DISS. ETH NO. 26650

Chemical Decision Making: An exploration into the modeling and quantification of modern drug discovery

A thesis submitted to attain the degree of

Doctor of Sciences of ETH Zurich
(Dr. sc. ETH Zurich)

presented by

Alexander Luke Button

Bachelor of Science Awarded with First Class Honours in Chemistry from the University of Adelaide

born on *10.10.1991*

citizen of
Australia
United Kingdom of Great Britain and Northern Ireland

accepted on the recommendation of

Prof. Dr. Gisbert Schneider
Prof. Dr. Jonathan Hall

2020

"*Give me a lever long enough and a fulcrum far enough and I will move the world.*"

Archimedes

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. Dr. Gisbert Schneider, for his boundless support, enthusiasm, and encouragement. Over the course of this project, my understanding and capabilities, not only in the field of pharmaceutical science, but in life in general, have vastly improved thanks to you. The opportunity to be a member of the Schneider lab has truly been a privilege. Thank you Gisbert, for always challenging me to become a better scientist.

I would like to thank Dr. Jan Hiss for his support inside and outside of the lab. It is hard to express the incalculable ways in which Jan has been helpful. I would like to thank Dr. Daniel Merk for the enriching collaboration. I have the upmost respect for your professionalism and honesty, and am very grateful to have had the opportunity to work with you. Thank you Sarah Haller, for all of the laboratory assistance you gave and for always creating a warm and welcoming office environment.

I would like to extend the deepest and warmest thanks to the entire Schneider lab. Thank you for all the fun, laughter, and mistakes that we made together. I would like to thank Cyrill Brunner and Berend Huisman. Thank you guys for sticking with me as we attempted the impossible. Working with you has been an incredibly rewarding experience. Thank you to Alice and Benedikt, for accomplishing the impossible on my behalf. I say without hyperbole that you were the best Master's students I could have hoped for. I would like to thank Ryan, JJ (XueZhang), Domi, Michael, Damian, Francesca, and Gisela for not only being good lab-mates, but also great friends. The time that we have spent together has meant the world to me.

I would like to thank my family, my mother and father, and my sister Lizzie. Thank you for the unending love and support that you have always given me, and that I hope I have returned in kind. Lastly, I would like to thank Jolanda. She knows why.

Contents

List of Figures

List of Tables

List of Abbreviations

A

AI	Artificial Intelligence
ALK	Anaplastic Lymphoma Kinase
ANN	Artificial Neural Network
ATP	Adenosine Triphosphate
AUC	Area Under Curve

B

BGEN	Building Block Generation

C

CAMP	Cyclic Adenosine Monophosphate
CATS	Chemically Advanced Template Search
CBS	4-sulfamoylbenzoic acid
ChEMBL	Chemical Database of Bioactive Molecules with Drug-like Properties
CHO	Chinese Hamster Ovary
CPU	Central Processing Unit

D

DINGOS	Design of Innovative NCEs Generated through Optimization Strategies
DMAP	4-Dimethylaminopyridine
DMF	Dimethylformamide
DMSO	Dimethyl Sulfoxide
DOI	Domain Of Interest

E

ECFP	Extended Connectivity Fingerprint
EDC	1-Ethyl-3-(3-Dimethylaminopropyl)Carbodiimide
EGFR	Epidermal Growth Factor Receptor
ESI	Electrospray Ionization

F

FDA	Food and Drug Administration
FGI	Functional Group Interconversions

G

GTPγS	Guanosine 5'-O-[gamma-thio]triphosphate

H

HBA	Hydrogen Bond Acceptor
HBD	Hydrogen Bond Donor

HEK293	Human Embryonic Kidney 293 Cells
HPLC	High Performance Liquid Chromatography
HRMS	High Resolution Mass Spectroscopy
HTRF	Homogeneous Time Resolved Fluorescence

I

IQR	Inter-Quartile Range

J

-	-

K

KNIME	Konstanz Information Miner

L

LCMS	Liquid Chromatography-Mass Spectrometry
LSTM	Long short-term memory network

M

MACCS	Molecular Access System
MCTS	Monte Carlo Tree Search
MeCN	Acetonitrile
MeOH	Methanol
MLP	Multi-Layer Perceptron
MOE	Molecular Operating Environment
MS	Mass Spectroscopy
MW	Molecular Weight

N

NCE	New Chemical Entity
NMR	Nuclear Magnetic Resonance
NSCLC	Non-Small Cell Lung Cancer

O

-	-

P

ppm	parts per million
PPAR	Peroxisome Proliferator-Activated Receptor

Q

QSAR	Quantitative Structure Activity Relationship

R

RDKit	Rational Discovery Kit
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic Curves
RXR	Retinoid X Receptor

S

SAR	Structure Activity Relationship
S.E.M.	Standard Error of Mean
SMARTS	SMILES Arbitrary Target Specification
SMILES	Simplified Molecular-Input Line-Entry System
SPiDER	Self-organizing map-based Prediction of Drug Equivalence Relationship
SPR	Surface Plasmon Resonance

T

THF	Tetrahydrofuran
TMS	Tetramethylsilane

U

USPTO	United States Patent and Trademark Office
UV-VIS	UltraViolet-visible Spectroscopy

V

-	-

W

-	-

X

-	-

Y

-	-

Z

ZINC	ZINC Is Not Commercial

Abstract

Drug discovery is the process of producing compounds with desired biological properties towards diseases of interest. Considering the enormous number of potential structures one could produce, the achievements seen in modern pharmaceutical science are astounding. Despite this, the immense amount of resources required to develop a drug compound presents a bottleneck in the development of pharmaceutical compounds. Computational methods, such as machine learning, offer a potential solution to this problem. Their major advantage is that, by proposing structures *in silico*, they can narrow the search space, reducing the number of compounds required to be synthesized and tested. However, the use of such methods is often encumbered with issues of synthetic feasibility.

To aid in this issue, we present the algorithm DINGOS (Design of Innovative NCEs Generated through Optimization Strategies). DINGOS combines the predictive capabilities of computational methods with a rule-based model of synthesizability. DINGOS is a ligand-based scoring method, generating structures which are similar to the template ligand, while within the provided definition of synthesizability. The functionality of the DINGOS algorithm was demonstrated in a case-study, in which designs were proposed for four compounds (alectinib, cariprazine, osimertinib, and pimavanserin). For each template, a set of 300 designs was proposed, out of which, above 50% were predicted as active by target prediction. Three out of four selected candidate structures were successfully synthesized with the synthetic pathway predicted by DINGOS. Of three synthesized compounds, one showed activity towards the 5-HT_{2B} serotonin receptor. The modular nature of the DINGOS algorithm was demonstrated in a series of follow-up studies. In the first follow-up study, DINGOS was paired with a custom automated synthesis system in order to generate *de novo* designs within an autonomous active learning drug design cycle. Two rounds of automated synthesis were performed, generating a total of 22 novel compounds, of which five showed micromolar activity towards carbonic anhydrase II. The method was further extended, first by modifying the core predictive component of the algorithm, allowing for the internal representation of the scoring function to be flexibly changed, and in the second, DINGOS was combined with a MCTS (Monte Carlo tree search) algorithm in order to allow for non-greedy optimization of the designed structures. The results of this work showcase the potential of the DINGOS algorithm as a method for generating synthetically feasible *de novo* designs, and highlights DINGOS' ability to be adapted to a variety of different drug design problems, narrowing the gap between *in silico* and experimental drug design.

Zusammenfassung

Die Entdeckung und Entwicklung von Medikamenten beschreibt den Prozess, chemische Verbindungen herzustellen, die gewünschte biologische Eigenschaften gegenüber der untersuchten Krankheiten aufweisen. Angesichts der beträchtlichen Anzahl möglicher Strukturen, die hergestellt werden können, sind die Errungenschaften der modernen pharmazeutischen Wissenschaft beeindruckend. Dennoch stellt der grosse zeitliche Aufwand und der hohe Bedarf an Ressourcen eine grosse Herausforderung in der Entwicklung von pharmazeutischen Wirkstoffen dar. Numerische Methoden, wie zum Beispiel *Machine Learning*, ermöglichen neue Lösungsansätze für diese Problematik. Ihre Stärke besteht darin, dass sie neuartige chemische Strukturen *in silico* vorschlagen können und so den Suchraum eingrenzen. Dadurch wird die Anzahl chemischer Verbindungen reduziert, die synthetisiert und getestet werden müssen. Allerdings bringt diese Herangehensweise keine Garantie für die Synthetisierbarkeit der vorgeschlagenen Moleküle mit sich.

Um diese Problematik anzugehen, präsentieren wir den Algorithmus DINGOS (*Design of Innovative NCEs Generated through Optimization Strategies*). DINGOS vereint die Fähigkeit der numerischen Methoden, geeignete chemische Sturkturen zu empfehlen, mit einem regelbasierten Modell, das die Synthetisierbarkeit jener Strukturen überprüft. DINGOS beruht auf einem ligandenbasierten Bewertungsverfahren, das einerseits Strukturen erzeugt, die dem ursprünglichen Liganden ähnlich sind, und auf der anderen Seite sicherstellt, dass diese der Definition von Synthetisierbarkeit entsprechen, die vom Benutzer vorgegebenen wurde. Die Funktionsweise des DINGOS Algorithmus wurde anhand einer Fallstudie demonstriert, in der *de novo* Designs für vier Wirkstoffe (Alectinib, Cariprazin, Osimertinib und Pimavanserin) vorgeschlagen wurden. Für jede dieser vier Vorlagen wurde ein Satz von 300 Designs von DINGOS erzeugt, wovon durch *Target Prediction* für 50% Aktivität gegenüber dem Zielmolekül des Vorlage-Liganden vorhergesagt wurde. Drei der vier ausgewählten Struktur-Kandidaten wurden erfolgreich mit Hilfe desjenigen Syntheseweges synthetisiert, der von DINGOS vorgeschlagen wurde. Davon wiederum zeigte ein *de novo* Wirkstoff Aktivität gegenüber dem 5-HT_{2B} Serotonin-Rezeptor. Der modulare Charakter des DINGOS-Algorithmus wurde in einer Reihe von Folgestudien demonstriert. In der ersten Folgestudie wurde DINGOS mit einem hauseigenen automatisierten Synthesesystem kombiniert, um *de novo* Designs innerhalb eines autonomen, aktiven Wirkstoffdesignlernzykluses zu generieren. Die automatisierte Synthese generierte in zwei Durchgängen insgesamt 22 neue chemische Verbindungen, von denen fünf mikromolare Aktivität gegenüber Carbonic Anhydrase II zeigten. Desweiteren wurde einerseits eine Modifizierung der prädikativen Kernkomponente von DINGOS durchgeführt, um eine flexible Anpassung des internen Bewertungsverfahrens zu ermöglichen. Andererseits wurde DINGOS mit dem *Monte Carlo Tree Search* Verfahren erweitert, was eine nicht-gierige (*non-greedy*) Optimierung der erstellten Strukturen ermöglicht. Die Ergebnisse dieser Arbeit präsentieren das Potenzial des DINGOS Algorithmus als Methode zur Generierung synthetisch realisierbarer *de novo*-Designs und unterstreichen seine Fähigkeit, sich an eine Vielzahl unterschiedlicher Probleme des Wirkstoffdesigns anzupassen und die Lücke zwischen *in silico* und experimentellem Wirkstoffdesign zu verkleinern.

Chapter 1

Introduction - Computer Assisted Drug Designs, Machine Learning, and Research Problem

1.1 The Problem of Drug Design

The task of drug discovery, in its simplest form, is the problem of how best to filter the vast expanse of chemical space to leave only those chemical entities with the desired biological activity and absorption, distribution, metabolism, and excretion (ADMET) properties [1, 2]. While the exact value is disputed, estimates place the number of possible drug-like structures somewhere between 10^{18} and 10^{200} [3, 4]. To put this in context, mankind has synthesized an estimated 10^{8} molecules [5], or between 10^{-8} and 10^{-190} % of the total number of potential candidate structures. In early stage drug discovery, one is often primarily concerned with the pharmacodynamic properties of a molecule, usually manifesting as a binding interaction towards an identified target protein of interest [6]. Efficient exploration of chemical space as a means of locating regions with favourable binding interactions is a problem of much interest. Numerous techniques have been developed in order to accomplish this task [7–9].

1.1.1 Natural Product Inspired Drug Design

Natural products have historically been a rich source for drug structures [10, 11], with the use of chemical species derived from natural sources dating back thousands of years. A textbook example of this would be the compound of acetylsalicylic acid (Aspirin) Figure 1.1, which was originally developed from willow bark [12], while a more contemporary example would be the anticancer compound paclitaxel (taxol) Figure 1.2, which was originally extracted from *Taxus brevifolia* [13, 14]. Taxol acts by inhibiting tubulin formation in cellular mitosis [15], preventing cell division. One reason why natural products are often an attractive starting points in drug discovery, is that, owing to their co-evolution with many protein species, they possess an optimized ligand-protein binding interaction [16].

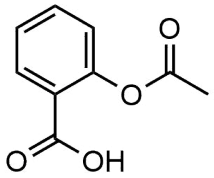


FIGURE 1.1: The chemical structure of the anti-inflammatory compound Aspirin.

By understanding and emulating the natural products structure, we can create novel, synthetic compounds that modulate the target protein's activity through binding. As natural products are formed by biogenesis, rather than chemical synthesis, their structures tend towards a high degree of complexity, making direct chemical synthesis difficult [17]. Taxol, for example, possesses a molecular weight of 854 Dalton and eleven stereo-centers. Despite being isolated in 1971, a total synthesis of taxol was not achieved until 1994 [18, 19].

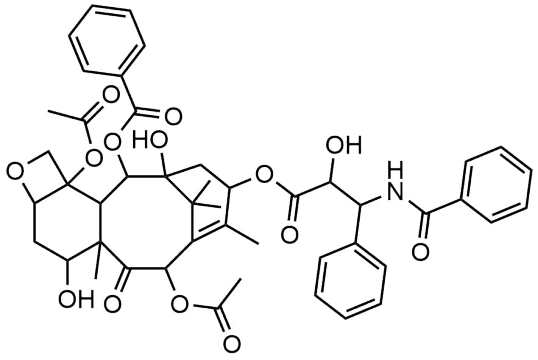


FIGURE 1.2: The chemical structure of the anitcancer compound taxol.

Unaltered natural products, such as taxol, make up an important class of drug compounds [20]; however, recently, medicinal chemists have also explored the use of natural product inspired analogs [21]. Due to the relative difficulty of synthesizing such natural products, typical drug discovery projects attempt to create analogs that retain key moieties, but with an overall reduction in complexity and an increase in synthetic tractability. Often these compounds are designed to improve upon the efficacy and toxicity of the natural product [22].

1.1.2 Screening Libraries

Another popular method in drug discovery is that of screening libraries [23]. While natural product inspired drug design focuses on select, often complex, bioactive ligands, screening libraries test a large array of diverse chemical species. In this approach, a library of compounds is carefully compiled and screened against the target protein(s) of interest. Any compounds showing activity towards the target are selected as drug "hits" and further optimized to a "lead" drug candidate structure. Compounds within these libraries are often selected due to their drug-likeness and absence of potential toxic or reactive moieties [24]. In phenotypic screening, rather than being screened against a target protein, compound screening is performed against target cells, and the phenotypical response is measured. In phenotypical screening, the

bioavailability of the compounds is of particular importance, as the molecule must not only bind but also be absorbed into the target cell [25]. The screening libraries may be formed from internal libraries, known drug compounds, or even commercially available sets of molecules. In a 2013 study [26], Sun *et al.* performed a drug repurposing screening experiment in which a library of 4096 known drug compounds was screened against the fungal strain *Exserohilum rostratum.* The anti-fungal agents posaconazole and lanoconazole were shown to also be potent inhibitors of *Exserohilum rostratum* growth.

1.1.3 Structure–Activity Relationship

Once an active compound has been established, iterative alterations are made to improve upon the properties of the drug candidate. One common way of accomplishing this is to make a series of closely related derivatives and measure their activity. By exploring the effect induced by the introduction of the various structural variations, one can establish what we call a structure-activity relationship (SAR), which gives us an indication of which structural modifications lead to increased activity towards the target. By following this procedure, we can make informed modifications to the structure in order to optimize activity. In a study by Wang *et al.* [27], the group performed a series of SAR studies based on the tubulin polymerization inhibitor colchicine. Starting from previous lead compounds, the group synthesized and tested a series of N-aryl-6-methoxy-1,2,3,4-tetrahydroquinoline derivatives. It was found that the inclusion of a quinazoline moiety lead to high cytoxicity, resulting in the production of a series of highly potent compounds (Figure 1.3).

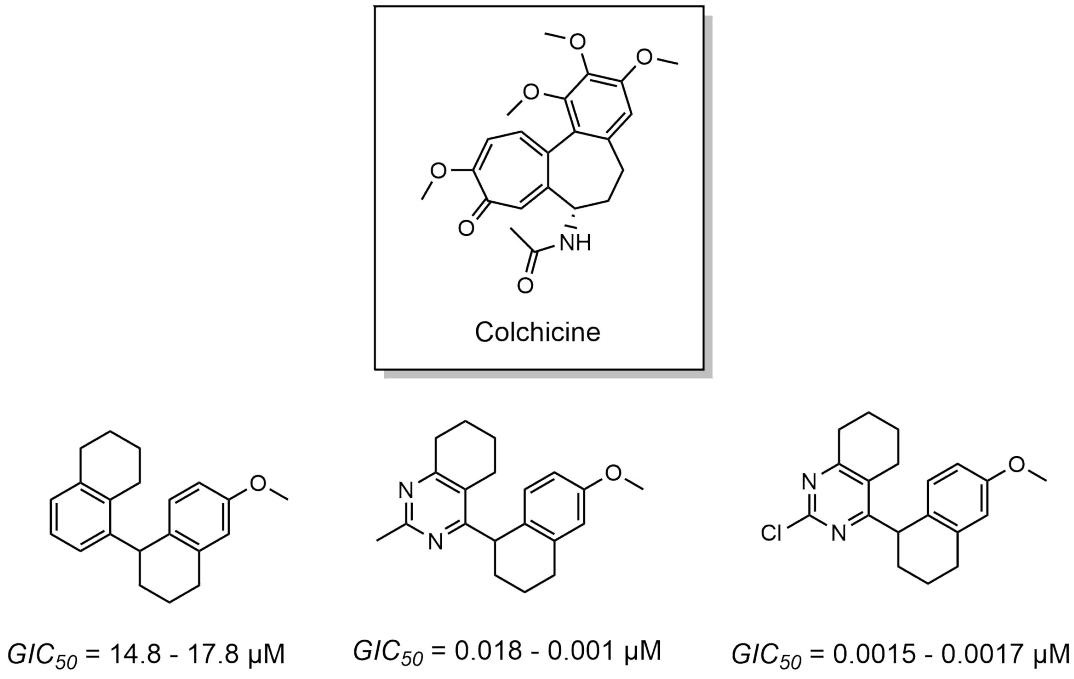


FIGURE 1.3: The natural ligand, Colchicine, is shown in the top box. Below are three select compounds from the study by Wang *et al.* [28]. In comparing the three structures, we see that minor structural modifications lead to a significant increase in efficacy.

Outside of isolated SAR studies, the information relating structural features to activity can then be used to construct a quantitative model of activity, known as a quantitative structure–activity relationship (QSAR). Through the use of QSAR models, compounds can be preferentially chosen based on their predicted activity, reducing the number of synthesized compounds required for lead optimization [29].

1.2 Computer Assisted Drug Design

Recently, scientists have begun exploring the potential of using computer-based methods to propose active structures [30]. These methods attempt to describe the problem algorithmically, with decisions relating to the chemical structure being based purely on quantitative metrics. Inspired from QSAR studies (see Section 1.1.3), many computer assisted drug design problems use a bioactive ligand(s) as a template in order to inform structure generation. One common method amongst drug design algorithms is that of *de novo* drug design, in which the molecules are generated entirely from simplistic base units. These units can be individual atoms, or can be composed of larger components, such as reactant molecules. One major advantage of computer-assisted methods is that computer programs can process an immense amount of information, incorporating structural [31, 32], biophysical [33] and even more abstract information such as dynamic shape [34] and energetic [35] profiles. Secondly, computer-assisted methods offer the potential to be scaled up in order to increase their overall efficiency. With advances in parallel and GPU computing, an accurate but computationally expensive computational method could be sped up in order to complete the *de novo* design in a shorter time frame [36, 37]. By generating and evaluating the molecules *in silico*, we reduce the resource requirements in order to obtain a reasonable drug candidate.

1.2.1 Descriptor Representation

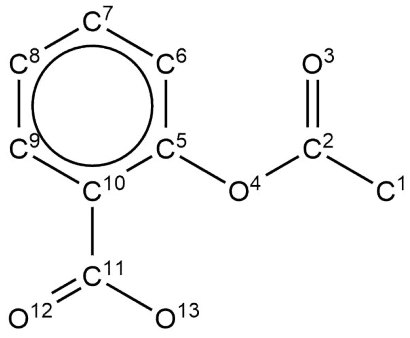
While most chemical structures as drawn only depict the connectivity of the molecular graph and the atomic number of each atom, a chemist can, through their expertise, infer a large amount of relevant, implicit information about the molecule, such as: the potential bonding interactions, electronic properties, conformational dynamics, tautomerism, protonation states, reactivity, etc. In order for a computer-assisted method to achieve a similar understanding, we first need a computer readable format with which to describe the molecules. A common format routinely used by cheminformaticians is that of molecular descriptors [38–40]. A molecular descriptor encodes chemical information about a molecule in a descriptive computer readable format, such as a vector, which can then be read by the computer algorithm. The elements of the descriptor can either describe purely structural features, such as the presence or absence of particular groups, or describe properties relating to the molecule, such as its electronic properties or conformational shape. A multitude of different methods exists for converting molecules into chemical descriptors, each one differing in what specific information is included explicitly. The choice of representation is often motivated by the nature of the problem one is working on. In the work by Wang *et al.* [41], the group used molecules' solvent accessible surfaces in order to construct a molecular descriptor for solubility prediction.

1.2.2 Molecular Similarity

The chemical similarity principle says that *compounds which share a similar structure are likely to show similar properties* [42]. This principle forms the central tenet of ligand-based scoring methods, in which we evaluate a compound according to its similarity to a given template lignad. Using information related to known bioactives, we aim to obtain a similar activity by emulating structural elements. This process is analogous to the methods seen in natural product design and SAR studies, where structures related to a known active are explored with the hopes of generating a novel compound with the desired properties. Applying this principle quantitatively presents us with the challenge of defining the concept of "similarity" rigorously. A typical solution to this is to use the elements of the molecular descriptor to define this similarity (see Section 1.2.1). Descriptors define the molecules as a vector in high-dimensional space. Metric functions commonly seen in mathematics, such as the euclidean distance, can be used to define the similarity of two molecules as the complement of the distance between their respective descriptor values [43]. This definition qualitatively matches what we would want in such a metric. If two molecules share all of the same features described by the chosen molecular descriptor, they will possess all the same elements, and hence the distance between them will be zero. We would say that the two molecules are identical under the given descriptor representation. Under a different descriptor representation the elements may differ, leading to a non-zero distance value. The greater the degree to which the descriptor elements differ, the larger the observed distance, corresponding to a higher degree of dissimilarity reflected in the metric [44, 45].

1.2.3 SMILES and SMARTS

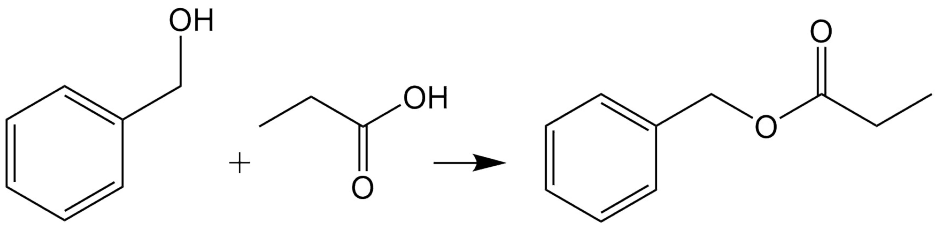
Another representation of molecules that is routinely used are those of the text-based approaches. In this form, molecules are represented as a series of characters within a text string. These characters can convey atomic information, such as the element type and charge state, as well as topological information, such as the bond order and arrangement. A very popular form of this is the Simplified Molecular-Input Line-Entry System (SMILES) string developed by Weininger [46] and implemented in the chemical computing software RDKit [47]. Figure 1.4 shows a depiction of a molecule and its corresponding SMILES strings. Recently, several research groups have exploited the similarities between of format with natural language to leverage algorithms developed for language translation to generate novel molecules [48].



CC(=O)Oc1ccccc1C(=O)O

FIGURE 1.4: Schematic representation of the SMILES text format adapted from Arús-Pous *et al.* [49]. The molecular graph of Aspirin is shown, with its corresponding SMILES representation underneath. Each atom in the graph is label according to its position in the SMILES string.

In a similar way to how molecular structures can be represented as SMILES, RDKit also provides a text based format for representing chemical reactions. This format is called SMILES arbitrary target specification (SMARTS). It makes use of an extended version of the SMILES notation. Molecules are represented as their corresponding SMILES, with individual chemical species being separated by the "." character, and reactants and products being separated by the "»" character. A depiction of this can be seen in Figure 1.5.



OC(C1=CC=CC=C1) . CCC(O)=O >> O=C(OCCC)C1=CC=CC=C1

FIGURE 1.5: Schematic representation of the SMARTS format. Shown is the reaction SMARTS text defining an esterification reaction between phenylmethanol and propionic acid. The two reactants are separated by the "." character, while the reactants and propyl benzoate product are separated by the "»" character.

Additionally, SMARTS provides the option to map specific atoms within the reactants to those within the product molecules. Importantly, the molecules within a reaction SMARTS encode the subgroup of a molecule. Any unspecified substituents effectively form an R-group, allowing us to represent a large number of explicit reactions with one generic reaction SMARTS. Further control is provided in order to control the environment of the individual R-groups.

1.3 Machine Learning

Recently, machine learning has received a huge amount of attention within the field of computer-science, with companies such as Google, Apple, and Microsoft investing heavily into machine learning research, following early innovation in academia [50–52]. The combined effect of this has been a dramatic increase in the number of machine learning methods available. In adapting these techniques, we have seen significant advancements made in a variety of problems ranging from image recognition [53, 54] to natural language processing [55]. One such example of this was the 2016 victory of the AlphaGo algorithm over world champion Go player Lee Sedol [56, 57]. This represented a substantial leap forward for the field, being several decades ahead of the commonly-predicted date, which was based solely on increases in computational power rather than algorithmic innovation.

1.3.1 Neural Networks

In contrast to other algorithmic methods in computer science, in which a computational model follows a set of predefined instructions in order to find a solution to the given problem [58, 59], in machine learning a model is trained to provide solutions based on observed trends in data. In order to train these models, we provide the machine learning method with a set of existing examples, and from these examples, the model learns inferred associations and constructs a statistical relationship, mapping evidence probabilistically to a given solution. One such model which is central to the field of machine learning is the artificial neural network (ANN) [60, 61]. ANNs combine a series of simplistic perceptron models in order to learn complex patterns from data [62]. The perceptron model was originally developed by Frank Rosenblatt in 1958 [63]. The model weights inputs by some real number, and through the use of a specific function called the activation function, converts the weighted input values into a binary output. By tuning this weight term, the perceptron model can learn to predict a sample point's binary label based on its value. The problem solving ability of this model was originally seen as analogous to that of biological neurons [64, 65], however, it was soon shown that there existed problems that the perceptron model could not solve [66]. ANNs were developed in order to resolve this issue. The ANN model consists of multiple individual perceptrons connected together to form a network. The perceptrons, or neurons, are organized in a series of layers, called hidden layers, with the output values of one layer being combined to form the inputs of the subsequent layer. A schematic depiction of this can be seen in Figure 1.6. By combining the collective outputs of each individual perceptron model, these ANNs are capable of approximating any real-valued function [67]. The weights of the network can be optimized so that the predictions of the ANN best match the desired outputs, and by carefully selecting the parameters of the network, we can improve the predictive capabilities of the model. The process of selecting these parameters is known as hyper-parameter optimization. Since their inception, many variants of the ANN architecture have been developed [68, 69]. The model presented here is often denoted as a "feed-forward" neural network.

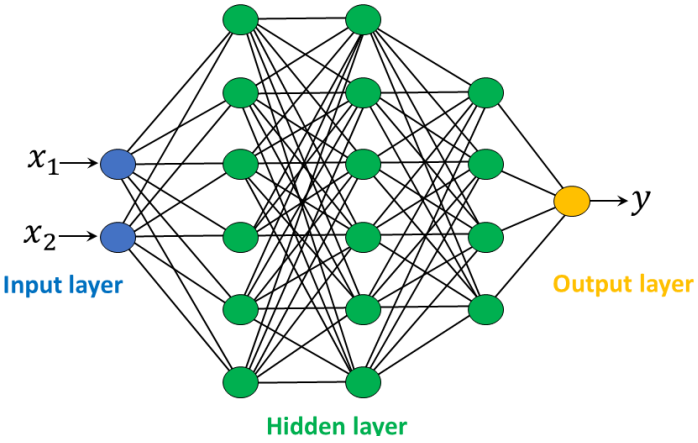


FIGURE 1.6: Schematic representation of an artificial neural network adapted from Bahi *et al.* [70]. The input values x_1, x_2, depicted on the left-hand side in blue are fed into the hidden layers (green). The connections represent the outputs of the various neurons, each weighted by an individual weight parameter. The combination of weighted outputs results in the prediction y, depicted in yellow on the right-hand side.

1.3.2 Training and Loss

As previously stated, in machine learning we train by presenting them with a series of examples. The model parameters are updated such that the predicted outputs best match their real-world equivalents. This process is called training. In order to quantify the degree to which a model's predictions adhere to the data, we define a loss function. This is a measure of the error in a given model's predictive capability. As an example, in categorical prediction, a given model is used to assign a particular sample to a category, such as assigning the label cat, dog, or bird to a given image. This prediction can be defined as a binary string of length three with each on-bit representing the assignment of the associated category (e.g. 100 for cats, 010 for dogs, and 001 for birds). A common loss function for a task such as this is the Hamming loss [71] (see Equation 1.1), which quantifies the disagreement between this predicted binary string and the true binary string of the training data.

$$
\begin{aligned}
H(y, \hat{y}) &= \frac{1}{N} \sum_{i=1}^{N} I(y_i, \hat{y}_i) \\
I(a, b) &= \begin{cases} 1 & \text{if } a \neq b \\ 0 & \text{if } a = b \end{cases}
\end{aligned}
\tag{1.1}
$$

EQUATION 1: Expression for the hamming loss between two binary strings. y represents the true binary string, while $\hat{y}$ represents the binary string predicted by the machine learning model. Each y_i corresponds to the bit value of the i^{th} element of the string y, and each $\hat{y}_i$ corresponds to the bit value of the i^{th} $\hat{y}$ string element. N represents the string length.

While one could use the Hamming loss for a problem such as this, many alternative loss functions also exist, such as binary cross-entropy (see Equation 1.2) [72, 73]. Our choice of loss function can have a profound influence of the model training [74].

$$H(y,p) = -\frac{1}{N}\sum_{i=1}^{N} y_i * \log(p_i) + (1-y_i) * \log(1-p_i) \tag{1.2}$$

EQUATION 2: Expression for the binary cross-entropy between two binary strings. p represents the string of probabilities associated with our predicted binary string, while y represents the true binary string. Each y_i corresponds to the bit value of the i^{th} string element, and each p_i to the model probability of assigning an on-bit (1) in this position.

One main problem seen in training a machine learning method is the problem of over-training. Here, the parameters have been set so that the model can perfectly predict each of the training examples, however, it has not learned the general trend; rather, it has simply learned the values of the training data. As we want our models to properly learn the associations between inputs and outputs, rather than effectively memorize the training data, this poses a serious problem. One common way to test for over-training is through a method called K-fold cross-validation [75]. In K-fold cross-validation, the training data is split into two partitions, one containing $\frac{1}{K}$ of the data, and another with the remaining $\frac{(K-1)}{K}$. The larger partition is used to train the model, with the smaller partition retained for testing the predictive accuracy (see Figure 1.7). This process is repeated K times, each with a different split of the data. Performing evaluations in this way allows us to evaluate how generalized the predictive accuracy of the model is (see Equation 1.3).

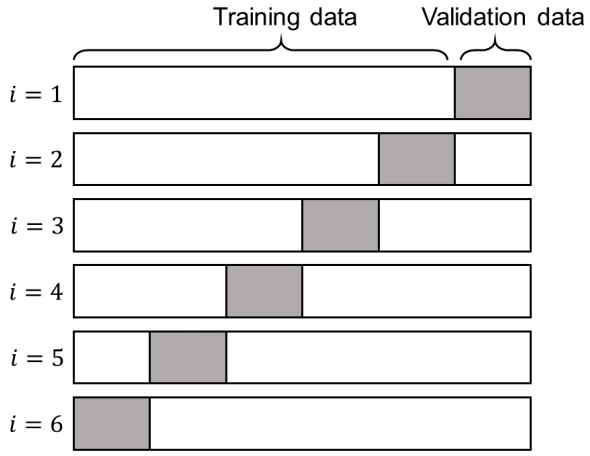


FIGURE 1.7: Representation of K-fold cross validation with 6-folds. The entire dataset is split into the training data (white), representing $5/6^{\text{th}}$ of the data, and the validation data (gray), representing the remaining $1/6^{\text{th}}$. The model is trained on the majority of the data, using the remaining validation data for evaluating the performance of the model. This procedure is repeated six times, each with a different section of the data being used for training and validation. The training and validation losses are averaged across all 6 of the folds to give the final cross-validated losses

$$CV_{train} = \frac{1}{K}\sum_{i=1}^{K} L(X_i^{train}, Y_i^{train})$$
$$CV_{val} = \frac{1}{K}\sum_{i=1}^{K} L(X_i^{val}, Y_i^{val}) \tag{1.3}$$

EQUATION 3: K-fold cross-validation. Using the loss function $L(x, y)$, the training and validation loss is evaluated and averaged across each of the K folds, thus giving the training and validation CV scores of the model.

1.3.3 Retrosynthetic Analysis

As previously mentioned in Section 1.3, one emblematic example of the power of machine learning was the victory of Google's AlphaGo algorithm against the current world champion [56, 57]. The model combines elements of machine learning with those of decision theory in order to create a powerful decision model capable of learning from past experiences. The core principles of the model are based on Markov decision theory [76]. Markov decision theory is a subset of probability theory that is concerned with the problem of making sequential decisions, each with an associated uncertainty, in order to produce a desired outcome. The problem is modeled as a series of transitions from various states. In the case of the game of Go the states are the different board configurations, and the actions are moves that can be made from these board states. In AlphaGo, rather than training a model that maps inputs to outputs, we seek to accurately model the overall outcome of selecting a particular action from a given state. That is, to accurately determine the conditional expectation value of the outcome from a given state s with given action a. Recently, methods such as these have been used to solve problems in chemistry such as in the field of predictive retrosynthetic analysis [77, 78]. In 2018, Segler *et al.* [79], inspired by the AlphaGo method, created a similar technique that makes used of tree searching algorithms in order to perform retrosynthesis. Here, the molecules represent the states, and the various possible retrosynthetic cleavages represent the actions one could take. The desired outcome of the program was to successfully reduce the structure into a set of known reagents, thus providing a full synthetic pathway to the target molecule. The procedure was governed by a machine learning model trained on a database of 12.4 million reactions, in order to bias the selection towards feasible retrosynthetic steps. Using their method, they were successfully able to solve 95% of the synthetic routes from a diverse set of 497 molecules.

1.3.4 Molecule Generation

One recent development in the field of *de novo* drug design has been the use of natural language models for molecule generation, with one prominent example of this being the long short-term memory network. This method produces a conditional probability distribution that predicts characters based on a provided input sequence. Starting with an arbitrary "START" character, the models predict individual characters sequentially, producing a sequence. This sequence is then used as the input to predict further characters thus, generating full strings of text. This method was applied to a set of molecular SMILES, and the models was used to sample the conditional probability distributions between SMILES characters, thus generating novel molecules. By selecting key examples for the model to focus on, we can bias this

distribution to only generate molecules that possess features that we are interested in. In the work by Merk *et al.* [80], a series of compounds active against PPARγ were used as templates in order to generate a set of novel actives. By applying the model, Merk *et al.* were able to generate a set of novel bioactive compounds, thus capturing the desired underlying property (bioactivity) while making a leap into less well-characterised chemical space (Figure 1.8).

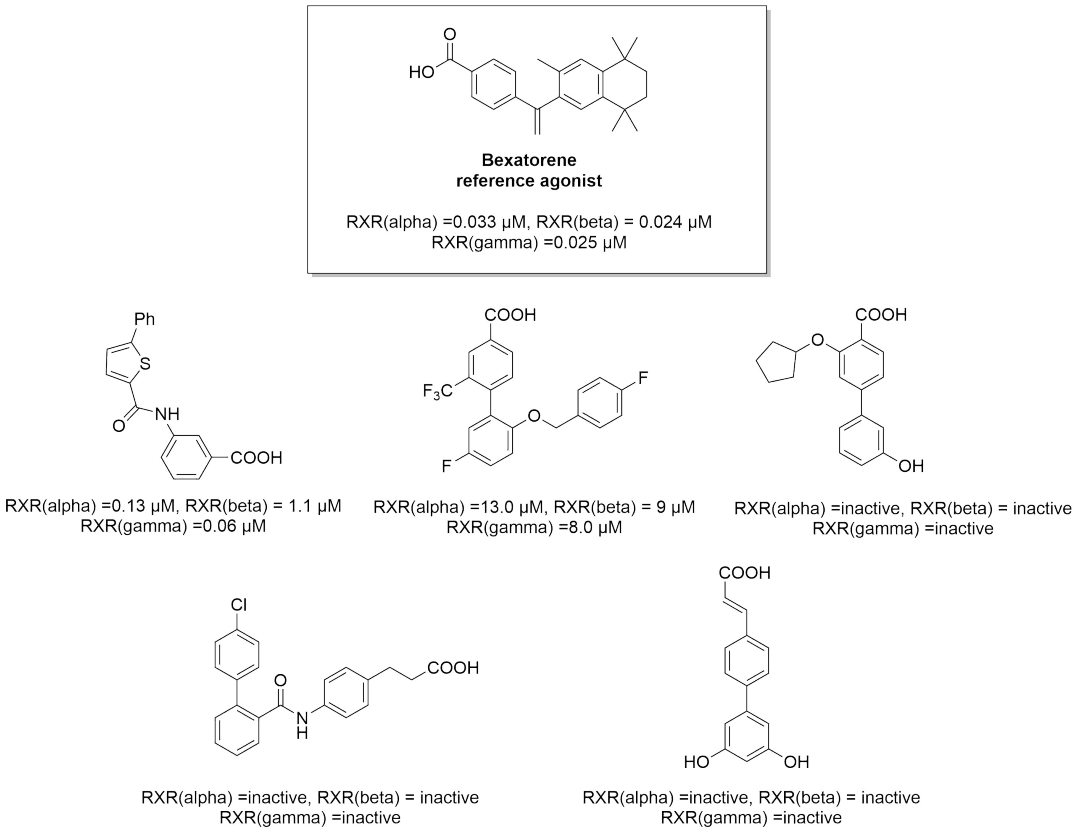


FIGURE 1.8: Five *de novo* designed molecules synthesized by Merk *et al,* [80]. Compounds were generated by a long short-term memory network (LSTM) which had been trained on a series of 541,555 bioactive compounds extract from ChEMBL [81], and then fine-tuned on a series of 25 compounds with known RXR and PPAR agonism.

1.4 Current Problems

1.4.1 Machine Learning in Computer-Assisted Drug Discovery

Despite the success of recent computer assisted methods using machine learning, a major limitation of these methods is the difficulty associated with producing the molecules proposed. In order to confirm the accuracy of the models experimentally, the novel, predicted compounds must be synthesized and tested. Not only does this impose a rate limiting step, taking many weeks or months to manufacture a suitably large sample size; but, due to the implicit nature of machine learning based methods, the feasibility of producing such species is not guaranteed. The compounds shown in Section 1.3.4 were selected based on their synthetic feasibility, with all synthetic decisions being made independent of the model. While this did lead to a set of bioactives, selection methods such as these have the potential to bias the results, making

a full evaluation of the predictions made by the model difficult. In conventional drug design projects, this does not pose a significant issue, as each step is governed not only by the underying structural hypothesis but also by the synthetic feasability of the considered compounds. In machine learning based drug design, however, the decisions are implicitly learned from the data. Therefore, chemical considerations are not a part of the decision making process, which leads to a high number of infeasible design structures. While recent advancements in simulated retrosynthesis and product generation offer promising solutions to this problem, this issue still remains one of the primary drawbacks of computer-based design. It would be beneficial to restrict models such as these to only produce outputs with a given degree of synthetic tractability. This would increase the throughput of such methods, and provide the opportunity to more easily evaluate, and hence improve, a model's performance on a given problem.

Chapter 2

Aims of this Thesis

The goal of this project was to develop a technique for generating *de novo* designed molecules, that also guarantees that all molecules produced adhered explicitly to a user stated definition of synthesizability. These definitions represent the expert knowledge governing the synthesizability of the designs produced. By combining the machine learning components for generation with the user provided rule-based logic for synthesizability, the aim is to synergistically combine the predictive capabilities of machine learning with the control and reliability of rule-based methods. Due to the diversity of drug design problems and the differing requirements of each specific drug project, it was of importance that the method developed be flexible with respect to the target problem. To this end, we place particular importance on the modularity of the *de novo* design method, as this would allow the technique to be adapted to each specific drug design project. We intend that our method forms a bridge between *in silico* predictions and the experimental consideration of a typical drug design project.

The aims of this thesis are organized thusly:

- Develop a computational method to generate novel *de novo* designs similar to a template ligand that incorporates expert knowledge on synthesizaiblity. Of particular importance is the modularity of the method, specifically the databases, predictive model, and synthesizability criteria.
- Experimentally explore the synthesizability and similarity to the template ligand of the structure proposed by the developed method in a laboratory environment.
- Extend the method to be fully integrated within an active learning drug design cycle, using automated synthesis (continuous flow) and surface plamson resonance for the synthesis and biotesting respectively.
- Implement a generalization of the initial method, allowing the algorithm to incorporate arbitrary molecular representations and non-greedy *de novo* design solutions.

Chapter 3

DINGOS - Design of Innovative NCEs Generated through Optimization Strategies

This section was adapted from the following publication:
[82] **Automated *de novo* molecular design by hybrid machine intelligence and rule-driven chemical synthesis**
Authors: *Alexander Button, Daniel Merk, Jan A. Hiss and Gisbert Schneider*
Journal : *Nature Machine Intelligence.*

The work within this PhD thesis resulted in the publication, Button *et al.* [82]. In this publication, we established and defined the DINGOS (Design of Innovative NCEs Generated through Optimization Strategies) algorithm, as well as presenting the first proof-of-concept case study of its operation. This chapter was directly adapted from this work, with some sections taken directly from the publication. These include the methods descriptions, some figures along with their corresponding figure captions, and the break down of the DINGOS algorithm. Alexander Button programmed the software and performed the computational experiments. Alexander Button, Dr. Jan Hiss and Prof. Dr. Gisbert Schneider designed the algorithm and analysed the data. Dr. Daniel Merk supervised the chemical part of the study and, together with Alexander Button, synthesized the compounds. Prof. Dr. Gisbert Schneider designed the study.

DINGOS was developed as a de novo design algorithm for generating synthetically feasible new chemical entities (NCEs) that are similar with regard to a predefined similarity index to a given template ligand of interest. To accomplish this, DINGOS combines more traditional rule-based methods with those of artificial intelligence (AI). DINGOS is a ligand-based scoring method and optimises the NCEs generated for their similarity to the provided template ligand. In this chapter, we present the DINGOS algorithm, and test its capabilities, both experimentally and in silico, in a proof-of-principle case study. DINGOS was used to generate a set of 300 novel de novo designed molecules for four separate template ligands. These included: alectinib, cariprazine, osimertinib, and pimavanserin. These four compounds represented a set of drug molecules that had been recently approved by the U.S. Food and Drug Administration (FDA). The de novo populations were ranked according to their similarity towards their respective template ligand. The similarity of the DINGOS designs were compared to DINGOS' compound database (used to construct the designs) and to the bioactive database ChEMBL. For each template ligand, it was shown that DINGOS successfully produced designs that were more similar to their respective templates than

the compound database; however, with the exception of the pimavanserin de novo population, it was found that the DINGOS designs were less similar than those from the ChEMBL database. One design was selected from each of the four DINGOS de novo populations for synthesis. Of these four compounds, three were successfully synthesised (the cariprazine, osimertinib, and pimavanserin de novo designs), and of those, one compound, the pimavanserin design, showed binding affinity towards the desired target. The relative activity of this compound was equivalent to 1 μM of serotonin against the 5-HT_{2B} serotonin receptor.

3.1 Introduction

3.1.1 The DINGOS algorithm

DINGOS was developed as a *de novo* design tool for producing new chemical entities (NCEs). The intention was that these NCEs would be both similar to known template ligands of interest and possess a high degree of synthetic feasibility. In recent years we've seen an explosion in the number *de novo* design tools, particularly ones utilizing methods from machine learning [48, 83, 84]. One main drawback for such methods is the need for the intervention of a chemist in order to select compounds and plan their synthesis. Not only is this an experimental bottleneck, but also may lead to a number of proposed structures being rejected due to issues of synthetic feasibility and internal bias. To deal with this problem, we developed DINGOS (Design of NCEs Generated through Optimization Strategies), a *de novo* design tool that forms NCEs by performing *in silico* virtual synthesis from an existing chemical database (Figure 3.1).

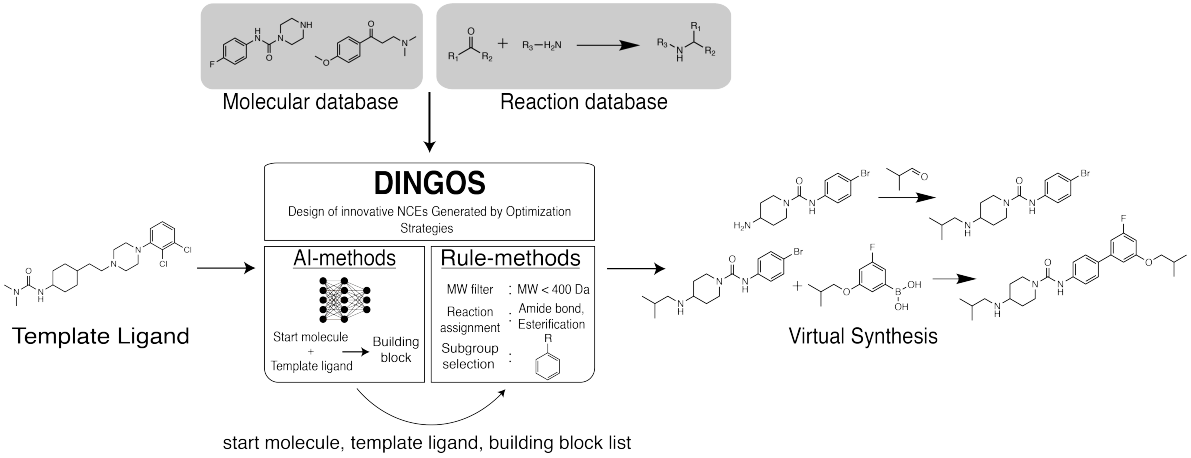


FIGURE 3.1: Overview of the DINGOS algorithm. A template ligand of interest, whose chemical and/or biological properties we would like to emulate, is provided along with a molecule and reaction database. The molecule database contains the structure of all potential reactant molecules. The reaction database contains the chemical logic required to perform the virtual syntheses. The DINGOS algorithm contains two main components, the AI component, which is used to recommend the reactant pairs, and the rule-based component, which is used to perform the virtual synthesis and make compound evaluations. From the provided template ligand, DINGOS generates a set of *de novo* designed molecules along with their proposed synthetic pathways. Reprinted with permission from Button *et al.* [82]

DINGOS combines AI methods with rule-based methods in order to generate compounds. The AI methods are used to recommend meaningful reactant pairs for chemical synthesis, while the rule-based methods are used to perform the virtual synthesis. The incorporation of these rule-based methodologies allows for the desired chemical logic to be explicitly encoded within the algorithm. In this way, DINGOS forces the designs to adhere to the user defined definition of synthesizability. Additionally, *de novo* designs are constructed from a database of known building block molecules, ensuring that all structures presented are chemically valid and, if desired, commercially available. The advantage of performing compound generation in this way is that the generated structures are accompanied with a synthetic pathway composed of real building block molecules from the compound database. DINGOS is a ligand-based scoring [85] algorithm, meaning that compounds are generated in order to maximize their similarity to a known template ligand of interest. In order to quantify this similarity, molecules are represented as molecular descriptors [86], which are vector representations of the overall chemical structure, and the molecular descriptors are compared using an appropriately chosen metric function [87]. This metric function calculates the distance between the molecules in descriptor space, in which the distance represents the chemical similarity of the two chemical species. By carefully choosing our descriptor and metric functions, we are able to bias the *de novo* design towards desired chemical and biochemical properties of interest. The procedure of combining molecules in order to optimize for the chemical similarity towards a template ligand is applied to individual single-step reactions (Figure 3.2). Two molecules are reacted together forming a new product molecule. This molecule then becomes the new starting molecule and the procedure is repeated until either a

set of user defined stop criteria is met or we cease to see further improvements in the molecules distance score (local convergence).

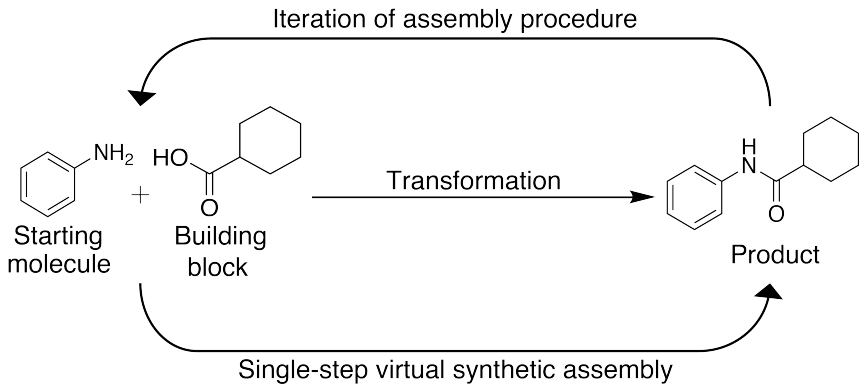


FIGURE 3.2: A schematic representation of the single-step iterative assembly method used by DINGOS in order to perform its *de novo* design. Reprinted with permission from Button et al. [82]

The DINGOS algorithm is defined in the following four main steps and illustrated in Figure 3.3. This description, along with the corresponding flow diagram and the corresponding figure caption, are taken directly from the publication Button *et al.* [82]:

"**Step 1: Generation of the molecular building block library {mol}.**
The compound database can be any set of molecules, with the only requirement being that the molecular entries have a valid SMILES [88] representation. *De novo* drug design by fragment growing involves the assembly of complex molecules from smaller more simplistic components. For this rationale to be reflected within DINGOS, the molecular weight of the building blocks should be considerably smaller than that of the template drug. To ensure this, a molecular weight range is specified and molecules outside of this mass range are not considered during the assembly procedure. Additional filtering criteria based on molecular subgroups or properties can be applied.

Step 2: generation of the set of starting molecules {S}.
The molecular descriptor of each member of the molecular building block library {mol} is calculated, and the distance between those building blocks and the template descriptor is evaluated. Building blocks are then sorted according to their increasing distance to the template. A subset of the M closest molecules {S} is then selected from this sorted set. Each element of {S} is used as the starting molecule for an individual assembly procedure. The selection of several unique starting points for each NCE encourages a high degree of structural diversity within the producwas chosnt set and is meant to promote designs with scaffolds that differ structurally from that of the template ligand for the purpose of chemical scaffold-hopping [89–91].

Step 3: construction of optimal intermediates and products P_{opt}.
The ith product molecule P_i is formed from the ith element S_i of the start mol set {S}. Thereby, S_i and the template T serve as inputs for the machine learning model M, which takes the descriptor values of S_i and T and predicts a descriptor value. This

predicted descriptor corresponds to the building block fingerprint B* representing the ideal building block for transforming S_i to T, which has been learned by the model M during training. A distance calculation between B* and {mol} is performed, and a subset {B} of the N most similar molecules is produced. All valid chemical transformations between S_i and {B} are applied, generating a set of intermediate products {P_1,..,P_k } of size K. The element most similar to T is chosen as the optimal intermediate product P_{opt}. If none of the termination criteria is met (step 4), then P_{opt} is selected as the starting molecule for the next growing step ($S_i = P_{opt}$).

Step 4: termination.

The growing of S_i is continued until at least one of the stop criteria is met. There are three conditions under which the construction is halted: (1) the molecular weight of the product exceeds the molecular weight limit, (2) the number of applied reaction steps exceeds that of the reaction step limit and (3) the distance of P_{opt} to the template T is greater than that of the starting molecule S_i. On halting the construction process, the current optimal product P_{opt} is saved as the i^{th} final product ($P_{final} \equiv P_{opt}$) and P_{final} is added to the output product set {P}. In the event of criterion (3) being met, the starting molecule of the current step is saved as the final product ($P_{final} \equiv S_i$) instead of P_{opt} . The current P_{opt} is not considered for any further assembly steps, as it has been shown to be less similar to the template ligand than the starting molecule. Step 3 is then repeated for the next element of {S}, S_{i+1} ."

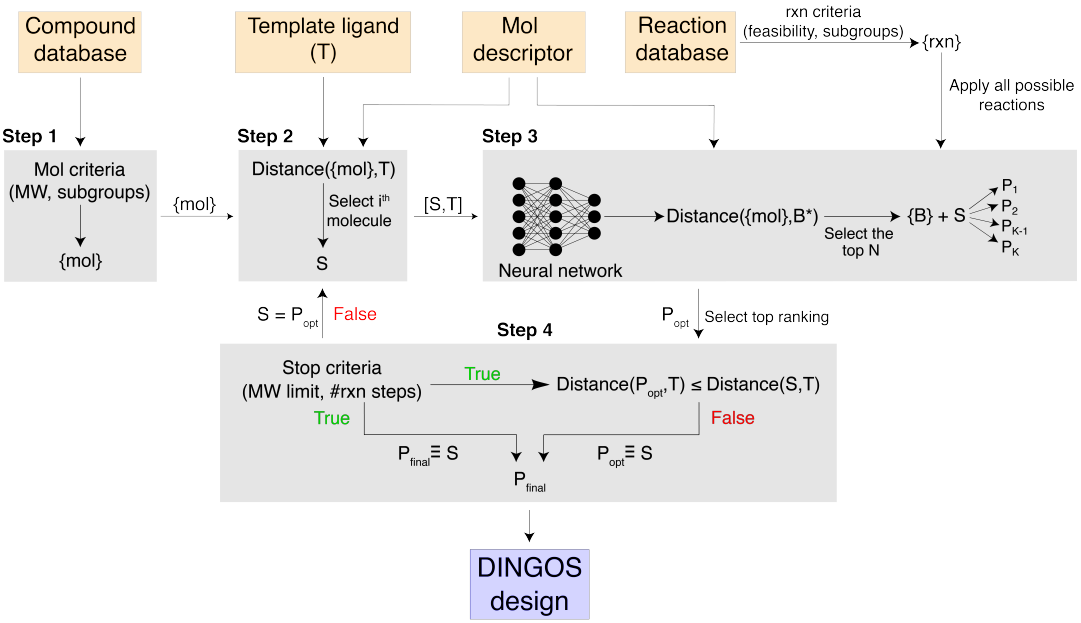


FIGURE 3.3: Orange boxes, inputs; blue box, output. Step 1: the compound database is filtered according to a set of predefined criteria to obtain the set of building blocks {mol}. Step 2: this input molecule set {mol} is sorted according to its distance to the template molecule T. The most similar molecule S is selected as the starting molecule. The starting tuple [S,T] serves as the input for the trained machine learning model (here, a feedforward neural network [92]). Step 3: the network predicts the descriptor value of the building block B*, and the N most similar molecules form the building block set {mol}. The set of possible reactions {rxn} between {B} and S is selected from the reaction database and all reactions are performed, thus generating a set of intermediate products {P}. From the intermediate product set, the top-ranking molecule P_{opt} is selected. Step 4: if P_{opt} is more dissimilar to the template T than the starting molecule S, then the starting molecule is selected as the final product. If P_{opt} is more similar to T and none of the stop criteria are met, then P_{opt} is selected as the new start molecule for a further iteration of the DINGOS molecule growing algorithm. Otherwise, if the stop criteria are met, then P_{opt} is selected as the final product P final. Reprinted with permission from Button *et al.* [82]

DINGOS was intended to be directly integrated within existing drug design projects. Owing to the different synthetic and design requirements of each individual drug design project, it is unlikely that one overarching *de novo* design strategy would be universally optimal. Hence, the DINGOS algorithm was design to be modular in nature, allowing for the individual components to be easily replaced while maintaining the overall functionality of the method. In this chapter, we present our first proof-of-concept case study of the DINGOS method, highlighting its performance and function on a simple test problem. For this purpose, we chose one consistent set of parameters. The modular aspects of the algorithm, however, are further explored in Chapters 5 and 6.

3.1.2 Rule-Based Method

A key goal in developing DINGOS was to produce a *de novo* design method that generates synthetically feasible structures. While in recent years many achievements have been made in the field of reaction feasibility [93, 94], this problem is far from resolved. In order to practically tackle this problem, we established a set of 64 *in silico* reactions (shown in Appendix A.2.1), that incorporated all of the desired chemical logic for synthesis. These *in silico* chemical reactions were used as our 'rules' for combining building block molecules and for forming the corresponding product molecule. In the case where multiple reactions, and hence products, were possible, all potential products were produced from a single building block pair. This set was written in the SMARTS notation, which is a text-based format for encoding chemical reactions within the RDKit software suite. While this set is by no means comprehensive, or infallible, it represents the total range of chemical logic available to the DINGOS algorithm. A key advantage of the modular nature of DINGOS is that these reaction can easily be replaced, allowing DINGOS to perfectly match the synthetic capabilities of the user. Furthermore, improved knowledge gained from both experiment and expertise can be used to update the reaction set, allowing the DINGOS to be continually improved without the need for reimplementation of the underlying algorithm.

3.1.3 Machine Learning Model

A crucial part of step 3 in the DINGOS algorithm is the recommendation of appropriate building block molecules for the *de novo* assembly. A machine learning model was developed in order to performed this recommendation. We chose to make use of the relatively simple multi-layered perceptron model (MLP) for the building block recommendation (see Figure 3.4). The 167 public molecular access keys (MACCS keys) [95] were chosen as the molecular descriptor for this study. The MACCS keys are a structural, binary fingerprint, in which each on-bit (1) represents the presence of a particular substructure. By combing the various on- and off-bits we gain a picture of the overall molecular structure. In comparison to other molecular descriptors, such the extended connectivity fingerprint, the MACCS keys are a more coarse-grained representation. The choice of the MACCS keys were motivated by two features of this representation. Firstly, the MACCS keys represent a relatively low dimensional representation, encompassing only 167 bits. By comparison, the more common-place Morgan fingerprint commonly uses 2048 bits in its encoding. The use of a low-dimensional representation allows for more effective training within the machine learning model, and reduces the influence of the 'curse of dimensionality' (The increased data requirement needed to fully represent a high dimensional space) [96, 97]. Secondly, since the MACCS keys encode substructural elements explicitly, this means that reaction relevant subgroups, such as the alcohols for esterification, are directly present within the representation. For molecular descriptors incorporating non-structural features, such as 3D shape, these requirements would have to be learned implicitly from the data, which could potentially inhibit the learning of the algorithm.

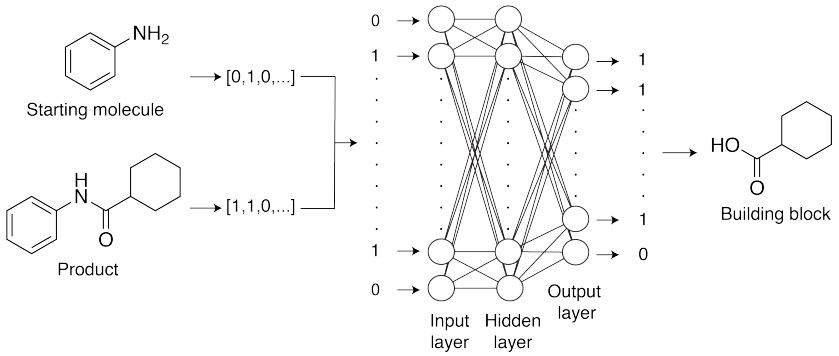


FIGURE 3.4: Schematic representation of the multi-layered perceptron model used in the DINGOS algorithm case-study. The start and product molecule (left-hand side) are converted into the MACCS keys descriptor. Their descriptor value serves as the input for the network (middle). The descriptor value of the building block molecule is predicted (right-hand side). Reprinted with permission from Button *et al.* [82]

For training data, we selected the US Patent and TradeMark Office database (USPTO). Each entry consisted of a single chemical reaction written as a SMARTS string. Reaction type and yield was not stated in the dataset. For training, the structures within the chemical reactions were converted to their descriptor representation. To form the training data, each product molecule was paired with one of the reactants within the reaction entry, this reactant molecule served as the starting molecule for the reaction. The remaining reactant, in the case of two-component reactions, was set as the building block. In the case of one-component reactions, a null-fingerprint (fingerprint of all off-bits) was chosen as the building block's descriptor value. In the case of two-component reactions, two pairs were formed with the product, one for each of the two reactant molecules. The model was trained and gave an average loss of 0.0988 $\pm$ 0.0002 (mean $\pm$ standard deviation) for the training set and 0.1029 $\pm$ 0.0006 for the validation set (Figure 3.5).

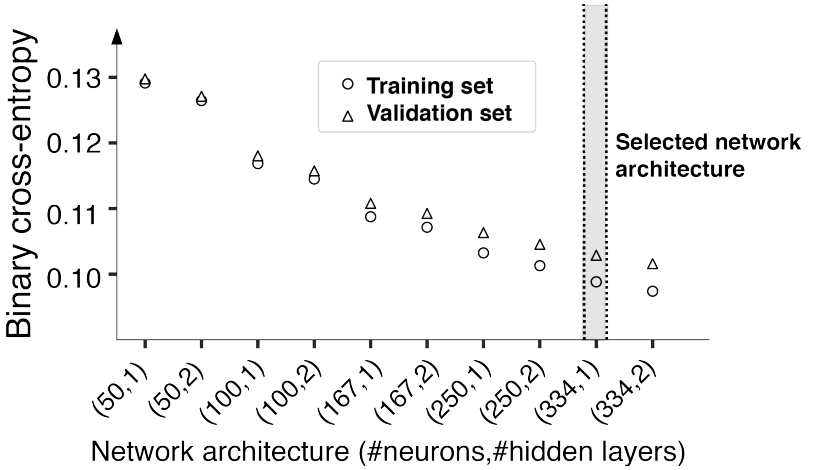


FIGURE 3.5: Plot showing the results of the hyperparameter optimisation of the multi-layered perceptron model. An architecture of one hidden layer with 334 hidden neurons was selected. Reprinted with permission from Button *et al.* [82]

3.2 Methods

All methods used in this section were taken directly from the publication Button *et al.* [82].

3.2.1 RDKIT Software

All in silico chemical manipulations were performed using the open-source cheminformatics software RdKit [47](version 2018.09.1). For the initial DINGOS study version 2017.03.3 was used. In the revised version of the DINGOS code (Chapter 5 and 6) version 2018.09.1 was used. Reactions were performed using the RunReactants() function. Conversion of the molecules into a canonical SMILES format was achieved with MolToSmiles(). The MACCS keys molecular fingerprints were generated with the RDKit function MACCSkeys().

3.2.2 Template Molecule Processing

For the DINGOS study, the template molecules used for *de novo* design were washed using the KNIME implementation of the MOE 'wash' method (Molecular Operating Environment, version 2011.10, The Chemical Computing Group) [98].

3.2.3 SPiDER

All target predictions were performed using the software SPiDER (self-organizing map–based prediction of drug equivalence relationship) [99]. SPiDER utilizes self-organising maps to evaluate the pseudo-probability of activity of a given molecule against a pool of 251 predefined targets. Predictions are reported as p-values representing the pseudo-probability of misclassification. Compounds with p-values less than 0.1 were considered as predicted active.

3.2.4 Compound Database

A set of commercially available chemical compounds were identified through Reaxys (www.reaxys.com, version 2018.3.14) [100]. All molecular structures were converted into a standardized canonical SMILES format using RDKit's molecule converter. Salts and minor components were removed as well as all incomplete and inaccurate structures. This yielded a dataset of 245,296 molecules, which were used as the compound database (construction set) for DINGOS.

3.2.5 Run Parameters

During the *de novo* assembly, the building block set was restricted to molecular weights less than 400 g mol^{-1}. A product limit of 300 compounds and a product molecular weight limit of 600 g mol^{-1} was set, this represented an upper limit of the templates' molecular weights. The number of reaction steps was set to a maximum of four, and the number of building blocks considered at each assembly step was 20. The MACCS keys descriptor was used to represent the molecules, and the distance between molecules was calculated as the Hamming distance, which is the complement of the Hamming loss (see Equation 1.1). All parameters were kept consistent across each run, and all calculations were performed on a single CPU within one hour.

3.2.6 Training Data

To obtain training data for our neural network model, 1.8 million entries were extracted from the US Patent and TradeMark Office (USPTO) database [101]. This dataset contained cases that fell outside the bounds of our considered problem set (peptides, large molecules and so on). To remove these cases, the product molecules were filtered based on molecular weight. An upper molecular weight limit of 400 g mol^{-1} was enforced on each of the starting reactants. This ensured that all products were formed from a combination of small molecular building blocks. The same sanitation procedure used for the compound database was applied to remove salts, minor components and erroneous cases. Reactions were filtered by number of reactants; an upper limit of two reactants per reaction was imposed. To extend the data set for training, examples were generated in which reactant positions were exchanged. This yielded a dataset of 897,286 examples.

3.2.7 Multi-layered Perceptron Model

For the training of the machine learning model, the binary cross-entropy (see Equation 1.2) [102] of the binary fingerprints was used as the loss function, with the Adam optimizer being used in order to perform the back-propagation [103]. The input layer of the network consisted of 334 neurons, twice the size of the MACCS keys. This represented a concatenation of the starting and product molecules' descriptor values. The output layer consisted of 167 neurons, each corresponding to the bits of the predict building block descriptor. The network was trained for 50 epochs with a batch size of 256 and a learning rate of 0.001. For the activation the sigmoid activation function was chosen. In order to select the architecture of the model, hyperparameter optimization was performed with 10-fold cross validation (Figure 3.5). From the considered architectures, we selected a single hidden-layer consisting of 334 neurons.

3.2.8 ChEMBL Dataset

The molecule sets used for the distance analysis (Section 3.3.4) were prepared with the same procedure used for the compound database. Entries from the ChEMBL [104] dataset that did not have valid activity data were omitted. A molecular weight limit of 1,000 g mol^{-1} was enforced to ensure that only small molecule drug structures were considered. For the physico-chemical analysis, four separate sets of compounds were extracted from ChEMBL, each sharing the biological targets of the four template compounds. Only compounds with inhibition constants (K_i values) less than 10 nM were considered.

3.2.9 Synthesis of compound 5 - alectinib *de novo* design

3-(((4-(2,2-dimethylmorpholino)-2-ethoxyphenyl) amino)methyl)-1H-indole-5-carbonitrile (Compound 5)

4-(2,2-dimethylmorpholino)-2ethoxyaniline (Compound **5**, 125 mg, 0.50 mmol, 1.00 equiv.) and 3-formyl-1H-indole-5carbonitrile (10, 85 mg, 0.50 mmol, 1.00 equiv.) were dissolved in dichloroethane (5 mL), a 4 Å molecular sieve and acetic acid (0.25 mL) were added and the mixture was stirred at room temperature for 60 min. Sodium triacetoxyborohydride (210 mg, 1.00 mmol, 2.00 equiv.) was then added and the mixture was stirred at room temperature for another 4 h. The mixture was filtered, water (25 mL) was added, the phases were separated, and the aqueous layer was extracted three times with ethyl acetate (3 x 25 mL). The combined organic layers were dried over magnesium sulfate and the solvents were evaporated in vacuum. The crude product was purified by column chromatography using methylene chloride/ methanol 98:2 as the mobile phase to obtain the title compound as yellow oil; MS(ESI+) m/z 405.4 ([M + H]+). Compound **5** was sensitive to water and especially light, and was not stable enough for *in vitro* characterization.

3.2.10 Synthesis of compound 6 - cariprazine *de novo* design

N-(4-chlorophenyl)-4-(2-isopropylbenzyl)piperazine-1-carboxamide (Compound 6)

N-(4-chlorophenyl)-piperazine-1-carboxamide hydrochloride (Compound **6**, 138 mg, 0.50 mmol, 1.00 equiv.) and 2-isopropylbenzaldehyde (12, 96 mg, 0.65 mmol, 1.30 equiv.) were dissolved in dichloroethane, 4 Å molecular sieve was added and the mixture was stirred at room temperature for 30 min. Sodium triacetoxyborohydride (211 mg, 1.00 mmol, 2.00 equiv.) was slowly added and the mixture was stirred at 50 °C for 48 h. The reaction mixture was then filtered, added to saturated sodium carbonate solution (25 mL), the phases were separated, and the aqueous layer was extracted with ethyl acetate (3 x 25 mL). The combined organic layers were dried over magnesium sulfate and the solvents were evaporated in vacuum. The crude product was purified by column chromatography using methylene chloride/methanol (98:2) as the mobile phase to obtain the title compound as a colourless solid (119 mg, 64%). 1H NMR (400 MHz, chloroform-d) δ = 1.16 (d, J = 6.9, 6H), 2.37–2.44 (m, 4H), 3.28 (hept., J = 6.9, 1H), 3.35–3.40 (m, 4H), 3.47 (s, 2H), 7.06 (td, J = 7.2, 1.6, 1H), 7.13–7.18 (m, 3H), 7.20–7.28 (m, 4H) ppm. 13C NMR (101 MHz, chloroform-d) δ = 24.06, 28.51, 44.29, 52.62, 60.66, 121.08, 125.67, 127.85, 128.04, 128.86, 130.42, 134.21, 137.61, 148.60, 154.68 ppm. MS(ESI+) m/z 372.3 ([M + H]+). HRMS(ESI+) m/z calculated 372.1837 for $C_{21}H_{27}ClN_3O$, found 372.1841 ([M + H]+). HPLC, retention time: 2.740 min.

3.2.11 Synthesis of compound 7 - osimertinib *de novo* design

N-(2-((2-(dimethylamino)ethyl)amino)benzyl)-2-(1-(5-methoxybenzo[d] oxazol-2-yl)piperidin-3-yl)acetamide (Compound 7)

2-(1-(5-methoxybenzo[d]oxazol-2-yl) piperidin-3-yl)acetic acid (Compound **7**, 77 mg, 0.25 mmol, 1.00 equiv.), N1-(2-(aminomethyl) phenyl)-N2,N2-dimethylethane-1,2-diamine (14, 58 mg, 0.30 mmol, 1.20 equiv.) and 4-DMAP (31 mg, 0.25 mmol, 1.00 equiv.) were dissolved in chloroform (abs., 10.0 mL) and EDC (47 mg, 53 μl, 0.30 mmol, 1.20 equiv.) was slowly added. The mixture was stirred under reflux for 2 h. After cooling to room temperature, 15 mL saturated sodium carbonate solution was added, phases were separated, and the aqueous layer was extracted with ethyl acetate (2 x 15 mL). The combined organic layers were dried over magnesium sulfate and the solvents were evaporated in vacuum. The crude product was purified by column chromatography using methylene chloride/methanol (95:5) and acetone/triethylamine (98:2) as mobile phases to obtain the title compound as reddish oil (59 mg, 51%). 1H NMR (400 MHz, chloroform-d): δ = 1.34 (td, J = 8.8, 4.2, 1H), 1.50–1.69 (m, 2H), 1.81–1.88 (m, 1H), 2.01 (dd, J = 13.1, 5.5, 1H), 2.11 (d, J = 0.5, 1H), 2.11–2.18 (m, 1H), 2.22 (dd, J = 13.1, 7.8, 1H), 2.41 (s, 6H), 2.76 (t, J = 6.3, 2H), 3.15 (dd, J = 13.2, 7.8, 1H), 3.27 (t, J = 6.4, 2H), 3.28–3.35 (m, 1H), 3.71 (s, 3H), 3.74–3.81 (m, 1H), 3.85 (dd, J = 13.2, 3.6, 1H), 4.28 (dd, J = 14.6, 5.5, 1H), 4.38 (dd, J = 14.6, 6.2, 1H), 6.47 (dd, J = 8.7, 2.6, 1H), 6.56 (dd, J = 8.1, 1.2, 1H), 6.62 (td, J = 7.4, 1.1, 1H), 6.69 (d, J = 2.5, 1H), 7.00 (d, J = 8.7, 1H), 7.09–7.17 (m, 2H) ppm. 13 C NMR (101 MHz, chloroform-d): δ = 23.56, 29.28, 30.42, 30.93, 32.84, 39.87, 40.60, 41.04, 44.81, 46.35, 50.44, 53.80, 55.96, 57.62, 101.11, 107.00, 108.54, 110.43, 116.92, 129.36, 130.80, 143.14, 143.97, 146.05, 157.01, 163.14 ppm. MS(ESI+) m/z 466.1 ([M + H]+). HRMS(ESI+) m/z calculated 466.2813 for $C_{26}H_{36}N_5O_3$, found 466.2811 ([M + H]+). HPLC, retention time: 15.650 min.

3.2.12 Synthesis of compound 8 - pimavanserin *de novo* design

N-(4-bromophenyl)-4-(isobutylamino)piperidine-1-carboxamide (Compound 16)

4-amino-N-(4-bromophenyl)-piperidine-1-carboxamide hydrochloride (Compound **16**, 100 mg, 0.33 mmol, 1.00 equiv.) and isobutyric aldehyde (40 μl, 32 mg, 0.43 mmol, 1.30 equiv.) were dissolved in dichloroethane (5.0 mL), acetic acid (0.50 mL) and 4 Å molecular sieve were added and the mixture was stirred at room temperature for 30 min. Sodium triacetoxyborohydride (95 mg, 0.43 mmol, 1.30 equiv.) was slowly added and the mixture was stirred at room temperature for 16 h. The reaction mixture was filtered, added to saturated sodium carbonate solution (25 mL), the phases were separated and the aqueous layer was extracted with ethyl acetate (3 x 25 mL). The combined organic layers were dried over magnesium sulfate and the solvents were evaporated in vacuum. The crude product was purified by column chromatography using methylene chloride/methanol (95:5) as the mobile phase to obtain the title compound as a colourless solid (89 mg, 76%). 1H NMR (400 MHz, DMSO-d6) δ = 0.97 (d, J = 6.7, 6H), 1.44–1.59 (m, 2H), 1.90–2.10 (m, 3H), 2.73–2.86 (m, 4H), 4.20 (d, J = 13.6, 2H), 7.38–7.43 (m, 2H), 7.43–7.49 (m, 2H), 8.55 (s, 2H), 8.78 (s, 1H) ppm. 13C NMR (101 MHz, DMSO-d6) δ = 20.59, 26.07, 28.33, 42.83, 51.44, 55.42, 121.82, 131.53, 140.45, 144.62, 155.03 ppm. MS(ESI+) m/z 354.2, 356.2 ([M + H]+).

N-(3'-fluoro-5'-isobutoxy-[1,1'-biphenyl]-4-yl)-4-(isobutylamino) piperidine-1-carboxamide (Compound 8)
16 (65 mg, 0.18 mmol, 1.00 equiv.), 3-fluoro-5isobutyloxyphenylboronic acid (Compound **8**, 78 mg, 0.37 mmol, 2.00 equiv.) and caesium carbonate (180 mg, 0.55 mmol, 3.00 equiv.) were dissolved in a mixture of dioxane (9.0 mL) and DMF (1.0 mL) and the mixture was stirred for 30 min at room temperature. Tetrakis(triphenylphosphine)-palladium(0) (42 mg, 0.04 mmol, 0.20 equiv.) was then added and the mixture was stirred for 12 h under reflux. After cooling to room temperature, the reaction mixture was filtered, water (25 mL) was added and the mixture was extracted with ethyl acetate (3 x 25 mL). The combined organic layers were dried over magnesium sulfate and the solvents were evaporated in vacuum. The crude product was purified by column chromatography using methylene chloride/methanol (9:1) as the mobile phase and recrystallized from chloroform/hexane to obtain the title compound as a colourless solid (26 mg, 33%). 1 H NMR (400 MHz, methanol-d4) δ = 0.95 (d, J = 5.0, 6H), 0.97 (d, J = 5.0, 6H), 1.52 (qd, J = 12.5, 4.4, 2H), 1.88–2.02 (m, 3H), 2.06–2.12 (m, 2H), 2.83 (d, J = 7.2, 2H), 2.85–2.92 (m, 1H), 3.25 (s, 1H), 3.70 (d, J = 6.4, 2H), 4.25 (dt, J = 13.8, 2.6, 2H), 4.48 (s, 1H), 6.53 (dt, J = 10.8, 2.3, 1H), 6.79 (ddd, J = 9.9, 2.3, 1.5, 1H), 6.84 (t, J = 1.9, 1H), 7.33–7.38 (m, 2H), 7.42–7.46 (m, 2H) ppm. 13C NMR (101 MHz, methanol-d4) δ = 12.38, 17.13, 18.11, 18.76, 18.93, 26.17, 27.96, 28.18, 42.42, 74.45, 87.72, 98.97, 104.83, 116.50, 116.97, 120.70, 126.77, 155.41, 168.24, 171.31 ppm. MS(ESI+) m/z 442.4 ([M + H]+). HRMS(ESI+) m/z calculated 442.2864 for $C_{26}H_{37}FN_3O_2$, found 442.2865 ([M + H]+). HPLC, rentention time: 18.367 min.

3.2.13 *In vitro* testing of compound 6 against the D_{2L}, D_{2S}, and D_{2L} Dopamine receptors

Compound **6** was studied on dopamine receptors D_{2L}, D_{2S} and D_3. D_{2L} activation and antagonism were studied in a functional assay using membranes containing human recombinant D_{2L} receptors (expressed in Chinese hamster ovary (CHO) cells) wherein binding of radiolabelled [35S]GTPγS was determined. Results represent relative activity compared to 1 mM dopamine. Activity on D_{2S} was assessed in a cell-based (HEK293 (human embryonic kidney) cells) impedance assay and a cellular (CHO cells) homogeneous time resolved fluorescence (HTRF) assay with cyclic adenosine monophosphate (cAMP) readout served for D_3 testing. Biological assays were performed by Eurofins (www.eurofins. com) on a fee-for-service basis.

3.2.14 *In vitro* testing of compound 7 against the EGFR receptor

The inhibitory potency of compound **7** on EGFR was studied on recombinant enzyme (expressed in insect cells) with poly-Glu-Tyr as substrate in the presence of radiolabelled [γ32P]ATP (adenosine triphosphate). Substrate phosphorylation was quantified by scintillation measurements. Biological assays were performed by Eurofins (www.eurofins. com) on a fee-for-service basis.

3.2.15 *In vitro* testing of compound 8 against the 5-HT_{2A}, 5-HT_{2B}, and 5-HT_{2C} Serotonin Receptors

The activity of compound **8** on serotonin receptors 5-HT_{2A}, 5-HT_{2B}, and 5-HT_{2C} was determined in cellular functional assays (HEK293 cells for 5-HT_{2A} and 5-HT_{2C} and CHO cells for 5-HT_{2B}) with detection of IP1 by HTR fluorescence resonance

energy transfer. Serotonin receptor activation and antagonism were assessed, and the results represent relative activity compared to 1 μM serotonin. Biological assays were performed by Eurofins (www.eurofins. com) on a fee-for-service basis.

3.3 Results and Discussion

3.3.1 Proof-of-Concept Case Study

DINGOS has two main algorithmic objectives. The first is to produce compounds that are synthetically feasible, that is, ones that can be successfully synthesized given the resources of a standard organic chemistry laboratory. The second objective is that DINOGS produces molecules that are similar to a provided template of interest. This "similarity" is defined by the choice of molecular descriptor and design metric. The underlying hypothesis of similarity-driven *de novo* design is that, with an appropriate measure of similarity, we can expect that two molecules sharing a high degree of similarity will, similarly, share chemical and biological properties [105]. Our choice of similarity metric represents our underlying design hypothesis, with the molecules produced representing evidence that either supports or rejects this hypothesis. In order to evaluate the performance of the DINGOS algorithm, a case-study was constructed. Descriptor and metric functions were chosen, along with a set of four bioactive ligands of interest. A series of chemical reactions were compiled within the rule-based method module, and a predictive machine learning model was trained for the building block recommendation system. The goal of the case-study was to see if DINGOS could successfully generate designs that were similar to the respective templates, synthesizable with the provided synthetic pathways, and, ideally, active against the biological targets of the respective template ligands.

3.3.2 Template Ligands

In order to test the DINGOS method, we required a set of template ligands. These template ligands represent the molecules that we would like to emulate with DINGOS. For this, we compiled a set of four drug molecules: alectinib, an ALK inhibitor used as a treatment for non-small cell lung cancer [106], cariprazine, a dopamine D2/D3 partial agonist used as a treatment for schizophrenia and bipolar disorder [107], osimertinib, an EGFR inhibitor used as a treatment for non-small cell lung cancer [108], and pimavanserin, an inverse agonist against the serotonin receptor proposed as a potential anti-psychotic [109](see Figure 3.6).

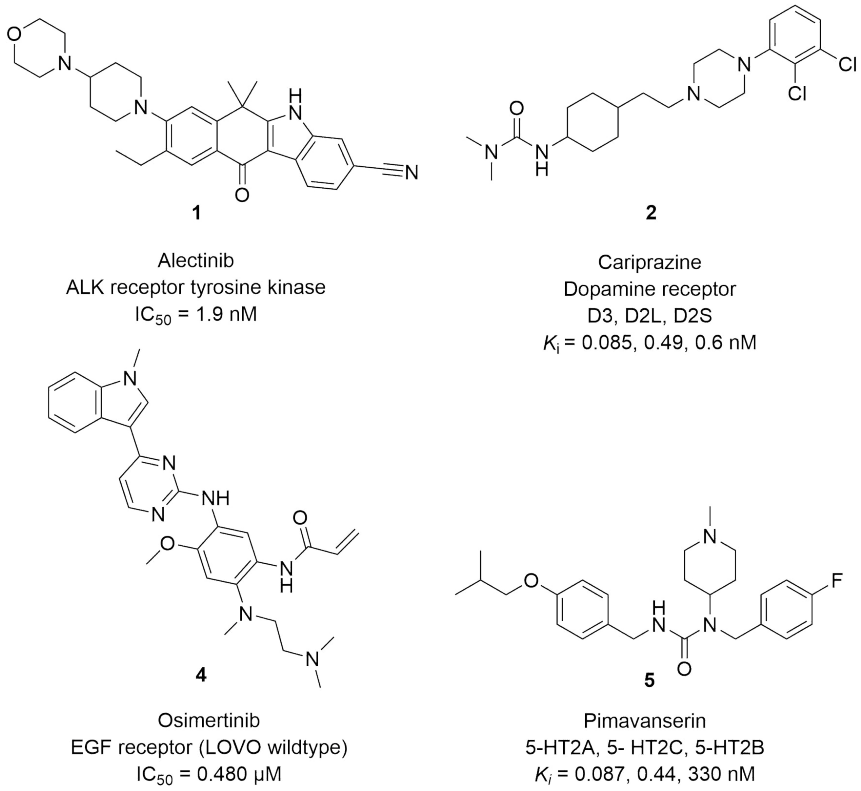


FIGURE 3.6: Structures of the four FDA drug compounds used as the template ligands. These consist of alectinib (**1**), cariprazine (**2**), osimertinib (**3**), and pimavanserin (**4**)

3.3.3 *In silico* Analysis

DINGOS produced four sets of 300 *de novo* designed molecules. DINGOS' overarching goal was to produce populations of molecules that were both similar to the provided templates and synthetically feasible. Similarity in the context of this work is defined in terms of the number of shared bits between the binary fingerprints of two molecules; however, we are also interested in the overall drug-like and bioactive character of the molecules and to what degree this agrees with that of the template ligands. In this section, we interrogate the similarity of the designs, considering the distance values obtained as well as physico-chemical properties and predictive activities.

3.3.4 Distance Analysis

The structures generated by DINGOS are optimized for their distance to the template ligand. The underlying hypothesis is that structures with small distance values are more likely to share biological properties with the template ligand than those that are dissimilar [110]. To investigate DINGOS' performance with respect to distance optimization, the distance values of the *de novo* designs relative to those of their respective templates were calculated. In evaluating the performance, we are posed with a problem, namely, defining what a "close distance" is in this context. In order to determine this, we consider two extreme cases. In the "worst case" scenario, we consider the case where all of the designs produced would be less similar to the template than those of the compound database used. This is the worst-case, as we would see superior performance by simply screening our compound database, and hence,

failed to produce any molecules with improved similarity to our desired template ligands. To this end, we sorted the compound database according to its distance to the four template ligands and compared the distance values obtained. It should be noted that for this analysis the molecular weight and reaction filters were not applied as to not unfavourably bias the results towards the DINGOS algorithm. In addition to this analysis, the distance values of the DINGOS designs were also compared to those of the ChEMBL database [104]. This represented our "best-case" scenario. The ChEMBL database is a chemical database encompassing over 1.8 million molecules along with their curated bioactivity values. This represents a significant proportion of our current biochemical knowledge. Outperforming the ChEMBL database would mean that DINGOS produced designs that were more similar than any molecules previously reported within ChEMBL. Figure 3.7 shows a comparison of the distance distributions obtained between the DINGOS designs and the 300 top ranked compounds from both the compound database and the ChEMBL set (300 encompasses the entire set of DINGOS designs). As can be seen, the DINGOS designs produced a larger range of values, with the construction and ChEMBL distributions possessing a lower interquartile range. A comparison of the median values showed that for each of the four FDA template ligands, a median value lower than that of the compound database was obtained, while for the ChEMBL set, the median distance values were consistently larger.

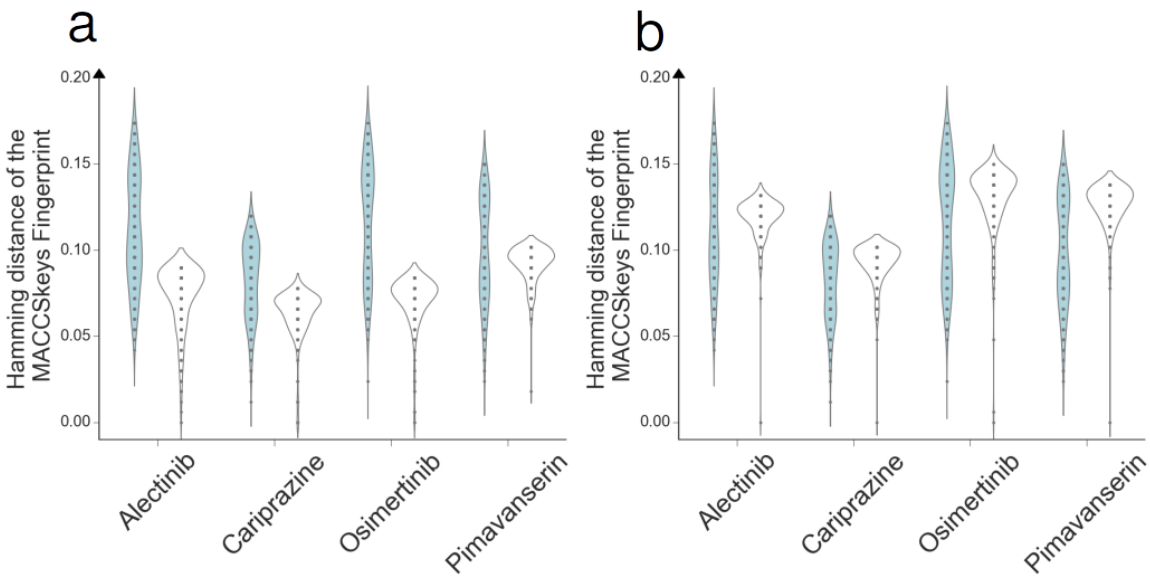


FIGURE 3.7: Distance distributions of the DINGOS designs (light blue) compared against the top 300 ranked compounds from the compound database (white) **a** and the ChEMBL database (white) **b**. Reprinted with permission from Button *et al.* [82]

The DINGOS distance distribution possessed a large range of values. A further comparison was made between the top 20 ranking molecules. These represented a focused set of ranked compounds. The comparison of the distance distributions can be seen in Figure 3.8. As can be seen, just as in Figure 3.7, DINGOS consistently

obtained a lower median distance than that of the compound database. These results show that DINGOS was indeed capable of producing designs that improved upon the similarity of its initial compound database. In comparing to the ChEMBL database, it was observed that ChEMBL possessed a lower median distance value for three of the four drug templates (alectinib, cariprazine, osimertinib).

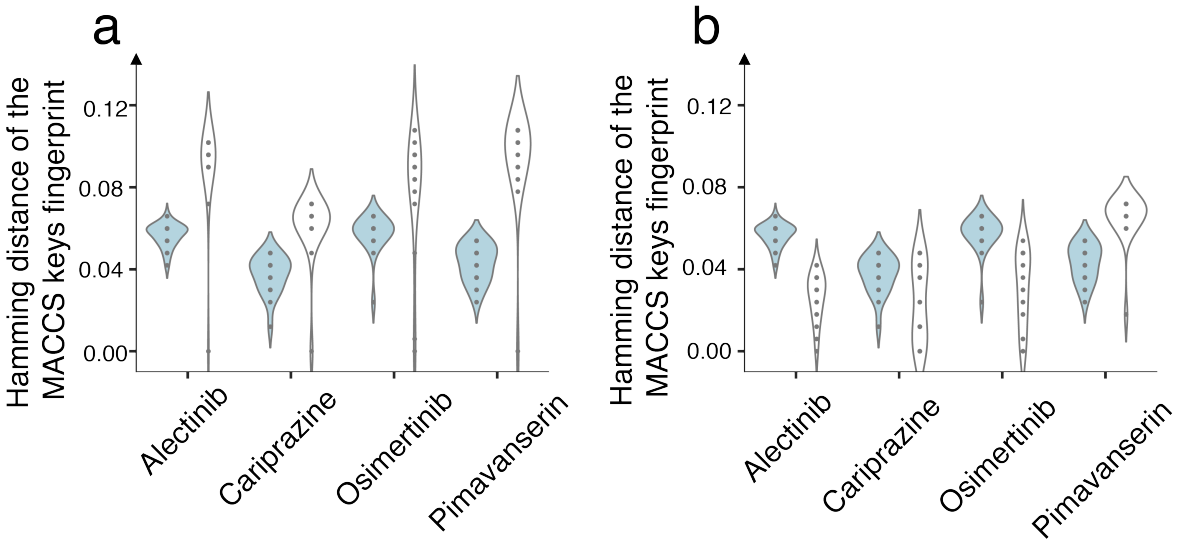


FIGURE 3.8: Distance distributions of the top 20 ranked DINGOS designs (light blue) compared against the top 20 ranked compounds from the compound database (white) **a** and the ChEMBL database (white) **b**. Reprinted with permission from Button *et al.* [82]

While the distributions were of a comparable range, the largest difference in median distance observed was 0.03. These results indicate that, despite DINGOS' capabilities at assembling more similar designs, the compounds produced were still more dissimilar than those extracted from the ChEMBL database. One notable exception was the case of pimavanserin, in which a lower median distance value compared to ChEMBL was observed. Although only occurring in one out of four cases, this result highlights DINGOS potential to produce novel chemical entities that are similar to a desired template ligand. The results of this analysis are summarized in Table 3.1.

TABLE 3.1: A summary of the median distance values obtained from the distance analysis presented in Figures 3.7 and 3.8. Median distance values are reported along with the inter-quartile ranges (IQR) shown in brackets

Template	DINGOS Top 300	ChEMBL Top 300	Sigma Aldrich Top 300	DINGOS Top 20	ChEMBL Top 20	Sigma Aldrich Top 20
alectinib	0.11 (0.05)	0.08 (0.02)	0.12 (0.01)	0.06 (0.006)	0.03 (0.02)	0.10 (0.01)
cariprazine	0.08 (0.04)	0.07 (0.01)	0.10 (0.01)	0.04 (0.01)	0.02 (0.03)	0.065 (0.006)
osimertnib	0.13 (0.06)	0.08 (0.01)	0.14 (0.02)	0.06 (0.006)	0.03 (0.03)	0.09 (0.01)
pimavanserin	0.10 (0.05)	0.10 (0.01)	0.13 (0.01)	0.05 (0.01)	0.069 (0.007)	0.1 (0.02)

3.3.5 Physico-Chemical Properties

In the previous section, it was shown that DINGOS was capable of producing compounds that were similar to the provided template ligands. In this section we now seek to determine if this similarity corresponds with shared physico-chemical properties with the respective template ligand. The goal of similarity-based *de novo* design is that generating similar structures will result in molecules that share properties of interest with our template ligand. While we are often concerned with the biological properties of the molecules, it is also of interest for these compounds to reflect the various physico-chemical attributes of the provided template ligand. For this study, we chose to measure the Lipinski rule-of-5 properties, which were originally prescribed in Lipinski *et al.* [111] as being important for bioavailability. To further determine if these are desirable attributes for our *de novo* designs, we compared these results to a series of bioactive sets extracted from ChEMBL. Four sets of compounds were extracted, each sharing the biological targets of the four template ligands. Only compounds with K_i values less than 10 nM were considered as active.

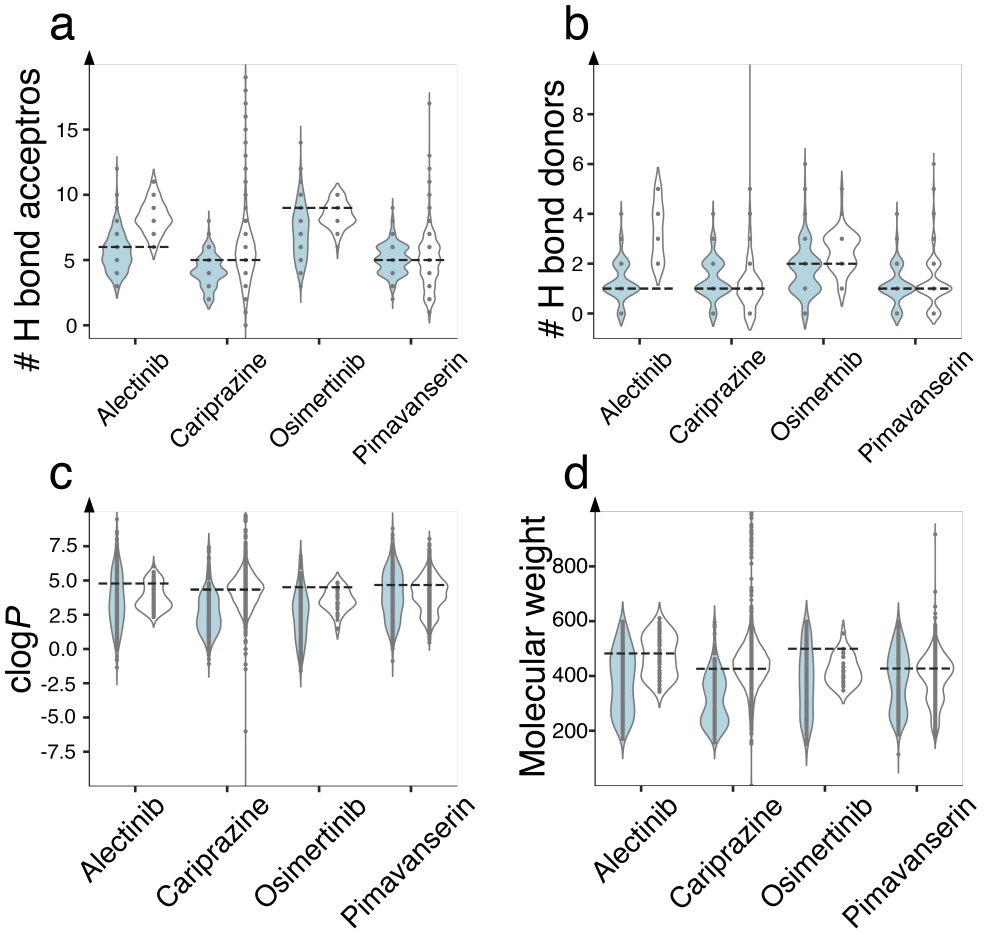


FIGURE 3.9: Distributions of the rule-of-5 properties of the *de novo* designs. a) Number of hydrogen-bond acceptors (HBA), b) number of hydrogen-bond donors (HBC), c) calculated logP (clog$P$), and d) molecular weight. The DINGOS designed populations (300 molecules, light blue), bioactive ligands extracted from ChEMBL (white) that show activity against the respective template's target, and the value of the corresponding template compound (dashed line). Each plot in the figure represents the estimated probability distribution of the property values with dots representing the explicit data points. These distributions are mirrored to aid in readability of the plots. Reprinted with permission from the publication Button *et al.* [82]

As can be seen from Figure 3.9, both the DINGOS designs and the extracted bioactives agreed well with the physico-chemical properties of the template ligands. Table 3.2 summarizes these results. The one notable exception is the case of alectinib, in which the DINGOS designs showed a stronger adherence to the template ligand with respect to the number of hydrogen bond donor and acceptor groups. These results show that the DINGOS algorithm did indeed conserve the physico-chemical properties of the template ligands in the *de novo* designs. We see that the DINGOS designs showed a stronger adherence to the number of hydrogen bond donors and acceptors, while they underestimated the ClopP and molecule weight.

TABLE 3.2: Summary of the median values obtained in the physicochemical analysis. The median Lipinski rule-of-5 values were calculated for each of the DINGOS *de novo* and ChEMBL sets, as well as the corresponding template ligands. Median values are reported along with their IQR shown in brackets.

Template	Dataset	HBA	HBD	clog$P$	MW
alectinib	Template	6	1	4.8	482
cariprazineb	Template	5	1	4.3	426
osimertinibb	Template	9	2	4.5	499
pimavanserin	Template	5	1	4.7	427
alectinib	DINGOS	5 (1)	1 (1)	3.5 (2.9)	389 (190)
cariprazineb	DINGOS	4 (2)	1 (1)	2.4 (2.1)	309 (145)
osimertinibb	DINGOS	7 (2)	2 (1)	2.7 (2.7)	376 (202)
pimavanserin	DINGOS	5 (2)	1 (0)	3.8 (2.7)	399 (180)
alectinib	ChEMBL	8 (1)	3 (2)	3.6 (1.5)	470 (125)
cariprazineb	ChEMBL	5 (2)	1 (0)	4.4 (1.4)	442 (86)
osimertinibb	ChEMBL	8 (1)	3 (1)	3.4 (1.1)	418 (88)
pimavanserin	ChEMBL	5 (2)	1 (1)	4.1 (1.7)	410 (122)

3.3.6 Target Prediction

Ultimately, the goal of any drug design project is to produce bioactive ligands. While the activity of a given compound can only truly be confirmed experimentally for larger sets of molecules, the process of synthesizing and biotesting can quickly become prohibitively expensive. One solution to this problem is to employ predictive methods in order to estimate the activity of a given compound *in silico*. Considering the large number of compounds (1200) that would require full synthesis, we chose to perform target prediction to estimate the bioactivity of the populations. For this study, the algorithm SPiDER was chosen. SPiDER is a similarity based target prediction method that was developed by Reker *et al.* [99]. Since its development, it has been utilized in a series of successful drug design studies [112–114]. SPiDER performs target prediction by clustering a set of 12,000 known bioactive ligands in descriptor space along with the query molecule whose target activity we're interested in. This set of 12,000 ligands is a well-curated list of bioactive molecules representing 251 distinct biological targets. SPiDER uses the cluster assignment of the query molecule, along with the distribution of in-cluster ligands, to assign a series of p-value scores for each target. These scores give the pseudo-probability of inactivity. The lower the p-value for a given target, the lower the probability that the molecule is inactive against this target. In essence, this is a guilt by association, if the ligands within the query molecule's cluster all share the same biological target, SPiDER asserts that there is a high probability that the query molecule also shares this activity. SPiDER uses both the CATS descriptor [115], a topological pharmacophoric based descriptor for evaluating bioactive scaffolds, and the MOE descriptor [98], a descriptor encompassing a list of physicochemical properties, to encode the molecules in descriptor space. For the purpose of evaluating the DINGOS designs, the pharmacophoric and physicochemical descriptors utilized by SPiDER are an orthogonal measure to the purely structure-based descriptor used for generating the *de novo* designs. While target prediction results in an increase in efficiency, it also introduces error in our

measurement originating from inaccuracies in the chosen predictive model. To account for this, we first verified the predictive model for our problem of interest. The bioactive sets used in the physico-chemical analysis section 3.3.5 were used as our control set. SPiDER predictions were performed and the proportion of compounds that were accurately predicted as active was measured. These results are summarized in Table 3.3. It was found that SPiDER achieved an accuracy between 86-100% on all four of the set, thus illustrating its applicability for this given task.

TABLE 3.3: Proportion of known bioactive compounds predicted active against the target subgroups of the template compounds. Bioactive compounds showing a $K_i < 10$nM against the templates' targets were extracted from ChEMBL. Predictions were performed with the target prediction software SPiDER. All four sets showed between 86-100% predictive accuracy, thus validating SPiDER as a prediction tool against these templates.

	Target Subgroup	Number of Bioactives	Number of Predicted Active	Percent Active (%)
alectinib	Tyrosine Kinase	79	72	91
cariprazine	Dopamine Receptor	3539	3057	86
osimertinib	Tyrosine Kinase	50	50	100
pimavanserin	Serotonin Receptor	1898	1777	94

Having verified SPiDER for the activity prediction of the template ligands primary biological targets, we performed activity prediction on the DINGOS designed sets. For each of the *de novo* designed sets, above 50% of the compounds were predicted as active against the respective template ligands' biological target (as shown in Figure 3.10). In the case of pimavanserin, a vast majority of the compounds, 90%, were predicted as active. While this high degree of predicted activity was encouraging, a concern was that this could be due to DINGOS solely generating compounds with a single bioactive scaffold, and not reflective of true *de novo* design. In order to investigate this, we performed scaffold analysis on the *de novo* populations.

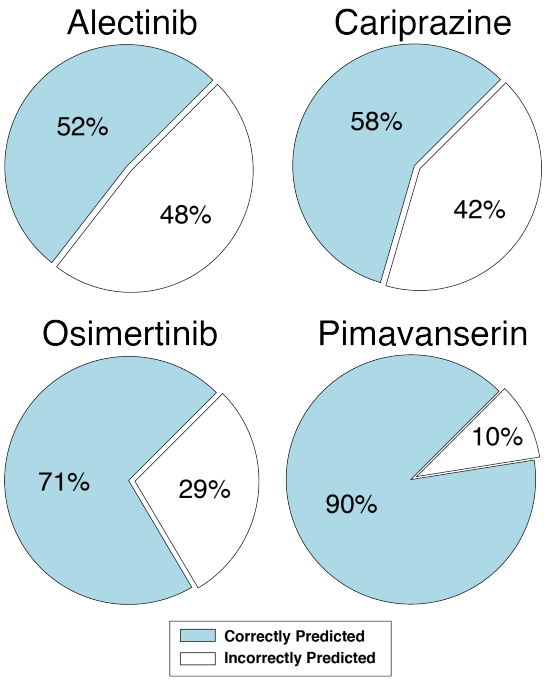


FIGURE 3.10: Proportion of DINGOS generated compounds predicted active against the respective template's target. SPiDER software was used for all target predictions. Predictions were presented as p-values, pseudo-probabilities of misclassification. A compound was considered active against the desired target, if it showed a p-value of less than 0.1. All four sets showed a predicted active proportion of above 50%. For pimavanserin, 90% of the compounds produced were predicted active against the serotonin target receptor family. Reprinted with permission from Button *et al.* [82]

3.3.7 Scaffold Analysis

Scaffold analysis was performed in order to evaluate the overall structural novelty of the *de novo* designs. While DINGOS' primary goal is to produce compounds with maximal similarity to the template ligand, it is also of interest that these compounds possess a high degree of diversity and novelty. One feature of the DINGOS algorithm intended to promote the diversity, and hence novelty, of the designs is the selection of a unique starting molecule for each design. By selecting a unique starting molecule, we force the molecules to incorporate varied structural elements. There are multiple reasons why inherent novelty is of interest. One primary reason is that by promoting diversity we gain a greater understanding of DINGOS' capabilities for generating similar designs. If, in fact, DINGOS produced only near-identical structures with only minute structural differences and all sharing a common scaffold, we would likely observe disproportionately close distances in the design population. These "highly similar" *de novo* designed structures, while scoring well, would not represent the true *de novo* design we're interested in, that is, generating structurally novel molecules that are similar to our template ligand of interest. An additional reason motivating our interest in diversity, is that it enables us to better interrogate our underlying design hypothesis. The diversity in the generated designs provides us with multiple structural variants on our template ligand. By synthesizing and testing these, we can better understand the nature of the template ligand's bioactivity. In a scenario

in which diversity was not present within the designs, we may indeed only end up confirming the existence of single, known bioactive scaffolds. In order to quantify the degree of diversity in our *de novo* designs, we calculated the percentage of unique Murocko scaffolds for each of the sets of *de novo* designs. Additionally, we measure the percentage of unique scaffolds for the corresponding construction and ChEMBL sets. These results are summarized in Table 3.4.

TABLE 3.4: Percentage of unique Murcko scaffolds from the DINGOS, ChEMBL and compound database populations. All scaffold were calculated in the RDKit suite. Reprinted with permission from the publication Button *et al.* [82]

Template ligand	DINGOS designs (%)	ChEMBL dataset (%)	Compound database (%)
alectinib	88	73	63
cariprazine	43	45	25
osimertinib	87	61	85
pimavanserin	59	57	36

As can be seen, the DINGOS designs possessed a higher degree of diversity than both the ChEMBL and compound databases, with the exception of the case of cariprazine, in which the ChEMBL and DINGOS designs differed by 2%. Cariprazine and pimavanserin had a significantly lower percent of unique scaffolds compared to that of alectinib and osimertinib ($> 20\%$). Both cariprazine and pimavanserin represent more simplistic structures in comparison to alectinib and osimertinib, and this may have resulted in DINGOS producing multiple designs sharing the same scaffold. To further understand this discrepancy, we examined the five most frequent scaffolds observed for each of the DINGOS sets (see Figure 3.11). We observed that the scaffolds generated for the cariprazine/pimavanserin sets generally contained more reduced scaffolds (lower molecular weight, fewer atoms) in contrast to the alectinib/osimertinib sets. We also saw that the cariprazine/pimavanserin sets favoured 1-2 top scaffolds in the designs, the top scaffold represented 12 and 9% of the population respectively, while the alectinib/osimertinib scaffolds were more homogeneously distributed, the top scaffolds were only representative of 2 and 1% of the *de novo* sets.

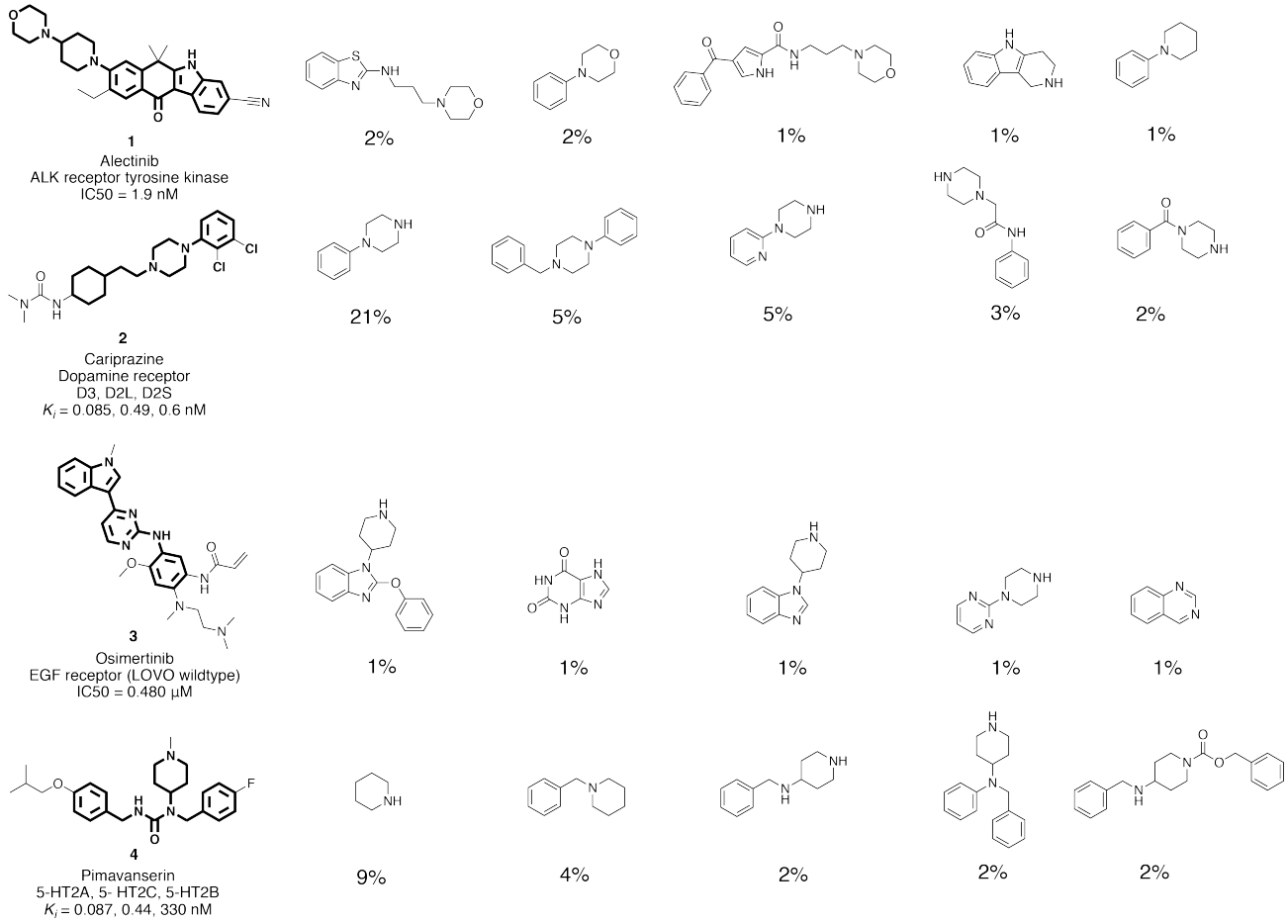


FIGURE 3.11: The four drug templates, alectinib (**1**), cariprazine (**2**), osimertinib (**3**) and pimavanserin (**4**) with their Murcko scaffolds (atom scaffolds) highlighted in bold as well as the top five most frequently observed Murcko scaffolds of the DINOGS designs, ordered according to occurrence. The occurrence of each distinct molecular scaffold was measured for the four *de novo* sets, and the number of occurrences for the five most frequent scaffolds was determined. The occurrences are presented as a percentage of the *de novo* sets (300 molecules each). Reprinted with permission from the publication Button *et al.* [82]

3.3.8 Synthetic Feasibility and Bioactivity

Apart from proposing designs that are similar to a given template ligand, DINGOS other main objective was to produce compounds that are synthesizable, thus enabling production and biological evaluation of the proposed structures. This synthesizability is key to determining the validity of our design hypothesis and improving our understanding of the target system. In order to investigate this aspect of the DINGOS algorithm, we selected one compound from each of the *de novo* design sets for synthesis. The synthesized compounds were then tested for activity against the corresponding template ligand's biological target.

3.3.9 Design Synthesis and Biotesting

For selection of the drug candidates, we used the 20 top ranking molecules considered in Section 3.3.4. These sets were further reduced by removing all compounds that were not predicted active by SPiDER, that is, compounds with a p-value greater than 0.1 (10%) for the intended target. From the remaining sets, one compound was selected for synthesis. This choice was motivated by the availability of the corresponding building blocks. As part of the output, DINGOS not only gives the structure of the *de novo* design, but also the corresponding synthetic pathway, with all pathways being based on commercially available molecules from our compound database. For the synthesis, we imposed the restriction that all compounds must be synthesized in accordance with the synthetic pathways proposed by DINGOS. As reaction conditions and solvent choice is not a part of the DINGOS algorithm, this was left to the discretion of the chemist. The selected compounds, along with their syntheses, are shown in Figure 3.12.

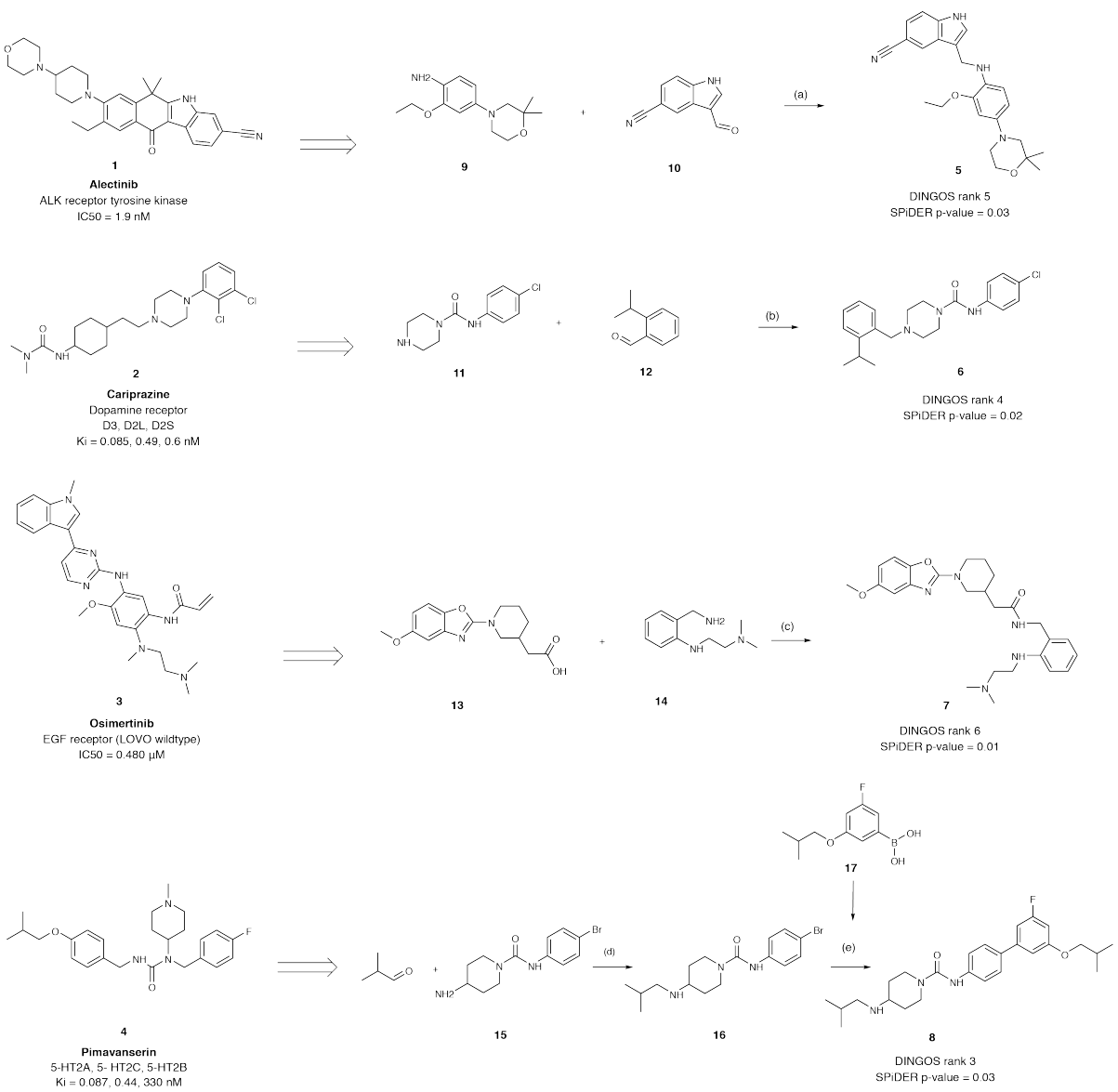


FIGURE 3.12: Synthetic reagents and conditions: **a** DCE (1,2-dichloroethane), $NaB(OAc)_3H$, room temperature, 5 h, product instable; **b** DCE, NaB(OAc) 3 H, 50 °C, 48 h, 64%; **c** $CHCl_3$, EDC (1-ethyl-3-(3-dimethylaminopropyl)carbodiimide), 4-DMAP (4-(dimethylamino)pyridine), reflux, 2 h, 51%; **d** DCE, $NaB(OAc)_3H$, room temperature, 16 h, 76%; **e** dioxane/DMF, Cs_2CO_3, $Pd(PPh_3)_4$, reflux, 12 h, 33%. IC_{50}, half-maximum inhibitory concentration. K_i inhibition constant; ALK, anaplastic lymphoma kinase; EGF, epidermal growth factor; LOVO, human LoVo cancer cells; 5-HT, 5-hydroxytrytamine (serotonin) receptor. Reprinted with permission from the publication Button *et al.* [82]

Three of the four selected compounds we successfully synthesized with yields ranging from 33-64%. These compounds corresponded to the cariprazine, osimertinib, and pimavanserin designs. For the alectinib, we obtained the correct mass signal (m/z 405.4 $[M + H]^+$), however, the sample proved to be light and water sensitive, and hence, could not be fully structurally characterized. The NMR spectra for the synthesized compounds are shown in Appendix A.2.2. Having successfully synthesized

three of the four designs, we now sought to evaluate their biological activity towards the target of the respective template ligands. All three of the designs were tested *in vitro* at a concentration of 10 μM. The cariprazine design was tested against the D_{2S}, D_{2L}, and D_3 dopamine receptor subgroups, the osimertinib design against the EGFR, and the pimavanserin designs against the 5-HT_{2A}, 5-HT_{2B}, and 5-HT_{2C} serotonin receptor subgroups. Figure 3.13 summarizes the bioactivity results. As can be seen, both the cariprazine and osimertinib design did not show agonism or antagonism against the intended target. The pimavanserin design showed preferential antagonistic activity against the 5-HT_{2B} subgroup. Following these results a dose-dependent study was conducted, revealing partial antagonism equivalent to a relative activity of 1 μM of serotonin. While partial activity was observed in the pimavanserin design, this value is significantly lower than that of the pimavanserin template ligand (K_i = 0.087 nM). The inactivity observed in the DINGOS designs refutes the underlying designs hypothesis used in this case-study, that is, that the hamming loss of the molecules' MACCSkeys binary fingerprint is proportional to the likelihood of their shared bioactivity.

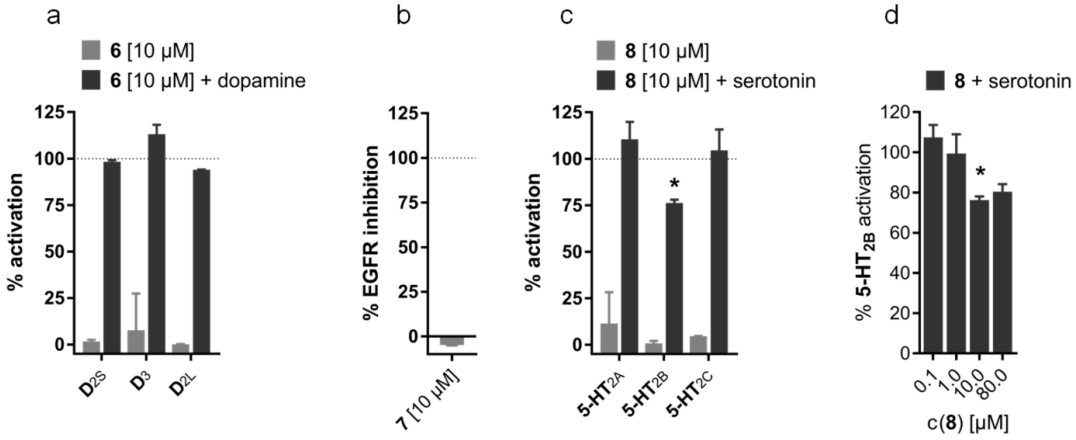


FIGURE 3.13: *In vitro* pharmacological activities of designs **6**-**8** at 10 μM concentration on the molecular targets of their respective templates **2**-**4**: **a**) cariprazine mimetic **6** showed no agonistic or antagonistic activity on dopamine receptors D_{2S}, D_{2L} or D_3. **b**) Design **7** derived from Osimeritinib (**3**) was inactive on EGFR. **c**) pimavanserin mimetic **8** significantlyantagonized 5-HT_{2B} activation by serotonin and **d**) dose-response characterization indicated dose-dependent partial antagonistic activity of compound **8**. All results are the mean±S.E.M., N=2, *p<0.05. Reprinted with permission from the publication Button *et al.* [82]

3.3.10 Large *in silico* Design Study

Having established DINGOS with the initial case-study, we now sought to investigate the influence of the choice of template, and to determine for which templates DINGOS would perform best. In order to do this, we compiled a set of 927 drug molecules extracted from the DrugBank database [116]. These molecule were filtered to be within a molecular weight range of 300-600 g mol^{-1}. The same run parameters were used as those in the case-study, again using the MACCS keys and Hamming distance as the descriptor and metric functions respectively. Figure 3.14 shows the median distance values obtained for both the entire set of 300 designs, the top 20 ranked

molecules, and the top, most similar molecule. For this analysis, we removed all zero step designs, that is, designs in which the starting molecule was selected as the optimal product. As the compound database contained many of the existing templates and close analogues, this was done in order to more fairly evaluate DINGOS' capability at generating similar designs.

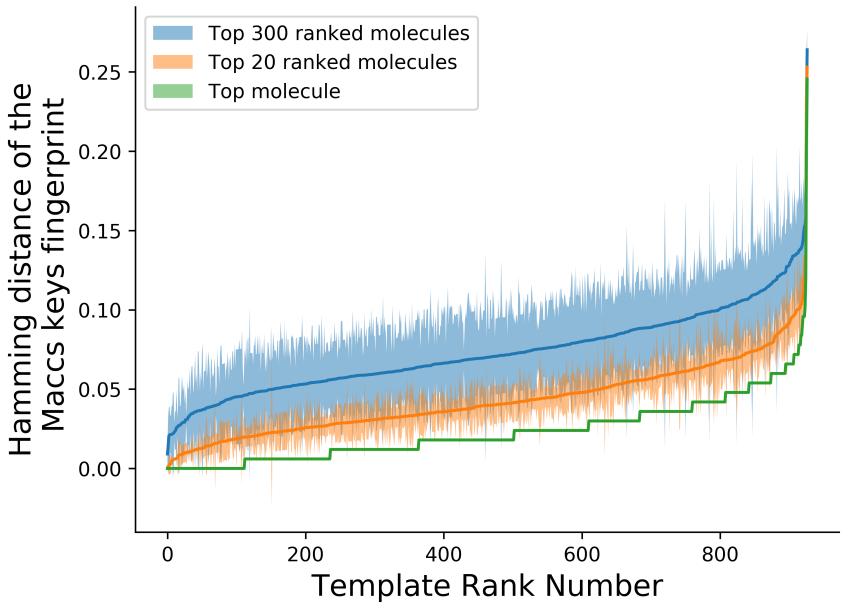


FIGURE 3.14: Comparison of the median distances for the top 300 (blue), top 20 (orange), and top (green) ranked DINGOS *de novo* designs for each of the 927 separate drug-template *de novo* design populations, ordered by their rank on the x-axis. The area around the median curve represents the IQR. Distinct steps are observed amongst the top molecule due to the discrete nature of the binary descriptor.

Broadly, we see that the Hamming distance varied between a value of 0.0 and 0.25, with the zero distance values representing an "optimal" design, that is, as similar as possible under the given representation. Table 3.5 shows the first, second, and third quartiles of the three distance distributions. Amongst the four template ligands considered in the case-study, the *de novo* populations possessed a median distance value between 0.08 and 0.11. Comparing these results with those of the DrugBank set, we found that 65% of the template ligands (601 template ligands) resulted in *de novo* populations with a median distance value less than 0.8. For the top 20 designs, in which a median value of 0.03-0.06 was observed in the case-study, 31% of the DrugBank templates (286 template ligands) possessed a median distance less than 0.03. These templates represent drug-design challenges in which DINGOS would have out-performed the results seen in the case-study.

TABLE 3.5: A summary of the 1st, 2nd and 3rd quartile values for the drug-template distance distributions presented in Figure 3.14

Size of set	1st Quartile	2nd Quartile	3rd Quartile
Top 300 ranked mols	0.06	0.07	0.09
Top 20 ranked mols	0.03	0.04	0.06
Top molecule	0.01	0.02	0.04

The initial flat line of the 'Top molecule' plot indicates designs that possessed a zero distance value, i.e. designs that were "optimal" under the MACCS keys. From the set of 927 template ligands, 12% of the *de novo* designed populations (112 template ligands) contained an optimal design. Figure 3.15 shows a selection of three of these optimal designs, along with their corresponding template ligands. Despite the zero distance, structural variation between the designs and template ligand can be seen. These differences highlight the limitations of the descriptor used.

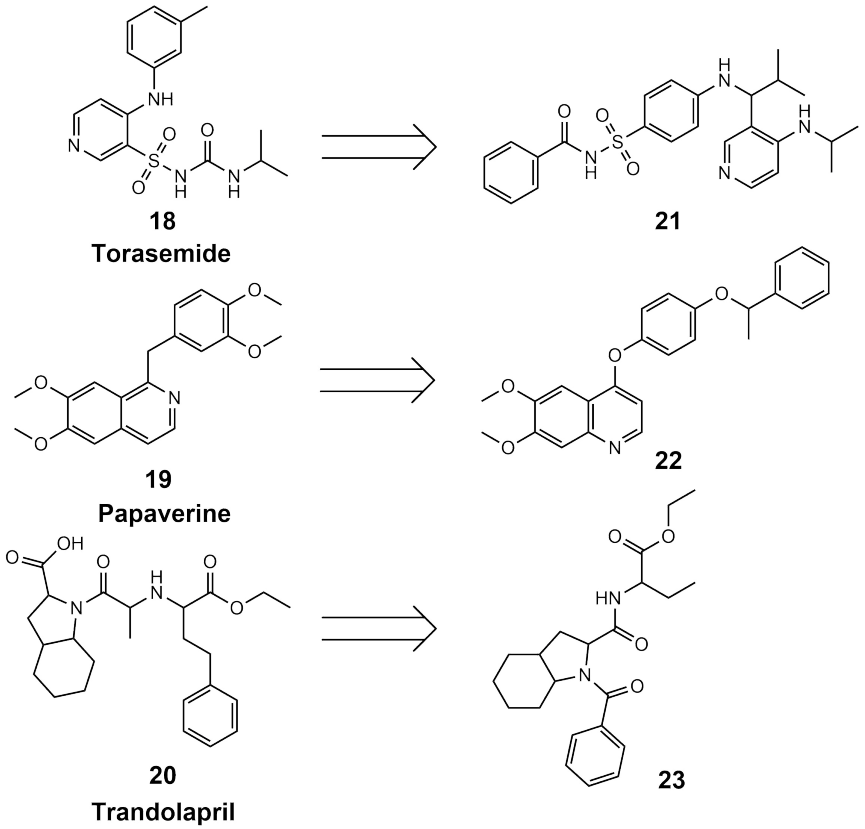


FIGURE 3.15: Structural comparison between template ligands and the "optimal" DINGOS designs, defined as a Hamming distance of zero. This has been observed in the case of the torasemide [117], papaverine [118], trandolapril [119] template ligands. One can see that a zero distance does not imply identical structure.

3.4 Conclusion

In silico analysis showed that DINOGS is capable of producing designs with improved similarity towards the template ligand. Analysis of the physico-chemical properties and Murcko scaffolds showed a strong adherence to the physico-chemical properties of the template ligand, along with a high degree of scaffold diversity in the *de novo* populations. Target prediction showed that above 50% of the designs were predicted as active against the intended target, with pimavanserin showing 90% predicted actives. One design was selected from each population and of the four, three were successfully synthesized with the proposed synthetic pathway. The synthesized compounds were then biotested, and of these three, one compound, the pimavanserin mimetic, showed partial antagonism of the serotonin 5-HT_{2B} receptor. Despite this, comparable activity to the template ligand was not observed, refuting the chosen design hypothesis. Having established DINGOS with this initial case-study, in further sections of this thesis we seek to extend the DINGOS method towards larger scale and higher degree drug design problems. Within this section, a total of four syntheses were attempted. In the next chapter, we expand upon this, integrating DINGOS with contemporary continuous-flow chemistry to generate a larger set of *de novo* designs. These results enable us to better determine the overall synthesizability of the designs, and to further determine the applicability of the underlying design hypothesis. A major limitation of the current implementation of the DINGOS method is the rigidity of the machine learning model used. The machine learning model used in this study was trained explicitly on the MACCS keys descriptor values of the training data. This means that in order to change our underlying design hypothesis, we would be required to develop and train an entirely new machine learning model. While this is indeed possible, the time investment, along with the difficulties associated with training a model on other, higher dimensional representations, makes this a significant limitation. In Chapter 5, we investigate potential solutions to this problem and use these solutions to interrogate the influence of the underlying design hypothesis.

Chapter 4

Amalgamating DINGOS with Automated Synthesis

The work presented within this chapter was conducted in collaboration with Berend Huisman, Cyrill Brunner, Benedikt Winkler, and Alice Lessing. Data analysis and operations of the DINGOS algorithm were performed by Alexander Button. Chemical synthesis, analysis and operations of the continuous flow system were performed by Berend Huisman and Benedikt Winkler. The continuous flow system was developed by Berend Huisman. Berend Huisman, Benedikt Winkler, and Alice Lessing all contributed to the testing and implementation of the continuous flow system used in this chapter. All biotesting and bioanalysis was performed by Cyrill Brunner. All figures, passages, and methods provided by the collaborators are explicitly stated as such.

A primary motivation for the development of the molecular design software DINGOS is the need to find novel, optimised, and bioactive compounds for targets of interest. Accompanying the structure and distances scores, DINGOS also provides the synthetic pathways used for the in silico de novo design output. The existence of these explicit synthetic routes makes DINGOS an attractive complement to the field of automated synthesis. Automated synthesis, in contrast to conventional organic synthesis, aims at synthesizing synthetic molecules through the use of a mechanical apparatus, with minimal direct human involvement. Automated systems can run continuously with lower resource and time expenditure per reaction. The goal of this is to improve upon the efficiency and safety of conventional synthesis, and facilitate the synthesis of diverse compounds-of-interest. Various automated procedures have been proposed with the two main methods being automated batch synthesis and continuous micro-flow. DINGOS is a potentially promising tool for integration within such an automated workflow, as all synthetic routes are explicitly provided, and all chemical components consist of real, commercially available building block molecules. In this chapter, we explore a proof of principle case study, in which DINGOS was used in conjunction with an established continuous micro-flow system to generate novel carbonic anhydrase inhibitors. With acetozolamide selected as the template ligand, two rounds of de novo design, synthesis, and bio-testing were performed, generating 15 and seven compounds respectively. Overall, five bioactive compounds were produced, with three of the de novo designs showing low micromolar binding affinity toward carbonic anhydrase II.

4.1 Introduction

4.1.1 Continuous Flow System

Continuous flow offers the opportunity to greatly increase the synthetic output within a given drug design project. Despite this, however, current continuous flow systems are still racked with many limitations. Currently, the field of automated synthesis is still in its infancy, with many chemical reactions which are commonplace within conventional organic chemistry not having yet been successfully established in automated synthesis [120]. There are many reasons for these discrepancies, such as the limited range of viable solvent conditions, limits on the catalysts which can be incorporated, and restrictions on possible temperature and pressure ranges within a closed system. While solutions to these limitations are presently being explored, the possibility to account for these restrictions and exclude them from consideration would be advantageous. The DINGOS algorithm, due to its modularity, offers such a potential solution. As the procedures within DINGOS can be readily altered, this allows us to tailor our design strategies to suit the specific constraints of a given automated setup. In the work presented here, we explored the possibility of integrating the DINGOS algorithm with autonomous continuous flow. Within the course of this project, we were provided access to the custom continuous flow system developed by Berend Huisman. An image of the system is shown in Figure 4.1.

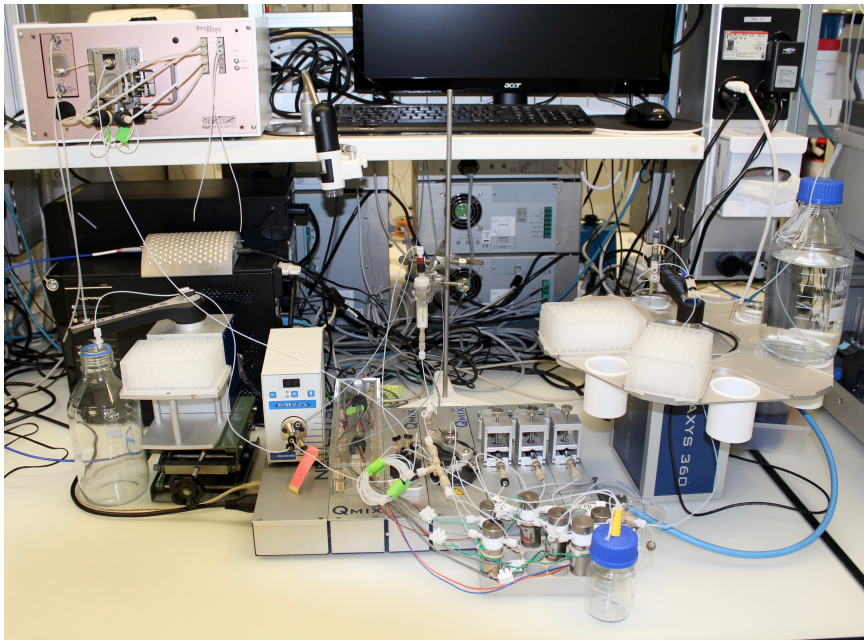


FIGURE 4.1: Photograph of the continuous flow system used for this experiment. Courtesy of Berend Huisman.

Combining the DINGOS algorithm within such a continuous flow system, not only provides key insights into the feasibility of the synthetic pathways suggested by DINGOS, but also provides the opportunity to assess the challenges associated with incorporating such an automated system within our design procedure, and, ultimately, provide further validation of DINGOS' ability to generate NCEs with similar bioactivity to their template ligand. In the work presented here, both the DINGOS algorithm and the continuous flow system were used to develop an active drug design learning cycle.

4.1.2 Active Drug Design Learning Cycle

Active learning is a machine-learning approach in which an algorithm trained on existing data, such as DINGOS, generates new data points which are then used to inform and further improve the algorithm [121, 122]. By cycling through this procedure, we are able to greatly improve the predictive accuracy of our models [123, 124], and generate previously unconsidered data points which can provide a greater degree of insight for our underlying hypothesis. Despite these advantages, one major shortcoming is the cost of producing of new data points. For purely computational problems, such as those seen in computer chess and AlphaGo gameplay [56, 57], this is governed solely by the costs of the algorithm. In problems involving real-world feedback, such as in drug design, there is a significant cost in the production of new data points. In the context of drug design, the synthesis and biotesting of novel structures present the major bottleneck, severely limiting the throughput of our active learning cycle, and, subsequently, the rate of improvement of model quality. In the context of a drug design active learning cycle, the combination of DINGOS and automated synthesis offer a promising solution to this problem. In order to facilitate improved throughput, we require proper communication between the model (DINGOS) and the automated procedure (continuous flow). Here, the overall synthesizability of the designs within our continuous flow system is critical for the possibility of prompting an effective learning procedure. In the work presented here, we investigate the ways in which these two components, the continuous flow system and the *de novo* design tool DINGOS, might be successfully integrated, and how these might lead us on the first steps towards a fully automated, active learning driven drug design cycle. In order to facilitate testing of the de novo designs, surface plasmon resonance (SPR) was used to test the binding affinity of the *de novo* designs.

4.2 Methods

Parts of this section were provided by the collaborators, describing the methods as they were performed. All methods descriptions that were provided by the collaborators and their work are explicitly stated as such.

4.2.1 General Chemistry

Description provided by Berend Huisman

All chemicals and solvents were reagent grade and used without further purification, unless specified otherwise. All reactions were conducted using a Cetoni flow chemistry setup (Cetoni GmbH, Korbussen, DE) in absolute solvents, unless specified otherwise. NMR spectra were recorded on a Bruker Ultrashield spectrometer (Bruker Corporation, Bremen, DE). Chemical shifts (δ) are reported in ppm relative to TMS (tetramethylsilane) as reference; approximate coupling constants (J) are reported in Hz. Compound purity was evaluated by high-performance liquid chromatography (HPLC) on a Shimadzu LCMS-2020 system (Shimadzu, Kyoto, JP) equipped with a C18 reverse phase column (Macherey-Nagel, Nucleodur C18 HTec, 5 μm) using a gradient (H_2O/MeCN 70:30 + 0.1% formic acid isocratic for 3 min to H_2O/MeCN 5:95 + 0.1% formic acid over 12 min and H_2O/MeCN 5:95 + 0.1% formic acid isocratic for an additional 2 min) at a flow rate of 0.5 mL min^{-1} and UV detection at 254 nm and 280 nm. All final compounds had a purity of >95% (area-under-the-curve

for UV254 and UV280 peaks). Preparative HPLC was carried out on a Shimadzu LCMS-2020 system (Shimadzu, Kyoto, JP) equipped with a C18 (Macherey-Nagel, VP 150/21 Nucleodur C18 HTec, 5 μm) or C8 (Macherey-Nagel, VP 250/21 Nucleosil 300-5 C8) reverse phase column using a gradient (H_2O/MeCN 70:30 + 0.1% formic acid to H_2O/MeCN 5:95 + 0.1% formic acid over 17 min or H2O/MeCN 95:5 + 0.1% formic acid to H_2O/MeCN 30:70 + 0.1% formic acid over 28 min). High-resolution mass spectra were recorded on a Bruker maXis ESI-Qq-TOF-MS (electrospray ionization quadruple time-of-flight mass spectrometry) instrument (Bruker Corporation, Bremen, DE).

4.2.2 Amide Bond Formation Procedure in Continuous Flow

Description provided by Berend Huisman and Benedikt Winkler

Dimethylformamide. The amine reactants were dissolved to a concentration of 0.2 M in DMF. A 2 equiv. of triethylamine was added to each of the amine solutions. A 0.2 M solution of the corresponding acid chloride was dissolved in DMF, and 1 mL of the amine and acid chloride solutions were each added to separate wells within the 96-well plate. 0.75 mL of each solution was aspirated in the syringe pumps and diluted to 1 mL in the running solution (DMF). The two reactant solutions were mixed within the reaction chip and then pumped through the reaction coil. Continuous flow synthesis was performed with a residence time of 5 minutes and at a temperature of 75 °C. A final solution of 2 mL was collected in the 96-well plate.

Dimethylformamide/Acetonitrile. The amine reactants were dissolved to a concentration of 0.2 M in DMF. A 0.2 M solution of the corresponding acid chloride was dissolved in MeCN, and 1 mL of the amine and acid chloride solutions were each added to separate wells within the 96-well plate. 0.75 mL of each solution was aspirated in the syringe pumps and diluted to 1 mL in the running solution (DMF/MeCN). The two reactant solutions were mixed within the reaction chip and then pumped through the reaction coil. Continuous flow synthesis was performed with a residence time of 10 minutes and at a temperature of 70 °C. A final solution of 2 mL was collected in the 96-well plate.

Tetrahydrofuran. The amine reactants were dissolved to a concentration of 0.2 M in THF. A 0.2 M solution of the corresponding acid chloride was dissolved in THF, and 1 mL of the amine and acid chloride solutions were each added to separate wells within the 96-well plate. 0.75 mL of each solution was aspirated in the syringe pumps and diluted to 1 mL in the running solution (THF). The two reactant solutions were mixed within the reaction chip and then pumped through the reaction. Continuous flow synthesis was performed with a residence time of 5 minutes and at a temperature of 55 °C. A final solution of 2 mL was collected in the 96-well plate.

Tetrahydrofuran/Acetonitrile. The amine reactants were dissolved to a concentration of 0.2 M in MeCN. A 0.2 M solution of the corresponding acid chloride was dissolved in THF, and 1 mL of the amine and acid chloride solutions were each added to separate wells within the 96-well plate. 0.75 mL of each solution was aspirated in the syringe pumps and diluted to 1 mL in the running solution (THF/MeCN). The two reactant solutions were mixed within the reaction chip and then pumped through the reaction coil. Continuous flow synthesis was performed with a residence time of 5 minutes and at a temperature of 55 °C. A final solution of 2 mL was collected in

the 96-well plate.

Tetrahydrofuran/Acetonitrile/Dimethylformamide. The amine reactants were dissolved in a minimal amount of DMF, and then diluted in MeCN to a concentration of 0.2 M. A 0.2 M solution of the corresponding acid chloride was dissolved in THF, and 1 mL of the amine and acid chloride solutions were each added to separate wells within the 96-well plate. 0.75 mL of each solution was aspirated in the syringe pumps and diluted to 1 mL in the running solution (THF/MeCN/DMF). The two reactant solutions were mixed within the reaction chip and then pumped through the reaction coil. Continuous flow synthesis was performed with a residence time of 5 minutes and at a temperature of 55 °C. A final solution of 2 mL was collected in the 96-well plate.

4.2.3 Reductive Amination Procedure in Continuous Flow

Description provided by Berend Huisman and Benedikt Winkler

One-step reductive amination without catalyst. The amine reactants were dissolved with 1 equiv. of formic acid to a concentration of 0.2 M in MeCN. A 0.2 M solution of the corresponding aldehyde/ketone was dissolved in MeCN, and 1 mL of the amine and aldehyde/ketone solutions were each added to separate wells within the 96-well plate. 0.75 mL of each solution was aspirated in the syringe pumps and diluted to 1 mL in the running solution (MeCN). The two reactant solutions were mixed within the reaction chip and then pumped through the reaction coil. Continuous flow synthesis was performed with a residence time of 10 minutes and at a temperature of 75 °C. A final solution of 2 mL was collected in the 96-well plate.

One-step reductive amination with a heterogeneous catalyst. A series of 0.5 M amine and 0.33 M aldehyde solutions were prepared in a 1:4 ratio of MeOH/THF. 1 mL of the amine and aldehyde/ketone solutions were each added to separate wells within the 96-well plate. The reactant solutions were taken up into the reaction chip, and the reaction mixture was then pumped through a column containing the heterogeneous catalyst (1:1:0.76 w/w of Celite/$NaBH_4$/LiCl). A final solution of 14 mL was collected in the 96-well plate.

Two-step reductive amination with Methanol. A 1 M solution of the amine and aldehyde/ketone reactants were prepared separately in MeOH. A 1.88 M stock solution of the reducing agent ($NaBH_3CN$) in MeOH was prepared. 1 mL of the amine and aldehyde/ketone solutions were each added to separate wells within the 96-well plate and the reducing agent stock solution was connected to the inlet of the third syringe pump. 0.75 mL of each solution was loaded in the syringe pumps and diluted to 1 mL total volume with the running solvent (MeOH). After mixing both solutions in the reaction chip, the reaction mixture was pumped through the coil. Continuous flow synthesis was performed with a residence time of 20 minutes and at a temperature of 57 °C and afterwards, the reaction mixture was mixed with 2 ml of the reducing agent solution using a T-piece, resulting in 4 mL being collected in the 96-well plate.

4.2.4 Imidazole Arylation Procedure in Continuous Flow

Description provided by Berend Huisman and Benedikt Winkler

Imidazole Arylation. The imidazole reactant was dissolved in DMF to a concentration of 0.015 M. A 0.015 M solution of copper(II)acetate with 2 equiv. of boronic acid, 1 equiv. of triethylamine, and 2 equiv. of pyridine was prepared in DMF. 1 mL of the imidazole and copper(II)acetate solutions were each added to separate wells within the 96-well plate. 0.75 mL of each solution was aspirated in the syringe pumps and diluted to 1 mL in the running solution (DMF). The two solutions were mixed within the reaction chip and then pumped through the reaction coil. Continuous flow synthesis was performed with a residence time of 30 minutes and at a temperature of 130 °C. A final solution of 2 mL was collected in the 96-well plate.

4.2.5 Biophysical Evaluation

Description provided by Cyrill Brunner

Surface plasmon resonance (SPR) measurements were performed on an automated SPR-16 instrument (Bruker Daltonics AG, Hamburg, Germany) for bovine carbonic anhydrase II (bCAII). All experiments were performed at 25 °C with a constant flow rate of 5 μL min^{-1} for the immobilization and 25 μL min^{-1} for both the screening and affinity determination. The assay running buffer contained 10 mM phosphate buffered saline (PBS, P3813, Sigma-Aldrich) pH 7.4 with 0.05% TWEEN 20 (P9416, Sigma-Aldrich) and 5% dimethyl sulfoxide (DMSO, Merck) and was vacuum filtered and degassed for 15 minutes. bCAII was immobilized by amine coupling on a High Capacity Sensor (Bruker Daltonics, Hamburg, Germany). The surface was activated with an injection of 200 mM N-ethyl-3-(3-dimethylaminopropyl)-carbodiimide (EDC) and 50 mM N-hydroxy-succinimide (NHS)) for 8 minutes. bCAII was than immobilized at 20 μg mL^{-1} (according to UV-VIS spectrometric determination, BioSPEC-nano, Shimadzu Scientific Instruments) with four minutes contact time to reach a final immobilization level of approximately 4000 RU (screening) and 3000 RU (kinetic analysis). The surface was inactivated with 1.0 M Ethanolamine-HCl pH 8.5 (Bruker Daltonics, Hamburg, Germany) for 8 minutes.

The synthesized designs were evaluated *in vitro* for their biological activity on carbonic anhydrase II. All compounds were dissolved in DMSO, diluted and measured at three concentrations (10, 50 and 100 μM) with a final DMSO-concentration of 5%. All screening experiments were conducted as duplicates. The functionality of the immobilized bovine CA II was assessed by determining the dissociation constant of the positive control 4-sulfamoylbenzoic acid (CBS).

Hits from the screening were further analyzed for their binding affinities and kinetic constants. A twelve point dilution series of each compound in a range from 500 μM to 2 nM was prepared and injected in technical triplicates at 50 μl with 240 seconds of dissociation time. The effect of the DMSO-content on the signal was controlled by five DMSO controls at 4.6%, 4.8%, 5%, 5.2% and 5.4% DMSO at the beginning and the end of the measurement. CBS served as positive control and was measured in the same concentration range in technical quadruplicates.

4.2.6 Compound Database

The compound database used for this study was comprised of all compounds available within the in-house laboratory chemical listing at the time at which the study was conducted (February 2019). This gave a set of 574 molecules. With the exception of the 4-sulfamoylbenzoyl chloride, no compounds were ordered for the specific purpose of this study. All compounds were represented as canonical SMILES with salts and minor components removed. All molecular structures were provided to the DINGOS algorithm in RdKit (version 2017.03.3) using the MolToSmiles() function.

4.2.7 Flow Reactions

The reactions provided to the algorithm consist of a set of 19 reactions used in the first design cycle (Table 4.1). For the second design cycle, the reaction set was modified in order to improve the synthetic feasibility of the designs within the flow set-up. The reductive amination reaction was changed so that only aldehyde carbonyls were considered as valid reactant inputs. The sulfonamide formation reaction was split into three separate reactions, with each reaction taking either a primary amine, secondary amine, or ammonia as its input reactant. The Ugi and arylation of imidazole reactions were removed, as they were found not to perform well within the current flow set-up (Table 4.5). All reactions were written in the SMARTS format, with atom-mappings and reactive centers explicitly specified. All in-silico chemical transformations were performed in RDKIT using the RunReactants() function.

4.2.8 DINGOS Algorithm

The DINGOS algorithm used in this study was the same as that used in the initial DINGOS publication [82]. The building block molecular weight range was set to be between 0-250 g mol^{-1}, while the product molecular weight limit was set to 600 g mol^{-1}. The number of building blocks recommended within a single assembly step was set to 20 and the number of starting molecules, and hence the number of products produced, was set to the size of the in-lab compound database. For the design objective, the MACCSkeys were used as the descriptor representation and the Hamming distance was used in order to evaluate similarity. These were chosen in order to be consistent with the implementation of DINGOS considered in in the previous chapter (see Section 3.2.5). The compound acetazolamide was used as the template ligand and the reaction step limit was set to one reaction step.

4.3 Results and Discussion

4.3.1 Reaction and Building Block Set

The first step towards integrating DINGOS within the continuous flow system was establishing the scope of possible chemical reactions within the current setup. A list of 19 reactions were compiled that were considered feasible within the developed system (Table 4.1). These were based on reactions that had been successfully performed on other continuous flow systems [120]. This reaction set formed the reaction database used by DINGOS, forcing all designs produced to be within the scope of these reactions, and hence, in principle synthetically tractable within the flow system. One feature of the DINGOS method is that it uses custom sets of molecules for all compound generation. By selecting only available sets of molecules, this ensures

that all the required components exist and will be readily available. One particular case of interest is in utilizing in-house sets of building blocks. This allows for the development cycle to reduce the time and monetary investment of ordering new compounds, allowing for synthesis to be performed at a lower cost, and lowering the economic threshold for drug development. In the context of continuous flow, our goal was to produce large sets of molecules, while reducing the overall time and cost of production. Considering this, we chose to restrict our building block set solely to the molecules available within our own laboratory.

TABLE 4.1: Reactions considered within the continuous flow system. These reactions were compiled based on a literature review of reactions currently established on continuous flow [120]. Some reactions, such as those of Pictet-Spengler and the Ugi reactions were separated based on the size of the reaction relevant ring structure. FGI = Functional Group Interconversions. The corresponding reaction SMARTS can be found in Appendix A.3.1

Reaction Name
Pictet-Spengler-6-membered-ring
Pictet-Spengler-5-membered-ring
Aminothiazol formation
Paal-Knorr-pyrole formation
Triaryl-imidazol-1 2-diketone
Triaryl-imidazol-alpha hydroxy ketone
Fischer indole
Ester formation Acid Chloride
Thioester formation Acid Chloride
Reductive amination-Primary amine-Ketone
Amide formation Acid Chloride
Sulfonamide formation Sulfonyl Chloride
Aryl-Imidazole formation
FGI Acyl chloride
FGI sulfonyl chloride
Ugi-6-ring-aliphatic
Ugi-6-ring-aromatic
Ugi-5-ring-aliphatic
Hantzsch

4.3.2 The Drug Design Problem

Having established the reaction and building block sets, we now focused on selecting a model drug design problem. To this end, carbonic anhydrase II was chosen as the target system, with acetozolamide as the template ligand. This selection was made for two main reasons: Firstly, within the continuous flow setup used in this study, automated purification was currently not available. In a fully automated system, we ideally would incorporate an automated purification and testing component; however, such components are encumbered with their own set of technical difficulties. The system considered in this work did not contain such components, and hence, all purification and testing was performed manually. This meant that for multi-step reactions, intermediate products would require manual purification between reaction steps. As this would not adhere to fully automated synthesis, we chose to restrict the *de novo* designs to only a single reaction step. Acetozolamide is a relatively small drug

compound (222 Da), which would enable for effective *de novo* design within a single reaction step. The second reason is that biotesting of carbonic anhydrase is fairly robust and straightforward [125]; considering the large number of *de novo* designs we intended to test, selecting a system such as this reduces the overall experimental cost.

4.3.3 Run Parameters

For this experiment, the same version of DINGOS as in Chapter 3 was used. The MACCS keys were used as the descriptor function for the *de novo* designs, with the building blocks being recommended by the same AI model established in Section 3.1.3. A molecular weight limit of 250 g mol^{-1} was imposed on the building blocks, and 400 g mol^{-1} was used for the final products. This was done in order to make the overall designs more in line with the molecular weight of the template ligand. Along with *de novo* designs, the DINGOS output also included returned starting molecules. After removing all returned starting molecules, we obtained a set of 123 *de novo* designs.

4.3.4 First Round *De Novo* Design

DINGOS produced 123 *de novo* designs. Considering the small size of the compound database used, we thought it pertinent to investigate the overall novelty of the compounds produced, and to see whether or not the designs already existed within established synthetic and/or screening libraries. To determine this, we made use of the ZINC database (ZINC15, 2015) [126]. ZINC is an online docking database developed by John Irwin and the Shoichet lab. It contains a library of over 230 million molecules, comprised of 393 distinct chemical catalogs. These include both pharmaceutical databases, such as ChEMBL20 [81], as well as synthetic and commercial catalogs, such as the Enamine [127] and the Sigma Aldrich catalogs. ZINC provides the option to screen sets of compounds based on their structure. When querying compounds within the database, the compounds are reported as either immediately commercially available, or via synthesis on-demand. In addition to this, ZINC provides publicly-reported activities or biotests which have been performed on the report compounds. Of the 123 molecules suggested, 68 were found to be within the ZINC database. Two of these compounds had reported activities, both having been tested in a cell-based displacement assay against the Trace amine-associated receptor 1 [128], rather than the carbonic anhydrase target annotated for our template structure. None of the reported compounds had been tested before against carbonic anhydrase. Despite the limited size of the compound database used, DINGOS was capable of producing a large number of previously unreported *de novo* designs, none of which had been considered in the context of our biological target.

4.3.5 First Round of Automated Synthesis

All syntheses were performed by Benedikt Winkler and Berend Huisman.

The set of *de novo* designs was sorted according to its distance to the acetazolamide template, and the top 40 compounds were selected for synthesis (Figure 4.2). Of the 40 top-ranked designs, 22 were formed by amide bond formation (Figure 4.3), 16 by reductive amination (Figure 4.4), and two by arylation of an imidazole (Figure 4.5). Having obtained the *de novo* set, we then attempted to synthesize the

compounds with our continuous flow system using the reaction pathways prescribed by DINGOS.

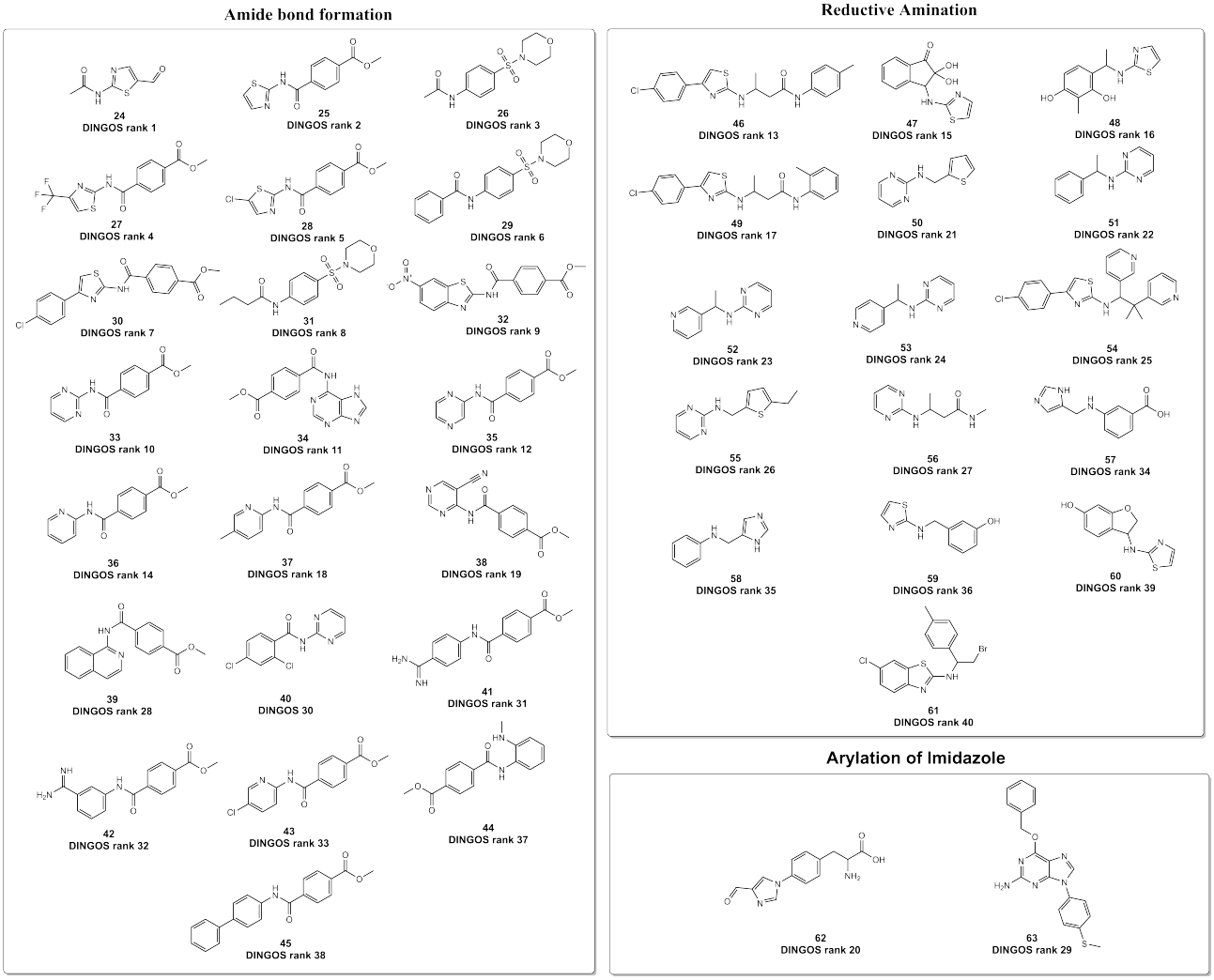


FIGURE 4.2: Overview of the 40 DINGOS designs selected for synthesis. Of the 40 proposed compounds, 22 were formed by amide bond formation, 16 by reductive amination, and two by imidazole arylation. Compounds are segregated by their reaction type, and sorted by their distance rank.

4.3.6 Amide Bond Formation

One challenge presented by continuous flow is that all species must remain in solution in order for the reaction to commence. Any solid material will accumulate throughout the system and likely lead to clogging within the system. This restricted our choice of reaction conditions, as the reaction could only be performed in a solvent-system that fully dissolved both building blocks.

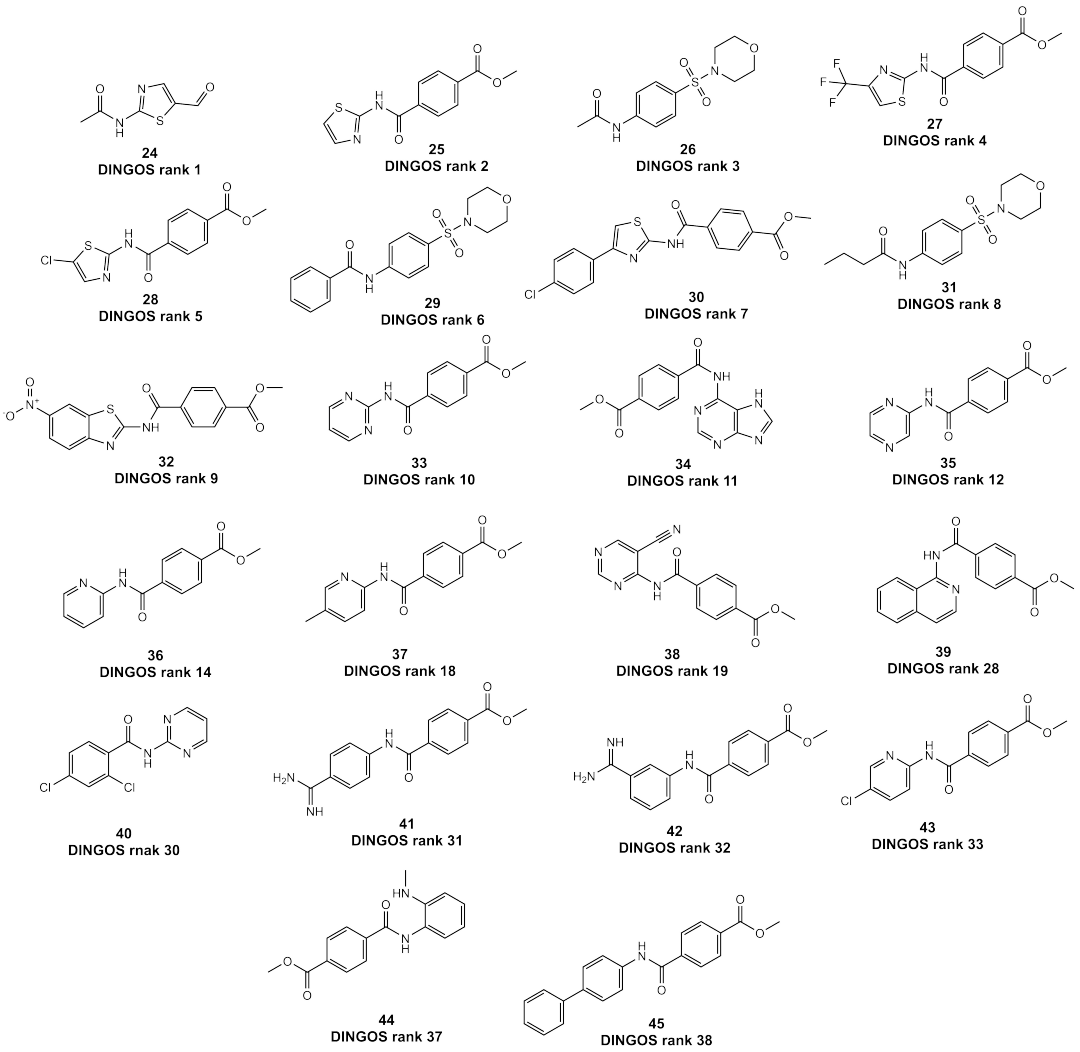


FIGURE 4.3: Overview of the 22 proposed DINGOS designs formed by amide bond formation. Compounds are sorted by their distance rank.

As a preliminary step in the synthesis of the 22 amide bond products (Figure 4.3), a series of trial experiments were performed in order to determine the appropriate solvent conditions in which to dissolve the building blocks. N,N dimethylformaldehyde (DMF), tetrahydrofuran (THF), and acetonitrile (MeCN) were all considered. Table 4.2 summarizes the resulting solubilities of the proposed designs.

TABLE 4.2: Resultant solubilities of the 22 proposed amide bond DINGOS designs. A design is specified as insoluble, if one or more of the building blocks could not be successfully dissolved in the solvent system considered. DMF=N,N dimethylformaldehyde, THF=tetrahydrofuran, MeCN= acetonitrile

	DMF	DMF/MeCN	THF	THF/MeCN	THF/MeCN/DMF
24	Insoluble	Insoluble	Insoluble	Insoluble	Insoluble
25	**Soluble**	**Soluble**	Insoluble	Insoluble	**Soluble**
26	**Soluble**	-	Insoluble	Insoluble	**Soluble**
27	-	-	Insoluble	Insoluble	**Soluble**
28	Insoluble	Insoluble	Insoluble	Insoluble	Insoluble
29	**Soluble**	-	Insoluble	Insoluble	**Soluble**
30	-	**Soluble**	Insoluble	Insoluble	-
31	-	-	Insoluble	Insoluble	**Soluble**
32	-	-	Insoluble	Insoluble	**Soluble**
33	**Soluble**	-	**Soluble**	**Soluble**	-
34	Insoluble	Insoluble	Insoluble	Insoluble	Insoluble
35	-	-	**Soluble**	**Soluble**	-
36	-	-	**Soluble**	**Soluble**	-
37	-	-	-	**Soluble**	-
38	Insoluble	Insoluble	Insoluble	Insoluble	Insoluble
39	-	-	-	**Soluble**	-
40	**Soluble**	-	-	**Soluble**	-
41	Insoluble	Insoluble	Insoluble	Insoluble	Insoluble
42	Insoluble	Insoluble	Insoluble	Insoluble	Insoluble
43	-	-	-	**Soluble**	-
44	-	-	-	**Soluble**	-
45	-	-	-	**Soluble**	-

As can be seen from Table 4.2, there were no consistent solvent conditions which worked for all of the designs. In total, six of the 22 amide products were removed due to the insolubility of the corresponding building blocks, resulting in a set of 16 viable designs (Figure 4.4). It was found that the THF/MeCN solvent mixture was capable of dissolving the aniline/pyridine based building blocks, while the THF/MeCN/DMF solvent mixture was capable of dissolving the tertiary sulfonamide and 2-aminothiazole building blocks. Having established the working solvent conditions, all 16 of the reactions were performed in continuous flow with a residence time of 5 minutes at a temperature of 55 °C. Of the nine aniline/pyridine based products performed in THF/MeCN, seven were successfully synthesized and purified. The remaining sulfonamide and 2-aminothiazole based products were reacted in THF/MeCN/DMF, and, of these seven products, five products were successfully synthesized, with two of the 2-aminothiazole based products failing to yield the desired product. Table 4.3 summarizes these results.

TABLE 4.3: Resultant reactions of the 22 proposed amide DINGOS designs. The reactions were performed in four separate solvent conditions: DMF, DMF/MeCN, THF, THF/MeCN, and THF/MeCN/DMF. A design was considered successful if the formation of the desired compound could be confirmed. DMF=N,N dimethylformaldehyde, THF=tetrahydrofuran, MeCN= acetonitrile

	DMF	DMF/MeCN	THF	THF/MeCN	THF/MeCN/DMF
24	-	-	-	-	-
25	**Successful**	**Successful**	-	-	**Successful**
26	Failed	-	-	-	**Successful**
27	-	-	-	-	**Successful**
28	-	-	-	-	-
29	Failed	-	-	-	**Successful**
30	-	Failed	-	-	-
31	-	-	-	-	**Successful**
32	-	-	-	-	Failed
33	Failed	-	**Successful**	**Successful**	-
34	-	-	-	-	-
35	-	-	**Successful**	**Successful**	-
36	-	-	**Successful**	**Successful**	-
37	-	-	-	**Successful**	-
38	-	-	-	-	-
39	-	-	-	Failed	-
40	Failed	-	-	**Successful**	-
41	-	-	-	-	-
42	-	-	-	-	-
43	-	-	-	**Successful**	-
44	-	-	-	Failed	-
45	-	-	-	**Successful**	-

4.3.7 Reductive Amination

Reductive amination in continuous flow has been performed previously by multiple groups [129–131]. These methods vary from those proposed by Zhang *et al.* [132], in which Leuckart-Wallach conditions (formic acid in the presence of an amine) were used, to those employed by Seeberger *et al.* [120] in which a solid-phase heterogeneous column composed of Sodium Borohydrate and Lithium Chloride was used to catalyze the reaction. In the work by Cronin *et al.* [133] they made use of a two-step continuous flow system. In this setup, the formation of the desired amine occurred in two steps. In the first step, the intermediate imine was formed, and then subsequently reduced to give the desired amine product. These three procedures were investigated for their use in the formation of the DINGOS amine products. All building blocks were successfully dissolved in methanol, and the three reaction procedures were attempted. Reactions were performed with a residence time of 20 minutes and at a temperature of 57 °C. Table 4.4 summarizes the findings of the various reaction conditions.

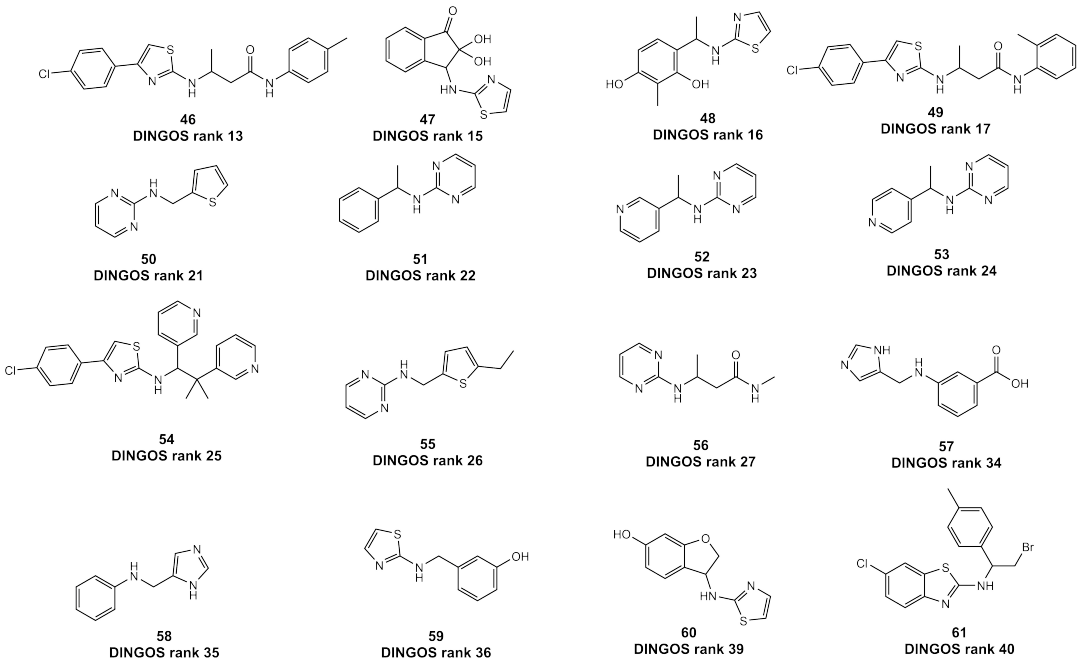


FIGURE 4.4: Overview of the 16 proposed DINGOS designs formed by reductive amination. Compounds are sorted by their distance rank.

TABLE 4.4: Resultant reactions of the 16 proposed reductive amination DINGOS designs. A design was considered successful if the formation of the desired compound could be confirmed.

	One-step No Catalyst	One-step Heterogeneous Catalyst	Two-step MeOH
46	-	-	Failed
47	-	-	Failed
48	-	-	Failed
49	-	-	Failed
50	Failed	Failed	**Successful**
51	-	-	Failed
52	-	-	Failed
53	-	-	Failed
54	-	-	Failed
55	Failed	Failed	**Successful**
56	-	-	Failed
57	Failed	Failed	Failed
58	Failed	Failed	**Successful**
59	Failed	Failed	Failed
60	-	-	Failed
61	-	-	Failed

As can be seen, of the considered conditions, only the two-step system using MeOH yielded successful results. With this system, we were successfully able to produce the desired product for three of the five aldehyde based building blocks. The method was, however, unfortunately shown to be unsuccessful for all of the ketone

based building blocks. In the work outlined in Zhang *et al.* and Cronin *et al.*, all experiments only considered reductive amination through aldehyde based reactants, and in the publication by Seeberger, attempts to perform reductive amination with ketone based building blocks gave no product. Taking this into consideration, we decided to discard these designs. In total, of the 16 considered products, only three were successfully synthesized and purified.

4.3.8 Imidazole Arylation

Two of the proposed designs were suggested to be formed by imidazole arylation (Figure 4.5). Previous work by Van Der Eycken *et al.* [134] had shown successful N-arylation in continuous flow, however, they did not consider imidazoles in their experiments.

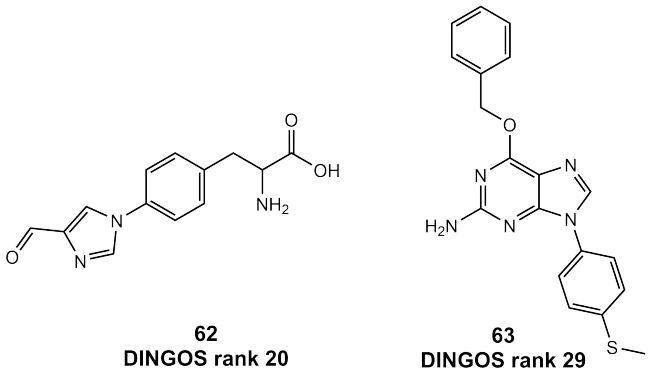


FIGURE 4.5: Overview of the two proposed DINGOS designs formed by imdiazole arylation.

In the outlined procedure, they successfully performed copper catalyzed N-arylation in both dichloromethane (DCM) and DMF. As a preliminary step, the solubility of the building blocks in DMF was established. Unfortunately, only one of the product's building blocks, that of **63**, was found to be soluble. Because of this, we discarded **62**, and focused soley on the remaining design. In the procedure outlined by Van Der Eycken, 0.15 M of the copper(II)acetate catalyst was used. We were, however, unsuccessful at dissolving the required amount of copper catalyst in DMF. Multiple attempts were made to form the required catalyst solution, however ultimately, the catalyst could only be dissolved by diluting it by a factor of 10. To match equivalence, the synthetic procedure was performed at the reduced concentration, however, no product was observed. Hence, we were unable to synthesize either of the two Ar-imidazole designs.

4.3.9 Bioactivity of First-Round Designs

All biotesting was performed by Cyrill Brunner. The prepublished results presented here were directly communicated by Cyrill Brunner.

Of the set of 40 *de novo* designs, we were successfully able to synthesize 15 of the compounds within the continuous flow system. These experiments were performed over a time span of four weeks. From this set of 15 compounds, we performed biotesting in order to determine their binding affinity towards carbonic anhydrase II. The bovine carbonic anhydrase II protein (bCAII) was used as the target protein, with

the biotesting experiments being performed using Surface Plasmon Resonance (SPR). 4-sulfamoylbenzoic acid (CBS) was used as our positive control. A compound was considered as active if it showed an SPR response above 50% of the CBS reference at 50 μM. The experiment was repeated at concentrations of 10, 50, and 100 μM. Of the 15 designs, no response above 50% CBS was detected, indicating that the compounds produced did not bind to the target protein.

4.3.10 Refinement of Designs Using Information Gained from the First Cycle

Having completed our first cycle of the active learning procedure, we then sought to improve upon the results obtained. Our first cycle presented issues with both the synthesis and binding affinity of the compounds produced. Using the information gained, both synthetically and biologically, we updated our procedure in order to improve the results.

4.3.11 Improving Binding Affinity by Inclusion of a Potential Terminal Sulfonamide

The tested DINGOS designs did not show significant binding towards carbonic anhydrase. For the *de novo* design, the template ligand used by DINGOS was the 20 nM carbonic annydrase inhibitor acetozoloamide. It is known that the binding mode of acetozoloamide occurs primarily through a coordination of the terminal sulfonamide to the zinc(II) cation of the carbonic anhydrase protein [135]. Multiple studies have confirmed that this interaction is indeed important for carbonic anhydrase II binding. Of the 40 DINGOS designs, none of the produced compounds possesses this key terminal sulfonamide group. Upon examination of the in-house building block set used as the compound database, it was discovered that given the reagents and reactions provided, formation of terminal sulfonamide possessing designs would not have been possible. This was because no sulfonamide containing building blocks capable of reacting were present within this set, nor were any terminal sulfonamide forming reactions in our reaction database. The set did contain, however, sulfonyl chlorides, which were responsible for the formation of the three secondary sulfonamides produced by DINGOS. In comparing the structures of the *de novo* designs, it was observed that 31 of the 123 designs contained terminal methyl ester moieties, formed by the selection of the methyl 4-(chlorocarbonyl)benzoate acid chloride building block. It was hypothesized that this group was selected by DINGOS as a structural replacement for the sulfonamide group found in acetozoloamide. To this end, we introduced the compound 4-sulfamoylbenzoyl chloride into our compound database. This was done in order to, firstly, see whether or not DINGOS would preferentially select this compound for inclusion within the *de novo* designs, and secondly, whether this inclusion would result in an improvement in binding affinity for the bCAII target.

4.3.12 Improving Synthetic Feasibility by Updating the Reaction to Incorporate Constrains of the Flow

Of the 40 designs that we intended to synthesize, we were only able to successfully synthesize 15. This was primarily due to issues with solubility of the building blocks and discrepancies between the reaction definitions in conventional organic chemistry, and those that are feasible within a continuous flow system. A significant proportion of the designs, 12 of the 40 proposed, were rejected, due to the infeasibility of the reductive amination reaction. This was due to the presence of ketone based building

blocks, for which the established amination reactions were not applicable. Additionally, we were unable to perform either of the proposed imidazole arylations within our system. While the issues observed could have potentially been resolved with system modifications, we decided instead to modify the reaction logic within DINGOS' reaction database. Modifications of this nature kept more inline with the notion of rapid compound generation. These changes were facilitated by the modular nature of the DINGOS algorithm. To resolve the observed issues, we modified the reductive amination reaction to exclude all ketone carbonyl building blocks, and removed the imidazole arylation reaction entirely. A review of the reaction outputs produced by DINGOS also revealed that the Ugi reaction was not possible with the building blocks provided, and hence, this reaction was subsequentially removed. As we intended in this round to bias our results towards sulfonamides, we decided to split the more generalized sulfonamide formation reaction into three variants. One that accepted sulfonic acids, one that accepted sulfonyl chlorides, and one that would allow for the conversion of sulfonyl chlorides to a terminal sulfonyl amide through the use of ammonia. Table 4.5 shows the updated reaction list.

TABLE 4.5: Updated reaction set used for the second active learning cycle. The reductive amination reaction was modified to neglect ketone carbonyls, while the sulfonamide formation reaction was split into three separate reactions: one for sulfonic acid, one for sulfonyl chlorides, and one that converts sulfonyl chloride to a sulfonamide via ammonia. The imidazole arylation and Ugi reactions were removed. FGI=Functional Group Interconversions. The corresponding reaction SMARTS can be found in Appendix A.3.2

Reaction Name
Pictet-Spengler-6-membered-ring
Pictet-Spengler-5-membered-ring
Aminothiazol formation
Paal-Knorr-pyrole formation
Triaryl-imidazol-1 2-diketone
Triaryl-imidazol-alpha hydroxy ketone
Fischer indole
Ester formation Acid Chloride
Thioester formation Acid Chloride
Reductive amination-Aldehyde
Amide formation Acid Chloride
Sulfonamide formation Sulfonyl Chloride Secondary amine
Sulfonamide formation Sulfonyl Chloride Primary amine
Sulfonamide formation Sulfonyl Chloride Ammonia
FGI Acyl chloride
FGI sulfonyl chloride
Hantzsch
Ugi-5-ring-aliphatic
Hantzsch

A test run was performed in order to ensure that the modified reactions set still yielded meaningful designs. DINGOS was successfully able to generate 128 *de novo* designs, with the sulfonamide and reductive amination reactions still being represented in the *de novo* outputs. The compounds were again queried in the ZINC database. Of the 128 designs generated by DINGOS, 84 of the compounds were

found to be reported within the ZINC database, and of these reported compounds, it was found that two of them had reported bioactivities. In contrast to the previous design cycle, these compounds had in fact been tested against carbonic anhydrase. Both compounds were found to be inactive against CAII, with reported IC_{50} values of above 50 μM.

4.3.13 Second Round of Automated Synthesis

All syntheses were performed by Berend Huisman.

We now performed the second *de novo* design cycle, incorporating both the updated reaction set and modified compound database. Within this second cycle, DINGOS was successfully able to produce 109 *de novo* designs. Of the 109 designs, 67 of the product molecules were formed using the 4-sulfamoylbenzoyl chloride building block. Of the designs generated in the first cycle, three contained sulfonamides and 31 a methyl ester group. In contrast, the second round designs contained 77 sulfonamides and eight methyl esters, supporting the claim that DINGOS had been selecting the methyl moiety in order compensate for the lack of sulfonamide building blocks. As with the previous design cycle, we queried the ZINC database in order to determine the overall novelty of the *de novo* designs. Of the 109 designed compounds, 68 were reported within the ZINC database. Of these 68 compounds, two had been tested, and in fact, had been shown to be active against carbonic anhydrase II.

As with the previous cycle, the top 40 designs were sorted according to their distance scores. A comparison was made with the compounds from the first design cycle. This can be seen in Figure 4.6. As can be seen, both updating the reactions and introducing the 4-sulfamoylbenzoyl building block lead to an improvement in the designs' distance values.

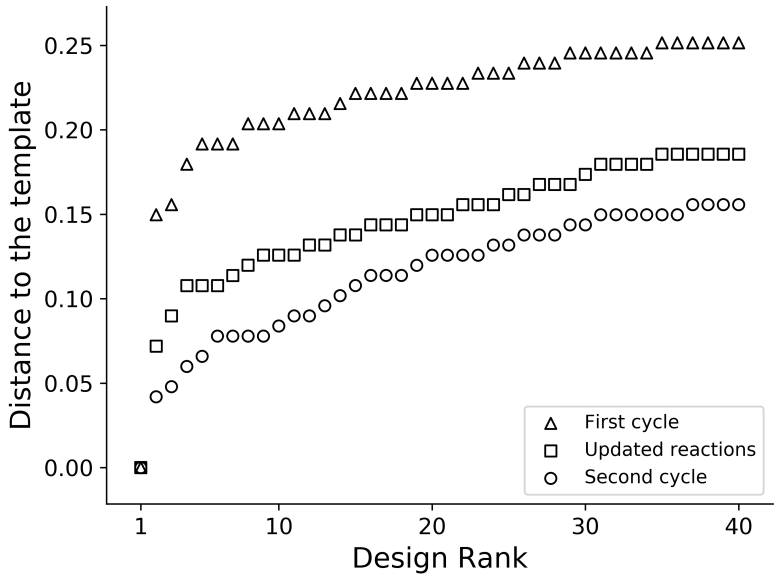


FIGURE 4.6: Comparative plot of the distance values obtained for the top 40 ranked molecules for the first and second active learning drug design cycle, as well as with the updated reaction set. Distance was measures as the Hamming distance of the MACCS keys descriptors.

The top 40 ranked designs from the first design cycle had a median distance value of 0.25, while the median distance of those from the second cycle was 0.12, a reduction of approximately one half. Of the 40 top ranked molecules, we selected eleven of the designs for synthesis. For all eleven of these compounds, DINGOS proposed synthesis by amide bond formation. All building blocks were successfully dissolved in MeCN/THF. The same reaction conditions as used in the previous cycle were used here. Of the eleven proposed designs, seven were successfully synthesized and purified (Figure 4.7).

4.3.14 Bioactivity of Second-Round Designs

All biotesting was performed by Cyrill Brunner. The prepublished results presented here were directly communicated by Cyrill Brunner.

The seven compounds that were successfully synthesized in the second *de novo* design cycle were then tested against bCAII. The same experimental conditions as those in the previous design cycle were used. Of the seven compounds tested, five were found to show activity below 50 μM. Of these five actives, three showed activity below 10 μM (Figure 4.7).

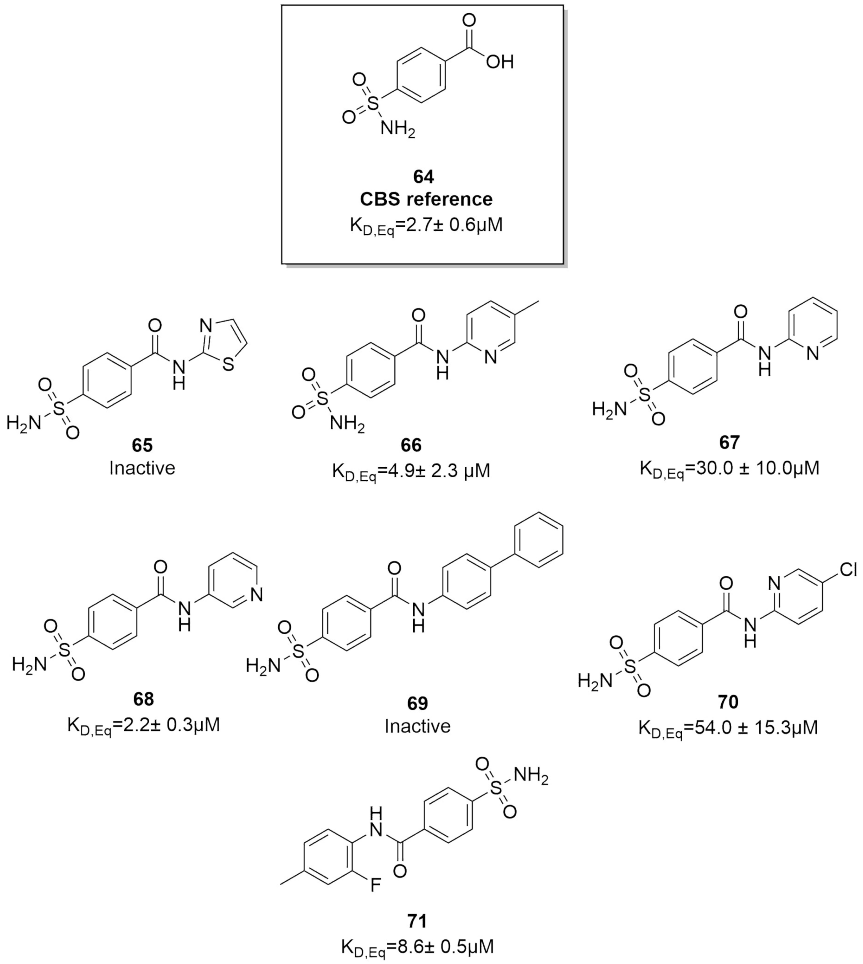


FIGURE 4.7: Overview of the seven compounds synthesized in the second active learning cycle. All seven compounds were tested against bCAII in a SPR binding assay, with CBS as a reference. Of the seven designs, **66**, **67**, **68**, **70**, **71** showed binding affinity towards the target. Compounds **66**, **68**, and **71** possessed a comparable affinity to that of the CBS reference.

The introduction of the terminal sulfonamide group did in fact result in an increased binding affinity towards bCAII. Despite this, we did observe sulfonamide containing structures that showed no binding towards bCAII. Figure 4.7 shows the structure of the seven synthesized compounds. In comparing their structure, one immediately striking result is the discrepancy between compound **68** and compound **67**. Compounds **68** and **67** differed only in the position of the pyridine nitrogen (meta in **68** and ortho in **67**). Despite their similar structures, **68**'s $K_{D,Eq.}$ was 10 times lower than that of **67**. A similar discrepancy was observed in **66**, which only differed from **67** in the presence of a single methyl group at the para position of the pyridine ring, but similarly showed a 10-fold increase in binding affinity. NMR spectra of the five bioactive compounds are shown in Appendix A.3.3.

4.4 Conclusion

In summary, in this work we established an active learning *de novo* design cycle. Compounds were generated by DINGOS, transferred to our continuous flow system

for synthesis, and tested for their binding affinity via SPR measurements. Within this cycle, DINGOS was successfully integrated within our custom continuous flow system. Two full *de novo* design cycles were performed, producing 15 and seven *de novo* designs as prescribed by the DINGOS algorithm. The first cycle highlighted some of the key issues in adapting a method such as this to automated synthesis. In contrast to conventional synthetic chemistry, continuous flow places stricter limitations on the choice of solvent systems and reagents. This was a major factor of consideration in synthesizing the amide based *de novo* designs. In order to successfully transfer the reactions into continuous flow, a series of preliminary solubility experiments were first required in order to determine the appropriate solvent conditions. Further limitations relating to the scope of applicable reactants in flow resulted in the exclusion of 12 of the 16 amine designs, and all of the Ar-imidazole ones. Biotesting revealed that of the compounds synthesized within the first cycle, none of the designs showed binding affinity towards the bCAII target. It was observed that the designs lacked a terminal sulfonamide, which has been associated with CAII binding. Between the two cycles, modifications were made to both the reaction set and the compound database in order to improve upon the observed activity and synthesizability. The lack of binding affinity observed in the first cycle motivated the inclusion of the 4-sulfamoylbenzoyl chloride building block into the compound database, in order to bias the *de novo* designs to be more in line with the conventional mode of CAII binding. The reaction database was modified in order to exclude failed reaction conditions observed in the first design cycle. These modification resulted in an increase in the proportion of the successful syntheses, as well as the overall binding affinity of the generated designs. The relative ease with which the modifications were performed highlight the advantages of DINGOS' modular design in such an application. Ultimately, we were able to produce five novel, active compounds, with three of them showing a comparable binding affinity to our positive control (CBS). Two of the *de novo* designs, **68** and **67**, showed a 10-fold difference in binding affinity, despite only differing structurally by a single atom. The presence of activity clefts such as these highlights the importance of locally exploring new structures. DINGOS was shown capable of being integrated within such an active learning procedure; however, as with the previous study, we were still restricted to a fairly constrained design hypothesis, that is, the Hamming distance of the MACCS keys. Despite the success seen within this work, this is still a major limitation of the DINGOS algorithm. In the following chapter, we investigate a potential solution to this problem, and a way of generalizing DINGOS towards a broader range of design hypotheses.

Chapter 5

Generalization of DINGOS Towards Arbitrary Descriptor Spaces through Generative Machine Translation

A core component of the DINGOS algorithm is the building block recommendation system. This component selects the most appropriate building block to combine with a given starting molecule in order to generate a de novo design that is similar to a provided template ligand. Building blocks are selected by first predicting the molecule's descriptor value, and then using this value to query the compound database. While this method did lead to the generation of similar designs in Chapter 3 and 4, it presented difficulties associated with changing the descriptor representation. In order to change the descriptor, an entirely new predictive model would be required, limiting the flexibility of the DINGOS algorithm. In this chapter, we investigate the use of generative machine translation as a potential solution. A model which generates the building blocks molecules directly was developed, and the capabilities of this newly generalized DINGOS method were explored. DINGOS was used to generate de novo designs for eleven separate sets of bioactive molecules extracted from the ChEMBL database. For these experiments, we defined our domain of interest, which were all distances below the most similar inactive compound. The goal was to produce compounds within this region. DINGOS was shown capable of generating designs within the 'domain of interest' for each of the eleven template ligands. In the case of the H3 histamine receptor, DINGOS was able to generate compounds that were more similar to the respective template ligand than any reported bioactive in ChEMBL.

5.1 Introduction

5.1.1 Descriptor Agnostic Building Block Recommendation

The version of DINGOS presented in Chapter 3 performed building block recommendations by first generating a molecular fingerprint and then finding the molecules from the compound database with the most similar fingerprints. This fingerprint is generated by a multilayer-perceptron model, which was trained on a set of existing, patented chemical reactions (USPTO database). For the training procedure, the molecules were converted into the MACCS keys binary representation. Because of this, the trained model can only select appropriate building blocks by a comparison of their MACCS keys. Ideally, one would want to be able to change representation without the need for entirely new predictive models, however, this requires a model

which is independent of the choice of representation.
Recent work adapting models for text generation have shown promise for generating molecular structures [84, 136]. All of these studies make use of the SMILES format, in which the chemical structure of a molecule is represented as a sequence of characters. Of particular note was the work of Schwaller *et al.* [137]. Schwaller and co-authors trained a model to predict the most probable products from a given set of reactant molecules. Based on this work, we sought to develop a predictive model to perform the reverse process, that is, generate a building block from a pair of starting and product molecules. As the model would be generating molecular structures, rather than descriptors values, the AI model would be completely independent of our choice of descriptor representation.

5.1.2 Transformer Model

The model by Schwaller *et al.* was adapted from the transformer architecture [138] developed by Google brain in 2017 for machine translation. In machine translation, a model is trained in order to convert text from one target language, say English, to another, say French. One can say that the French sentences are generated by the model using the English sentences as an input. Most machine translation models consists of two main components, an encoder, which converts the input text into a latent representation, and a decoder, which converts this latent representation into the output text [139]. The transformer architecture is an encoder-decoder that makes use of an attention mechanism in order to perform text generation. In machine translation, attention mechanisms are a method for determining which elements of a sequence the model should pay attention to during translation. The attention mechanism converts a sentence into a series of attention vectors, one for each word in the sentence. These vectors encode the importance of each word for the given translation task relative to each of the other words in the sentence. An example of this can be seen in the work by Bahdanau *et al.* [140], in which they combined an attention mechanism with a recurrent neural network (RNN) based encoder-decoder to perform English-French translations. As can be seen from Figure 5.1, each of the individual words plays a different level of importance towards the overall translation.

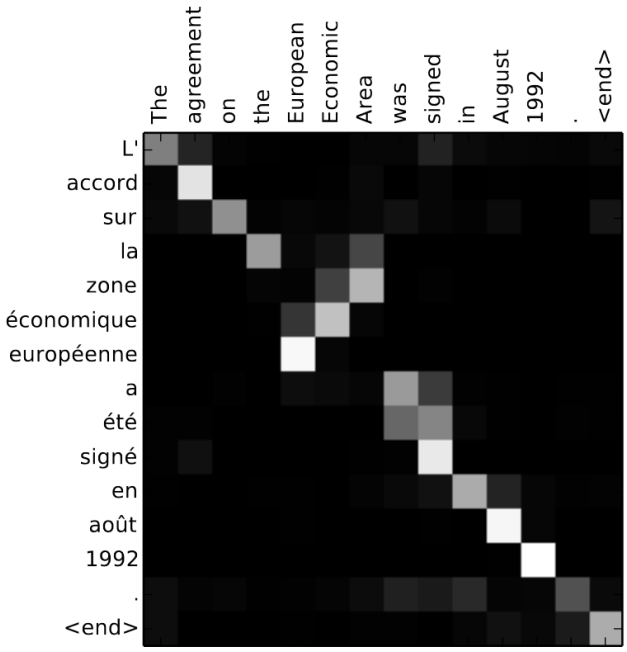


FIGURE 5.1: Attention map taken from Bahdanau *et al.* [140]. The map shows the results of the machine translation model RNNsearch. Along the horizontal axis is the English input sentence, while along the vertical axis is the generated French translation. The brightness of the boxes in the heat map indicate the value of the attention weights for the English-French word pairs, with white representing a value of 1 and black a value of 0. The higher the attention weight the more importance that was placed on the given word.

In the transformer, input and output sentences are each converted into separate attention vectors which are then combined to produce a series of encoder-decoder attention vectors. These encoder-decoder attention vectors then serve as the inputs for a feed-forward network component. The ultimate output of the transformer is a vector of output probabilities, with each element giving the probability of selecting a particular word for the translation (see Figure 5.2).

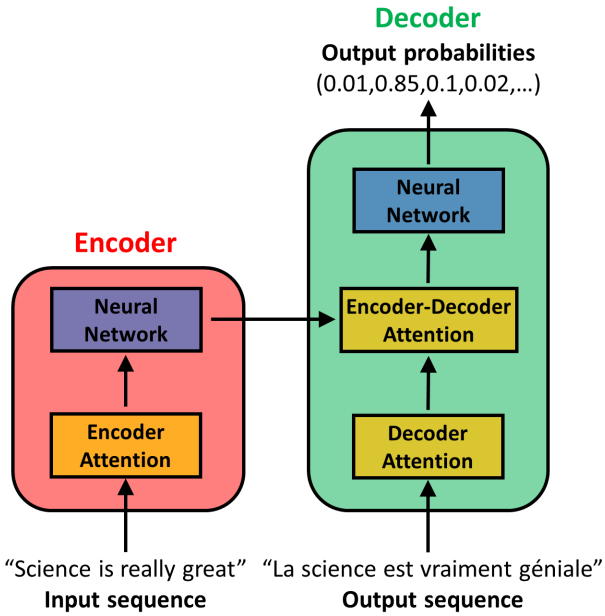


FIGURE 5.2: Schematic depiction of the transformer model adapted from Vaswani *et al.* [138]. The encoder, depicted in red, is comprised of an attention layer followed by a feed-forward neural network. The encoder converts the English sentence into its corresponding encoder-attention vectors, which then serve as inputs for the feed-forward neural network. The neural network then generates the latent representation which is then accepted by the decoder. The decoder is similarly comprised of an attention layer which accepts the translated sentence in French and converts it into the decoder-attention vectors. This initial decoder attention layer is followed by an additional encoder-decoder attention layer which accepts the latent representation from the encoder and the decoder-attention vectors as inputs. A final feed-forward neural network layer converts the output of the encode-decoder attention layer to a vector of output probabilities, with each entry representing the probability of selecting a specific word in French.

5.2 Methods

5.2.1 Descriptor Set

A set of eight descriptors were used for this study. This set was comprised of the MACCS keys, extended connectivity fingerprint (ECFP), featMorgan, Atom pair, RDKit, Avalon, Torsional, and Layered descriptors using the functions available within RDKit. The descriptors are summarized in Table 5.1.

TABLE 5.1: Summary of the eight molecular descriptors used for the bioactivity analysis.

Descriptor Name	Description	References
MACCS keys	A 1-dimensional structural representation consisting of a set of individual molecular fragments or 'keys'. The molecule is queried against this set of keys and assigned an on-bit (1) if they are found within the molecule and an off-bit (0) if not.	Durant *et al.* (2001) [95]
ECFP	Iteratively expands each atom's neighbours in order to create a list of substructures. Each substructure is converted into an atomic invariant format using the Daylight atomic invariants.	Morgan [141], Rogers, Hahn [142]
featMorgan	Uses the same algorithm as the ECFP descriptor, however, instead of using the Daylight atomic invariants, it uses a predefined feature list to convert the substructure list. This list includes common pharmacophoric groups such as donor and acceptor.	Morgan [141], Rogers, Hahn [142]
Atom Pair	Measures the frequency of atom pairs within a molecule. Each atom is defined by its atom property (atomic type, number of π-electrons, and number of non-hydrogen neighbors) and by the shortest number of bond separating the pair.	Carhart *et al.* [143]
RDKit	Converts the molecule into a set of all linear molecular subgraphs. Atom-type, bond-type, atomic number and the aromaticity are used to calculate the descriptor value.	RDKit [47]
Avalon	Similar to the RDKit Daylight fingerprint, the Avalon fingerprint enumerates a predefined list of path and atom feature classes in order to convert the molecular graph into a property list.	Gedeck *et al.* [144]
Torsional	Converts the molecules into a set of four atom fragments (inspired by the torsional angle). The atom type, number of π-electrons, and non-hydrogen neighbors are used to describe the atoms within a fragment. The frequency of each fragment is used to calculate the descriptor value.	Nilakantan *et al.* [145]
Layered	Using the same algorithm as the RDKit descriptor, however, includes additional information, such as the number and size of rings within the molecule.	RDKit [47]

5.2.2 Training Data for the Generative Model

The datasets used were taken from the publication Schwaller *et al.* [137] [1]. The authors experimented with two datasets, one taken from Jin *et al.* [146], and a modified version in which the reaction set was doubled by randomizing the canonical representation of the SMILES. These two datasets were called the MIT and MIT-augm datasets, respectively. The datasets were preprocessed in order to remove salts, solvents, and common reagents. The remaining datasets were further filtered in order

[1] https://github.com/pschwllr/MolecularTransformer

to limit the reactions to at most two-component. The initial reaction SMARTS were of the form "reactant.reactant»product". In order to be used for building block generation, the training data first needed to be converted to the appropriate format, that is "starting molecule.product»building block". In converting to this format, we were presented with a choice. As the distinction between the starting molecule and the building block is arbitrary, a decision was required when converting the reaction SMARTS. To resolve this, we decided to generate two versions of each dataset; *a*) one labeled "-SINGLE", in which the most similar molecule to the product (as defined by Tanimoto distance of the extended connectivity fingerprint(ECFP)) was selected as the starting molecule, and *b*) one labeled "-DOUBLE", in which every two-component reaction was split into two separate reactions SMARTS, with each reactant serving as the starting molecule. This preprocessing procedure was applied to both the MIT and MIT-augm datasets, yielding a total of four datasets: MIT-SINGLE, MIT-augm-SINGLE, MIT-DOUBLE, MIT-augm-DOUBLE.

5.2.3 Compound Database for the DINGOS Algorithm

The Sigma Aldrich catalog was used as the compound database for this study. The catalog was extracted from the PubChem database (www.pubchem.ncbi.nlm.nih.gov, version 2019.01.23) [147] as canonical SMILES. All salts and minor components were removed. The molecular SMILES were converted to their corresponding RDKit mol objects with the MolToSmiles() function. Any entries that did not yield a valid mol object were removed. The resulting database contained 217,758 molecules.

5.2.4 Data for the Bioactivity Analysis

All data was extracted from ChEMBL(ChEMBL25, 2018) [81]. Each set consisted of reported compounds that had been tested against the following ChEMBL targets: Histamine H1, Histmaine H2, Histamine H3, Histamine H4, Dopamine D1, Dopamine D2, Dopamine D3, Dopamine D4, TTK dual-specificity protein kinase, Cannabinoid CB1, and Coagulation factor Xa. The 'Canonical SMILES' field was converted into the corresponding RDKit mol object. All structures that did not lead to a valid RDKit mol object were removed from the set. Descriptor values were calculated for each of the descriptors in the descriptor set (see Methods 5.2.1). Entries that failed to produce a valid descriptor were removed. All entries without a valid entry in the "STANDARD UNITS", "STANDARD TYPE" and/or "STANDARD VALUE" fields were removed, as without these entries a valid bioactivity value could not be discerned. Repeated SMILES were discarded if the relative standard deviation (see Appendix A.4.3) in their bioactivity value was above 0.25.

5.2.5 Bioactive Analysis

The distance analysis was performed with compounds with a reported IC_{50} or K_i value in nM units. For the analysis, an activity threshold of 1 μM was set, with all molecules with a reported activity below this threshold being considered as active. We define the "distance threshold" for each set as the minimum observed distance within the set of inactive compounds. This threshold was calculated for of the eight descriptors. Using the distance threshold, we defined our "domain of interest", which was the region encapsulating all compounds with a distance value less than that of the distance threshold. By definition all of these compounds were active. The distance values were calculated as the Tanimoto distance (1-Tanimoto coefficient). Receiver Operating Characteristic Curves (ROCs) were calculated by moving the distance

threshold along the x-axis and measuring the true and false positive rates. The area under curve (AUC) of each curve was measured.

5.2.6 *De Novo* Design Paramters

For the DINGOS algorithm, the number of starting molecules was set to 300, the number of reaction steps was set to a limit of four, and the building block recommendation pool was set to 20. A building block molecular weight limit of 400 g mol^{-1} and a product weight limit of 600 g mol^{-1} was used. The Tanimoto distance metric was used to evaluate the similarity of the designs to the template ligand. In the case-study, a single template ligand was selected for each target system. For each target, the template and descriptor that gave the highest number of compounds in the "domain of interest" was selected.

5.2.7 Reaction set

The reaction set used in this chapter was adapted from the set used in Chapter 3.1.2 with the following modifications. The Williamson ether reaction was separated into two separate reactions, one forming ether and another thioether products. The esterifciation and amide bond formation reactions were separated into ones that accepted carboxylic acid reactants and ones accepting acid chlorides. The sulfonamide forming reaction was separated into ones accepting sulfonic acid and ones accepting sulfonyl chloride. Reductive amination was separated between ketone and aldehyde reactants. Additionally, two reactions were added to the set, Nucleophilic substitution and the Grignard reaction. This yielded a set of 73 separate reactions (shown in Appendix A.4.1)

5.3 Results and Discussion

5.3.1 Building Block Generation

The model of Schwaller *et al.* was originally developed for machine translation [138]. In machine translation, a predictive model is used to convert a piece of text in one target language to another language. By exploiting the text based SMILES format, Schwaller *et al.* were able to train a model that translates the reactant text into that of the corresponding product. The associations between the inputs and outputs (reactants and products) are learned directly from the reaction data. This means that, in theory, it would be possible to produce a similar model that generates the building block molecule simply by interchanging the product and building block molecules within the training data. The datasets used in the original publication were extracted and converted to the format of our target problem (see Methods 5.2.2). This yielded four datasets: MIT-SINGLE, MIT-augm-SINGLE, MIT-DOUBLE, MIT-augm-DOUBLE. Four separate models were trained, one on each of the datasets, using the same hyper-parameters as in Schwaller *et al.* A testing set of 40000 previously unseen reaction SMARTS were used in order to quantify the predictive accuracy of the respective models. These models do not generate a single SMILES, but rather, predict a set of the most likely SMILES. In the original paper, the group measured the predictive accuracy up to the top five predicted SMILES, providing a broader measure of the models performance. Similarly, we measured the predictive accuracy of the top five predicted building blocks SMILES. A prediction was considered correct, if the model was capable of generating the building block

SMILES exactly. As can be seen from Table 5.2, for more than 85% of the testing data, all four of the models were successfully able to generate the correct SMILES as their top prediction. Of the models considered, the MIT-DOUBLE and the MIT-augm-DOUBLE models possessed the highest predictive accuracy amongst the top five predictions.

TABLE 5.2: Predictive accuracy (percentage of correct predictions) of generating the correct building block SMILES. A set of 40000 testing SMILES were used in order to evaluate the four models considered. The accuracy was measured among the top five predicted building block molecules with Top5 representing an accurate prediction within the top five predicted SMILES.

	Top1	Top2	Top3	Top4	Top5
MIT-augm-DOUBLE	86.4	92.6	94.3	95	95.2
MIT-augm-SINGLE	86.3	92.2	93.9	94.6	94.8
MIT-DOUBLE	86.7	92.5	94.2	94.9	95.2
MIT-SINGLE	85.9	91.6	93.3	94	94.3

In order to determine if the structural information was retained within the model, we performed a second test. Here, we took the top predicted building blocks from the previous analysis and combined it with the product molecule to determine if the model was capable of successfully reproducing the correct starting molecule.

TABLE 5.3: Predictive accuracy (percentage of correct predictions) of reproducing the correct starting molecule from the top predicted building block. A set of 40000 testing SMILES were used in order to evaluate the four models considered. The accuracy was measured among the top five predicted starting molecules with Top5 representing an accurate prediction within the top five predicted SMILES. Of the four trained models, MIT-DOUBLE obtained the highest accuracy (bold).

	Top1	Top2	Top3	Top4	Top5
MIT-augm-SINGLE	31.6	37.5	40	41.4	42.1
MIT-augm-DOUBLE	66.8	74	76.2	77.2	77.6
MIT-SINGLE	28.5	33.8	36.2	37.6	38.2
MIT-DOUBLE	**67.9**	**74.9**	**77.5**	**78.8**	**79.2**

As can be seen from Table 5.3, a strong difference in performance was observed between the "SINGLE" and "DOUBLE" models. This can likely be attributed to the fact that the "DOUBLE" models have seen both building blocks in the input space. Of the four models, the "MIT-DOUBLE" model obtained the highest predictive accuracy amongst the top five predictions. This model was selected as the AI component for the DINGOS algorithm, and was used for the remainder of the study. In all subsequent chapters the building generation model is referred to as the "BGEN" model, and the updated version of DINGOS as "DINGOS-BGEN"

5.3.2 Comparative Performance with the Previous Model

DINGOS was designed to produce synthesizable molecules which adhered to a user defined hypothesis. In Chapter 3, we stated that a limitation of the method was the

use of the multi-layer perceptron (MLP) model for descriptor prediction. While this did prove capable of generating structurally-related molecules, the need to produce an entirely new model in order to use different descriptors might be seen as a drawback. DINGOS-BGEN allows one to use any descriptor without the need for retraining. One concern in incorporating this new method is that the BGEN model may be less capable than the MLP model at recommending appropriate building blocks for *de novo* design. To investigate the influence that the use of the BGEN model had on the overall designs, a comparison was made between the DINGOS-BGEN algorithm and the original DINGOS model (Chapter 3). In order for the comparison to be meaningful, the same design parameters were used as in the previous study. The MACCS keys were used as the molecular descriptor and the Hamming distance (1 - Hamming loss) was used in order to quantify the similarity of the designs. Figure 5.3 shows a comparison of the distance values for the top 300 and top 20 *de novo* designs.

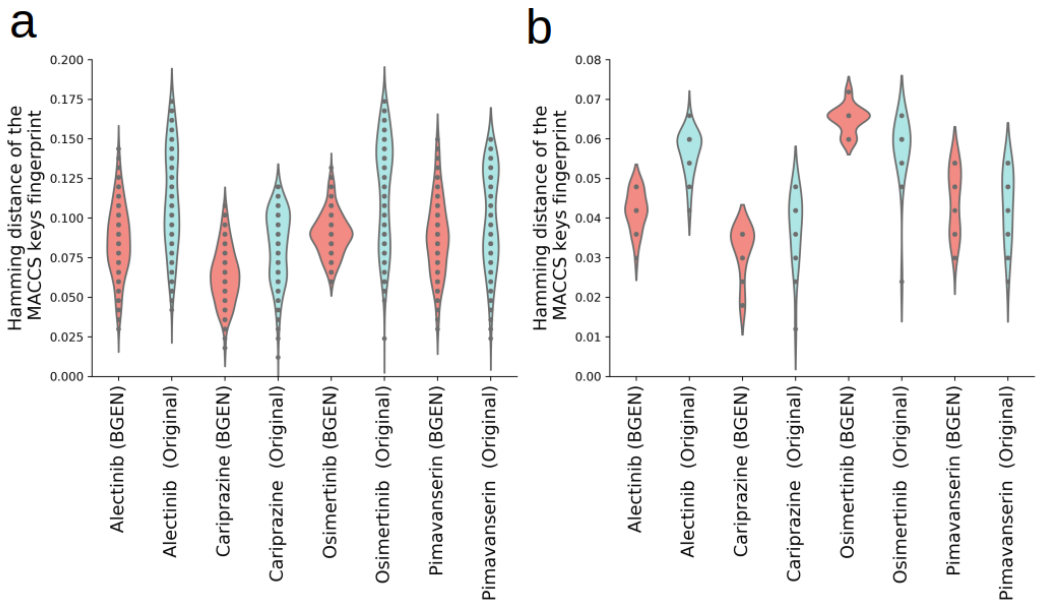


FIGURE 5.3: Distance comparison between the DINGOS-BGEN model (Red) and the DINGOS model presented in Chapter 3 (Blue). The same parameters and templates as those in Chapter 3 were used. Distances were calculated as the Hamming distance of the MACCS keys. The distances were compared for **a**) the top 300 and **b**) top 20 most similar *de novo* designs. In the case of the top 300 designs, one can see that DINGOS-BGEN improved upon the similarity for each template.

Comparing the median distances of the top 300 designs, we see that the DINGOS-BGEN algorithm outperformed the previous model for all four of the template ligands, while for the top 20, a lower median distance value of the DINGOS-BGEN algorithm was only obtained for the alectinib template (Table 5.4). Of the remaining drug templates, the difference in median distance of the top 20 designs was at most 0.01, within the interquartile ranges of the distributions. These results showed that the same, or better, quality of performance from DINGOS could be obtained through the use of the BGEN model.

TABLE 5.4: A summary of the median distance values obtained from the distance analysis presented in Figure 5.3. Median distance values are reported along with the inter-quartile ranges (IQRs) shown in brackets.

Template	Model	Median distance Top 300	Median distance Top 20
Alectinib	BGEN	0.08 (0.03)	0.04 (0.007)
Alectinib	Original	0.11 (0.05)	0.06 (0.006)
Cariprazine	BGEN	0.07 (0.02)	0.04 (0.006)
Cariprazine	Original	0.08 (0.04)	0.04 (0.01)
Osimertinib	BGEN	0.09 (0.02)	0.07 (0.001)
Osimertinib	Original	0.13 (0.06)	0.06 (0.006)
Pimavanserin	BGEN	0.09 (0.03)	0.05 (0.018)
Pimavanserin	Original	0.1 (0.06)	0.04 (0.01)

5.3.3 Choice of Descriptor Representation

Having now established the performance of the DINGOS-BGEN method at the previously considered case-problem, we sought to test the capabilities of the method with different ligand-scoring drug design problems. To this end, we compiled a set of eight well-known molecular descriptors, which are summarized in Table 5.1. In Chapter 3 the Hamming distance was chosen to be consistent with the loss function used in training the multi-layer perception model; however, while the Hamming loss is commonplace within the machine learning community, it is less common in the field of cheminformatics. The reason for this is that descriptor representations of molecules tend to be very sparse, owing to the large number of potential structural features, and because of this, the on-bits of the descriptor tend to be under represented in the similarity measurements [148, 149]. Unlike the Hamming distance, the Taniomoto distance only considers the shared on-bits, excluding all off-bit elements (see Equation 5.1).

$$D_{tanimoto}(A, B) = 1 - \frac{c}{a + b - c} \tag{5.1}$$

EQUATION 4: a is the number of on bits in the binary string A, b is number of on bits in binary string B, and c is number of on bits in both A and B

Because of this, the Tanimoto distance tends to give preferential weighting to these structural features found within the molecule. We chose to use the Tanimoto distance for all following work.

5.3.4 Case Study - Data Driven *De Novo* Design for Bioactive Compounds

The ultimate goal of the *de novo* drug design is to produce molecules with the desired biological properties. DINGOS is a template based ligand scoring method, and therefore, its designs are governed by the choice of scoring function (i.e. similarity metric) and template. In order to decide on the appropriate descriptor-template system, analysis was performed on a series of compounds with reported activity values. The goal of this analysis was to determine under which descriptor representation the compounds' distance values were proportional to their activity measurements.

Eleven sets of reported compounds were extracted from ChEMBL, each from a separate biological target of interest (see Methods 5.2.4). Two activity measurements were considered: K_i and IC_{50}. For each set, the compounds with the five lowest recorded activity values were chosen as template ligands, and their distance to the set was measured. This analysis was performed for each of the different descriptor representations. Two threshold values were defined, one for bioactivity, which was defined as being any point with an activity value less than 1 μM, and one for distance, which was the smallest distance observed amongst the inactive compounds (i.e. those below the bioactivity threshold). An example of this analysis in shown in Figure 5.4.

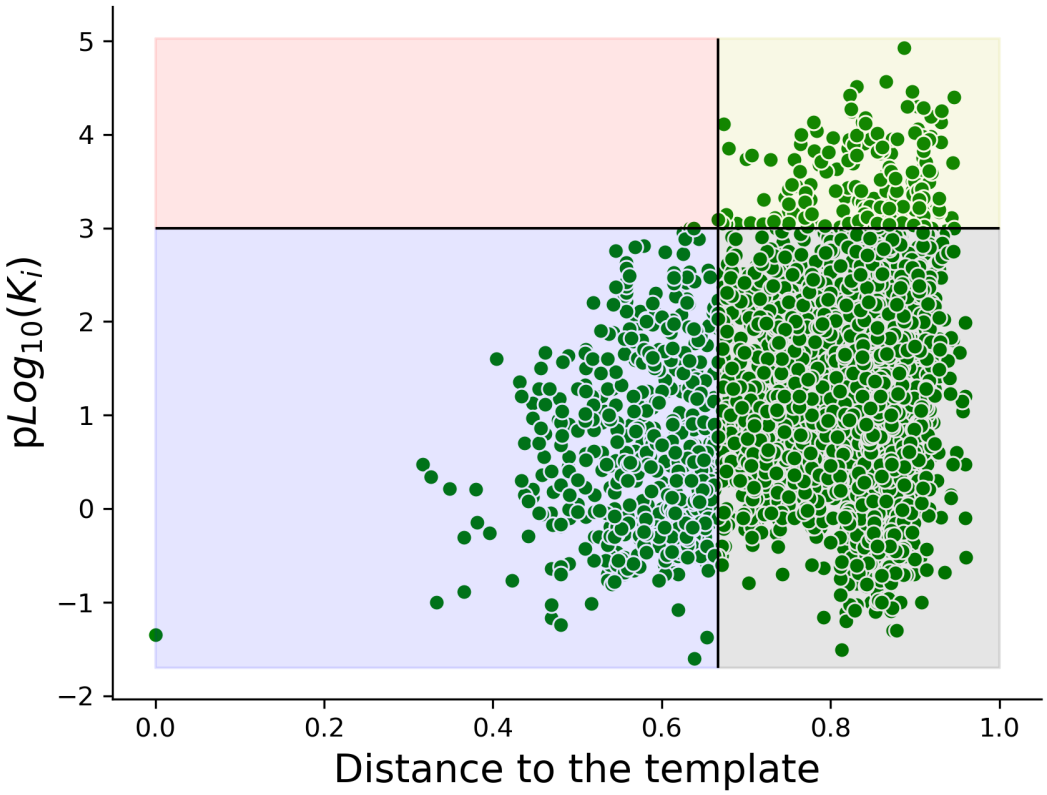


FIGURE 5.4: Distance values of the extracted H3 histamine bioactives (ChEMBL ID:264) relative to the template **72** (ChEMBL ID:2413837). Distances were calculated as the Tanimoto distance between ECFP4 descriptor values. The horizontal line depicts the activity threshold (1 μM), while the vertical line depicts the distance of the most similar observed inactive, forming our distance threshold. The plot is divided into four sections: (blue) Actives below the distance threshold (domain of interest), (red) inactives below the distance threshold, (yellow) inactives above the distance threshold, and (grey) actives above the distance threshold.

If the actives and inactives within each set were perfectly separable by distance, then only the lower left and upper right regions would be populated; however, as can be seen from Figure 5.4, the lower right corner (region in which all compounds are active and above the distance threshold) contains a significant proportion of the population. The lower left region represents a range of distances in which no inactive has yet to be observed. This region represents our "domain of interest" in which any compound showing a distance to the template within this range shows at least 1 μM activity. Table 5.5 summarizes the descriptor-ligand systems that gave the highest

number of points in the lower left region, that is, the highest number of positive examples within our "domain of interest".

DINGOS was used to produce *de novo* designs for each of the descriptor-ligand systems shown in Table 5.5. The goal was to investigate whether DINGOS would be capable of populating the various "domains of interest" with synthesizable examples, and hence, providing examples for which one could produce and test the underlying hypothesis.

TABLE 5.5: Summary of the descriptor-template systems that gave the largest number of points in the domain of interest (DOI). Of the considered systems, the H3 histamine receptor had the largest number of positive examples in the domain of interest. In contrast, the H2 histamine receptor, despite its ChEMBL set containing over 312 reported compounds, only showed three within the domain of the interest. For each set the ROC-AUC of the distance values was calculated.

Target name	Descriptor	Activity (nM)	Number of mols in DOI	Threshold value	ROC-AUC	Set size
Cannabinoid CB1	ECFP4	0.065 (Ki)	111	0.68	0.65	2420
Coagulation factor X	featMorgan	0.004 (Ki)	157	0.66	0.63	3534
Dopamine D1	torsional	0.2 (Ki)	153	0.73	0.62	904
Dopamine D2	ECFP4	0.058 (Ki)	196	0.65	0.61	5530
Dopamine D3	atom pair	0.043 (Ki)	436	0.55	0.78	3992
Dopamine D4	atom pair	0.03 (Ki)	72	0.47	0.65	1980
Dual specificity protein kinase TTK	ECFP4	0.2 (IC50)	292	0.79	0.81	877
Histamine H1	atom pair	0.108 (Ki)	67	0.57	0.67	1082
Histamine H2	RDKit	18 (Ki)	3	0.75	0.74	312
Histamine H3	ECFP4	0.045 (Ki)	502	0.67	0.64	3064
Histamine H4	layered	0.04 (Ki)	122	0.39	0.58	1056

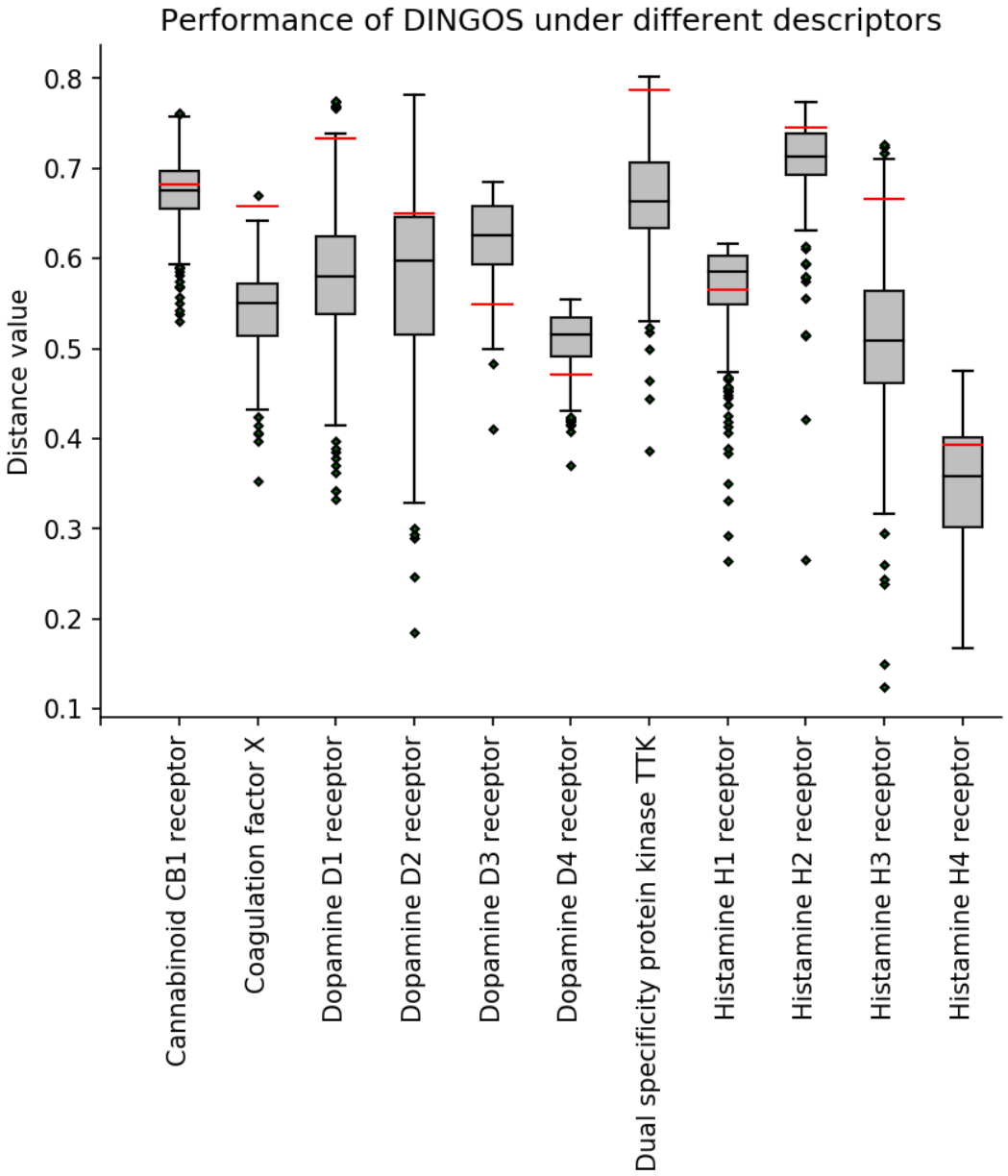


FIGURE 5.5: Median distance values of the DINGOS-BGEN *de novo* designs generated for the eleven ChEMBL template-descriptor systems. The distance threshold value defining the domain of interest for each system is depicted as a red line. Each template was chosen based on the bioactivity analysis, taking the ligand that resulted in the highest number of actives in the domain of interest. For each template system, DINGOS successfully generated *de novo* designs in the domain of interest.

As can be seen from Figure 5.5, for each system DINGOS was shown capable of generating designs below the threshold, producing between 15 and 287 designs within the domain of interest. The Dopamine D1 and dual-specificity protein TTK kinase systems generated the most number of designs below this threshold. For seven of the eleven cases, the median distance value of the *de novo* design populations was less than that of the threshold value (Table 5.6). The *de novo* compounds below the distance threshold represent a testable set of examples that could be synthesized

in order to interrogate the validity of the domain of interest, which could then be exploited to find novel, bioactive ligands.

TABLE 5.6: Summary distance analysis of the DINGOS *de novo* design experiment presented in Figure 5.5. Each template was chosen based on the bioactivity analysis, taking the ligand that resulted in the highest number of actives in the domain of interest.

Target name	Descriptor	Threshold value	Median distance (IQR)	Number of mols below the threshold
Cannabinoid CB1	ECFP4	0.68	0.68 (0.04)	187
Coagulation factor X	featMorgan	0.66	0.55 (0.06)	282
Dopamine D1	torsional	0.73	0.58 (0.09)	287
Dopamine D2	ECFP4	0.65	0.6 (0.13)	228
Dopamine D3	atom pair	0.55	0.63 (0.07)	15
Dopamine D4	atom pair	0.47	0.52 (0.04)	41
Dual specificity protein kinase TTK	ECFP4	0.79	0.66 (0.07)	287
Histamine H1	atom pair	0.57	0.59 (0.05)	98
Histamine H2	RDKit	0.75	0.71 (0.05)	245
Histamine H3	ECFP4	0.67	0.51 (0.1)	280
Histamine H4	layered	0.39	0.36 (0.1)	208

5.3.5 H3 Histamine Inhibitors - Exploration of Previously Unexplored Compound Space

Of particular note from the previous analysis was the H3 histamine system. Of these systems considered, the H3 histamine system had the highest number of reported actives in the domain of interest (502). Of the 300 *de novo* designs generated, 280 of the them possessed a distance below the H3 set's threshold value. Eight of the designs were found to be more similar to the template ligand than any other compounds found within the template's bioactive set. These designs are shown in Figure 5.6. One can see a gradual change in structural similarity away from the template ligand, with the top two designs (**73** and **74**) possessing only a two and three atom difference respectively. For these top two designs, the main diaryl scaffold was formed during the *de novo* design, rather than being present in the start molecule. For these two designs, two separate C-C bond forming reactions were used, Suzuki coupling and Negishi. This is in contrast to traditional combinatorial library approaches [150], in which a single reaction is enumerated in order to generate compounds.

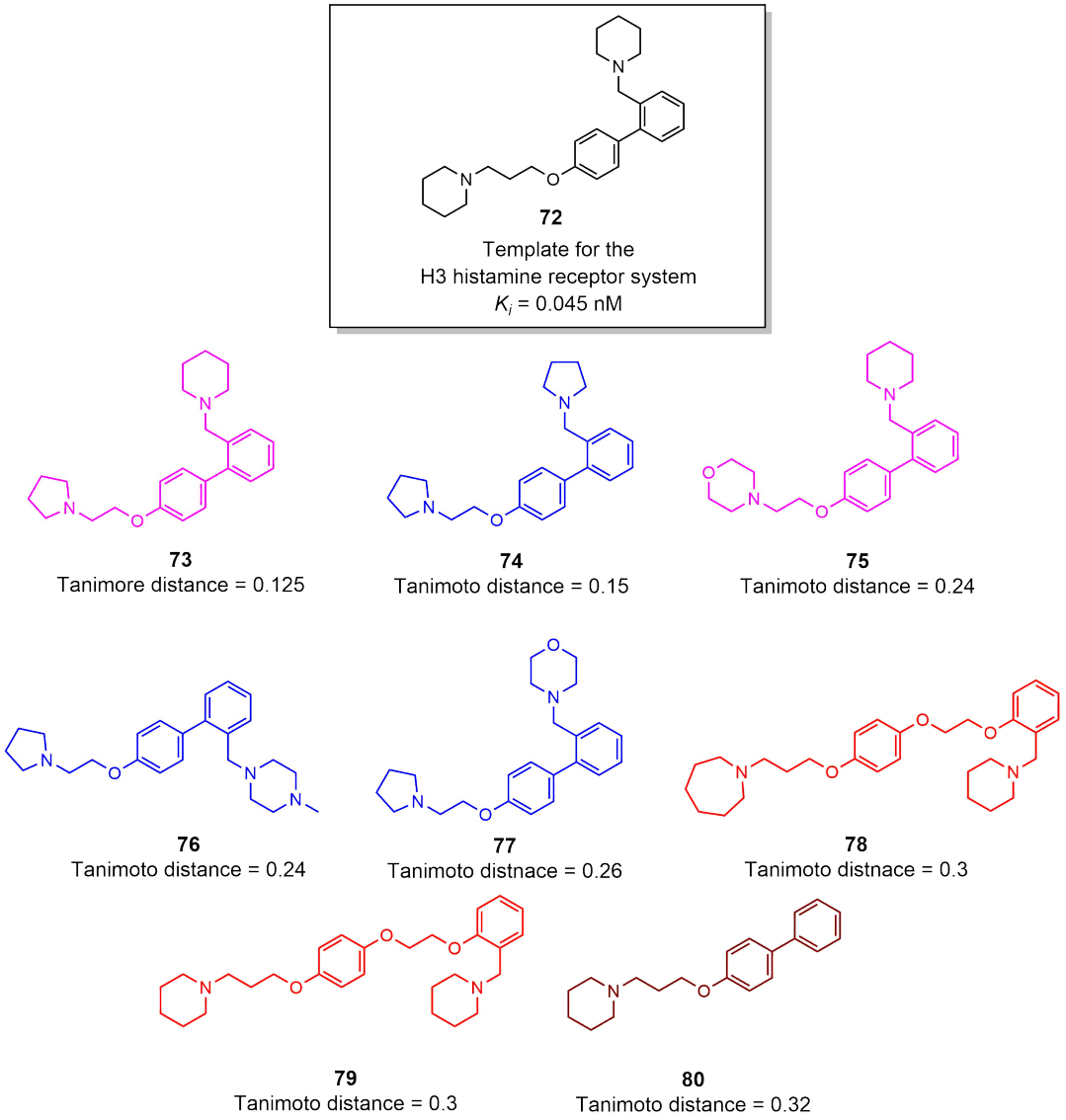


FIGURE 5.6: Eight most similar *de novo* designs produced for the H3 histamine system. Distance was measured as the Tanimoto distance of the ECFP4 descriptor. Compounds depicted in magenta are formed via a Suzuki coupling reaction, those in blue via Negishi coupling, in red nucleophilic aromatic substitution, and in brown Williamson ether.

5.3.6 TTK Kinase Inhibitors - Dataset with the Highest ROC-AUC Value

For each of the datasets, ROC-AUC values were calculated using the distance threshold as the discriminator. This value represented how well distance can be used to discriminate between the active and inactive points of the given dataset.

It was found that the TTK dual specificity protein kinase had the highest ROC-AUC value of any of the datasets considered, indicating that this was the system that was most well described by its distance to the template molecule. Of the 877 points reported for the TTK kinase, 38% were found to be in the lower left and upper right regions. These points are those in which activity can be safely separated by distance (see Figure 5.7). In contrast, for the H3 histamine system, only 10% of

the 3064 reported points confirmed the distance hypothesis. This is reflected by the significantly lower ROC-AUC value observed. For the TTK system, DINGOS was able to generate 287 designs with a distance value below that of the threshold (0.79). These designs ranged in distance from 0.39 to 0.80 with a median value of 0.66 (0.08).

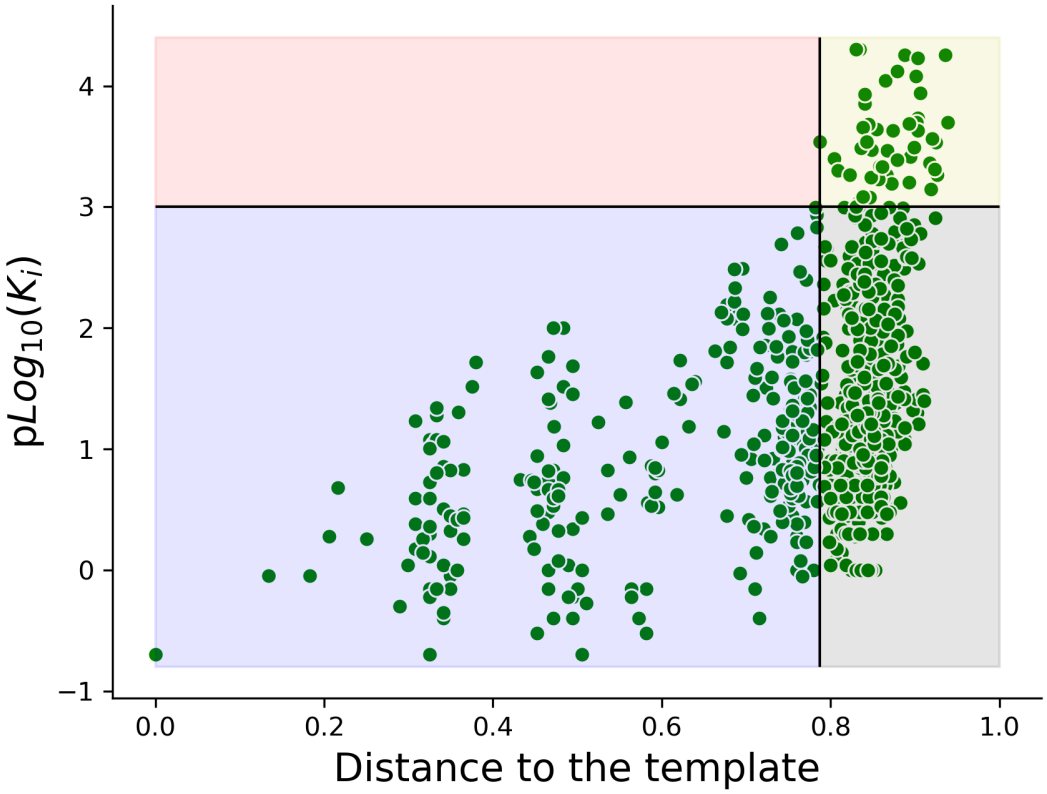


FIGURE 5.7: Distance values of the extracted dual-specificity protein kinase TTK (ChEMBL ID: 3983) relative to the template **81** (ChEMBL ID: 3951371). Distances were calculated as the Tanimoto distance between ECFP4 descriptor values. The horizontal line depicts the activity threshold (1 μM), while the vertical line depicts the distance threshold. The plot is divided into four sections: (blue) Actives below the distance threshold (domain of interest), (red) inactives below the distance threshold, (yellow) inactives above the distance threshold, and (grey) actives above the distance threshold.

Figure 5.8 shows the top eight most similar *de novo* designs from the TTK DINGOS experiment. In comparing the top design **82** to the template ligand, we see that it is considerably smaller than the template, with a molecular of 389 g mol_{-1} compared to that of 479 g mol_{-1}. This could be due to the choice of molecular descriptor, as ECFP, in contrast to say the MACCS keys, penalizes dissimilar substructural arrangements, rather than simply rewarding matching subgroups. As we go down the distance rankings we see larger structures emerging. Of particular note is that of the compound **88**, which had a molecular weight of 493 g mol_{-1}. Of the top eight designs, this was the only compound to possess the same benzimidazole scaffold as the template. Additionally, the design also contain the cyclopropane amide substituent, and the 3-floro aliphatic chain. It did, however, differ in the position of the 3-floro chain and the pyridine substituent about the benzimidazole ring, as well as lacking one methyl group and replacing a secondary amine with an amide bond.

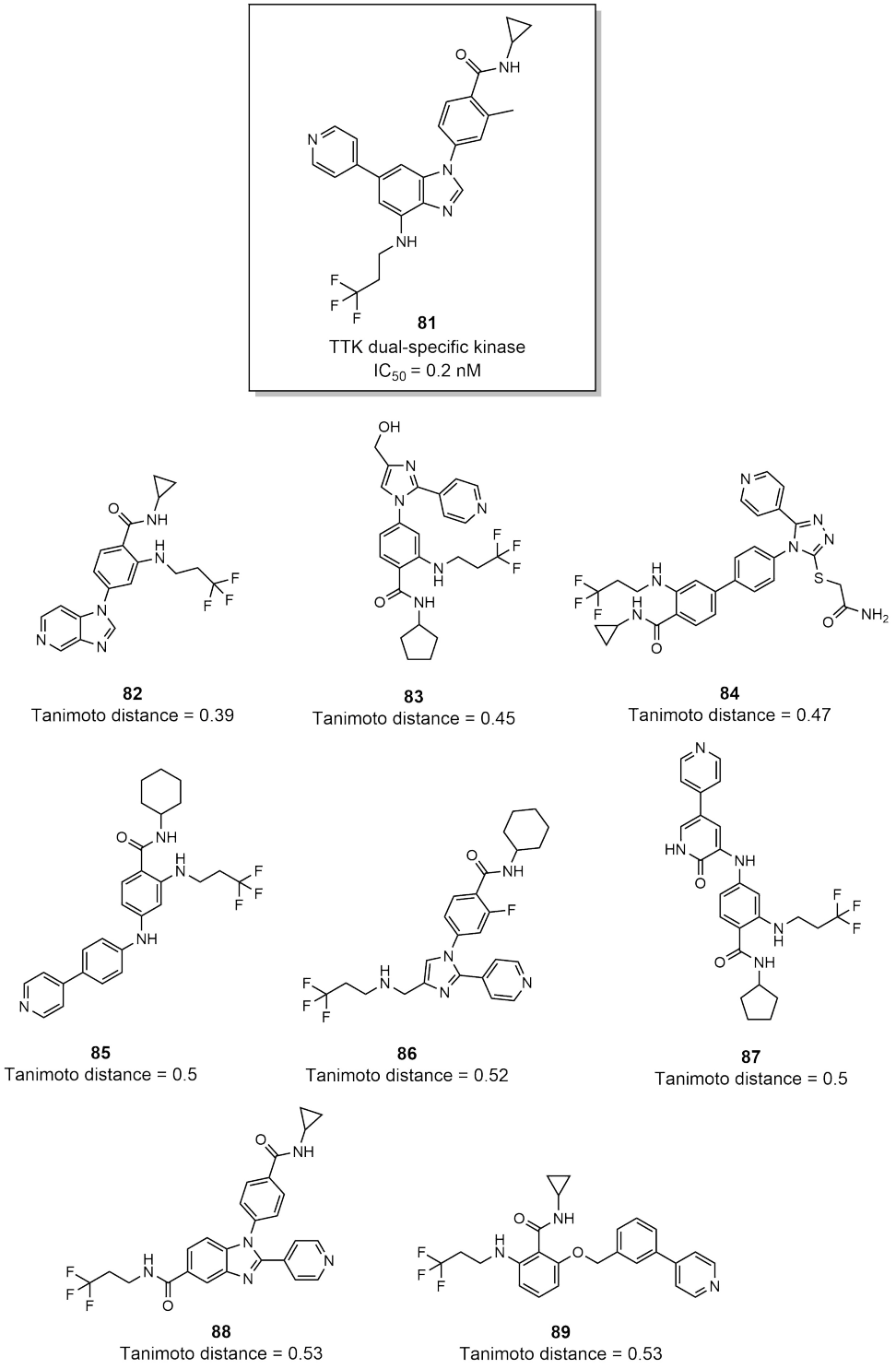


FIGURE 5.8: Eight most similar *de novo* designs produced for the dual-specificity protein kinase TTK system. The distance was measured as the Tanimoto distance of the ECFP4 descriptors.

5.4 Conclusion

Through the use of machine translation, we were successfully able to further generalise the DINGOS algorithm (DINGOS-BGEN). In contrast to the previous version

presented in Chapter 3, this modified version of DINGOS can accept any descriptor representation without the need for retraining of the machine learning model. The *de novo* design experiments performed in Section 3.3.1 were repeated with the DINGOS-BGEN model, and the distance values obtained were compared. It was found that the *de novo* sets produced by the DINGOS-BGEN model had a lower median distance value for each of the template ligands. In a further case-study, we performed analysis on a series of compounds extracted from the ChEMBL database. The goal was to determine the appropriate choice of template and descriptor for *de novo* design, and to see whether or not DINGOS was capable of generating "desirable" NCEs based on this analysis. For each system, we calculated a threshold distance value. Compounds below this threshold represented molecules in which low distances and bioactivity were observed. For each descriptor-template system, DINGOS was successfully able to populate this region. The highest proportion of the designs in this region were generated by the dopamine D1 and TTK protein kinase system, in which 287 designs were produced below this threshold. In the case of the H3 histamine receptor, the eight most similar compounds were more similar to the template ligand than any of the reported H3 active compounds.

Chapter 6

Extension of the DINGOS Algorithm Towards Non-Greedy Solutions in the Exploration of Chemical Space through the Use of Monte Carlo Tree Search

In the previous chapter we extended DINGOS to be capable of accepting any arbitrary descriptor and metric function. This greatly expanded the range of problems that DINGOS was capable of tackling. In effect, DINGOS can now be used for any problem that can be phrased as a distance optimization problem. Despite these improvements, other limitations are still present within the DINGOS algorithm. One main limitation is that the optimization always follows a greedy strategy [151], i.e. at each assembly step, only the highest scoring solution is considered. While incremental improvement through greedy optimization may yield optimal results in some cases, often a globally optimal solution requires making locally suboptimal steps [152–154]. In order to investigate this, the DINGOS algorithm was further extended to allow for non-greedy optimization. We chose the Monte Carlo tree search (MCTS) algorithm [155] in order to do this. This algorithm was chosen as it is an efficient method for performing local, explorative searches. Using the descriptor-template systems considered in Chapter 5.3.4, a series of simulated experiments were performed in order to parameterize the extended DINGOS-MCTS algorithm.

6.1 Introduction

6.1.1 Monte Carlo Tree Search

Decision theory is a sub-discipline of probability, in which we are concerned with transitions from particular states s in which there is a degree of uncertainty about the value of each state [76]. The act of moving from a given state to another can be viewed as a "decision" or "action". One particular problem in decision theory is that of the Markov decision process, in which we are required to make a series of sequential actions with the goal being to successfully transition to the most optimal state. It can often be useful to model processes such as these through a reward function $R(s)$ and policy. A reward function evaluates the value of being in a particular state, while a policy determines the probability of selecting a particular action from a given state. Depending on the problem of interest, the number of potential states and actions can be immense, leading to the need for effective searching algorithms [156, 157]. One

recent example is that of AlphaGo [56]. The game of Go is estimated to have some 10^{172} potential moves from the starting position [158, 159]. In order to effectively navigate this space, the team at Google employed a combination of Monte Carlo tree search(MCTS) and deep learning and were successfully able to defeat world champion Lee Sedol, in 2016 [56, 57]. MCTS is a search strategy in which the algorithm employs Monte Carlo sampling in order to estimate the value of a given state [155]. A series of simulated actions are performed, and the resulting reward values are averaged to give the expected reward for taking a particular action at the given state. This information is used to build up a model of the decision problem, allowing the algorithm to estimate which decisions will be most globally advantageous.

The MCTS algorithm is typically broken down into four distinct phases:

Selection

Starting at the initial state S_o, the MCTS algorithm transitions to one of the next available states S_i in the sequence. This process is repeated until some stop criterion is met. Selections are made by a policy called the *tree policy*.

Expansion

One or more new states are added from S_i. These states are selected from all possible states, governed by a separate policy, called the *expansion policy*.

Simulation

From the new state(s), a simulation is performed. In the simulation phase, a more efficient policy (faster), called the default- or rollout-policy, is used to simulate searching from the given state. The goal is to approximate the outcome one would expect from being in this state.

Backpropagation

Once the simulation is complete, the reward function is used to evaluate the reward of the final simulated state. This reward is then backpropagated through the sequence in order to reflect the ultimate outcome of selecting the early stage states.

A schematic representation of this process can be see in Figure 6.1.

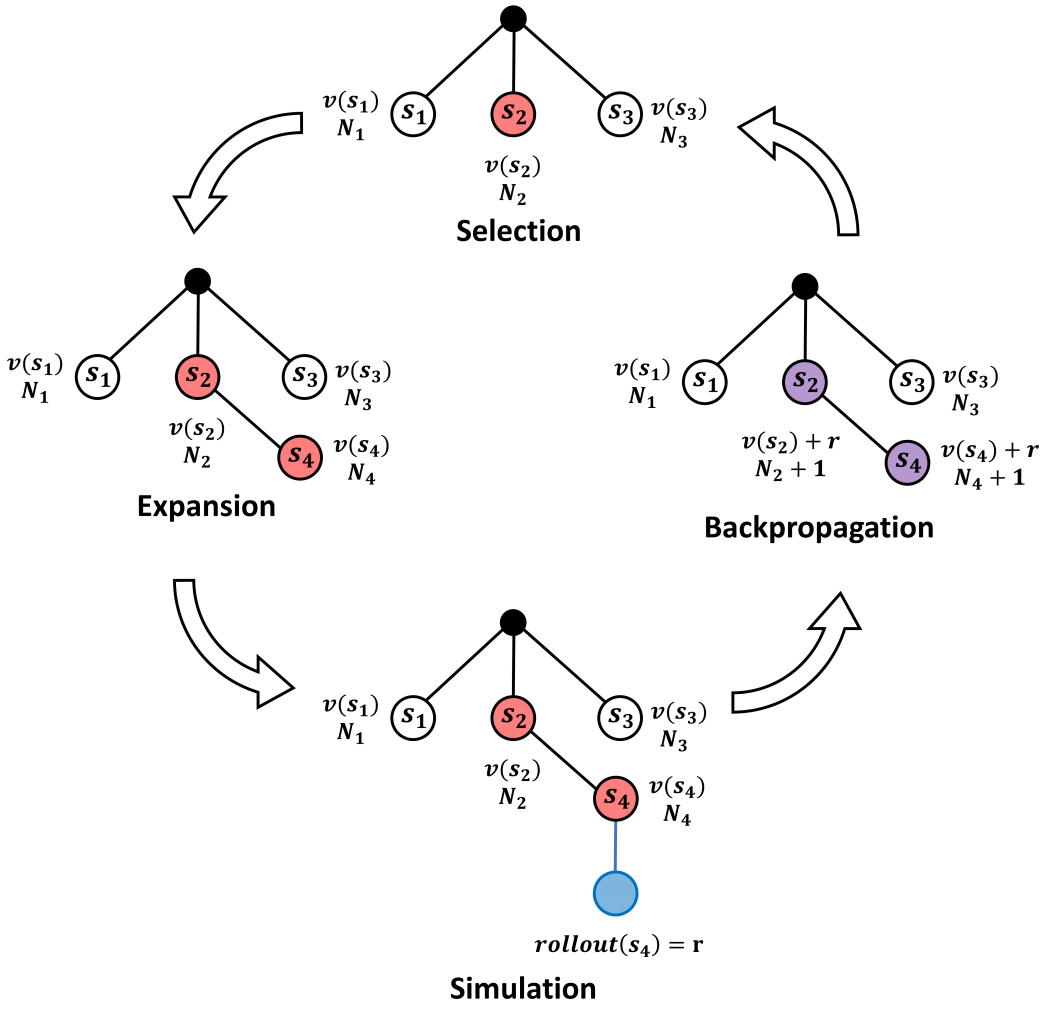


FIGURE 6.1: Schematic representation of the Monte Carlo tree search algorithm. Circles depict the individual states (s_1,s_2,s_3,s_4) and the connecting edges depict the actions between states. Each state possess a visit count N, which measures the number of times a particular state has been visited, and an intrinsic value given by $v(s)$. The four distinct steps in the MCTS algorithm are depicted: selection, expansion, simulation, and backpropagation. The circles in red represent states that have been selected for this iteration. The blue circle and edge represent the simulated state obtained via rollout. The value of this state is given by $rollout(s)$. The circles in purple represent the states for which the rollout value, r, is backpropagated. The value of the state and the visit counts are updated accordingly. This image was adapted from Cameron *et al.* [155].

6.1.2 *De Novo* Drug Design as a Decision Process

The MCTS extension presented here offers the potential to greatly expand the capabilities of the DINGOS algorithm. The primary advantage of this method is in the the ability to use future values to evaluate decisions. However, despite the potential benefits, the introduction of this extension greatly increases the complexity of the algorithm. We can frame the problem of *de novo* drug design as a Markov decision problem. In this picture, molecules represent the states, and the reactions performed our actions. The reward for which we are trying to optimize is the similarity of the

designs to our template of interest. In this chapter, we combine the MCTS approach with the DINGOS algorithm established in Chapter 5. The modified version of the DINGOS algorithm is referred to as the "DINGOS-MCTS" algorithm throughout this chapter.

6.1.3 Algorithm Components

Decision function
All selections are made by a decision function, which is a function that takes into account the 'value' of a given state, that is, the expected reward from selecting this state, as well as the number of times that this state has been visited. The constant parameter C is used to control the influence of these two attributes (see Equation 6.1).

$$Decision\ score = \bar{X}(s) + C * G(N) \tag{6.1}$$

The C parameter controls the explorative and exploitative behaviour of the search strategy. If C equals 0, then the decision function is equal solely to the expected reward, and a purely exploitative run is performed (each step is selected only based on which state gives the highest expected reward). If C is significantly larger than $X(s)$, then the N term will dominate and a purely explorative run will be performed, i.e. states are selected only by their visit counts. In practice, neither purely explorative or exploitative searching is desirable, but rather some combination of the two.

Expected reward
The expected reward, $\bar{X}(s)$, is calculated as the average reward obtained over all searches. The initial reward is obtained directly from the state being considered, while the rewards for future searches are obtained via the backpropagation. In effect, the expected reward incorporates the average of the state reward with those of the future states. This "reward" is evaluated using the value-policy. The value-policy is a function that calculates the quality of a given state as some positive reward. Adapting a technique used in AlphaGo, the state value was calculated as a combination of the state itself and the value of the rollout-policy. This is shown in Equation 6.2.

$$\begin{aligned} \bar{X}(s) &= \frac{1}{N(s)} \sum_{i=1}^{N} V(s_i) \\ V(s) &= \lambda * v(s) + (1 - \lambda) * rollout(s) \end{aligned} \tag{6.2}$$

Here, $v(s)$ is the intrinsic value of the state. In the context of a ligand-based scoring method, this is the distance of the design to the template ligand. The function $rollout(s)$ gives the reward obtained by applying the rollout policy at state s. The lambda parameter (λ) controls the degree to which the overall state value $V(s)$ is determined by its intrinsic value and its rollout (future) value. If lambda equals 1, we only consider the state values. If lambda equals 0, we only consider the reward of future values. In many MCTS problems rollout is more important than intrinsic state value, as we are often primarily concerned by the terminal states of the tree (states that cannot be further expanded). In the context of similarity-driven drug design, this is not the case, as intermediate solutions are as valid as terminal state solutions. It is, therefore, worthwhile to consider the intrinsic value of the state value as well as the rollout value.

Explorative function

The explorative function, $G(N)$, governs to what degree the algorithm values the novelty of a particular state over its expected reward. Various explorative terms can be used, with this defining multiple different decision functions. For this study, we chose the *upper confidence bound for tree* (UCT), which is defined as

$$G(N) = \sqrt{\frac{2 \log N}{n}} \tag{6.3}$$

The UCT function takes two terms, n, which is the number of total searches of the parent state, and N, which is the number of searches of the state being considered. This represents a measure of the number of times the state has been considered, weighted by the visit count of the parent. The less often a state has been visited, the higher the value of $G(s)$, and hence, the more likely this state is to be selected.

Terminal states

A terminal state is one that cannot be expanded upon further. This can either be defined by exhaustive constraints, i.e. there being no other potential moves available, or by some external stop criteria. Once a terminal state is reached, no further actions can be taken from this state. In the DINGOS-MCTS algorithm, a state is defined as terminal if it reaches one of the DINGOS stop criteria (Molecular weight limit, reaction step limit, etc.)

Rollout

The rollout-policy is a function that allows us to efficiently evaluate the future rewards expected from a given state position. The idea of a rollout-policy is that it is a more simplistic, less-accurate model, capable of simulating searches quickly. By performing fast evaluation at each state s, we are able to estimate the potential value of previously unexplored states. For this work, we chose to use the DINGOS algorithm for the rollout-policy. The idea here is that we would use a reduced form of the algorithm to generate a "simulated-design". This compound is the design that would be obtained if the DINGOS algorithm was applied using the chosen state as the starting molecule. In the context of MCTS, in which the algorithm favours many iterations, the full DINGOS algorithm is too slow to expect convergence in a reasonable time scale. To solve this, we restricted the DINGOS compound database to only considers the R most similar molecules to the template. This value R was called the "rollout-depth". Restricting the number of potential building blocks sped up the algorithm, but also reduced the number of potential products that could be formed.

State expansion

At each state position there are an enormous number of potential intermediate products that could be formed. Considering all of the potential start molecule-building block pairs would be infeasible. To resolve this, we employed the use of an expansion-policy. The expansion-policy is a function that selects a small subset of potential intermediate products from a given state. In the context of this work, we employed the DINGOS algorithm, restricting the method to a single reaction step, and only producing M intermediate products. It should be noted that we allow for the further expansion of states that have already been expanded upon. In effect, this means that given enough time, the MCTS algorithm will consider all potential start molecule-building block pairs.

6.2 Methods

6.2.1 DINGOS Parameters

The DINGOS-MCTS algorithm was adapted from the extension of DINGOS presented in Chapter 5 (DINGOS-BGEN). The same reaction and compound database was used. A molecular weight limit of 400 and 600 g mol^{-1} was used for the building block and product molecules respectively. A reaction limit of four was used. For the expansion state experiments, the building block recommendation pool was set to be ten times that of the expansion limit.

6.3 Results and Discussion

6.3.1 Algorithm Parameterization

The performance of the MCTS approach is determined by multiple parameters. In order to determine what value these parameters should take, we performed a parameter optimization experiment. For the purpose of making a meaningful comparison, we chose the bioactive ChEMBL set from Section 5.2.4. Due to the combinatorial nature of parameter optimization, we chose only to optimize the C and lambda parameters. A rollout depth of 1000 and an expansion limit of 1 was used for each of the *de novo* design runs. All experiments were run for 12 hours. In order to facilitate comparison with the results from Section 5.3.4, we evaluated the parameters according to the median distance value of the top 300 most similar designs. This was done in order to provide a fair comparison between the 300 compounds generated by the DINGOS-BGEN algorithm. For the experiments, we considered C parameter values of 0, 0.01, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75, and 1.0, and lambda values of 0.0, 0.25, 0.5, 0.75, and 1.0. Figure 6.2 shows a heatmap of the values obtained. A summary of the parameters that gave the lowest median distance value amongst the 300 most similar designs is shown in Table 6.1.

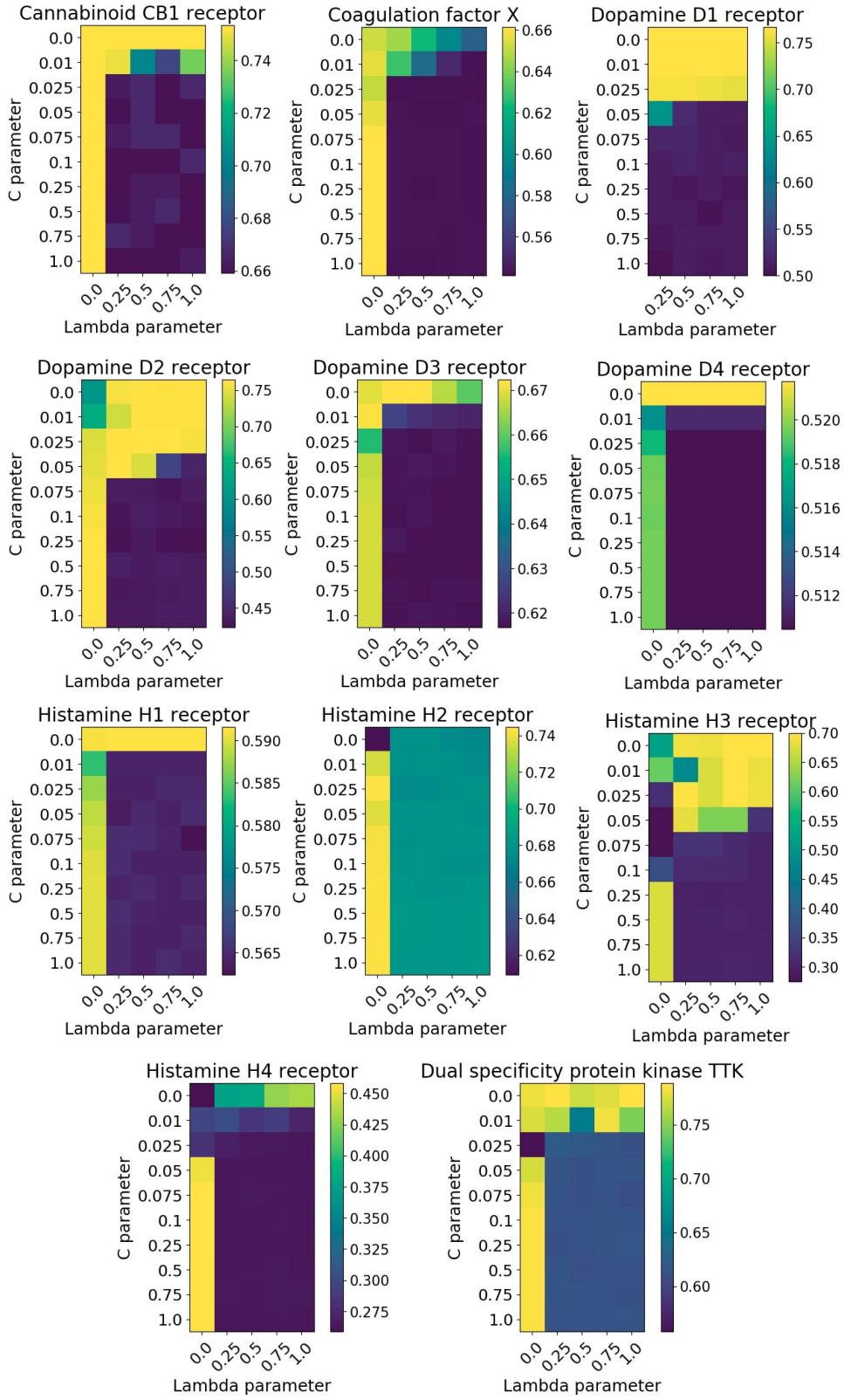


FIGURE 6.2: Heatmaps showing the median distance values from the C-lambda parameterization experiments. The maps are coloured according to median distance of the 300 most similar *de novo* designs generated in each run, with blue indicating low distance, and yellow indicating high distance. The y-axis shows the C parameter values (0, 0.01, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1.0), while the x-axis shows the lambda values (0.0, 0.25, 0.5, 0.75, 1.0).

In comparing the results, no consistent combination of parameters gave optimal performance for all of the considered systems. For the TTK kinase, and the H2, H3, and H4 histamine receptor systems, the minimum median value was obtained for a lambda value of zero, that is, one in which no simulated searching was performed. For two of these systems, that of the H2 and H4 histamine receptor, optimal performance was obtained with a C parameter of zero, representing purely explorative runs. For the remaining systems, we see from Figure 6.2 that a lambda value of zero had a detrimental influence on the results produced (shown by the first column). We also observed that above a certain C parameter value, a comparable degree of performance was observed e.g. 0.025 for the coagulation factor Xa.

TABLE 6.1: Optimal parameters obtained from the C-lambda parameterization experiment, performed for each of the template ligands used Section 5.3.4. Optimal was defined as the parameters that gave the lowest median distance amongst the 300 most similar designs.

Target name	C	Lambda	Median distance (300)	Median distance (20)	Top distance
Cannabinoid CB1 receptor	1	0.25	0.66	0.62	0.57
Coagulation factor X	0.025	1	0.54	0.5	0.48
Dopamine D1 receptor	1	0.25	0.5	0.39	0.33
Dopamine D2 receptor	0.25	1	0.42	0.29	0.18
Dopamine D3 receptor	0.5	1	0.62	0.56	0.52
Dopamine D4 receptor	0.75	0.75	0.51	0.44	0.41
Dual specificity protein kinase TTK	0.025	0	0.56	0.5	0.39
Histamine H1 receptor	0.075	1	0.56	0.45	0.29
Histamine H2 receptor	0	0	0.61	0.59	0.27
Histamine H3 receptor	0.075	0	0.27	0.12	0.1
Histamine H4 receptor	0	0	0.26	0.21	0.18

Having now parameterized the DINGOS-MCTS algorithm, we sought to investigate the influence that the rollout depth and expansion limit had on the designs produced.

6.3.2 Investigating Rollout-Depth Limits

The rollout depth defines the number of molecules available to our rollout-policy. There is a trade-off in setting the rollout depth. A large value leads to a more accurate rollout-policy, however, this policy is also more computationally expensive, leading to a reduction in the number of iterations performed within the set runtime. In order to investigate the influence of this, we performed a series of *de novo* design experiments using a rollout depth of 1000, 5000, 10000, 50000, and 100000. For the remaining parameters the optimal values from Section 6.3.1 (see Table 6.1) were used. For runs where optimal performance was obtained with a lambda equal to zero, we chose the best parameter set with a non-zero lambda value.

Table 6.2 shows the median distance values for the runs that gave the lowest median distance for the top 300 molecules. In cases of multiple "best" runs, the run with the smallest rollout depth was selected, as this would be the most computationally inexpensive run.

TABLE 6.2: Summary of the results of the rollout-depth experiments. A series of *de novo* design experiments were performed with DINGOS-MCTS for each of the template ligands used Section 5.3.4. A rollout-depth value of 1000, 5000, 10000, 50000, and 100000 was used for the *de novo* design. Only the optimal parameters are shown for each template ligand. Optimal was defined as the parameters that gave the lowest median distance amongst the 300 most similar designs.

Run name	Rollout depth	Median distance (300)	Median distance (20)	Top distance
Cannabinoid CB1 receptor	10000	0.64	0.58	0.53
Coagulation factor Xa	10000	0.48	0.42	0.34
Dopamine D1 receptor	5000	0.49	0.4	0.34
Dopamine D2 receptor	5000	0.41	0.29	0.18
Dopamine D3 receptor	50000	0.57	0.48	0.42
Dopamine D4 receptor	1000	0.51	0.44	0.41
Dual specificity protein kinase TTK	1000	0.56	0.5	0.39
Histamine H1 receptor	100000	0.53	0.43	0.29
Histamine H2 receptor	1000	0.61	0.59	0.27
Histamine H3 receptor	5000	0.26	0.12	0.05
Histamine H4 receptor	1000	0.26	0.21	0.18

For the TTK protein kinase, Dopamine D4, and Histamine H2 and H4 targets, the optimal median distance value was obtained using the same rollout depth used in the C-lambda parameterization. One notable result is the improvement observed for the top designs of the Coagulation Factor Xa target. The optimal result from the C-lambda parameterization gave a distance of 0.48, however, by increasing the rollout depth from 1000 to 100000, a distance of 0.34 was obtained. Shown in Figure 6.3 is a comparison of these two top designs.

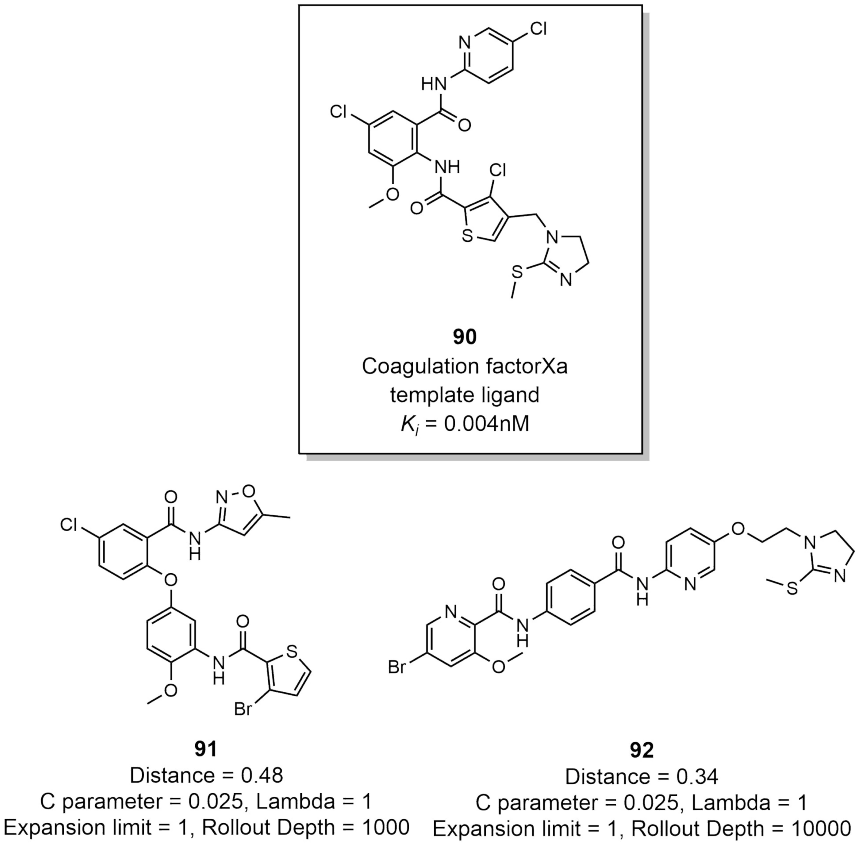


FIGURE 6.3: DINGOS-MCTS designs for the coagulation factorXa (ChEMBL ID: 244) template ligand (ChEMBL ID: 226775). Compounds **91** and **92** are the most similar *de novo* designs generated from Section 6.3.1 (C-lambda parameterization) and Section 6.3.2 (rollout parameterization) respectively. Compound **92**, which possessed a rollout depth ten times greater than that of **91**, was to be closer to the template by 0.14.

6.3.3 Investigating Expansion Limits

In the previous analysis, only a single state expansion was considered, however, it is also possible to allow for multiple expanded states to occur at each expansion event. This enables efficient exploitation of desirable states, however, depending on the choice of C and lambda, may bias the results towards these early selected states. In order to investigate the influence of the expansion limit, a series of *de novo* design experiments using a state expansion limit 1, 20, 100, 1000, and 10000 were performed. One important note, at each expansion event, a simulation is performed on each of the expanded states, and each of the simulation scores are back-propagated through the tree separately. The increase in the number of expansions increases the number of simulated rollouts that are performed, and, therefore, increases the computational load of each search. Taking this into consideration, we set the rollout depth to 1000 for all runs. The optimal C and lambda values from the parameterization experiment in Section 6.3.1 were used. Table 6.3 shows the distance values for the runs that gave the lowest median distance value for the top 300 molecules. Again, as in Section 6.3.2, when multiple "best" runs were observed, the run with the lowest expansion limit was selected.

TABLE 6.3: Summary of the results of the expansion-limit experiments. A series of *de novo* design experiments were performed with DINGOS-MCTS for each of the template ligands used Section 5.3.4. Expansion-limit values of 1, 20, 100, 1000, and 10000 were used for the *de novo* design. Only the optimal parameters are shown for each template ligand. Optimal was defined as the parameters that gave the lowest median distance amongst the 300 most similar designs.

Run name	Expansion limit	Median distance (300)	Median distance (20)	Top distance
Cannabinoid CB1 receptor	20	0.65	0.62	0.57
Coagulation factor Xa	20	0.53	0.5	0.48
Dopamine D1 receptor	100	0.48	0.38	0.33
Dopamine D2 receptor	1	0.42	0.29	0.18
Dopamine D3 receptor	20	0.61	0.56	0.53
Dopamine D4 receptor	1	0.51	0.44	0.41
Dual specificity protein kinase TTK	20	0.42	0.37	0.34
Histamine H1 receptor	20	0.55	0.43	0.29
Histamine H2 receptor	1000	0.56	0.5	0.27
Histamine H3 receptor	100	0.16	0.07	0
Histamine H4 receptor	100	0.17	0.15	0.13

For the Dopamine D2 and D4 targets, the minimum median distance was obtained with the original expansion limit. Of particular note is the case of the H3 histamine receptor, in which increasing the expansion limit to 1000 resulted in the DINGOS-MCTS algorithm producing two designs with zero distance to the template. The structure of these two designs is shown in Figure 6.4.

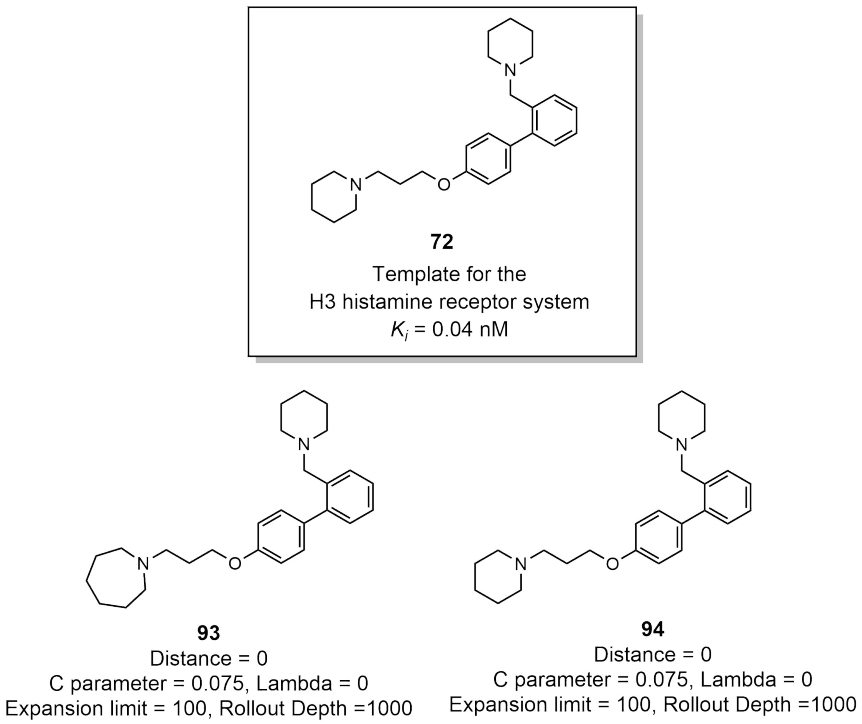


FIGURE 6.4: DINGOS-MCTS designs for the H3 histamine (ChEMBL ID:264) template ligand (ChEMBL ID:2413837). Both designs possessed a zero distance value, and as one can see, compound **94** was identical to the target ligand.

As can be seen, compound **94** has exactly the same structure as the template ligand. Compound **93** does vary structurally from in the template ligand in that one of the piperidine rings is actually an azepane ring. This difference is at the limit of what can be detected by the ECFP descriptor.

6.3.4 Populating the Domain of Interest

Table 6.4 summarizes the parameters that gave optimal results from the three parameter experiments. These results were compared with those of Section 5.3.4. The number of molecules produced by the DINGOS-BGEN algorithm is limited to 300 (the number of starting molecules). The DINGOS-MCTS algorithm is not limited in such a way, and hence, each *de novo* set did not have a fixed number of compounds. In order to make a meaningful comparison, we selected the 300 most similar designs from the DINGOS-MCTS *de novo* set. Figure 6.5 shows a comparison of the distance values obtained between the two DINGOS algorithms. These are summarized in Table 6.5.

TABLE 6.4: Optimal parameters obtained from the DINGOS-MCTS *de novo* design experiments considered in this chapter. Optimal was defined as the parameters that gave the lowest median distance amongst the 300 most similar designs.

Target name	Descriptor	C	lambda	rollout depth	expansion limit
Cannabinoid CB1 receptor	ECFP4	1	0.25	1000	1
Coagulation factorXa	featMorgan	0.025	1	10000	1
Dopamine D1	torsional	1	0.25	1000	1
Dopamine D2	ECFP4	0.25	1	5000	1
Dopamine D3	atom pair	0.5	1	50000	1
Dopamine D4 receptor	atom pair	0.75	0.75	1000	1
Dual specificity protein kinase TTK	ECFP4	0.025	0	1000	20
Histamine H1	atom pair	0.075	1	100000	1
Histamine H2	rdkit	0	0	1000	10000
Histamine H3	ECFP4	0.075	0	1000	100
Histamine H4	layered	0	0	1000	100

As can be seen, the DINGOS-MCTS algorithm outperformed the DINGOS-BGEN algorithm for each of the target systems. The largest difference in performance was observed for the H3 histamine receptor target, in which a difference in median distance of 0.35 was observed. By contrast, the smallest observed median distance difference, which was for the H4 histamine receptor, was 0.01, which is within the inter-quartile range of the two populations.

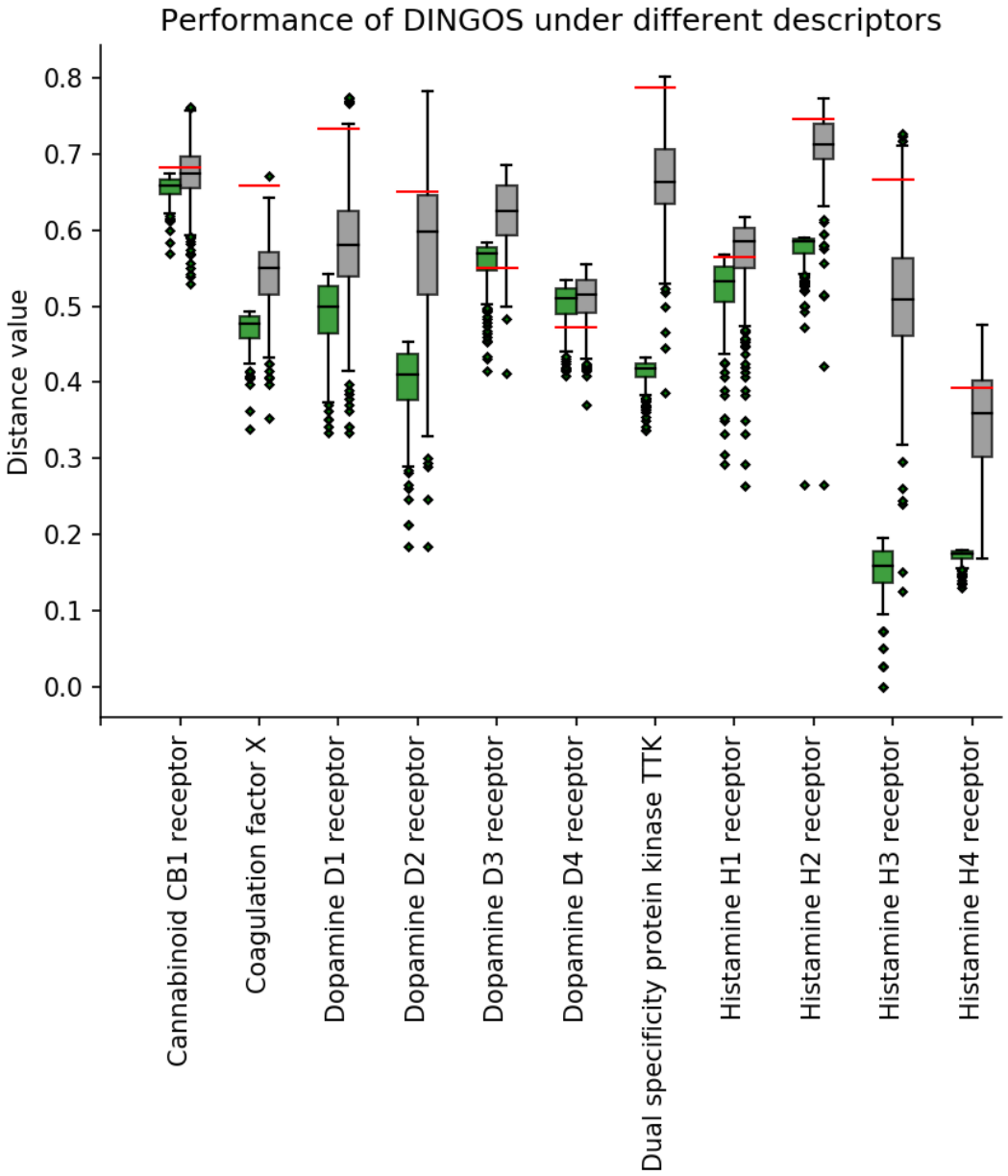


FIGURE 6.5: Distance comparison between DINGOS-BGEN and DINGOS-MCTS *de novo* designs. *De novo* designs were produced for the eleven template ligands considered in Section 5.3.4. For the DINGOS-MCTS algorithm, the parameters shown in Table 6.4 were used for the *de novo* design. For the comparison, the 300 most similar designs were selected from those generated by DINGOS-MCTS.

We can also compare the number of molecules which were produced below the distance threshold value. Table 6.5 summarizes these results. Of the eleven target systems considered, eight of the DINGOS-MCTS runs resulted in 300 molecules below the threshold (red line). In contrast, none of the DINGOS-BGEN runs results in the entire population being below the threshold, with the largest being 287 generated for the D_1 dopamine and TTK kinase systems.

TABLE 6.5: Comparison of the number of molecules generated below the distance threshold by the DINGOS-BGEN and DINGOS-MCTS algorithms, respectively. For the comparison, the 300 most similar *de novo* designs generated by the DINGOS-MCTS algorithm were used.

Target name	Descriptor	Threshold value	Mols below the threshold (BGEN)	Mols below the threshold (MCTS)
Cannabinoid CB1	ECFP4	0.68	187	300
Coagulation factor Xa	featMorgan	0.66	282	300
Dopamine D1	torsional	0.73	287	300
Dopamine D2	ECFP4	0.65	228	300
Dopamine D3	atom pair	0.55	15	83
Dopamine D4	atom pair	0.47	41	40
Dual specificity protein kinase TTK	ECFP4	0.79	287	300
Histamine H1	atom pair	0.57	98	290
Histamine H2	rdkit	0.75	245	300
Histamine H3	ECFP4	0.67	280	300
Histamine H4	layered	0.39	208	300

6.4 Conclusion

The results presented in this chapter represent a first step towards exploring the potential capabilities of the DINGOS-MCTS algorithm. Monte Caro Tree Searching was incorporated within the DINGOS algorithm in order to expanded DINGOS' *de novo* design procedure to non-greedy solutions. A parameterization experiment was performed for the C and lambda parameters to determine their influence on the *de novo* design. This parameterization was performed on the eleven bioactive template ligands derived from our analysis of the ChEMBL database performed in Section 5.3.4. The optimal choice of parameter was found to be template dependent, with different parameters being selected for different template ligands. Using the parameters derived from the parameterization, a series of experiments were performed in order to investigate the influence of the rollout depth and expansion limit. While improvements were shown by increasing the the values of these parameters, this was not consistent across all target systems, with four of the target systems showing no improvement with increased rollout depth, and one showing no improvement with an increased expansion limit. The experiments revealed that some targets were more susceptible to particular parameters than others. The Coagulation Factor Xa showed improvements upon increasing the rollout depth, while very little improvement was seen when increasing the expansion limit. In contrast, by increasing the expansion limit on the H3 histamine receptor, we obtain a significant decrease in the median distance, as well as obtaining an "optimal" *de novo* design. One aspect where an improvement was observed was in the number of high scoring designs produced. For eight of the eleven systems, all of the 300 most similar *de novo* designs were found to be below the distance threshold. In contrast, this was not observed for any of the *de novo* sets generated by the DINGOS-BGEN model. Due to the explorative nature of the tree searching algorithm, the MCTS extension lead to a higher proportion of

high scoring designs. In the context of the bioactive generation, DINGOS-MCTS was capable of populating our domain of interest with hundreds of designs. This large number of designs provides a large sample of examples that could be syntheiszed and tested to better understand the nature of our target problem.

Chapter 7

Conclusion

The goal of this work was to develop techniques for generating *de novo* drug designs that are both similar to a provided template ligand and synthetically tractable. We intended for this method to be highly modular and flexible in order for it to be smoothly adapted and integrated within a given drug design project. The primary motivation for achieving this goal was to bridge the gap between the *in silico* and experimental components in drug design and discovery. Computational methods offer a promising solution to many of the problems presented in drug design and discovery; however, experimentally generating and verifying results from these methods still remains a challenging enterprise.

To address this, we developed the DINGOS algorithm. DINGOS combines machine learning with aspects of rule-based methods in order to perform *de novo* drug design. Within these rules, we force the generated designs to adhere to the desired chemical logic. This, in effect, provides the chemists with the capability to directly incorporate their expert knowledge into the design procedure. In Chapter 3, we presented the DINGOS algorithm and the initial case-study that was used to evaluate its performance. DINGOS uses a virtual synthetic procedure to react sets of molecule *in silicio* to generate NCEs. The pairs of molecules that are reacted are selected by ligand-based scoring combined with predictive artificial intelligence. The four drug compounds alectinib, cariprazine, osimertinib, and pimavanserin were selected as template ligands, and with these, four sets of *de novo* designs were produced by DINGOS. Distance analysis showed a general improvement in similarity towards the template ligand over the compound database used. The designs showed adherence to the physico-chemical properties of the template, as well as above 50% of the compounds being predicted as active in target prediction studies. Four designs from the *de novo* sets produced by DINGOS were selected for synthesis. We were successfully able to synthesize three of the four designs with the synthetic pathway proposed by DINGOS. The remaining design was found to be light and water sensitive and so could not be structurally characterised. Each synthesized design was tested against the biological target of the respective template lignand, and of the tested designs, one compound showed a partial agonism against the 5-HT_{2B} equivalent to 1 μM of serotonin. This study showed that DINGOS is capable of achieving its intended purpose, that is, proposing synthetically feasible molecules that are similar to a given template ligand.

In Chapter 4, a follow up study was conducted in which DINGOS was integrated within an active learning drug design cycle. Here, the goal was to further interrogate the overall synthetic feasibility of the *de novo* designs produced, while simultaneously testing the modular aspects of the DINGOS algorithms. DINGOS was combined with surface plasmon resonance and a custom continuous flow synthesis system to

produce *de novo* designs autonomously. The carbonic anhydrase inhibitor acetozolamide was selected as the template ligand, and the compound and reaction databases were restricted to adhere to the capabilities of the in-lab environment. Two cycles of the active learning procedure were successfully performed, generating a total of 22 designs. Within the first cycle, 15 of the 40 proposed designs were successfully synthesized. The unsuccessful designs were primarily due to limitations of the system restricting the choice of solvent and reagents. Of the 15 synthesized designs, no binding was observed towards the carbonic anhydrase protein. Modifications were made to the compound and reaction database in order to improve the synthesizability within the flow and the overall binding affinity of the designs. In the second cycle, seven of the eleven proposed designs, were successfully synthesized and of these, five showed binding towards the target protein. Three of these were shown to bind with a low micromolar K_d value. Here, the modular nature of the DINGOS algorithm proved advantageous, as it enabled us to rapidly update our procedure to better serve the target problem and system. Modifications made through DINGOS lead to an increase in the synthesizability of the designs from 38% to 64%, and resulted in the generation of five novel carbonic anhydrase binders.

In Chapter 5, modifications were made to the core machine learning component of the DINGOS algorithm in order to allow for flexible substitution of the descriptor representation. The initial implementation performed building block recommendation by predicting the molecular descriptor values. This prevented a change in descriptor representation, limiting the scope of potential design problems. This predictive model was replaced with a language-based model that generated the molecular structures directly. By using a descriptor agnostic model, DINGOS was effectively generalized to any descriptor representation. In order to test the capabilities of the updated DINGOS algorithm, a series of eleven bioactives sets were extracted from the ChEMBL database, with each set representing a different biological target. For each set, we defined a "domain of interest", which was a distance range that we were interested in populating with novel, *de novo* designs. Each set's domain of interest was defined for a particular descriptor representation. For each set, DINGOS was shown capable of populating this region, and in some cases up to 96% of the designs produced were within this region. These results show that with the modified machine learning component, DINGOS can successfully perform descriptor agnostic *de novo* design.

In the final chapter (Chapter 6), we explored one potential extension to the DINGOS algorithm. Taking inspiration from advances seen in algorithms such as AlphaGo, we combined the DINGOS algorithm with Monte Carlo tree searching, in order to allow for non-greedy optimization of the *de novo* designs. Here, a decision tree method was implemented, with DINGOS acting as the expansion and rollout policy, generating and simulating new states through DINGOS' *de novo* design procedure. The C and lambda parameter, controlling exploration and rollout respectively, were parameterized for each of the sets of bioactive molecules considered in the previous chapter. No consistent choice of parameters was found which resulted in globally-optimal results. Once fully parameterized, two experiments were performed in order to investigate the influence of the rollout policy and expansion limit. Again, no general trend was observed, with the results appearing to be template specific. In comparing the results obtained with those of the previous chapter, the extended version of the DINGOS algorithm was found to outperform the previous implementation for each system considered, with eight of the sets being entirely comprised of compounds within our domain of interest. In one case, DINGOS was shown capable

of exactly reproducing the template ligand.

Through the work presented here, we have illustrated the capabilities of the DINGOS algorithm. DINGOS provides an experimentally focused computer-assisted *de novo* design method, which, owing to its modular nature, can be tailored to the specific needs of a given drug design project. *De novo* designs were produced for four drug templates, resulting in the synthesis and biotesting of three novel *de novo designs*. A *de novo* active learning cycle was developed, combining DINGOS with a custom continuous flow system, providing the first steps towards fully automated drug design. Having established the DINGOS methods, the underlying algorithm was modified, extending the *de novo* design procedure to any arbitrary descriptor representation and to non-greedy solutions.

Chapter 8

Future Directions

Despite the results presented in this thesis, there are still many interesting places in which the DINGOS algorithm could be extended. In all of the studies considered, the problem of drug design was described in terms of ligand-based scoring. While methods such as these having proven successful in various studies, the analysis performed in Chapter 5 demonstrated that this success cannot be taken for granted. It would be interesting to apply DINGOS to problems other than solely ligand-based scoring. Instead of scoring based on similarity, DINGOS could employ, due to its modularity, alternative scoring functions, including advanced crystallographic or hydrogen bonding models, or predictive models designed to estimate compound bioactivity or ADMET properties (Absorption, Distribution, Metabolism, Excretion, and Toxicity).

In addition to modifying the scoring function, it would also be of interest to further develop the synthetic component. Currently, synthetisizability is controlled by a set number of rules. While this does provide direct control over the design's structure, it imposes a developmental bottleneck, as improving and extending the reaction database requires the laborious task of encoding all desired "synthetic rules". Recent advancements in the field of computer-assisted retro-synthetic analysis and feasibility prediction may offer a solution. Equally, it would be interesting to see out how these predictive reaction components compare to the more explicit rule-based elements used in this work, and what their influence is on the overall structure of the designs.

The work presented in Chapter 6 represents a preliminary implementation and parameterization of the Monte Carlo tree search algorithm. In order to fully come to terms with the potential capabilities of this method, further studies investigating the influence of the various control parameters are required. A key element that was not realised in this thesis was that of adaptive self-improvement. Both AlphaGo and its successor AlphaZero achieved superhuman performance by using the results of previous playthroughs to update their respective machine learning components. Despite the potential benefits, implementing a strategy such as this is encumbered with its own technical considerations. It would be of interest to develop such a self-learning procedure within DINGOS-MCTS.

Bibliography

1. Reymond, J.-L. & Awale, M. Exploring chemical space for drug discovery using the chemical universe database. *ACS chemical neuroscience* **3.** 23019491[pmid], 649–657. ISSN: 1948-7193. https://www.ncbi.nlm.nih.gov/pubmed/23019491 (2012).
2. Reymond, J.-L., van Deursen, R., Blum, L. C. & Ruddigkeit, L. Chemical space as a source for new drugs. *Med. Chem. Commun.* **1,** 30–38. http://dx.doi.org/10.1039/C0MD00020E (1 2010).
3. Drew, K., Baiman, H., Khwaounjoo, P., Yu, B. & Reynisson, J. Size estimation of chemical space: How big is it? *The Journal of pharmacy and pharmacology* **64,** 490–5 (Apr. 2012).
4. Fink, T., Dr, H. & Reymond, J.-L. Virtual Exploration of the Small-Molecule Chemical Universe below 160 Daltons. *Angewandte Chemie International Edition* **44,** 1504 –1508 (Feb. 2005).
5. Renner, S. *et al.* Recent trends and observations in the design of high-quality screening collections. *Future Medicinal Chemistry* **3.** PMID: 21554080, 751–766. eprint: https://doi.org/10.4155/fmc.11.15. https://doi.org/10.4155/fmc.11.15 (2011).
6. Hughes, J. P., Rees, S., Kalindjian, S. B. & Philpott, K. L. Principles of early drug discovery. *British journal of pharmacology* **162.** 21091654[pmid], 1239–1249. ISSN: 1476-5381. https://www.ncbi.nlm.nih.gov/pubmed/21091654 (2011).
7. Drews, J. Drug Discovery: A Historical Perspective. *Science* **287,** 1960–1964. ISSN: 0036-8075. eprint: https://science.sciencemag.org/content/287/5460/1960.full.pdf. https://science.sciencemag.org/content/287/5460/1960 (2000).
8. Congreve, M., Murray, C. W. & Blundell, T. L. Keynote review: Structural biology and drug discovery. *Drug Discovery Today* **10,** 895 –907. ISSN: 1359-6446. http://www.sciencedirect.com/science/article/pii/S1359644605034847 (2005).
9. Cheng, T., Li, Q., Zhou, Z., Wang, Y. & Bryant, S. H. Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review. *The AAPS Journal* **14,** 133–141. ISSN: 1550-7416. https://doi.org/10.1208/s12248-012-9322-0 (2012).
10. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *Journal of Natural Products* **79.** PMID: 26852623, 629–661. eprint: https://doi.org/10.1021/acs.jnatprod.5b01055. https://doi.org/10.1021/acs.jnatprod.5b01055 (2016).
11. Shen, B. A New Golden Age of Natural Products Drug Discovery. *Cell* **163,** 1297 –1300. ISSN: 0092-8674. http://www.sciencedirect.com/science/article/pii/S0092867415015500 (2015).

12. Desborough, M. J. R. & Keeling, D. M. The aspirin story - from willow to wonder drug. *British Journal of Haematology* **177,** 674–683. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjh.14520. https://onlinelibrary.wiley.com/doi/abs/10.1111/bjh.14520 (2017).

13. Wani, M. C., Taylor, H. L., Wall, M. E., Coggon, P. & McPhail, A. T. Plant antitumor agents. VI. Isolation and structure of taxol, a novel antileukemic and antitumor agent from Taxus brevifolia. *Journal of the American Chemical Society* **93.** PMID: 5553076, 2325–2327. eprint: https://doi.org/10.1021/ja00738a045. https://doi.org/10.1021/ja00738a045 (1971).

14. De Bree, E. *et al.* Treatment of ovarian cancer using intraperitoneal chemotherapy with taxanes: From laboratory bench to bedside. *Cancer Treatment Reviews* **32,** 471 –482. ISSN: 0305-7372. http://www.sciencedirect.com/science/article/pii/S030573720600137X (2006).

15. Yvon, A. M., Wadsworth, P. & Jordan, M. A. Taxol suppresses dynamics of individual microtubules in living human tumor cells. *Molecular biology of the cell* **10.** 10198049[pmid], 947–959. ISSN: 1059-1524. https://www.ncbi.nlm.nih.gov/pubmed/10198049 (1999).

16. Harvey, A. L., Edrada-Ebel, R. & Quinn, R. J. The re-emergence of natural products for drug discovery in the genomics era. *Nature Reviews Drug Discovery* **14,** 111–129. ISSN: 1474-1784. https://doi.org/10.1038/nrd4510 (2015).

17. Zhao, C. *et al.* A general strategy for diversifying complex natural products to polycyclic scaffolds with medium-sized rings. *Nature Communications* **10,** 4015. ISSN: 2041-1723. https://doi.org/10.1038/s41467-019-11976-2 (2019).

18. Holton, R. A. *et al.* First total synthesis of taxol. 1. Functionalization of the B ring. *Journal of the American Chemical Society* **116,** 1597–1598. eprint: https://doi.org/10.1021/ja00083a066. https://doi.org/10.1021/ja00083a066 (1994).

19. Nicolaou, K. C. *et al.* Total synthesis of taxol. *Nature* **367,** 630–634. ISSN: 1476-4687. https://doi.org/10.1038/367630a0 (1994).

20. Clark, A. M. Natural Products as a Resource for New Drugs. *Pharmaceutical Research* **13,** 1133–1141. ISSN: 1573-904X. https://doi.org/10.1023/A:1016091631721 (1996).

21. Li, G. & Lou, H.-X. Strategies to diversify natural products for drug discovery. *Medicinal Research Reviews* **38,** 1255–1294. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/med.21474. https://onlinelibrary.wiley.com/doi/abs/10.1002/med.21474 (2018).

22. Guo, Z. The modification of natural products for medical use. *Acta Pharmaceutica Sinica B* **7.** Traditional Chinese medicine and natural products research, 119 –136. ISSN: 2211-3835. http://www.sciencedirect.com/science/article/pii/S2211383516300466 (2017).

23. Janzen, W. Screening Technologies for Small Molecule Discovery: The State of the Art. *Chemistry and Biology* **21,** 1162 –1170. ISSN: 1074-5521. http://www.sciencedirect.com/science/article/pii/S1074552114002440 (2014).

24. Brenk, R. *et al.* Lessons Learnt from Assembling Screening Libraries for Drug Discovery for Neglected Diseases. *ChemMedChem* **3,** 435–444. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cmdc.200700139. https://onlinelibrary.wiley.com/doi/abs/10.1002/cmdc.200700139 (2008).

25. Jones, L. H. & Bunnage, M. E. Applications of chemogenomic library screening in drug discovery. *Nature Reviews Drug Discovery* **16,** 285–296. ISSN: 1474-1784. https://doi.org/10.1038/nrd.2016.244 (2017).

26. Sun, W. *et al.* Rapid identification of antifungal compounds against Exserohilum rostratum using high throughput drug repurposing screens. *PloS one* **8.** 23990907[pmid], e70506–e70506. ISSN: 1932-6203. https://www.ncbi.nlm.nih.gov/pubmed/23990907 (2013).

27. Wang, X.-F. *et al.* N-Aryl-6-methoxy-1,2,3,4-tetrahydroquinolines: A novel class of antitumor agents targeting the colchicine site on tubulin. *European Journal of Medicinal Chemistry* **67,** 196 –207. ISSN: 0223-5234. http://www.sciencedirect.com/science/article/pii/S0223523413004091 (2013).

28. Wang, X.-F. *et al.* N-Aryl-6-methoxy-1,2,3,4-tetrahydroquinolines: a Novel Class of Antitumor Agents Targeting the Colchicine Site on Tubulin. *European journal of medicinal chemistry* **67C,** 196–207 (June 2013).

29. Wang, T. C., Cheng, L. P., Huang, X. Y., Zhao, L. & Pang, W. Identification of potential tubulin polymerization inhibitors by 3D-QSAR molecular docking and molecular dynamics. *RSC Adv.* **7,** 38479–38489. http://dx.doi.org/10.1039/C7RA04314G (61 2017).

30. Hartenfeller, M. & Schneider, G. Enabling future drug discovery by de novo design. *WIREs Computational Molecular Science* **1,** 742–759. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.49. https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.49 (2011).

31. Putin, E. *et al.* Reinforced Adversarial Neural Computer for de Novo Molecular Design. *Journal of Chemical Information and Modeling* **58.** PMID: 29762023, 1194–1204. eprint: https://doi.org/10.1021/acs.jcim.7b00690. https://doi.org/10.1021/acs.jcim.7b00690 (2018).

32. Spänkuch, B. *et al.* Drugs by Numbers: Reaction-Driven De Novo Design of Potent and Selective Anticancer Leads. *Angewandte Chemie International Edition* **52,** 4676–4681. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201206897. https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201206897 (2013).

33. Cunningham, J. M., Koytiger, G., Sorger, P. K. & AlQuraishi, M. Biophysical prediction of protein-peptide interactions and signaling networks using machine learning. *Nature Methods* **17,** 175–183. ISSN: 1548-7105. https://doi.org/10.1038/s41592-019-0687-1 (2020).

34. Fleetwood, O., Kasimova, M. A., Westerlund, A. M. & Delemotte, L. Molecular Insights from Conformational Ensembles via Machine Learning. *Biophysical Journal.* ISSN: 0006-3495. http://www.sciencedirect.com/science/article/pii/S0006349519344017 (2019).

35. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Delta-Machine Learning Approach. *Journal of Chemical Theory and Computation* **11.** PMID: 26574412, 2087–2096. eprint: https://doi.org/10.1021/acs.jctc.5b00099. https://doi.org/10.1021/acs.jctc.5b00099 (2015).

36. Owens, J. *et al.* GPU computing. *Proceedings of the IEEE* **96,** 879–899 (May 2008).

37. Gawehn, E., Hiss, J. A., Brown, J. B. & Schneider, G. Advancing drug discovery via GPU-based deep learning. *Expert Opinion on Drug Discovery* **13.** PMID: 29668343, 579–582. eprint: https://doi.org/10.1080/17460441.2018.1465407. https://doi.org/10.1080/17460441.2018.1465407 (2018).

38. Roberto, T. & Viviana, C. *Handbook of Molecular Descriptors* (2000).

39. Roberto, T. & Viviana, C. *Molecular Descriptors for Chemoinformatics* (2009).

40. Danishuddin & Khan, A. U. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discovery Today* **21,** 1291 –1302. ISSN: 1359-6446. http://www.sciencedirect.com/science/article/pii/S1359644616302318 (2016).

41. Wang, J. *et al.* Development of Reliable Aqueous Solubility Models and Their Application in Druglike Analysis. *Journal of Chemical Information and Modeling* **47.** PMID: 17569522, 1395–1404. eprint: https://doi.org/10.1021/ci700096r. https://doi.org/10.1021/ci700096r (2007).

42. Schulz, K.-P. M. Johnson, G. Maggiora (Eds.): Concepts and Applications of Molecular Similarity, Wiley Interscience, New York, Chichester 1990. ISBN 0-471175-7, 393 Seiten. Preis: 51.35. *Berichte der Bunsengesellschaft für physikalische Chemie* **96,** 1087–1087. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bbpc.19920960825. https://onlinelibrary.wiley.com/doi/abs/10.1002/bbpc.19920960825 (1992).

43. Cereto-Massagué, A. *et al.* Molecular fingerprint similarity search in virtual screening. *Methods* **71.** Virtual Screening, 58 –63. ISSN: 1046-2023. http://www.sciencedirect.com/science/article/pii/S1046202314002631 (2015).

44. Riniker, S. & Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics* **5,** 26. ISSN: 1758-2946. https://doi.org/10.1186/1758-2946-5-26 (2013).

45. Skinnider, M. A., Dejong, C. A., Franczak, B. C., McNicholas, P. D. & Magarvey, N. A. Comparative analysis of chemical similarity methods for modular natural products with a hypothetical structure enumeration algorithm. *Journal of Cheminformatics* **9,** 46. ISSN: 1758-2946. https://doi.org/10.1186/s13321-017-0234-y (2017).

46. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **28,** 31–36. eprint: https://pubs.acs.org/doi/pdf/10.1021/ci00057a005. https://pubs.acs.org/doi/abs/10.1021/ci00057a005 (1988).

47. *RDKit: Open source cheminformatics* http://www.rdkit.org/.

48. Merk, D., Friedrich, L., Grisoni, F. & Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Molecular Informatics* **37,** 1700153. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/minf.201700153. https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201700153 (2018).

49. Arús-Pous, J. *et al.* Randomized SMILES strings improve the quality of molecular generative models. *Journal of Cheminformatics* **11,** 71. ISSN: 1758-2946. https://doi.org/10.1186/s13321-019-0393-0 (2019).

50. Hinton, G. E. & Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science* **313,** 504–507. ISSN: 0036-8075. eprint: https://science.sciencemag.org/content/313/5786/504.full.pdf. https://science.sciencemag.org/content/313/5786/504 (2006).

51. Salakhutdinov, R., Mnih, A. & Hinton, G. *Restricted Boltzmann Machines for Collaborative Filtering* in *Proceedings of the 24th International Conference on Machine Learning* (Association for Computing Machinery, Corvalis, Oregon, USA, 2007), 791–798. ISBN: 9781595937933. https://doi.org/10.1145/1273496.1273596.

52. Srivastava, R. K., Greff, K. & Schmidhuber, J. in *Advances in Neural Information Processing Systems 28* (eds Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M. & Garnett, R.) 2377–2385 (Curran Associates, Inc., 2015). http://papers.nips.cc/paper/5850-training-very-deep-networks.pdf.

53. Hoashi, H., Joutou, T. & Yanai, K. *Image Recognition of 85 Food Categories by Feature Fusion* in *2010 IEEE International Symposium on Multimedia* (2010), 296–301.

54. Meiyin Wu & Li Chen. *Image recognition based on deep learning* in *2015 Chinese Automation Congress (CAC)* (2015), 542–546.

55. Collobert, R. & Weston, J. *A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning* in *Proceedings of the 25th International Conference on Machine Learning* (Association for Computing Machinery, Helsinki, Finland, 2008), 160–167. ISBN: 9781605582054. https://doi.org/10.1145/1390156.1390177.

56. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529,** 484–489. ISSN: 1476-4687. https://doi.org/10.1038/nature16961 (2016).

57. Silver, D. *et al.* Mastering the game of Go without human knowledge. *Nature* **550,** 354–359. ISSN: 1476-4687. https://doi.org/10.1038/nature24270 (2017).

58. Yao, J., Lin, C., Xie, X., Wang, A. J. & Hung, C. *Path Planning for Virtual Human Motion Using Improved A* Star Algorithm* in *2010 Seventh International Conference on Information Technology: New Generations* (2010), 1154–1158.

59. Wegner, L. M. Quicksort for Equal Keys. *IEEE Transactions on Computers* **C-34,** 362–367. ISSN: 2326-3814 (1985).

60. Park, D. C., El-Sharkawi, M. A., Marks, R. J., Atlas, L. E. & Damborg, M. J. Electric load forecasting using an artificial neural network. *IEEE Transactions on Power Systems* **6,** 442–449. ISSN: 1558-0679 (1991).

61. Nayak, R., Jain, L. & Ting, B. in *Computational Mechanics–New Frontiers for the New Millennium* (eds Valliappan, S. & Khalili, N.) 887 –892 (Elsevier, Oxford, 2001). ISBN: 978-0-08-043981-5. http://www.sciencedirect.com/science/article/pii/B9780080439815501322.

62. Kim, T.-h. *Pattern Recognition Using Artificial Neural Network: A Review* in *Information Security and Assurance* (eds Bandyopadhyay, S. K., Adi, W., Kim, T.-h. & Xiao, Y.) (Springer Berlin Heidelberg, Berlin, Heidelberg, 2010), 138–148. ISBN: 978-3-642-13365-7.

63. Rosenblatt & Frank. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review* **65,** 386– (Jan. 1958).

64. Fan, F., Cong, W. & Wang, G. A new type of neurons for machine learning. *International Journal for Numerical Methods in Biomedical Engineering* **34.** e2920 CNM-Apr-17-0102.R1, e2920. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cnm.2920. https://onlinelibrary.wiley.com/doi/abs/10.1002/cnm.2920 (2018).

65. Lippmann, R. An introduction to computing with neural nets. *IEEE ASSP Magazine* **4,** 4–22. ISSN: 1558-1284 (1987).

66. Block, H. A review of "perceptrons: An introduction to computational geometry. *Information and Control* **17,** 501 –522. ISSN: 0019-9958. http://www.sciencedirect.com/science/article/pii/S0019995870904092 (1970).

67. Gori, M. in *Machine Learning* (ed Gori, M.) 236 –338 (Morgan Kaufmann, 2018). ISBN: 978-0-08-100659-7. http://www.sciencedirect.com/science/article/pii/B9780081006597000051.

68. Mikolov, T., Kombrink, S., Burget, L., Černocký, J. & Khudanpur, S. *Extensions of recurrent neural network language model* in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2011), 5528–5531.

69. Lawrence, S., Giles, C. L., Ah Chung Tsoi & Back, A. D. Face recognition: a convolutional neural-network approach. *IEEE Transactions on Neural Networks* **8,** 98–113. ISSN: 1941-0093 (1997).

70. Bahi, M. & Batouche, M. *Deep Learning for Ligand-Based Virtual Screening in Drug Discovery* in (Oct. 2018), 1–5.

71. Gao, W. & Zhou, Z.-H. On the Consistency of Multi-Label Learning. *Journal of Machine Learning Research - Proceedings Track* **19,** 341–358 (Jan. 2011).

72. Zhang, Z. & Sabuncu, M. in *Advances in Neural Information Processing Systems 31* (eds Bengio, S. *et al.*) 8778–8788 (Curran Associates, Inc., 2018). http://papers.nips.cc/paper/8094-generalized-cross-entropy-loss-for-training-deep-neural-networks-with-noisy-labels.pdf.

73. West, N. E. & O'Shea, T. *Deep architectures for modulation recognition* in *2017 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)* (2017), 1–6.

74. Rosasco, L., Vito, E. D., Caponnetto, A., Piana, M. & Verri, A. Are Loss Functions All the Same? *Neural Computation* **16,** 1063–1076. eprint: https://doi.org/10.1162/089976604773135104. https://doi.org/10.1162/089976604773135104 (2004).

75. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. **14** (Mar. 2001).

76. Baeuerle, N. & Rieder, U. Markov Decision Processes. *Jahresbericht der Deutschen Mathematiker-Vereinigung* **112,** 217–243 (Dec. 2010).

77. Badowski, T., Gajewska, E. P., Molga, K. & Grzybowski, B. A. Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning. *Angewandte Chemie International Edition* **59,** 725–730. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201912083. https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201912083 (2020).

78. Baylon, J. L., Cilfone, N. A., Gulcher, J. R. & Chittenden, T. W. Enhancing Retrosynthetic Reaction Prediction with Deep Learning Using Multiscale Reaction Classification. *Journal of Chemical Information and Modeling* **59.** PMID: 30642173, 673–688. eprint: https://doi.org/10.1021/acs.jcim.8b00801. https://doi.org/10.1021/acs.jcim.8b00801 (2019).

79. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555,** 604–610. ISSN: 1476-4687. https://doi.org/10.1038/nature25978 (2018).

80. Merk, D., Friedrich, L., Grisoni, F. & Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Molecular Informatics* **37,** 1700153. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/minf.201700153. https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201700153 (2018).

81. EMBL-EBI, E. M. B. L. *ChEMBL* https://www.ebi.ac.uk/chembl/ (2020).

82. Button, A., Merk, D., Hiss, J. A. & Schneider, G. Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis. *Nature Machine Intelligence* **1,** 307–315. https://doi.org/10.1038/s42256-019-0067-7 (2019).

83. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. *Drug Discovery Today* **23,** 1241 –1250. ISSN: 1359-6446. http://www.sciencedirect.com/science/article/pii/S1359644617303598 (2018).

84. Gupta, A. *et al.* Generative Recurrent Networks for De Novo Drug Design. *Molecular Informatics* **37,** 1700111. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/minf.201700111. https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201700111 (2018).

85. Hamza, A., Wei, N.-N. & Zhan, C.-G. Ligand Based Virtual Screening Approach Using a New Scoring Function. *Journal of Chemical Information and Modeling* **52.** PMID: 22486340, 963–974. eprint: https://doi.org/10.1021/ci200617d. https://doi.org/10.1021/ci200617d (2012).

86. Roy, K., Kar, S. & Das, R. N. in *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment* (eds Roy, K., Kar, S. & Das, R. N.) 47 –80 (Academic Press, Boston, 2015). ISBN: 978-0-12-801505-6.

87. Kausar, S. & Falcao, A. O. Analysis and Comparison of Vector Space and Metric Space Representations in QSAR Modeling. *Molecules (Basel, Switzerland)* **24.** 31052325[pmid], 1698. ISSN: 1420-3049. https://www.ncbi.nlm.nih.gov/pubmed/31052325 (2019).

88. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **28,** 31–36. eprint: https://pubs.acs.org/doi/pdf/10.1021/ci00057a005. https://pubs.acs.org/doi/abs/10.1021/ci00057a005 (1988).

89. Grisoni, F. *et al.* Scaffold hopping from natural products to synthetic mimetics by holistic molecular similarity. *Communications Chemistry* **1,** 44. ISSN: 2399-3669. https://doi.org/10.1038/s42004-018-0043-x (2018).

90. Merk, D., Grisoni, F., Friedrich, L., Gelzinyte, E. & Schneider, G. Scaffold hopping from synthetic RXR modulators by virtual screening and de novo design. *Med. Chem. Commun.* **9,** 1289–1292. http://dx.doi.org/10.1039/C8MD00134K (8 2018).

91. Grisoni, F., Merk, D., Byrne, R. & Schneider, G. Scaffold-Hopping from Synthetic Drugs by Holistic Molecular Representation. *Scientific Reports* **8,** 16469. ISSN: 2045-2322. https://doi.org/10.1038/s41598-018-34677-0 (2018).

92. Bebis, G. & Georgiopoulos, M. Feed-forward neural networks. *IEEE Potentials* **13,** 27–31. ISSN: 1558-1772 (1994).

93. Engkvist, O. *et al.* Computational prediction of chemical reactions: current status and outlook. *Drug Discovery Today* **23,** 1203 –1218. ISSN: 1359-6446. http://www.sciencedirect.com/science/article/pii/S1359644617305068 (2018).

94. Socorro, I. M., Taylor, K. & Goodman, J. M. ROBIA A Reaction Prediction Program. *Organic Letters* **7.** PMID: 16048337, 3541–3544. eprint: https://doi.org/10.1021/ol0512738. https://doi.org/10.1021/ol0512738 (2005).

95. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Computer Sciences* **42.** PMID: 12444722, 1273–1280. eprint: https://doi.org/10.1021/ci010132r. https://doi.org/10.1021/ci010132r (2002).

96. Bellman, R. *Dynamic Programming* ISBN: 9780486428093. https://books.google.ch/books?id=fyVtp3EMxasC (Dover Publications, 2003).

97. Bellman, R. & Collection, K. M. R. *Adaptive Control Processes: A Guided Tour* https://books.google.ch/books?id=POAmAAAAMAAJ (Princeton University Press, 1961).

98. *Molecular Operating Environment* 2017. https://www.chemcomp.com/.

99. Reker, D., Rodrigues, T., Schneider, P. & Schneider, G. Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proceedings of the National Academy of Sciences* **111,** 4067–4072. ISSN: 0027-8424. eprint: https://www.pnas.org/content/111/11/4067.full.pdf. https://www.pnas.org/content/111/11/4067 (2014).

100. *Reaxys (Elsevier).* 2018. https://www.reaxys.com/.

101. Lowe, D. Chemical reactions from US patents (1976-Sep2016). https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873 (2017).

102. Murphy, K. *Machine Learning: A Probabilistic Perspective* ISBN: 9780262304320. https://books.google.ch/books?id=RC43AgAAQBAJ (MIT Press, 2012).

103. Kingma, D. P. & Ba, J. *Adam: A Method for Stochastic Optimization* 2014. arXiv: 1412.6980 [cs.LG].

104. Institute, E. B. ChEMBL Database 2017. https://www.ebi.ac.uk/chembl/ (2017).

105. Maggiora, G. M. On Outliers and Activity CliffsWhy QSAR Often Disappoints. *Journal of Chemical Information and Modeling* **46.** PMID: 16859285, 1535–1535. eprint: https://doi.org/10.1021/ci060117s. https://doi.org/10.1021/ci060117s (2006).

106. Sakamoto, H. *et al.* CH5424802, a Selective ALK Inhibitor Capable of Blocking the Resistant Gatekeeper Mutant. *Cancer Cell* **19,** 679 –690. ISSN: 1535-6108. http://www.sciencedirect.com/science/article/pii/S1535610811001486 (2011).

107. Kiss, B. *et al.* Cariprazine (RGH-188), a Dopamine D3 Receptor-Preferring, D3/D2 Dopamine Receptor Antagonist–Partial Agonist Antipsychotic Candidate: In Vitro and Neurochemical Profile. *Journal of Pharmacology and Experimental Therapeutics* **333,** 328–340. ISSN: 0022-3565. eprint: http://jpet.aspetjournals.org/content/333/1/328.full.pdf. http://jpet.aspetjournals.org/content/333/1/328 (2010).

108. Chu, C.-Y., Choi, J., Eaby-Sandy, B., Langer, C. J. & Lacouture, M. E. Osimertinib: A Novel Dermatologic Adverse Event Profile in Patients with Lung Cancer. *The Oncologist* **23,** 891–899. eprint: https://theoncologist.onlinelibrary.wiley.com/doi/pdf/10.1634/theoncologist.2017-0582. https://theoncologist.onlinelibrary.wiley.com/doi/abs/10.1634/theoncologist.2017-0582 (2018).

109. Touma, K. & Touma, D. Pimavanserin (Nuplazid) for the treatment of Parkinson disease psychosis: A review of the literature. *Mental Health Clinician* **7,** 230–234 (Sept. 2017).

110. Yang, X., Wang, Y., Byrne, R., Schneider, G. & Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chemical Reviews* **119.** PMID: 31294972, 10520–10594. eprint: https://doi.org/10.1021/acs.chemrev.8b00728. https://doi.org/10.1021/acs.chemrev.8b00728 (2019).

111. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3-25.1. *Advanced Drug Delivery Reviews* **46.** Special issue dedicated to Dr. Eric Tomlinson, Advanced Drug Delivery Reviews, A Selection of the Most Highly Cited Articles, 1991-1998, 3 –26. ISSN: 0169-409X. http://www.sciencedirect.com/science/article/pii/S0169409X00001290 (2001).

112. Friedrich, L., Rodrigues, T., Neuhaus, C. S., Schneider, P. & Schneider, G. From Complex Natural Products to Simple Synthetic Mimetics by Computational De Novo Design. *Angewandte Chemie International Edition* **55,** 6789–6792. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201601941. https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201601941 (2016).

113. Schneider, P, Röthlisberger, M, Reker, D. & Schneider, G. Spotting and designing promiscuous ligands for drug discovery. *Chemical communications (Cambridge, England)* **52** (Nov. 2015).

114. Friedrich, L., Rodrigues, T., Neuhaus, C., Schneider, P. & Schneider, G. From Complex Natural Products to Simple Synthetic Mimetics by Computational De Novo Design. *Angewandte Chemie International Edition in English* **55** (Apr. 2016).

115. Reutlinger, M. *et al.* Chemically Advanced Template Search (CATS) for Scaffold-Hopping and Prospective Target Prediction for "Orphan' Molecules. *Molecular informatics* **32,** 133–138 (Feb. 2013).

116. Wishart DS Knox C, G. A. S. S. H. M. S. P. C. Z. W. J. *Drugbank: a comprehensive resource for in silico drug discovery and exploration. 34 (Database issue):D668-72. 16381955* `https://www.drugbank.ca/` (2020).

117. Friedel, H. A. & Buckley, M. M. T. Torasemide. *Drugs* **41,** 81–103. ISSN: 1179-1950. `https://doi.org/10.2165/00003495-199141010-00008` (1991).

118. Benej, M. *et al.* Papaverine and its derivatives radiosensitize solid tumors by inhibiting mitochondrial metabolism. *Proceedings of the National Academy of Sciences* **115,** 10756–10761. ISSN: 0027-8424. eprint: `https://www.pnas.org/content/115/42/10756.full.pdf`. `https://www.pnas.org/content/115/42/10756` (2018).

119. Chobanian, A. V., Haudenschild, C. C., Nickerson, C & Hope, S. Trandolapril inhibits atherosclerosis in the Watanabe heritable hyperlipidemic rabbit. *Hypertension* **20,** 473–477 (1992).

120. Plutschack, M. B., Pieber, B., Gilmore, K. & Seeberger, P. H. The Hitchhiker's Guide to Flow Chemistry. *Chemical Reviews* **117.** PMID: 28570059, 11796–11893. eprint: `https://doi.org/10.1021/acs.chemrev.7b00183`. `https://doi.org/10.1021/acs.chemrev.7b00183` (2017).

121. Cohn, D., Atlas, L. & Ladner, R. Improving generalization with active learning. *Machine Learning* **15,** 201–221. ISSN: 1573-0565. `https://doi.org/10.1007/BF00993277` (1994).

122. Cohn, D., Ghahramani, Z. & Jordan, M. *Active Learning with Statistical Models* in. **4** (Feb. 1996), 705–712.

123. Siddiquie, B. & Gupta, A. *Beyond active noun tagging: Modeling contextual interactions for multi-class active learning* in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2010), 2979–2986.

124. Sivaraman, S. & Trivedi, M. M. A General Active-Learning Framework for On-Road Vehicle Recognition and Tracking. *IEEE Transactions on Intelligent Transportation Systems* **11,** 267–276 (2010).

125. Papalia, G. A. *et al.* Comparative analysis of 10 small molecules binding to carbonic anhydrase II by different investigators using Biacore technology. *Analytical Biochemistry* **359,** 94 –105. ISSN: 0003-2697. `http://www.sciencedirect.com/science/article/pii/S000326970600621X` (2006).

126. *ZINC15* `https://zinc.docking.org/`.

127. *Enamine* `https://enamine.net/`.

128. Jing, L. & Li, J.-X. Trace amine-associated receptor 1: A promising target for the treatment of psychostimulant addiction. *European journal of pharmacology* **761.** 26092759[pmid], 345–352. ISSN: 1879-0712. `https://www.ncbi.nlm.nih.gov/pubmed/26092759` (2015).

129. Suveges, N. S., de Souza, R. O. M. A., Gutmann, B. & Kappe, C. O. Synthesis of Mepivacaine and Its Analogues by a Continuous-Flow Tandem Hydrogenation/Reductive Amination Strategy. *European Journal of Organic Chemistry* **2017,** 6511–6517. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejoc.201700824. https://onlinelibrary.wiley.com/doi/abs/10.1002/ejoc.201700824 (2017).

130. Laroche, B., Ishitani, H. & Kobayashi, S. Direct Reductive Amination of Carbonyl Compounds with H2 Using Heterogeneous Catalysts in Continuous Flow as an Alternative to N-Alkylation with Alkyl Halides. *Advanced Synthesis & Catalysis* **360,** 4699–4704. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adsc.201801457. https://onlinelibrary.wiley.com/doi/abs/10.1002/adsc.201801457 (2018).

131. Soloshonok, V. A. & Ono, T. First example of continuous-flow reaction conditions for biomimetic reductive amination of fluorine-containing carbonyl compounds. *Journal of Fluorine Chemistry* **129.** Fluorine in Biomedicinal Chemistry, 785 –787. ISSN: 0022-1139. http://www.sciencedirect.com/science/article/pii/S0022113908001498 (2008).

132. Ma, T., Zhang, H.-Y., Yin, G., Zhao, J. & Zhang, Y. Catalyst-free reductive amination of levulinic acid to N-substituted pyrrolidinones with formic acid in continuous-flow microreactor. *Journal of Flow Chemistry* **8,** 35–43. ISSN: 2063-0212. https://doi.org/10.1007/s41981-018-0005-6 (2018).

133. Dragone, V., Sans, V., Rosnes, M., Kitson, P. & Cronin, L. 3D-printed devices for continuous-flow organic chemistry. *Beilstein journal of organic chemistry* **9,** 951–9 (May 2013).

134. Singh, B. K., Stevens, C. V., Acke, D. R., Parmar, V. S. & der Eycken], E. V. V. Copper-mediated N- and O-arylations with arylboronic acids in a continuous flow microreactor: a new avenue for efficient scalability. *Tetrahedron Letters* **50,** 15 –18. ISSN: 0040-4039. http://www.sciencedirect.com/science/article/pii/S0040403908018431 (2009).

135. Supuran, C. T. Carbonic anhydrases: novel therapeutic applications for inhibitors and activators. *Nature Reviews Drug Discovery* **7,** 168–181. ISSN: 1474-1784. https://doi.org/10.1038/nrd2467 (2008).

136. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Science Advances* **4.** eprint: https://advances.sciencemag.org/content/4/7/eaap7885.full.pdf. https://advances.sciencemag.org/content/4/7/eaap7885 (2018).

137. Schwaller, P. *et al.* Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science* **5,** 1572–1583. eprint: https://doi.org/10.1021/acscentsci.9b00576. https://doi.org/10.1021/acscentsci.9b00576 (2019).

138. Vaswani, A. *et al.* in *Advances in Neural Information Processing Systems 30* (eds Guyon, I. *et al.*) 5998–6008 (Curran Associates, Inc., 2017). http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

139. Klein, G., Kim, Y., Deng, Y., Senellart, J. & Rush, A. *OpenNMT: Open-Source Toolkit for Neural Machine Translation* in *Proceedings of ACL 2017, System Demonstrations* (Association for Computational Linguistics, Vancouver, Canada, July 2017), 67–72. https://www.aclweb.org/anthology/P17-4012.

140. Bahdanau, D., Cho, K. & Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *ArXiv* **1409** (Sept. 2014).

141. Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **5,** 107–113. eprint: https://doi.org/10.1021/c160017a018. https://doi.org/10.1021/c160017a018 (1965).

142. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **50.** PMID: 20426451, 742–754. eprint: https://doi.org/10.1021/ci100050t. https://doi.org/10.1021/ci100050t (2010).

143. Carhart, R. E., Smith, D. H. & Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences* **25,** 64–73. eprint: https://pubs.acs.org/doi/pdf/10.1021/ci00046a002. https://pubs.acs.org/doi/abs/10.1021/ci00046a002 (1985).

144. Gedeck, P., Rohde, B. & Bartels, C. QSAR - How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *Journal of Chemical Information and Modeling* **46.** PMID: 16995723, 1924–1936. eprint: https://doi.org/10.1021/ci050413p. https://doi.org/10.1021/ci050413p (2006).

145. Nilakantan, R., Bauman, N., Dixon, J. S. & Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *Journal of Chemical Information and Computer Sciences* **27,** 82–85. eprint: https://pubs.acs.org/doi/pdf/10.1021/ci00054a008. https://pubs.acs.org/doi/abs/10.1021/ci00054a008 (1987).

146. Jin, W., Coley, C., Barzilay, R. & Jaakkola, T. in *Advances in Neural Information Processing Systems 30* (eds Guyon, I. *et al.*) 2607–2616 (Curran Associates, Inc., 2017). http://papers.nips.cc/paper/6854-predicting-organic-reaction-outcomes-with-weisfeiler-lehman-network.pdf.

147. *PubChem* https://pubchem.ncbi.nlm.nih.gov/.

148. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics* **7.** 26052348[pmid], 20–20. ISSN: 1758-2946. https://www.ncbi.nlm.nih.gov/pubmed/26052348 (2015).

149. Khanna, V. & Ranganathan, S. Molecular Similarity and Diversity Approaches in Chemoinformatics. *Drug Development Research* **72,** 74 –84 (Feb. 2011).

150. Mini-review. Application of combinatorial library methods in cancer research and drug discovery. *Anti-Cancer Drug Design* **12** (1997).

151. Cerrone, C., Cerulli, R. & Golden, B. Carousel Greedy: A Generalized Greedy Algorithm with Applications in Optimization. *Computers and Operations Research* **85** (Apr. 2017).

152. Nair, A., McGrew, B., Andrychowicz, M., Zaremba, W. & Abbeel, P. *Overcoming Exploration in Reinforcement Learning with Demonstrations* in *2018 IEEE International Conference on Robotics and Automation (ICRA)* (2018), 6292–6299.

153. Thomas, G., Chien, M., Tamar, A., Ojea, J. A. & Abbeel, P. *Learning Robotic Assembly from CAD* in *2018 IEEE International Conference on Robotics and Automation (ICRA)* (2018), 3524–3531.

154. Andrychowicz, M. *et al.* in *Advances in Neural Information Processing Systems 30* (eds Guyon, I. *et al.*) 5048–5058 (Curran Associates, Inc., 2017). `http://papers.nips.cc/paper/7090-hindsight-experience-replay.pdf`.

155. Coulom, R. *Efficient Selectivity and Backup Operators in Monte-Carlo Tree Search* in *Computers and Games* (eds van den Herik, H. J., Ciancarini, P. & Donkers, H. H. L. M. J.) (Springer Berlin Heidelberg, Berlin, Heidelberg, 2007), 72–83. ISBN: 978-3-540-75538-8.

156. Hansen, E., Zhou, R. & Feng, Z. *Symbolic Heuristic Search Using Decision Diagrams* in. **2371** (Aug. 2002), 83–98.

157. Kearns, M., Mansour, Y. & Ng, A. A Sparse Sampling Algorithm for Near-Optimal Planning in Large Markov Decision Processes. *Machine Learning* **49** (June 2001).

158. Weijun Zhu, Qinglei Zhou & Linfeng Jiao. *An algorithm for searching states of game of Go based on symbolic model checking* in *2016 2nd IEEE International Conference on Computer and Communications (ICCC)* (2016), 1201–1205.

159. Li, F. & Du, Y. From AlphaGo to Power System AI: What Engineers Can Learn from Solving the Most Complex Board Game. *IEEE Power and Energy Magazine* **16,** 76–84 (2018).

160. Lei, Q. *et al.* YLT-11, a novel PLK4 inhibitor, inhibits human breast cancer growth via inducing maladjusted centriole duplication and mitotic defect. *Cell Death & Disease* **9,** 1066. ISSN: 2041-4889. `https://doi.org/10.1038/s41419-018-1071-2` (2018).

161. Holland, A. J. & Cleveland, D. W. Polo-like kinase 4 inhibition: a strategy for cancer therapy? *Cancer cell* **26.** 25117704[pmid], 151–153. ISSN: 1878-3686. `https://www.ncbi.nlm.nih.gov/pubmed/25117704` (2014).

162. Liu, Z. *et al.* Synthesis and biological evaluation of (E)-4-(3-arylvinyl-1H-indazol-6-yl)pyrimidin-2-amine derivatives as PLK4 inhibitors for the treatment of breast cancer. *RSC Adv.* **7,** 27737–27746. `http://dx.doi.org/10.1039/C7RA02518A` (44 2017).

163. Sampson, P. *et al.* The Discovery of Polo-Like Kinase 4 Inhibitors: Design and Optimization of Spiro[cyclopropane-1,3[3 H]indol]-2(1 H)-ones as Orally Bioavailable Antitumor Agents. *Journal of medicinal chemistry* **58** (May 2014).

Appendix A

Supplementary Information

A.1 DINGOS Code Availability

The code of the original DINGOS method from Chapter 3, as well as the trained multilayer-perceptron model, CAS number of the training data and reaction SMARTS has been published in Button *et al.* [82] and is provided in the Code Ocean capsule https://doi.org/10.24433/CO.6930970.v1.

The modified version of the DINGOS code shown in Chapter 5 and Chapter 6 will be released alongside the subsequent publications.

A.2 Chapter 3

A.2.1 Reaction Set Used in Chapter 3

Reaction Name	Reaction SMARTS|
Bischler-Napieralski	[$(C([CH2,CH3])),CH:10](=[O:11])−[NH+0:9]−[C$(C(N)(C)(C)(C)),C$([CH](N)(C)(C)),C$([CH2](N)(C)):8]−[C$(C(c)(C)(C)(C)),C$([CH](c)(C)(C)),C$([CH2](c)(C)):7]−[c:6]1[cH:1][c:2][c:3][c:4][c:5]1»[C:10]−1=[N+0:9]−[C:8]−[C:7]−[c:6]2[c:5][c:4][c:3][c:2][c:1]−12
Pictet-Gams	[$(C([CH2,CH3])),CH:10](=[O:11])−[NH+0:9]−[C$([CH](N)(C)(C)),C$([CH2](N)(C)):8]−[C$([C](c)(C)(C)),C$([CH](c)(C)):7]([O$(OC),OH])−[c:6]1[cH:1][c:2][c:3][c:4][c:5]1»[c:10]−1[n:9][c:8][c:7][c:6]2[c:5][c:4][c:3][c:2][c:1]−12
Pictet-Spengler-6-membered-ring	[NH3+,NH2]−[C$(C(N)(C)(C)(C)),C$([CH](N)(C)(C)),C$([CH2](N)(C)):8]−[C$(C(c)(C)(C)(C)),C$([CH](c)(C)(C)),C$([CH2](c)(C)):7]−[c:6]1[c:1][c:2][c:3][c:4][cH:5]1.[CH:10](−[CX4:12])=[O:11]»[c,C:12]−[CH:10]−1−[N]−[C:8]−[C:7]−[c:6]2[c:1][c:2][c:3][c:4][c:5]−12
Pictet-Spengler-5-membered-ring	[NH3+,NH2]−[C$(C(N)(C)(C)(C)),C$([CH](N)(C)(C)),C$([CH2](N)(C)):8]−[C$(C(c)(C)(C)(C)),C$([CH](c)(C)(C)),C$([CH2](c)(C)):7]−[c:6]1[c:1][c:2][nH:3][cH:5]1.[CH:10](−[CX4:12])=[O:11]»[c,C:12]−[CH:10]−1−[N]−[C:8]−[C:7]−[c:6]2[c:1][c:2][nH:3][c:5]−12
Bischler-Indole	[NH2,NH3+1:8]−[c:5]1[cH:4][c:3][c:2][c:1][c:6]1.[Br:18][C$([CH2](C)(Br)),C$([CH](C)(C)(Br)):17]−[C:15](=[O:16])−[c:10]1[c:11][c:12][c:13][c:14][c:9]1»[c:13]1[c:12][c:11][c:10]([c:9][c:14]1)−[c:15]1[c:17][c:4]2[c:3][c:2][c:1][c:6][c:5]2[nH+0:8]1
Benzimidazol formation	[OH,O−]−[C$(C([CX4])),C$([CH]):2]=[O:3].[NH2,NH3+:12]−[c:9]1[c:8][c:7][c:6][c:5][c:10]1−[N$([NH](c)([CX4])),N$([NH2,NH3+1](c)):11]»[c:2]1[n+0:12][c:9]2[c:8][c:7][c:6][c:5][c:10]2[n:11]1

Aminothiazol formation	[C$([CH]([Br,Cl,I])(C)([CX4])),C$([CH2]([Br,Cl,I])(c)),C$([CH2]([Br,Cl,I])(C)):1](−[Br,Cl,I:3])−[C$(C(C)([c,C])),C$([CH](C)):2]=[O:4]>[NH2:8]−[C:7](−[NH2:9])=[S:10]>[NH2+0: 8]−[c:7]1[n:9][c:1][c:2][s:10]1
Benzoxazol formation	[OH,O−]−[C$(C([CX4])),C$([CH]):2]=[O:3].[NH2,NH3+:12]−[c:9]1[c:8][c:7][c:6][c:5][c:10]1−[OH:11]»[c:2]1[o+0:12][c:9]2[c:8][c:7][c:6][c:5][c:10]2[n:11]1
Benzothiazol formation	[OH,O−]−[C$(C([CX4])),C$([CH]):2]=[O:3].[NH2,NH3+:12]−[c:9]1[c:8][c:7][c:6][c:5][c:10]1−[SH:11]»[c:2]1[s+0:12][c:9]2[c:8][c:7][c:6][c:5][c:10]2[n:11]1
Rap-Stoermer	[Cl:1][CH2:2]−[C$([CH](C)),C$(C(C)(C)):3]=[O:4].[OH:12]−[c:11]1[c:6][c:7][c:8][c:9][c:10]1−[CH:13]=[O:14]»[C:3](=[O:4])−[c:2]1[c:13][c:10]2[c:9][c:8][c:7][c:6][c:11]2[o:12]1
Niementowski	[N$([NH](C)([CX4])),N$([NH2,NH3+1](C)):2]−[C$(C(N)(C)),C$([CH](N)):1]=[O:3].[NH2,NH3+1:13]−[c:8]1[c:7][c:6][c:5][c:4][c:9]1−[C:10](−[OH,O−:12])=[O:11]»[O:11]=[c:10]−1[n:2][c:1][n:13][c:8]2[c:7][c:6][c:5][c:4][c:9]−12
Quinazolinone formation	[NH2,NH3+]−[C$([CX4](N)([c,C])([c,C])([c,C])),C$([CH](N)([c,C])([c,C])),C$([CH2](N)([c,C])),C$([CH3](N)):2].[NH2:12]−[c:7]1[c:6][c:5][c:4][c:3][c:8]1−[C:9](−[OH,O−:11])=[O:10]»[C:2]−[n+0]−1[c:13][n:12][c:7]2[c:6][c:5][c:4][c:3][c:8]2[c:9]−1=[O:10]
Chinonlin-2-one Intramol	[C$(C(C)(=O)([CX4])),C$([CH](C)(=O)):10](=[O:13])−[C$([CH]([CX4])),C$([CH2]):9]−[C:8](=[O:12])−[NH:7]−[c:5]1[cH:6][c:1][c:2][c:3][c:4]1»[c:10]−1[c:9][c:8](−[OH:12])−[n:7]−[c:5]2[c:4][c:3][c:2][c:1][c:6]−12
Tetrazol formation	[C$(C(#N)([CX4])),C$([CH](#N)):1]#[N:2]»[c$(c(n)(n)([CX4])),c$([cH](n)(n)):1]−1[n:2][nH:4][n:6][n:5]−1
Tetrahydro-Indole formation	[N$([NH2]([CX4])),N$([NH3+1]([CX4])):1].[O:5]−[C$([CH]([CX4])(C)(O)),C$([CH2]([CX4])(O)):3][C$(C([CX4])(=O)([CX4])),C$([CH]([CX4])(=O)):4]=[O:6]>[O:15]=[C:9]−1−[CH2:10]−[CH2:11]−[CH2:12]−[CH2:13]−[CH2:14]−1>[c:4]1[c:3][n+0:1][c:10]2−[C:11]−[C:12]−[C:13]−[C:14]−[c:9]12
3-nitrile pyridine	[C$(C(=O)([CX4])([CX4])),C$([CH](=O)([CX4])):2](=[O:6])−[C$([CH]([CX4])),C$([CH2]):3]−[C$(C(=O)([CX4])([CX4])),C$([CH](=O)([CX4])):4]=[O:7].[NH2:8]−[C:9](=[O:10])−[CH2:11][C:12]#[N:13]»[OH:10]−[c:9]1[n:8][c:4][c:3][c:2][c:11]1[C:12]#[N:13]
Triazole formation	[C$(C(#N)([CX4])):2]#[N:3].[NH2,NH3+1:4]−[NH:5]−[C$(C(N)(=O)([CX4])),C$([CH](N)(=O)):6]=[O:7]»[c:6]−1[n:5][c:2][n:3][n:9]−1
Huisgen 1-3 Dipolar Cycloaddition	[C$(C(#C)([CX4])):2]#[C$(C(#C)([CX4])):1].[N$(N(N)([CX4])):5]~[N]~[N]»[c:2]1[c:1][n:5][n][n]1
Huisgen 1 3 Dipolar Cycloaddition double bond	[C$(C(=C)([CX4])):2]=[C$(C(=C)([CX4])):1].[N$(N(N)([CX4])):5]~[N]~[N]»[C:2]1[C:1][N:5][N]=[N]1

Diels-Alder	[C$(C(=C)([CX4,OX2,NX3])([CX4,OX2,NX3])),C$([CH](=C)([CX4,OX2,NX3])),C$([CH2](=C)):1]=[C$(C(=C)([CX4,OX2,NX3])([CX4,OX2,NX3])),C$([CH](=C)([CX4,OX2,NX3])),C$([CH2](=C)):2].[C$(C(=C)([CX4,OX2,NX3])([CX4,OX2,NX3])),C$([CH](=C)([CX4,OX2,NX3])),C$([CH2](=C)):3]=[C$([C](=C)(C)([CX4,OX2,NX3])),C$([CH](=C)(C)):4]−[C$([C](=C)(C)([CX4,OX2,NX3])),C$([CH](=C)(C)):5]=[C$(C(=C)([CX4,OX2,NX3])([CX4,OX2,NX3])),C$([CH](=C)([CX4,OX2,NX3])),C$([CH2](=C)):6]»[C:1]1[C:2][C:3][C:4]=[C:5][C:6]1
Diels-Alder-Alkyne	[C$(C(#C)([CX4,OX2,NX3])),C$([CH](#C)):1]#[C$(C(#C)([CX4,OX2,NX3])),C$([CH](#C)):2].[C$(C(=C)([CX4,OX2,NX3])([CX4,OX2,NX3])),C$([CH](=C)([CX4,OX2,NX3])),C$([CH2](=C)):3]=[C$([C](=C)(C)([CX4,OX2,NX3])),C$([CH](=C)(C)):4]−[C$([C](=C)(C)([CX4,OX2,NX3])),C$([CH](=C)(C)):5]=[C$(C(=C)([CX4,OX2,NX3])([CX4,OX2,NX3])),C$([CH](=C)([CX4,OX2,NX3])),C$([CH2](=C)):6]»[C:1]1=[C:2][C:3][C:4]=[C:5][C:6]1
Spiro-piperidine formation	[N$(N([CX4])),N$([NH]):5]−1−[C$(C(C)(N)([CX4])([CX4])),C$([CH](C)(N)([CX4])),C$([CH2](C)(N)):4]−[C$(C(C)(C)([CX4])([CX4])),C$([CH](C)(C)([CX4])),C$([CH2](C)(C)):3]−[C:2](=[O:7])−[C$(C(C)(C)([CX4])([CX4])),C$([CH](C)(C)([CX4])),C$([CH2](C)(C)):1]−[C$(C(C)(N)([CX4])([CX4])),C$([CH](C)(N)([CX4])),C$([CH2](C)(N)):6]−1.[C$([CH](C)([CX4])([CX4])),C$([CH2](C)([CX4])),C$([CH3]):18]−[C:16](=[O:17])−[c:14]1[c:9][c:10][c:11][c:12][c:13]1−[OH:15]»[N:5]−1−[C:4]−[C:3][C:2]2([C:1]−[C:6]−1)[C:18]−[C:16](=[O:17])−[c:14]1[c:9][c:10][c:11][c:12][c:13]1−[O:15]2
Pyrazol formation	[NH2,NH3+:3]−[N$([NH](N)([CX4])):2].[C$([CH](C)(C)([CX4])),C$([CH2](C)(C)):6](−[C$(C(=O)(C)([CX4])),C$([CH](=O)(C)):5]=[O:9])−[C$(C(=O)(C)([CX4])),C$([CH](=O)(C)):7]=[O:10]»[c:7]1[n:3][n:2][c:5][c:6]1
Phthalazinone	[NH2,NH3+1:2]−[N$([NH](N)([CX4])):1].[OH,O−:12]−[C:10](=[O:11])−[c:5]1[c:4][c:9][c:8][c:7][c:6]1−[C$(C(c)(=O)([CX4])),C$([CH](c)(=O)):13]=[O:14]»[N:1]−1−[N:2]=[C:13][c:6]2[c:7][c:8][c:9][c:4][c:5]2−[C:10]−1=[O:11]
Paal-Knorr-pyrole formation	[C$(C(=O)(C)([CX4])),C$(C[H](=O)(C)):1](=[O:2])−[$([CH](C)(C)([CX4])),$([CH2](C)(C)):3]−[$([CH](C)(C)([CX4])),$([CH2](C)(C)):4]−[C$(C(=O)(C)([CX4])),C$(C[H](=O)(C)):5]=[O:6].[N$([NH2,NH3+1]([CX4])):7]»[c:5]1[c:4][c:3][c:1][n+0:7]1
Triaryl-imidazol-1 2-diketone	[CH:7](=[O:8])−[c:1]1[c:2][c:3][c:4][c:5][c:6]1.[O:24]=[C:23](−[C:22](=[O:25])−[c:15]1[c:10][c:11][c:12][c:13][c:14]1)−[c:20]1[c:21][c:16][c:17][c:18][c:19]1>[NH4].[O−]C(=O)C>[nH:27]−1[c:7]([n:26][c:23]([c:22]−1[c:15]1[c:10][c:11][c:12][c:13][c:14]1)−[c:20]1[c:21][c:16][c:17][c:18][c:19]1)−[c:1]1[c:2][c:3][c:4][c:5][c:6]1
Triaryl-imidazol-alpha hydroxy ketone	[CH:7](=[O:8])−[c:1]1[c:2][c:3][c:4][c:5][c:6]1.[O:24]=[C:23](−[CH:22](−[OH :25])−[c:15]1[c:10][c:11][c:12][c:13][c:14]1)−[c:20]1[c:21][c:16][c:17][c:18][c:19]1>[NH4].[O−]C(=O)C>[nH:27]−1[c:7]([n:26][c:23]([c:22]−1[c:15]1[c:10][c:11][c:12][c:13][c:14]1)−[c:20]1[c:21][c:16][c:17][c:18][c:19]1)−[c:1]1[c:2][c:3][c:4][c:5][c:6]1
Fischer indole	[C$([CH2](C)([CX4])),C$([CH3](C)):4]−[C$(C([CX4])(=O)([CX4])),C$([CH]([CX4])(=O)):3]=[O:1].[NH2:6]−[NH:7]−[c:9]1[c:10][c:11][c:12][c:13][cH:8]1»[c:3]1[nH:7][c:9]2[c:10][c:11][c:12][c:13][c:8]2[c:4]1

Friedlaender chinoline formation	[C$([CH2](C)([CX4])),C$([CH3](C)):4]−[C$(C([CX4])(=O)([CX4])),C$([CH]([CX4])(=O)):2]=[O:1].[NH2:12]−[c:10]1[c:9][c:8][c:7][c:6][c:11]1−[C$(C(c)(=O)([CX4])),C$([CH](c)(=O)):13]=[O:14]»[c:2]1[c:13][c:11]2[c:6][c:7][c:8][c:9][c:10]2[n:12][c:4]1
Peachmann coumarine	[OH:7]−[c:6]1[cH:1][c:2][c:3][c:4][c:5]1.[O$(O(C)([CX4])):12]−[C:11](=[O:15])−[C$([CH](C)(C)([CX4])),C$([CH2](C)(C)):10]−[C:8]=[O:16]»[C:8]−1=[C:10]−[C:11](=[O:15])−[O]−[c:6]2[c:5][c:4][c:3][c:2][c:1]−12
Benzofuran formation	[C$(C(#C)([CX4])),C$([CH](#C)):2]#[CH:1].[OH:11]−[c:8]1[c:7][c:6][c:5][c:4][c:9]1[I:10]»[c:2]1[c:1][c:9]2[c:4][c:5][c:6][c:7][c:8]2[o:11]1
Imidazol-Acetamid	[C$(C(=O)(N)([CX4])),C$([CH](=O)(N)):5](=[O:6])−[NH:4]−[C:2](−[NH2:1])=[NH:3].[Br:12][C$([CH](Br)(C)([CX4])),C$([CH2](Br)(C)):9]−[C$(C(=O)(C)([CX4])),C$([CH](=O)(C)):8]=[O:10]»[C:5](=[O:6])−[NH:4]−[c:2]1[n:3][c:9][c:8][nH:1]−1
Dieckmann 5-ring	[O$(O(C)([CX4])):8][C:7](=[O:9])[CH:6][C:5][C:4][C:3][C:2]([O$(O(C)([CX4])):10])=[O:1]»[O:8][C:7](=[O:9])[C:6]1[C:5][C:4][C:3][C:2]1=[O:1]
Dieckmann 6-ring	[O$(O(C)([CX4])):8][C:7](=[O:9])[CH:6][C:5][C:11][C:4][C:3][C:2]([O$(O(C)([CX4])):10])=[O:1]»[O:8][C:7](=[O:9])[C:6]1[C:5][C:11][C:4][C:3][C:2]1=[O:1]
Flavone formation	[Cl:9][C:7](=[O:8])−[c:3]1[c:2][c:1][c:6][c:5][c:4]1.[C$([CH2](C)([CX4])),C$([CH3](C)):18]−[C:16](=[O:17])−[c:14]1[c:13][c:12][c:11][c:10][c:15]1−[OH:19]»[O:17]=[C:16]−1−[C:18]=[C:7](−[O:8]−[c:15]2[c:10][c:11][c:12][c:13][c:14]−12)−[c:3]1[c:2][c:1][c:6][c:5][c:4]1
Oxadiazole formation	[OH,O−:3]−[C$(C(=O)(O)[CX4]),C$([CH](=O)(O)):2]=[O:1].[N:12]#[C:11][c:10]1[c:5][c:6][c:7][c:8][c:9]1»[c:2]1[n:12][c:11]([n:13][o:1]1)−[c:10]1[c:5][c:6][c:7][c:8][c:9]1
Michael addition	[C$(C(C)(=O)([CX4,OX2&H0])),C$(C(C)(#N)),N$([N+1](C)(=O)([O−1])):1][C$([CH]([C,N])([C,N])([CX4])),C$([CH2]([C,N])([C,N])):2][C$(C(C)(=O)([CX4,OX2&H0])),C$(C(C)(#N)),N$([N+1](C)(=O)([O−1])):3].[C$(C(C)(#N)),C$(C(C)([CX4,OX2&H0])([CX4,OX2&H0])([OX2&H0])),C$([CH](C)([CX4,OX2&H0])([OX2&H0])),C$([CH2](C)([OX2&H0])),C$(C(C)(=O)([OX2&H0])):6][CH:5]=[C$(C(=C)([CX4])([CX4])),C$([CH](=C)([CX4])),C$([CH2](=C)):4]»[C:6][C:5][C:4][C:2]([C:1])[C:3]
Cross Claissen	[C$([C](O)([CX4])([CX4])([CX4])),C$([CH](O)([CX4])([CX4])),C$([CH2](O)([CX4])):4]−[O:3]−[C$(C(=O)([CX4])),C$([CH](=O)):2]=[O:5].[C$([CH](C)([CX4])([CX4])),C$([CH2](C)([CX4])),C$([CH3](C)):7]−[C$(C(=O)([CX4])),C$([CH](=O)):8]=[O:9]»[C:7](−[C:2]=[O:5])−[C:8]=[O:9]
Williamson ether	[Br,Cl,I:1][C$(C([Br,Cl,I])([CX4,c])([CX4,c])([CX4,c])),C$([CH]([Br,Cl,I])([CX4,c])([CX4,c])),C$([CH2]([Br,Cl,I])([CX4,c])),C$([CH3]([Br,Cl,I])),c$(c([Br,Cl,I])([c,n,o])([c,n,o])):2].[OH:3][C$(C(O)([CX4,c])([CX4,c])([CX4,c])),C$([CH](O)([CX4,c])([CX4,c])),C$([CH2](O)([CX4,c])),C$([CH3]([OH])),c$(c([OH])([c,n,o])([c,n,o])):4]»[C,c:2][O:3][C,c:4]
Ester foramtion	[Cl,OH,O−:3][C$(C(=O)([CX4,c])),C$([CH](=O)):2]=[O:4].[O$([OH]([CX4,c])),O$([OH]([CX4,c])([CX4,c])),S$([SH]([CX4,c])),S$([SH]([CX4,c])([CX4,c])):6]»[*:6]−[C:2]=[O:4]
Reductive amination-Ketone	[C$(C(=O)([CX4,c])([CX4,c])),C$([CH](=O)([CX4,c])):1]=[O:2].[N$([NH2,NH3+1]([CX4,c])),N$([NH]([CX4,c])([CX4,c])):3]»[N+0:3][C:1]

Suzuki coupling	[Br:1][c$(c(Br)),n$(n(Br)),o$(o(Br)),C$([CH](Br)(=C)):2].[C$(C(B)([CX4])([CX4])([CX4])),C$([CH](B)([CX4])([CX4])),C$([CH2](B)([CX4])),C$([CH2](B)),C$(C(B)(=C)),c$(c(B)),o$(o(B)),n$(n(B)):3][B$(B([C,c,n,o])([OH,$(OC)])([OH,$(OC)])),B$([B−1]([C,c,n,o])(N)([OH,$(OC)])([OH,$(OC)])):4]»[C,c,n,o:2][C,c,n,o:3]
Piperidine and Indole	[cH:9]1[c:8][n:7][c:5]2[c:4][c:3][c:2][c:1][c:6]12.[N:17]−1−[CX4:16]−[CH:15]−[C:14](=[O:20])−[CX4:19]−[CX4:18]−1»[N:17]−1−[C:18]−[C:19]−[C:14](=[C:15]−[C:16]−1)−[c:9]1[c:8][n:7][c:5]2[c:4][c:3][c:2][c:1][c:6]12
Negishi	[Br,I:1][C$(C([Br,I])([CX4])([CX4])([CX4])),C$([CH]([Br,I])([CX4])([CX4])),C$([CH2]([Br,I])([CX4])),C$([CH3]([Br,I])),C$([C]([Br,I])(=C)([CX4])),C$([CH]([Br,I])(=C)),C$(C([Br,I])(#C)),c$(c([Br,I])):2].[Br,I:3][C$(C([Br,I])([CX4])([CX4])([CX4])),C$([CH]([Br,I])([CX4])([CX4])),C$([CH2]([Br,I])([CX4])),C$([CH3]([Br,I])),C$([C]([Br,I])(=C)([CX4])),C$([CH]([Br,I])(=C)),C$(C([Br,I])(#C)),c$(c([Br,I])):4]»[C,c:2][C,c:4]
Mitsunobu imide	[C$(C(C)([CX4])([CX4])([CX4])),C$([CH](C)([CX4])([CX4])),C$([CH2](C)([CX4])),C$([CH3](C)):1][C:2](=[O:3])−[NH:4]−[C:5]([C$(C(C)([CX4])([CX4])([CX4])),C$([CH](C)([CX4])([CX4])),C$([CH2](C)([CX4])),C$([CH3](C)):10])=[O:7].[OH:11]−[C$(C(O)([CX4])([CX4])([CX4])),C$([CH](O)([CX4])([CX4])),C$([CH2](O)([CX4])),C$([CH3](O)):9]»[C:9][N+0:4](−[C:2]([C:1])=[O:3])−[C:5]([C:10])=[O:7]
Mitsunobu carboxlyic acid	[OH,O−]−[C$(C(=O)(O)([CX4,c])):2]=[O:3].[OH:8]−[C$([CH](O)([CX4,c])([CX4,c])),C$([CH2](O)([CX4,c])),C$([CH3](O)):6]»[C:6][O]−[C:2]=[O:3]
Mitsunobu sulfonic amide	[OH:1]−[C$([CH](O)([CX4,c])([CX4,c])),C$([CH2](O)([CX4,c])),C$([CH3](O)):3].[N$([NH](S)([CX4])),N$([NH2,NH3+1](S)):9][S$(S(N)([CX4])):6](=[O:7])=[O:8]»[C:3][N+0:9][S:6](=[O:8])=[O:7]
Heck	[C$([CH](=C)([CX4])),C$([CH2](=C)):2]=[C$(C(=C)([CX4])([CX4])),C$([CH](=C)([CX4])),C$([CH2](=C)):3].[Br,I:7][C$([CX4]([Br,I])),c$([c]([Br,I])):4]»[C,c:4][C:2]=[C:3]
Amide formation	[Cl,OH,O−:3][C$(C(=O)([CX4,c])),C$([CH](=O)):2]=[O:4].[N$([NH2,NH3+1]([CX4,c])),N$([NH]([CX4,c])([CX4,c])):6]»[N+0:6]−[C:2]=[O:4]
Thiolether formation-Alkene	[C$(C(=C)([CX4])([CX4])),C$([CH](=C)([CX4])),C$([CH2](=C)):1]=[C$(C(=C)([CX4])([CX4])),C$([CH](=C)([CX4])),C$([CH2](=C)):2].[SH:4]−[CX4:5][Br,Cl,I]»[C:1]−[C:2]−[S:4][C:5]
Thiolether formation-Carboxlyic acid	[C$([C](=O)([CX4])),C$([CH](=O)):2](=[O:1])[OH,Cl,O−:6].[SH:4]−[CX4:5][Br,Cl,I]»[CH2:2]−[S:4][C:5]
Ketone formation	[I:1][C$(C(I)([CX4,c])([CX4,c])([CX4,c])),C$([CH](I)([CX4,c])([CX4,c])),C$([CH2](I)([CX4,c])),C$([CH3](I)):2].[C$(C(=O)([Cl,OH,O−])([CX4,c])),C$([CH]([Cl,OH,O−])(=O)):3](=[O:6])[Cl,OH,O−:5]»[C:2]−[C:3]=[O:6]
Sulfonamide formation	[Cl:5][S$(S(=O)(=O)(Cl)([CX4])):2](=[O:3])=[O:4].[NH2+0,NH3+:6]−[C$(C(N)([CX4,c])([CX4,c])([CX4,c])),C$([CH](N)([CX4,c])([CX4,c])),C$([CH2](N)([CX4,c])),C$([CH3](N)),c$(c(N)):7]»[C,c:7]−[NH+0:6][S:2](=[O:4])=[O:3]
Ar-Imidazole formation	[c:5]1[c:4][nH:3][c:2][n:1]1.[OH,$(OC):13]−[B:12](−[OH,$(OC):14])−[c:6]1[c:7][c:8][c:9][c:10][c:11]1»[c:4]1[c:5][n:3]([c:2][n:1]1)−[c:6]1[c:7][c:8][c:9][c:10][c:11]1

Alkyne alkylation	[*:1][C:2]#[CH:3].[Br,I:4][C$(C([CX4,c])([CX4,c])([CX4,c])),C$([CH]([CX4,c])([CX4,c])),C$([CH2]([CX4,c])),C$([CH3]),c$(c):5]»[C,c:5][C:3]#[C:2][*:1]
Alkyne acylation	[C$(C(C)([CX4])([CX4])([CX4])),C$([CH](C)([CX4])([CX4])),C$([CH2](C)([CX4])),C$([CH3](C)):1][C:2]#[CH:3].[Br,I:4][C$(C(=O)([Br,I])([CX4])),C$([CH](=O)([Br,I])):5]=[O:6]»[C:1][C:2]#[C:3][C:5]=[O:6]
FGI Acyl chloride	[OH,O−:4]−[C$(C(=O)([OH,O−])([CX4])),C$([CH](=O)([OH,O−])):2]=[O:3]»[Cl:5][C:2]=[O:3]
FGI bromination	[OH:2]−[$([CX4]),c:1]»[Br:3][C,c:1]
FGI chlorination	[OH:2]−[$([CX4]),c:1]»[Cl:3][C,c:1]
FGI sulfonyl chloride	[OH,O−:3][S$(S([CX4])):2](=[O:4])=[O:5]»[Cl:6][S:2](=[O:5])=[O:4]
FGA alpha bromination	[OH+0,O−:5]−[C:3](=[O:4])−[C$([CH]([CX4])),C$([CH2]):2]»-[OH+0,O−:5]−[C:3](=[O:4])−[C:2]([Br:6])
FGA alpha chlorination	[OH+0,O−:5]−[C:3](=[O:4])−[C$([CH]([CX4])),C$([CH2]):2]»-[OH+0,O−:5]−[C:3](=[O:4])−[C:2]([Cl:6])
FGI Rosenmund-von Braun	[Cl,I,Br:7][c:1]1[c:2][c:3][c:4][c:5][c:6]1»[N:9]#[C:8][c:1]1[c:2][c:3][c:4]-[c:5][c:6]1
FGI nitrilation	[OH,NH2,NH3+:3]−[CH2:2]−[C$(C([CX4,c])([CX4,c])([CX4,c])),-C$([CH]([CX4,c])([CX4,c])),C$([CH2]([CX4,c])),C$([CH3]),c$(c):1]»-[C,c:1][C:2]#[N:4]

A.2.2 NMR Spectra

Presented here are the 1H-NMR spectra of the *de novo* designs synthesized in Chapter 3. The NMR spectra were provided courtesy of Dr. Daniel Merk.

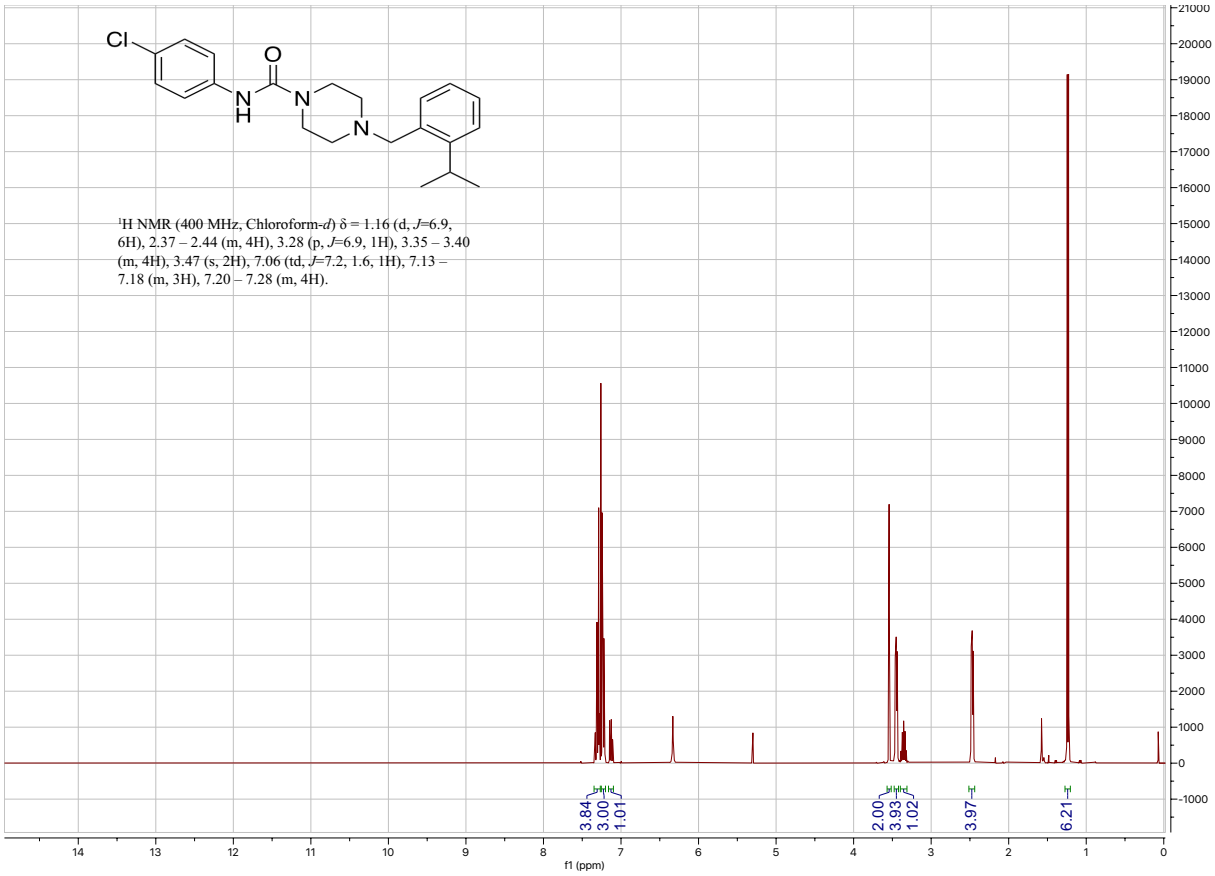


FIGURE A.1: 1-H NMR spectrum of the cariprazine *de novo* design (compound **6**).

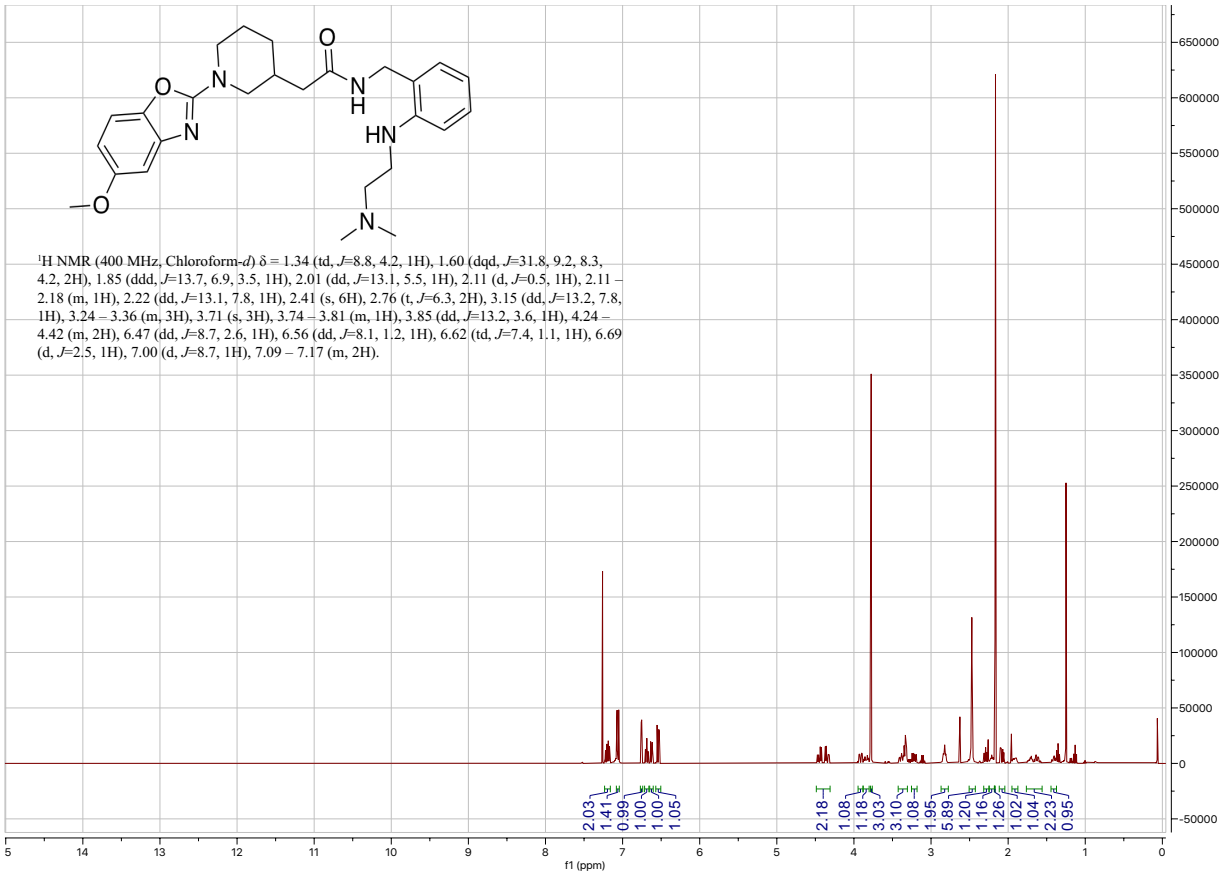


FIGURE A.2: 1-H NMR spectrum of the osimertinib *de novo* design (compound **7**).

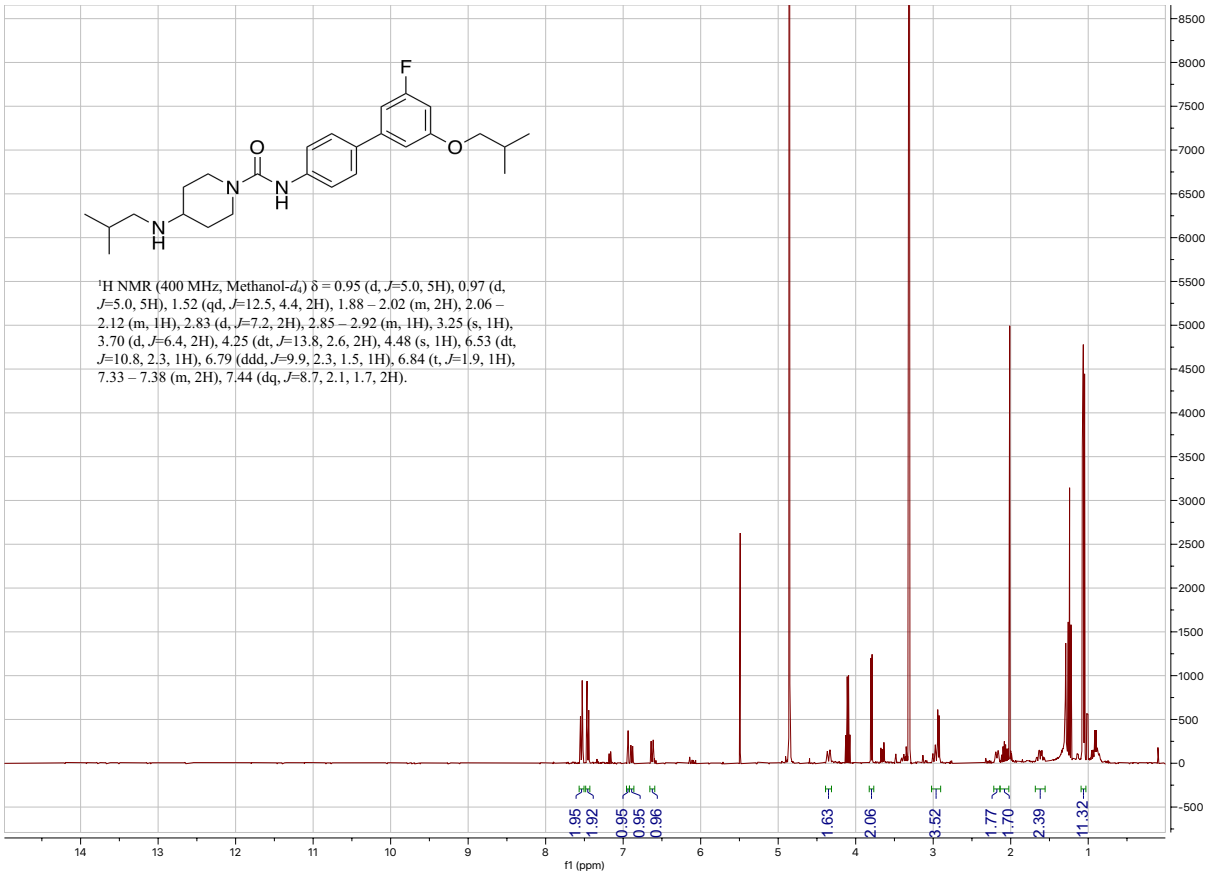


FIGURE A.3: 1-H NMR spectrum of the pimavanserin *de novo* design (compound **8**).

A.3 Chapter 4

A.3.1 Reaction Set Used in Chapter 4 for the First Design Cycle

Reaction Name	Reaction SMARTS
Pictet-Spengler-6-membered-ring	[NH3+,NH2]−[C$(C(N)(C)(C)(C)),C$([CH](N)(C)(C)),C$([CH2](N)(C)):8]−[C$(C(c)(C)(C)(C)),C$([CH](c)(C)(C)),C$([CH2](c)(C)):7]−[c:6]1[c:1][c:2][c:3][c:4][cH:5]1.[CH:10](−[CX4:12])=[O:11]»[c,C:12]−[CH:10]−1−[N]−[C:8]−[C:7]−[c:6]2[c:1][c:2][c:3][c:4][c:5]−12
Pictet-Spengler-5-membered-ring	[NH3+,NH2]−[C$(C(N)(C)(C)(C)),C$([CH](N)(C)(C)),C$([CH2](N)(C)):8]−[C$(C(c)(C)(C)(C)),C$([CH](c)(C)(C)),C$([CH2](c)(C)):7]−[c:6]1[c:1][c:2][nH:3][cH:5]1.[CH:10](−[CX4:12])=[O:11]»[c,C:12]−[CH:10]−1−[N]−[C:8]−[C:7]−[c:6]2[c:1][c:2][nH:3][c:5]−12
Aminothiazol formation	[C$([CH]([Br,Cl,I])(C)([CX4])),C$([CH2]([Br,Cl,I])(c)),C$([CH2]([Br,Cl,I])(C)):1](−[Br,Cl,I:3])−[C$(C(C)([c,C])),C$([CH](C)):2]=[O:4]>[NH2:8]−[C:7](−[NH2:9])=[S:10]>[NH2+0:8]−[c:7]1[n:9][c:1][c:2][s:10]1
Paal-Knorr-pyrole formation	[C$(C(=O)(C)([CX4])),C$(C[H](=O)(C)):1](=[O:2])−[$([CH](C)(C)([CX4])),$([CH2](C)(C)):3]−[$([CH](C)(C)([CX4])),$([CH2](C)(C)):4]−[C$(C(=O)(C)([CX4])),C$(C[H](=O)(C)):5]=[O:6].[N$([NH2,NH3+1]([CX4])):7]»[c:5]1[c:4][c:3][c:1][n+0:7]1

Triaryl-imidazol-1 2-diketone	[CH:7](=[O:8])−[c:1]1[c:2][c:3][c:4][c:5][c:6]1.[O:24]=[C:23](−[C:22](=[O:25])−[c:15]1[c:10][c:11][c:12][c:13][c:14]1)−[c:20]1[c:21][c:16][c:17][c:18][c:19]1>[NH4].[O−]C(=O)C>[nH:27]−1[c:7]([n:26][c:23]([c:22]−1[c:15]1[c:10][c:11][c:12][c:13][c:14]1)−[c:20]1[c:21][c:16][c:17][c:18][c:19]1)−[c:1]1[c:2][c:3][c:4][c:5][c:6]1
Triaryl-imidazol-alpha hydroxy ketone	[CH:7](=[O:8])−[c:1]1[c:2][c:3][c:4][c:5][c:6]1.[O:24]=[C:23](−[CH:22](−[OH:25])−[c:15]1[c:10][c:11][c:12][c:13][c:14]1)−[c:20]1[c:21][c:16][c:17][c:18][c:19]1>[NH4].[O−]C(=O)C>[nH:27]−1[c:7]([n:26][c:23]([c:22]−1[c:15]1[c:10][c:11][c:12][c:13][c:14]1)−[c:20]1[c:21][c:16][c:17][c:18][c:19]1)−[c:1]1[c:2][c:3][c:4][c:5][c:6]1
Fischer indole	[C$([CH2](C)([CX4])),C$([CH3](C)):4]−[C$(C([CX4])(=O)([CX4])),C$([CH]([CX4])(=O)):3]=[O:1].[NH2:6]−[NH:7]−[c:9]1[c:10][c:11][c:12][c:13][cH:8]1»[c:3]1[nH:7][c:9]2[c:10][c:11][c:12][c:13][c:8]2[c:4]1
Ester formation Acid Chloride	[Cl:3][C$(C(=O)([CX4,c])),C$([CH](=O)):2]=[O:4].[O$([OH]([CX4,c])),O$([OH]([CX4,c])([CX4,c])):6]»[O:6]−[C:2]=[O:4]
Thioester formation Acid Chloride	[Cl:3][C$(C(=O)([CX4,c])),C$([CH](=O)):2]=[O:4].[S$([SH]([CX4,c])),S$([SH]([CX4,c])([CX4,c])):6]»[S:6]−[C:2]=[O:4]
Reductive amination-Primary amine-Ketone	[C$(C(=O)([CX4,c])([CX4,c])),C$([CH](=O)([CX4,c])):1]=[O:2].[N$([NH2,NH3+1]([CX4,c])),N$([NH]([CX4,c])([CX4,c])):3]»[N+0:3][C:1]
Amide formation Acid Chloride	[Cl:3][C$(C(=O)([CX4,c])),C$([CH](=O)):2]=[O:4].[N$([NH2,NH3+1]([CX4,c])),N$([NH]([CX4,c])([CX4,c])):6]»[N+0:6]−[C:2]=[O:4]
Sulfonamide formation Sulfonyl Chloride	[Cl:5][S$(S(=O)(=O)(Cl)([CX4])):2](=[O:3])=[O:4].[NH2+0,NH3+:6]−[C$(C(N)([CX4,c])([CX4,c])([CX4,c])),C$([CH](N)([CX4,c])([CX4,c])),C$([CH2](N)([CX4,c])),C$([CH3](N)),c$(c(N)):7]»[C,c:7]−[NH+0:6][S:2](=[O:4])=[O:3]
Ar-Imidazole formation	[c:5]1[c:4][nH:3][c:2][n:1]1.[OH,$(OC):13]−[B:12](−[OH,$(OC):14])−[c:6]1[c:7][c:8][c:9][c:10][c:11]1»[c:4]1[c:5][n:3]([c:2][n:1]1)−[c:6]1[c:7][c:8][c:9][c:10][c:11]1
FGI Acyl chloride	[OH,O−:4]−[C$(C(=O)([OH,O−])([CX4])),C$([CH](=O)([OH,O−])):2]=[O:3]»[Cl:5][C:2]=[O:3]
FGI sulfonyl chloride	[OH,O−:3][S$(S([CX4])):2](=[O:4])=[O:5]»[Cl:6][S:2](=[O:5])=[O:4]
UGI-6-ring-aliphatic	[NH2,NH3+1:1][C:2]1=[N:3][CH2:4][CH2:5][CH2:6][CH2:7]1.O=[CH:8][c:9]2[cH:14][cH:13][cH:12][cH:11][cH:10]2>[N:15]#[C:16]C>[CH3:16][NH:15]c3[n:3]4[c:2]([CH2:7][CH2:6][CH2:5][CH2:4]4)[n+0:1][c:8]3[c:9]5[cH:14][cH:13][cH:12][cH:11][cH:10]5
UGI-6-ring-aromatic	[NH2,NH3+1:1][c:2]1[cH:7][cH:6][cH:5][cH:4][n:3]1.O=[CH:8][c:9]2[cH:14][cH:13][cH:12][cH:11][cH:10]2>[N:15]#[C:16]C>[CH3:16][NH:15]c([c:8]([c:9]3[cH:14][cH:13][cH:12][cH:11][cH:10]3)[n+1:1]4)[n:3]5[c:2]4[cH:7][cH:6][cH:5][cH:4]5
UGI-5-ring-aliphatic	[NH2,NH3+1:1][C:2]1=[N:3][CH2:4][CH2:5][CH2:6]1.O=[CH:7][c:8]2[cH:13][cH:12][cH:11][cH:10][cH:9]2>[N:14]#[C:15]C>[CH3:15][NH:14]c3[n:3]4[c:2]([CH2:6][CH2:5][CH2:4]4)[n+0:1][c:7]3[c:8]5[cH:13][cH:12][cH:11][cH:10][cH:9]5
Hantzsch	[C:1][NH:2][C:3]([NH2NH3+1:5])=[S:4].Br[CH2:6][C:7]([Cc:8])=O»[Cc:8][c:7]1[c:6][s:4][c:3]([NH:2][C:1])[n+0:5]1

A.3.2 Reaction Set Used in Chapter 4 for the Second Design Cycle

Reaction Name	Reaction SMARTS
Pictet-Spengler-6-membered-ring	[NH3+,NH2]−[C$(C(N)(C)(C)(C)),C$([CH](N)(C)(C)),C$([CH2](N)(C)):8]−[C$(C(c)(C)(C)(C)),C$([CH](c)(C)(C)),C$([CH2](c)(C)):7]−[c:6]1[c:1][c:2][c:3][c:4][cH:5]1.[CH:10](−[CX4:12])=[O:11]»[c,C:12]−[CH:10]−1−[N]−[C:8]−[C:7]−[c:6]2[c:1][c:2][c:3][c:4][c:5]−12
Pictet-Spengler-5-membered-ring	[NH3+,NH2]−[C$(C(N)(C)(C)(C)),C$([CH](N)(C)(C)),C$([CH2](N)(C)):8]−[C$(C(c)(C)(C)(C)),C$([CH](c)(C)(C)),C$([CH2](c)(C)):7]−[c:6]1[c:1][c:2][nH:3][cH:5]1.[CH:10](−[CX4:12])=[O:11]»[c,C:12]−[CH:10]−1−[N]−[C:8]−[C:7]−[c:6]2[c:1][c:2][nH:3][c:5]−12
Aminothiazol formation	[C$([CH]([Br,Cl,I])(C)([CX4])),C$([CH2]([Br,Cl,I])(c)),C$([CH2]([Br,Cl,I])(C)):1](−[Br,Cl,I:3])−[C$(C(C)([c,C])),C$([CH](C)):2]=[O:4]>[NH2:8]−[C:7](−[NH2:9])=[S:10]>[NH2+0:8]−[c:7]1[n:9][c:1][c:2][s:10]1
Paal-Knorr-pyrole formation	[C$(C(=O)(C)([CX4])),C$(C[H](=O)(C)):1](=[O:2])−[$([CH](C)(C)([CX4])),$([CH2](C)(C)):3]−[$([CH](C)(C)([CX4])),$([CH2](C)(C)):4]−[C$(C(=O)(C)([CX4])),C$(C[H](=O)(C)):5]=[O:6].[N$([NH2,NH3+1]([CX4])):7]»[c:5]1[c:4][c:3][c:1][n+0:7]1
Triaryl-imidazol-1 2-diketone	[CH:7](=[O:8])−[c:1]1[c:2][c:3][c:4][c:5][c:6]1.[O:24]=[C:23](−[C:22](=[O:25])−[c:15]1[c:10][c:11][c:12][c:13][c:14]1)−[c:20]1[c:21][c:16][c:17][c:18][c:19]1>[NH4].[O−]C(=O)C>[nH:27]−1[c:7]([n:26][c:23]([c:22]−1[c:15]1[c:10][c:11][c:12][c:13][c:14]1)−[c:20]1[c:21][c:16][c:17][c:18][c:19]1)−[c:1]1[c:2][c:3][c:4][c:5][c:6]1
Triaryl-imidazol-alpha hydroxy ketone	[CH:7](=[O:8])−[c:1]1[c:2][c:3][c:4][c:5][c:6]1.[O:24]=[C:23](−[CH:22](−[OH:25])−[c:15]1[c:10][c:11][c:12][c:13][c:14]1)−[c:20]1[c:21][c:16][c:17][c:18][c:19]1>[NH4].[O−]C(=O)C>[nH:27]−1[c:7]([n:26][c:23]([c:22]−1[c:15]1[c:10][c:11][c:12][c:13][c:14]1)−[c:20]1[c:21][c:16][c:17][c:18][c:19]1)−[c:1]1[c:2][c:3][c:4][c:5][c:6]1
Fischer indole	[C$([CH2](C)([CX4])),C$([CH3](C)):4]−[C$(C([CX4])(=O)([CX4])),C$([CH]([CX4])(=O)):3]=[O:1].[NH2:6]−[NH:7]−[c:9]1[c:10][c:11][c:12][c:13][cH:8]1»[c:3]1[nH:7][c:9]2[c:10][c:11][c:12][c:13][c:8]2[c:4]1
Ester formation Acid Chloride	[Cl:3][C$(C(=O)([CX4,c])),C$([CH](=O)):2]=[O:4].[O$([OH]([CX4,c])),O$([OH]([CX4,c])([CX4,c])):6]»[O:6]−[C:2]=[O:4]
Thioester formation Acid Chloride	[Cl:3][C$(C(=O)([CX4,c])),C$([CH](=O)):2]=[O:4].[S$([SH]([CX4,c])),S$([SH]([CX4,c])([CX4,c])):6]»[S:6]−[C:2]=[O:4]
Reductive amination-Aldehyde	[C$([CH](=O)([CX4,c])):1]=[O:2].[N$([NH2,NH3+1]([CX4,c])),N$([NH]([CX4,c])([CX4,c])):3]»[N+0:3][C:1]
Amide formation Acid Chloride	[Cl:3][C$(C(=O)([CX4,c])),C$([CH](=O)):2]=[O:4].[N$([NH2,NH3+1]([CX4,c])),N$([NH]([CX4,c])([CX4,c])):6]»[N+0:6]−[C:2]=[O:4]
Sulfonamide formation Sulfonyl Chloride Secondary amine	[Cl:5][S$(S(=O)(=O)(Cl)([CX4,c])):2](=[O:3])=[O:4].[C$(C(N)([CX4,c])([CX4,c])([CX4,c])),C$([CH](N)([CX4,c])([CX4,c])),C$([CH2](N)([CX4,c])),C$([CH3](N)),c$(c(N)):8]−[NH+0,NH2+:6]−[C$(C(N)([CX4,c])([CX4,c])([CX4,c])),C$([CH](N)([CX4,c])([CX4,c])),C$([CH2](N)([CX4,c])),C$([CH3](N)),c$(c(N)):7]»[C,c:7]−[N+0:6]([C,c:8])[S:2](=[O:4])=[O:3]
Sulfonamide formation Sulfonyl Chloride Primary amine	[Cl:5][S$(S(=O)(=O)(Cl)([CX4,c])):2](=[O:3])=[O:4].[NH2+0,NH3+:6]−[C$(C(N)([CX4,c])([CX4,c])([CX4,c])),C$([CH](N)([CX4,c])([CX4,c])),C$([CH2](N)([CX4,c])),C$([CH3](N)),c$(c(N)):7]»[C,c:7]−[NH+0:6][S:2](=[O:4])=[O:3]

Sulfonamide formation Sulfonyl Chloride Ammonia	[Cl:5][S$(S(=O)(=O)(Cl)([CX4,c])):2](=[O:3])=[O:4]>[NH3]>[NH2+0:6][S:2](=[O:4])=[O:3]
FGI Acyl chloride	[OH,O−:4]−[C$(C(=O)([OH,O−])([CX4])),C$([CH](=O)([OH,O−])):2]=[O:3]»[Cl:5][C:2]=[O:3]
FGI sulfonyl chloride	[OH,O−:3][S$(S([CX4])):2](=[O:4])=[O:5]»[Cl:6][S:2](=[O:5])=[O:4]
Hantzsch	[C:1][NH:2][C:3]([NH2,NH3+1:5])=[S:4].Br[CH2:6][C:7]([C,c:8])=O»[C,c:8][c:7]1[c:6][s:4][c:3]([NH:2][C:1])[n+0:5]1

A.3.3 NMR Spectra of the Bioactive DINGOS Designs

Presented here are the 1H-NMR spectra of the bioactive *de novo* designs synthesized in Chapter 4. The NMR spectra were provided courtesy of Berend Huisman.

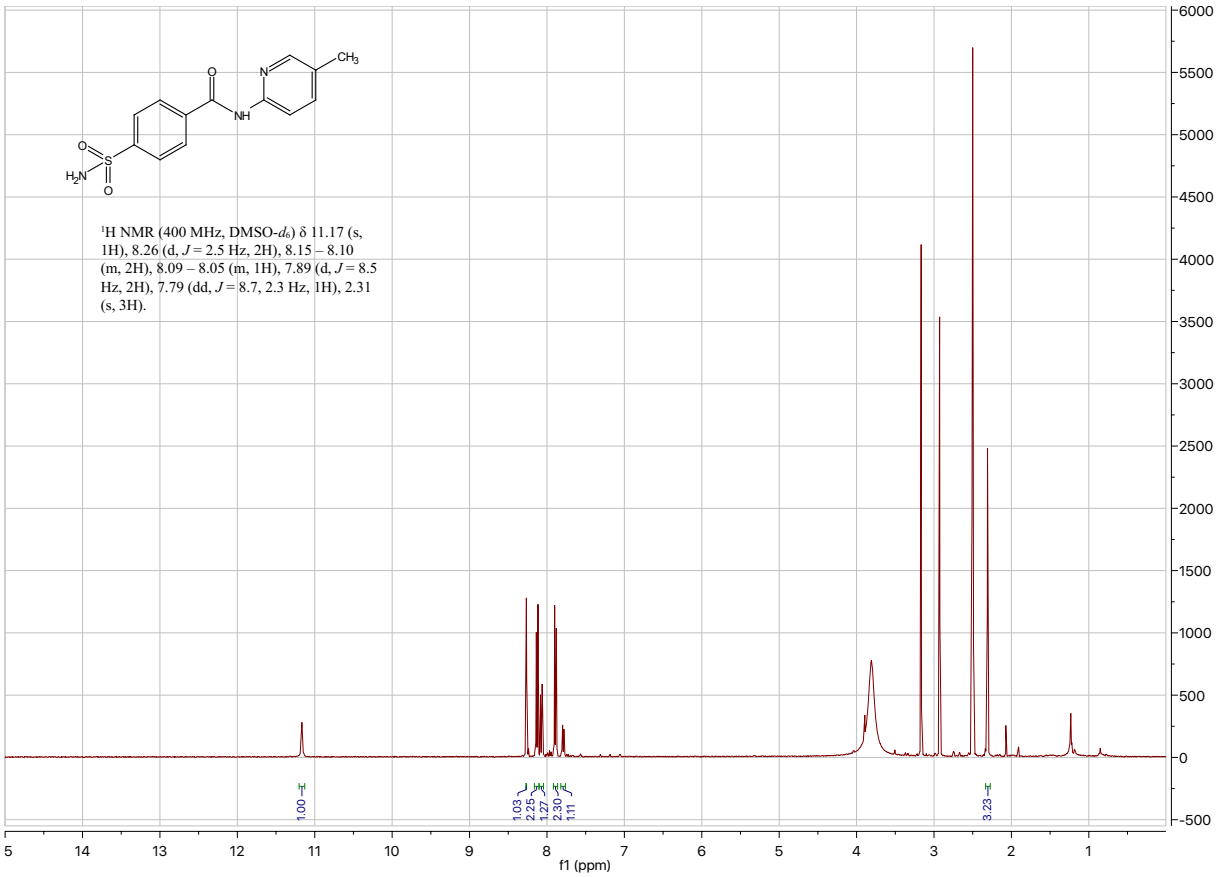


FIGURE A.4: 1-H NMR spectrum of the compound **66**.

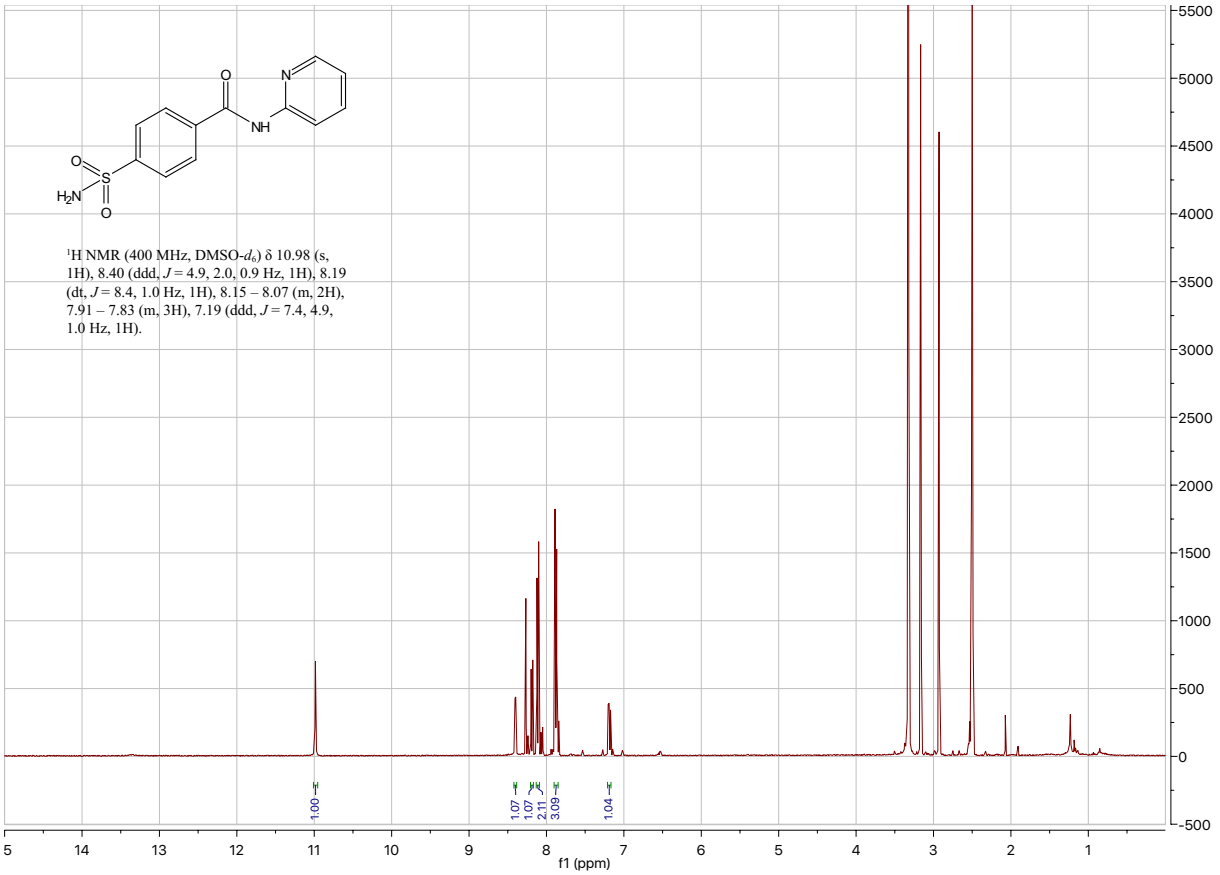


FIGURE A.5: 1-H NMR spectrum of the compound **67**.

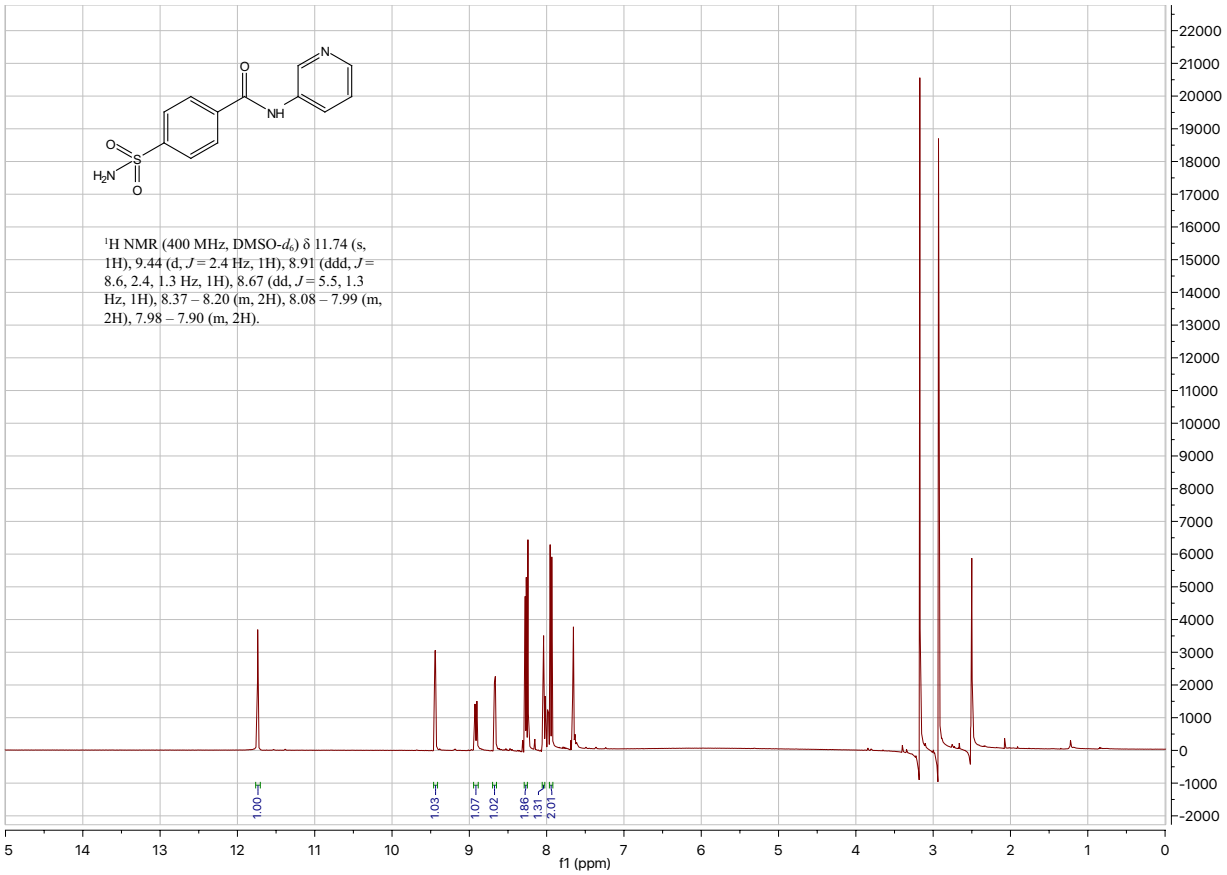


FIGURE A.6: 1-H NMR spectrum of the compound **68**.

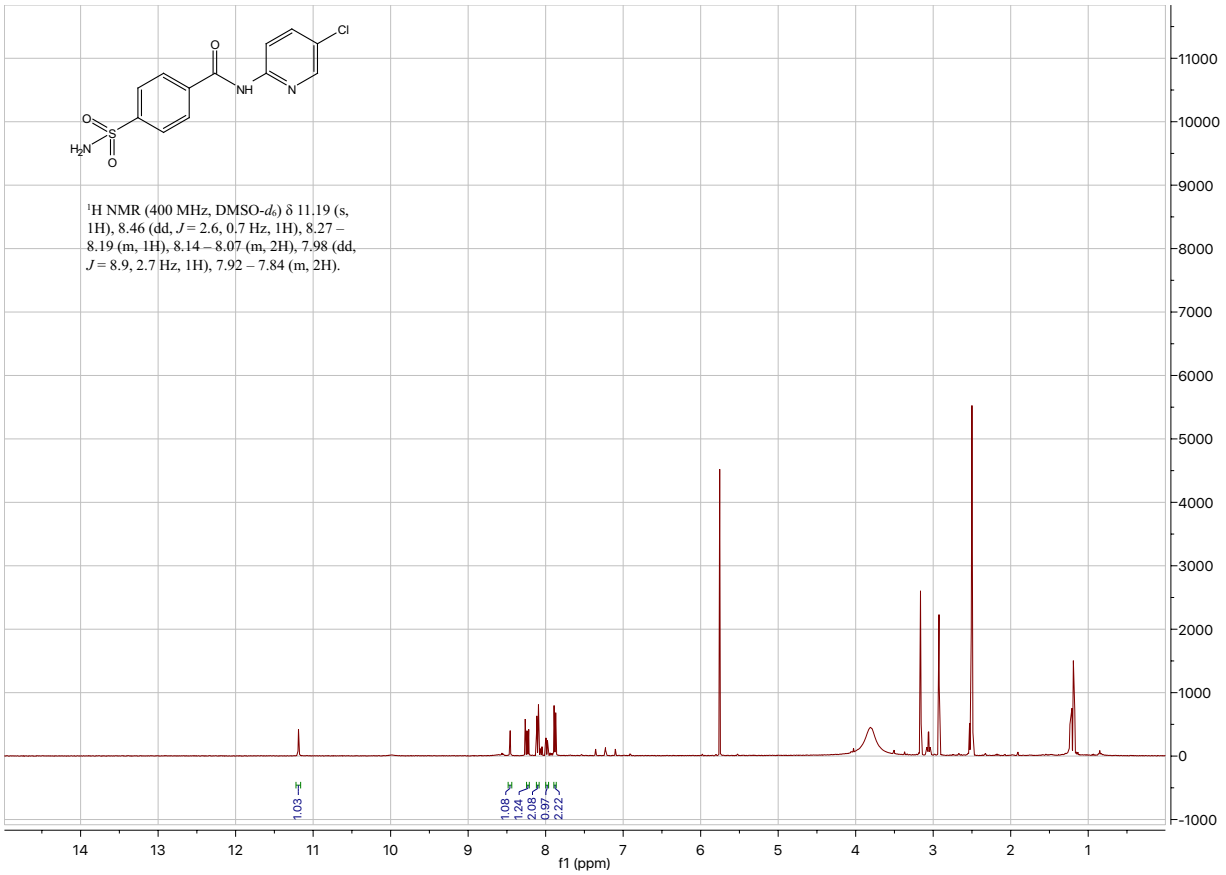


FIGURE A.7: 1-H NMR spectrum of the compound **70**.

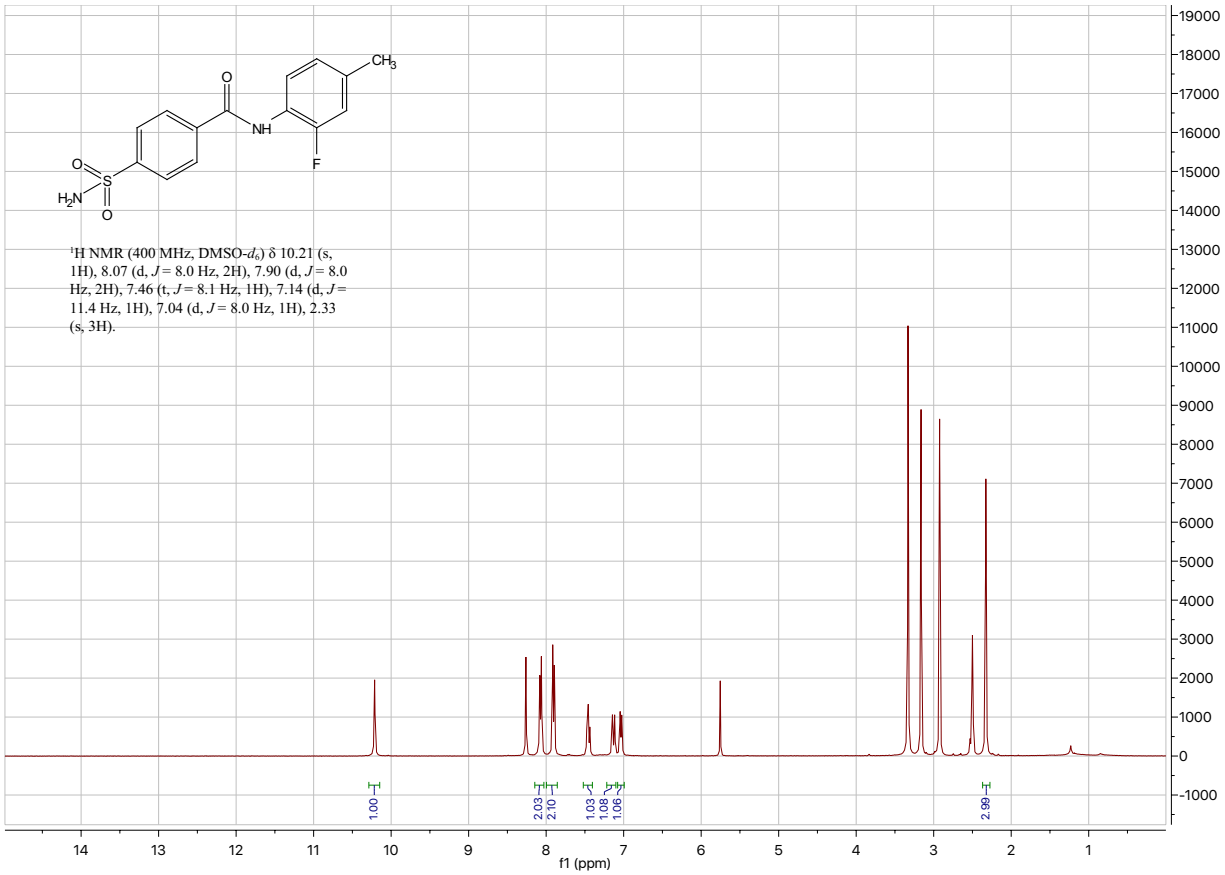


FIGURE A.8: 1-H NMR spectrum of the compound **71**.

A.4 Chapter 5

A.4.1 Reaction Set Used in Chapter 5

Reaction Name	Reaction SMARTS
Bischler-Napieralski	[$(C([CH2,CH3])),CH:10](=[O:11])−[NH+0:9]−[C$(C(N)(C)(C)(C)),C$([CH](N)(C)(C)),C$([CH2](N)(C)):8]−[C$(C(c)(C)(C)(C)),C$([CH](c)(C)(C)),C$([CH2](c)(C)):7]−[c:6]1[cH:1][c:2][c:3][c:4][c:5]1»[C:10]−1=[N+0:9]−[C:8]−[C:7]−[c:6]2[c:5][c:4][c:3][c:2][c:1]−12
Pictet-Gams	[$(C([CH2,CH3])),CH:10](=[O:11])−[NH+0:9]−[C$([CH](N)(C)(C)),C$([CH2](N)(C)):8]−[C$([C](c)(C)(C)),C$([CH](c)(C)):7]([O$(OC),OH])−[c:6]1[cH:1][c:2][c:3][c:4][c:5]1»[c:10]−1[n:9][c:8][c:7][c:6]2[c:5][c:4][c:3][c:2][c:1]−12
Pictet-Spengler-6-membered-ring	[NH3+,NH2]−[C$(C(N)(C)(C)(C)),C$([CH](N)(C)(C)),C$([CH2](N)(C)):8]−[C$(C(c)(C)(C)(C)),C$([CH](c)(C)(C)),C$([CH2](c)(C)):7]−[c:6]1[c:1][c:2][c:3][c:4][cH:5]1.[CH:10](−[CX4:12])=[O:11]»[c,C:12]−[CH:10]−1−[N]−[C:8]−[C:7]−[c:6]2[c:1][c:2][c:3][c:4][c:5]−12
Pictet-Spengler-5-membered-ring	[NH3+,NH2]−[C$(C(N)(C)(C)(C)),C$([CH](N)(C)(C)),C$([CH2](N)(C)):8]−[C$(C(c)(C)(C)(C)),C$([CH](c)(C)(C)),C$([CH2](c)(C)):7]−[c:6]1[c:1][c:2][nH:3][cH:5]1.[CH:10](−[CX4:12])=[O:11]»[c,C:12]−[CH:10]−1−[N]−[C:8]−[C:7]−[c:6]2[c:1][c:2][nH:3][c:5]−12

Bischler-Indole	[NH2,NH3+1:8]−[c:5]1[cH:4][c:3][c:2][c:1][c:6]1.[Br:18][C$([CH2](C)(Br)),C$([CH](C)(C)(Br)):17]−[C:15](=[O:16])−[c:10]1[c:11][c:12][c:13][c:14][c:9]1»[c:13]1[c:12][c:11][c:10]([c:9][c:14]1)−[c:15]1[c:17][c:4]2[c:3][c:2][c:1][c:6][c:5]2[nH+0:8]1
Benzimidazol formation	[OH,O−]−[C$(C([CX4])),C$([CH]):2]=[O:3].[NH2,NH3+:12]−[c:9]1[c:8][c:7][c:6][c:5][c:10]1−[N$([NH](c)([CX4])),N$([NH2,NH3+1](c)):11]»[c:2]1[n+0:12][c:9]2[c:8][c:7][c:6][c:5][c:10]2[n:11]1
Aminothiazol formation	[C$([CH]([Br,Cl,I])(C)([CX4])),C$([CH2]([Br,Cl,I])(c)),C$([CH2]([Br,Cl,I])(C)):1](−[Br,Cl,I:3])−[C$(C(C)([c,C])),C$([CH](C)):2]=[O:4]>[NH2:8]−[C:7](−[NH2:9])=[S:10]>[NH2+0:8]−[c:7]1[n:9][c:1][c:2][s:10]1
Benzoxazol formation	[OH,O−]−[C$(C([CX4])),C$([CH]):2]=[O:3].[NH2,NH3+:12]−[c:9]1[c:8][c:7][c:6][c:5][c:10]1−[OH:11]»[c:2]1[o+0:12][c:9]2[c:8][c:7][c:6][c:5][c:10]2[n:11]1
Benzothiazol formation	[OH,O−]−[C$(C([CX4])),C$([CH]):2]=[O:3].[NH2,NH3+:12]−[c:9]1[c:8][c:7][c:6][c:5][c:10]1−[SH:11]»[c:2]1[s+0:12][c:9]2[c:8][c:7][c:6][c:5][c:10]2[n:11]1
Rap-Stoermer	[Cl:1][CH2:2]−[C$([CH](C)),C$(C(C)(C)):3]=[O:4].[OH:12]−[c:11]1[c:6][c:7][c:8][c:9][c:10]1−[CH:13]=[O:14]»[C:3](=[O:4])−[c:2]1[c:13][c:10]2[c:9][c:8][c:7][c:6][c:11]2[o:12]1
Niementowski	[N$([NH](C)([CX4])),N$([NH2,NH3+1](C)):2]−[C$(C(N)(C)),C$([CH](N)):1]=[O:3].[NH2,NH3+1:13]−[c:8]1[c:7][c:6][c:5][c:4][c:9]1−[C:10](−[OH,O−:12])=[O:11]»[O:11]=[c:10]−1[n:2][c:1][n:13][c:8]2[c:7][c:6][c:5][c:4][c:9]−12
Quinazolinone formation	[NH2,NH3+]−[C$([CX4](N)([c,C])([c,C])([c,C])),C$([CH](N)([c,C])([c,C])),C$([CH2](N)([c,C])),C$([CH3](N)):2].[NH2:12]−[c:7]1[c:6][c:5][c:4][c:3][c:8]1−[C:9](−[OH,O−:11])=[O:10]»[C:2]−[n+0]−1[c:13][n:12][c:7]2[c:6][c:5][c:4][c:3][c:8]2[c:9]−1=[O:10]
Chinonlin-2-one Intramol	[C$(C(C)(=O)([CX4])),C$([CH](C)(=O)):10](=[O:13])−[C$([CH]([CX4])),C$([CH2]):9]−[C:8](=[O:12])−[NH:7]−[c:5]1[cH:6][c:1][c:2][c:3][c:4]1»[c:10]−1[c:9][c:8](−[OH:12])−[n:7]−[c:5]2[c:4][c:3][c:2][c:1][c:6]−12
Tetrazol formation	[C$(C(#N)([CX4])),C$([CH](#N)):1]#[N:2]»[c$(c(n)(n)([CX4])),c$([cH](n)(n)):1]−1[n:2][nH:4][n:6][n:5]−1
Tetrahydro-Indole formation	[N$([NH2]([CX4])),N$([NH3+1]([CX4])):1].[O:5]−[C$([CH]([CX4])(C)(O),C$([CH2]([CX4])(O)):3][C$(C([CX4])(=O)([CX4])),C$([CH]([CX4])(=O)):4]=[O:6]>[O:15]=[C:9]−1−[CH2:10]−[CH2:11]−[CH2:12]−[CH2:13]−[CH2:14]−1>[c:4]1[c:3][n+0:1][c:10]2−[C:11]−[C:12]−[C:13]−[C:14]−[c:9]12
3-nitrile pyridine	[C$(C(=O)([CX4])([CX4])),C$([CH](=O)([CX4])):2](=[O:6])−[C$([CH]([CX4])),C$([CH2]):3]−[C$(C(=O)([CX4])([CX4])),C$([CH](=O)([CX4])):4]=[O:7].[NH2:8]−[C:9](=[O:10])−[CH2:11][C:12]#[N:13]»[OH:10]−[c:9]1[n:8][c:4][c:3][c:2][c:11]1[C:12]#[N:13]
Triazole formation	[C$(C(#N)([CX4])):2]#[N:3].[NH2,NH3+1:4]−[NH:5]−[C$(C(N)(=O)([CX4])),C$([CH](N)(=O)):6]=[O:7]»[c:6]−1[n:5][c:2][n:3][n:9]−1
Huisgen 1-3 Dipolar Cycloaddition	[C$(C(#C)([CX4])):2]#[C$(C(#C)([CX4])):1].[N$(N(~N)([CX4])):5]~[N]~[N]»[c:2]1[c:1][n:5][n][n]1
Huisgen 1 3 Dipolar Cycloaddition double bond	[C$(C(=C)([CX4])):2]=[C$(C(=C)([CX4])):1].[N$(N(~N)([CX4])):5]~[N]~[N]»[C:2]1[C:1][N:5][N]=[N]1

Diels-Alder	[C$(C(=C)([CX4,OX2,NX3])([CX4,OX2,NX3])),C$([CH](=C)([CX4,OX2,NX3])),C$([CH2](=C)):1]=[C$(C(=C)([CX4,OX2,NX3])([CX4,OX2,NX3])),C$([CH](=C)([CX4,OX2,NX3])),C$([CH2](=C)):2].[C$(C(=C)([CX4,OX2,NX3])([CX4,OX2,NX3])),C$([CH](=C)([CX4,OX2,NX3])),C$([CH2](=C)):3]=[C$([C](=C)(C)([CX4,OX2,NX3])),C$([CH](=C)(C)):4]−[C$([C](=C)(C)([CX4,OX2,NX3])),C$([CH](=C)(C)):5]=[C$(C(=C)([CX4,OX2,NX3])([CX4,OX2,NX3])),C$([CH](=C)([CX4,OX2,NX3])),C$([CH2](=C)):6]»[C:1]1[C:2][C:3][C:4]=[C:5][C:6]1
Diels-Alder-Alkyne	[C$(C(#C)([CX4,OX2,NX3])),C$([CH](#C)):1]#[C$(C(#C)([CX4,OX2,NX3])),C$([CH](#C)):2].[C$(C(=C)([CX4,OX2,NX3])([CX4,OX2,NX3])),C$([CH](=C)([CX4,OX2,NX3])),C$([CH2](=C)):3]=[C$([C](=C)(C)([CX4,OX2,NX3])),C$([CH](=C)(C)):4]−[C$([C](=C)(C)([CX4,OX2,NX3])),C$([CH](=C)(C)):5]=[C$(C(=C)([CX4,OX2,NX3])([CX4,OX2,NX3])),C$([CH](=C)([CX4,OX2,NX3])),C$([CH2](=C)):6]»[C:1]1=[C:2][C:3][C:4]=[C:5][C:6]1
Spiro-piperidine formation	[N$(N([CX4])),N$([NH]):5]−1−[C$(C(C)(N)([CX4])([CX4])),C$([CH](C)(N)([CX4])),C$([CH2](C)(N)):4]−[C$(C(C)(C)([CX4])([CX4])),C$([CH](C)(C)([CX4])),C$([CH2](C)(C)):3]−[C:2](=[O:7])−[C$(C(C)(C)([CX4])([CX4])),C$([CH](C)(C)([CX4])),C$([CH2](C)(C)):1]−[C$(C(C)(N)([CX4])([CX4])),C$([CH](C)(N)([CX4])),C$([CH2](C)(N)):6]−1.[C$([CH](C)([CX4])([CX4])),C$([CH2](C)([CX4])),C$([CH3]):18]−[C:16](=[O:17])−[c:14]1[c:9][c:10][c:11][c:12][c:13]1−[OH:15]»[N:5]−1−[C:4]−[C:3][C:2]2([C:1]−[C:6]−1)[C:18]−[C:16](=[O:17])−[c:14]1[c:9][c:10][c:11][c:12][c:13]1−[O:15]2
Pyrazol formation	[NH2,NH3+:3]−[N$([NH](N)([CX4])):2].[C$([CH](C)(C)([CX4])),C$([CH2](C)(C)):6](−[C$(C(=O)(C)([CX4])),C$([CH](=O)(C)):5]=[O:9])−[C$(C(=O)(C)([CX4])),C$([CH](=O)(C)):7]=[O:10]»[c:7]1[n:3][n:2][c:5][c:6]1
Phthalazinone	[NH2,NH3+1:2]−[N$([NH](N)([CX4])):1].[OH,O−:12]−[C:10](=[O:11])−[c:5]1[c:4][c:9][c:8][c:7][c:6]1−[C$(C(c)(=O)([CX4])),C$([CH](c)(=O)):13]=[O:14]»[N:1]−1−[N:2]=[C:13][c:6]2[c:7][c:8][c:9][c:4][c:5]2−[C:10]−1=[O:11]
Paal-Knorr-pyrole formation	[C$(C(=O)(C)([CX4])),C$(C[H](=O)(C)):1](=[O:2])−[$([CH](C)(C)([CX4])),$([CH2](C)(C)):3]−[$([CH](C)(C)([CX4])),$([CH2](C)(C)):4]−[C$(C(=O)(C)([CX4])),C$(C[H](=O)(C)):5]=[O:6].[N$([NH2,NH3+1]([CX4])):7]»[c:5]1[c:4][c:3][c:1][n+0:7]1
Triaryl-imidazol-1 2-diketone	[CH:7](=[O:8])−[c:1]1[c:2][c:3][c:4][c:5][c:6]1.[O:24]=[C:23](−[C:22](=[O:25])−[c:15]1[c:10][c:11][c:12][c:13][c:14]1)−[c:20]1[c:21][c:16][c:17][c:18][c:19]1>[NH4].[O−]C(=O)C>[nH:27]−1[c:7]([n:26][c:23]([c:22]−1[c:15]1[c:10][c:11][c:12][c:13][c:14]1)−[c:20]1[c:21][c:16][c:17][c:18][c:19]1)−[c:1]1[c:2][c:3][c:4][c:5][c:6]1
Triaryl-imidazol-alpha hydroxy ketone	[CH:7](=[O:8])−[c:1]1[c:2][c:3][c:4][c:5][c:6]1.[O:24]=[C:23](−[CH:22](−[OH:25])−[c:15]1[c:10][c:11][c:12][c:13][c:14]1)−[c:20]1[c:21][c:16][c:17][c:18][c:19]1>[NH4].[O−]C(=O)C>[nH:27]−1[c:7]([n:26][c:23]([c:22]−1[c:15]1[c:10][c:11][c:12][c:13][c:14]1)−[c:20]1[c:21][c:16][c:17][c:18][c:19]1)−[c:1]1[c:2][c:3][c:4][c:5][c:6]1
Fischer indole	[C$([CH2](C)([CX4])),C$([CH3](C)):4]−[C$(C([CX4])(=O)([CX4])),C$([CH]([CX4])(=O)):3]=[O:1].[NH2:6]−[NH:7]−[c:9]1[c:10][c:11][c:12][c:13][cH:8]1»[c:3]1[nH:7][c:9]2[c:10][c:11][c:12][c:13][c:8]2[c:4]1

Friedlaender chinoline formation	[C$([CH2](C)([CX4])),C$([CH3](C)):4]−[C$(C([CX4])(=O)([CX4])),C$([CH]([CX4])(=O)):2]=[O:1].[NH2:12]−[c:10]1[c:9][c:8][c:7][c:6][c:11]1−[C$(C(c)(=O)([CX4])),C$([CH](c)(=O)):13]=[O:14]»[c:2]1[c:13][c:11]2[c:6][c:7][c:8][c:9][c:10]2[n:12][c:4]1
Peachmann coumarine	[OH:7]−[c:6]1[cH:1][c:2][c:3][c:4][c:5]1.[O$(O(C)([CX4])):12]−[C:11](=[O:15])−[C$([CH](C)(C)([CX4])),C$([CH2](C)(C)):10]−[C:8]=[O:16]»[C:8]−1=[C:10]−[C:11](=[O:15])−[O]−[c:6]2[c:5][c:4][c:3][c:2][c:1]−12
Benzofuran formation	[C$(C(#C)([CX4])),C$([CH](#C)):2]#[CH:1].[OH:11]−[c:8]1[c:7][c:6][c:5][c:4][c:9]1[I:10]»[c:2]1[c:1][c:9]2[c:4][c:5][c:6][c:7][c:8]2[o:11]1
Imidazol-Acetamid	[C$(C(=O)(N)([CX4])),C$([CH](=O)(N)):5](=[O:6])−[NH:4]−[C:2](−[NH2:1])=[NH:3].[Br:12][C$([CH](Br)(C)([CX4])),C$([CH2](Br)(C)):9]−[C$(C(=O)(C)([CX4])),C$([CH](=O)(C)):8]=[O:10]»[C:5](=[O:6])−[NH:4]−[c:2]1[n:3][c:9][c:8][nH:1]−1
Dieckmann 5-ring symmetry 1	[O$(O(C)([CX4])):8][C:7](=[O:9])[CH:6][C:5][C:4][C:3][C:2]([O$(O(C)([CX4])):10])=[O:1]»[O:8][C:7](=[O:9])[C:6]1[C:5][C:4][C:3][C:2]1=[O:1]
Dieckmann 6-ring symmetry 1	[O$(O(C)([CX4])):8][C:7](=[O:9])[CH:6][C:5][C:11][C:4][C:3][C:2]([O$(O(C)([CX4])):10])=[O:1]»[O:8][C:7](=[O:9])[C:6]1[C:5][C:11][C:4][C:3][C:2]1=[O:1]
Flavone formation	[Cl:9][C:7](=[O:8])−[c:3]1[c:2][c:1][c:6][c:5][c:4]1.[C$([CH2](C)([CX4])),C$([CH3](C)):18]−[C:16](=[O:17])−[c :14]1[c:13][c:12][c:11][c:10][c:15]1−[OH:19]»[O:17]=[C:16]−1−[C:18]=[C:7](−[O:8]−[c:15]2[c:10][c:11][c:12][c:13][c:14]−12)−[c:3]1[c:2][c:1][c:6][c:5][c:4]1
Oxadiazole formation	[OH,O−:3]−[C$(C(=O)(O)[CX4]),C$([CH](=O)(O)):2]=[O:1].[N:12]#[C:11][c:10]1[c:5][c:6][c:7][c:8][c:9]1»[c:2]1[n:12][c:11]([n:13][o:1]1)−[c:10]1[c:5][c:6][c:7][c:8][c:9]1
Michael addition	[C:1][OH,SH:2].[C$([CH2]),C$([CH]C),C$(C(C)(C)):3]=[C:4][C$(C(C)(C)(=O)),C$([CH](C)(=O)):5](=[O,S:6])»[C:1][O,S:2][C:3][C:4][C:5](=[O,S:6])
Cross Claisen	[C$([CH2](C)(C)),C$([CH3](C)):2][C:3](=[O:4])[O:5][C,c:6].[C$([CH2]([C,c])(C)),C$([CH3](C)):7][C:9](=[O:10])[O:11][C,c:12]»[C$([CH2]([C,c])(C)),C$([CH3](C)):7][C:9](=[O:10])[C$([CH](C)(C)(C)),C$([CH2](C)(C)):2][C:3](=[O:4])[O:5][C,c:6]
Williamson ether Alcohol	[Br,Cl,I:1][C$(C([Br,Cl,I])([CX4])([CX4])([CX4])),C$([CH]([Br,Cl,I])([CX4])([CX4])),C$([CH2]([Br,Cl,I])([CX4])),C$([CH3]([Br,Cl,I])):2].[OH:3][C$(C(O)([CX4,c])([CX4,c])([CX4,c])),C$([CH](O)([CX4,c])([CX4,c])),C$([CH2](O)([CX4,c])),C$([CH3]([OH])),c$(c([OH])([c,n,o])([c,n,o])):4]»[C:2][O:3][C,c:4]
Williamson ether Thiol	[Br,Cl,I:1][C$(C([Br,Cl,I])([CX4])([CX4])([CX4])),C$([CH]([Br,Cl,I])([CX4])([CX4])),C$([CH2]([Br,Cl,I])([CX4])),C$([CH3]([Br,Cl,I])):2].[SH:3][C$(C(S)([CX4,c])([CX4,c])([CX4,c])),C$([CH](S)([CX4,c])([CX4,c])),C$([CH2](S)([CX4,c])),C$([CH3]([SH])),c$(c([SH])([c,n,o])([c,n,o])):4]»[C:2][S:3][C,c:4]
Nucleophilic substitution	[C$([CH3]),C$([CH2]([C,c])),C$([CH]([C,c])([C,c])),C$(C([C,c])([C,c])([C,c])),c$(c([c,n,o])([c,n,o])):1][Cl,Br,I:2].[C$([CH3]),C$([CH2]([C,c])),C$([CH]([C,c])([C,c])),C$(C([C,c])([C,c])([C,c])),c$(c([c,n,o])([c,n,o])):3][OH,SH,NH2,NH3+1:4]»[C,c:1][O,S,N+0:4][C,c:3]

Grignard reaction	[C$([CH3]),C$([CH2]([C,c])),C$([CH]([C,c])([C,c])),C$(C([C,c])([C,c])([C,c])),c$(c([c,n,o])([c,n,o])):1][Cl,Br,I:2].[C$([CH3]),C$([CH2]([C,c])),C$([CH]([C,c])([C,c])),C$(C([C,c])([C,c])([C,c])),c$(c([c,n,o])([c,n,o])):3][C:4](=[O:5])[C$([CH3]),C$([CH2]([C,c])),C$([CH]([C,c])([C,c])),C$(C([C,c])([C,c])([C,c])),c$(c([c,n,o])([c,n,o])):6]»[C:4]([OH:5])([C,c:1])([C,c:3])([C,c:6])
Ester formation Acid Chloride	[Cl:3][C$(C(=O)([CX4,c])),C$([CH](=O)):2]=[O:4].[O$([OH]([CX4,c])),O$([OH]([CX4,c])([CX4,c])):6]»[O:6]−[C:2]=[O:4]
Ester formation Carbox	[OH,O−:3][C$(C(=O)([CX4,c])),C$([CH](=O)):2]=[O:4].[O$([OH]([CX4,c])),O$([OH]([CX4,c])([CX4,c])):6]»[O:6]−[C:2]=[O:4]
Thioester formation Acid Chloride	[Cl:3][C$(C(=O)([CX4,c])),C$([CH](=O)):2]=[O:4].[S$([SH]([CX4,c])),S$([SH]([CX4,c])([CX4,c])):6]»[S:6]−[C:2]=[O:4]
Thioester formation Carbox	[OH,O−:3][C$(C(=O)([CX4,c])),C$([CH](=O)):2]=[O:4].[S$([SH]([CX4,c])),S$([SH]([CX4,c])([CX4,c])):6]»[S:6]−[C:2]=[O:4]
Reductive amination-Ketone	[C$(C(=O)([CX4,c])([CX4,c])):1]=[O:2].[N$([NH2,NH3+1]([CX4,c])),N$([NH]([CX4,c])([CX4,c])):3]»[N+0:3][C:1]
Reductive amination-Aldehyde	[C$([CH](=O)([CX4,c])):1]=[O:2].[N$([NH2,NH3+1]([CX4,c])),N$([NH]([CX4,c])([CX4,c])):3]»[N+0:3][C:1]
Suzuki coupling	[Br:1][c$(c(Br)),n$(n(Br)),o$(o(Br)),C$([CH](Br)(=C)):2].[C$(C(B)([CX4])([CX4])([CX4])),C$([CH](B)([CX4])([CX4])),C$([CH2](B)([CX4])),C$([CH2](B)),C$(C(B)(=C)),c$(c(B)),o$(o(B)),n$(n(B)):3][B$(B([C,c,n,o])([OH,$(OC)])([OH,$(OC)])),B$([B−1]([C,c,n,o])(N)([OH,$(OC)])([OH,$(OC)])):4]»[C,c,n,o:2][C,c,n,o:3]
Piperidine and Indole	[cH:9]1[c:8][n:7][c:5]2[c:4][c:3][c:2][c:1][c:6]12.[N:17]−1−[CX4:16]−[CH:15]−[C:14](=[O:20])−[CX4:19]−[CX4:18]−1»[N:17]−1−[C:18]−[C:19]−[C:14](=[C:15]−[C:16]−1)−[c:9]1[c:8][n:7][c:5]2[c:4][c:3][c:2][c:1][c:6]12
Negishi	[Br,I:1][C$(C([Br,I])([CX4])([CX4])([CX4])),C$([CH]([Br,I])([CX4])([CX4])),C$([CH2]([Br,I])([CX4])),C$([CH3]([Br,I])),C$([C]([Br,I])(=C)([CX4])),C$([CH]([Br,I])(=C)),C$(C([Br,I])(#C)),c$(c([Br,I])):2].[Br,I:3][C$(C([Br,I])([CX4])([CX4])([CX4])),C$([CH]([Br,I])([CX4])([CX4])),C$([CH2]([Br,I])([CX4])),C$([CH3]([Br,I])),C$([C]([Br,I])(=C)([CX4])),C$([CH]([Br,I])(=C)),C$(C([Br,I])(#C)),c$(c([Br,I])):4]»[C,c:2][C,c:4]
Mitsunobu imide	[C$(C(C)([CX4])([CX4])([CX4])),C$([CH](C)([CX4])([CX4])),C$([CH2](C)([CX4])),C$([CH3](C)):1][C:2](=[O:3])−[NH:4]−[C:5]([C$(C(C)([CX4])([CX4])([CX4])),C$([CH](C)([CX4])([CX4])),C$([CH2](C)([CX4])),C$([CH3](C)):10])=[O:7].[OH:11]−[C$(C(O)([CX4])([CX4])([CX4])),C$([CH](O)([CX4])([CX4])),C$([CH2](O)([CX4])),C$([CH3](O)):9]»[C:9][N+0:4](−[C:2]([C:1])=[O:3])−[C:5]([C:10])=[O:7]
Mitsunobu carboxlyic acid	[OH,O−]−[C$(C(=O)(O)([CX4,c])):2]=[O:3].[OH:8]−[C$([CH](O)([CX4,c])([CX4,c])),C$([CH2](O)([CX4,c])),C$([CH3](O)):6]»[C:6][O]−[C:2]=[O:3]
Mitsunobu sulfonic amide	[OH:1]−[C$([CH](O)([CX4,c])([CX4,c])),C$([CH2](O)([CX4,c])),C$([CH3](O)):3].[N$([NH](S)([CX4])),N$([NH2,NH3+1](S)):9][S$(S(N)([CX4])):6](=[O:7])=[O:8]»[C:3][N+0:9][S:6](=[O:8])=[O:7]
Heck	[C$([CH](=C)([CX4])),C$([CH2](=C)):2]=[C$(C(=C)([CX4])([CX4])),C$([CH](=C)([CX4])),C$([CH2](=C)):3].[Br,I:7][C$([CX4]([Br,I])),c$([c]([Br,I])):4]»[C,c:4][C:2]=[C:3]

Amide formation Carbox	[OH,O−:3][C$(C(=O)([CX4,c])),C$([CH](=O)):2]=[O:4].[N$([NH2,NH3+1]([CX4,c])),N$([NH]([CX4,c])([CX4,c])):6]»[N+0:6]−[C:2]=[O:4]
Amide formation Acid Chloride	[Cl:3][C$(C(=O)([CX4,c])),C$([CH](=O)):2]=[O:4].[N$([NH2,NH3+1]([CX4,c])),N$([NH]([CX4,c])([CX4,c])):6]»[N+0:6]−[C:2]=[O:4]
Thiolether formation-Alkene	[C$(C(=C)([CX4])([CX4])),C$([CH](=C)([CX4])),C$([CH2](=C)):1]=[C$(C(=C)([CX4])([CX4])),C$([CH](=C)([CX4])),C$([CH2](=C)):2].[SH:4]−[CX4:5][Br,Cl,I]»[C:1]−[C:2]−[S:4][C:5]
Thiolether formation-Carboxlyic acid	[C$([C](=O)([CX4])),C$([CH](=O)):2](=[O:1])[OH,Cl,O−:6].[SH:4]−[CX4:5][Br,Cl,I]»[CH2:2]−[S:4][C:5]
Ketone formation	[I:1][C$(C(I)([CX4,c])([CX4,c])([CX4,c])),C$([CH](I)([CX4,c])([CX4,c])),C$([CH2](I)([CX4,c])),C$([CH3](I)):2].[C$(C(=O)([Cl,OH,O−])([CX4,c])),C$([CH]([Cl,OH,O−])(=O)):3](=[O:6])[Cl,OH,O−:5]»[C:2]−[C:3]=[O:6]
Sulfonamide formation Sulfonic Acid	[OH,O−:5][S$(S(=O)(=O)(O)([CX4])):2](=[O:3])=[O:4].[NH2+0,NH3+:6]−[C$(C(N)([CX4,c])([CX4,c])([CX4,c])),C$([CH](N)([CX4,c])([CX4,c])),C$([CH2](N)([CX4,c])),C$([CH3](N)),c$(c(N)):7]»[C,c:7]−[NH+0:6][S:2](=[O:4])=[O:3]
Sulfonamide formation Sulfonyl Chloride	[Cl:5][S$(S(=O)(=O)(Cl)([CX4])):2](=[O:3])=[O:4].[NH2+0,NH3+:6]−[C$(C(N)([CX4,c])([CX4,c])([CX4,c])),C$([CH](N)([CX4,c])([CX4,c])),C$([CH2](N)([CX4,c])),C$([CH3](N)),c$(c(N)):7]»[C,c:7]−[NH+0:6][S:2](=[O:4])=[O:3]
Ar-Imidazole formation	[c:5]1[c:4][nH:3][c:2][n:1]1.[OH,$(OC):13]−[B:12](−[OH,$(OC):14])−[c:6]1[c:7][c:8][c:9][c:10][c:11]1»[c:4]1[c:5][n:3]([c:2][n:1]1)−[c:6]1[c:7][c:8][c:9][c:10][c:11]1
Alkyne alkylation	[*:1][C:2]#[CH:3].[Br,I:4][C$(C([CX4,c])([CX4,c])([CX4,c])),C$([CH]([CX4,c])([CX4,c])),C$([CH2]([CX4,c])),C$([CH3]),c$(c):5]»[C,c:5][C:3]#[C:2][*:1]
Alkyne acylation	[C$(C(C)([CX4])([CX4])([CX4])),C$([CH](C)([CX4])([CX4])),C$([CH2](C)([CX4])),C$([CH3](C)):1][C:2]#[CH:3].[Br,I:4][C$(C(=O)([Br,I])([CX4])),C$([CH](=O)([Br,I])):5]=[O:6]»[C:1][C:2]#[C:3][C:5]=[O:6]
FGI Acyl chloride	[OH,O−:4]−[C$(C(=O)([OH,O−])([CX4])),C$([CH](=O)([OH,O−])):2]=[O:3]»[Cl:5][C:2]=[O:3]
FGI bromination	[OH:2]−[$([CX4]),c:1]»[Br:3][C,c:1]
FGI chlorination	[OH:2]−[$([CX4]),c:1]»[Cl:3][C,c:1]
FGI sulfonyl chloride	[OH,O−:3][S$(S([CX4])):2](=[O:4])=[O:5]»[Cl:6][S:2](=[O:5])=[O:4]
FGA alpha bromination	[OH+0,O−:5]−[C:3](=[O:4])−[C$([CH]([CX4])),C$([CH2]):2]»[OH+0,O−:5]−[C:3](=[O:4])−[C:2]([Br:6])
FGA alpha chlorination	[OH+0,O−:5]−[C:3](=[O:4])−[C$([CH]([CX4])),C$([CH2]):2]»[OH+0,O−:5]−[C:3](=[O:4])−[C:2]([Cl:6])
FGI Rosenmund-von Braun	[Cl,I,Br:7][c:1]1[c:2][c:3][c:4][c:5][c:6]1»[N:9]#[C:8][c:1]1[c:2][c:3][c:4][c:5][c:6]1
FGI nitrilation	[OH,NH2,NH3+:3]−[CH2:2]−[C$(C([CX4,c])([CX4,c])([CX4,c])),C$([CH]([CX4,c])([CX4,c])),C$([CH2]([CX4,c])),C$([CH3]),c$(c):1]»[C,c:1][C:2]#[N:4]

A.4.2 *De Novo* Designs with the YLT-11 Template Ligand

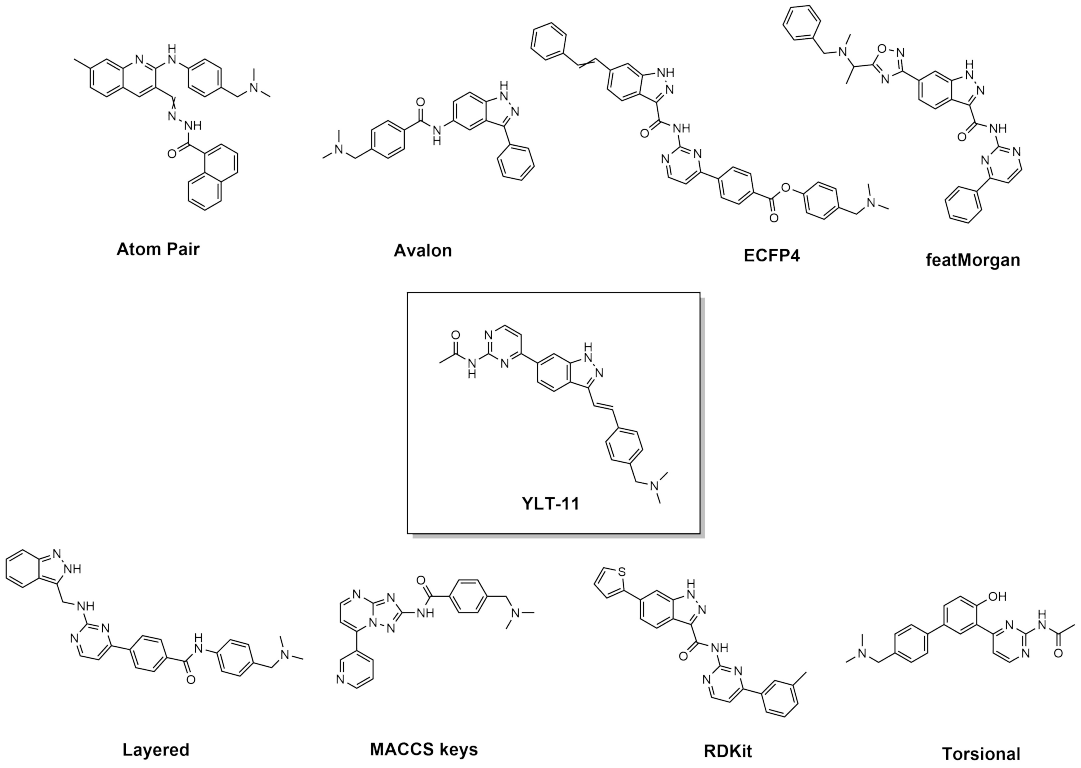
The primary goal of DINGOS is to create synthesizable structures that are related to a template ligand as a means of determining the structural dependence for the

observed bioactivity, and to potentially produce novel bioactive "hits". To emulate the scenario in which a chemist is working with a known bioactive and wishes to produce similar compounds, we chose the PLK4 kinase inhibitor YLT-11 [160] as a template ligand. The PLK4 kinase, despite being a potential therapeutic target for antitumour therapy [161], possesses few small-molecule inhibitors under clinical trial, with the underlying mechanism of inhibition still not fully known.

YLT-11 was originally developed by Liu *et al.* as part of a structure activity relationship study for the PLK4 kinase [162]. Based of existing reported actives, the scaffold (E)-4-(3-arylvinyl-1H-indazol-6-yl)pyrimidin-2-amine was chosen for the study and the corresponding R groups were modified, thus yielding YLT-11. Lei *et al.* [160] selected YLT-11 for further testing and analysis, due to its selectivity towards PLK4 over the other subgroups (>200-fold selectivity against PLK1, PLK2 and, PLK3). Lei *et al.* confirmed preferential activity towards triple-negative breast cancers (TNBC) cells over mammalian cell lines, showing an IC_{50} of between 68-120 nM. Molecular docking of YLT-11 showed binding in the ATP-binding pocket of PLK4, indicating that it was likely an ATP-competitive inhibitor. Hydrogen bonding between the indazole ring nitrogens and the catalytic Cysteine and Glutamic acid residues, as well as between the pyrimidine ring and Lysine residue, was observed. These results are consistent with hydrogen bonding patterns determined from other co-crystallized PLK4 inhibitors [163].

Eight individual *de novo* design experiments were performed, each with a different descriptor representation (see Methods 5.2.1). The same parameters as in Methods 5.2.6 were used, with the exception of the building block recommendation pool, which was set to 1000. This yielded eight sets of 300 *de novo* designed molecules. Figure A.9 shows the top ranked designs from each of the *de novo* descriptor populations. Of the eight designs, four contained the indazole and pyrimidine subgroups.

FIGURE A.9: Most similar *de novo* designs obtained for the YLT-11 template, using the eight separate descriptor representations. Of the eight designs, both the indazole and pyrimidine subgroups were only observed for the the layered, RDKit, ECFP4, and featMorgan designs.



A.4.3 Relative Standard Deviation

For a sample $X_1, X_2, ..., X_N$ we define the relative standard deviation as

$$\sigma_r = \frac{\sigma}{\mu}$$

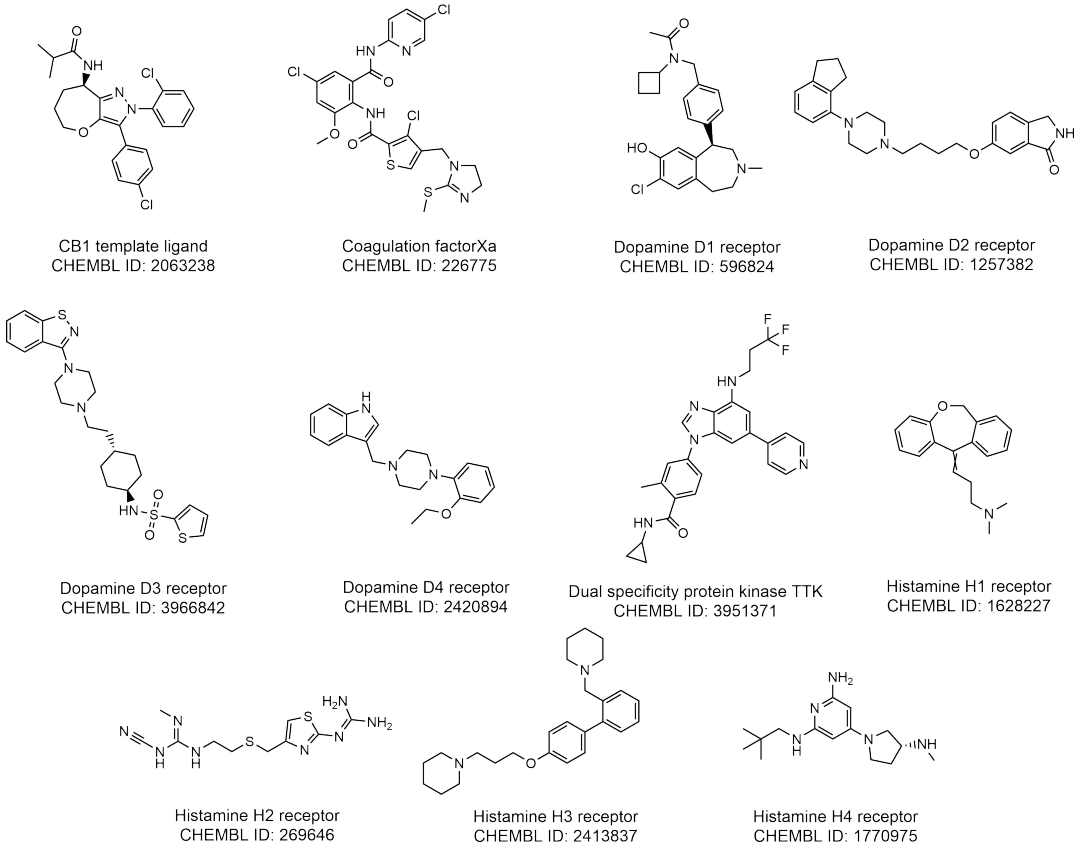
Where μ and σ are defined as follows:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i \quad \text{and} \quad \sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu)^2}$$

A.4.4 Template Ligands Extracted from ChEMBL

The template ligands that were selected in Chapter 5 via the bioactivity analysis are shown in Figure A.10.

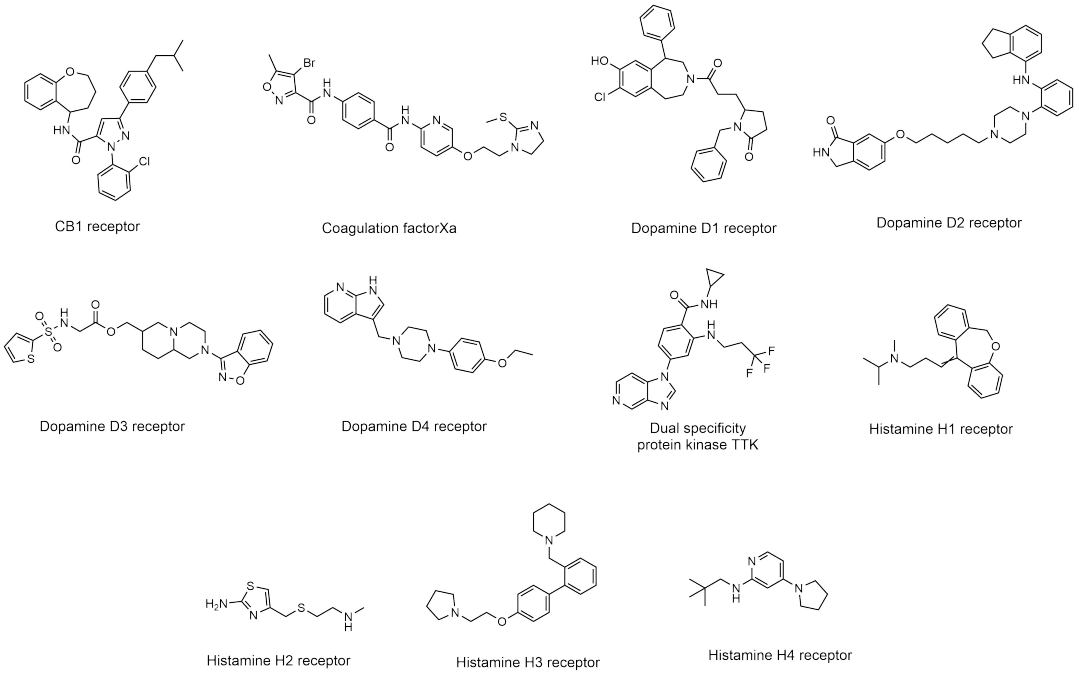
FIGURE A.10: Structures of the eleven template ligands selected for *de novo* design in Chapter 5. Ligands are shown with their corresponding ChEMBL IDs and biological targets.



A.4.5 Top *De Novo* Designs Generated by DINGOS-BGEN

Figure A.11 shows the eleven most similar *de novo* designs produced from the DINGOS-BGEN experiments presented Section 5.3.4.

FIGURE A.11: Structures of the eleven top *de novo* designs generated in Chapter 5 by the DINGOS-BGEN model for the eleven template ligands (see Figure A.10). Compounds are labeled by the targets of their corresponding template ligand.

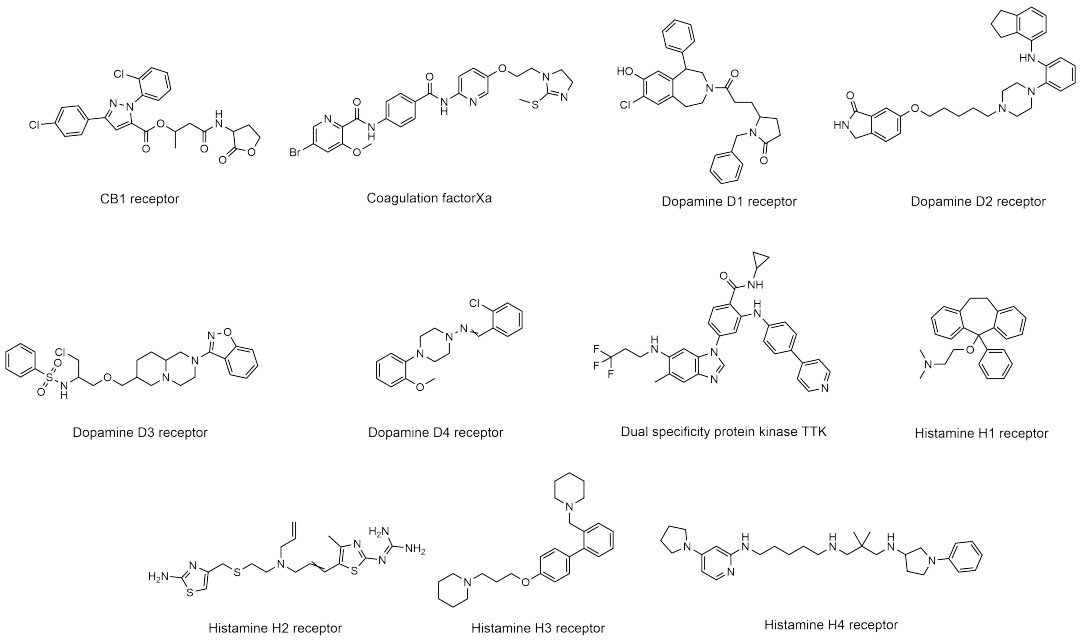


A.5 Chapter 6

A.5.1 Top *De Novo* Designs Generated by DINGOS-MCTS

The eleven most similar *de novo* designs produced in Section 6.3.4 are shown in Figure A.12.

FIGURE A.12: Most similar *de novo* designs produced in Section 6.3.4. The structures shown in Figure A.10 were used as the template ligands. Compounds are labeled by the targets of their corresponding template ligand.



Curriculum vitae

Alexander Luke Button

Date & Place of Birth:	10 October 1991, Australia
	Australian Citizen, British Citizen by Descent
Address:	Unterfeldstrasse 104, 8050 Zurich
Phone:	078 854 0026
E-mail:	alexander.button@pharma.ethz.ch

Academic History

Bachelors of Science (Nanoscience and Materials) with Honours in Chemistry, First Class (2009-2013, University of Adelaide)

- Majors in Chemistry and Applied Mathematics
- Obtained First Class Honours which is the highest mark possible (grade: 1)

Honours year project – 'A mechanistic study into the destabilization of DNA Triplexes' (February 2013 – November 2013, University of Adelaide)

- Performing molecular dynamics simulations of DNA triplex structures
- Synthesizing DNA triplexes in biologically relevant conditions
- Analysing triplex samples by Ion-mobility mass spectrometry (IM-MS), UV-Vis absorption and gel electrophoresis

Doctoral thesis (October 2015 – February 2020, ETH Zurich)

Industrial experience

CSIRO (Commonwealth Scientific and Industrial Research Organisation) Summer Scholarship (November 2012 – February 2013, CSIRO)

- Development of algorithms based on Quantum Monte Carlo methods (QMC)
- Calculating electronic energies by various methods (Hartree-Fock, DFT, QMC)
- Submitting code to the national supercomputer
- Collecting data and presenting in meetings and seminars

Research and teaching experience

Research Assistant at the University of Adelaide with Dr. David Huang (2014, University of Adelaide)

- Constructing kinetic models of DNA binding
- Implementing Monte Carlo computer simulations using MatLab

Practical demonstrator (February – November 2013, University of Adelaide)

- Supervised and directed students while they undertook chemistry practical's
- Marked practical reports written on the experiments

Achievements

14th German Conference on Chemoinformatics

- Won the GCC 2018, 14th German Conference on Chemoinformatics poster prize

Undergraduate degree

- Received outstanding academic achievement award two years in a row (2011, 2012)
- Additional studies extra to my course criteria: Mathematics 1B, Physics 2A

Own Publications

1. Schneider, P., Müller, A. T., Gabernet, G., Button, A. L., Posselt, G., Wessler, S., Hiss, J. A. and Schneider, G. (2017) Hybrid network model for "deep learning" of chemical data: Application to antimicrobial peptides. Mol. Inf. **36**, 1600011.
2. Button, A. L., Hiss, J. A., Schneider, P. and Schneider, G. (2017) Scoring of de novo designed chemical entities by macromolecular target prediction. *Mol. Inf*. **36**, 1600110.
3. Li, J., Begbie, A., Boehm, B. J., Button, A., Whidborne C., Pouferis, Y., Huang, D. M., Pukala, T. L. (2018) Ion Mobility-Mass Spectrometry Reveals Details of Formation and Structure for GAA·TCC DNA and RNA Triplexes. J. Am. Soc. Mass Spectrom. **30**, 103-112
4. Boehm, B., Whidborne, C., Button, A., Pukala, T., Huang, D. (2018). DNA triplex structure, thermodynamics, and destabilisation: insight from molecular simulations. *Physical Chemistry Chemical Physics*, **20**, 14013-14023.
5. Button, A., Merk, D., Hiss, J. A. and Schneider, G. (2019) Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis. *Nature Mach. Intell*. **1**, 307–315

Personal Skills

Computing

Programming – C, C++, Python, Matlab, mysql, Django, Labview
Software – KNIME, Gaussian, Gaussview, LAMMPS, AMBER (including ANTECHAMBER and NAB), PuTTY, Masslynx, Matlab

Languages

- English (Native Speaker)
- German (B1)
- Mandarin (Beginner)

References

Adelaide University

1) Dr. David Huang
   Theoretical Chemist – *Honours Academic Supervisor*
   University of Adelaide
   david.huang@adelaide.edu.au

2) Dr. Tara Pukala
   Biological Chemist – *Honours Academic Supervisor*
   University of Adelaide
   tara.pukala@adelaide.edu.au

Virtual Nanoscience Laboratory – CSIRO
Dr. Amanda Barnard
Theoretical and Computational Physics
Virtual Nanoscience Laboratory – CSIRO
Amanda.Barnard@csiro.au