universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Predicting Enantiomeric Excess in a Cross-Coupling Reaction with Machine Learning“

verfasst von / submitted by

Nadja Katharina Singer, B.Sc.

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2019 / Vienna, 2019

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 910
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Computational Science
Betreut von / Supervisor:	Dr. Philipp Marquetand, Privatdoz.

ACKNOWLEDGMENTS

First, I want to thank my supervisor Priv.-Doz. Dr. Philipp Marquetand for giving me the possibility to conduct my master studies as well as supporting and encouraging me throughout the time.

Next, I want to thank my collaborators of the group of Prof. Nuno Maulide, in particular Alexander Preinfalk, for the fundamental experimental data and the pleasant cooperation. I would especially like to thank the double agent Dr. Boris Maryasin for patiently enduring all my questions and teaching me his methods.

Furthermore, I want to thank Univ.-Prof. Dr. Dr. h.c. Leticia González and the whole group, including Dr. Markus Oppel and Simon Kropf, who made things work, when they didn't, Dr. Sebastian Mai for his ongoing feedback, the members of the girls room, Sandra Gómez and Julia Westermayr, and everybody else that I had the honor and pleasure of working with.

Additionally, I want to thank my fellow student of the computational science master program Charlotte Bode, Wolfgang Ost, Konrad von Kirchbach, and Jakob Weber for interesting and weird discussions at lunch.

Finally, I want to thank my family and my significant other, Carsten Seyfferth, for their everlasting support and encouragement.

CONTENTS

ACRONYMS

AI	Artificial Intelligence
ASO	Average Steric Occupancy
BO	Bayesian Optimization
BOA	Born-Oppenheimer Approximation
CC	Coupled Cluster
CI	Configuration Interaction
DFT	Density Functional Theory
DFTB	Density Functional Tight Binding
ee	Enantiomeric Excess
ELU	Exponential Linear Unit
er	Enantiomeric Ratio
FF	Force Field
GC	Genetic Z-Matrix Crossing
HB	Hyperband
HF	Hartree-Fock
HPLC	High-Performance Liquid Chromatography
MAE	Mean Absolute Error
MD	Molecular Dynamics
ML	Machine Learning
MM	Molecular Mechanics
MSE	Mean Squared Error
MTD	Meta-Dynamics
PCA	Principal Component Analysis
PES	Potential Energy Surface
QC	Quantum Chemistry
QM	Quantum Mechanics
RBF	Radial Basis Function
ReLU	Rectifier Linear Unit
RFR	Random Forest Regression
RMSD	Root-Mean-Square Deviation
SELU	Scaled Exponential Linear Unit
SQM	Semiempirical Quantum Mechanics

SVR Support Vector Regression

TISE Time-Independent Schrödinger Equation

ee	Enantiomeric Excess
F	Mole Fraction
F_{R}	Mole Fraction of the (R)-Enantiomer
F_{S}	Mole Fraction of the (S)-Enantiomer
a	Area
a_{favored}	Area under the Signal of the Favored Enantiomer
$a_{\mathrm{disfavored}}$	Area under the Signal of the Disfavored Enantiomer
k	Reaction Rate Constant
k_{favored}	Reaction Rate Constant of the Favored Enantiomer
$k_{\mathrm{disfavored}}$	Reaction Rate Constant of the Disfavored Enantiomer
$A(T)$	Pre-Exponential Factor
$\Delta G^{\neq}$	Activation Gibbs Free Energy
R	Gas Constant
T	Temperature
$\delta\Delta G^{\neq}$	Activation Gibbs Free Energy Difference
$\Delta G^{\neq}_{\mathrm{favored}}$	Activation Gibbs Free Energy of the Favored Enantiomer
$\Delta G^{\neq}_{\mathrm{disfavored}}$	Activation Gibbs Free Energy of the Disfavored Enantiomer
$G^{\mathrm{TS}}_{\mathrm{favored}}$	Transition State Free Energy of the Favored Enantiomer
$G^{\mathrm{TS}}_{\mathrm{disfavored}}$	Transition State Free Energy of the Disfavored Enantiomer
ΔG^{TS}	Transition State Free Energy Difference
$\hat{H}$	Hamiltonian Operator
Ψ	Total Non-Relativistic Wavefunction
$\vec{R}$	Nuclei Coordinate Vector
$\vec{r}$	Electron Coordinate Vector
$\hat{H}_{\mathrm{el}}$	Electronic Hamiltonian Operator
Ψ_{el}	Electronic Time-Independent Wavefunction
$\bar{R}$	Parametric Dependence on the Nuclei Coordinate Vector
E_{el}	Electronic Energy
N_{el}	Number of Electrons
E_{HF}	Hartree-Fock Energy
Z	Atomic Number
E_{PES}	Potential Energy Surface
V_{nuc}	Nuclear Repulsion Potential
W	Wigner Distribution
x	Coordinates of the Normal Mode
p	Momentum of the Normal Mode
$\hbar$	Reduced Planck Constant
N_{nuc}	Number of Atoms

w_i	Unnormalized Wigner Distribution of Mode i
μ	Reduced Mass
η	Angular Frequency
i	Normal Mode
P_i	Random Number in [-5,5]
Q_i	Random Number in [-5,5]
R_i	Atomic Coordinates
ν_i	Vibrational Frequencies
m_i	Normal Mode Vector
E_{tot}	Total Energy
E_{bias}	Biasing Potential
$E_{\text{bias}}^{\text{RMSD}}$	Biasing Potential Based on the RMSD
n	Number of Reference Structures
κ	Pushing Strength
α	Width of the Gaussian Potential
Δ	Collective Variable
r_j	Component of the Cartesian Space Vector of j
MSE	Mean Squared Error
y_i	Observable
$\tilde{y}_i$	Prediction
N	Number of Samples
MAE	Mean Absolute Error
R^2	Regression Score Function
$\bar{y}_i$	Mean Observable
l	Layer
$n^{(l)}$	Number of Nodes in Layer l
$y_\beta^{(l)}$	Value of the Node β in Layer l
$\sigma^{(l)}$	Activation Function of Layer l
Σ	Weighted Sum
$\Theta_{\alpha\beta}^{(l)}$	Weight of the Connection between Node α of Layer $(l-1)$ and Node β of Layer l
$\mathbf{P}$	(Hyper)Plane
δ	Margin
ϵ	Distance from Plane $\mathbf{P}$ to Margin Border
K	Kernel Function
$\varphi(y)$	Transformed Data Point y
$\mathcal{V}$	Inner Product Space
C	Penalty Parameter of Error Term

1 INTRODUCTION

Stereochemistry is the study of molecules that only differ in the three-dimensional orientation of their atoms in space, but not in their constitution. A special focus in stereochemistry lies in chiral molecules, which are molecules that are non-superposable on their mirror images, also called enantiomers. An example found in nature are D- and L-amino acids, where the latter are the prevailing form.[1] The importance of stereochemistry became especially apparent to the public with the tremendous Contergan scandal in the late 1950s and early 1960s. While the (S)-enantiomer of Thalidomide, the active substance of the drug Contergan, is bioactive, the (R)-form is teratogenic, which lead to thousands of children born with deformities.[2,3] This is just one example of two enantiomers with critically different effects as drugs. Still today the significance of pharmacodynamic[a] and -kinetic[b] differences[4,5] between the enantiomers of chiral drugs is a crucial part of research.[6]

Naturally, from this knowledge about the differences between enantiomers, a need for efficient synthetic methods for stereoselective reactions building complex organic molecules evolved.

An important method for preparing complex organic molecules, like drugs, are cross-coupling reactions. Especially palladium-catalyzed reactions have proven to be an irreplaceable tool in the C-C bond forming processes. The pioneer work of Richard Heck,[7–9] Ei-ichi Negishi,[10,11] and Akira Suzuki[12–14] was even rewarded with the Nobel Prize in Chemistry in 2010. Their work in the early 1970s paved the way for many related reactions like Corriu-Kumada,[15,16] Stille,[17–19] and Hiyama[20,21] cross-couplings to just name a few.[22] Today cross-coupling reactions are still evolving and increasing their synthetic potential. One reason is that they were found to be a useful tool also in stereoselective C-C bond forming reactions and with that they are of interest for pharmaceutical research. The preparation of chiral cross-coupling products was already shown by Hayashi, Kumada, and co-workers in the 1970s and 80s.[23–28] Recently Thaler and co-workers reported highly diastereoselective Negishi cross-couplings on cycloalkylzinc reagents with 1,2-, 1,3-, and 1,4-stereocontrol.[29]

As state of today our collaborators, the Maulide group at the University of Vienna, work on maximizing the enantiomeric excess (*ee*) of Negishi cross-coupling products.[30] The Negishi reaction couples organic halides or triflates with organozinc compounds using palladium or nickel catalysts. The products of the Negishi reactions executed by the Maulide group only have a single center of chirality and therefore only have two possible enantiomers, the (R)- or (S)-enantiomer. The *ee* describes the relation of these two enantiomers relative to each other and can therefore be used as a measure of success. The *ee* is one if the product is enantiopure and zero if the product is a racemate. Consequently maximizing the *ee* is the target of many studies. The maximization

a Pharmacodynamics is the study of pharmacological actions of drugs on living systems, like reactions with and binding to cell constituents.[4,5]

b Pharmacokinetics is the study of the fate of a drug inside the metabolism from uptake to transformation to elimination.[4,5]

of the *ee* is achieved by varying the ligands of the palladium catalyst, among other reaction parameters.

Trying out different reactants and synthesizing hundreds of compounds for reaction optimization is of course not the most effective way in the digital age. Particularly, there are reams of modern quantum chemistry (QC) software packages that can numerically determine properties of materials and molecules. Unfortunately, when it comes to large-scale screening the computational costs are still the limiting factor for quantum calculations. However in the time of big data and artificial intelligence (AI), machine learning (ML) methods can provide accurate predictions of chemical properties at a significantly reduced computational effort. That is why there are various examples of successful integration of ML in chemistry.

The study of Gómez-Bombarelli *et al.*[31] is an excellent example of how a workflow can be established to combine high-throughput virtual screening with an experimental approach to find novel molecules. This virtual screening uses modern ML methods and QC calculations to explore molecules outside of the known molecular space. Further, an example for the usage of AI in chemistry and additionally in combination with cross-coupling reactions is provided by Doyle and Dreher.[32] They showed how the data of nanomole-scale high-throughput robot experiments can be used to train an AI algorithm, in this case a random forest model, to accurately predict the yields of Buchwald-Hartwig couplings. While they needed to complete their experimental data with rather expensive QC calculations, the work of Lilienfeld and Corminboeuf[33] relates easily accessible descriptor variables directly to catalytic performance. By developing Volcano plots they were able to identify the most promising, thermodynamically attractive, and readily available catalyst candidates. Finally, there is to mention the work of Denmark and co-workers[34] who were able to predict higher-selectivity catalysts by a computer-driven workflow and ML. They rank chiral phosphoric acid-catalysts for thiol additions to N-acylimines by the *ee* value. Compared to the yield of an experiment, which was used as the observable by Doyle and Dreher,[32] the *ee* can be reproduced synthetically more reliably, which makes accurate predictions even more valuable. The group of Denmark was able to predict the *ee* of coupling reactions for unknown reactants, unknown catalyst, and even *ee* values beyond the range of observed values in the training set. It should not go unmentioned here that the catalysts used by Denmark and co-workers[34] for the reaction are very similar, containing the same core structure, and only differing in substituents.

All the mentioned works mark an important contribution to modern chemistry in conjunction with AI. Nevertheless, there is still room for improvement, especially in regards to the choice of ML models and their hyperparameter optimization as well as the used descriptors and the generalizability of the workflow and model.

The aim of this thesis is to develop a ML workflow which can efficiently predict the *ee* of palladium-catalyzed Negishi couplings. The underlying reaction of organic bromine reactants with organozinc compounds using palladium catalysts to form chiral products is pictured in Figure 1.1.

In this thesis, a ML model is developed using experimental data. The experimental data as a basis for the ML model was provided by the Maulide group (see Figures A.1 to A.7 and Table A.1 in the *Appendix*).[30] The goal of

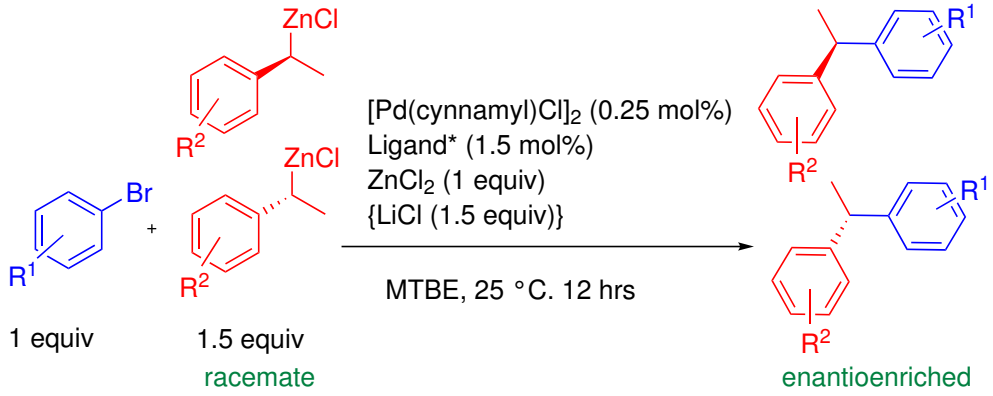


Figure 1.1: Basic reaction scheme for the Negishi cross-coupling of interest. The reaction is a coupling of a bromine reactant and a racemic organozinc compound with the goal to obtain an enantioenriched product. The catalyst used is based on palladium with different ligands. Zinc chloride was used as additive, as well as lithium chloride in some cases. The solvent is methyl *tert*-butyl ether (MTBE).

the model is to connect the structure and other properties of the reactants, obtained from experiments and calculations, to the *ee*. Therefore, the collected data is transformed to a machine readable format and used to train different ML models with the *ee* as observable. The present work is a first step in developing a high throughput screening workflow for novel reactants and catalysts in the mentioned Negishi cross-coupling reaction yielding chiral products (Fig. 1.1). It is meant to be used as an efficient prescreening tool to provide promising reactants that will be validated first by QC computations and finally by synthesis in the future.

This work is, to the author's knowledge, the first to predict the *ee* of Negishi coupling reactions. It tries to use the idea behind the work of Denmark and co-workers [34] for a different reaction with a wider structural variety of catalysts. The variety of the catalyst ligands can be seen in Figures A.1 to A.5 in the *Appendix*.

An overview over the theory of all relevant topics regarding organic and theoretical chemistry as well as ML is given in Chapter 2. Chapter 3 explains the developed workflow from experimental data to descriptor calculation and finally to the prediction of the *ee* for unknown catalysts. It also includes the computational details. In Chapter 4 the Average Steric Occupancy (ASO) descriptor is closely inspected and different ML models are validated and compared. Additionally, predictions for the ee of unknown reactions are made. The work is summarized in Chapter 5 and an outlook is given.

2 BACKGROUND

The following chapter provides an overview over the theory relevant for this thesis. At the beginning, a brief summary of the organic background important for the work is given, including the Negishi cross-coupling reaction and the *ee*. It is followed by a description of the methods used to optimize molecular structures and to generate a large conformer-rotamer-ensemble with the ultimate goal of calculating a suitable descriptor. Furthermore, this descriptor as well as the ML methods used to predict the *ee* are explained.

2.1 Organic Chemistry Background

This section aims to summarize the general mechanism of the Negishi cross-coupling reaction and explain the formulas underlying the *ee*.

2.1.1 Negishi Cross-Coupling

As mentioned before in Chapter 1, the Negishi cross-coupling reaction is an important C-C bond forming process for complex organic molecules and is additionally applicable to chiral synthesis. The basic reaction scheme is shown in Figure 2.1. The C-C coupling process can be applied to every possible combination of carbon atom types (sp, sp^2, or sp^3) and tolerates many different functional groups at the reagents. The reaction is mostly performed with palladium catalysts, but also nickel based catalysts are possible.[35]

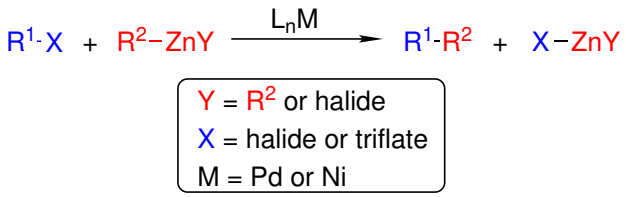


Figure 2.1: General scheme of the Negishi reaction, where the two organic groups R^1 and R^2 are coupled together.

The general catalytic mechanism is shown in Figure 2.2. The first step of the reaction is the oxidative addition, where the bond between the organic group R^1 and the heteroatom X breaks and two new bonds are formed with the metal M, increasing its oxidation state by two units. The second step is the transmetalation, which is the characteristic step of the reaction. The second organic group R^2 needs to be transferred from the zinc atom to the metal M. The exact mechanism of this transmetalation is part of recent research in experiment and theory with multiple possible proposed pathways.[36–39] The last step of the catalytic cycle is the reductive elimination, where the two organic groups R^1 and R^2 leave the metal M to form a common bond. By leaving the metal M its oxidation state is reduced by two units, coming back to its initial state.[35]

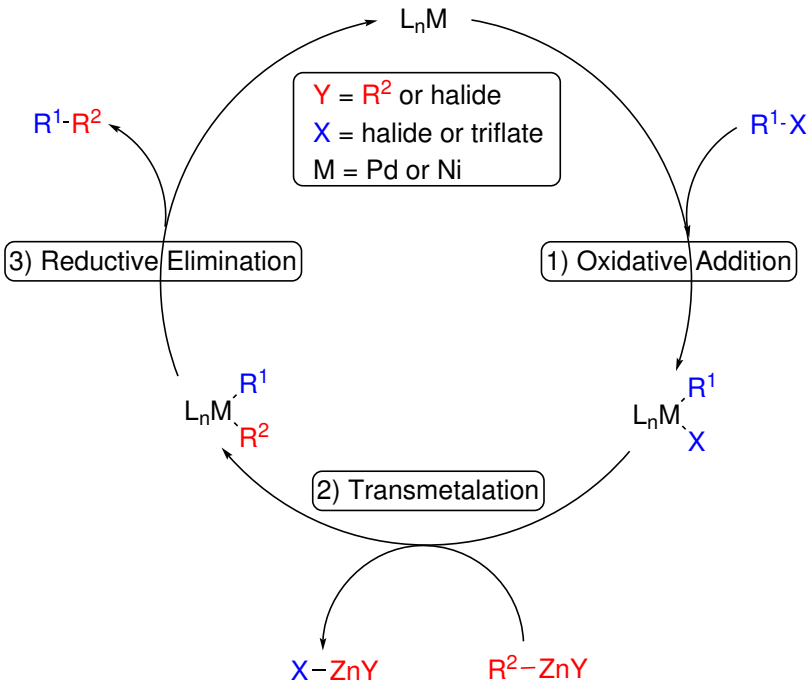


Figure 2.2: General catalytic cycle of the Negishi coupling, with R being sp, sp^2, or sp^3 carbon type organic compounds.

2.1.2 Enantiomeric Excess

The importance of enantiomers and the *ee* was already discussed in Chapter 1, while this section aims to define the underlying formulas and link the *ee* to theoretical chemistry. The *ee* is generally defined by the absolute difference between the mole fraction F of the (R)- and the (S)-enantiomer as

$$\text{ee} = |F_\text{R} - F_\text{S}| \quad \text{with} \quad F_\text{R} + F_\text{S} = 1\,. \tag{2.1}$$

In practice, it is often used as a percent *ee* and expressed by

$$\%\text{ee} = (|F_\text{R} - F_\text{S}| \times 100)\,. \tag{2.2}$$

Therefore, a chiral substance with 100 % *ee* is called enantiopure and only contains one of the two possible enantiomers. A chiral substance with 0 % *ee* is called a racemate and contains a 50:50 mixture of both enantiomers. Everything in between 0–100 % *ee* is called enantioenriched. In experiment, the *ee* is mostly measured by chiral high-performance liquid chromatography (HPLC), where the *ee* can be expressed using the area under the signals a.[a]

$$\text{ee} = \frac{a_\text{favored} - a_\text{disfavored}}{a_\text{favored} + a_\text{disfavored}} \tag{2.3}$$

To link the *ee* to theoretical chemistry it can also be expressed in terms of reaction rate constants k leading to the favored and disfavored enantiomer.[40]

$$\text{ee} = \frac{k_\text{favored} - k_\text{disfavored}}{k_\text{favored} + k_\text{disfavored}} \tag{2.4}$$

a This is how the experimental *ee* was determined for this work.[30]

Via transition state theory the rate constant k can be expressed as

$$k = A(T)e^{-\Delta G^{\neq}/RT} \tag{2.5}$$

with $\Delta G^{\neq}$ being the activation Gibbs free energy, R the gas constant, T the temperature, and $A(T)$ the pre-exponential factor, which is assumed to be equivalent for both enantiomeric pathways. The substitution of Equation (2.5) into Equation (2.4) leads to

$$\text{ee} = \frac{e^{-\delta\Delta G^{\neq}/RT} - 1}{e^{-\delta\Delta G^{\neq}/RT} + 1} \quad \text{with} \quad \delta\Delta G^{\neq} = \Delta G^{\neq}_{\text{favored}} - \Delta G^{\neq}_{\text{disfavored}}. \tag{2.6}$$

This equation links the *ee* with ΔG^{TS}, the difference in transition state free energies of the enantiomers, by

$$\delta\Delta G^{\neq} = G^{\text{TS}}_{\text{favored}} - G^{\text{TS}}_{\text{disfavored}} = \Delta G^{\text{TS}}, \tag{2.7}$$

which can be done because the reactants that lead to the transition state are identical and therefore have identical free energy. Rearrangement of Equation (2.6) with (2.7) gives an expression to compute ΔG^{TS} from the *ee* and vice versa.

$$\Delta G^{\text{TS}} = RT \ln\left(\frac{1+\text{ee}}{1-\text{ee}}\right) \tag{2.8}$$

This shows that from the transition state energies of a reaction the *ee* is theoretically computable if the transition state is known.[40]

2.2 Theoretical Chemistry

Since the formulation of the Schrödinger equation in 1926, its exact solution can only be calculated for two particle systems. Therefore, approximations have to be proposed for larger systems. The most common one is the Born-Oppenheimer approximation[41] (BOA). It assumes that the motions of the nuclei and the electrons can be separated due to the huge difference between the masses and therefore also the velocities of both particles. This allows to separate the Hamiltonian $\hat{H}$ and the total wave function $\Psi(\vec{R},\vec{r})$ of the time-independent Schrödinger equation (TISE) into nuclear and electronic parts, leading to the definition of the electronic TISE as

$$\hat{H}_{\text{el}}\Psi_{\text{el}}(\bar{R},\vec{r}) \quad = \quad E_{\text{el}}(\bar{R})\Psi_{\text{el}}(\bar{R},\vec{r}). \tag{2.9}$$

The remaining task is to solve this electronic TISE for N_{el}-electron systems, which is a difficult problem as the corresponding wave function $\Psi_{\text{el}}(\bar{R},\vec{r})$ is a high-dimensional object. There is a range of different methods tackling this challenge and trying to balance accuracy versus computational cost. In the Hartree-Fock (HF) approach the N_{el}-electron (many-body) wave function is approximated by an antisymmetrized product of N_{el} one-electron wave functions, usually referred to as a Slater determinant. However this approximation leads to an energy E_{HF} always larger than the energy of the exact solution within the BOA. The difference between these energies is called correlation energy. Post-HF methods like Møller-Plesset perturbation theory, coupled cluster (CC), configuration interaction (CI), and quadratic CI try to approximate the correlation energy. In principle the exact wave functions and energies

of all states could be obtained by some of the mentioned techniques, but approximations have to be made due to unfavorable computational scaling.[42] Another approach tackling the many-body problem is the Quantum Monte Carlo method, which uses statistical techniques to approximately solve the problem.[42] A different ab initio method is density functional theory (DFT). This approach maps the many-body problem to a single-body problem, where the electron density instead of a wave function is used to obtain information about the chemical systems. The electron density is an observable, can be measured experimentally, and is defined as the probability of finding any of the N_{el} electrons within a defined volume element.[43] Aside from the alluded quantum mechanical (QM) methods there is also a range of very fast molecular mechanics (MM) methods to determine the energy and other properties of especially large systems. One of the approaches is to use empirical force fields (FFs), which are empirical potentials, consisting of a large number of parameters and functions referring, e.g., to electrostatic interactions, to van der Waals terms, and to stretching, bending, and torsional forces.[42] However, the fast purely classical MM treatment has its limits and the more accurate QM methodology is not always applicable. A solution for the dilemma are various semiempirical methods, which approximate integrals of the ab initio QM methods by experimental or MM data. One of these methods is GFN*n*-xTB.[44]

2.2.1 GFN*n*-xTB

GFN*n*-xTB[44] (Geometry, Frequency, Noncovalent, Extended Tight Binding of version *n*) is a semiempirical tight binding quantum chemical method primarily designed to yield fast and reasonable geometries, vibrational frequencies and noncovalent interactions especially for systems with a large number of atoms, including elements up to radon ($Z = 86$). Semiemprical quantum mechanical (SQM) methods in general try to provide an alternative route between costly ab initio QM methods like HF or DFT, and the effective FF methods, which treat atoms in a classical mechanics way. These SQM methods are faster by at least 2 orders of magnitude than QM methods due to drastic integral approximations.[45] The GFN*n*-xTB method is related to the self-consistent density functional tight binding (DFTB) scheme, avoids element-pair-specific parameters, and is parametrized using reference data at the hybrid DFT level. Variant 2 of the algorithm, GFN2-xTB,[46] additionally includes multiple electrostatics and density-dependent dispersion contributions.

2.2.2 Conformer Search

The electronic TISE using the BOA (Eq. (2.9)) can be used to calculate the energy of a specific electronic state with fixed nuclei. The potential energy $E_{\mathrm{PES}}(\bar{R})$ used to create potential energy surfaces (PESs) is the sum of the nuclear repulsion $V_{\mathrm{nuc}}(\bar{R})$ and the electronic energy $E_{\mathrm{el}}(\bar{R})$.

$$E_{\mathrm{PES}}(\bar{R}) = V_{\mathrm{nuc}}(\bar{R}) + E_{\mathrm{el}}(\bar{R}) \tag{2.10}$$

By solving the electronic TISE pointwise for each nuclear configuration separately and adding nuclear repulsion, a PES is generated. Knowledge over the PES enables the calculation of free energy barriers from which in turn

the *ee* can be obtained (see Eq. 2.6). Pointwise evaluation of energy surfaces is not a very elegant way to determine equilibrium geometries, vibrational frequencies, and other data related to nuclear coordinates, but many of these procedures have been automated by algorithms.[42] Oftentimes only the global minimum of the PES, and therefore the energetically most favorable structure, is of interest in QC. However, one goal in this workflow is to also find conformers and rotamers, which are described by local minima on the PES. Rotamers are structures with very similar total energies that only differ in the orientation around a single bond, while conformers differ by more rotations leading to different total energies. There are various different methods for conformer search, which all have to make compromises in terms of accuracy versus computational cost. The following describes two techniques that are used in this work to create a library of geometries for each molecule.

Wigner Sampling

The Wigner quasiprobability distribution or Wigner-Ville distribution links the Schrödinger wave function to a probability distribution in phase space.[47] The basic idea is to derive an expression that gives a statistical description for the atomistic coordinates and corresponding momenta of quantum systems. The equation for the Wigner distribution W based on uncoupled harmonic oscillators of the normal modes i is given by

$$W(x,p) = \frac{1}{(\pi\hbar)^{3N_{\mathrm{nu}}-6}} \prod_{i}^{3N_{\mathrm{nu}}-6} w_i \quad \text{with} \quad w_i = e^{-\left(\frac{2\mu_i\eta_i x_i^2}{\hbar}\right)} e^{-\left(\frac{2p_i^2}{\hbar\mu_i\eta_i}\right)} \tag{2.11}$$

with x_i being the coordinates corresponding to the normal mode i, p_i the momentum, μ_i the reduced mass, and η_i the angular frequency. The implementation in SHARC[48] generates geometries by sampling a simplified quantum-harmonic Wigner distribution. For each normal mode i, two random numbers P_i and Q_i are chosen uniformly from the interval [-5,5] [b] [c] to calculate the ground state quantum Wigner distribution w_i with

$$w_i = e^{-(P_i^2+Q_i^2)}. \tag{2.12}$$

If w_i is larger than a uniform random number from [0,1], then P_i and Q_i are accepted and the coordinates are updated by

$$R_i = R_{i-1} + \frac{Q_i}{\sqrt{2\nu_i}} m_i\,, \tag{2.13}$$

where m_i are the normal mode vectors and ν_i the vibrational frequencies.[49,50] The geometries generated by this algorithm are not conformers. The structures are not optimized, therefore aren't local minima of the PES, and can even be unphysical, because of the displacement simply along normal modes. Nevertheless, they can be used as a very fast and easy initial guess for the steric environment of a single conformer. The Wigner sampling could also be used for an already obtained conformer-ensemble to sample geometries for every conformer.

b Unit: Dimension-less mass and frequency weighted normal mode coordinates.

c In theory the interval $[-\infty,\infty]$ of the Gaussian shaped Wigner distribution should be sampled, but in practice interval [-5,5] is chosen to increase the probability of w_i getting accepted and to ensure the harmonic oscillator approximation holds.

iMTD-GC - Iterative Meta-dynamics Sampling and Genetic Z-matrix Crossing

The iMTD-GC workflow is an iterative algorithm for generating conformer-rotamer libraries by combining meta-dynamics (MTD) with molecular dynamics (MD), and genetic Z-matrix crossing (GC) based on GFN*n*-xTB calculations.[51–53]

First, MTD is used as a powerful method for efficiently exploring the PES, and therefore the conformer space, by adding a biasing potential E_{bias} to the PES.

$$E_{\text{tot}} = E_{\text{PES}} + E_{\text{bias}} \tag{2.14}$$

This biasing potential is the core of MTD. It is a history-dependent potential that fills the minima of the PES over time. This enables the algorithm to overcome even large reaction barriers by discouraging the system from revisiting the same spots on the PES. The used biasing potential is a Gaussian potential based on the standard root-mean-square deviation (RMSD) in Cartesian space given by

$$E_{\text{bias}}^{\text{RMSD}} = \sum_{i=1}^{n} \kappa_i e^{-\alpha \Delta_i^2} \quad \text{with} \quad \Delta_i = \sqrt{\frac{1}{N_{\text{nuc}}} \sum_{j=1}^{N_{\text{nuc}}} \left(r_j - r_j^{\text{ref},i}\right)^2} \tag{2.15}$$

with Δ being the collective variable RMSD, n the number of reference structures associated with the pushing strength κ, α the width of the biasing potential, N_{nuc} the number of atoms, r_j a component of the Cartesian space vector of the actual molecule, and $r_j^{\text{ref},i}$ the corresponding element in the reference structure i.

Second, MD is added to the algorithm to sample low lying conformers more extensively to find conformers within small energy barriers that might have been overshot by the MTD algorithm. MD is a technique to describe the dynamical evolution of a molecular system by numerically solving Newton's equations of motion.

Finally, the GC algorithm is especially important for the generation of rotamers. It creates new structures by projecting structural elements present in already generated structures onto reference structures. Responsible for the genetic character is the fact that frequent structural elements are inherited more often than others.

This iMTD-GC workflow provides an extensive conformer-rotamer-ensemble for every molecule, which will be important for the descriptor presented in Section 2.3.3.

2.3 Machine Learning

ML models are algorithms that build a mathematical model based on sample data (training data) to perform a task without being explicitly programmed for it. Generally there are two types of ML: supervised and unsupervised. Supervised models are trained with training data including the wanted output (observable), while unsupervised models are able to recognize patterns in the data without any prior knowledge and observable. The training process of ML models utilizes some kind of loss function. Loss or cost functions, are

functions that represent a 'cost' associated with the problem, e.g., the error, and are minimized during the optimization. The loss function used for the neural network is the mean squared error (MSE)

$$\mathrm{MSE} = \frac{\sum_{i=1}^{N}(y_i - \tilde{y}_i)^2}{N}, \tag{2.16}$$

but various other functions are common, like mean absolute error (MAE), L2, L1, and mean absolute percentage or logarithmic error.

In chemistry, the aim is usually to transform the chemical knowledge about atoms, molecules, or reactions into something machine readable using descriptors and then predict certain characteristics from it. The research areas include, e.g., prediction of the wave function, energies, energy surfaces, or other chemical properties, like the *ee*.

2.3.1 **Model Selection**

This work focuses on supervised ML models that are suitable for regression problems. In the following, neural networks, support vector regression, and random forest regression are briefly discussed. The models are evaluated by the MAE

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N}(y_i - \tilde{y}_i)}{N} \tag{2.17}$$

and by the regression score function R^2

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \tilde{y}_i)^2}{\sum_{i}^{N}(y_i - \bar{y})^2} \tag{2.18}$$

with y_i being the observed value, $\bar{y}$ being their mean, $\tilde{y}$ being the predicted value, and N being the number of samples.

Neural Networks

The focus of the following explanation is on deep feed-forward neural networks (NNs), which are used in this work.[54] Other architectures of NNs are, for example, convolutional or recurrent NNs. Neural networks are in general structured into layers l (see Fig. 2.3) consisting of neurons, or nodes, $n^{(l)}$ interconnected by weighted edges $\mathbf{\Theta}^{(l)}_{\alpha\beta}$. In feed-forward NNs the initial signal $\mathbf{y}^{(0)}$ is only transmitted in one direction through the NN, from input to output, by recursively applying the relation:

$$y^{(l)}_{\beta} = \sigma^{(l)}\Big(\Sigma\Big) \quad \text{with} \quad \Sigma = \mathbf{\Theta}^{(l)}_{0\beta} + \sum_{\alpha}^{n^{(l-1)}} \mathbf{\Theta}^{(l)}_{\alpha\beta} y^{(l-1)}_{\alpha}. \tag{2.19}$$

The first layer in these networks is the input layer $l = 0$, which collects all the data provided by the user. In the case of chemistry, this can, e.g., be any kind of information about atoms, molecules, or reactions. The kind of input used in this work will be further explained in Section 2.3.3. The input layer has a number of nodes suitable for the input data (e.g. Fig. 2.3: 3 nodes) and

is followed by an arbitrary number of hidden layers (e.g. Fig. 2.3: 2 hidden layers) with an arbitrary number of nodes per layer (e.g. Fig. 2.3: 3 and 4 nodes per hidden layer). Each node in the network is fully connected to the nodes of the previous and following layer. Every one of these connections between two nodes α and β, is performed by a weight $\mathbf{\Theta}^{(l)}_{\alpha\beta}$. These weights are initialized randomly and have to be adjusted during the NN training, e.g., with backpropagation. The backpropagation algorithm optimizes the weights from output back to input by calculating the gradient of the loss function. A node can just represent the weighted sum Σ of the previous layer or can be enhanced by an activation function σ (e.g. Fig. 2.3: σ for hidden layer 2). Activation functions can in general be any kind of function acting on the weighted sum Σ, but the most widely used ones are rectifier, exponential and scaled exponential linear units (ReLU, ELU, and SELU), softplus, and sigmoid, to just name a few. Finally, the output layer is built to fit the number of nodes suitable for the observable (e.g. Fig. 2.3: 2 nodes) and can also contain an activation function. Additionally, there is the possibility to introduce bias nodes $y^{(l)}_0$ that provide an adjustable offset to the layers. The only parameters of NNs that have to be known prior to training are called hyperparameters and control the training process. They include for example the choice of activation function σ, number of layers l, and number of nodes in each layer $n^{(l)}$.

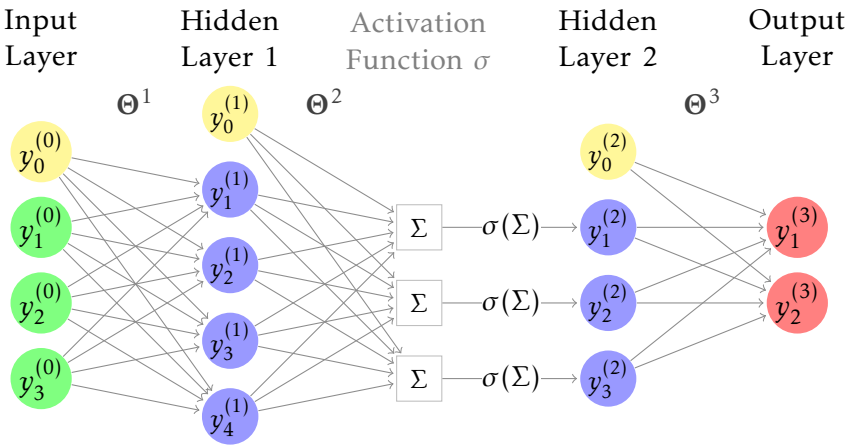


Figure 2.3: Example of a feed-forward neural network containing:
1) an input layer $\mathbf{y}^{(0)}$ with 3 nodes and 1 bias node $y^{(0)}_0$
2) a hidden layer $\mathbf{y}^{(1)}$ with 4 nodes and 1 bias node $y^{(1)}_0$
3) another hidden layer $\mathbf{y}^{(2)}$ with an activation function σ acting on the weighted sum Σ, 3 nodes, and 1 bias node $y^{(2)}_0$
4) an output layer $\mathbf{y}^{(3)}$ with 2 nodes.
An activation function is exemplarily added to hidden layer 2, but could in theory additionally be added to hidden layer 1 and the output layer.

Support Vector Regression

Support vector regression (SVR)[55] originates from the support vector machine (SVM)[56] used for solving classification problems (see Fig. 2.4). The idea is to separate two classes (filled and unfilled circles) by fitting a plane[d] **P** between

d Plane or hyperplane, depending on the dimensionality of the problem.

them. This plane is defined by the maximum distances between the plane and the nearest member of each class, also called support vector (marked in red). In other words, a plane that separates the classes with a maximal margin $\delta = [-\epsilon, +\epsilon]$ is wanted. Even classes that are not linearly separable can be separated using the kernel trick.[57] The trick uses kernel functions K, which can be expressed as inner product in another space $\mathcal{V}$.

$$K(y_i, y_j) = \langle \varphi(y_i), \varphi(y_j) \rangle_{\mathcal{V}} \tag{2.20}$$

This simplifies the complexity of the problem, because only the computationally cheaper inner product $\langle ., . \rangle_{\mathcal{V}}$ of the transformed data points $\varphi(y)$ has to be known, but not the actual higher dimensional coordinates of the data. It enables the method to operate in a high-dimensional, implicit feature space, where the classes are separable, without ever computing in this dimension.

To solve regression problems, instead of separating classes with a plane, the SVR method tries to find a plane such that hopefully all (kernel transformed) data points lie on the plane or at least within the margin δ. All data points outside the margin increase the error depending on the distance to the margin border and the penalty parameter of the error term C. C is the extent of punishment and therefore a tradeoff between the algorithm complexity and number of samples outside of the margin. When C is small the empirical error of the original data is small and the complexity of the model is low, but with a risk of underfitting. When the C value is large, the computational complexity is high and the risk of overfitting increases. Therefore, it is very important to choose the appropriate penalty parameter of the error term.[58]

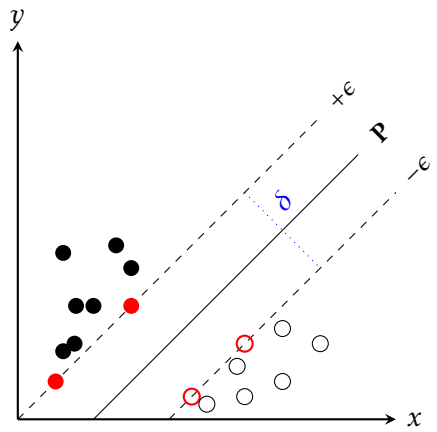


Figure 2.4: Two classes, the filled and unfilled circles, are separated by a plane **P** determined by the support vectors, or nearest neighbors, marked in red. The margin δ is defined by the interval $[-\epsilon, +\epsilon]$.

Random Forest Regression

Random forests or random decision forests operate by constructing a multitude of decision trees and outputting the mean prediction over all trees.[59] A decision tree can build regression or classification models. For that it breaks down the input data incrementally into smaller subsets, while at the same time building the tree structure. The tree starts with the root node, which corresponds to the best predictor, because it contains the whole set of data. This

root node is then split by taking the (categorical or numerical) feature with the highest variance and splitting it into two (or multiple) sets of data. The split decision is made such that the resulting subsets have minimal variance. These nodes are called decision nodes. The decision nodes are then incrementally split until the number of samples after a split is below a certain threshold. These final nodes are called leaf nodes. To predict an observable from an unknown sample, the sample is processed through the tree, according to the defined splits, until it reaches a leaf node. The average value of the samples of this leaf node is then the predicted value. In random forest regression (RFR) models each individual tree, of the multitude of trees made, learns from a random subset of the training data, which can also include multiple of the same training data points. The other important concept of RFR is that not all features are taken into account for a split, but only the variances of a random subset of features is the decisive factor. The RFR model in this work uses the MSE as the function to measure the quality of a split.

2.3.2 **Hyperparameter Optimization**

Every one of the prior mentioned ML models has some kind of hyperparameter controlling the training. These hyperparameters have to be chosen carefully and have an important influence on the performance of a model. The traditional way of performing hyperparameter optimization is called grid search. It basically means trying out every possible given combination of hyperparameters and comparing the performance. This is exhaustive and suffers from the curse of dimensionality. Although the evaluated configurations are typically independent of each other and can be computed in parallel, the computation can still be too demanding, depending on the number of possible configurations. A way to reduce the computational effort is to randomly select some configurations out of all possible configurations. This random search performs well if only a small number of hyperparameters affect the final performance. A third method is Bayesian optimization (BO). Since the objective function that optimizes the hyperparameters is unknown, the BO method builds a random model for the objective. The model is updated by iteratively evaluating only promising hyperparameter configurations. The promising configurations are determined by the first initial random model. This advanced evaluation for the objective function generation can still be too expensive in some cases. That's why Hyperband[60] (HB) was invented. The method defines cheap-to-evaluate approximate versions of the objective function. For the maximum budget (e.g. amount of computational time) the approximation equals the best possible solution, while for smaller budget the quality decreases. This makes it possible to evaluate randomly sampled configurations first on a small budget and evaluate only the most promising configurations on maximum budget.

BOHB algorithm

The BOHB[61] algorithm combines Bayesian optimization with Hyperband. It relies on HB to determine how many configurations to evaluate with which budget, but replaces the initial random selection of configurations. It is re-

placed by a model that uses BO to select, which new configurations to evaluate next, based on the already evaluated configurations and their performance.

2.3.3 **Descriptors**

As already mentioned before, the input data for ML, of course, needs to be machine readable. In case of chemistry this means one needs to find chemical and physical properties, or a number representation of chemical structures to describe the most important features of the system (atom, molecule, or reaction). These representations are called descriptors and their choice is a crucial part in a computational chemistry workflow. The descriptors need to include all the information necessary to answer a given question. In the case of chemistry data there are currently different molecular descriptors in use, regarding structure, charge density, vibronic frequencies, or reactivity, to just name a few. There is no perfect descriptor that solves every problem, but different descriptors are suitable for different problems, and have to be chosen and combined carefully. In this work only one descriptor, the ASO, is used.

Average Steric Occupancy

The ASO is a descriptor representing the steric environment, and accessible conformers and rotamers of a molecule. It was developed and implemented by Denmark and co-workers [62] in 2018 and first used by Zahrt et al. in 2019. [34] To compute the ASO, a complete set of aligned conformers (and rotamers) of compounds is required. First, a spherical grid of points is computed to enclose the entire conformer library of molecules. In a second step, for each conformer it is determined if a grid point is within the van der Waals radius of an atom, yielding the value of one for the grid point, or if it is outside the van der Waals radius, yielding the value zero. The ASO of a molecule is then computed as the average grid over all corresponding conformers. This gives a descriptor value of $0 \leq ASO \leq 1$ at each grid point. The ASO is therefore able to characterize the shape of the molecule weighted by how often the molecule occupies different regions of space. The ASO can also be used to describe reactions by concatenating the ASOs of all reacting molecules.

3 METHODS

This chapter outlines the basic idea behind the combined experimental and theoretical chemistry approach with ML, explains the exact workflow that was developed, and summarizes the computational details.

3.1 Idea and Scope of this Thesis

The idea behind the combination of experimental and theoretical chemistry with ML is outlined in Figure 3.1. The main approach is to start with a set of experimental data (Fig. 3.1: (1)), which in this work consists of structural formulas of reactants of a Negishi coupling and the corresponding *ee* values. This data is considered to be the most accurate data for the *ee* and will be used to validate all further calculations. The amount of experimental data is most likely not sufficient for a ML approach, as experiments are notably time-consuming. Hence, a QC model (Fig. 3.1: (2)) has to be developed that is able to generate a larger training set by providing the *ee* values for unknown reactions. The QC model in this work consists of the transmetalation transition states for the (R)- and the (S)-enantiomer. Using Equation (2.8) (Section 2.1.2) the *ee* can be calculated from the energy difference between both transition states ΔG^{TS}. Once the QC model is developed the generation of training data is certainly faster than the execution of the same amount of experiments, but still more costly than a ML approach. The next step is to use the obtained training data from experiments and QC to train a suitable ML algorithm (Fig. 3.1: (3)) that is then able to predict the wanted observable rapidly for a large number of unknown reactions. The most promising reactions predicted by ML can subsequently be tested by the QC model (Fig. 3.1: (4)) and only the most excellent reactions are validated by experiment (Fig. 3.1: (5)). This workflow makes it possible to perform a high throughput screening for countless reactions. For the screening, a large database of unknown reactions has to be generated that is used on one hand to complete the training set with the help of the QC model and on the other hand for predictions with ML. This database can in theory contain any kind of reactant suitable for the reaction, but should be limited to reactants and catalysts that can be synthesized in practice.

The scope of this work is to handle the experimental data provided by our collaborators, the Maulide group,[30] and find suitable descriptors (Fig. 3.1: (1)). The presented work exclusively uses the experimental data. A significant improvement of the model requires a QC study of the reaction mechanism. Kinetic and thermodynamic properties of the systems from the training set should be thoroughly investigated computationally. The calculated energetic and structural characteristics of the intermediates and transition states (this data is hardly achievable experimentally) will significantly develop the model and its potential to predict experimental observation. It is important to mention that the studied systems are extremely demanding also for the computational study. The thorough investigation of, e.g., conformational space of all stationary points as well as additional benchmarking calculations to

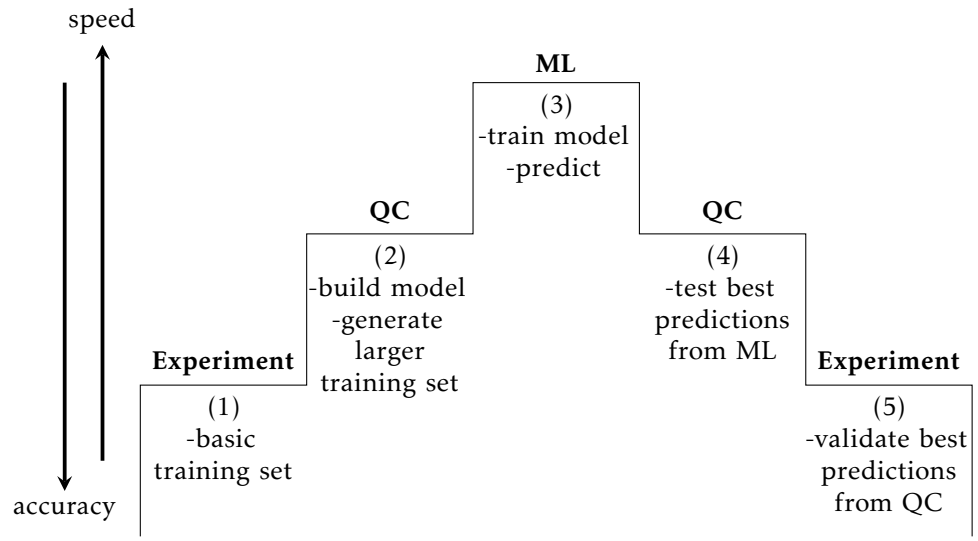


Figure 3.1: Workflow for a combined experimental and theoretical chemistry approach with machine learning.

determine the reasonable QC approach are necessary. This part of work is currently in progress in our group.

The author is aware that the amount of data used in this work is in general not sufficient for a ML problem as complex as this. In order to benefit from statistics, not only one, but multiple ML algorithms are trained (Fig. 3.1: (3)) and compared in Section 4.2.1. The unknown reactions used for testing the ML model are designed with the help of our collaborating experimentalists, which will in the future also perform the experimental validation of the predictions (Fig. 3.1: (5)).

3.2 Workflow

This section outlines the actual workflow from getting experimental data to predicting the *ee* for unknown reactions. The computational details are summarized in Section 3.3. The prerequisites for the proposed workflow are a training data set consisting of (two-dimensional) structural formulas of the three main reactants of the Negishi couplings (organic halide, organozinc compound, and ligand of the Pd-catalyst), the corresponding *ee* value, and a test data set consisting of unknown, but plausible reactions.

1) FROM 2D TO 3D A common format that organic chemists use for their 2D structural formulas is the ChemDraw *.cdx* format. The two-dimensional structures are converted into three-dimensional *.xyz* structures using the Chem3D program from the ChemOffice package. Simultaneously a pre-optimization of the geometry is performed using a FF energy minimization method. The whole process was automated using AutoIt.[63]

2) CONFORMER SEARCH In the next step, the conformer-rotamer-ensemble for every molecule is calculated. The conformer search is performed for all organic halide reactants, the organozinc enantiomers, and the catalyst ligands of every reaction using the iMTD-GC (RMSD) algorithm implemented in

`crest`.[51–53] For the following, the conformers and rotamers of the organozinc enantiomers are combined into one racemic conformer-rotamer-ensemble. In this work also a different method for obtaining the set of initial structures is performed based on Wigner sampling. The differences are discussed in Section 4.1.1. For the Wigner approach, the molecule geometries are optimized and the frequencies calculated using GFN2-xTB,[44,46] while the *wigner.py* script of SHARC[48] uses the calculated frequencies to generate the ensemble of geometries.[a]

3) ALIGN CONFORMERS In the next step, all conformers and rotamers within the three compound groups have to be aligned with the same orientation to assure a correct calculation of the ASO descriptor. The catalyst ligands all have at least one three-bonding phosphor atom in their structures. Hence, they are translated to have the common phosphor atom in the origin. If there are two phosphor atoms, the one to be in the origin is picked randomly. This is sufficiently exact as almost all ligands are symmetric. The ligands are then aligned by the three bonds between the phosphor center and the neighboring atoms (C, O, or N). The bromine reactants are translated to have bromine in the origin and are aligned by the atoms marked in Figure 3.2-A. The organozinc reactants are translated to have zinc in the origin and aligned by the atoms marked in Figure 3.2-B. The algorithm used for the alignment is the Kabsch algorithm that calculates the optimal rotation matrix by minimizing the RMSD.[64,65] Finally, all structures are transformed from *.xyz* into *.mol2* format using OpenBabel.[66]

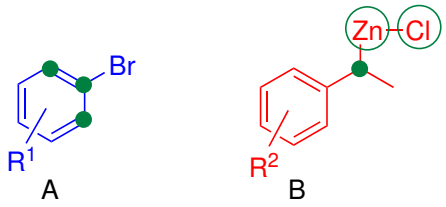


Figure 3.2: The bromine reactants **A** and the organozinc reactants **B** are shown. The molecules are each aligned by the three atoms marked in green.

4) CALCULATE GRID After the conformers for the three main reactants (bromine reactant, organozinc compound, and catalyst ligand) are aligned, a spherical grid is calculated for each reactant group. These homogeneous grids enclose the conformers with an additional 3 Å buffer and are calculated using the ccheminfolib package.[34,62]

5) CALCULATE ASO The next step is to calculate the ASO descriptor for each one of the molecules. This is also done using the the ccheminfolib package[34,62] and the prior calculated associated grids. For each compound group the same grid has to be used to ensure that the number of grid points (features) is always the same.[b] In a last step, the ASOs of the three main reactants of a reaction are concatenated. For these concatenated ASOs it

a The Wigner sampling approach does not produce real conformers as explained in Section 2.2.
b The input vector of a ML model has to have the same length for every sample.

is specifically important to, while aligning, ensure the same orientation of compounds within each of the three reactant groups.

6) FEATURE REDUCTION The concatenated ASOs are already in the correct shape to operate as input for a ML algorithm, but due to the nature of the ASO each reaction sample can have hundreds of thousands of features with a large amount of them being zero. Therefore, a feature reduction, the principal component analysis (PCA), is performed. The PCA is a linear dimensionality reduction using singular value decomposition (SVD) of the data to project it to a lower dimensional space and is implemented in scikit-learn.[67] Before the feature reduction can be performed, the data set is split into training and validation set. The training set is used for training the ML models, while the validation set is used for evaluating them. The feature reduction is fitted only to the training set and then used to transform training and validation set separately. This is important so that the validation set is not biased towards the training data and the feature reduction can subsequently be applied to an unknown set of data for screening purposes. In this work, the validation set is used for testing purposes and no separate test set is used due to the small amount of data available.

7) TRAINING AND VALIDATION The reduced training set is then used to train multiple ML algorithms. A feed forward NN with BOHB hyperparameter optimization is trained as well as a SVR model with grid search hyperparameter optimization and a RFR model. The validation set is used to compare the performance of different models, which is further discussed in Section 4.2.1.

8) PREDICTIONS To predict the *ee* for unknown reactions the steps 1) to 3) have to be repeated for unknown reactants. Step 5) is repeated using the previously calculated grids from the training data, subsequently in step 6) the already calculated PCA model is applied to the data. Finally the unknown reactions can be tested using the trained ML models from step 7).

3.3 Computational Details

The 2D to 3D conversion and geometry optimization using Chem3D is automized using AutoIt v3.[63] It is a freeware BASIC-like scripting language designed for automating the Windows GUI and is used, because the in-house scripting language ChemScript of the ChemOffice package does not support the performed task. The geometry optimization is performed by minimizing the energy with the Chem3D MMFF limited to 2000 iterations and otherwise default parameters.

The conformer search was carried out with an 7 kcal/mol energy window using `crest`[51] version 2.6.3 with the iMTD-GC (RMSD) algorithm.[52,53] The Conformer-Rotamer Ensemble Sampling Tool, `crest`, is an utility program for the `xtb` program.[44,46,68] The `xtb` version 6.1.3 is used with `crest` and for preliminary calculations for the Wigner sampling. The Wigner sampling for 10 initial geometries is performed using the *wigner.py* script implemented in SHARC.[48] The Kabsch algorithm for the conformer alignment is implemented

in Python and the subsequent conversion from *.xyz* to *.mol2* format is executed using OpenBabel.[66]

For the ASO calulation the ccheminfolib[34,62] package is used. First, the grid is calculated with 0.5 Å and 1.0 Å grid spacing, which is compared in Section 4.1.1. Then the ASO is calculated using the obtained grids. The ccheminfolib[34,62] package is modified for these calculations. The van der Waals radii of zinc (2.12 Å[69]) and iron (1.98 Å[69]) are included in the *cchemlib/atomtypes.py* file. Additionally the *cdesclib/calculator.py* is changed to make the ASO calculation of the fine 0.5 Å grid possible with a reasonable amount of RAM access. The process of calculating the average over all conformers needs to be changed to be more RAM efficient compromising for performance.

The feature reduction using PCA is applied using the implementation in scikit-learn.[67] The 451,477 features of one reaction mixture are reduced to 150 features with a variance of close to 100 %, depending on the splitting in training and validation data set.

The NN is implemented using Keras[70] with a Tensorflow[71] backend. The BOHB[61] hyperparameter optimization is performed using HpBandSter[61] with a budget of 9 to 81 epochs,[c] a batch size of 5 samples, and 396 iterations. The fixed architecture of the NN is an input layer with 150 nodes (determined by the feature reduction) and an output layer with 1 node (for the ee) and ReLU activation function. The learning rate is optimized with BOHB[61] in the range of $1 \times 10^{-6}$to 1×10^{-2} and the optimizer is *Adam* or *SGD* (momentum: 0.0–0.99). The number of hidden layers is optimized from 2 to 5 with 1–100 nodes each, a dropout of 0.0–0.6 between each layer, and the activation function being ReLU, ELU, SELU, softmax, softplus, sigmoid, or softsign. The hyperparameter search resulted in a NN architecture of 3 hidden layers, with 68 nodes and ReLU activation, 20 nodes and ReLU activation, and 25 nodes and softsign activation, respectively. The dropout after each of these hidden layers was 0.140, 0.486, and 0.388, respectively. The optimizer is chosen to be Adam with a 1.583×10^{-3} learning rate.

The SVR and the corresponding grid search hyperparameter optimization, as well as the RFR model with 200 estimators, are implemented using scikit-learn.[67] The SVR grid search is performed with linear, radial basis function (RBF), and polynomial kernels (second- and third-order polynomials) for the penalty parameter of the error term $C = [0.0001, 0.001, 0.01, 0.1, 1, 10]$ and for $\epsilon = [0.01, 0.1, 0.5, 1, 2]$. The grid search for all kernels resulted in the hyperparameter $\epsilon = 0.01$ and furthermore in $C = 0.001$ for the linear, $C = 1.0$ for the RBF, and $C = 0.01$ for the two polynomial kernels. Ten replicate runs are performed for each ML model.

c Epochs = Number of times the whole training set is used in the process of training the model.

4 RESULTS

In this chapter, the workflow described in Section 3.2 is applied to a set of 190 reactions with known *ee* and predictions are made for 20 reactions with unknown *ee*. First, in Section 4.1, the ASO descriptor is compared for two different types of geometry sampling methods and furthermore the construction of concatenated ASOs for reaction mixtures is described. Second, in Section 4.2, different ML algorithms used for predicting the *ee* are validated and compared. Additionally, the *ee* predictions made for the set of 20 unknown reactions are discussed. Finally, in Section 4.3, the limitations of the proposed workflow are explained.

4.1 Descriptor Calculation

To calculate the ASO descriptor a set of geometries that represents the steric environment is necessary for each molecule. To obtain this set of geometries, there are a range of methods. These methods include mode-following methods, like Wigner sampling, or conformer searches at different levels of theory. The following compares the two different methods used and shows how the ASO is utilized for reaction mixtures.

4.1.1 Wigner Sampling versus `crest` Conformer Search

The first method used to generate a set of geometries is Wigner sampling. It is preceded by a geometry optimization and frequency calculation with `xtb` and therefore promises fast results. The method is used to sample just the catalyst ligands of the reaction and not the reactants. Using the Wigner sampling, the ASO is tested for the first time to confirm that it is able to capture the correct properties responsible for predicting the *ee*. Even though the Wigner structures are not real conformers and only the ASO of the catalyst ligand is used as input for the ML model, the algorithm already predicts *ee* values better than random, which can be seen in Figure 4.1. The predictions for 27 test samples are calculated with a NN with ReLU activation, dropout rate of 0.0522, Adam optimizer with a learning rate of 0.0006, 20 epochs, and 4 hidden layers with 3, 4, 63, and 16 nodes, respectively. In Figure 4.1 the absolute errors of the test predictions are compared to the errors of random predictions and of a uniform prediction, with the uniform value being the average of the observed *ee* of the test samples. Clearly, the errors of the random predictions have the highest values, followed by the uniform prediction, and then the test prediction with the overall lowest errors. This shows that there is a correlation between the ASO and the *ee* values.

With that in mind another method to sample real and optimized conformers and rotamers is wanted. The typical compromise between accuracy and computational cost leads to the use of the semiempirical iMTD-GC algorithm

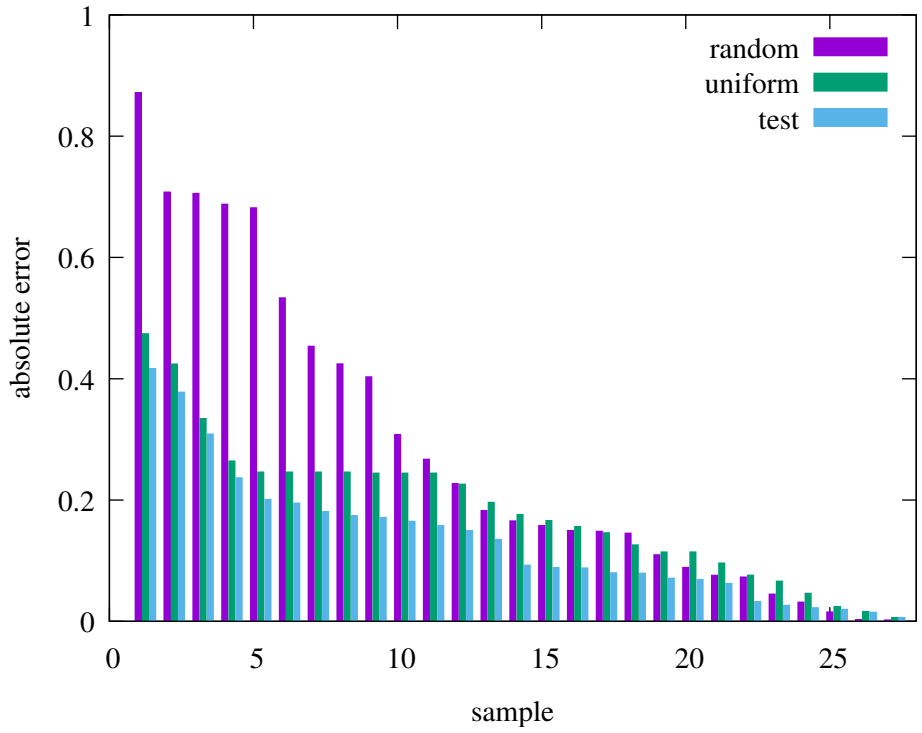
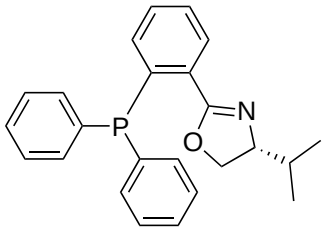


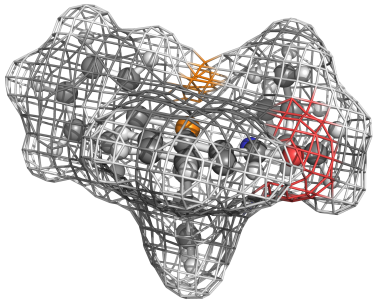
Figure 4.1: Plot of the absolute error of the enantiomeric excess from test predictions, from randomly predicted values, and from a uniformly predicted value (equaling the average of the observed *ee*) against 27 test samples.

implemented in `crest`. Using the example of the ligand PHOX2[a] the Wigner sampling is compared to the `crest` conformer search.

The ligand can be seen in Figure 4.2 as structural formula and in 3D with the surface[b] of the molecule pictured as a mesh.



(a) Structural Formula

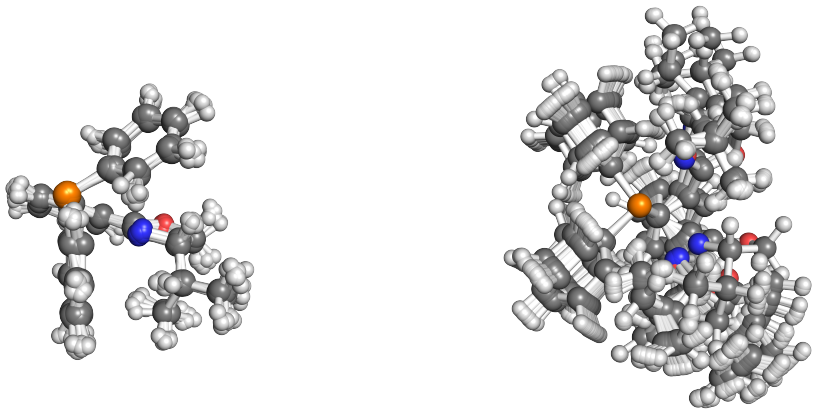


(b) Surface Plot

Figure 4.2: Structural formula and a three-dimensional plot with the surface picture by a mesh of the ligand PHOX2.

a All molecules and their labels can be seen in Figures A.1 to A.7 in the *Appendix*.
b The surface pictured is the *Connolly* or solvent-excluded surface.

Figure 4.3 shows the aligned geometries for both described methods. While the 10 Wigner geometries (Fig. 4.3a) mainly show deviations for the hydrogen atoms, the 154 `crest` conformers and rotamers (Fig. 4.3b) clearly occupy a wider range of space. The number of `crest` conformers and rotamers is, depending on the ligand, between 7 and 9,915. It is important to gain knowledge about the occupied space as the steric hindrance of the ligand is proposed to be (at least partially) responsible for the induction of the enantioselectivity in the transmetalation transition state of the Negishi cross-coupling. Therefore, the steric information of all molecules, including the reactants, involved in the transmetalation complex are of interest for the ML algorithm and should be represented in the descriptor (see Section 4.1.2).

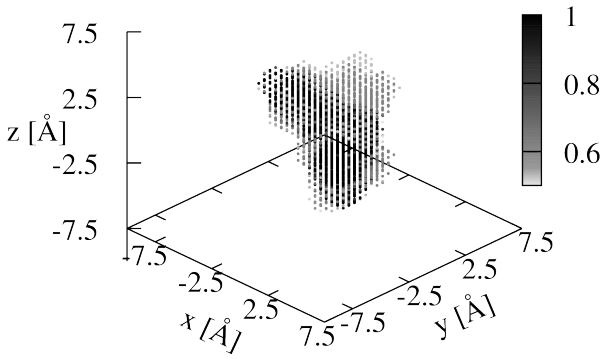


(a) Wigner Sampling **(b)** `crest` Sampling

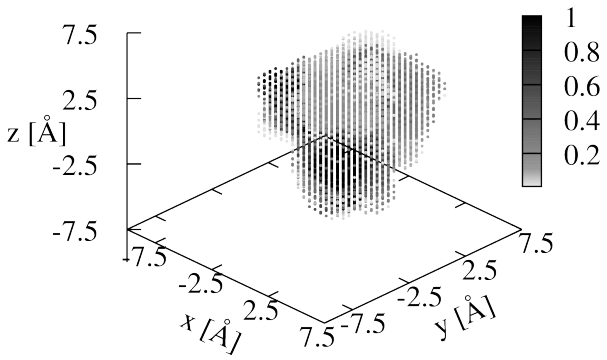
Figure 4.3: Geometries obtained from Wigner sampling **(a)** versus the conformer-rotamer-ensemble obtained from `crest` **(b)** of the ligand PHOX2.

To investigate the ASO of PHOX2 further, it is plotted in three dimensions in Figure 4.4. Figure 4.4 exemplarily shows the ASO for the `crest` sampling with a 0.5 Å grid spacing, which is the method that is used throughout the following work. For clarity, only the ASO values greater than 0.5 are plotted in Figure 4.4a and only values greater than zero are plotted in Figure 4.4b. The whole spherical grid, including the regions that are zero, can be seen in Figure A.8 in the *Appendix*. In general, regions with a darker color express a rigid part of the molecule, whereas lighter parts show the flexible regions that are occupied less often by the conformers. Figure 4.4a is plotted to show the regions that are occupied in more than half of the conformer structures. This plot can be related to the surface plot in Figure 4.2b and also to the van der Waals structure of PHOX2 in Figure 4.4c. The three six-ring structures on the left, front, and bottom of the surface plot are reflected on the same positions in the ASO in Figure 4.4a. While these rings are more rigid and therefore darker in color, the right side seems incomplete. The missing part is the five-ring of the structure, which can only be seen in Figure 4.4b on the right side, where additionally lower ASO values are plotted. The plot shows that the five-ring is rather flexible, as well as the six-ring in the front.

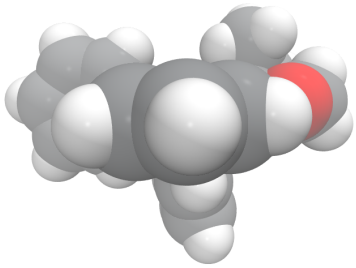
After the ASO calculation, the 3D grid (see Fig. 4.4) is unfolded to form a vector (see Fig. 4.5), which can be used as ML input. The vector can then be



(a) ASO ≥ 0.5



(b) ASO > 0



(c) Van der Waals plot

Figure 4.4: Van der Waals plot and three-dimensional plot of the non-zero parts of the ASO of PHOX2. The total grid including ASO = 0 can be seen in Figure A.8 in the *Appendix*.

plotted against the (grid point) index, which is shown for PHOX2 in Figure 4.5 to compare the Wigner sampling to the `crest` conformer search. The figure only shows the non-zero interval of the ASOs. The Wigner sampling ASO for 1.0 Å grid spacing (Fig. 4.5a) mainly has values of either one or zero, which means the grid points are either always or never within the van der Waals radius of the conformer atoms, corresponding to very similar geometries. The similarity of the structures can be confirmed by Figure 4.3a. In comparison, the ASO of the `crest` conformers (Fig. 4.5b and 4.5c) shows more variability in the conformers with its frequent values within the interval $0 < ASO < 1$. This variability can also be seen in Figure 4.3b. The ASO of the `crest` conformers is calculated for two different grid spacing sizes. Figure 4.5b is calculated with a 1.0 Å grid spacing, like Figure 4.5a, while Figure 4.5c uses a finer 0.5 Å grid spacing. A finer grid spacing of 0.5 Å enables the ASO to capture even smaller variations among the conformers. These small variations might induce the enantioselectiviy of the catalyst and are therefore important for capturing the reactivity. The grid point indices of Figure 4.5a and 4.5b are different, even though the same grid spacing is used, because less ligands are used for the Wigner test and the ligands are not all oriented in the same way as explained in Section 3.2.[c]

Concluding, the ASO is able to capture the differences in steric information of both sampling methods, Wigner and `crest`. To capture the steric environment of a molecule correctly and sufficiently the ASO needs to be supplied with a rather extensive conformer-ensemble and should be calculated with a suitable grid spacing. In the following work the `crest` conformer search with a 0.5 Å grid spacing is used.

4.1.2 **Concatenated ASOs**

To clarify, the input of a ML algorithm generally has to have the same dimensions for all samples used for training, validation and testing. For the ASO, this requirement is ensured by generating a grid that is big enough to fit even the largest molecule and using this grid for calculating the ASO of all molecules. In the case of Negishi coupling reactions the three main components important for the reaction are the organozinc compound, the bromine reactant and the ligand of the palladium catalyst. The ASOs are calculated for every reactant of a sample reaction separately. The obtained ASOs are then simply concatenated to represent the reaction mixture. The ASOs for concatenating should therefore always be in the same order and the molecules should all be aligned as explained in Section 3.2. This ensures, that the part of the molecule important for the reaction is always in the exact same spot of the grid. In principal the spherical ASO is rotation invariant and the orientation of the molecules should not matter, but it is definitely not rotation invariant, when multiple ASOs are concatenated. In general, the same grid can be used for calculating the ASOs for all reactants, but the bromine and organozinc reactants are all smaller and less bulky, compared to the ligands. Hence, the decision is made to use a separate grid for every reactant group to save computational time. Thus, the bromine reactants have a grid of 41,472

c The phosphor atom is not translated to the origin and the three neighboring atoms are not aligned for all structures. The conformers of a single molecule are nevertheless aligned.

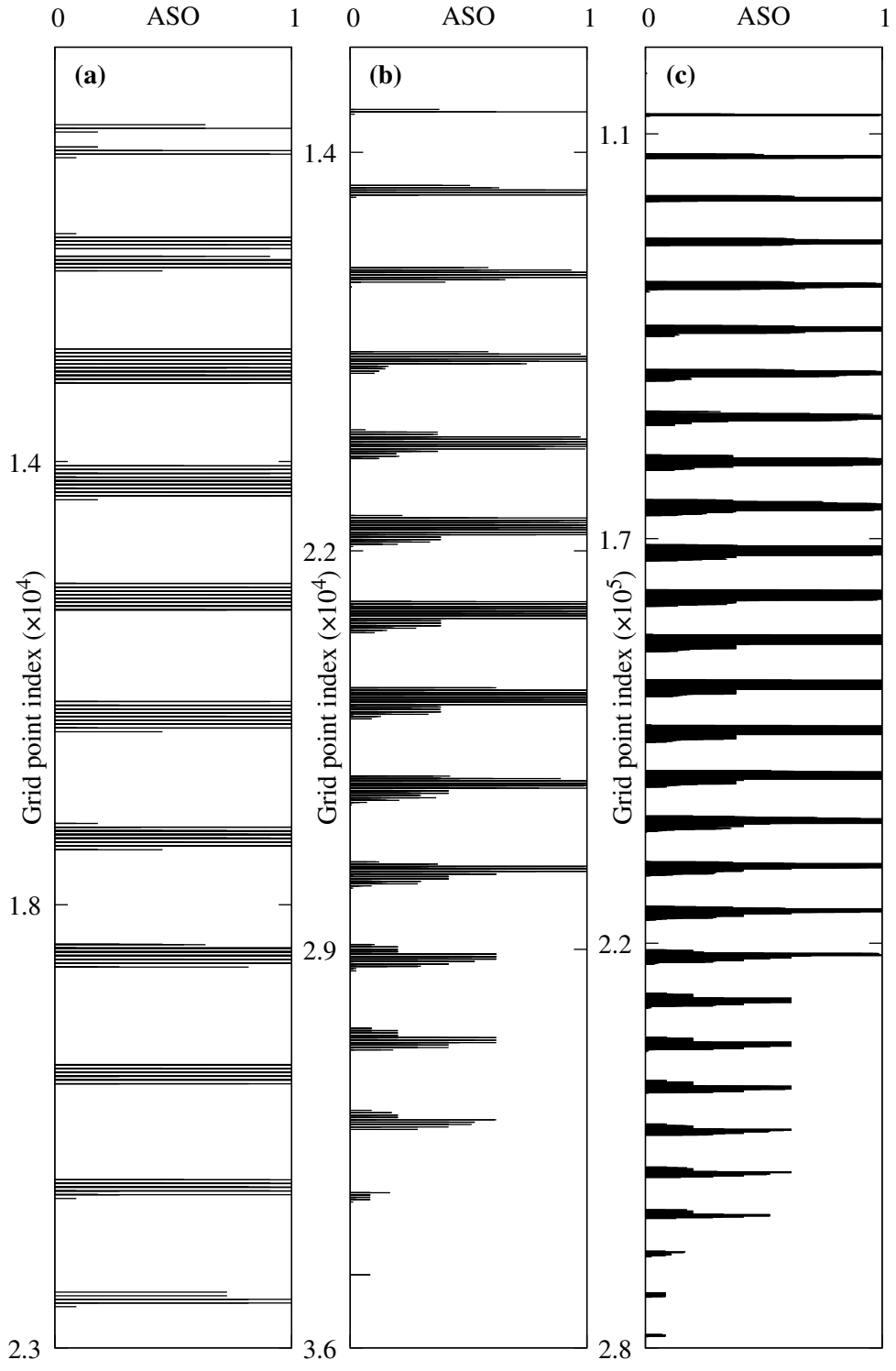


Figure 4.5: Plots of the ASO of the ligand PHOX2 (see Fig. 4.2) against the grid point index. The plots are zoomed in to mainly show the non-zero parts of the descriptor. The ASO is calculated from:
(a) Wigner sampling and 1.0 Å grid spacing,
(b) `crest` conformer search and 1.0 Å grid spacing, and
(c) `crest` conformer search and 0.5 Å grid spacing.

points, the organozinc compounds have 37,260 grid points, and the ligands have the biggest grid with 372,745 points.

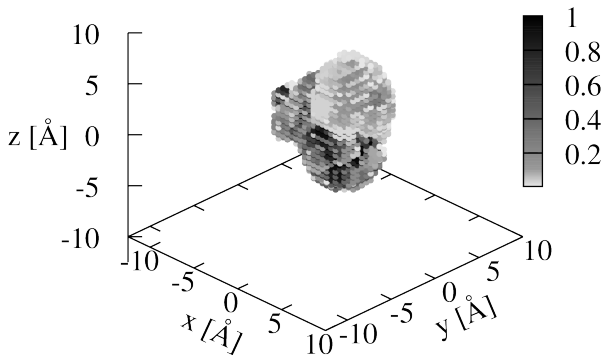
In the following, the concatenated ASOs of three different reactions are exemplarily investigated and compared. First, three different ligands MI16, CAT2, and PPSF17[d] are chosen and the three-dimensional ASO of the ligands is compared in Figure 4.6. Just by looking at the size of the plots, the size difference of the ligands becomes apparent. Additionally, ligand MI16 is more rigid as ligands CAT2 and PPSF17, which can be seen from the color difference. The ASO of MI16 shown in Figure 4.6a is able to represent the structure of the molecule with the two rather rigid methyl groups on the left side and two more flexible ring structures on the right. Contrary to that, the ASOs of the ligands CAT2 and PPSF17 (Fig. 4.6b and 4.6c) show almost spherical structures with mostly light regions, representing very flexible molecules.[e] This is due to the flexible *tert*-butyl groups, which can be seen in the *Appendix* in Figure A.3 for PPSF17 and Figure A.5 for CAT2.

Three reactions using the three chosen catalyst ligands are shown in an oversimplified manner in Figure 4.7. The first reaction (Fig. 4.7a) is chosen to show the concatenated ASOs of one of the smallest ligands. The other two reactions (Fig. 4.7b and 4.7c) involve rather big ligands that have quite similar structures.

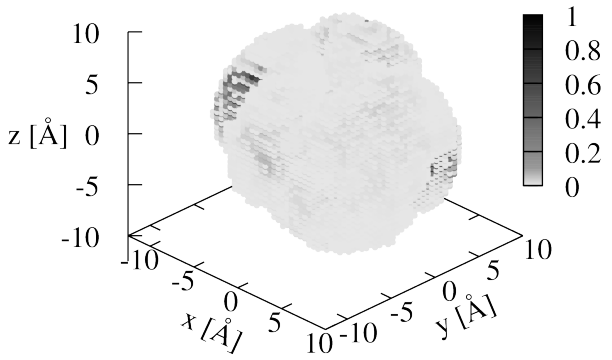
Finally, the ASOs of the three reaction mixtures are plotted against the grid point index in Figure 4.8. Every of the plots in Figure 4.8 corresponds to one of the reactions from Figure 4.7 with grid point indicies from 0 to 41,471 representing the bromine reactant, indicies from 41,472 to 78,731 representing the organozinc compound, and indicies 78,732 to 451,476 representing the ligands. Comparing Figure 4.8a to Figures 4.8b and 4.8c, the ASO clearly captures the already mentioned size difference of the ligands by spreading over a larger part of the grid. Even though the ligands of Figures 4.8b and 4.8c are similar in structure, differences can also be seen in the ASO. Even structural changes as little as the substitution of a proton to a CF_3 group can be seen, when comparing the organozinc compound of Figures 4.8a and 4.8c to Figure 4.8b.

This shows, that the ASO is really able to capture small differences in the steric environment of a molecule and can be used for a single molecule as well as all reactants of a sample reaction. Nevertheless, the alignment of the conformers is crucial for a reasonable ASO calculation.

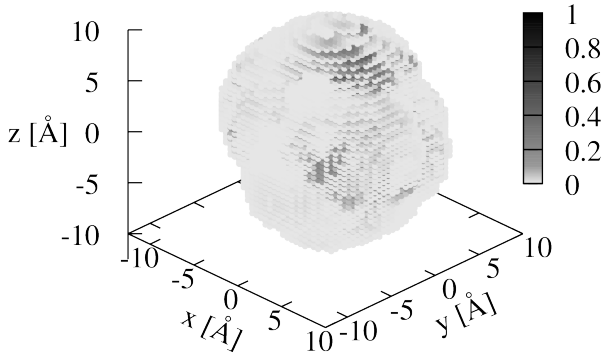
d All molecules and their labels can be seen in Figures A.1 to A.7 in the *Appendix*.
e Flexible molecules at least on the outside, when the phosphor atom is in the center.



(a) MI16

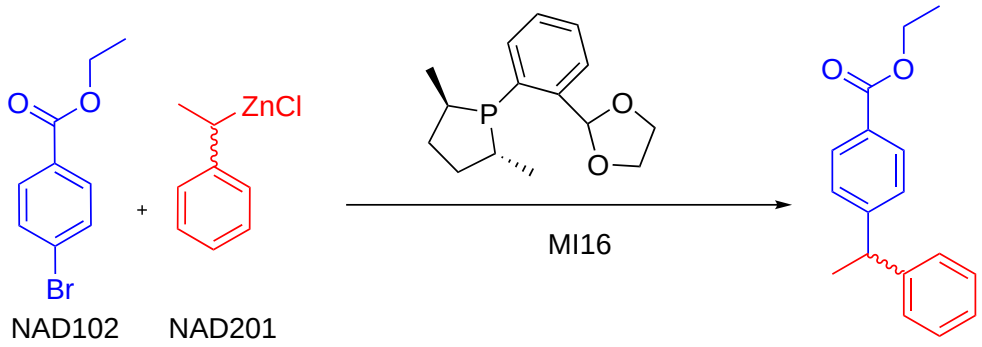


(b) CAT2

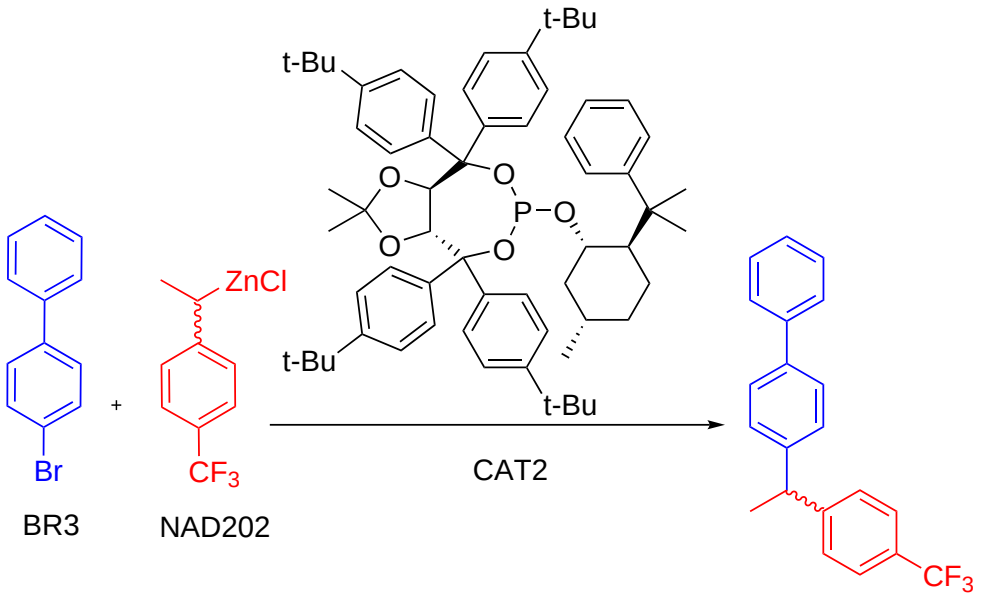


(c) PPSF17

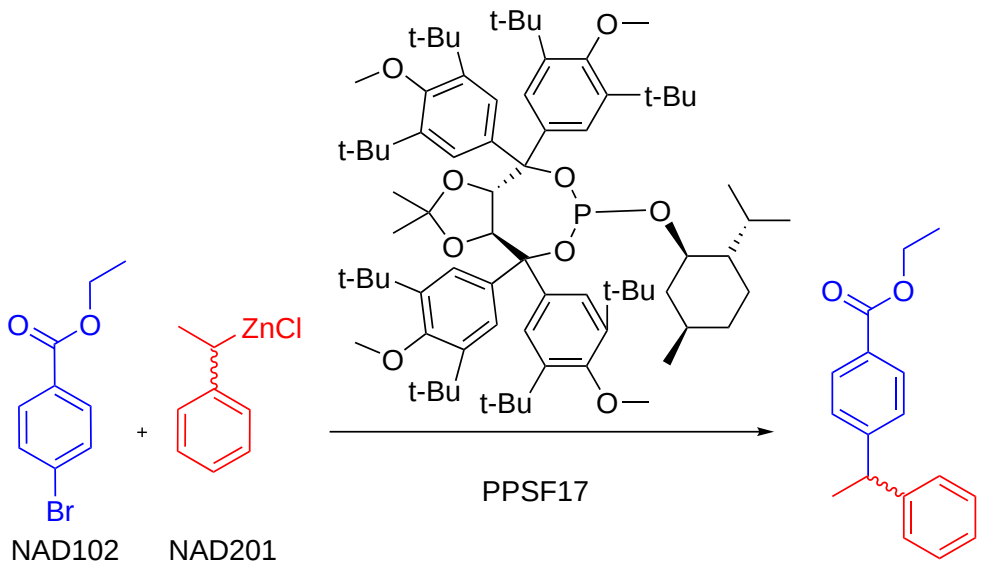
Figure 4.6: Three-dimensional plot of the ASO of the three different ligands, MI16, CAT2, and PPSF17. For clarity, the zero values are not plotted.



(a) Reaction NAD102_NAD201_MI16



(b) Reaction BR3_NAD202_CAT2



(c) Reaction NAD102_NAD201_PPSF17

Figure 4.7: Three simplified examples of reactions used for training the ML models.

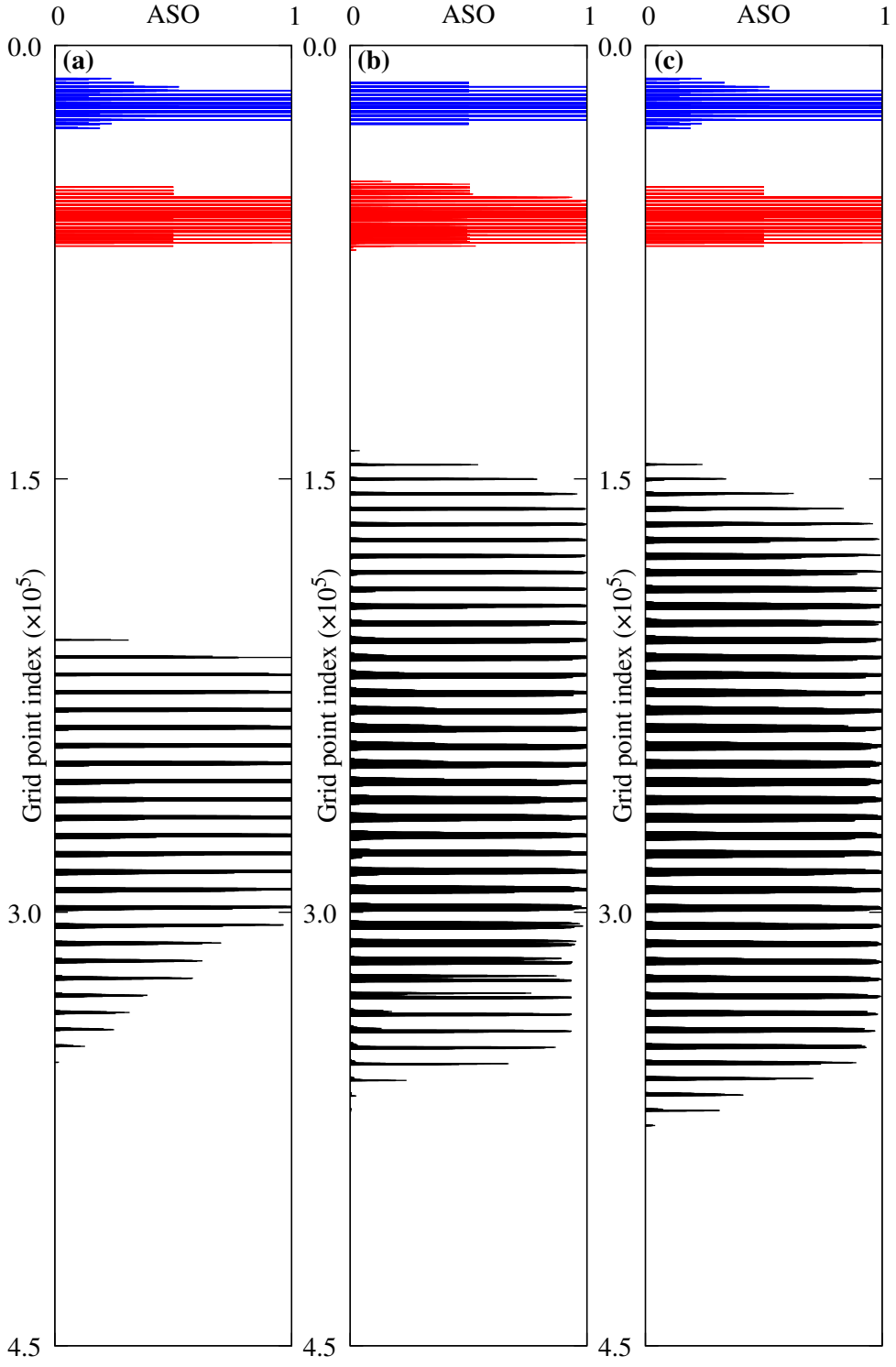


Figure 4.8: The concatenated ASOs plotted against the grid point index $[0, 451476]$, showing the reaction mixtures of Figure 4.7. The plot is colored according to the type of reactant it represents (blue: bromine reactant, red: organozinc compound, black: catalyst ligand).
(a) Reaction mixture NAD102_NAD201_MI16
(b) Reaction mixture BR3_NAD202_CAT2
(c) Reaction mixture NAD102_NAD201_PPSF17

4.2 Enantiomeric Excess Prediction with Machine Learning

This section consists of two parts. In the first part, the different ML algorithms that are used are compared, while in the second part predictions for 20 unknown reactions are discussed. It should be noted again that the training data set used in this thesis is not large enough for the algorithms used, since no data from QC is available yet. Nevertheless, the section is written as if the training set was sufficient, which means that a functioning workflow is implemented, but the results should be interpreted carefully.

4.2.1 ML Model Evaluation and Comparison

It is common practice in ML to set up multiple different models, train them, evaluate them, and in the end only choose the best performing model, which is then used for predictions. In this work six different models are trained and evaluated on a set of 190 data points. Four of them are SVR models with linear, 2nd and 3rd degree polynomial, and RBF kernel functions. The other models are a RFR model and a NN. Each of the models is trained on a set of 152 training data points and evaluated using the remaining 38 validation data points. The procedure is done several times for different splits of the data in training and validation set. This *k*-fold cross-validation is used to detect overfitting and selection bias caused by the split. Additionally, it gives an insight in how the model will generalize on an unknown, independent data set. The best performing model should then be tested again on an unknown testing set and can finally be used for predicting novel data. In this work, the best performing model is not tested again as for the lack of training data, but the six mentioned models are compared and additionally all used for predictions in Section 4.2.2. More than just one model is used in order to benefit from statistics as the data set used is insufficient for providing a satisfactory ML model.

For all models, a 10-fold cross-validation is performed and the averages over the ten replicate runs of the MAE and the coefficient of determination[f] R^2 are reported in Table 4.1. The R^2 score equals one for the perfect model, zero for a model predicting random numbers, and below-zero for arbitrarily worse models. Additionally, the error of the R^2 score is noted to show the stability of the model in terms of generalizability. The model with the least variability in performance and therefore the lowest error is the most stable model.

The lowest and therefore best MAE of 0.124 is obtained for the SVR model with RBF kernel. The worst model in terms of MAE is the RFR model with 0.146. The highest and therefore best R^2 score of 0.593 is obtained also for the SVR model with RBF kernel with an error of 0.036. The worst model in terms of R^2 score is the NN with 0.454 and an error of 0.049. The error as an indicator for the stability of the model shows the RFR model to be the most stable model with a value of 0.026 and the linear kernel SVR to be the most unstable one with a value of 0.059.

For the SVR models, the hyperparameters of every fold of the 10-fold cross-validation are reported to be $\epsilon = 0.01$ and $C = 0.001$ for the linear, $C = 1.0$

f The coefficient of determination is also called regression score function.

Table 4.1: Average Mean Absolute Error (MAE) and average coefficient of determination R^2 for 10 replicate runs of support vector regression (SVR) models with linear, 2nd and 3rd degree polynomial (POLY2, POLY3), and radial basis function (RBF) kernel functions, of a random forest regression (RFR) model, and a neural network (NN).

Model	MAE	R^2
SVR_linear	0.135	0.532 ± 0.059
SVR_POLY2	0.136	0.517 ± 0.042
SVR_POLY3	0.127	0.573 ± 0.045
SVR_RBF	0.124	0.593 ± 0.036
RFR	0.146	0.526 ± 0.026
NN	0.135	0.454 ± 0.049

for the RBF, and $C = 0.01$ for the two polynomial (POLY2, POLY3) kernels, respectively. The MAE is the lowest with 0.124 and therefore best for the SVR model with RBF kernel, while the SVR model with 2nd degree polynomial kernel performs the worst with an MAE of 0.136. The R^2 score is the best for the SVR model with RBF kernel with 0.593 and the worst for the SVR model with 2nd degree polynomial kernel with 0.517. The most unstable SVR model is the one with linear kernel.

Compared to the SVR models, the RFR model performs worse in terms of MAE and also the R^2 score is rather low with 0.526. The error nevertheless indicates a more stable model.

For the NN various different hyperparameter searches for different splits of training and validation data and different random numbers initializing the weights of the NN are performed with a different resulting hyperparameters each time. Finally, one of the architectures (reported in Section 3.3) is used for the 10-fold cross-validation. The difference to the previous models is that the NN is trained on 81 epochs, which means the training data set is used 81 times in the process of training. The weights of the NN are initialized randomly and using the available training set just once is not enough for training the weights sufficiently. The learning curve for such a process can be seen in Figure 4.9 showing the MSE against the number of epochs. One can clearly see that the MSE during the first 10 epochs decreases drastically. After 10 epochs, there is on average hardly any change for the validation MSE. The plot of the learning curve additionally enables to flag if a model is overfitting. If the validation curve starts to increase, while the training curve is decreasing the model is overfitting in favor of the training data. This cannot be seen here. The MAE and the R^2 score are reported in Table 4.1. The MAE with 0.135 is reported to be mediocre, similar to the SVR model with linear and 2nd degree polynomial kernel, while the R^2 score with 0.454 is lower than all previously reported values. This makes the NN one of the worst reported models just from looking at the MAE and the R^2 score. The model does perform the best for the data split that is used in the hyperparameter search, but fails drastically for other data splits. This phenomenon of overfitting can also be seen in the R^2 error of 0.049.

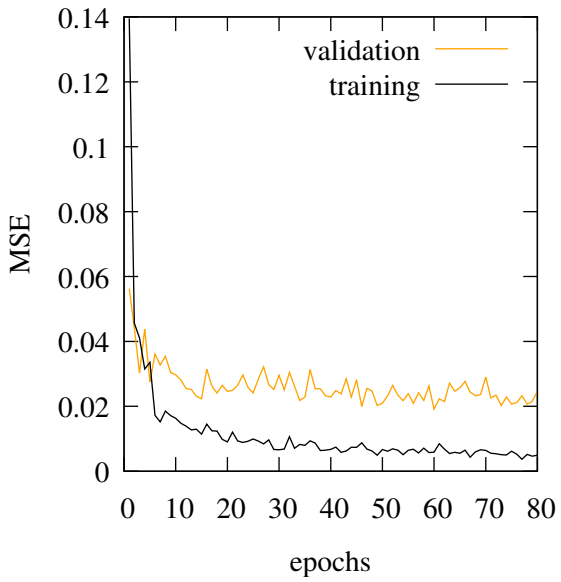
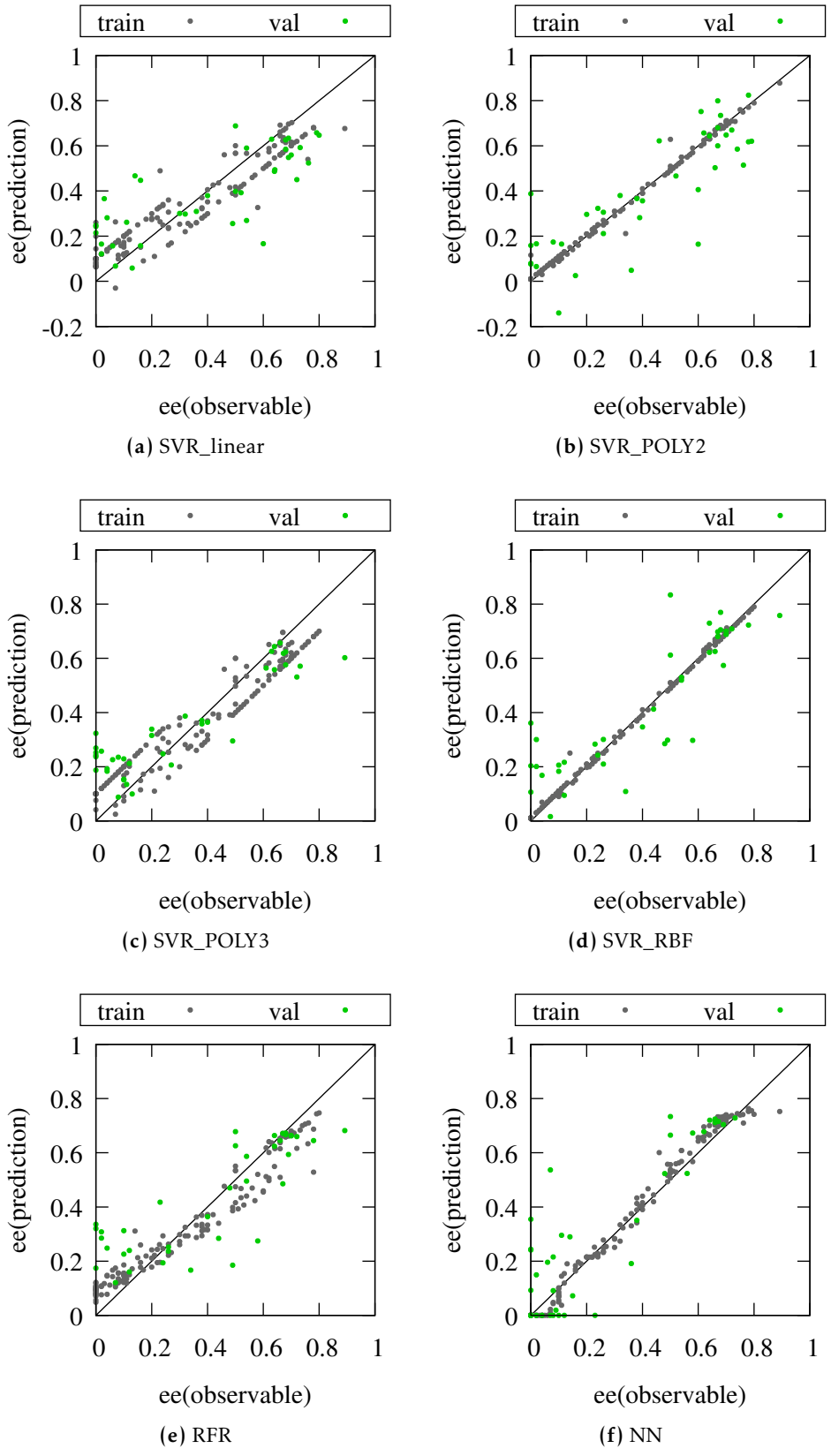


Figure 4.9: Plot of the learning curve of the neural network. The Mean Squared Error (MSE) is plotted against the epochs for the training and the validation set.

On top of evaluating the MAE values and the R^2 scores the models are evaluated visually. In Figure 4.10 the observed *ee* of the training and validation set is plotted against the predicted *ee* of the model for the best replicate run. Figure 4.10 confirms again, that the 190 data samples are not sufficient to satisfactorily solve the ML problem. The plots for the SVR model with linear (Fig. 4.10a) and 3rd degree polynomial kernel (Fig. 4.10c) show an unsatisfactory prediction of the training as well as the validation set, while the other two SVR models (Fig. 4.10b and 4.10d) show an almost perfect prediction of the training set with large errors for the validation set. Even though the SVR model with RBF kernel is the best model from analyzing the MAE and R^2 score, it seems to overfit in favor of the training data and struggles to predict the independent validation set. The RFR model (Fig. 4.10e) has similar validation performance to the SVR models, but shows deviations for the prediction of the training set. Contrary to the MAE and R^2 score evaluation the plot of the NN predictions looks the best. While the NN has drastic problems predicting *ee* values from 0.0 to 0.6, it is sufficiently exact for *ee* values above this threshold. However this is not a huge problem for the used system. To recap, the goal is to find reactions with a high *ee* and therefore products that are as enantioenriched as possible. Hence, the reactions with low *ee* do not need to be as exact as they are not of interest anyway. The more interesting data is the high selectivity data from values 0.6 and up, which is eligible for validation by experiments.

Figure A.9 in the *Appendix* shows the plots containing the data of all 10 runs and the same trends as in Figure 4.10 can be seen: None of the models is sufficiently exact in predicting the *ee*. The predictions for *ee* values below 0.6 are especially bad for all models, while an accumulation of correctly predicted values above 0.6 can be observed. In particular, the predictions of 0.0 *ee* seem to exceed the abilities of the ML models.



(a) SVR_linear **(b)** SVR_POLY2

(c) SVR_POLY3 **(d)** SVR_RBF

(e) RFR **(f)** NN

Figure 4.10: Plots of the observed *ee* of the training and validation set against the predicted *ee* for the the best run of each of six different models. All 10 replicate runs can be seen in Figure A.9. The model are support vector regression (SVR) models with linear, 2nd and 3rd degree polynomial (POLY2, POLY3), and radial basis function (RBF) kernel functions, a random forest regression (RFR) model, and a neural network (NN).

The problem of predicting low *ee* values arises probably from the lack of data and the use of only one descriptor. While the ASO descriptor is able to capture steric information, it completely lacks any electrostatic background. Taking electronic descriptors additionally into account is planned for the future.

4.2.2 Enantiomeric Excess Prediction for Unknown Reactions

In this section, *ee* predictions are reported for a set of 20 unknown reactions. The set of reactions was determined in collaboration with the experimentalists and can be seen in Figure 4.11. The ligands of the reactions are shown in Figure A.10 in the *Appendix*. Ligands TEST13 to TEST20 are taken from the work of Wang *et al.*[72] The decision to use the same organozinc and bromine reactant for all ligands is made, because the combination of NAD102 and NAD201 is the most commonly used one in the experimental data and therefore the best represented one during the training of the ML models. All predictions made in the following are speculative.

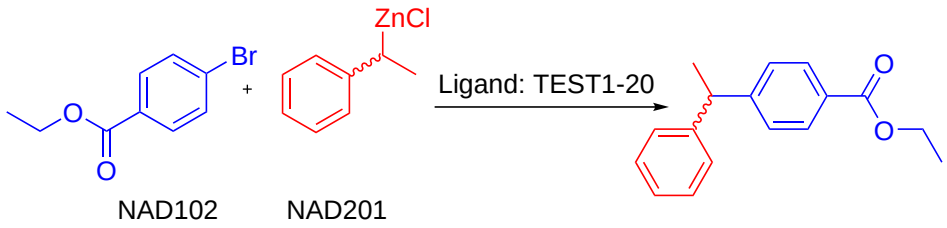


Figure 4.11: Basic reaction scheme of the unknown reactions used to predict the enantiomeric excess. The ligands TEST1 to TEST20 are pictured in Figure A.10.

In the following, predictions are made with the six previously used models of SVR with linear, 2nd and 3rd degree polynomial (POLY2, POLY3), and RBF kernel functions, the RFR model, and the NN. The average of the predictions over 10 replicate runs is given in Table 4.2 for every model. The average predictions for the *ee* vary from 0.000 to 0.519 for all ligands and all models. This is the region of predictions that is determined in Section 4.2.1 to be insufficiently predictable by all used models. Nevertheless, the predictions are discussed. The reactions mixtures in Table 4.2 are given in the order that is predicted by all SVR models as well as the RFR model. Which means that the reaction mixture with ligand TEST1 has the lowest *ee* and the reaction mixture with ligand TEST12 the highest *ee*. Only the NN predicts a different order of reaction mixtures with ligand TEST13 having the lowest *ee* and TEST16 having the highest *ee*.

To recap, non of the tested reactions seems to be in the desired range of *ee* values that are of interest for further research. Additionally all of the ligands are predicted in a range, where the models are insufficient, which makes the predicted *ee* values imprecise. Nevertheless, it would be interesting to have the reactions with ligands TEST1, TEST12, TEST13, and TEST16 validated by experiments, as they are predicted to have the highest and lowest *ee* values of the SVR, NN, and RFR models. Especially TEST13 is of interest as it is the lowest predicted *ee* for the NN, but the second highest value for the SVR and

Table 4.2: Average predictions for the enantiomeric excess of unknown reactions over 10 replicate runs. The names of the reaction mixtures are shortened by NAD102_-NAD201_... to just be naming the corresponding ligand TEST{...}.

Reaction Mixture	SVR				RFR	NN
	linear	POLY2	POLY3	RBF		
TEST1	0.030	0.055	0.033	0.118	0.197	0.155
TEST3	0.051	0.081	0.109	0.131	0.209	0.224
TEST2	0.079	0.104	0.122	0.135	0.218	0.224
TEST5	0.107	0.118	0.129	0.139	0.225	0.225
TEST4	0.120	0.138	0.139	0.142	0.235	0.109
TEST7	0.138	0.152	0.144	0.145	0.238	0.216
TEST6	0.149	0.159	0.153	0.147	0.243	0.227
TEST9	0.160	0.169	0.162	0.151	0.247	0.258
TEST8	0.177	0.173	0.171	0.155	0.252	0.040
TEST20	0.199	0.182	0.178	0.161	0.255	0.134
TEST19	0.214	0.189	0.189	0.166	0.260	0.127
TEST18	0.227	0.195	0.199	0.173	0.264	0.080
TEST15	0.245	0.202	0.206	0.185	0.268	0.032
TEST14	0.260	0.205	0.210	0.199	0.273	0.016
TEST17	0.274	0.212	0.215	0.199	0.277	0.005
TEST16	0.306	0.225	0.223	0.200	0.283	0.519
TEST11	0.323	0.233	0.230	0.201	0.290	0.316
TEST10	0.338	0.249	0.239	0.203	0.298	0.020
TEST13	0.359	0.281	0.245	0.204	0.307	0.000
TEST12	0.386	0.359	0.280	0.314	0.317	0.298

RFR models. Also ligand TEST16 is very interesting as it is the only ligand, where the *ee* is never predicted to be 0.0 for all 10 runs of the NN. All other ligands are at least once predicted to have zero *ee* by the NN.

4.3 Limitations

The goal of this section is to discuss the limitations of the presented workflow, which should be kept in mind for further development.

The first limitation of the workflow is the prediction of the *ee* itself. The *ee* is not a distinct value. It can for example be 50 % for a 75:25 mixture of (R)- to (S)-enantiomer, but it is also 50 % for a 25:75 mixture. In principle it would make more sense to predict excatly this enantiomeric ratio (*er*) of the (R)- to the (S)-enantiomer to be able to easily confirm which enantioenriched mixture the reaction produces. This idea breaks down, because of the experimental data provided. The experimentalists do in fact measure the *er* by HPLC, but they cannot assign the signals to the corresponding enantiomers without additional elaborate experiments. This is the reason why the *ee* is used instead of the *er*. Nevertheless, there are some reference experiments in the literature that can be used to assign the favorable enantiomer.

Another limitation lies in the way the conformers are aligned for the calculation of the ASO. In the current in-house implementation of the Kabsch algorithm the catalyst ligands have to have at least one phosphor atom that has three binding partners. Also the other two reactants of the reaction mixture need to have the marked atoms pictured in Figure 3.2. These limitations of the implementation can easily be changed if needed, which is not necessary for the provided data set. Additionally, the integration of other ligands apart from phosphor based ones is not foreseen for this work. A limitation that cannot be changed so easily is that ligands with two or more phosphor atoms are aligned randomly by one of the two phosphor atoms. It is not certain, which of the phosphor atoms actually bonds to the palladium center forming the catalyst complex. Therefore, the correct phosphor atom for the alignment cannot be identified. The approximation of using a random phosphor atom is sufficiently exact in this work as almost all of the affected molecules are symmetric.

A third limitation is the ASO as a descriptor. On the one hand the ASO is able to capture the steric environment of a molecule, on the other hand it neglects the influence of the elements within the molecule (except for the different van der Waals radii). Hence, one could possibly obtain the same ASO and therefore the same prediction, by exchanging atoms with similar van der Waals radii. It is the user's responsibility to provide reasonable geometries for the calculation and choose a suitable grid spacing. A real disadvantage of the ASO is that the grid used for the calculation has to be the same for training and validation. Consequently, the test molecules cannot be arbitrarily bigger than the molecules used for training, as they have to fit into the same grid. Of course a bigger grid can be used, if the test molecules are known prior to training.

5 CONCLUSION & OUTLOOK

In this thesis, the first steps in developing a high throughput screening workflow for predicting the enantiomeric excess (*ee*) of Negishi cross-coupling reactions are made. Negishi couplings are important C-C bond forming organic reactions that can be used as a tool for preparing stereoselective complex organic molecules. The need for stereoselective reactions arises from the different pharmacodynamic and -kinetic properties of enantiomers, which can be critical for the use of such enantiomers as drugs. To determine the optimal setup for reactions maximizing the *ee* and to ideally obtain an enantiopure product, a lot of experiments have to be carried out varying the reaction parameters. Machine learning (ML) can be a useful tool in automatizing this process and enabling a high throughput screening method for determining the optimal reaction parameters.

This work used experimental data of 190 Negishi coupling reactions to implement a fundamental workflow also applicable to larger sets of data. The workflow includes the generation of a conformer-rotamer-ensemble for every reactant capturing the steric environment, which is then represented by the Average Steric Occupancy (ASO) descriptor. Additionally it includes the setup of different ML algorithms, such as a neural network, random forest regression, and support vector regression models, to predict the *ee*. While the workflow in general is functioning, the results of the predictions are still insufficient. For a significant improvement of the model several further studies are required.

First, the generation of a larger training set enabled by a quantum chemical study of the reaction mechanism is required, which is currently work in progress in the group.

Second, another descriptor capturing the electrostatics might be necessary. This is indicated by the models inability to predict zero *ee*, even though reactions with zero *ee* are contained relatively often in the training set. For example, Denmark and co-workers additionally to the ASO descriptor used electronic descriptors derived from the perturbation that a catalyst substituent exerts on the electrostatic potential map of a quaternary ammonium ion.[34] This approach is not applicable in this work, as the ligands used do not differ just by substituents on a core structure, but cover a larger volume of the chemical space (see Figures A.1 to A.5 in the *Appendix*). Another grid-based descriptor, similar to the ASO, capturing electronic properties is wanted. The introduction of new descriptors will induce the need to adapt the feature reduction.

Third, the conformer space of every molecule is excessively sampled by the `crest` conformer search algorithm. This leads to a large database of over 260,000 ligand conformer structures for around 130 different ligands in total that all need to be considered, when calculation the ASO descriptor. The pure amount of structures is challenging in terms of RAM accessibility and computational costs. Possibilities to reduce the computational effort could be to only consider the conformers of the `crest` sampling (and not the rotamers) or reduce the energy window for the generation of conformers. Benchmarking regarding these two options is necessary.

Fourth, the generation of a large database of reactions is required. This database is on the one hand required for the generation of more training data via the quantum chemical study and on the other hand for the high throughput screening, which is the ultimate goal. Currently the database of reactions is composed of the 190 experiments as well as 20 hypothetical test reactions that were designed in consultation with the experimentalists. The future goal is to develop a tool that is able to generate an extensive database especially for the catalyst ligands. This goal can be achieved in different ways. In the work of Denmark and co-workers a general core structure for the catalyst is complemented by synthetically feasible substituents obtained from a database of readily available commercial sources or fragments that required no more than four well-established synthetic steps. The substituents are chosen by surveying catalogs of reagents that are compatible with the reaction conditions necessary to install the substituents and are in silico added to the predefined core structures using python scripts.[34] Gómez-Bombarelli *et al.* generated a database of 1.6 million candidates for OLED materials using in-house software relying on the RDKit package[73] by starting from a pool of fragments. These fragments are combined following a defined recipe and the resulting structures are pre-screened by a list of disallowed substructures, a synthetic accessibility score, and after the ML process the most favorable materials are rated by experimentalists on feasibility.[31] Lilienfeld and Corminboeuf in their work used a database with Simplified Molecular Input Line Entry System (SMILES) formats of 91 ligands in combination with six transition metals to form a database of catalysts (each with two ligands).[33] Generally there is a range of different formats and possibilities to generate a larger database, but some kind of chemical knowledge is necessary to choose core structures and substituents for ligands, the fragments composing the molecules, or the ligands of the catalysts at all.

Finally, the ML models might need to be adapted depending on their performance on more data and descriptors, but in general the implemented hyerparameter search and proposed models should be able to process these changes. Of course, there is a range of different ML algorithms that is not tested in this work and might be able to perform better on the presented training data.

Concluding, there is still a lot to be studied to reach the goal of predicting the *ee* of Negishi cross-coupling reactions with palladium catalysts just by the reactants and catalysts from a large database. Nevertheless, the first step in developing a high throughput screening method is made.

BIBLIOGRAPHY

[1] G. A. Khoury, R. C. Baliban, C. A. Floudas: Proteome-Wide Post-Translational Modification Statistics: Frequency Analysis and Curation of the Swiss-prot Database, *Sci. Rep.*, **1** (2011).

[2] K.-D. Thomann: Die Contergan-Katastrophe: Die trügerische Sicherheit der "harten" Daten, *Dtsch. Arztebl. International*, **104**, A (2007).

[3] T. Eriksson, S. Björkman, B. Roth, P. Höglund: Intravenous Formulations of the Enantiomers of Thalidomide: Pharmacokinetic and Initial Pharmacodynamic Characterization in Man, *J. Pharm. Pharmacol.*, **52**, 807 (2000).

[4] J. Duffus: Glossary for Chemists of Terms Used in Toxicology (IUPAC Recommendations 1993), *Pure Appl. Chem.*, **65**, 2003 (1993).

[5] M. Nordberg, J. Duffus, D. M. Templeton: Glossary of Terms Used in Toxicokinetics (IUPAC Recommendations 2003), *Pure Appl. Chem.*, **76**, 1033 (2004).

[6] A. G. Hutt, J. O'Grady: Drug Chirality: A Consideration of the Significance of the Stereochemistry of Antimicrobial Agents, *J. Antimicrob. Chemother.*, **37**, 7 (1996).

[7] R. F. Heck: Acylation, Methylation, and Carboxyalkylation of Olefins by Group VIII Metal Derivatives, *J. Am. Chem. Soc.*, **90**, 5518 (1968).

[8] R. F. Heck, J. P. Nolley: Palladium-Catalyzed Vinylic Hydrogen Substitution Reactions with Aryl, Benzyl, and Styryl Halides, *J. Org. Chem.*, **37**, 2320 (1972).

[9] H. A. Dieck, R. F. Heck: Organophosphinepalladium Complexes as Catalysts for Vinylic Hydrogen Substitution Reactions, *J. Am. Chem. Soc.*, **96**, 1133 (1974).

[10] S. Baba, E. Negishi: A Novel Stereospecific Alkenyl-Alkenyl Cross-Coupling by a Palladium- or Nickel-Catalyzed Reaction of Alkenylalanes with Alkenyl Halides, *J. Am. Chem. Soc.*, **98**, 6729 (1976).

[11] E. Negishi, A. O. King, N. Okukado: Selective Carbon-Carbon Bond Formation Via Transition Metal Catalysis. 3. a Highly Selective Synthesis of Unsymmetrical Biaryls and Diarylmethanes by the Nickel- or Palladium-Catalyzed Reaction of Aryl- and Benzylzinc Derivatives with Aryl Halides, *J. Org. Chem.*, **42**, 1821 (1977).

[12] N. Miyaura, A. Suzuki: Stereoselective Synthesis of Arylated (E)-Alkenes by the Reaction of Alk-1-enylboranes with Aryl Halides in the Presence of Palladium Catalyst, *J. Chem. Soc., Chem. Commun.*, 866 (1979).

[13] N. Miyaura, K. Yamada, A. Suzuki: A New Stereospecific Cross-Coupling by the Palladium-Catalyzed Reaction of 1-Alkenylboranes with 1-Alkenyl or 1-Alkynyl Halides, *Tetrahedron Lett.*, **20**, 3437 (1979).

[14] N. Miyaura, T. Yanagi, A. Suzuki: The Palladium-Catalyzed Cross-Coupling Reaction of Phenylboronic Acid with Haloarenes in the Presence of Bases, *Synth. Commun.*, **11**, 513 (1981).

[15] R. J. P. Corriu, J. P. Masse: Activation of Grignard Reagents by Transition-Metal Complexes. A New and Simple Synthesis of Trans-Stilbenes and Polyphenyls, *J. Chem. Soc., Chem. Commun.*, 144a (1972).

[16] K. Tamao, Y. Kiso, K. Sumitani, M. Kumada: Alkyl Group Isomerization in the Cross-Coupling Reaction of Secondary Alkyl Grignard Reagents with Organic Halides in the Presence of Nickel-phosphine Complexes as Catalysts, *J. Am. Chem. Soc.*, **94**, 9268 (1972).

[17] D. Azarian, S. S. Dua, C. Eaborn, D. R. M. Walton: Reactions of Organic Halides with R3 MMR3 Compounds (M = Si, Ge, Sn) in the Presence of Tetrakis(triarylphosphine)palladium, *J. Organomet. Chem.*, **117**, C55 (1976).

[18] M. Kosugi, K. Sasazawa, Y. Shimizu, T. Migita: Reactions of Allyltin Compounds III. Allylation of Aromatic Halides with Allyltributyltin in the Presence of Tetrakis(triphenylphosphine)palladium(0), *Chem. Lett.*, **6**, 301 (1977).

[19] D. Milstein, J. K. Stille: A General, Selective, and Facile Method for Ketone Synthesis from Acid Chlorides and Organotin Compounds Catalyzed by Palladium, *J. Am. Chem. Soc.*, **100**, 3636 (1978).

[20] Y. Hatanaka, T. Hiyama: Cross-Coupling of Organosilanes with Organic Halides Mediated by a Palladium Catalyst and Tris(diethylamino)sulfonium Difluorotrimethylsilicate, *J. Org. Chem.*, **53**, 918 (1988).

[21] Y. Hatanaka, T. Hiyama: Stereochemistry of the Cross-Coupling Reaction of Chiral Alkylsilanes with Aryl Triflates: A Novel Approach to Optically Active Compounds, *J. Am. Chem. Soc.*, **112**, 7793 (1990).

[22] C. C. C. Johansson-Seechurn, M. O. Kitching, T. J. Colacot, V. Snieckus: Palladium-Catalyzed Cross-Coupling: A Historical Contextual Perspective to the 2010 Nobel Prize, *Angew. Chem. Int. Ed.*, **51**, 5062 (2012).

[23] T. Hayashi, M. Tajika, K. Tamao, M. Kumada: High Stereoselectivity in Asymmetric Grignard Cross-Coupling Catalyzed by Nickel Complexes of Chiral (aminoalkylferrocenyl)phosphines, *J. Am. Chem. Soc.*, **98**, 3718 (1976).

[24] T. Hayashi, M. Konishi, T. Hioki, M. Kumada, A. Ratajczak, H. Niedbala: Preparation of (R)-N,N-dimethyl-1-[2-(diphenylphosphino)ferrocenyl]-2-propanamines and Asymmetric Grignard Cross-Coupling Catalyzed by Nickel Complexes with the Phosphine Ligands, *Bull. Chem. Soc. Jpn.*, **54**, 3615 (1981).

[25] T. Hayashi, M. Konishi, M. Fukushima, T. Mise, M. Kagotani, M. Tajika, M. Kumada: Asymmetric Synthesis Catalyzed by Chiral Ferrocenylphosphine-transition Metal Complexes. 2. Nickel- and Palladium-Catalyzed Asymmetric Grignard Cross-Coupling, *J. Am. Chem. Soc.*, **104**, 180 (1982).

[26] T. Hayashi, T. Hagihara, Y. Katsuro, M. Kumada: Asymmetric Cross-Coupling of Organozinc Reagents with Alkenyl Bromides Catalyzed by a Chiral Ferrocenylphosphine-palladium Complex, *Bull. Chem. Soc. Jpn.*, **56**, 363 (1983).

[27] T. Hayashi, M. Konishi, M. Fukushima, K. Kanehira, T. Hioki, M. Kumada: Chiral (.beta.-aminoalkyl)phosphines. Highly Efficient Phosphine Ligands for Catalytic Asymmetric Grignard Cross-Coupling, *J. Org. Chem.*, **48**, 2195 (1983).

[28] T. Hayashi: Catalytic Asymmetric Cross-Coupling, *J. Organomet. Chem.*, **653**, 41 (2002).

[29] T. Thaler, B. Haag, A. Gavryushin, K. Schober, E. Hartmann, R. M. Gschwind, H. Zipse, P. Mayer, P. Knochel: Highly Diastereoselective Csp3–Csp2 Negishi Cross-Coupling with 1,2-, 1,3- and 1,4-substituted Cycloalkylzinc Compounds, *Nat. Chem.*, **2**, 125 (2010).

[30] N. Maulide, A. Preinfalk, M. Simaan: In Preparation.

[31] R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. Baldo, R. P. Adams, A. Aspuru-Guzik: Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach, *Nat. Mater.*, **15**, 1120 EP (2016).

[32] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle: Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning, *Science*, **360**, 186 (2018).

[33] B. Meyer, B. Sawatlon, S. Heinen, O. A. von Lilienfeld, C. Corminboeuf: Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts, *Chem. Sci.*, **9**, 7069 (2018).

[34] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, S. E. Denmark: Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning, *Science*, **363**, eaau5631 (2019).

[35] M. García-Melchor, A. A. C. Braga, A. Lledós, G. Ujaque, F. Maseras: Computational Perspective on Pd-Catalyzed C–C Cross-Coupling Reaction Mechanisms, *Acc. Chem. Res.*, **46**, 2626 (2013).

[36] J. A. Casares, P. Espinet, B. Fuentes, G. Salas: Insights into the Mechanism of the Negishi Reaction: ZnRX versus ZnR_2Reagents, *J. Am. Chem. Soc.*, **129**, 3508 (2007).

[37] B. Fuentes, M. García-Melchor, A. Lledós, F. Maseras, J. Casares, G. Ujaque, P. Espinet: Palladium Round Trip in the Negishi Coupling of trans-[PdMeCl($PMePh_2$)$_2$] with ZnMeCl: An Experimental and DFT Study of the Transmetalation Step, *Chem. Eur. J.*, **16**, 8596 (2010).

[38] M. García-Melchor, B. Fuentes, A. Lledós, J. A. Casares, G. Ujaque, P. Espinet: Cationic Intermediates in the Pd-Catalyzed Negishi Coupling. Kinetic and Density Functional Theory Study of Alternative Transmetalation Pathways in the Me–Me Coupling of $ZnMe_2$ and trans-[PdMeCl($PMePh_2$)$_2$], *J. Am. Chem. Soc.*, **133**, 13519 (2011).

[39] C. Jimeno, S. Sayalero, T. Fjermestad, G. Colet, F. Maseras, M. Pericàs: Practical Implications of Boron-to-Zinc Transmetalation for the Catalytic Asymmetric Arylation of Aldehydes, *Angew. Chem. Int. Ed.*, **47**, 1098 (2008).

[40] S. T. Schneebeli, M. L. Hall, R. Breslow, R. Friesner: Quantitative DFT Modeling of the Enantiomeric Excess for Dioxirane-Catalyzed Epoxidations, *J. Am. Chem. Soc.*, **131**, 3965 (2009).

[41] M. Born, R. J. Oppenheimer: Zur Quantentheorie der Molekeln, *Ann. Phys.*, **389**, 457 (1927).

[42] B. M. Rode, T. S. Hofer, M. D. Kugler: *The Basics of Theoretical and Computational Chemistry*, Wiley VCH Verlag GmbH (2007).

[43] W. Koch, M. C. Holthausen: *A Chemist's Guide to Density Functional Theory: An Introduction*, Wiley-VCH (2000).

[44] S. Grimme, C. Bannwarth, P. Shushkov: A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1–86), *J. Chem. Theory Comput.*, **13**, 1989 (2017).

[45] W. Thiel: Semiempirical Quantum–Chemical Methods, *WIREs Comput. Mol. Sci.*, **4**, 145 (2013).

[46] C. Bannwarth, S. Ehlert, S. Grimme: GFN2-xTB—an Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions, *J. Chem. Theory Comput.*, **15**, 1652 (2019).

[47] E. Wigner: On the Quantum Correction for Thermodynamic Equilibrium, *Phys. Rev.*, **40**, 749 (1932).

[48] S. Mai, M. Richter, M. Heindl, M. F. S. J. Menger, A. Atkins, M. Ruckenbauer, F. Plasser, M. Oppel, P. Marquetand, L. González: SHARC2.0: Surface Hopping Including Arbitrary Couplings — Program Package for Non-adiabatic Dynamics, `https://sharc-md.org` (2018).

[49] R. Schinke: *Photodissociation Dynamics*, Cambridge University Press (1993).

[50] J. P. Dahl, M. Springborg: The Morse Oscillator in Position Space, Momentum Space, and Phase Space, *J. Chem. Phys.*, **88**, 4535 (1988).

[51] S. Grimme, C. Bannwarth, S. Dohm, A. Hansen, J. Pisarek, P. Pracht, J. Seibert, F. Neese: Fully Automated Quantum-Chemistry-Based Computation of Spin-Spin-Coupled Nuclear Magnetic Resonance Spectra, *Angew. Chem. Int. Ed.*, **56**, 14763 (2017).

[52] S. Grimme: Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-binding Quantum Chemical Calculations, *J. Chem. Theory Comput.*, **15**, 2847 (2019).

[53] P. Pracht, S. Grimme: In Preparation.

[54] J. Schmidhuber: Deep Learning in Neural Networks: An Overview, *Neural Networks*, **61**, 85 (2015).

[55] H. Drucker, C. C. Burges, L. Kaufman, A. J. Smola, V. N. Vapnik: Support Vector Regression Machines, in *Advances in Neural Information Processing Systems 9, Nips 1996*, 155–161, MIT Press (1997).

[56] C. Cortes, V. Vapnik: Support-Vector Networks, *Mach. Learn.*, **20**, 273 (1995).

[57] S. Theodoridis, K. Koutroumbas: *Pattern Recognition*, Academic Press (2008).

[58] H. Wang, D. Xu: Parameter Selection Method for Support Vector Regression Based on Adaptive Fusion of the Mixed Kernel Function, *J. Contr. Sci. Eng.*, **2017**, 1 (2017).

[59] T. K. Ho: Random Decision Forests, in *Proceedings of the Third International Conference on Document Analysis and Recognition (volume 1) - Volume 1*, ICDAR '95, 278, IEEE Computer Society, Washington, DC, USA (1995).

[60] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, A. Talwalkar: Hyperband: Bandit-Based Configuration Evaluation for Hyperparameter Optimization, in *International Conference on Learning Representations* (2017).

[61] S. Falkner, A. Klein, F. Hutter: BOHB: Robust and Efficient Hyperparameter Optimization at Scale, in *Proceedings of the 35th International Conference on Machine Learning*, 1437–1446, PMLR, Stockholmsmässan, Stockholm Sweden (2018).

[62] Denmark Lab Chemoinformatics: ccheminfolib, Project ID 8113486, https://gitlab.com/SEDenmarkLab/ccheminfolib, GitLab (2018).

[63] Jonathan Bennett and the AutoIt Team: Autoit, https://www.autoitscript.com/site/autoit (2019).

[64] W. Kabsch: A Solution for the Best Rotation to Relate Two Sets of Vectors, *Acta Crystallogr. Sect. A*, **32**, 922 (1976).

[65] W. Kabsch: A Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors, *Acta Crystallogr. Sect. A*, **34**, 827 (1978).

[66] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison: Open Babel: An Open Chemical Toolbox, *J. Cheminf.*, **3**, 33 (2011).

[67] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay: Scikit-learn: Machine Learning in Python, *JMLR*, **12**, 2825 (2011).

[68] University Bonn: xtb, version 6.1.3, please contact xtb@thch.uni-bonn.de for access to the program (2019).

[69] S. S. Batsanov: Intramolecular Contact Radii Close to the Van Der Waals Radii, *Zh. Neorg. Khim.*, **45**, 992 (2000).

[70] F. Chollet, et al.: Keras, https://keras.io (2015).

[71] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from https://tensorflow.org (2015).

[72] Y.-N. Wang, Q. Xiong, L.-Q. Lu, Q.-L. Zhang, Y. Wang, Y. Lan, W.-J. Xiao: Inverse-Electron-Demand Palladium-Catalyzed Asymmetric [4+2] Cycloadditions Enabled by Chiral P,S-Ligand and Hydrogen Bonding, *Angew. Chem. Int. Ed.*, **58**, 11013 (2019).

[73] RDKit: open source cheminformatics software, http://www.rdkit.org.

A

APPENDIX

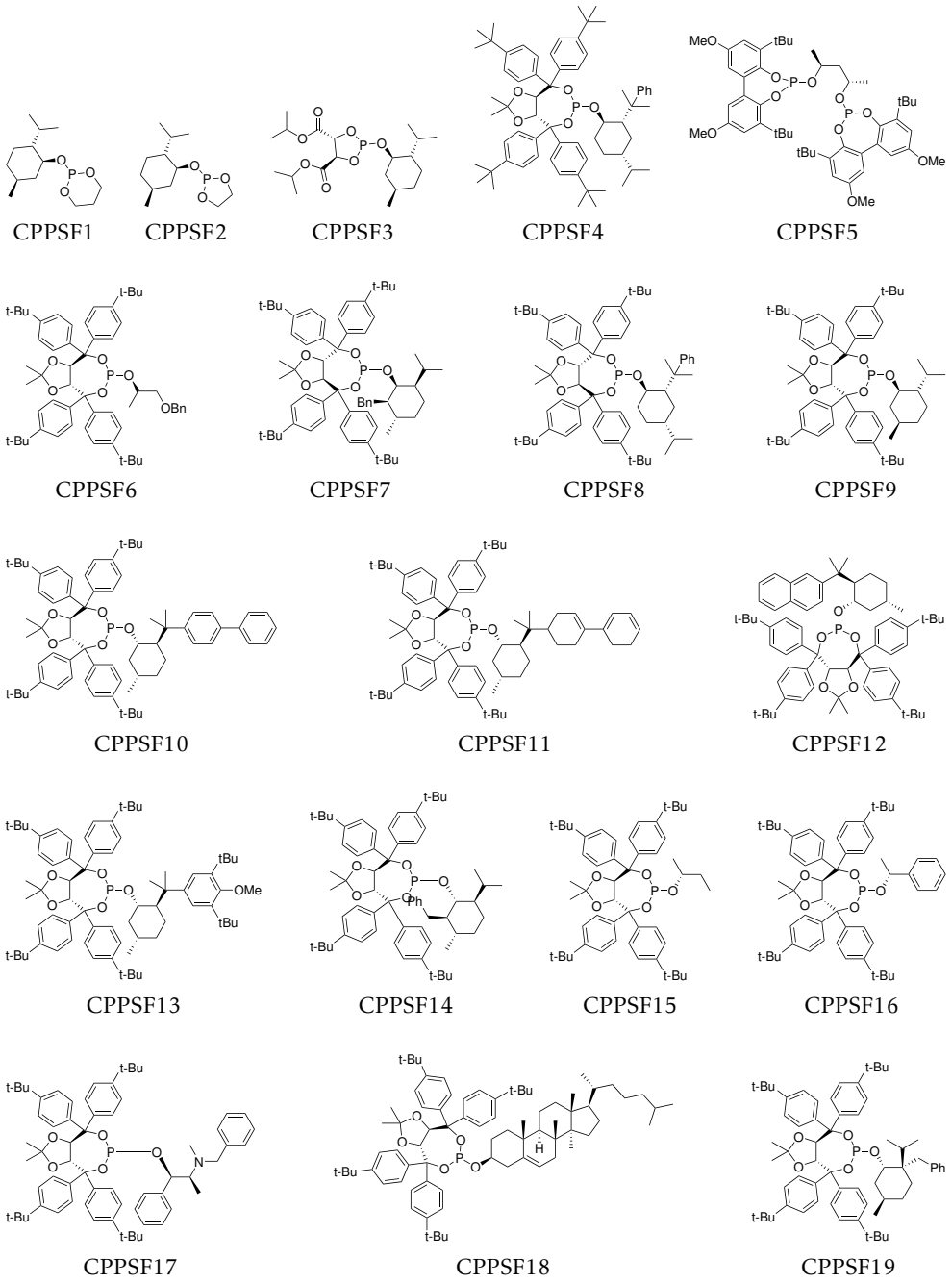


Figure A.1: Ligands CPPSF{1-19}.

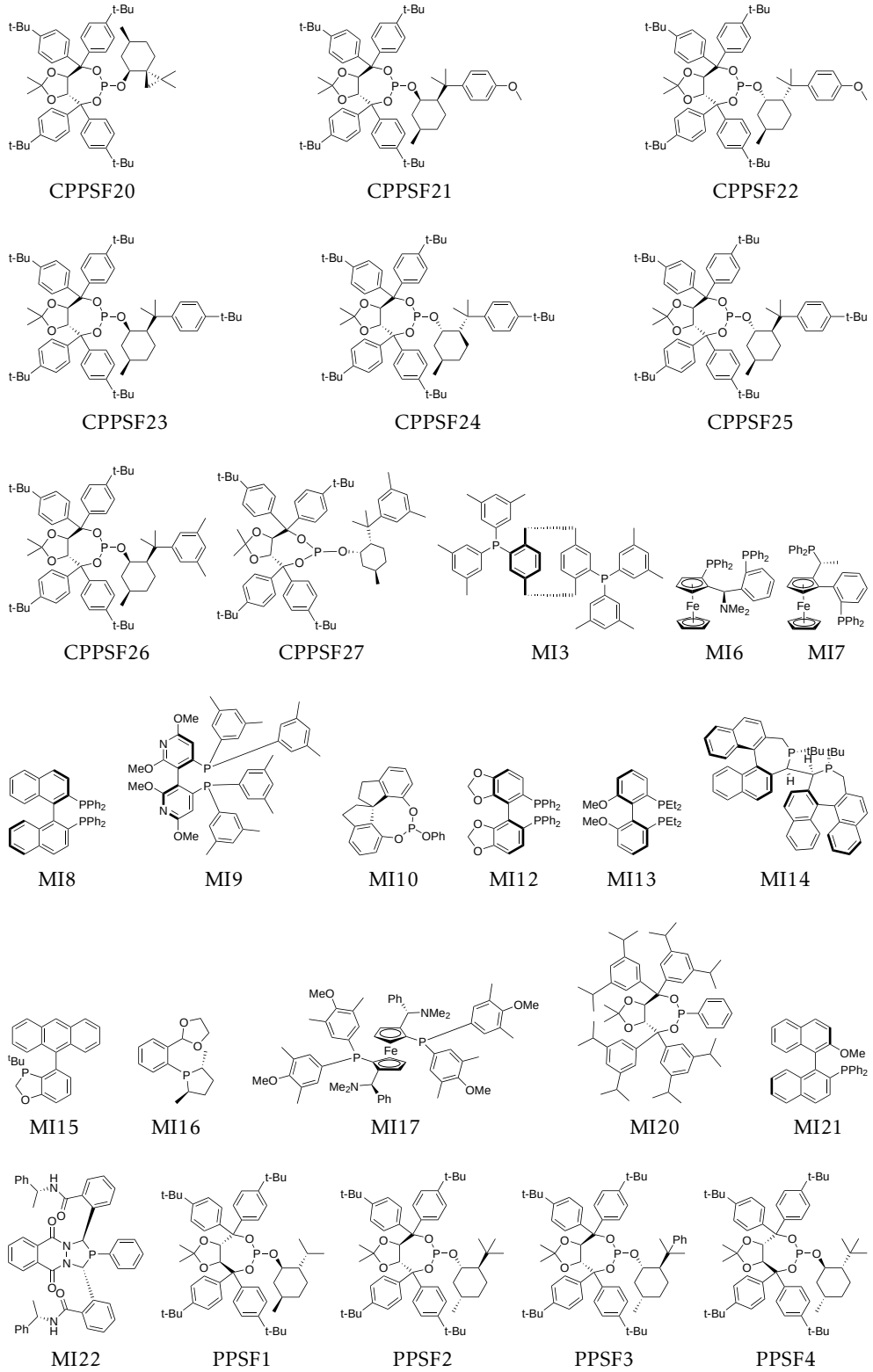


Figure A.2: Ligands CPPSF{20-27}, MI{3,6-10,12-17,20-22}, and PPSF{1-4}.

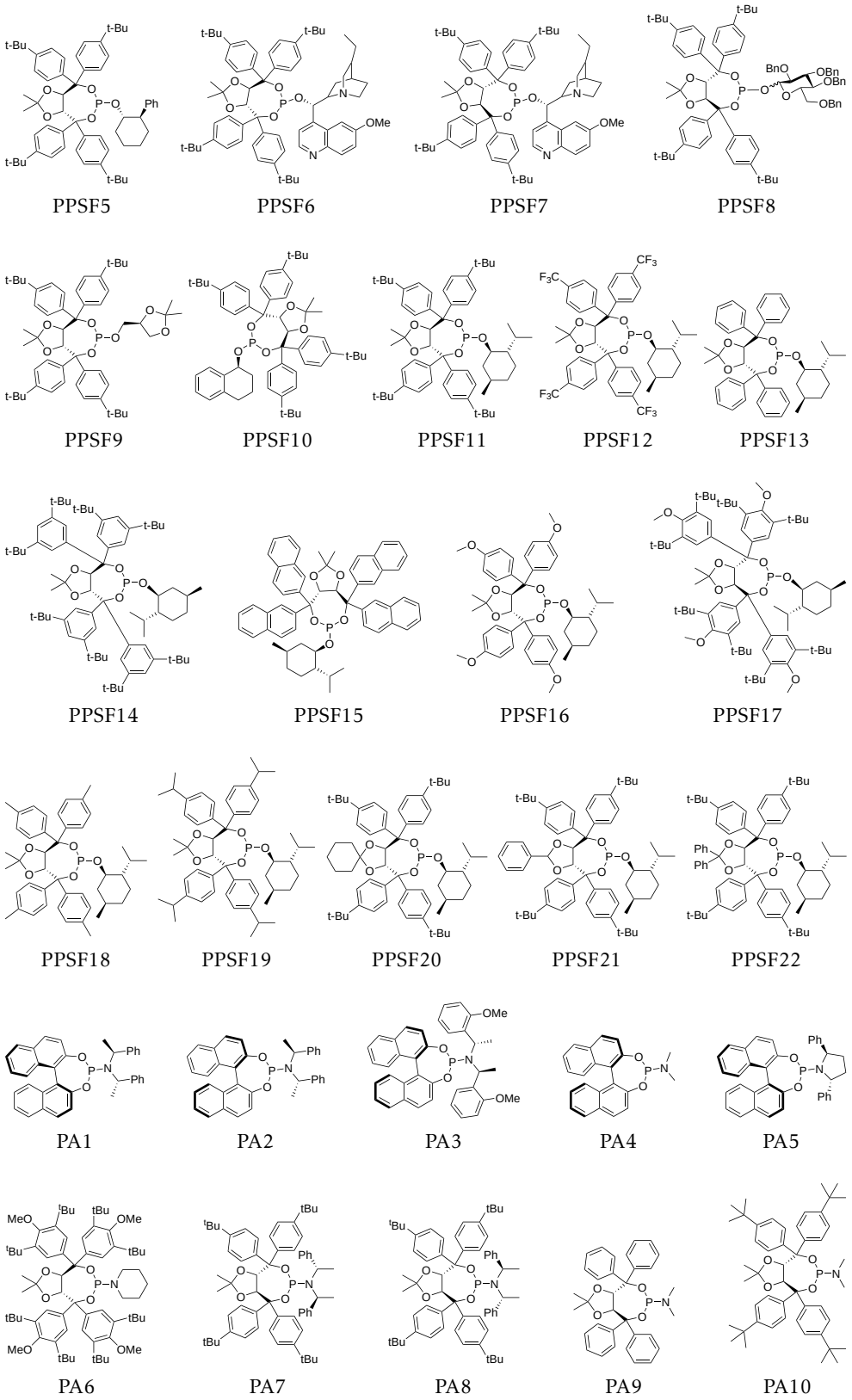


Figure A.3: Ligands PPSF{5-22} and PA{1-10}.

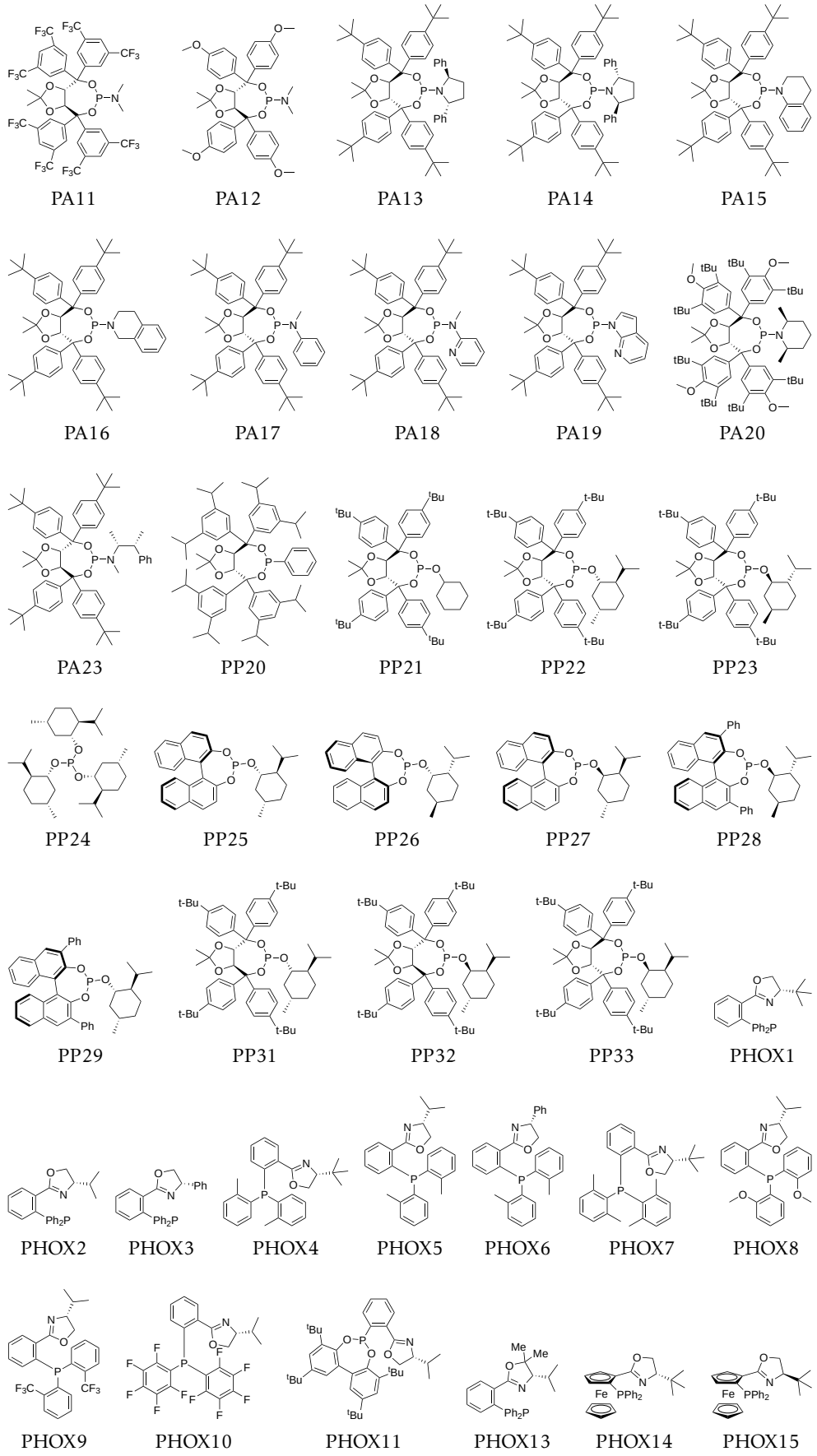


Figure A.4: Ligands PA{11-20,23}, PP{20-29,31-33}, and PHOX{1-11,13-15}.

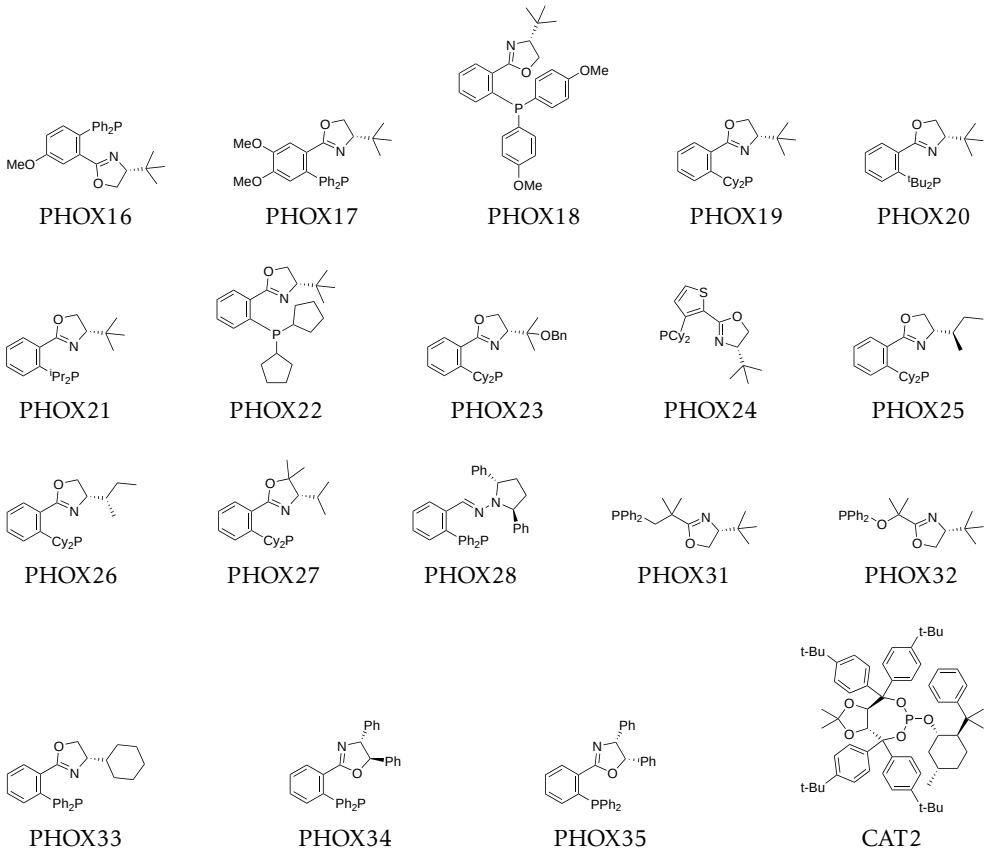


Figure A.5: Ligands PHOX{16-28,31-35} and CAT2.

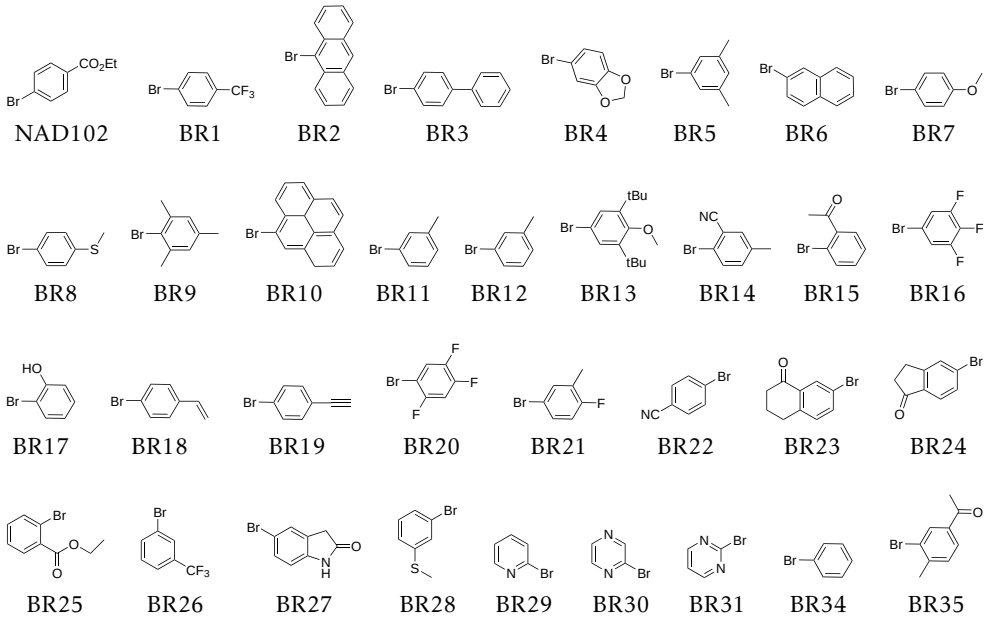
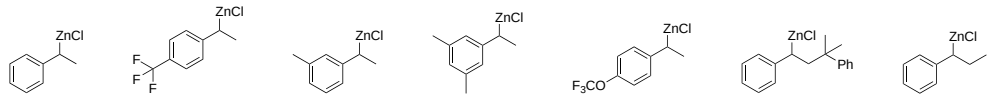


Figure A.6: Bromine reactants NAD102 and BR{1-31,34,35}.



NAD201 NAD202 NAD203 NAD204 NAD205 NAD206 NAD207

Figure A.7: Organozinc compounds NAD{201-207}.

Table A.1: Reaction mixtures and corresponding experimental enantiomeric excess values provided by the Maulide group.[30] The molecule labels correspond to Figures A.1 to A.7.

Reaction Mixture	Enantiomeric Excess
BR1_NAD201_CAT2	0.79
BR2_NAD201_CAT2	0.10
BR3_NAD201_CAT2	0.68
BR4_NAD201_CAT2	0.70
BR5_NAD201_CAT2	0.76
BR6_NAD201_CAT2	0.66
BR7_NAD201_CAT2	0.68
BR8_NAD201_CAT2	0.70
BR9_NAD201_CAT2	0.00
BR10_NAD201_CAT2	0.00
BR11_NAD201_CAT2	0.62
BR12_NAD201_CAT2	0.63
BR13_NAD201_CAT2	0.50
BR14_NAD201_CAT2	0.16
BR15_NAD201_CAT2	0.38
BR16_NAD201_CAT2	0.50
BR17_NAD201_CAT2	0.64
BR18_NAD201_CAT2	0.72
BR19_NAD201_CAT2	0.73
BR20_NAD201_CAT2	0.14
BR21_NAD201_CAT2	0.70
BR22_NAD201_CAT2	0.70
BR23_NAD201_CAT2	0.61
BR24_NAD201_CAT2	0.74
BR25_NAD201_CAT2	0.48
BR26_NAD201_CAT2	0.67
BR29_NAD201_CAT2	0.64
BR30_NAD201_CAT2	0.46
BR31_NAD201_CAT2	0.54
BR26_NAD201_CAT2	0.67
BR27_NAD201_CAT2	0.69
BR28_NAD201_CAT2	0.66
BR18_NAD202_CAT2	0.69
BR19_NAD202_CAT2	0.67
BR21_NAD202_CAT2	0.89
BR16_NAD202_CAT2	0.66

BR11_NAD202_CAT2	0.75
BR35_NAD202_CAT2	0.23
BR26_NAD202_CAT2	0.80
BR5_NAD202_CAT2	0.50
BR4_NAD202_CAT2	0.68
BR8_NAD202_CAT2	0.66
BR6_NAD202_CAT2	0.67
NAD102_NAD202_CAT2	0.71
BR34_NAD202_CAT2	0.78
BR3_NAD202_CAT2	0.66
BR7_NAD202_CAT2	0.70
BR19_NAD203_CAT2	0.68
BR5_NAD203_CAT2	0.64
NAD102_NAD203_CAT2	0.66
BR1_NAD203_CAT2	0.78
BR20_NAD203_CAT2	0.23
BR21_NAD203_CAT2	0.67
BR16_NAD203_CAT2	0.54
BR34_NAD204_CAT2	0.69
NAD102_NAD204_CAT2	0.50
BR1_NAD204_CAT2	0.50
NAD102_NAD205_CAT2	0.62
NAD102_NAD206_CAT2	0.68
NAD102_NAD207_CAT2	0.70
NAD102_NAD201_PHOX1	0.76
NAD102_NAD201_PHOX2	0.64
NAD102_NAD201_PHOX3	0.12
NAD102_NAD201_PHOX4	0.50
NAD102_NAD201_PHOX5	0.51
NAD102_NAD201_PHOX6	0.00
NAD102_NAD201_PHOX7	0.11
NAD102_NAD201_PHOX8	0.56
NAD102_NAD201_PHOX9	0.07
NAD102_NAD201_PHOX10	0.20
NAD102_NAD201_PHOX11	0.20
NAD102_NAD201_PHOX13	0.72
NAD102_NAD201_PHOX14	0.57
NAD102_NAD201_PHOX15	0.40
NAD102_NAD201_PHOX16	0.62
NAD102_NAD201_PHOX17	0.58
NAD102_NAD201_PHOX18	0.61
NAD102_NAD201_PHOX19	0.49
NAD102_NAD201_PHOX20	0.08
NAD102_NAD201_PHOX21	0.32
NAD102_NAD201_PHOX22	0.39
NAD102_NAD201_PHOX23	0.42
NAD102_NAD201_PHOX24	0.00
NAD102_NAD201_PHOX25	0.34
NAD102_NAD201_PHOX26	0.00
NAD102_NAD201_PHOX27	0.15

NAD102_NAD201_PHOX28	0.05
NAD102_NAD201_PHOX31	0.54
NAD102_NAD201_PHOX32	0.03
NAD102_NAD201_PHOX33	0.53
NAD102_NAD201_PHOX34	0.17
NAD102_NAD201_PHOX35	0.16
NAD102_NAD201_PA1	0.11
NAD102_NAD201_PA2	0.04
NAD102_NAD201_PA3	0.16
NAD102_NAD201_PA4	0.00
NAD102_NAD201_PA5	0.00
NAD102_NAD201_PA6	0.02
NAD102_NAD201_PA7	0.02
NAD102_NAD201_PA8	0.02
NAD102_NAD201_PA9	0.36
NAD102_NAD201_PA10	0.12
NAD102_NAD201_PA11	0.04
NAD102_NAD201_PA12	0.00
NAD102_NAD201_PA13	0.36
NAD102_NAD201_PA14	0.20
NAD102_NAD201_PA15	0.06
NAD102_NAD201_PA16	0.21
NAD102_NAD201_PA17	0.13
NAD102_NAD201_PA18	0.00
NAD102_NAD201_PA19	0.00
NAD102_NAD201_PA20	0.10
NAD102_NAD201_PA23	0.08
NAD102_NAD201_MI3	0.24
NAD102_NAD201_MI6	0.00
NAD102_NAD201_MI7	0.00
NAD102_NAD201_MI8	0.10
NAD102_NAD201_MI9	0.38
NAD102_NAD201_MI10	0.26
NAD102_NAD201_MI12	0.10
NAD102_NAD201_MI13	0.06
NAD102_NAD201_MI14	0.00
NAD102_NAD201_MI15	0.08
NAD102_NAD201_MI16	0.00
NAD102_NAD201_MI17	0.23
NAD102_NAD201_MI20	0.10
NAD102_NAD201_MI21	0.00
NAD102_NAD201_MI22	0.04
NAD102_NAD201_PP20	0.10
NAD102_NAD201_PP21	0.00
NAD102_NAD201_PP22	0.26
NAD102_NAD201_PP23	0.38
NAD102_NAD201_PP24	0.08
NAD102_NAD201_PP25	0.10
NAD102_NAD201_PP26	0.08
NAD102_NAD201_PP27	0.10

NAD102_NAD201_PP28	0.00
NAD102_NAD201_PP29	0.00
NAD102_NAD201_PP31	0.26
NAD102_NAD201_PP32	0.36
NAD102_NAD201_PP33	0.18
NAD102_NAD201_PPSF1	0.38
NAD102_NAD201_PPSF2	0.52
NAD102_NAD201_PPSF3	0.52
NAD102_NAD201_PPSF4	0.30
NAD102_NAD201_PPSF5	0.38
NAD102_NAD201_PPSF6	0.24
NAD102_NAD201_PPSF7	0.00
NAD102_NAD201_PPSF8	0.30
NAD102_NAD201_PPSF9	0.12
NAD102_NAD201_PPSF10	0.10
NAD102_NAD201_PPSF11	0.50
NAD102_NAD201_PPSF12	0.02
NAD102_NAD201_PPSF13	0.27
NAD102_NAD201_PPSF14	0.07
NAD102_NAD201_PPSF15	0.09
NAD102_NAD201_PPSF16	0.10
NAD102_NAD201_PPSF17	0.04
NAD102_NAD201_PPSF18	0.11
NAD102_NAD201_PPSF19	0.33
NAD102_NAD201_PPSF20	0.49
NAD102_NAD201_PPSF21	0.30
NAD102_NAD201_PPSF22	0.44
NAD102_NAD201_CPPSF1	0.00
NAD102_NAD201_CPPSF2	0.07
NAD102_NAD201_CPPSF3	0.07
NAD102_NAD201_CPPSF4	0.49
NAD102_NAD201_CPPSF5	0.00
NAD102_NAD201_CPPSF6	0.02
NAD102_NAD201_CPPSF7	0.00
NAD102_NAD201_CPPSF8	0.78
NAD102_NAD201_CPPSF9	0.58
NAD102_NAD201_CPPSF10	0.66
NAD102_NAD201_CPPSF11	0.60
NAD102_NAD201_CPPSF12	0.67
NAD102_NAD201_CPPSF13	0.62
NAD102_NAD201_CPPSF14	0.32
NAD102_NAD201_CPPSF15	0.00
NAD102_NAD201_CPPSF16	0.00
NAD102_NAD201_CPPSF17	0.22
NAD102_NAD201_CPPSF18	0.00
NAD102_NAD201_CPPSF19	0.40
NAD102_NAD201_CPPSF20	0.22
NAD102_NAD201_CPPSF21	0.24
NAD102_NAD201_CPPSF22	0.40
NAD102_NAD201_CPPSF23	0.26

NAD102_NAD201_CPPSF24	0.40
NAD102_NAD201_CPPSF25	0.60
NAD102_NAD201_CPPSF26	0.26
NAD102_NAD201_CPPSF27	0.44

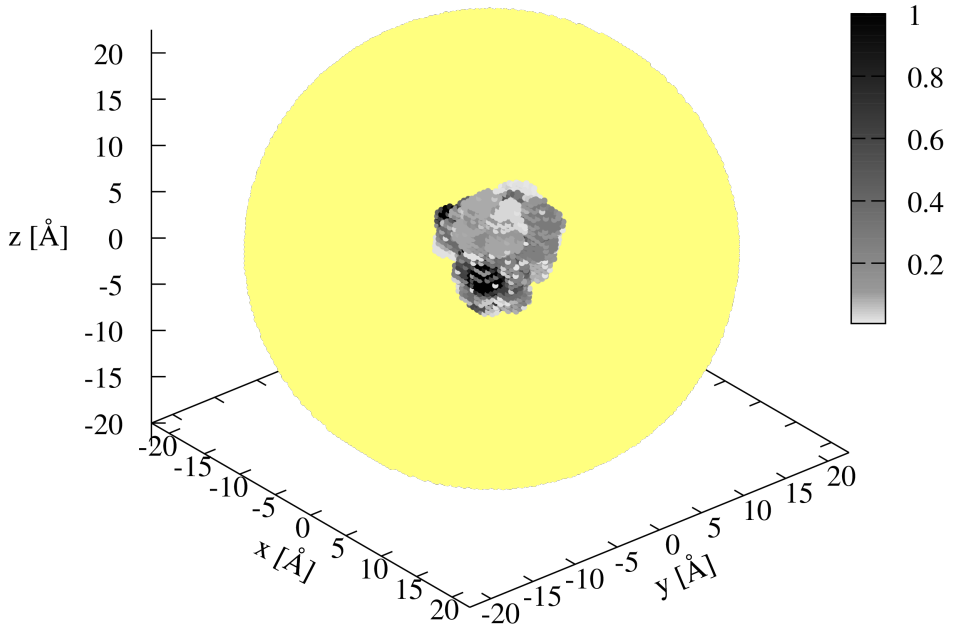
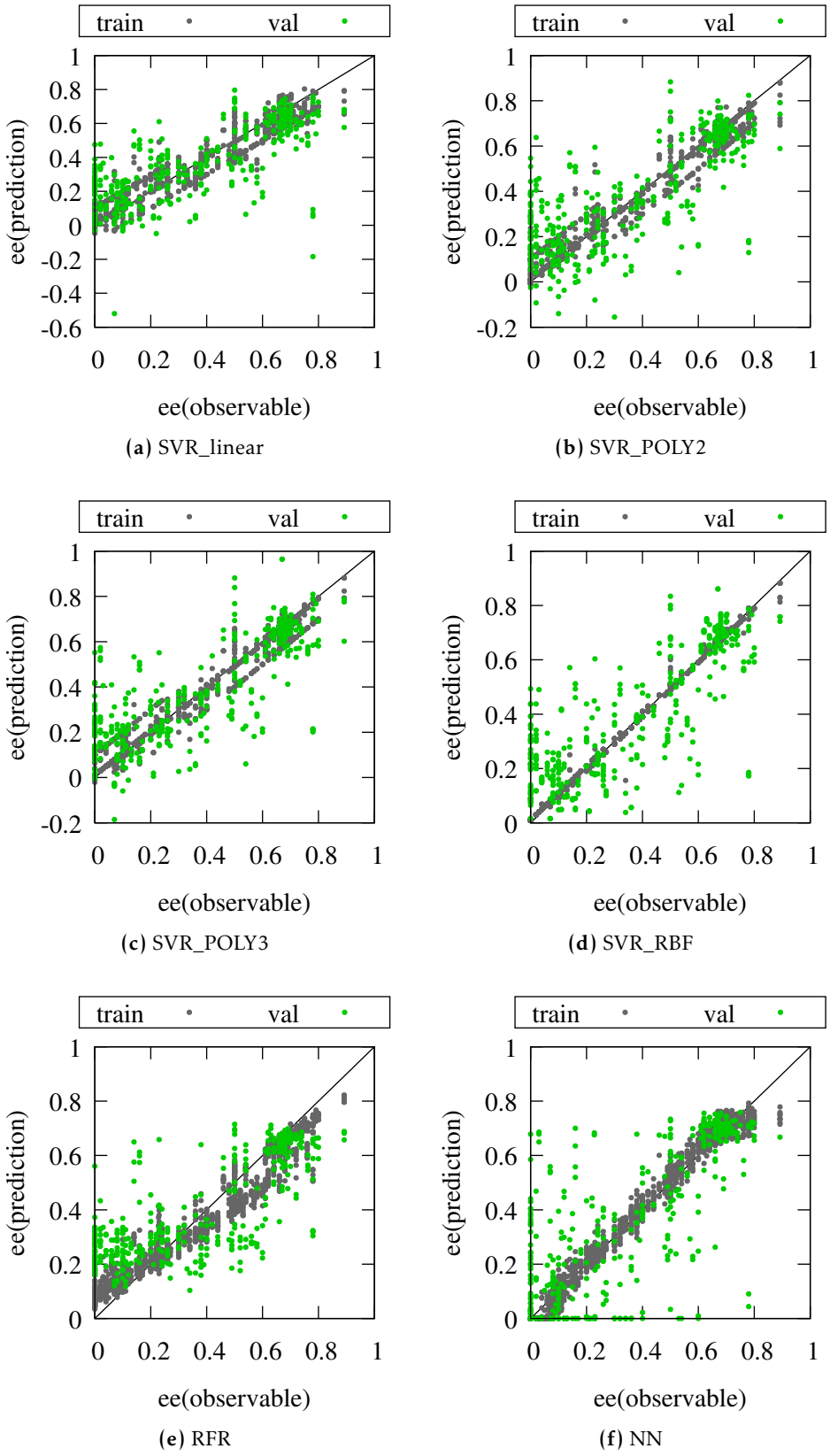


Figure A.8: Three-dimensional plot of the ASO > 0 of PHOX2 with the outline of the total grid in yellow in the background.



(a) SVR_linear **(b)** SVR_POLY2

(c) SVR_POLY3 **(d)** SVR_RBF

(e) RFR **(f)** NN

Figure A.9: Plots of the observed enantiomeric excess of the training and validation set against the predicted *ee* for six different models and 10 replicate runs.

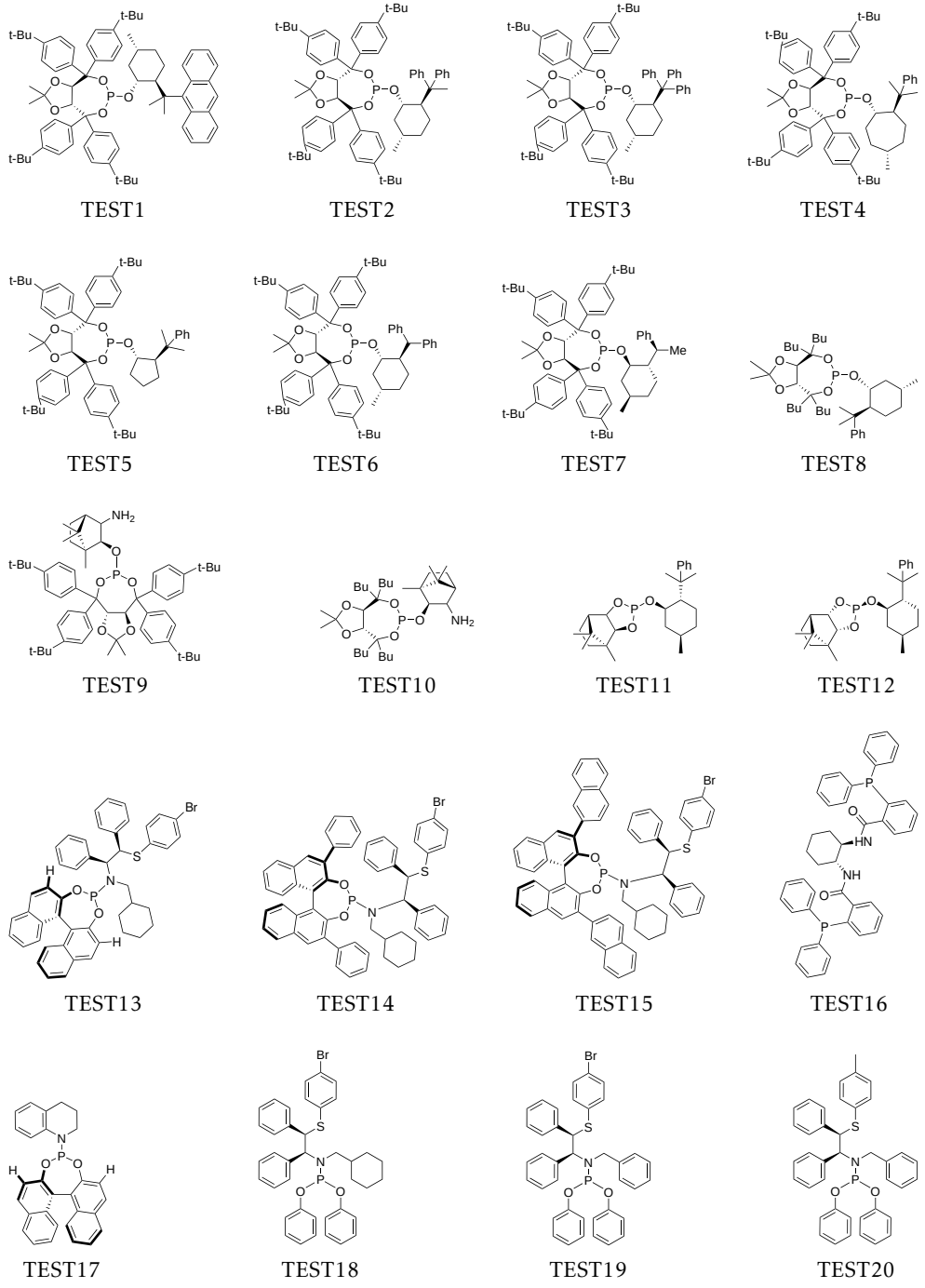


Figure A.10: Unknown ligands TEST1 to TEST20 of the reaction in Figure 4.11, which are used for predicting the enantiomeric excess in Section 4.2.2.

ABSTRACT

The design of stereoselective reactions has traditionally been driven by trial and error synthesis. The empirical nature of this approach can potentially be accelerated by a high throughput screening workflow with machine learning. This work proposes a workflow designed for the prediction of the enantiomeric excess of Negishi cross-coupling reactions, which are important C-C bond forming reactions with enantioselective potential. The workflow starts with handling experimental data, generating a conformer-rotamer-ensemble, transforming it with a descriptor to a machine readable format, and ends with training and validating different machine learning algorithms. The used descriptor is the Average Steric Occupancy (ASO), which represents the steric environment of molecules. The ASO is further analyzed in the work and the utilization of concatenated ASOs for reaction mixtures is discussed. Furthermore, different machine learning models, including neural networks, support vector machines, and random forest regression models, are evaluated and used for speculative predictions of the enantiomeric excess of novel reactions. Finally the limitations of the workflow are investigated and plans for further studies are proposed.

ZUSAMMENFASSUNG

Stereoselektive Reaktionen werden traditionell durch Versuch und Irrtum entwickelt. Der empirische Charakter dieser Methodik kann potenziell durch einen Hochdurchsatz-Selektions-Arbeitsablauf mit maschinellem Lernen beschleunigt werden. Diese Arbeit schlägt einen Arbeitsablauf vor, der für die Vorhersage des Enantiomerenüberschusses von Negishi-Kreuzkupplungsreaktionen entwickelt wurde. Negishi-Reaktionen sind wichtige C-C-Bindungsbildungsreaktionen mit enantioselektivem Potenzial. Der Arbeitsablauf beginnt mit der Handhabung experimenteller Daten, geht über die Generierung eines Konformer-Rotamer-Ensembles und dessen Transformation durch einen Deskriptor in ein maschinenlesbares Format und endet mit dem Trainieren und Validieren verschiedener Algorithmen des maschinellen Lernens. Der verwendete Deskriptor heißt Average Steric Occupancy (ASO) und gibt die sterische Umgebung von Molekülen wieder. Der ASO Deskriptor wird in der Arbeit weiter analysiert und die Verwendung von zusammengesetzten ASOs für Reaktionsgemische wird diskutiert. Darüber hinaus werden verschiedene maschinelle Lernmodelle, einschließlich neuronaler Netzwerke, Unterstützungsvektormaschinen und Zufallswald-Regressionsmodelle, ausgewertet und für spekulative Vorhersagen des Enantiomerenüberschusses unbekannter Reaktionen verwendet. Schließlich werden die Grenzen des vorgeschlagenen Arbeitsablaufs untersucht und ein Ausblick für weitere potentielle Studien wird gegeben.