methods depend on a definition of the connectivity, this information is usually superfluous for quantum mechanical modeling.

Chemical Representations for De Novo Design

The chemical representations deployed in *de novo* design methods need to identify the candidate and support structural modification. Thus, beside elemental composition, connectivity, and stereochemistry, a representation should allow quick identification and modification of the building components, i.e., atoms and bonds, fragments, or synthetic reaction steps. Further additional information may also be required, for example, geometrical information (i.e., coordinates, docking poses, multiple conformations) and molecular descriptors. Since the chemical representations needs to contain all necessary information to create a candidate chemical entity, and most *de novo* design methods apply evolutionary algorithms as optimization method (see Section 2.1), the chemical representation is commonly referred as to the *chromosome*.^{35,37}

Graphs efficiently represent the identity of a chemical structure,¹¹⁶ and have been used as chromosomes in *de novo* design methods.^{30,38,117} In fact, with graph vertices (or nodes) representing the building blocks and graphs edges the connection between them, the alteration of a building block or connection can easily be performed by modification of the corresponding vertex or edge. Accordingly, graph-based representations such as connection tables and SMILES have been widely exploited as chromosome.^{61,63,75,76,92} Alternatively, SMILES and 2D graphs have been used to represent only the molecular structure of the building blocks, i.e., the content of the vertices, while a dedicated data structure defined the surrounding graph-like chromosome.^{30,73,91,118,119} The analogous strategy with 3D representation has been used,^{58,66,74,88,120–122} but mostly in connection to *fragment-based drug discovery*,¹²³

which should not be confused with the *fragment-based building strategy* introduced above, although the latter is usually implemented as part of the former.⁷⁸

With an alternative use of graphs, the construction and modification of molecules can be seen as a short walk on a graph that spans all possible fragments and connections. Changing the path corresponds to modify the type of fragments and their ordered sequence, thus changing the molecular structure.^{17,76,124}

Information Content and Conversion

Different chemical representations contain different amounts and types of information, and conversion between representation may be accompanied by loss or creation of information.¹¹⁷ In particular, conversion from low information content chemical representations to highly informative ones (for example, from SMILES to 3D), is often mandatory to proceed with further modeling. The conversion is usually performed under the assumption that most of the additional information required to define the output (in the example, the atomic coordinates) can be inferred from the low information content input by means of standardized set of empiric criteria implemented in a conversion tool (for example, list of possible atom types, bond angles and lengths). While such artificial augmentation of the chemical information is likely to produce a proper output for the most common types of chemical compounds, the same cannot be said for peculiar chemical species. Unfortunately, this is the case for many transition metal species, and, in particular, for active species and reaction intermediates that are likely to be the focus of catalysis design. In order to safely handle such peculiar chemical entities, low information chemical representations should either be proven to handle the systems at hand properly or avoided in favour of all-atoms (i.e., no implicit atoms), three-dimensional representation, possibly allowing for extension of the information content. Examples of faulty treatment of implicit hydrogen atoms, stereochemistry, and molecular geometries are discussed in Chapter 5.

3.2 The General Purpose Fragment-Based Design Machinery

In light of the background presented in the previous section, a new method was introduced to enable automated generation and modification of chemical entities including peculiar ones that are not properly handled by other tools. The method exploits an internal graph-based representation (Section 3.2.1) designed to enable fragment-based construction and modification of any sort of chemical entity preferably from 3D building blocks, though lower dimensionality is also supported. The particularity of the method, and its strength, is that neither building blocks nor their connections are required to satisfy valence rules, geometrical constraints, or implicit atom formalism. To counterbalance the removal of such formalisms, all attachment points are labelled with codified chemical information (i.e., the class of the attachment point, see Section 3.2.2) that allows controlling the generation of new chemical entities (Section 3.2.3) by means of tuneable connection rules (Section 3.2.4). Assembling of 3D building blocks enables the preparation of the initial 3D geometries of tree-like graphs reducing to the minimum the need for empirical parameters and force fields (Section 3.2.5). Handling of any type of multi-fragment ring introduces a challenging conformational problem in the generation of graphs and preparation of 3D models. Thus, an empirical ring-closing potential (Section 3.2.6) is introduced to generate the 3D geometry of entities represented by cyclic graphs.

3.2.1 Graph Representation and Fragments

A chemical entity is represented by a graph $G(V,E)$ where $V=\{v_1, \dots, v_n\}$ is the set of vertices and $E=\{e_1, \dots, e_m\}$ is the set of edges (Figure 1). Each vertex is a container for a single molecular fragment, and each edge a connection between two different vertices (no self loops, no multiple edges). The graph can contain cycles of vertices, but it is always represented in term of a spanning tree $T(V, E')$, where $E' \subseteq E$ and T contains no cycle of vertices, and a set of fundamental cycles $Fc=\{C_l, \dots, C_k\}$. A single, root vertex is unambiguously identified by the graph generation algorithm (Section 3.2.3) and corresponds to the root of the spanning tree (green vertex in

Figure 1 C). Moreover, every non-root vertex is reachable starting from the root by a direct path in T (black arrows in Figure 1 C). Finally, each fundamental cycle includes only one chord of T (red arrows in Figure 1 C), which is an edge belonging to E but not to E' (for an overview on the nomenclature see ref. 125).

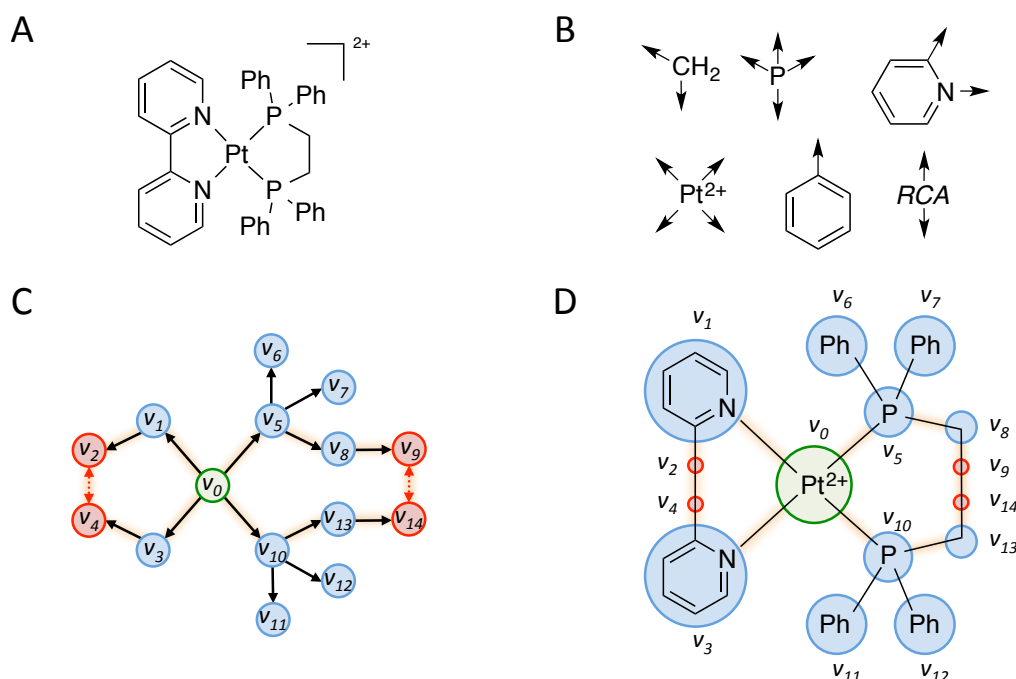


Figure 1: The graph representation of complex **A** built from the set of fragments **B** is displayed in **C**. The graph has two chords (red arrow in **C**) each connecting two ring-closing vertices (red circles in **C**) that contain only a dummy atom (ring-closing attractor, RCA). The correspondence between graph components and molecular representation is shown in **D**.

A molecular fragment contains atoms, dummy atoms, bonds and attachment points. The molecular structure of a fragment is represented by a graph where each vertex is an atom (or dummy atom) and each edge is a bond or a formal connection. All atoms are explicitly represented in the fragment. Both atoms and bonds are characterized by properties, which include element symbol, 3D coordinates, charge, and bond order. The CDK^{126,127} implementation of AtomContainer is used to represent the molecular structure of a fragment. In addition, fragments contain one or more attachment points (APs) that represent the possibility of binding other fragments. Each AP contains

information defining the atom (or dummy atom) prone to host the new connection, the class of the attachment point (AP class, see Section 3.2.2), and geometrical information defining the preferred spatial location of a hypothetically connected atom (AP vector).

Fragments and APs are not required to respect valence rules or stick to prefixed geometries. Therefore, peculiar chemical systems can be represented within fragments or formed by their connection. The only requirement for the fragments is that all atoms have to be connected to the rest of the fragment, thus creating a single network that spans all the atoms of the fragment. These intra-fragment bonds, as well as inter-fragment bonds that result from definition of edges in the graph G , are connections that do not represent a precise bond type, though bond orders may be specified to facilitate atom typing and conversion to standard molecular representations.

Overall, the representation consists in two layers of graphs: the outer layer (G), which handles information as to the collection of building blocks and the connections between them, and an inner layer of graphs, collecting all the graph representations of the fragments. Conversion of this data structure into a 3D molecular model is discussed in detail in Section 3.2.5.

3.2.2 Generation of Fragments and Cutting Rules

Molecular fragments (2D or 3D) are generated by automated fragmentation of existing molecules and computationally modelled structures. The process allows mining of chemical and geometrical knowledge on molecular building blocks from existing libraries, and projecting it into the generation of new candidates. To this end, molecules are fragmented according to user-defined cutting rules that identify specific target bonds. The cutting rules, which are codified by means of SMARTS¹⁰⁷ (see Figure 2), allow identification of chemical features surrounding the target bond and discriminate between the two sides of such bond. Removal of a matched bond generates two APs, each with one of the two complementary AP classes that derive from the unique identifier of the cutting rule. This way, information on the chemical

environments generating the AP is annotated into the AP class of each fragment. Moreover, if 3D structures are used as input, the geometrical relation between the two disconnected fragments (bond distance and angle, but not torsion) is also recorded in the APs in the form of two AP vectors, one per each AP generated (Figure 2).

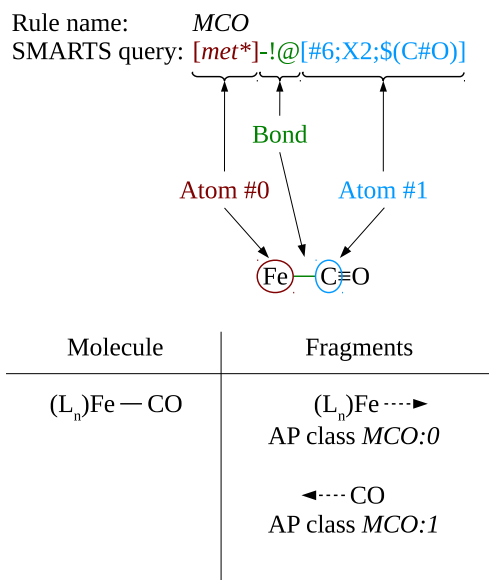


Figure 2: Examples of a cutting rule derivation of AP classes (*met** = list of transition metal symbols). Reprinted with permission from ref. 128. Copyright 2014 American Chemical Society.

Notably, while standard retrosynthetic rules are often applied to fragmentation of organic features,¹²⁹ the new method allows for definition of cutting rules matching multihapto binding sites. Moreover, the method distinguishes between multihapto and multidentate ligands, thus permitting independent handling of each metal-interacting group. In addition to standard fragmentation strategies, such as cutting all rotatable bonds or retrosynthetic approaches,^{85,86} the new method enables fragmentation of metal-organic connections based on principles of organometallic chemistry (for instance, the isolobal analogy).¹³⁰ Nevertheless, only constitutionally diverse arrangements of atoms can be discriminated as no geometrical information is currently supported by the cutting rules. Moreover, though the fragmentation procedure can handle large libraries of existing molecules, the list of cutting rules

remains an empirical classification of a specific set of target bonds, and the design, refinement, and management of cutting rules is let up to the intuition of the chemist.

3.2.3 Generation of Graphs

The generation of graphs proceeds via four main steps: (i) selection of a fragment for the root vertex, (ii) growth of the graph by recursive appending of vertices, (iii) growth termination by saturation of AP with single-AP fragments (capping procedure), and (iv) definition of the fundamental cycles. It should be noted that a cyclic graph is constructed as a tree, i.e., the spanning tree. Only after growth termination the set of fundamental cycles is defined thus making the graph cyclic.

The choice of the root fragment is performed among a given library of candidates. While any fragment can be used as root, the choice of a proper root ensures the presence of fundamental chemical features that may be required for the manifestation of the functional property.³⁰ For instance, the central metal atom is typically chosen for the design of transition metal compounds.

The growth of the graph is controlled by connection rules, which are further explained in Section 3.2.4, and parameters. The latter define (i) the probability of attaching a fragment at a certain level of the tree (i.e., the number of edges from the root vertex), (ii) the probability of projecting a branch of vertices from an AP to all the APs related by constitutional symmetry, which is defined according to AP class and connected environment, and (iii) threshold values for molecular weight, number of non-hydrogen atoms, and number of edges.

A particular feature of the building algorithm is the need for a growth termination step, which is referred as to the *capping* procedure. Due to the freedom characterizing the chemical features of the building blocks and the decision to avoid implicit atom formalism, APs corresponding to open valences and vacant coordination sites need to be saturated. To this end, appropriate single-AP fragments, i.e., capping groups (typically $-H$ and $-CH_3$), are appended on free APs having specific AP classes. On the contrary, other unused APs may not be suitable for the capping procedure nor

should be let free. Graphs with such forbidden ends are therefore rejected according to user-defined settings.

Handling of rings is divided in two cases that can coexist. If rings are entirely embedded within single fragments, then they are handled each as a single, not modifiable unit. In this case, although the inner graphs are indeed cyclic, the outer graph may be acyclic if there is no cycle of vertices. Alternatively, if one or more rings involve more than one fragment, then these rings define cycles of vertices in the outer graph. These cycles of vertices are defined by the set of fundamental cycles each of which includes only one chord of the spanning tree. Thus, to define the cycles of vertices chords must be identified. In this method, chords can connect only special purpose vertices, which are called *ring-closing vertices* (RCV, red nodes in Figure 1, page 39). An RCV is a regular vertex containing a special type of fragment that carries only one dummy atom, i.e., the *ring-closing attractor* (RCA in Figure 1), and two APs, one of which is reserved for graph growth and the other only available for chord formation. A chord can connect two RCVs only if these are compatible, and if the chord respects AP class connection rules (see below, Section 3.2.4) and satisfies ring size requirements (i.e., minimum and maximum number of atom members). In addition, when 3D building blocks are used, pairs of RCVs are connected only if the ring closability condition is satisfied; that is, the path connecting the RCVs in the spanning tree corresponds to a geometrically closable chain of atoms terminating with the RCAs. This condition is currently evaluated without alteration of the bond angles along the atom chain, though methods exist that allow inclusion of a tuneable amount of bond angle and bond length adaptation.¹³¹⁻¹³⁵ An additional condition is bound to the possibility of having multiple fundamental cycles sharing one or more bonds. This scenario introduced the need to evaluate the simultaneous closability of interdependent chains. Nevertheless, preliminary attempts to evaluate this condition by searching for a common ring closing conformation for all interdependent chains have been found more computationally demanding than filtering candidates based on successful conversion to 3D (see Section 3.2.5 and Paper IV).

Last, after the definition of the set of fundamental cycles, the graph is finalized by replacing unused RCAs with proper fragments, according to the capping procedure. The graph is then fully defined and ready for further operations.

3.2.4 Connection rules and Compatibility Matrix

Connection rules define which pairs of AP classes are allowed to connect fragments. The connection rules are collected in a square, non-symmetric Boolean matrix called the *compatibility matrix* (see Figure 3), where a *True* entry represents compatible pairs of AP classes.

AP class on fragments to be attached






		AP class on fragments to be attached					
		←··CO		←··NR ₂		 ketone:0	
		←··Fe	←··OR				
		MCO:0	MCO:1	alcohol:0	amine:0		
AP class on the growing molecule	Fe···	MCO:0	T/F Fe-Fe	T/F Fe-CO	T/F Fe-OR	T/F Fe-NR ₂	T/F 
	OC···	MCO:1	T/F OC-Fe	T/F OC-CO	T/F OC-OR	T/F OC-NR ₂	T/F 
	RO···	alcohol:0	T/F RO-Fe	T/F RO-CO	T/F RO-OR	T/F RO-NR ₂	T/F 
	R ₂ N···	amine:0	T/F R ₂ N-Fe	T/F R ₂ N-CO	T/F R ₂ N-OR	T/F R ₂ N-NR ₂	T/F 

Figure 3: Structure of the compatibility matrix and representation of the chemical features resulting by connection of fragments. Reprinted with permission from ref. 128. Copyright 2014 American Chemical Society.

The complete list of AP classes is used to index both rows and columns of the matrix, but while the row index refers to the class of the AP on the growing molecule, that of

columns refers to the class of the AP on the incoming fragment. As a consequence, the matrix is asymmetric and each entry of the compatibility matrix represents the chemical feature resulting from the connection of those represented by the row and column AP classes in the given order. The asymmetry of the compatibility matrix reflects the directionality of the edges belonging to the spanning tree. Instead, the evaluation of AP class compatibility in the formation of chords, which are undirected edges (see section 3.2.1), requires a symmetric compatibility rule. In fact, chords may involve any pair of compatible RCVs that belong to the same graph and the directionality of the growth is not relevant. Therefore, a second, symmetric compatibility matrix is dedicated only to the evaluation of AP class compatibility in the context of chord formation. Moreover, in such context the compatibility is evaluated between the APs that hold the RCVs rather than those of the RCVs. In fact, the RCVs are chemically empty (i.e., contain only a dummy atom) and the chemical feature created by chord formation is that defined by the combination of the AP classes of the APs holding the RCVs.

3.2.5 From Graph to 3D Molecular Model

The construction of a 3D molecular model for the entity defined by the graph involves two main steps: preparation of a tree-like 3D model corresponding to the spanning tree (i.e., chords are ignored), and folding of the tree branches to achieve the best conformation with closed rings. This two-step strategy reduces to the minimum the need of force field parameters while exploiting all the 3D information stored in the 3D fragments and avoiding the use of templates for cyclic systems.

The preparation of a tree-like structure exploits the fact that each 3D fragment, in addition to the full spatial characterization of all its atoms, contains information as to where in space all the connected fragments should be placed: the AP vectors (Figure 4).

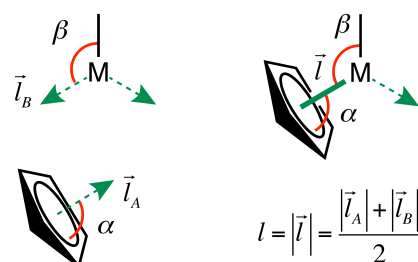


Figure 4: Two 3D fragments with AP vectors (green arrows), and the geometry resulting from connection of the fragments. Reprinted with permission from ref. 136. Copyright 2014 American Chemical Society.

These vectors allow the definition of the bond length and bond angles of all fragment–fragment connection, but do not provide information as to the torsion along such connections. Therefore, arbitrary conformations are used while assembling the tree-like 3D model starting from the root vertex and following the spanning tree.

In the resulting 3D molecular structure, all the rings defined in the graph by the set of fundamental cycles are not rings at all, but open chains terminated by RCAs. In order to obtain the final geometry, all such RCA-terminated chains must be folded so that the chain ends assume a relative orientation consistent with the definition of a bond between the atoms holding the RCAs. For a general system with any number of closing chains, any substitution pattern on such chains, any surrounding environment, and any number of possibly interdependent chains the solution of this folding problem is searched with a global optimization technique based on conformational search. To this end, the current implementation assumes, in first approximation, that closure of the rings requires only negligible deformation of bond angles. Thus, the conformational adaptation is reduced to a problem of torsions around the rotatable bonds, which include all fragment–fragment bonds and all bonds matching a user-defined list of rotatable bond types. The assumption that bond angles do not change significantly is justified by the use of 3D building blocks that, as the result of fragmentation of 3D structures, already have bond angles compatible with the ring conformation. Nevertheless, this assumption limits the creation of strained rings from 3D fragments that are not meant for such rings, and reduces the overall accuracy of

3D models of molecules containing rings of fragments. Thus, although not yet implemented in this method, the inclusion of a tuneable amount of bond angle deformation has been reported in the literature for other contexts^{131–135} and is going to be integrated in future improvements.

Within the approximation of fixed bond angles and lengths, the simultaneous closure of all rings with concomitant relaxation of all acyclic chains is achieved by coupling the potential smoothing and search algorithm for the torsional space (PSSROT procedure^{137–140} implemented in the Tinker package)¹⁴¹ with a specifically developed *ring-closing potential* (RCP), which is described in detail in Section 3.2.6. The outcome of this conformational search can be either a successfully folded geometry, where all rings have been closed, or a poorly folded conformation with incomplete ring closures. The latter case may be due to steric hindrance, impossibility of simultaneous ring closure of interdependent chains, or faulty solution of the global optimization problem. In all cases, this event is interpreted as a defective graph that, in *de novo* design context, is rejected.

The same type of conformational search, i.e., the PSSROT procedure, is applied also to relax the conformation of 3D models that have no cycle of vertices (acyclic graphs). This time, though, the definition of the potential energy simply ignores the components associated with ring closure and, in the simplest case, corresponds to the van der Waals term of the universal force field.¹⁴² Such a simplified, yet generally applicable force field can be used on any sort of chemical specie as PSSROT calculations operates only on torsions. In fact, the bond angles and lengths in the final 3D model are still those of the 3D building blocks or, for bond corresponding to fragment–fragment connections, the bond length calculated from the building blocks (Figure 4).

3.2.6 Ring-Closing Potential

The ring-closing potential (RCP) is an empirical force field that has been designed as part of this work to identify low-energy conformations in which the relative spatial arrangement of specific atom pairs is consistent with the definition of a new bond

between the two atoms in each pair. When the atoms in one such pair are the disconnected ends of an open chain, the formation of the new bond defines a new ring and the conformation of the chain is said to be a ring-closing conformation (RCC). The proper bonding orientation for each end of an open chain is intrinsically defined by the geometry of the atom–RCA connections terminating the chain (e.g., the relative position of RCA_A and SA_A in Figure 5). In fact, the RCA at one end of the chain (e.g., RCA_A in Figure 5) and the atom holding the corresponding RCA at the other end of the closing chain (e.g., SA_B in Figure 5) represent two images of the same atom. Therefore, the closure of the ring is obtained by overlapping the two pairs of images each pertaining to one of the two atoms involved in the new bond (i.e., SA_A and SA_B).

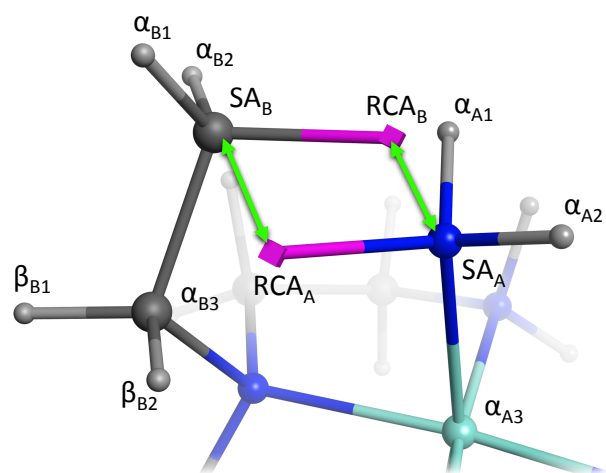


Figure 5: Example of a closing chain. Ring-closing attractors (RCA) are depicted as magenta squares and green arrows represent the attractive interactions of the ring-closing term.

Application of this concept in the context of PSSROT conformational search has led to the definition of the RCP as a sum of two contributions: the non-bonding interactions term ($E_{vdW}(t)$), and the ring-closing term ($E_{rct}(t)$):

$$E_{RCP}(t) = E_{vdW}(t) + E_{rct}(t), \quad (53)$$

where t is the potential surface smoothing parameter characterizing the PSSROT procedure. The first term corresponds to the van der Waals component of the universal force field (UFF),¹⁴² with the only difference that atom proximity reduction factors are applied as if the atoms holding the RCAs were connected (SA_A and SA_B in Figure 5). This practice reduces the interatomic repulsion between the two chain ends, which would otherwise repel each other. Instead, the ring-closing term defines a purely attractive interaction between the RCA on one end of the chain and the atom holding the RCA on the other end of the chain (green arrows in Figure 5). The functional form of this attractive interaction has been designed to be compatible with the PSSROT procedure, meaning that it is a solution of the semi-infinite, one-dimensional diffusion equation and the time (t) represents the smoothing parameter for the potential smoothing protocol.¹³⁹ The ring-closing term is then defined as follows (indexes as from Figure 5):

$$E_{rct}(t) = \sum_{C \in Fc} (f_{rct}(RCA_{A,C}, SA_{B,C}, t) + f_{rct}(RCA_{B,C}, SA_{A,C}, t)), \quad (54)$$

$$f_{rct}(RCA_i, SA_j, t) = \sum_{k=1}^3 \frac{a_{ij,k}}{(1+4Dt)^{3/2}} \exp\left(-\frac{b_{ij,k} r_{ij}^2}{1+4Dt}\right), \quad (55)$$

$$a_{ij,k} = a_k \alpha_{ij}, \quad (56)$$

$$b_{ij,k} = \frac{b_k}{\beta_{ij}^2}, \quad (57)$$

where the parameters α_{ij} and β_{ij} depend on the type of the RCAs and constitute the force field parameters, a_k and b_k define the shape of each of the three Gaussian functions, D is the diffusion constant and r_{ij} the distance between RCA_i and SA_j .

This implementation has been provided with a preliminary set of force field parameters developed to test the RCP in a real case scenario that is discussed in Chapter 7.

4. Design of Realistic Organometallic Compounds

As a test case for the fragment-based machinery described in Chapter 3, we evaluated the automated generation of organometallic ruthenium compounds with general formula $(L)_2(X)_2Ru=CR_2$, where R is H or aryl, L is a neutral ligand (i.e., two-electron donor) and X an anionic ligands (i.e., one-electron donor). Compound matching this formula are used as pre-catalysts for olefin metathesis (the active specie is generated by loss of one L).^{143–146} The wide possibility of altering the ligands set represents a good testing ground for the automated generation of molecules that should span structural variability while retaining synthetically accessible and type of coordination environment. In a word, only realistic candidates should be generated.

Two case studies were presented in Paper I. Case study I aimed at the automated generation of 2D compounds matching a predefined structural diversity scheme that imposed (i) high diversity for the dative ligands (L), (ii) use of selected and non-modifiable anionic ligands (X) and (iii) restricted substituents on the carbene. Thus, this test case also shows how the machinery can be used to exert control on the molecular building process. Next, case study II was focussed on a subclass of compounds, still based on the $(L)_2(X)_2Ru=CR_2$ general formula, that are known as the Hoveyda-Grubbs-type of catalysts.^{147,148} In this case, the machinery was used to build compounds using building blocks that were taken only from existing ligands.

4.1 Results and Discussion

For case study I, fragments were produced from more than 20000 molecules taken from Cambridge Structural Database¹⁴⁹ and fragmented according to cutting rules designed in two different strategies depending on whether metal atoms were involved or not in the matched bond. For bonds not involving metals, retrosynthetic reasoning was applied expanding the RECAP⁸⁶ set of rules to include chemical species of particular interest in transition metal chemistry. In particular, phosphines, *N*-heterocyclic carbenes, and multihapto ligands were added. Metal-involving cutting rules were instead designed to discriminate different types of ligands (one- and two-

electron donors) and hapticity. Moreover, to ensure high resolution of the fragmentation process, ligand types were further resolved, for instance, discriminating between M–amine, M–*N*-heteroaromatic, and M–phosphine bonds. For the fragments resulting by application of these cutting rules, two connection strategies were defined: first, a compatibility matrix reproducing a valence-only reconnection scheme (84.4% of *True* entries in the matrix, case A), and second, a compatibility matrix based on retrosynthetic approach for organic-only connections, and metal-ligand connections respecting the type of ligand and the designed diversity scheme (0.3% of *True* entries in the matrix, case B). The combination of the library of fragments and each of the two compatibility matrices defined two different organometallic fragment spaces from which molecules could be generated automatically. Examples of the generated molecules are given in Figure 6.

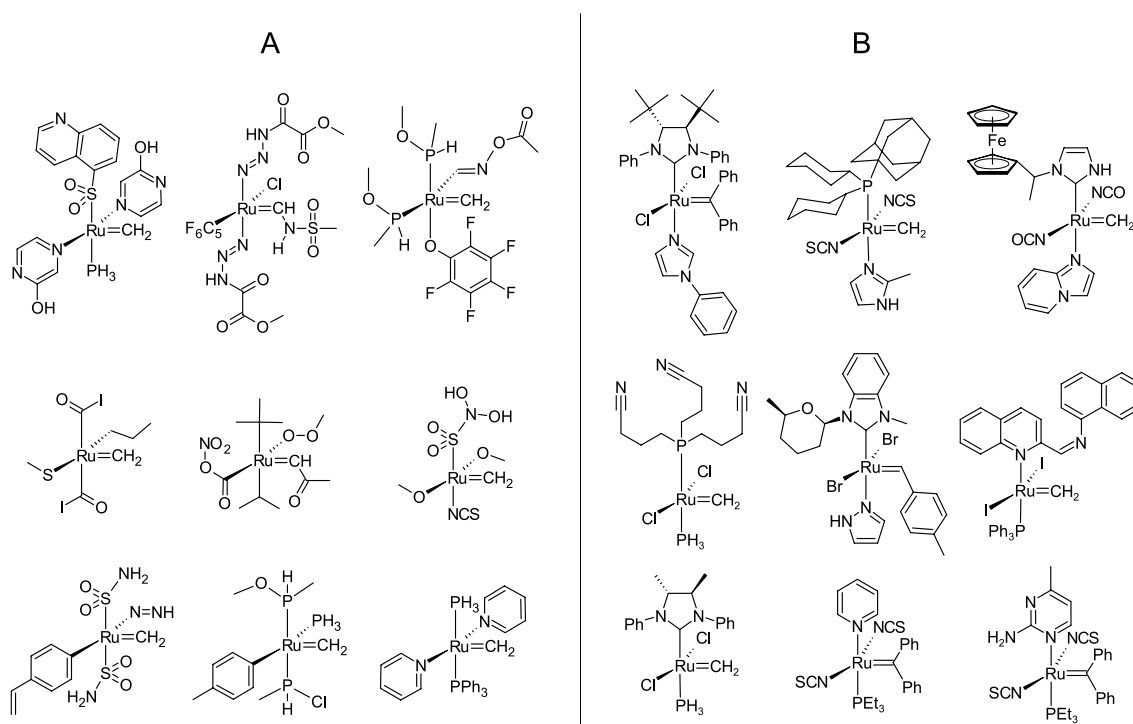


Figure 6: Example of molecules generated from the valence-only AP-compatibility scheme (A) and the strictly controlled organometallic fragment space (B).

The compounds produced by a strictly controlled definition of the compatibility matrix (B in Figure 6) demonstrate the capability of avoiding nonsensical connections

of fragments, such as many of those generated by valence-only AP-compatibility (A in Figure 6), while imposing the required metal-coordinating environment. Many of the ligands build with the sparse compatibility matrix (B) are, or resemble, commercially available or previously synthesized molecules, and span a wider structural diversity than that of the fragmented molecules. Moreover, while the precise selection of AP class compatibilities (B) has led exclusively to compounds with the required metal coordination environment, only a minority of the molecules generated from the valence-based connection rules (A) present the required coordination environment (Figure 7). Nevertheless, the results show that while the type of the coordinating ligand is always the same in B, different kinds of L and X ligands were used to build such metal-coordination environment (B in Figure 6).

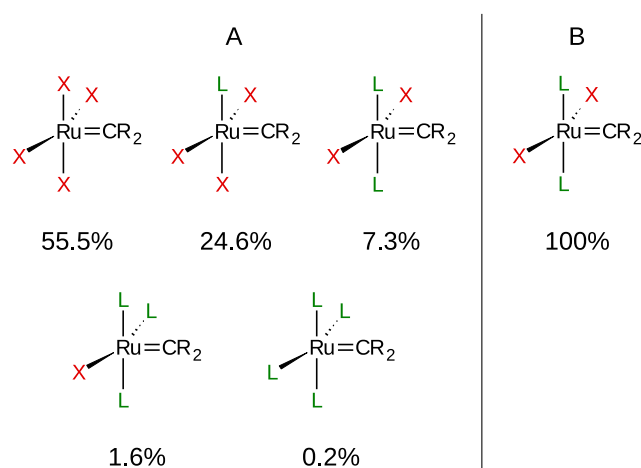


Figure 7: Formal classifications of metal-coordination environments generated by valence-only (A) or chemically controlled (B) connection rules. Each of the coordination environments depicted in A are meant to represent all possible stereoisomers with fixed $\text{Ru}=\text{CR}_2$. Reprinted with permission from ref. 136. Copyright 2014 American Chemical Society.

The evaluation of the actual feasibility of the complexes is not trivial. Systematic, automated, and empirical evaluation over a large dataset is not yet possible for the heterogeneous type of chemistry generated in these experiments. Visual inspection suggests that potential issues may derive from automatically generated ligands with too sterically hindered ligands, or displaying electronic effects hampering the

formation of dative ligands. Nevertheless, the design of the organometallic fragment space, i.e., filtration of fragments and management of cutting and connectivity rules, allows avoiding these issues according to the specific desires of the chemists.

Notably, some automatically generated ligand present more than one metal-coordinating moiety (B in Figure 6), which might introduce the issues of linkage isomerism, multidentate coordination, and metal-organic polymerization. While the deliberate formation of chelates is discussed in the dedicated chapter (see Chapter 7), potentially multidentate ligands can be formed accidentally from organic fragment that are obtained by fragmentation of purely organic molecules and hence do not bear any AP representing the possibility of coordinating metals. Design of a fragment space purged by potentially metal-coordinating groups is an easy workaround. Nevertheless, the reduced structural diversity resulting from such strategy may be too deleterious for some *de novo* design projects.

The issue of the potential multidentate ligands can be extrapolated in a more general context and interpreted as a problem that is due to the limited control over the content of the fragments. In fact, while the compatibility matrix regulates the type of chemical features generated by connection of fragments, the content of the fragments is not taken into account. On one hand this allows to handle any sort of chemical system, which is a requirement for the present method, but on the other hand it allows for incompatible or inter-reacting functional groups to appear in the same molecule. A typical example is the simultaneous presence of acid and basic groups on the same compounds. Although methods have been developed for the prediction of the protonation state of organic molecules and proteins,^{150–153} the general issue of incompatible moieties can be addressed first by proper design of the fragment space (i.e., filtering of fragments), and then by post-processing of the candidates, such as recalculation of the protonation state and rejection of candidates with incompatible moieties.

In case study II, the organometallic fragment space was designed as the combination of substituents harvested from crystallographic structures and commercially available

compounds. Therefore, in this case study the fragments are larger than in case study I; that is, fragments could be further fragmented. As stated in the introduction (Section 3.1.1), large building blocks reduce the resolution of the building process and simplify the problem of avoiding unrealistic candidates. Consequently, the molecules generated in case study II display chemical features that are very common among existing catalysts and are expected to be more synthetically accessible than those generated in case study I. Despite the reduced structural diversity with respect to the ligands built in case study I, and the high similarity of designed ligands with existing ones, most of the organometallic species generated have so far not been synthesized nor evaluated for catalytic activity.

4.2 Conclusion

The *de novo* design machinery demonstrated the capability of controlling the chemical traits of automatically generated transition metal compounds, combining retrosynthetic control for the organic connection and regulating the exchange of ligands by type (i.e., one- or two-electron donors) thus retaining the oxidation state and formal electron count of the metal center. Chemical knowledge and intuition were projected into the automated generation of molecules by means of a compatibility matrix controlling all possible connection, both for organic and for organometallic features. Although limited control could be exerted with respect to the content of the fragments assembled, the generation of molecules could be focussed on realistic candidates according to the intention of the designer.

5. Building of Accurate 3D Molecular Models

5.1 Introduction

While the computational estimation of the molecular properties of transition metal compounds can, to some extent, be performed from low dimensionality chemical representations, such as 2D or even line notations,^{154–156} accessing high accuracy molecular modeling methods requires the creation of 3D models. Therefore, the preparation of initial guess geometries from low dimensionality chemical representations has been a mandatory step since the early stages of molecular modeling, and a number of tools have been developed in response to this need.^{157–168} Unfortunately the peculiar geometrical features of many transition metal compounds have been shown to cause problems of accuracy and coverage.^{169,170}

The availability of robust methods for the generation of 3D models of transition metal compounds is crucial for *de novo* design methods that require (i) fully automated and efficient tools, (ii) capability of handling special chemical entities, and (iii) production of sufficiently accurate results. While lack of automation and capability of handling peculiar systems simply prevent the application of *de novo* design methods, low accuracy introduces more subtle effects. In particular, inconsistent treatment of some or all candidate compounds in a design project introduces biases in the evaluation of molecular properties. Moreover, inaccurate initial geometries increase the computational cost of further refinement steps and reduce the likeliness for automated molecular modeling to succeed and produce the intended system without encountering fatal errors or undesired rearrangement of the geometry.

It has been suggested that a cascade of increasing accuracy molecular modeling step (molecular mechanic, semiempirical, and quantum mechanic modeling) can provide access to accurate geometries.^{171–174} Nevertheless, this approach assumes that the empirical components, i.e., robust force fields and semiempirical parametrization, are well trained for the specific chemistry at hand, which is seldom the case for peculiar

transition metal species,¹⁷⁵ thus making re-parametrization a mandatory preliminary step.

Instead, retrieving geometrical information from available 3D models, i.e., crystallographic structures, has been suggested in methods for automated conversion of SMILES and 2D representation into 3D models.^{169,170} Nevertheless, use of low dimensionality chemical representation is often accompanied with systematic mistreatment of peculiar chemical species that are not compatible with the assumptions intrinsic in the low dimensionality representation (i.e., protonation status, bond type, and geometry). Thus, the method developed in this work (see Section 3.2) was designed to make use of 3D building blocks, bypass the limitations of low dimensionality chemical representations, and reduce to the minimum the need for force field parameters. For this reason, the method is referred as to the full-3D approach.

This chapter presents the application of the full-3D approach in the generation of 3D molecular models of challenging transition metal compounds with tree-like structure (i.e., all rings were entirely contained within fragments), and compares the performance with a series of approaches based on SMILES-to-3D conversion tools. Three case studies are discussed to highlight the capability of correctly handling unusual functionalities, geometries, and stereochemistry of reaction intermediates that are relevant in the design of transition metal and organometallic catalysts. In all cases, 3D models are also refined by molecular modeling methods (MM, semiempirical, and DFT) providing an evaluation of the effect that low accuracy initial geometries have on the production of refined models for property evaluation.

5.2 Results and Discussion

5.2.1 Case Study 1

A dataset of 82 ruthenium-carbene active species (**DS-1**) relevant for the design of ruthenium catalysts for olefin metathesis was retrieved from the literature.¹⁷⁶ The compounds have general formula $(L)(Cl)_2Ru=CH_2$, where L is one of various dative

ligands including phosphines, *N*- and *P*-heterocyclic carbenes, and aromatic heterocyclic compounds. A central issue associated with these compounds is the peculiar geometry of the metal center. In fact, although the metal has four ligands (coordination number = 4), the geometry is neither tetrahedral nor square planar. Instead, computational and experimental evidence show a geometry rather close to the disphenoidal case (*OC-4* in Figure 8).^{176–179} Moreover, six stereoisomers can be drawn with the same coordination geometry and the given formula (Figure 9). However, in this work the aim was to generate only one such stereoisomers (A in Figure 9).

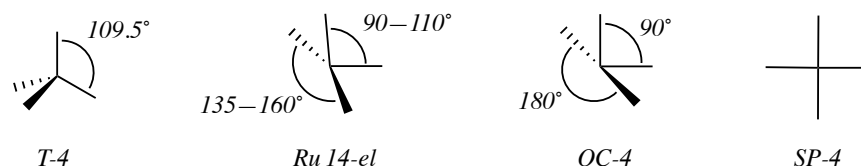


Figure 8: Comparison of the tetrahedral (*T-4*), disphenoidal (*OC-4*), and square planar (*SP-4*) coordination geometries for tetra-coordinate atoms, with the geometry of 14 electrons $(L)(Cl)_2Ru=CH_2$ centers in ref. 176 (*Ru 14-el*).

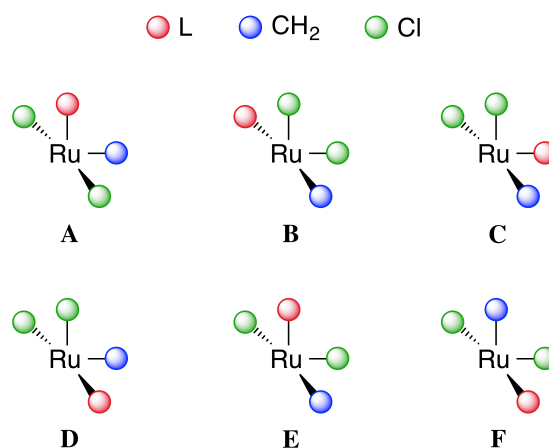


Figure 9: The six stereoisomers of a disphenoidal center with general formula $(L)(Cl)_2Ru=CH_2$. Reprinted with permission from ref. 136. Copyright 2014 American Chemical Society.

The performance of the full-3D method, which generated 3D models by assembling 3D fragments mostly taken from the Cambridge Structural Database, was compared

with those of combinations of Open Babel,¹⁸⁰ COSMOS,^{169,170} and Marvin¹⁸¹ tools, which were used to generate isomeric SMILES from 3D structures and then convert the SMILES back to 3D models (Table 1). In addition, since most of these SMILES-based protocols generated hydrogen-depleted 3D models, a “protonation” step was performed to attempt recovering the proper set of hydrogen atoms. Unfortunately, it is not to be expected that the protonation step can handle all the chemical features of organometallic compounds. In fact, Table 1 shows that none of the approach requiring the protonation step was consistent in managing the constitutional information provided as input.

Table 1: Definition of protocols used in Case Study 1 for generation of 3D models. Reprinted with permission from ref. 136. Copyright 2014 American Chemical Society.

ID	Generation of 3D from SMILES				3D models produced
	3D-to-SMILES	SMILES-to-3D	Protonation	Wrong number of H	
1	Marvin	Marvin	none ^a	0	82
2	Marvin	COSMOS	Marvin	8	74
3	Marvin	COSMOS	Open Babel	11	71
4	Open Babel	Marvin	Marvin	2	80
5	Open Babel	Marvin	Open Babel	15	67
6	Open Babel	COSMOS	Marvin	10	70 ^b
7	Open Babel	COSMOS	Open Babel	13	67 ^b
8	Open Babel ^c	Marvin	Marvin	2	80
9	Open Babel ^c	Marvin	Open Babel	15	67
10	Open Babel ^c	COSMOS	Marvin	10	70 ^b
11	Open Babel ^c	COSMOS	Open Babel	13	67 ^b
<i>Generation of 3D from 3D fragments</i>					
12	Full-3D ^d			0	82

^a Protonation step not needed.

^b For two molecules COSMOS did not recognize the string generated by Open Babel as a correct SMILES string.

^c Canonical smiles.

^d This work; see Section 3.2.5 for details.

Two geometrical descriptors of the geometry were deployed to evaluate the accuracy of the generated 3D structures with respect to the reference DFT structures reported

in the original paper:¹⁷⁶ the shape similarity index (T_S), and the mean angle difference for the metal center (MAD).¹⁶⁹ While the first takes into account the overall molecular shape (the higher the T_S , the better the quality of the overall geometry), the second focuses on the coordination geometry of the Ru atom and quantifies the differences of the set of angles formed by the ligands around the metal between the generated 3D models and the reference structures (the lower the MAD , the better the geometry of the metal).¹⁷⁶

Analysis of the distribution of shape similarity index for the twelve protocols indicates that in all cases, including the full-3D approach, conformations are significantly different from those of the reference structures. Nevertheless, the full-3D approach, performs appreciably better than all other protocols.

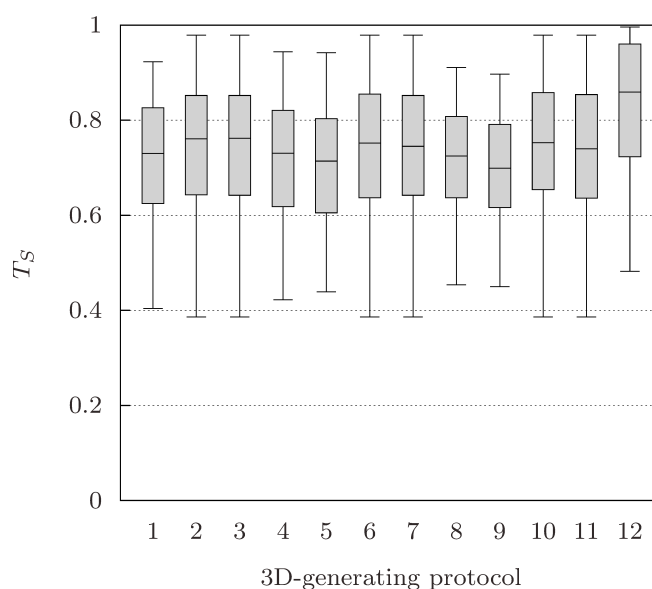


Figure 10: Boxplots (25th, 50th, 75th percentile, and whiskers up to 1.5 interquartile range) representing the distribution of molecular shape similarity index (T_S) of the 3D models of dataset **DS-1**. The 3D-generating processes are defined in Table 1. Reprinted with permission from ref. 136. Copyright 2014 American Chemical Society.

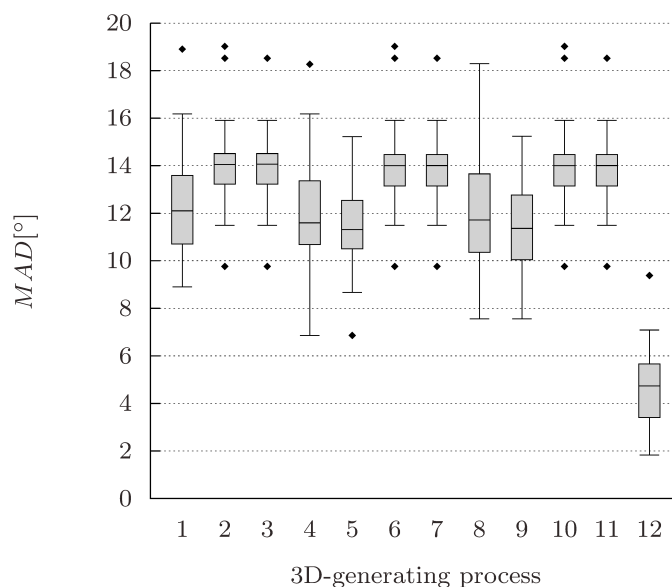


Figure 11: Boxplots (25th, 50th, 75th percentile, and whiskers up to 1.5 interquartile range) representing the distribution of mean angle differences (*MADs*) of the 3D models of dataset **DS-1**. The 3D-generating processes are defined in Table 1. Reprinted with permission from ref. 136. Copyright 2014 American Chemical Society.

The better performance of the full-3D method is partially explained by the improved geometry of the metal center. In fact, evaluation of the mean angle deviation (Figure 11) reveals the outstanding accuracy of the metal coordination geometries generated by the full-3D in comparison of all other tools. This result is clearly due to the use of a proper 3D building block for the metal center also combined with machinery capable of preserving such geometrical information. On the contrary, the coordination geometries produced by protocols 1-11 always tend to be tetrahedral in accordance to the misleading coordination number. To appreciate the consequence of the improved accuracy produced by the full-3D approach, the 3D models generated by all protocols were used as input for geometry optimization using (i) molecular mechanic method with the universal force field (label “UFF”),¹⁴² (ii) semiempirical method with parameters set PM6 (label “PM6”),¹⁸² (iii) and density functional theory with the OLYP functional^{183,184} and double- ζ basis set (i.e., LANL2DZ)^{185,186} either with (label “OLYP_A”) or without (label “OLYP_B”) preliminary geometry optimization with minimum basis set. The analysis of the subset of compounds that was properly

modelled in all cases by all three refinement methods (35 compounds, **DS-2** Figure 12), reveals that the average shape similarity index of the initial 3D models produced by the full-3D approach is higher than that of both UFF and PM6 refined models from all protocols.

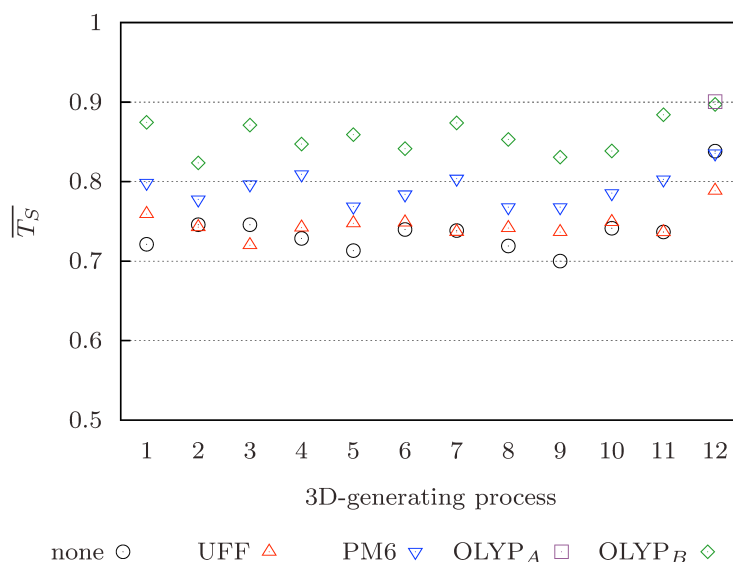


Figure 12: Average shape similarity index (\overline{T}_s) calculated for 3D models of **DS-2** before (none) and after refinement (UFF, PM6, OLYP_A, OLYP_B). Reprinted with permission from ref. 136. Copyright 2014 American Chemical Society.

The same applied for the overall mean angle deviation (Figure 13). Thus, the full-3D approach is capable of producing initial geometries that have, on average, better shape and coordination geometry than those achieved by UFF and PM6 refinements. Instead, the more demanding DFT refinement manages to improve dramatically both shape and coordination geometry in all cases, though, the best performance is obtained from input structures generated by the full-3D approach.

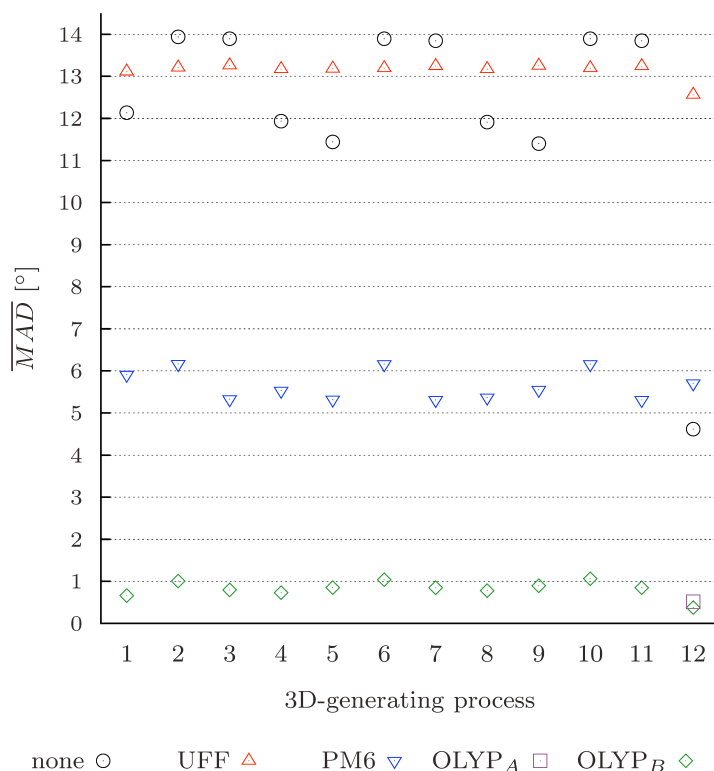


Figure 13: Overall mean angle difference (\overline{MAD}) calculated for 3D models of **DS-2** before (none) and after refinement (UFF, PM6, OLYP_A, and OLYP_B). Reprinted with permission from ref. 136. Copyright 2014 American Chemical Society.

Finally, the stereochemistry of the refined 3D models was also evaluated (Figure 14). While the full-3D approach generated models with the proper stereochemistry in the vast majority of cases, all other protocols returned significant amounts of structures with wrong stereochemistry. For protocols 1-11 the distorted geometry of the coordination sphere in the initial 3D models is the main cause of the wrong stereochemistry in the refined models. In fact, protocols 1-11 generated models with coordination geometries close to tetrahedral. Due to the presence of two equal ligands in (L)(Cl)₂Ru=CH₂, only one isomer can exist with tetrahedral geometry. Moreover, starting from such tetrahedral geometry any of the six stereoisomers can be reached by proper distortion of the angles. As highlighted by the different profiles obtained for UFF, PM6 and DFT refinements (Figure 14), accurate description of the potential allowed better recovery of the proper stereochemistry despite the distorted input.

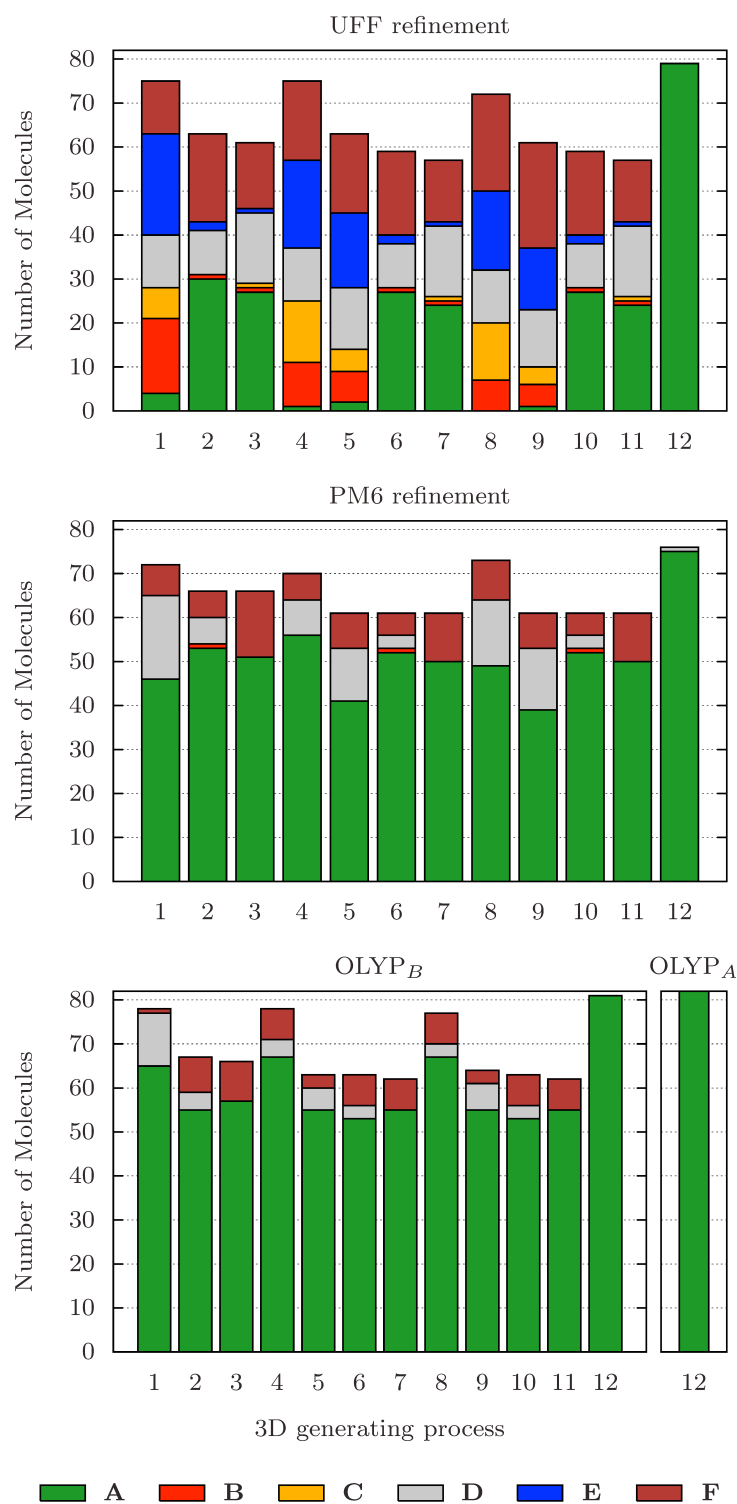


Figure 14: Distribution of stereoisomers for UFF-, PM6- and OLYP-refined 3D models of dataset **DS-1**. Reprinted with permission from ref. 136. Copyright 2014 American Chemical Society.

Nevertheless, only the combination of full-3D approach and OLYP_B refinement protocol led to the complete set of proper stereoisomers. In fact, the pre-optimization with minimal basis set, though capable of improving significantly the geometry and facilitating the double- ζ calculation, which was otherwise unfeasible for protocols 1-11, has the capability of introducing errors such as the change of connectivity occurred during the preliminary optimization of one model obtained from the full-3D and refined by OLYP_A. Surprisingly, the same initial geometry refined by OLYP_B produced the desired result (Figure 14), thus highlighting the mixed role of the pre-optimization.

5.2.2 Case Study 2

To evaluate the generality of the capabilities demonstrated in the previous case study, a structurally diverse dataset (**DS-3**) was created collecting 58 reaction intermediates from organometallic catalysed reactions including different metals, ligand types, geometries, oxidation states, and peculiar hydrogen atoms (Figure 15). The performance of full-3D approach was evaluated against the only other protocol that demonstrated the capability of handling hydrogen atoms consistently (protocol 1 in Table 1, page 58).

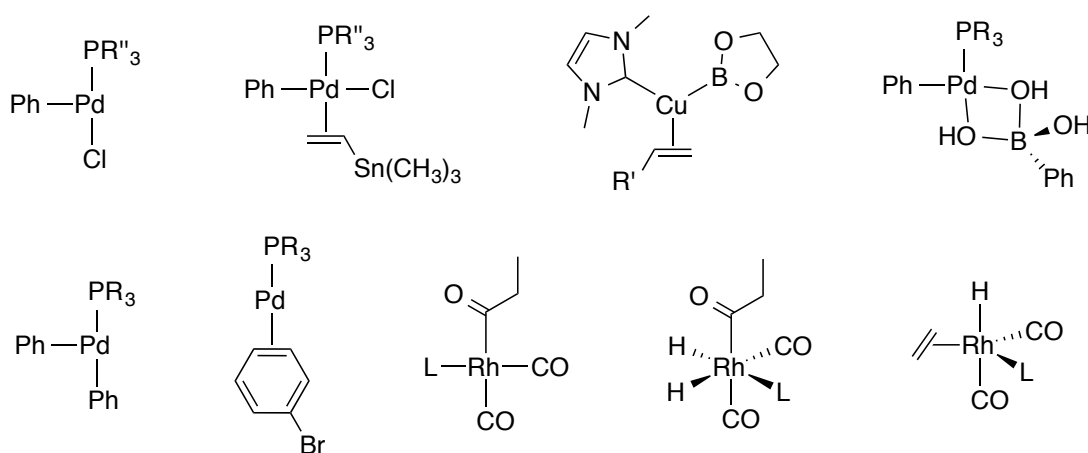


Figure 15: Schematic representation of the compounds included in **DS-3**. L: phosphines, phosphites esters, or *N*-heterocyclic carbenes; R: $-\text{CF}_3$, $-\text{Ph}$, or $-\text{tert}$ -butyl; R': $-\text{CN}$ or $-\text{Ph}$; R'': $-\text{Me}$, $-\text{Ph}$, or $-\text{tert}$ -butyl. Reprinted with permission from ref. 136. Copyright 2014 American Chemical Society.

The results confirmed the general conclusion from the previous case study; that is, both the overall shape (Figure 16) and the mean angle difference (Figure 17) of models generated by the full-3D approach are better than those from the SMILES-based approach. Moreover, also the poor stereochemical control of protocol 1 was confirmed with only 26% of proper stereoisomers produced for **DS-3**. Instead, the full-3D approach returned the proper result for all but one compound. The exception being a single compound where, despite the good coordination geometry and an apparently harmless conformational difference, the DFT refinement led to the disconnection of a phosphite ligand from the metal center. This example highlighted how the combination of conformational issues and approximate description of bonding interactions (i.e., OLYP functional does not account for dispersion and underestimates the metal-phosphite binding energy)¹⁸⁷ may lead to failure of the refinement process.

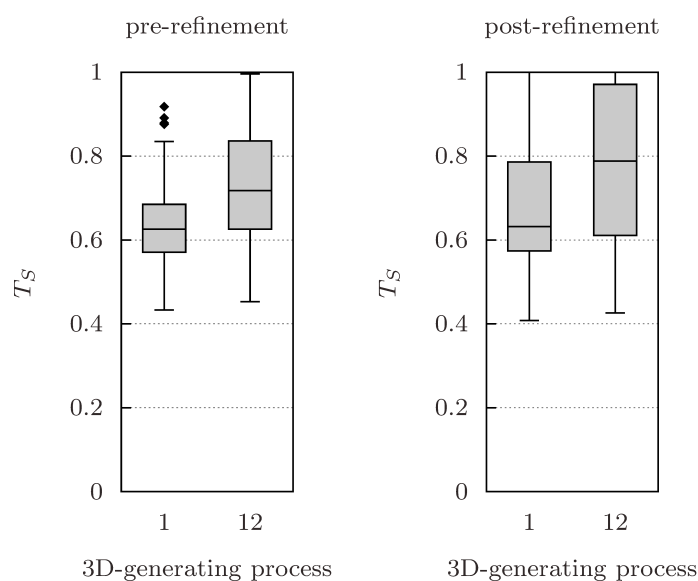


Figure 16: Distribution of the molecular shape similarity index (T_S) for 3D models of dataset **DS-3** before (left) and after (right) DFT-based refinement of the geometry. Reprinted with permission from ref. 136. Copyright 2014 American Chemical Society.

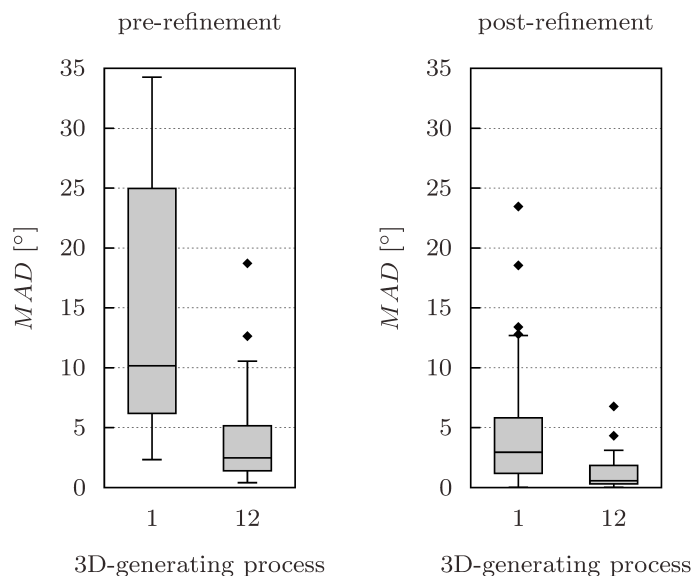


Figure 17: Distribution of the mean angle differences (*MADs*) for 3D models of dataset **DS-3** before (left) and after (right) DFT-based refinement of the geometry. Reprinted with permission from ref. 136. Copyright 2014 American Chemical Society.

The wide scope of the properties of ligands represented in the dataset highlighted one of the limitations of the full-3D approach. That is, the full-3D approach exploits the possibility of using the same 3D building block for many analogous of a given compound. With ligand characterized by very different electronic and steric properties the geometry of the metal may respond by displaying substantial geometrical changes, as shown here for the tetracoordinated Rh intermediates (Figure 15) that span from square planar to nearly disphenoidal geometry.¹⁸⁸ While the DFT refinement has properly taken care of this issue, the possibility for this to become a problem should be considered carefully in a *de novo* design application. An ideal solution would be to use, for the same fragment, different 3D geometries that are chosen according to the properties of the neighbour fragments. The use of multiple geometries depending on the chemical context is already feasible the presented method, thanks to the classification of the attachment points (see Section 3.2.2), but this examples show that care must be taken in allowing AP classes cross compatibilities that involve dramatic changes of steric and electronic properties.

5.2.3 Case Study 3

The last case study aimed to demonstrate the capability of the full-3D approach with respect to handling of chemical entities that cannot be properly represented in standard chemical representations. In particular, a dataset (**DS-4**) was made out of 22 contact ion pairs representing candidate active species in titanium-catalysed olefin polymerization (Figure 18).^{189–193} The challenge introduced by these compounds is the proper handling of the relative position of the two ions. In this context, the combination of graph representation and 3D building blocks allowed to define the ionic interaction as a connection between two fragments and thus to treat the relative position of the two ions as any other couple of fragments.

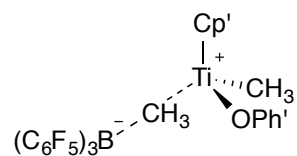


Figure 18: Schematic representation of the compounds included in **DS-4**. Reprinted with permission from ref. 136. Copyright 2014 American Chemical Society.

Once again, the shape similarity index and the mean angle difference were evaluated throughout the dataset against the set of reference structures both before and after DFT refinement. Although all 3D models presented the proper stereochemistry, the distribution of T_S and MAD (Figure 19), displayed values higher than expected. Deeper structural analysis identified in the conformational differences the rationale for the distribution of both T_S and MAD . In particular, the mean angle difference is altered by different conformations along the metal–pentahapto bond, i.e., the single formal bond between the metal and the centroid of the multihapto ligand. Instead, propeller isomerism of the $[B(C_6F_5)_3CH_3]^-$ anion and flipping (i.e., rotation of nearly 180°) of the aryloxy ligands contribute substantially to the value of T_S . Therefore, also in this case study, although the full-3D approach has demonstrated the good performance, the conformational problem stems as the major source of deviation from the reference structures.

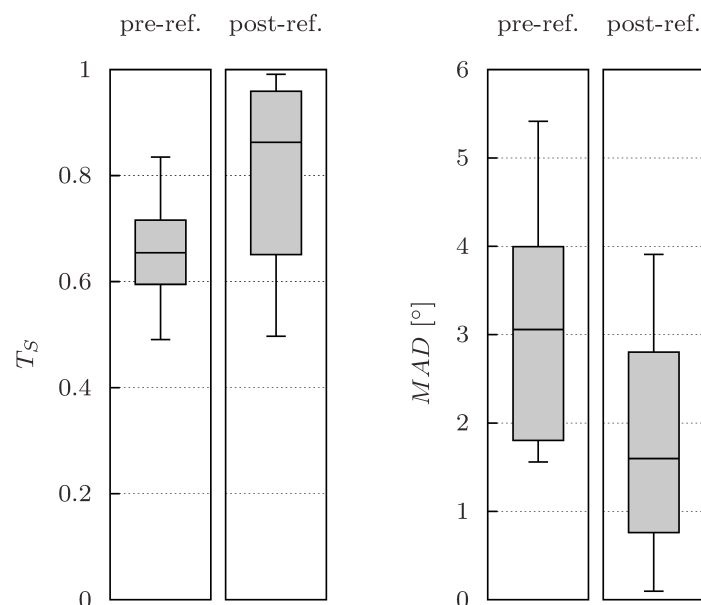


Figure 19: Distribution of molecular shape similarity index (T_S) and mean angle differences ($MADs$) before (“pre-ref.”) and after (“post-ref.”) DFT-based refinement of the geometry obtained using the full-3D method on dataset **DS-4**. Reprinted with permission from ref. 136. Copyright 2014 American Chemical Society.

5.3 Conclusion

Three case studies evaluated the capabilities of the automated construction of molecular models from 3D building locks. The method has been proven able to bypass the emerged limitation of tools converting low dimensionality chemical representation, such as SMILES, into 3D. In particular, superior results were obtained with respect to the accuracy of the metal coordination geometry and control of the stereochemistry of the metal center. Moreover, general applicability was demonstrated for chemical entities far beyond the capabilities of commonly deployed automated methods.

6. Fast Fitness from Ligand Field Molecular Mechanics

6.1 Introduction

Application of *de novo* design depends on the availability of a method for the automated evaluation of the fitness of candidates generated from scratch. Although the nature of the fitness function depends on the specific case study, the calculation is usually based on one or more numerical descriptors derived from evaluation of constitutional and topological information or 3D molecular models.^{175,194–198} In the latter case, molecular modeling is likely to be the most computationally demanding step in the path from the definition of a candidate compound to the knowledge of its fitness. While high accurate quantum mechanics, and in particular density functional theory, are widely accepted as the methods of choice for calculation of structures, electronic properties and energy profiles involving transition metal compounds, the computational cost is significantly high and, often, prohibitive for large scale design. Therefore, whenever possible fast fitness based on empirical or semiempirical molecular modeling tools is desirable.

Molecular mechanics, if properly trained and validated, can represent a good or even better compromise between accuracy and efficiency than DFT.^{31,199–201} In particular, while relative energies may not be accurate,^{200,202} empirical force fields are particularly useful when conformational search is required.^{201,203} Nevertheless, the varied chemistry of transition metals involves a wider range of coordination numbers and geometries than organic chemistry. In addition, the partially filled d orbitals open for multiple oxidation and spin states and determines geometrical effects like Jahn-Teller's and trans-influence. Since molecular mechanics does not treat electrons explicitly, all electronic effects represent a challenge. Although standard force field formalisms (see Section 2.2.2) can be tuned and successfully applied to reproduce a particular target geometry of a given type of metal–ligand set, with specific oxidation and spin states,^{204–210} this approach requires dedicated parameter sets which are not general, i.e., not transferable to other systems. Other approaches, instead, attempt to

develop additional potential energy terms specifically accounting for the peculiarities of the metal center, such as the description of the angles centred on the metal atom. Examples include the use of the “points on a sphere” model,³¹ and the implementations of valence bond theory in SHAPES²¹¹ and VALBOND.^{212–215} Among method aiming to improve the empirical definition of the potential energy terms for transition metal compounds, this work focuses on the ligand field molecular mechanics (LFMM)^{216–218} method, which provides explicit definition of spin and oxidation states without assumption on coordination geometry and number of ligands, all with the same set of parameters.

6.2 Ligand Field Molecular Mechanics (LFMM)

The LFMM model combines standard force fields, for the organic portion of the system, and a metal-dedicated empirical term based on ligand field theory (eq. 58).^{217,218} Therefore, the LFMM method can be described as a hybrid approach combining two empirical components.

$$E_{tot} = \sum_{not\ ML} E_{str} + \sum_{not\ LML} E_{bend} + \sum E_{tor} + \sum E_{nb} + \sum_{metals} E_{LFMM} \quad (58)$$

The organic portion of the system is described by standard force fields, such as Amber,²¹⁹ CHARMM,^{220,221} MMFF,²²² which are possibly modified to improve the performance with respect to ligands of specific transition metal compounds.²²³ The LFMM term accounts for the ligand field stabilization energy (LFSE), replaces bond stretching and bending terms that involve the metal atom with dedicated metal–ligand stretching (E_{ML}) and 1-3 ligand-ligand interaction term (E_{LL}), and accounts for an empirical estimation of the electron pairing energy (E_{pair} , eq. 59).

$$E_{LFMM} = E_{ML} + E_{LL} + E_{pair} + LFSE \quad (59)$$

The four LFMM terms consider only the metal and the first two layers of neighbour atoms (the second layer serves only to define the orientation of the ligands). In this metal-ligand core, the functional forms used to define the potential energy are

independent from those of the coupled organic force field. In particular, while metal–ligand stretching terms (E_{ML}) and 1-3 ligand-ligand interaction terms (E_{LL}) are based on standard potential formulations, such as harmonic, Lennard-Jones, and Morse potentials (see Paper III), the LFSE is obtained from the angular overlap model (AOM).²²⁴ The AOM model considers the overall result of local M–L contributions, each dependent on the overlap of the d orbitals of the metal with the ligand orbitals with proper symmetry (σ , π_x and π_y) with a strength of the interaction modulated by means of parameters proper of the M–L pair.²¹⁶ The electron pairing energy is estimated per each M–L pair according to the given spin state and empirical parameters. This local treatment of each M–L pair allows the model to handle any coordination environment without assumption on the coordination geometry, spin state, and the number and type of ligands.

6.3 Integration of LFMM in Tinker

The calculation of LFSE and electron pairing energy requires functionalities that are not implemented in molecular modeling tools. As a consequence, software supporting LFMM calculations is rare. The main implementation of the LFMM method, i.e., DommiMOE,²²⁵ is bound to the commercial software MOE (Molecular Operating Environment),²²⁶ but MOE's vendor provides no support or documentation on the LFMM implementation and, while powerful graphical user interface and numerous functionalities are available, further development of LFMM is affected by closed-source policy and bound to the programming language embedded in MOE. The original, pre-DommiMOE implementation of LFMM has been abandoned due to the lack of robust support for the organic force field,²¹⁶ and other implementations either depend on DommiMOE,²²⁷ or have not yet been released.^{31,228–231} Therefore, to increase the accessibility to the LFMM method and support its systematic use in fast fitness, the LFMM method was integrated into the popular and easily accessible Tinker package.²³²

The modular structure of Tinker enabled to follow an integration strategy based on implementing (i) an interface to the original Fortran code calculating the LFSE contribution to energy and gradient for a single metal center, (ii) the routines for the calculation of electron pairing energy, metal–ligand stretching, and ligand–ligand interaction, and (iii) support routines for handling of LFMM parameters. In addition, to ensure backwards compatibility of the parameter sets and allow further improvement of the force fields, the capabilities of the functional forms originally implemented in Tinker have been expanded. In particular, polynomial functional form with independent coefficients was implemented for bond stretching and angle bending energy terms, and the MMFF-style²²² torsional potential was implemented up to the 6-fold term.

The new LFMM-capable Tinker implementation is going to be distributed in the coming release of Tinker. Nevertheless, the new Tinker-LFMM implementation has been exploited in the *de novo* design project described in the next chapter.

7. Multidentate Ligands for Fe(II) Spin-Crossover Compounds: a Test Case for Ring-Closing Design

7.1 Introduction

Spin crossover (SCO) is the change in spin state exhibited by some transition metal complex as a consequence of the application of an external perturbation, such as temperature, pressure, light, and magnetic field.²³³ This spin transition is accompanied by change of magnetic properties and colour, which makes SCO compounds attractive for technological application such as display, sensing, and memory devices.^{234–237} The spin state preference can be explained with the ligand field theory as resulting from the competition between ligand field stabilization energy and electron spin-pairing energy; that is, weak ligand field cannot balance the spin-pairing energy and implies high spin (HS) ground state, but strong ligand fields justify low spin (LS) ground states. In thermal SCO the greater electronic and vibrational entropy of the HS state overcomes the higher enthalpy of the LS as the temperature increases.^{238,239} Thus, SCO compounds have HS ground state at high temperature and LS ground state at low temperatures. Octahedral complexes of 3d elements with d^5 , d^6 , and d^7 configuration display particularly strong tendency to undergo thermal spin transition. In fact, most known SCO compounds are complexes of Fe(II), Fe(III), and Co(II), but also Mn(II), Cr(II), and Co(III) have been reported.²⁴⁰ The most studied SCO compounds are combinations of Fe(II) with nitrogen-based ligands. The spin transition in these compounds is associated with the largest metal–ligand bond change that, in the solid state, can induce strong cooperatively between the metal centers and might lead to hysteresis, which is a tailored property for technological applications.^{241–243} Therefore, many Fe(II)- N_6 compounds have been synthesized and evaluated for SCO behaviour.²⁴⁴

While understanding and rationalization of the overall properties of SCO materials for technological application requires handling of a challenging crystal-engineering problem,²⁴⁵ design of new compounds capable of displaying SCO behaviour focus on

the ligand set surrounding the metal atom.²⁴⁶ To this end, Deeth and co-workers have developed an empirical LFMM force field that enables access to both HS and LS states for the same compound with the same set of empiric parameters, thus providing computationally inexpensive evaluation of the spin state energy difference.^{223,247} Although the force field covers only Fe(II)-amine compounds, the ligand field generated by six amine groups can, to some extent, be tuned by modification of the molecular frame supporting the metal-coordinating moieties. In general, the design of complexes with multidentate ligands should allow for variable denticity, different bridge types and lengths with mutable identity of the bridged atoms, and possibility of decorating both bridge and coordinating moieties. With the exception of the modification of decorating groups, all such structural alterations imply modification of both organic and metal-organic rings. Therefore, the design of novel Fe(II) SCO compounds represents a test field for the ring-closing machinery developed in the context of 3D fragment-based design (Chapter 3). This application takes advantage of the evolutionary algorithms for *de novo* optimization of transition metal compounds described in ref. 30 (see also Section 2.1).

7.2 Brief Computational Details

Artificial evolution experiments were performed to evolve new Fe(II) SCO compounds under the pressure of an empirical fitness function based on spin state energy differences calculated with the Tinker implementation of LFMM (see Chapter 6).²⁴⁸ In addition to the energy gap associated with the relaxed LS→HS transition, which is calculated between the two global minima of the HS and LS state potential energy surfaces (i.e., relaxed transition),²⁴⁹ the fitness consider also the two vertical transitions LS→HS and HS→LS at the fixed, lowest energy geometry of the LS and HS state respectively. These vertical transitions served as a rough indication of possible high lying minimum energy crossing points. Overall, the fitness function was designed empirically to match the reasonable expectation that best candidates should have LS ground state at 0 K with a relaxed spin transition within 5-6 kcal/mol

(i.e., the fitness favours the LS with a bias of 3 kcal/mol on the relaxed spin transition energy) and small contributions from the vertical spin transition.^{250,251}

New candidate Fe(II) SCO compounds were generated by assembling 3D fragments collected from crystallographic structures and computed models. In particular, as the aim is to identify compounds with LS ground state, 3D fragments were generated from purposely-modelled low spin geometries in order to obtain proper metal–ligand bond lengths and attachment point vector lengths. Other fragments were generated specifically from metal-organic five and six member rings thus providing the proper bond angles for formation of rings in the new structures.

7.3 Results and Discussion

Two different types of evolutionary experiments were performed: in one case, new compounds were built starting from root fragments containing only a single Fe(II) atom (strategy **A**), while in the other case large root fragments were used that contained the Fe(II) atom and portions of pre-coordinated ligands (strategy **B**). On one hand **A** could explore structures that were not accessible by **B**, which, instead, could exploit previously known portions of multidentate ligands. On the other hand, the complexity of the combinatorial problem, i.e., the generation of valid graphs, and that of the conformational problem, which is due to the simultaneous closure of a larger number of rings, made strategy **A** more challenging and computationally demanding than **B**. In particular, to produce the same amount of valid structures, experiments performed with strategy **A** required the evaluation of about four times the number of graphs used by those run with strategy **B**. Regardless, the generation of valid graphs in both cases has shown inefficiencies mainly due to the random generation of graphs and the elementary implementation of the method with respect of detection of duplicates and evaluation of ring-closing conditions (see Section 3.2.3). Artificial evolution experiments performed with the two strategies have evolved randomly generated populations of SCO candidates and identified new multidentate ligands with improved fitness (Figure 20 and Figure 21).

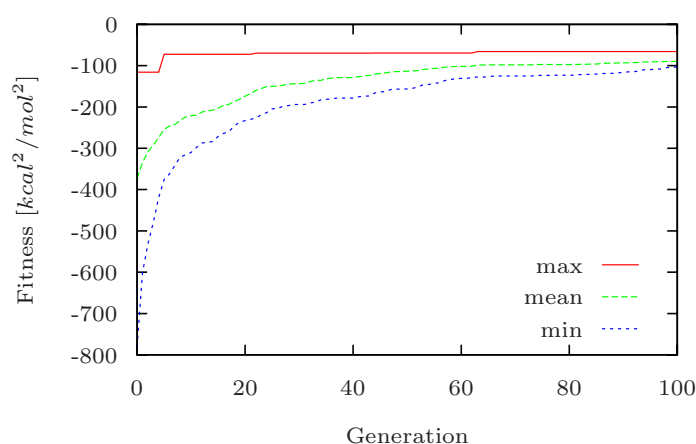


Figure 20: Maximum, mean, and minimum of the fitness in the instantaneous population through an experiment of type **A**.

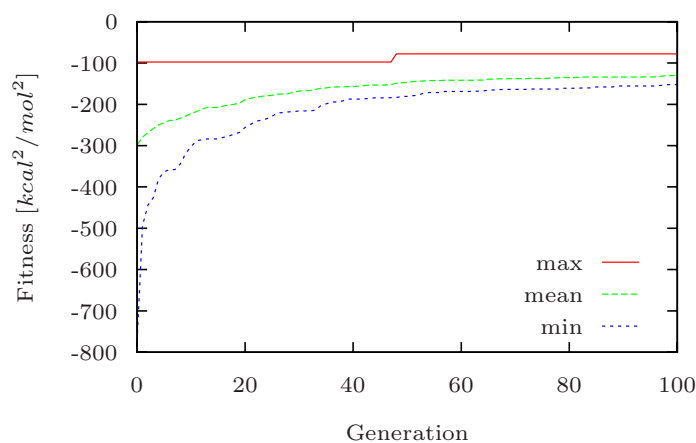


Figure 21: Maximum, mean, and minimum of the fitness in the instantaneous population through an experiment of type **B**.

The possibility of altering the molecular cyclicity resulted in a competition of different ligand denticity patterns. In particular, while four denticity patterns were permitted, namely three bidentate ligands ($\kappa^2, \kappa^2, \kappa^2$), two tridentate (κ^3, κ^3), one hexadentate ligand (κ^6), and combination of one tetradentate and one bidentate (κ^4, κ^2), experiments of both type **A** and type **B**, were dominated by pairs of tridentate ligand sets (Figure 22 and Figure 23). However, analysis of the occurrence of hexadentate ligands indicates that the competition between different denticities is biased by the ease of generating candidate with a given denticity set. The occurrence

of hexadentate ligands is low in experiments of type **A** (Figure 22), where construction of such compounds requires the simultaneous closure of six metal-organic rings, while in the experiment of type **B**, where only three ring closures are required to generate an hexadentate ligand, these candidates represent more than 30% of the population during the whole artificial evolution (Figure 23).

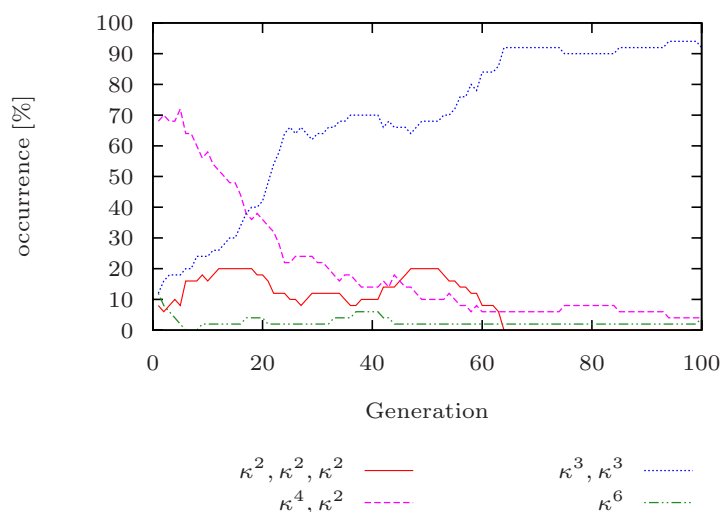


Figure 22: Characterization of the ligand denticity of the population throughout the experiment of type **A**. Each κ^n represents a single n-dentate ligand.

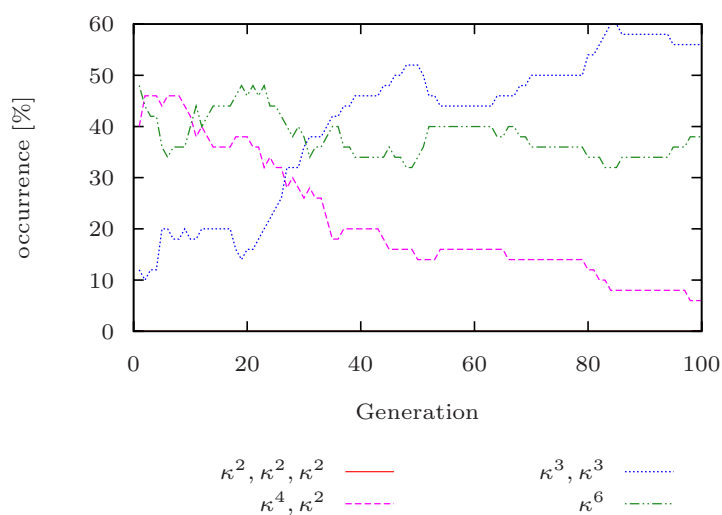


Figure 23: Characterization of the ligand denticity of the population throughout the experiment of type **B**. Each κ^n represents a single n-dentate ligand.

The fittest SCO candidates (Figure 24) are characterized by ligands that are new to the field of Fe(II) SCO compounds, but are (or resemble) molecules that have been synthesized before (Figure 25). The empirical evaluation of the relaxed spin transition agrees with that at DFT level indicating values that are within the range for potential SCO behaviour (Table 2). Instead the vertical transitions, which are not considered in the parametrization of the force field,²²³ are characterized by larger deviations.

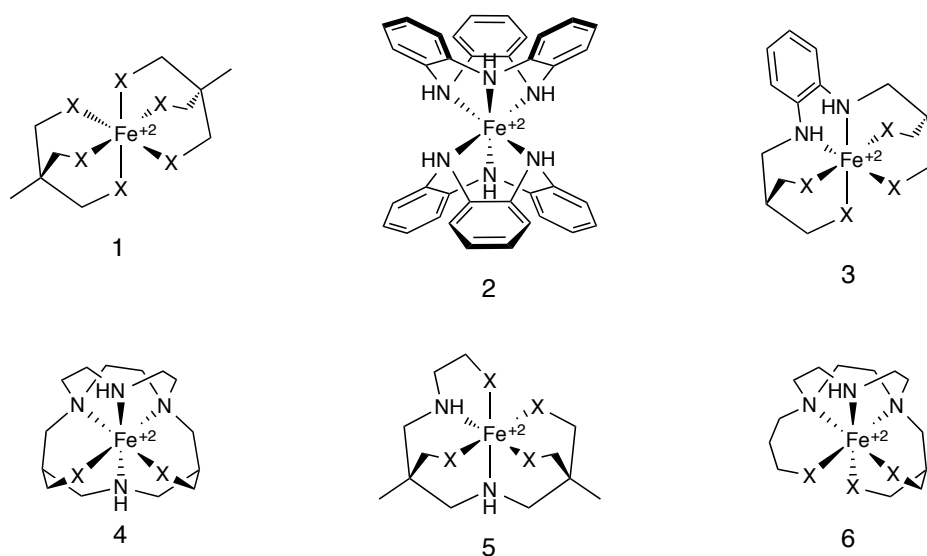


Figure 24: Sketches of representative molecules with high fitness discussed in the text (X = NH₂).

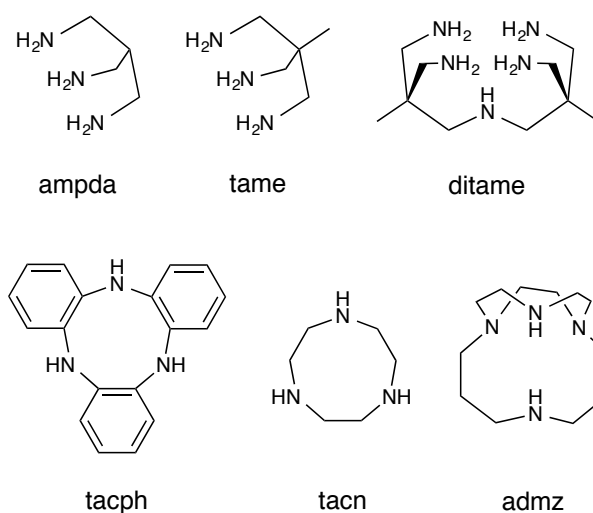


Figure 25: Sketches of multidentate skeletons characterizing the fittest candidates and mentioned in the text.

Table 2: Energies of the relaxed (ΔE_{rH-rL}) and vertical (ΔE_{vHL} and ΔE_{vLH}) spin transition energy (kcal/mol) for selected candidates and existing molecules.

Mol. ID	ΔE_{rH-rL}		ΔE_{vLH}		ΔE_{vHL}	
	LFMM	DFT ^(e)	LFMM	DFT ^(e)	LFMM	DFT ^(e)
1 ^(a)	0.42	3.05	29.06	32.68	20.15	18.60
2 ^(a)	0.84	3.36	29.38	33.51	20.86	30.35
3 ^(a)	1.30	4.51	30.07	34.30	21.62	20.47
4 ^(b)	4.93	5.47	33.29	32.13	19.35	18.78
5 ^(a)	0.94	4.55	31.46	33.21	24.38	20.74
6 ^(b)	-1.96	2.31	27.93	30.72	26.06	19.32
[Fe(tacn) ₂] ²⁺ ^(c)	1.40	5.29	42.14	35.88	25.22	18.63
[Fe(NH ₃) ₆] ²⁺ ^(d)	-6.27	-7.36	31.99	22.82	28.56	27.45
Mean Absolute Deviation		2.72		4.14		3.85

^a New candidate generated by an experiment of type **A**.

^b New candidate generated by an experiment of type **B**.

^c Previously existing compound; undergoes thermal SCO in solution.^{252,253}

^d Previously existing compound; HS ground state, no SCO behavior.

^e OPBE/cc-pVTZ PCM(water).

Overall, three frequently occurring tridentate skeletons characterize the populations of fittest molecules: ampda, tacn, and tacph (see Figure 25). The best SCO candidate identified by the artificial evolution experiments is a simple complex characterised by the tridentate ampda skeleton (**1**). Despite the fact that the ligand displayed in **1** is a commercial product that has been used to prepare complexes with other metals,^{254–256} and is often exploited as building block to prepare hexadentate ligands for Fe(II) SCO candidates,²⁵⁷ complex **1** has so far not been synthesized. DFT evaluations of relaxed and vertical transitions are in good agreement with the LFMM calculation, thus confirming the promising value of this candidate. In experiments of both type **A** and type **B**, the ampda skeleton has a dominating role and appears in five of the six fittest molecules (Figure 24). Compounds **4** and **6** present the combination of ampda with another of the preferred skeletons, i.e., tacn, which is found in many molecules generated by experiment **B**, and characterized an existing compounds known to undergo spin crossover in solutions.^{252,253} Instead, both types of experiments have

identified tacph (see Figure 25) as one of the skeletons leading to the highest fitness values. In fact, the LFMM-based fitness indicated **2** as the second best candidate overall. However, the crown-like conformation required to coordinate the metal is 12.9 kcal/mol (OPBE/cc-pVTZ PCM water, detail in Paper IV) higher than that of the crystallized *N,N'*-dimethylated analogous.²⁵⁸ The latter conformation does not allow tridentate N-coordination of the metal, thus, although the evolved populations of both **A** and **B** present several high fitness candidates with such skeleton differently decorated, this ligands lacks preorganization.²⁵⁹⁻²⁶¹ Moreover, LFMM overestimates the fitness of **2** by significantly underestimating one of the vertical transitions (Table 2). As a result, the actual value of the tacph skeleton seems not as promising as the LFMM evaluation would indicate.

Candidate **4** presents a tetradentate bicyclic core, i.e., admz in Figure 25, that has been used before to prepare complexes of first-row transition metals.²⁶² The presence of the two additional metal-coordinating primary amines imposes a more strained conformation that might compromise the stability of the hexadentate complex. Instead, compound **6** represents the monocyclic analogous that results from removal on a methylene bridge of **4**. This structural modification heals the strain issues of **4**, but the LFMM fitness of **6** is significantly lower than that of **4** due to the predicted HS ground state. On the contrary, the calculations at DFT level (Table 2) indicate LS ground state and, overall, suggest **6** as one of the most promising candidates. Further, investigation is thus required to clarify the actual value of this candidate.

The highlighted strain and lack of preorganization in the generated ligands represent issues that could be addressed by (i) improving the evaluation of the atom path closability conditions (see Section 3.2.3), for instance by reducing the allowance deployed to consider a chain closable, (ii) including an evaluation of the preorganization of the ligand, for instance by comparison the energy of free and metal-coordinating conformations,²⁶⁰ and (iii) evaluating the overall stability of the complex.^{31,259} While ii and iii could be included in the fitness function or used to pre-filter the candidates, the additional computational cost has to be considered. Instead, integrating more refined algorithms for identification of ring-closing conformations

(i.e., based on inverse kinematics),^{131–135} can significantly improve the efficiency of the method while allowing a more precise evaluation of the closability condition.

7.4 Conclusion

The design of multidentate ligands for Fe(II) SCO candidate compounds has involved the generation and modification of multicyclic systems thus serving as a challenging test case of the ring-closing machinery embedded in the 3D fragment-based design tool. The results demonstrate that new and unexpected multidentate ligands, which are or resemble existing molecules, are generated and evolved according to the definition of the fitness function. The best SCO candidates identified by the artificial evolution experiments are characterized by spin transition energies that are within the range suggested for possible SCO behaviour. Nevertheless, the stability of the designed metal complexes represents an issue that should be tackled both by evaluation of the preorganization of the ligands and by development of more refined algorithms for design of rings. The latter is also pointed out as a requirement for improving efficiency.

8. Concluding Remarks

The work presented in this thesis enhanced the capabilities of *de novo* design methods for applications in transition metal and organometallic chemistry, and provided insights on the issues related with automated handling of such species.

The need for generality was satisfied developing computational machinery capable of representing any sort of chemical entity without assumptions based on strict chemical formalisms. The resulting tool allowed proper handling of organometallic species (Paper I and II) and peculiar entities with bonding interactions not supported by other tools (Paper II), also combined with the capability of identifying novel cyclic systems potentially involving peculiar species (Paper IV).

Punctual control of the automated molecular generation and modification was provided by definition of a protocol based on annotated molecular fragments and connection rules which are collected in a so-called compatibility matrix. Retrosynthetic and purpose-based assembly of molecules could be achieved thus focussing the automated generation of candidates on realistic compounds with various ligand sets still retaining the properties of the metal coordinating environment (Paper I).

Preparation of accurate initial 3D molecular models by assembling of 3D building blocks was performed for peculiar tree-like chemical species demonstrating the superior performance of the approach that, contrarily to other methods, can precisely control the geometry of metal centers, their stereochemistry, and the spatial arrangement of chemical entities otherwise not supported by other automated tools (Paper II).

To empower the development and applications of fast fitness calculations, the ligand field molecular mechanics (LFMM) method was integrated into a promptly available software package (i.e., Tinker) thus providing broad accessibility to a method that is specifically meant for transition metal complexes (Paper III).

The capability of handling multicyclic systems has been evaluated by *de novo* design of Fe(II) spin-crossover compounds with multidentate amine ligands. The artificial evolution successfully identified promising ligands with unexpected skeletons that are, or resemble, existing molecules and, so far, have not been considered for spin-crossover compounds.

Finally, in the work presented herein, full automation was achieved for tasks such as the generation and modification of peculiar chemical entities, including the introduction and alteration of new rings from acyclic 3D building blocks, and the preparation of proper 3D guess structures, for tree-like molecules, that allow accurate evaluation of molecular properties. These capabilities are believed to boost the application of *de novo* design techniques to a broader range of chemical problems and, in particular, to the design of organometallic catalysts and functional transition metal compounds in general.

9. Directions for Future Work

The strength and drawback of *de novo* design approach is the vastness of the chemical space. In fact, while new molecules are easy to find, the numbers quickly become intractable. Paper I provided the means for restricting the exploration of the chemical space to a defined subspace, but the preparation of a representative sample of the chemical diversity enclosed in such a space is still a missing piece of the puzzle. Unfortunately the concept of chemical diversity is context-dependent,^{263,264} thus the sampling method should be flexible and provide a different set of criteria, i.e., descriptors, that are to be chosen accordingly to the needs of the specific application domain. Nevertheless, future developments will have to consider the introduction of diversity-oriented sampling of the chemical space identified by the library of building blocks and the connection rules.

It is worth mentioning an intended, yet still unexploited property of the double layer graph representation described in Section 3.2.1; that is, the outer graph can be used to represent multiple steps along a reaction path. In fact, while the root vertex indeed contains a fragment, the actual molecular representation of such fragment is hidden in the inner layer graph and perceived by the outer graph only in terms of its attachment points. The root fragment can be designed in such a way that all the bond formations and ruptures occurs within such vertex, meaning that all other vertices represent ancillary atoms altering the steric and electronic properties of the reacting system but not directly involved in the reaction. By providing alternative structural representations of the root fragment, making sure the attachment point are set consistently, then various reaction intermediate and transition states models can be obtained from a single graph. Multiple conversion of the graph into molecular representations, each time with a different root fragment, easily generates the initial guess structures for the desired set of reaction steps.

10. References

- (1) Schneider, G. *De Novo Molecular Design*; Wiley: Weinheim, Germany, 2014.
- (2) Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D.; Karelson, M.; Kahn, I.; Dobchev, D. A. Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chem. Rev.* **2010**, *110* (10), 5714–5789.
- (3) Dobson, C. M. Chemical Space and Biology. *Nature* **2004**, *432* (7019), 824–828.
- (4) Reymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48* (3), 722–730.
- (5) Weymuth, T.; Reiher, M. Inverse Quantum Chemistry: Concepts and Strategies for Rational Compound Design. *Int. J. Quantum Chem.* **2014**, *114* (13), 823–837.
- (6) Weymuth, T.; Reiher, M. Gradient-Driven Molecule Construction: An Inverse Approach Applied to the Design of Small-Molecule Fixating Catalysts. *Int. J. Quantum Chem.* **2014**, *114* (13), 838–850.
- (7) Mlinar, V. Utilization of Inverse Approach in the Design of Materials over Nano- to Macro-Scale. *Ann. Phys.* **2015**, *527* (3-4), 187–204.
- (8) Hall, R. J.; Mortenson, P. N.; Murray, C. W. Efficient Exploration of Chemical Space by Fragment-Based Screening. *Prog. Biophys. Mol. Biol.* **2014**, *116* (2–3), 82–91.
- (9) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martínez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *J. Chem. Inf. Model.* **2012**, *52* (4), 867–881.
- (10) Shoichet, B. K. Virtual Screening of Chemical Libraries. *Nature* **2004**, *432* (7019), 862–865.
- (11) Schneider, G.; Fechner, U. Computer-Based *de Novo* Design of Drug-like Molecules. *Nat. Rev. Drug Discov.* **2005**, *4* (8), 649–663.
- (12) Hu, Q.; Peng, Z.; Sutton, S. C.; Na, J.; Kostrowicki, J.; Yang, B.; Thacher, T.; Kong, X.; Mattaparti, S.; Zhou, J. Z.; Gonzalez, J.; Ramirez-Weinhouse, M.; Kuki, A. Pfizer Global Virtual Library (PGVL): A Chemistry Design Tool Powered by Experimentally Validated Parallel Synthesis Information. *ACS Comb. Sci.* **2012**, *14* (11), 579–589.
- (13) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864–2875.
- (14) Hartenfeller, M.; Schneider, G. *De Novo* Drug Design. In *Chemoinformatics and Computational Chemical Biology*; Bajorath, J., Ed.; Methods in Molecular Biology; Humana Press, 2011; pp 299–323.
- (15) Danziger, D. J.; Dean, P. M. Automated Site-Directed Drug Design: A General Algorithm for Knowledge Acquisition about Hydrogen-Bonding Regions at Protein Surfaces. *Proc. R. Soc. Lond. B Biol. Sci.* **1989**, *236* (1283), 101–113.

-
- (16) Lewis, R. A.; Dean, P. M. Automated Site-Directed Drug Design: The Formation of Molecular Templates in Primary Structure Generation. *Proc. R. Soc. Lond. B Biol. Sci.* **1989**, *236* (1283), 141–162.
- (17) Hartenfeller, M.; Schneider, G. Enabling Future Drug Discovery by *de Novo* Design. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (5), 742–759.
- (18) Dahiyat, B. I.; Mayo, S. L. *De Novo* Protein Design: Fully Automated Sequence Selection. *Science* **1997**, *278* (5335), 82–87.
- (19) Zanghellini, A. *De Novo* Computational Enzyme Design. *Curr. Opin. Biotechnol.* **2014**, *29*, 132–138.
- (20) Budin, N.; Majeux, N.; Tenette–Souaille, C.; Caflisch, A. Structure-Based Ligand Design by a Build-up Approach and Genetic Algorithm Search in Conformational Space. *J. Comput. Chem.* **2001**, *22* (16), 1956–1970.
- (21) Rodrigo, G.; Landrain, T. E.; Jaramillo, A. *De Novo* Automated Design of Small RNA Circuits for Engineering Synthetic Riboregulation in Living Cells. *Proc. Natl. Acad. Sci.* **2012**, *109* (38), 15271–15276.
- (22) Seiffert, J.; Huhle, A. A Full-Automatic Sequence Design Algorithm for Branched DNA Structures. *J. Biomol. Struct. Dyn.* **2008**, *25* (5), 453–466.
- (23) Lewis, D. W.; Willock, D. J.; Catlow, C. R. A.; Thomas, J. M.; Hutchings, G. J. *De Novo* Design of Structure-Directing Agents for the Synthesis of Microporous Solids. *Nature* **1996**, *382* (6592), 604–606.
- (24) Corà, F.; Catlow, C. R. A.; Lewis, D. W. *De Novo* Design of Microporous Transition Metal Oxides. *Chem. Commun.* **1998**, No. 18, 1943–1944.
- (25) Pophale, R.; Daeyaert, F.; Deem, M. W. Computational Prediction of Chemically Synthesizable Organic Structure Directing Agents for Zeolites. *J. Mater. Chem. A* **2013**, *1* (23), 6750–6760.
- (26) Bao, Y.; Martin, R. L.; Simon, C. M.; Haranczyk, M.; Smit, B.; Deem, M. W. *In Silico* Discovery of High Deliverable Capacity Metal–Organic Frameworks. *J. Phys. Chem. C* **2015**, *119* (1), 186–195.
- (27) Martin, R. L.; Simon, C. M.; Smit, B.; Haranczyk, M. *In Silico* Design of Porous Polymer Networks: High-Throughput Screening for Methane Storage Materials. *J. Am. Chem. Soc.* **2014**, *136* (13), 5006–5022.
- (28) Shu, Y.; Levine, B. G. Simulated Evolution of Fluorophores for Light Emitting Diodes. *J. Chem. Phys.* **2015**, *142* (10), 104104.
- (29) Venkatraman, V.; Foscatto, M.; Jensen, V. R.; Alsberg, B. K. Evolutionary *de Novo* Design of Phenothiazine Derivatives for Dye-Sensitized Solar Cells. *J. Mater. Chem. A* **2015**, *3* (18), 9851–9860.
- (30) Chu, Y.; Heyndrickx, W.; Occhipinti, G.; Jensen, V. R.; Alsberg, B. K. An Evolutionary Algorithm for *de Novo* Optimization of Functional Transition Metal Compounds. *J. Am. Chem. Soc.* **2012**, *134* (21), 8885–8895.
- (31) Comba, P.; Hambley, T. W.; Martin, B. *Molecular Modeling of Inorganic Compounds*; John Wiley & Sons: Weinheim, Germany, 2009.
- (32) Bauerschmidt, S.; Gasteiger, J. Overcoming the Limitations of a Connection Table Description: A Universal Representation of Chemical Species. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (4), 705–714.

-
- (33) Clark, A. M.; Williams, A. J.; Ekins, S. Machines First, Humans Second: On the Importance of Algorithmic Interpretation of Open Chemistry Data. *J. Cheminformatics* **2015**, *7* (1), 9.
- (34) Hastings, J.; Matos, P. de; Dekker, A.; Ennis, M.; Harsha, B.; Kale, N.; Muthukrishnan, V.; Owen, G.; Turner, S.; Williams, M.; Steinbeck, C. The ChEBI Reference Database and Ontology for Biologically Relevant Chemistry: Enhancements for 2013. *Nucleic Acids Res.* **2013**, *41* (D1), D456–D463.
- (35) Devi, R. V.; Sathya, S. S.; Coumar, M. S. Evolutionary Algorithms for *de Novo* Drug Design – A Survey. *Appl. Soft Comput.* **2015**, *27*, 543–552.
- (36) Lameijer, E.-W.; Bäck, T.; Kok, J. N.; Ijzerman, A. P. Evolutionary Algorithms in Drug Design. *Nat. Comput.* **2005**, *4* (3), 177–243.
- (37) Boussaïd, I.; Lepagnot, J.; Siarry, P. A Survey on Optimization Metaheuristics. *Inf. Sci.* **2013**, *237*, 82–117.
- (38) Brown, N.; McKay, B.; Gilardoni, F.; Gasteiger, J. A Graph-Based Genetic Algorithm and Its Application to the Multiobjective Evolution of Median Molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 1079–1087.
- (39) Nicolaou, C. A.; Brown, N.; Pattichis, C. S. Molecular Optimization Using Computational Multi-Objective Methods. *Curr. Opin. Drug Discov. Devel.* **2007**, *10* (3), 316–324.
- (40) Ekins, S.; Honeycutt, J. D.; Metz, J. T. Evolving Molecules Using Multi-Objective Optimization: Applying to ADME/Tox. *Drug Discov. Today* **2010**, *15* (11–12), 451–460.
- (41) Nicolaou, C. A.; Kannas, C.; Loizidou, E. Multi-Objective Optimization Methods in *de Novo* Drug Design. *Mini Rev. Med. Chem.* **2012**, *12* (10), 979–987.
- (42) Nicolaou, C. A.; Brown, N. Multi-Objective Optimization Methods in Drug Design. *Drug Discov. Today Technol.* **2013**, *10* (3), e427–e435.
- (43) Leach, A. *Molecular Modelling: Principles and Applications*, 2 edition.; Prentice Hall: Harlow, England, 2001.
- (44) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH, 2000.
- (45) Lewars, E. G. *Computational Chemistry. Introduction to the Theory and Applications of Molecular and Quantum Mechanics*; Kluwer Academic Publishers: Secaucus, NJ, USA, 2003.
- (46) Griffiths, D. J. *Introduction to Quantum Mechanics*, 2nd edition.; Pearson Prentice Hall: Upper Saddle River, NJ, 2004.
- (47) Ramachandran, K. I.; Deepa, G.; Namboori, K. *Computational Chemistry and Molecular Modeling: Principles and Applications*, 2008 edition.; Springer: Berlin, 2008.
- (48) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2007**, *120* (1-3), 215–241.
- (49) Grimme, S. Density Functional Theory with London Dispersion Corrections. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (2), 211–228.

-
- (50) Levitt, M.; Warshel, A. Computer Simulation of Protein Folding. *Nature* **1975**, 253 (5494), 694–698.
- (51) Kolinski, A. Protein Modeling and Structure Prediction with a Reduced Representation. *Acta Biochim. Pol.* **2004**, 51 (2), 349–371.
- (52) Nielsen, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L. Coarse Grain Models and the Computer Simulation of Soft Materials. *J. Phys. Condens. Matter* **2004**, 16 (15), R481.
- (53) Halgren, T. A.; Damm, W. Polarizable Force Fields. *Curr. Opin. Struct. Biol.* **2001**, 11 (2), 236–242.
- (54) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **2010**, 114 (8), 2549–2564.
- (55) Nishibata, Y.; Itai, A. Automatic Creation of Drug Candidate Structures Based on Receptor Structure. Starting Point for Artificial Lead Generation. *Tetrahedron* **1991**, 47 (43), 8985–8990.
- (56) Rotstein, S. H.; Murcko, M. A. GenStar: A Method for *de Novo* Drug Design. *J. Comput. Aided Mol. Des.* **1993**, 7 (1), 23–43.
- (57) Gehlhaar, D. K.; Moerder, K. E.; Zichi, D.; Sherman, C. J.; Ogden, R. C.; Freer, S. T. *De Novo* Design of Enzyme Inhibitors by Monte Carlo Ligand Generation. *J. Med. Chem.* **1995**, 38 (3), 466–472.
- (58) Glen, R. C.; Payne, A. W. R. A Genetic Algorithm for the Automated Generation of Molecules within Constraints. *J. Comput. Aided Mol. Des.* **1995**, 9 (2), 181–202.
- (59) Nachbar, R. B. Molecular Evolution: Automated Manipulation of Hierarchical Chemical Topology and Its Application to Average Molecular Structures. *Genet. Program. Evolvable Mach.* **2000**, 1 (1-2), 57–94.
- (60) Globus, A.; Lawton, J.; Wipke, T. Automatic Molecular Design Using Evolutionary Techniques. *Nanotechnology* **1999**, 10 (3), 290.
- (61) Douguet, D.; Thoreau, E.; Grassy, G. A Genetic Algorithm for the Automated Generation of Small Organic Molecules: Drug Design Using an Evolutionary Algorithm. *J. Comput. Aided Mol. Des.* **2000**, 14 (5), 449–466.
- (62) Mishima, K.; Kaneko, H.; Funatsu, K. Development of a New *De Novo* Design Algorithm for Exploring Chemical Space. *Mol. Inform.* **2014**, 33 (11-12), 779–789.
- (63) Kawai, K.; Yoshimaru, K.; Takahashi, Y. Generation of Target-Selective Drug Candidate Structures Using Molecular Evolutionary Algorithm with SVM Classifiers. *J. Comput. Chem. Jpn.* **2011**, 10 (3), 79–87.
- (64) Frenking, G.; Fröhlich, N. The Nature of the Bonding in Transition-Metal Compounds. *Chem. Rev.* **2000**, 100 (2), 717–774.
- (65) Kutchukian, P. S.; Shakhnovich, E. I. *De Novo* Design: Balancing Novelty and Confined Chemical Space. *Expert Opin. Drug Discov.* **2010**, 5 (8), 789–812.
- (66) Huang, Q.; Li, L.-L.; Yang, S.-Y. PhDD: A New Pharmacophore-Based *de Novo* Design Method of Drug-like Molecules Combined with Assessment of Synthetic Accessibility. *J. Mol. Graph. Model.* **2010**, 28 (8), 775–787.

-
- (67) Gillet, V. J.; Myatt, G.; Zsoldos, Z.; Johnson, A. P. SPROUT, HIPPO and CAESA: Tools for *de Novo* Structure Generation and Estimation of Synthetic Accessibility. *Perspect. Drug Discov. Des.* **1995**, *3* (1), 34–50.
- (68) Baber, J.; Feher, M. Predicting Synthetic Accessibility: Application in Drug Discovery and Development. *Mini-Rev. Med. Chem.* **2004**, *4* (6), 681–692.
- (69) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminformatics* **2009**, *1* (1), 8.
- (70) Fukunishi, Y.; Kurosawa, T.; Mikami, Y.; Nakamura, H. Prediction of Synthetic Accessibility Based on Commercially Available Compound Databases. *J. Chem. Inf. Model.* **2014**, *54* (12), 3259–3267.
- (71) Boda, K.; Seidel, T.; Gasteiger, J. Structure and Reaction Based Evaluation of Synthetic Accessibility. *J. Comput. Aided Mol. Des.* **2007**, *21* (6), 311–325.
- (72) Schneider, G.; Lee, M. L.; Stahl, M.; Schneider, P. *De Novo* Design of Molecular Architectures by Evolutionary Assembly of Drug-Derived Building Blocks. *J. Comput. Aided Mol. Des.* **2000**, *14* (5), 487–494.
- (73) Pegg, S. C.-H.; Haresco, J. J.; Kuntz, I. D. A Genetic Algorithm for Structure-Based *de Novo* Design. *J. Comput. Aided Mol. Des.* **2001**, *15* (10), 911–933.
- (74) Dey, F.; Caflisch, A. Fragment-Based *de Novo* Ligand Design by Multiobjective Evolutionary Optimization. *J. Chem. Inf. Model.* **2008**, *48* (3), 679–690.
- (75) Fechner, U.; Schneider, G. Flux (1): A Virtual Synthesis Scheme for Fragment-Based *de Novo* Design. *J. Chem. Inf. Model.* **2006**, *46* (2), 699–707.
- (76) Kutchukian, P. S.; Lou, D.; Shakhnovich, E. I. FOG: Fragment Optimized Growth Algorithm for the *de Novo* Generation of Molecules Occupying Druglike Chemical Space. *J. Chem. Inf. Model.* **2009**, *49* (7), 1630–1642.
- (77) Durrant, J. D.; Amaro, R. E.; McCammon, J. A. AutoGrow: A Novel Algorithm for Protein Inhibitor Design. *Chem. Biol. Drug Des.* **2009**, *73* (2), 168–178.
- (78) Loving, K.; Alberts, I.; Sherman, W. Computational Approaches for Fragment-Based and *de Novo* Design. *Curr. Top. Med. Chem.* **2010**, *10* (1), 14–32.
- (79) Mortier, J.; Rakers, C.; Frederick, R.; Wolber, G. Computational Tools for *in Silico* Fragment-Based Drug Design. *Curr. Top. Med. Chem.* **2012**, *12* (17), 1935–1943.
- (80) Wang, R.; Gao, Y.; Lai, L. LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design. *Mol. Model. Annu.* **2000**, *6* (7-8), 498–516.
- (81) Ehrlich, H.-C.; Volkamer, A.; Rarey, M. Searching for Substructures in Fragment Spaces. *J. Chem. Inf. Model.* **2012**, *52* (12), 3181–3189.
- (82) Ertl, P.; Lewis, R. IADE: A System for Intelligent Automatic Design of Bioisosteric Analogs. *J. Comput. Aided Mol. Des.* **2012**, *26* (11), 1207–1215.
- (83) Pirard, B.; Ertl, P. Evaluation of a Semi-Automated Workflow for Fragment Growing. *J. Chem. Inf. Model.* **2015**, *55* (1), 180–193.
- (84) Van Deursen, R.; Reymond, J.-L. Chemical Space Travel. *ChemMedChem* **2007**, *2* (5), 636–640.
- (85) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the Art of Compiling and Using “Drug-Like” Chemical Fragment Spaces. *ChemMedChem* **2008**, *3* (10), 1503–1507.

-
- (86) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. 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.* **1998**, *38* (3), 511–522.
- (87) Degen, J.; Rarey, M. FlexNovo: Structure-Based Searching in Large Fragment Spaces. *ChemMedChem* **2006**, *1* (8), 854–868.
- (88) Dean, P. M.; Firth-Clark, S.; Harris, W.; Kirton, S. B.; Todorov, N. P. SkelGen: A General Tool for Structure-Based *de Novo* Ligand Design. *Expert Opin. Drug Discov.* **2006**, *1* (2), 179–189.
- (89) Makino, S.; Ewing, T. J. A.; Kuntz, I. D. DREAM++: Flexible Docking Program for Virtual Combinatorial Libraries. *J. Comput. Aided Mol. Des.* **1999**, *13* (5), 513–532.
- (90) Vinkers, H. M.; de Jonge, M. R.; Daeyaert, F. F. D.; Heeres, J.; Koymans, L. M. H.; van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Van Aken, K.; Janssen, P. A. J. SYNOPSIS: SYNthesize and OPTimize System in Silico. *J. Med. Chem.* **2003**, *46* (13), 2765–2773.
- (91) Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. DOGS: Reaction-Driven *de Novo* Design of Bioactive Compounds. *PLoS Comput Biol* **2012**, *8* (2), e1002380.
- (92) Beccari, A. R.; Cavazzoni, C.; Beato, C.; Costantino, G. LiGen: A High Performance Workflow for Chemistry Driven *de Novo* Design. *J. Chem. Inf. Model.* **2013**, *53* (6), 1518–1527.
- (93) Schürer, S. C.; Tyagi, P.; Muskal, S. M. Prospective Exploration of Synthetically Feasible, Medicinally Relevant Chemical Space. *J. Chem. Inf. Model.* **2005**, *45* (2), 239–248.
- (94) Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K.-H.; Schneider, G.; Jacoby, E.; Renner, S. A Collection of Robust Organic Synthesis Reactions for In Silico Molecule Design. *J. Chem. Inf. Model.* **2011**, *51* (12), 3093–3098.
- (95) Warr, W. A. A Short Review of Chemical Reaction Database Systems, Computer-Aided Synthesis Design, Reaction Prediction and Synthetic Feasibility. *Mol. Inform.* **2014**, *33* (6-7), 469–476.
- (96) Hay, B. P.; Firman, T. K.; Lumetta, G. J.; Rapko, B. M.; Garza, P. A.; Sinkov, S. I.; Hutchison, J. E.; Parks, B. W.; Gilbertson, R. D.; Weakley, T. J. R. Toward the Computer-Aided Design of Metal Ion Sequestering Agents. *J. Alloys Compd.* **2004**, *374* (1–2), 416–419.
- (97) Hageman, J. A.; Westerhuis, J. A.; Frühauf, H.-W.; Rothenberg, G. Design and Assembly of Virtual Homogeneous Catalyst Libraries – Towards *in Silico* Catalyst Optimisation. *Adv. Synth. Catal.* **2006**, *348* (3), 361–369.
- (98) Hay, B. P.; Firman, T. K. HostDesigner: A Program for the *de Novo* Structure-Based Design of Molecular Receptors with Binding Sites That Complement Metal Ion Guests. *Inorg. Chem.* **2002**, *41* (21), 5502–5512.
- (99) Bryantsev, V. S.; Hay, B. P. *De Novo* Structure-Based Design of Bisurea Hosts for Tetrahedral Oxoanion Guests. *J. Am. Chem. Soc.* **2006**, *128* (6), 2035–2042.
- (100) Hay, B. P.; Jia, C.; Nadas, J. Computer-Aided Design of Host Molecules for Recognition of Organic Guests. *Comput. Theor. Chem.* **2014**, *1028*, 72–80.

-
- (101) Bryan, J. C.; Hay, B. P.; Sachleben, R. A.; Eagle, C. T.; Zhang, C.; Bonnesen, P. V. Design, Synthesis, and Structure of Novel Cesium Receptors. *J. Chem. Crystallogr.* **2003**, *33* (5-6), 349–355.
- (102) Vukovic, S.; Hay, B. P. *De Novo* Structure-Based Design of Bis-Amidoxime Uranophiles. *Inorg. Chem.* **2013**, *52* (13), 7805–7810.
- (103) Burello, E.; Rothenberg, G. *In Silico* Design in Homogeneous Catalysis Using Descriptor Modelling. *Int. J. Mol. Sci.* **2006**, *7* (9), 375–404.
- (104) Warr, W. A. Representation of Chemical Structures. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1* (4), 557–579.
- (105) Chepelev, L. L.; Dumontier, M. Chemical Entity Semantic Specification: Knowledge Representation for Efficient Semantic Cheminformatics and Facile Data Integration. *J. Cheminformatics* **2011**, *3* (1), 20.
- (106) Yang, C.; Tarkhov, A.; Maruszyk, J.; Bienfait, B.; Gasteiger, J.; Kleinoeder, T.; Magdziarz, T.; Sacher, O.; Schwab, C. H.; Schwoebel, J.; Terfloth, L.; Arvidson, K.; Richard, A.; Worth, A.; Rathman, J. New Publicly Available Chemical Query Language, CSRML, To Support Chemotype Representations for Application to Data Mining and Modeling. *J. Chem. Inf. Model.* **2015**, *55* (3), 510–528.
- (107) Daylight Theory Manual, In Daylight Chemical Information System, Inc. <http://www.daylight.com/dayhtml/doc/theory/index.html> (accessed Mar 21, 2013).
- (108) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36.
- (109) O'Boyle, N. M. Towards a Universal SMILES Representation - A Standard Method to Generate Canonical SMILES Based on the InChI. *J. Cheminformatics* **2012**, *4* (1), 22.
- (110) Gakh, A. A.; Burnett, M. N. Modular Chemical Descriptor Language (MCDL): Composition, Connectivity, and Supplementary Modules. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (6), 1494–1499.
- (111) Gakh, A. A.; Burnett, M. N.; Trepalin, S. V.; Yarkov, A. V. Modular Chemical Descriptor Language (MCDL): Stereochemical Modules. *J. Cheminformatics* **2011**, *3* (1), 5.
- (112) Heller, S. R.; McNaught, A. D. The IUPAC International Chemical Identifier (InChI). *Chem. Int.* **2009**, *31* (1), 7–9.
- (113) Cho, Y. S.; No, K. T.; Cho, K.-H. yaInChI: Modified InChI String Scheme for Line Notation of Chemical Structures. *SAR QSAR Environ. Res.* **2012**, *23* (3-4), 237–255.
- (114) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32* (3), 244–255.
- (115) Willighagen, E. Three-Dimensional (3D) Molecular Representations. In *Handbook of Chemoinformatics Algorithms*; Chapman and Hall/CRC, 2010; Vol. 20103144, pp 65–87.

-
- (116) Balaban, A. T. Applications of Graph Theory in Chemistry. *J. Chem. Inf. Comput. Sci.* **1985**, 25 (3), 334–343.
- (117) Nicolaou, C. A.; Apostolakis, J.; Pattichis, C. S. *De Novo* Drug Design Using Multiobjective Evolutionary Graphs. *J. Chem. Inf. Model.* **2009**, 49 (2), 295–307.
- (118) Wong, S. S. Y.; Luo, W.; Chan, K. C. C. EvoMD: An Algorithm for Evolutionary Molecular Design. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2011**, 8 (4), 987–1003.
- (119) Kawai, K.; Nagata, N.; Takahashi, Y. *De Novo* Design of Drug-Like Molecules by a Fragment-Based Molecular Evolutionary Approach. *J. Chem. Inf. Model.* **2014**, 54 (1), 49–56.
- (120) Teodoro, M.; Muegge, I. BIBuilder: Exhaustive Searching for *De Novo* Ligands. *Mol. Inform.* **2011**, 30 (1), 63–75.
- (121) Douguet, D.; Munier-Lehmann, H.; Labesse, G.; Pochet, S. LEA3D: A Computer-Aided Ligand Design for Structure-Based Drug Design. *J. Med. Chem.* **2005**, 48 (7), 2457–2468.
- (122) DeWitte, R. S.; Shakhnovich, E. I. SMOG: *de Novo* Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence. *J. Am. Chem. Soc.* **1996**, 118 (47), 11733–11744.
- (123) Murray, C. W.; Rees, D. C. The Rise of Fragment-Based Drug Discovery. *Nat. Chem.* **2009**, 1 (3), 187–192.
- (124) Kutchukian, P. S.; Virtanen, S. I.; Lounkine, E.; Glick, M.; Shakhnovich, E. I. Construction of Drug-Like Compounds by Markov Chains. In *De novo Molecular Design*; Schneider, G., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA, 2013; pp 311–323.
- (125) Jonathan L. Gross; Jay Yellen; Ping Zhang. *Handbook of Graph Theory, Second Edition*; Discrete Mathematics and Its Applications; Chapman and Hall/CRC, 2013.
- (126) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, 43 (2), 493–500.
- (127) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. *Curr. Pharm. Des.* **2006**, 12 (17), 2111–2120.
- (128) Foscatto, M.; Occhipinti, G.; Venkatraman, V.; Alsberg, B. K.; Jensen, V. R. Automated Design of Realistic Organometallic Molecules from Fragments. *J. Chem. Inf. Model.* **2014**, 54 (3), 767–780.
- (129) Mauser, H.; Stahl, M. Chemical Fragment Spaces for *de Novo* Design. *J. Chem. Inf. Model.* **2007**, 47 (2), 318–324.
- (130) Hoffmann, R. Building Bridges Between Inorganic and Organic Chemistry (Nobel Lecture). *Angew. Chem. Int. Ed. Engl.* **1982**, 21 (10), 711–724.
- (131) Bruccoleri, R. E.; Karplus, M. Chain Closure with Bond Angle Variations. *Macromolecules* **1985**, 18 (12), 2767–2773.
- (132) Coutsiias, E. A.; Seok, C.; Jacobson, M. P.; Dill, K. A. A Kinematic View of Loop Closure. *J. Comput. Chem.* **2004**, 25 (4), 510–528.

-
- (133) Coutsiias, E. A.; Seok, C.; Wester, M. J.; Dill, K. A. Resultants and Loop Closure. *Int. J. Quantum Chem.* **2006**, *106* (1), 176–189.
- (134) Chys, P.; Chacón, P. Random Coordinate Descent with Spinor-Matrices and Geometric Filters for Efficient Loop Closure. *J. Chem. Theory Comput.* **2013**, *9* (3), 1821–1829.
- (135) Zamuner, S.; Rodriguez, A.; Seno, F.; Trovato, A. An Efficient Algorithm to Perform Local Concerted Movements of a Chain Molecule. *PLoS ONE* **2015**, *10* (3), e0118342.
- (136) Foscatto, M.; Venkatraman, V.; Occhipinti, G.; Alsberg, B. K.; Jensen, V. R. Automated Building of Organometallic Complexes from 3D Fragments. *J. Chem. Inf. Model.* **2014**, *54* (7), 1919–1931.
- (137) Kostrowicki, J.; Scheraga, H. A. Application of the Diffusion Equation Method for Global Optimization to Oligopeptides. *J. Phys. Chem.* **1992**, *96* (18), 7442–7449.
- (138) Nakamura, S.; Hirose, H.; Ikeguchi, M.; Doi, J. Conformational Energy Minimization Using a Two-Stage Method. *J. Phys. Chem.* **1995**, *99* (20), 8374–8378.
- (139) Pappu, R. V.; Hart, R. K.; Ponder, J. W. Analysis and Application of Potential Energy Smoothing and Search Methods for Global Optimization. *J. Phys. Chem. B* **1998**, *102* (48), 9725–9742.
- (140) Hart, R. K.; Pappu, R. V.; Ponder, J. W. Exploring the Similarities between Potential Smoothing and Simulated Annealing. *J. Comput. Chem.* **2000**, *21* (7), 531–552.
- (141) Jay W. Ponder. TINKER: Software Tools for Molecular Design, 6.2ed.; Washington University School of Medicine: Saint Louis, MO, 2013.
- (142) Rappé, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114* (25), 10024–10035.
- (143) Hoveyda, A. H.; Zhugralin, A. R. The Remarkable Metal-Catalysed Olefin Metathesis Reaction. *Nature* **2007**, *450* (7167), 243–251.
- (144) Scholl, M.; Ding, S.; Lee, C. W.; Grubbs, R. H. 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.* **1999**, *1* (6), 953–956.
- (145) Huang, J.; Stevens, E. D.; Nolan, S. P.; Petersen, J. L. Olefin Metathesis-Active Ruthenium Complexes Bearing a Nucleophilic Carbene Ligand. *J. Am. Chem. Soc.* **1999**, *121* (12), 2674–2678.
- (146) Trnka, T. M.; Grubbs, R. H. The Development of L₂X₂RuCHR Olefin Metathesis Catalysts: An Organometallic Success Story. *Acc. Chem. Res.* **2001**, *34* (1), 18–29.
- (147) Kingsbury, J. S.; Harrity, J. P. A.; Bonitatebus, P. J.; Hoveyda, A. H. A Recyclable Ru-Based Metathesis Catalyst. *J. Am. Chem. Soc.* **1999**, *121* (4), 791–799.
- (148) Garber, S. B.; Kingsbury, J. S.; Gray, B. L.; Hoveyda, A. H. Efficient and Recyclable Monomeric and Dendritic Ru-Based Metathesis Catalysts. *J. Am. Chem. Soc.* **2000**, *122* (34), 8168–8179.

- (149) Allen, F. H. The Cambridge Structural Database: A Quarter of a Million Crystal Structures and Rising. *Acta Crystallogr. B* **2002**, *58* (3), 380–388.
- (150) Lee, A. C.; Crippen, G. M. Predicting pKa. *J. Chem. Inf. Model.* **2009**, *49* (9), 2013–2033.
- (151) Li, H.; Robertson, A. D.; Jensen, J. H. Very Fast Empirical Prediction and Rationalization of Protein pKa Values. *Proteins Struct. Funct. Bioinforma.* **2005**, *61* (4), 704–721.
- (152) Zevatskii, Y. E.; Samoilov, D. V. Modern Methods for Estimation of Ionization Constants of Organic Compounds in Solution. *Russ. J. Org. Chem.* **2011**, *47* (10), 1445–1467.
- (153) Alexov, E.; Mehler, E. L.; Baker, N.; M. Baptista, A.; Huang, Y.; Milletti, F.; Erik Nielsen, J.; Farrell, D.; Carstensen, T.; Olsson, M. H. M.; Shen, J. K.; Warwicker, J.; Williams, S.; Word, J. M. Progress in the Prediction of pKa Values in Proteins. *Proteins Struct. Funct. Bioinforma.* **2011**, *79* (12), 3260–3275.
- (154) Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Manganaro, A. QSPR Modeling of Enthalpies of Formation for Organometallic Compounds by SMART-Based Optimal Descriptors. *J. Comput. Chem.* **2009**, *30* (15), 2576–2582.
- (155) Toropov, A. A.; Toropova, A. P.; Benfenati, E. QSAR-Modeling of Toxicity of Organometallic Compounds by Means of the Balance of Correlations for InChI-Based Optimal Descriptors. *Mol. Divers.* **2010**, *14* (1), 183–192.
- (156) Drummond, M. L.; Sumpter, B. G. Use of Drug Discovery Tools in Rational Organometallic Catalyst Design. *Inorg. Chem.* **2007**, *46* (21), 8613–8624.
- (157) Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *J. Chem. Inf. Model.* **2012**, *52* (5), 1146–1158.
- (158) Rusinko, A.; Sheridan, R. P.; Nilakantan, R.; Haraki, K. S.; Bauman, N.; Venkataraghavan, R. Using CONCORD to Construct a Large Database of Three-Dimensional Coordinates from Connection Tables. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (4), 251–255.
- (159) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (4), 1000–1008.
- (160) Mizutani, M. Y.; Nakamura, K.; Ichinose, T.; Itai, A. Starting Point to Molecular Design: Efficient Automated 3D Model Builder Key3D. *Chem. Pharm. Bull. (Tokyo)* **2006**, *54* (12), 1680–1685.
- (161) Leite, T. B.; Gomes, D.; Miteva, M. A.; Chomilier, J.; Villoutreix, B. O.; Tufféry, P. Frog: A FRee Online druG 3D Conformation Generator. *Nucleic Acids Res.* **2007**, *35* (Web Server issue), W568–W572.
- (162) Lagorce, D.; Pencheva, T.; Villoutreix, B. O.; Miteva, M. A. DG-AMMOS: A New Tool to Generate 3D Conformation of Small Molecules Using Distance Geometry and Automated Molecular Mechanics Optimization for *in Silico* Screening. *BMC Chem. Biol.* **2009**, *9* (1), 6.

-
- (163) Miteva, M. A.; Guyon, F.; Tufféry, P. Frog2: Efficient 3D Conformation Ensemble Generator for Small Compounds. *Nucleic Acids Res.* **2010**, *38* (suppl 2), W622–W627.
- (164) Lagorce, D.; Villoutreix, B. O.; Miteva, M. A. Three-Dimensional Structure Generators of Drug-like Compounds: DG-AMMOS, an Open-Source Package. *Expert Opin. Drug Discov.* **2011**, *6* (3), 339–351.
- (165) Hawkins, P. C. D.; Nicholls, A. Conformer Generation with OMEGA: Learning from the Data Set and the Analysis of Failures. *J. Chem. Inf. Model.* **2012**, *52* (11), 2919–2936.
- (166) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D-Atomic Coordinates for Organic Molecules. *Tetrahedron Comput. Methodol.* **1990**, *3* (6, Part C), 537–547.
- (167) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93* (7), 2567–2581.
- (168) Gasteiger, J.; Sadowski, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. Chemical Information in 3D Space. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (5), 1030–1037.
- (169) Andronico, A.; Randall, A.; Benz, R. W.; Baldi, P. Data-Driven High-Throughput Prediction of the 3-D Structure of Small Molecules: Review and Progress. *J. Chem. Inf. Model.* **2011**, *51* (4), 760–776.
- (170) Sadowski, P.; Baldi, P. Small-Molecule 3D Structure Prediction Using Open Crystallography Data. *J. Chem. Inf. Model.* **2013**, *53* (12), 3127–3130.
- (171) Buda, C.; Flores, A.; Cundari, T. R. *De Novo* Prediction of the Ground State Structure of Transition Metal Complexes Using Semiempirical and *ab Initio* Quantum Mechanics. Coordination Isomerism. *J. Coord. Chem.* **2005**, *58* (7), 575–585.
- (172) Buda, C.; Burt, S. K.; Cundari, T. R.; Shenkin, P. S. *De Novo* Structural Prediction of Transition Metal Complexes: Application to Technetium. *Inorg. Chem.* **2002**, *41* (8), 2060–2069.
- (173) Buda, C.; Cundari, T. R. *De Novo* Prediction of Ground State Multiplicity and Structural Isomerism for Transition Metal Complexes. *J. Mol. Struct. THEOCHEM* **2004**, *686* (1–3), 137–145.
- (174) Ball, D. M.; Buda, C.; Gillespie, A. M.; White, D. P.; Cundari, T. R. Can Semiempirical Quantum Mechanics Be Used To Predict the Spin State of Transition Metal Complexes? An Application of *De Novo* Prediction. *Inorg. Chem.* **2002**, *41* (1), 152–156.
- (175) Comba, P.; Kerscher, M. Computation of Structures and Properties of Transition Metal Compounds. *Coord. Chem. Rev.* **2009**, *253* (5–6), 564–574.
- (176) Occhipinti, G.; Bjørsvik, H.-R.; Jensen, V. R. Quantitative Structure-Activity Relationships of Ruthenium Catalysts for Olefin Metathesis. *J. Am. Chem. Soc.* **2006**, *128* (21), 6952–6964.
- (177) Romero, P. E.; Piers, W. E.; McDonald, R. Rapidly Initiating Ruthenium Olefin-Metathesis Catalysts. *Angew. Chem. Int. Ed.* **2004**, *43* (45), 6161–6165.
- (178) Dubberley, S. R.; Romero, P. E.; Piers, W. E.; McDonald, R.; Parvez, M. Synthesis, Characterization and Olefin Metathesis Studies of a Family of

- Ruthenium Phosphonium Alkylidene Complexes. *Inorganica Chim. Acta* **2006**, 359 (9), 2658–2664.
- (179) Wenzel, A. G.; Grubbs, R. H. Ruthenium Metallacycles Derived from 14-Electron Complexes. New Insights into Olefin Metathesis Intermediates. *J. Am. Chem. Soc.* **2006**, 128 (50), 16048–16049.
- (180) O’Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminformatics* **2011**, 3 (1), 33.
- (181) Marvin, Version 5.11.1, 2012, ChemAxon, Web: [Http://www.chemaxon.com](http://www.chemaxon.com).
- (182) MOPAC2012, Stewart, James J. P., Stewart Computational Chemistry, Version 13.113L, Web: <http://OpenMOPAC.net>.
- (183) Handy, N. C.; Cohen, A. J. Left-Right Correlation Energy. *Mol. Phys.* **2001**, 99 (5), 403–412.
- (184) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, 37 (2), 785–789.
- (185) Wadt, W. R.; Hay, P. J. Ab Initio Effective Core Potentials for Molecular Calculations. Potentials for Main Group Elements Na to Bi. *J. Chem. Phys.* **1985**, 82 (1), 284–298.
- (186) Hay, P. J.; Wadt, W. R. Ab Initio Effective Core Potentials for Molecular Calculations. Potentials for K to Au Including the Outermost Core Orbitals. *J. Chem. Phys.* **1985**, 82 (1), 299–310.
- (187) Minenkov, Y.; Occhipinti, G.; Jensen, V. R. Metal–Phosphine Bond Strengths of the Transition Metals: A Challenge for DFT. *J. Phys. Chem. A* **2009**, 113 (43), 11833–11844.
- (188) Sparta, M.; Børve, K. J.; Jensen, V. R. Activity of Rhodium-Catalyzed Hydroformylation: Added Insight and Predictions from Theory. *J. Am. Chem. Soc.* **2007**, 129 (27), 8487–8499.
- (189) Macchioni, A. Ion Pairing in Transition-Metal Organometallic Chemistry. *Chem. Rev.* **2005**, 105 (6), 2039–2074.
- (190) Nomura, K.; Liu, J.; Padmanabhan, S.; Kitiyanan, B. Nonbridged Half-Metallocenes Containing Anionic Ancillary Donor Ligands: New Promising Candidates as Catalysts for Precise Olefin Polymerization. *J. Mol. Catal. Chem.* **2007**, 267 (1–2), 1–29.
- (191) Manz, T. A.; Phomphrai, K.; Medvedev, G.; Krishnamurthy, B. B.; Sharma, S.; Haq, J.; Novstrup, K. A.; Thomson, K. T.; Delgass, W. N.; Caruthers, J. M.; Abu-Omar, M. M. Structure–Activity Correlation in Titanium Single-Site Olefin Polymerization Catalysts Containing Mixed Cyclopentadienyl/Aryloxy Ligand. *J. Am. Chem. Soc.* **2007**, 129 (13), 3776–3777.
- (192) Manz, T. A.; Sharma, S.; Phomphrai, K.; Novstrup, K. A.; Fenwick, A. E.; Fanwick, P. E.; Medvedev, G. A.; Abu-Omar, M. M.; Delgass, W. N.; Thomson, K. T.; Caruthers, J. M. Quantitative Effects of Ion Pairing and Sterics on Chain Propagation Kinetics for 1-Hexene Polymerization Catalyzed by Mixed Cp’/ArO Complexes. *Organometallics* **2008**, 27 (21), 5504–5520.
- (193) Manz, T. A.; Caruthers, J. M.; Sharma, S.; Phomphrai, K.; Thomson, K. T.; Delgass, W. N.; Abu-Omar, M. M. Structure–Activity Correlation for Relative

- Chain Initiation to Propagation Rates in Single-Site Olefin Polymerization Catalysis. *Organometallics* **2012**, *31* (2), 602–618.
- (194) Fey, N. The Contribution of Computational Studies to Organometallic Catalysis: Descriptors, Mechanisms and Models. *Dalton Trans.* **2009**, *39* (2), 296–310.
- (195) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96* (3), 1027–1044.
- (196) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics, Volume 41 (2 Volume Set)*; John Wiley & Sons, 2009.
- (197) Hechinger, M.; Leonhard, K.; Marquardt, W. What Is Wrong with Quantitative Structure–Property Relations Models Based on Three-Dimensional Descriptors? *J. Chem. Inf. Model.* **2012**, *52* (8), 1984–1993.
- (198) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative Structure–Property Relationship Modeling of Diverse Materials Properties. *Chem. Rev.* **2012**, *112* (5), 2889–2919.
- (199) Boeyens, J. C. A.; Comba, P. Molecular Mechanics: Theoretical Basis, Rules, Scope and Limits. *Coord. Chem. Rev.* **2001**, *212* (1), 3–10.
- (200) Comba, P. Interpretation and Prediction of Properties of Transition Metal Coordination Compounds. In *Modeling of Molecular Properties*; Comba, P., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA, 2011; pp 107–121.
- (201) Zimmer, M. Bioinorganic Molecular Mechanics. *Chem. Rev.* **1995**, *95* (8), 2629–2649.
- (202) Bygott, A. M. T.; Sargeson, A. M. Critical Evaluation of Metal Complex Molecular Mechanics. Part 1. Cobalt(III) Hexaamines. *Inorg. Chem.* **1998**, *37* (19), 4795–4806.
- (203) Bartol, J.; Comba, P.; Melter, M.; Zimmer, M. Conformational Searching of Transition Metal Compounds. *J. Comput. Chem.* **1999**, *20* (14), 1549–1558.
- (204) Hay, B. P. Methods for Molecular Mechanics Modeling of Coordination Compounds. *Coord. Chem. Rev.* **1993**, *126* (1–2), 177–236.
- (205) Hu, L.; Ryde, U. Comparison of Methods to Obtain Force-Field Parameters for Metal Sites. *J. Chem. Theory Comput.* **2011**, *7* (8), 2452–2463.
- (206) Norrby, P.-O.; Liljefors, T. Automated Molecular Mechanics Parameterization with Simultaneous Utilization of Experimental and Quantum Mechanical Data. *J. Comput. Chem.* **1998**, *19* (10), 1146–1166.
- (207) Reichert, D. E.; Norrby, P.-O.; Welch, M. J. Molecular Modeling of Bifunctional Chelate Peptide Conjugates. 1. Copper and Indium Parameters for the AMBER Force Field. *Inorg. Chem.* **2001**, *40* (20), 5223–5230.
- (208) Hagelin, H.; Svensson, M.; Åkermark, B.; Norrby, P.-O. Molecular Mechanics (MM3*) Force Field Parameters for Calculations on Palladium Olefin Complexes with Phosphorus Ligands. *Organometallics* **1999**, *18* (22), 4574–4583.
- (209) Bernhardt, P. V.; Comba, P. Molecular Mechanics Calculations of Transition Metal Complexes. *Inorg. Chem.* **1992**, *31* (12), 2638–2644.
- (210) Bernardes, C. E. S.; Canongia Lopes, J. N.; da Piedade, M. E. M. All-Atom Force Field for Molecular Dynamics Simulations on Organotransition Metal

- Solids and Liquids. Application to $M(\text{CO})_n$ ($M = \text{Cr, Fe, Ni, Mo, Ru, or W}$) Compounds. *J. Phys. Chem. A* **2013**, *117* (43), 11107–11113.
- (211) Allured, V. S.; Kelly, C. M.; Landis, C. R. SHAPES Empirical Force Field: New Treatment of Angular Potentials and Its Application to Square-Planar Transition-Metal Complexes. *J. Am. Chem. Soc.* **1991**, *113* (1), 1–12.
- (212) Root, D. M.; Landis, C. R.; Cleveland, T. Valence Bond Concepts Applied to the Molecular Mechanics Description of Molecular Shapes. 1. Application to Nonhypervalent Molecules of the P-Block. *J. Am. Chem. Soc.* **1993**, *115* (10), 4201–4209.
- (213) Landis, C. R.; Cleveland, T.; Firman, T. K. Valence Bond Concepts Applied to the Molecular Mechanics Description of Molecular Shapes. 3. Applications to Transition Metal Alkyls and Hydrides. *J. Am. Chem. Soc.* **1998**, *120* (11), 2641–2649.
- (214) Firman, T. K.; Landis, C. R. Valence Bond Concepts Applied to the Molecular Mechanics Description of Molecular Shapes. 4. Transition Metals with π -Bonds. *J. Am. Chem. Soc.* **2001**, *123* (47), 11728–11742.
- (215) Tubert-Brohman, I.; Schmid, M.; Meuwly, M. Molecular Mechanics Force Field for Octahedral Organometallic Compounds with Inclusion of the Trans Influence. *J. Chem. Theory Comput.* **2009**, *5* (3), 530–539.
- (216) Burton, V. J.; Deeth, R. J.; Kemp, C. M.; Gilbert, P. J. Molecular Mechanics for Coordination Complexes: The Impact of Adding d-Electron Stabilization Energies. *J. Am. Chem. Soc.* **1995**, *117* (32), 8407–8415.
- (217) Deeth, R. J. The Ligand Field Molecular Mechanics Model and the Stereoelectronic Effects of d and s Electrons. *Coord. Chem. Rev.* **2001**, *212* (1), 11–34.
- (218) Deeth, R. J.; Anastasi, A.; Diedrich, C.; Randell, K. Molecular Modelling for Transition Metal Complexes: Dealing with d-Electron Effects. *Coord. Chem. Rev.* **2009**, *253* (5–6), 795–816.
- (219) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179–5197.
- (220) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins†. *J. Phys. Chem. B* **1998**, *102* (18), 3586–3616.
- (221) Mackerell, A. D.; Feig, M.; Brooks, C. L. Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulations. *J. Comput. Chem.* **2004**, *25* (11), 1400–1415.
- (222) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17* (5–6), 490–519.

-
- (223) Handley, C. M.; Deeth, R. J. A Multi-Objective Approach to Force Field Optimization: Structures and Spin State Energetics of d^6 Fe(II) Complexes. *J. Chem. Theory Comput.* **2012**, *8* (1), 194–202.
- (224) Schäffer, C. E.; Jørgensen, C. K. The Angular Overlap Model, an Attempt to Revive the Ligand Field Approaches. *Mol. Phys.* **1965**, *9* (5), 401–412.
- (225) Deeth, R. J.; Fey, N.; Williams–Hubbard, B. DommiMOE: An Implementation of Ligand Field Molecular Mechanics in the Molecular Operating Environment. *J. Comput. Chem.* **2005**, *26* (2), 123–130.
- (226) Chemical Computing Group Inc., Molecular Operating Environment (MOE), Version 2011.10, https://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm (accessed April 2015).
- (227) Smith, W.; Yong, C. W.; Rodger, P. M. DL_POLY: Application to Molecular Simulation. *Mol. Simul.* **2002**, *28* (5), 385–471.
- (228) Carlsson, A. E.; Zapata, S. The Functional Form of Angular Forces around Transition Metal Ions in Biomolecules. *Biophys. J.* **2001**, *81* (1), 1–10.
- (229) Piquemal, J.-P.; Williams-Hubbard, B.; Fey, N.; Deeth, R. J.; Gresh, N.; Giessner-Prettre, C. Inclusion of the Ligand Field Contribution in a Polarizable Molecular Mechanics: SIBFA-LF. *J. Comput. Chem.* **2003**, *24* (16), 1963–1970.
- (230) Woodley, S. M.; Battle, P. D.; Catlow, C. R. A.; Gale, J. D. Development of a New Interatomic Potential for the Modeling of Ligand Field Effects. *J. Phys. Chem. B* **2001**, *105* (29), 6824–6830.
- (231) Xiang, J. Y.; Ponder, J. W. An Angular Overlap Model for Cu(II) Ion in the AMOEBA Polarizable Force Field. *J. Chem. Theory Comput.* **2014**, *10* (1), 298–311.
- (232) Ponder, J. W. TINKER [Online], Version 6.3; <Http://dasher.wustl.edu/tinker/> (accessed March 2014).
- (233) Gütllich, P.; Goodwin, H. A. *Spin Crossover in Transition Metal Compounds I*; Springer Science & Business Media, 2004.
- (234) Bousseksou, A.; Molnár, G.; Demont, P.; Menegotto, J. Observation of a Thermal Hysteresis Loop in the Dielectric Constant of Spin Crossover Complexes: Towards Molecular Memory Devices. *J. Mater. Chem.* **2003**, *13* (9), 2069–2071.
- (235) Kahn, O.; Martinez, C. J. Spin-Transition Polymers: From Molecular Materials Toward Memory Devices. *Science* **1998**, *279* (5347), 44–48.
- (236) Bousseksou, A.; Molnár, G.; Salmon, L.; Nicolazzi, W. Molecular Spin Crossover Phenomenon: Recent Achievements and Prospects. *Chem. Soc. Rev.* **2011**, *40* (6), 3313–3335.
- (237) Gaspar, A. B.; Seredyuk, M. Spin Crossover in Soft Matter. *Coord. Chem. Rev.* **2014**, *268*, 41–58.
- (238) König, E.; Ritter, G.; Kulshreshtha, S. K. The Nature of Spin-State Transitions in Solid Complexes of Iron(II) and the Interpretation of Some Associated Phenomena. *Chem. Rev.* **1985**, *85* (3), 219–234.
- (239) Halcrow, M. A. The Foundation of Modern Spin-Crossover. *Chem. Commun.* **2013**, *49* (93), 10890–10892.
- (240) Gütllich, P. Spin Crossover – Quo Vadis? *Eur. J. Inorg. Chem.* **2013**, *2013* (5-6), 581–591.

- (241) Hauser, A.; Jeftić, J.; Romstedt, H.; Hinek, R.; Spiering, H. Cooperative Phenomena and Light-Induced Bistability in Iron(II) Spin-Crossover Compounds. *Coord. Chem. Rev.* **1999**, *190–192*, 471–491.
- (242) Halcrow, M. A. Spin-Crossover Compounds with Wide Thermal Hysteresis. *Chem. Lett.* **2014**, *43* (8), 1178–1188.
- (243) Brooker, S. Spin Crossover with Thermal Hysteresis: Practicalities and Lessons Learnt. *Chem. Soc. Rev.* **2015**, *44* (10), 2880–2892.
- (244) Halcrow, M. A. The Spin-States and Spin-Transitions of Mononuclear Iron(II) Complexes of Nitrogen-Donor Ligands. *Polyhedron* **2007**, *26* (14), 3523–3576.
- (245) Halcrow, M. A. Structure: function Relationships in Molecular Spin-Crossover Complexes. *Chem. Soc. Rev.* **2011**, *40* (7), 4119–4142.
- (246) Atmani, C.; El Hajj, F.; Benmansour, S.; Marchivie, M.; Triki, S.; Conan, F.; Patinec, V.; Handel, H.; Dupouy, G.; Gómez-García, C. J. Guidelines to Design New Spin Crossover Materials. *Coord. Chem. Rev.* **2010**, *254* (13–14), 1559–1569.
- (247) Deeth, R. J.; Anastasi, A. E.; Wilcockson, M. J. An *In Silico* Design Tool for Fe(II) Spin Crossover and Light-Induced Excited Spin State-Trapped Complexes. *J. Am. Chem. Soc.* **2010**, *132* (20), 6876–6877.
- (248) Foscatto, M.; Deeth, R. J.; Jensen, V. R. Integration of Ligand Field Molecular Mechanics in Tinker. *J. Chem. Inf. Model.* **2015**, Just Accepted, DOI: 10.1021/acs.jcim.5b00098.
- (249) Swart, M. Accurate Spin-State Energies for Iron Complexes. *J. Chem. Theory Comput.* **2008**, *4* (12), 2057–2066.
- (250) Ye, S.; Neese, F. Accurate Modeling of Spin-State Energetics in Spin-Crossover Systems with Modern Density Functional Theory. *Inorg. Chem.* **2010**, *49* (3), 772–774.
- (251) Houghton, B. J.; Deeth, R. J. Spin-State Energetics of Fe^{II} Complexes – The Continuing Voyage Through the Density Functional Minefield. *Eur. J. Inorg. Chem.* **2014**, *2014* (27), 4573–4580.
- (252) Turner, J. W.; Schultz, F. A. Intramolecular and Environmental Contributions to Electrode Half-Reaction Entropies of M(tacn)₂^{3+/2+} (M = Fe, Co, Ni, Ru; Tacn = 1,4,7-Triazacyclononane) Redox Couples. *Inorg. Chem.* **1999**, *38* (2), 358–364.
- (253) Turner, J. W.; Schultz, F. A. Solution Characterization of the Iron(II) Bis(1,4,7-Triazacyclononane) Spin-Equilibrium Reaction. *Inorg. Chem.* **2001**, *40* (20), 5296–5298.
- (254) Dalrymple, S. A.; Shimizu, G. K. H. Second-Sphere Coordination Networks: “Tame-Ing” (Tame=1,1,1-Tris(aminomethyl)ethane) the Hydrogen Bond. *J. Mol. Struct.* **2006**, *796* (1–3), 95–106.
- (255) Dalrymple, S. A.; Shimizu, G. K. H. Crystal Engineering of a Permanently Porous Network Sustained Exclusively by Charge-Assisted Hydrogen Bonds. *J. Am. Chem. Soc.* **2007**, *129* (40), 12114–12116.
- (256) Brown, K. N.; Hockless, D. C. R.; Sargeson, A. M. Synthesis and Electrochemistry of [Pt(tame)₂]⁴⁺: Crystallographic Analysis of bis[1,1,1-Tris(aminomethyl)ethane-N,N']platinum(II) Bis(tetrachlorozincate) Dihydrate. *J. Chem. Soc. Dalton Trans.* **1999**, No. 13, 2171–2176.

-
- (257) Seredyuk, M.; Gaspar, A. B.; Kusz, J.; Gütlich, P. Mononuclear Complexes of Iron(II) Based on Symmetrical Tripodand Ligands: Novel Parent Systems for the Development of New Spin Crossover Metallomesogens. *Z. Für Anorg. Allg. Chem.* **2011**, *637* (7-8), 965–976.
- (258) Panagopoulos, A. M.; Zeller, M.; Becker, D. P. Synthesis of an Ortho-Triazacyclophane: *N,N',N''*-Trimethyltribenzo-1,4,7-Triazacyclononatriene. *J. Org. Chem.* **2010**, *75* (22), 7887–7892.
- (259) Comba, P. Metal Ion Selectivity and Molecular Modeling. *Coord. Chem. Rev.* **1999**, *185–186*, 81–98.
- (260) Hay, B. P.; Zhang, D.; Rustad, J. R. Structural Criteria for the Rational Design of Selective Ligands. 2. Effect of Alkyl Substitution on Metal Ion Complex Stability with Ligands Bearing Ethylene-Bridged Ether Donors. *Inorg. Chem.* **1996**, *35* (9), 2650–2658.
- (261) Hancock, R. D.; Martell, A. E. Ligand Design for Selective Complexation of Metal Ions in Aqueous Solution. *Chem. Rev.* **1989**, *89* (8), 1875–1914.
- (262) Broge, L.; Pretzmann, U.; Jensen, N.; Søtofte, I.; Olsen, C. E.; Springborg, J. Cobalt(II), Nickel(II), Copper(II), and Zinc(II) Complexes with [35]Adamanzane, 1,5,9,13-Tetraazabicyclo[7.7.3]nonadecane, and [(2.3)2.21]Adamanzane, 1,5,9,12-Tetraazabicyclo[7.5.2]hexadecane. *Inorg. Chem.* **2001**, *40* (10), 2323–2334.
- (263) Maldonado, A. G.; Hageman, J. A.; Mastroianni, S.; Rothenberg, G. Backbone Diversity Analysis in Catalyst Design. *Adv. Synth. Catal.* **2009**, *351* (3), 387–396.
- (264) Sauer, W. H. B.; Schwarz, M. K. Molecular Shape Diversity of Combinatorial Libraries: A Prerequisite for Broad Bioactivity. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 987–1003.