Thesis for The Degree of Licentiate of Engineering

Intelligent data acquisition for drug design through combinatorial library design

Simon Johansson

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2023

Intelligent data acquisition for drug design through combinatorial library design

Simon Johansson

ISSN 1652-876X

Department of Computer Science and Engineering
Division of Data Science and AI
Chalmers University of Technology | University of Gothenburg
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2023.

To my parents

Intelligent data acquisition for drug design through combinatorial library design

Simon Johansson

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg

Abstract

A problem that occurs in machine learning methods for drug discovery is a need for standardized data. Methods and interest exist for producing new data but due to material and budget constraints it is desirable that each iteration of producing data is as efficient as possible. In this thesis, we present two papers methods detailing different problems for selecting data to produce. We investigate Active Learning for models that use the margin in model decisiveness to measure the model uncertainty to guide data acquisition. We demonstrate that the models perform better with Active Learning than with random acquisition of data independent of machine learning model and starting knowledge. We also study the multi-objective optimization problem of combinatorial library design. Here we present a framework that could process the output of generative models for molecular design and give an optimized library design. The results show that the framework successfully optimizes a library based on molecule availability, for which the framework also attempts to identify using retrosynthesis prediction. We conclude that the next step in intelligent data acquisition is to combine the two methods and create a library design model that use the information of previous libraries to guide subsequent designs.

Keywords

Cheminformatics, machine learning, drug discovery, generative models, active learning, determinantal point processes

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] **S.V. Johansson**, H. Gummesson Svensson, E. Bjerrum, A. Schliep, MH Chehreghani, C. Tyrchan, O. Engkvist, *Using Active Learning to Develop Machine Learning Models for Reaction Yield Prediction Molecular Informatics 41 (June 2022), 2200043.* doi.org/10.1002/minf.202200043

[**Paper II**] **S.V. Johansson**, M.H. Chehreghani, O. Engkvist, A. Schliep, *de novo generated combinatorial library design Submitted, under review.*

Other publications

The following publications were published during my PhD studies, or are currently in submission/under revision. However, they are not appended to this thesis, due to contents overlapping that of appended publications or contents not related to the thesis.

[**a**] O. Prykhodko, **S.V. Johansson**, P.C. Kotsias, J. Arús-Pous, E. Bjerrum, O. Engkvist, H. Chen, *A de novo molecular generation method using latent vector based generative adversarial network*
Journal of Cheminformatics 11 (December 2019), 1-13.
https://doi.org/10.1186/s13321-019-0397-9

[**b**] J. Arús-Pous, **S.V. Johansson**, O. Prykhodko, E. Bjerrum, C. Tyrchan, J-L. Reymond, H.Chen, O. Engkvist, *Randomized SMILES string improve the quality of molecular generative models*
Journal of Cheminformatics 11 (December 2019), 1-13.
https://doi.org/10.1186/s13321-019-0393-0

[**c**] D. Polykovskiy, A. Zhebrak B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, A. Kadurin, **S.V. Johansson**, H. Chen, S. Nikolenko, A. Aspuru-Guzik, A. Zhavoronkov, *Molecular sets (MOSES): a benchmarking platform for molecular generation models*
Frontiers in pharmacology 11 (December 2020), 565644.
https://doi.org/10.3389/fphar.2020.565644

[**d**] L.H. Mervin, **S.V. Johansson**, E. Semenova, K.A. Giblin, O. Engkvist, *Uncertainty quantification in drug design*
Drug Discovery Today 26 (February 2021), 474-489.
https://doi.org/10.1016/j.drudis.2020.11.027

[**e**] A. Thakkar, **S.V. Johansson**, K. Jorner, D. Buttar, J-L. Reymond, O. Engkvist, *Artificial intelligence and automation in computer aided synthesis planning*
Reaction Chemistry & Engineering 6 (November 2020), 27-51.
https://doi.org/10.1039/D0RE00340A

[**f**] **S.V. Johansson**, A. Thakkar, T. Kogej, E. Bjerrum, S. Genheden, T. Bastys, C. Kannas, A. Schliep, H. Chen, O. Engkvist, *AI-assisted synthesis prediction*
Drug Discovery Today: Technologies 32-33 (December 2019), 65-72.
https://doi.org/10.1016/j.ddtec.2020.06.002

Acknowledgment

First, I would like to thank my supervisor Alexander Schliep, for his support and understanding. I would also like to thank my industrial supervisor, Ola Engkvist, for his ideas and research expertise. I would like to thank my co-supervisor Morteza Hagir Chehreghani, for his technical knowledge. I want to thank my examiner, Graham Kemp, for our interesting discussions.

I want to thank my fellow PhD colleagues at Chalmers: Markus, David, Firooz, Mehrdad, Tobias, Cristopher, Emilio, Mena, Hannes, Niklas, Arman, Linus, Emil, Alexander, Filip, Daniel, Lena, Newton, Adam and Lovisa. I also want to thank Rocío, for her support both as a current and former colleague. I want to especially thank my colleagues with whom I have shared an office: Hanna, Deepthi, Fazeleh and Denitsa. I want to thank the rest of the DSAI division; faculty, administrative staff and post-docs.

I want to thank everyone in the Molecular AI department at AstraZeneca for many interesting discussions: Pallavi, Tomas, Lili, Michael, Rosa, Samuel, Gökçe, Peter, Jiazhen, Preeti, Jon Paul, Mikhail, Christos, Marco, Thierry Hannes, Thomas, Emma, Lakshidaa, Emma, Alessandro, Bob, Varvara, Vincenzo, Yasmine, Alexey, Annie, Michele, Helen and Lewis. A special thanks to my fellow industrial PhD colleagues Juan and Hampus, with whom I am a colleague twice over and whose shared experiences has given me immeasurable support.

I want to thank former colleagues for being part in a work environment that made me want to pursue my PhD: Atanas, Hongming, Josep, Amol, Esben, Panagiotis, Laurianne and Michael.

I want to thank my friends Iliyan, Jisoo, Stefan and Stefaan, whom despite never having physically met me has continuously provided support and ideas both intentionally and unintentionally.

I also want to thank my close long time friends Sebastian, Lukas, Björn, Rickard, Jonathan and Helmer, Ida, Honoka and Amanda, for reminding me that work is not everything and that it sometimes can be productive to have a break.

Finally, I want to thank my family; my father Kenneth, my mother Berit and my sister Sara, for the continuous support through everything.

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation

Contents

Part I

Summary

Chapter 1

Introduction

Data science and artificial intelligence have made remarkable strides in development in every field from recent developments in natural language processing to pharmacology. In drug discovery, the progress of machine learning research is often limited by the availability of data [1]. Reasons that the availability of data is low could be that the data is proprietary to a company, or unreported because it was an unsuccessful experiment. Additionally, chemistry data historically could be recorded formats that require processing before it could be used in machine learning, such as free text in lab notebooks. As the need for data has grown, interested parties have looked to producing chemistry data at larger scale to better train models.

Combinatorial chemical libraries are used to select and produce chemistry data for a group of molecules that share a common purpose; whether that be model training in machine learning, lead optimization in hit discovery or patent protection [2]. The appeal of combinatorial chemistry lies in the material efficiency since the reagents are shared between the reactions, thereby reducing the number of different compounds needed to be acquired before synthesis. The different goals of library design can be *focused* [3],[4],[5], i.e. oriented towards optimizing around a target area of the chemical space to find variations on a lead compound that optimize some chemical property or *diversity-driven*[6][7], i.e. covering a large part of the chemical space to reduce redundancy and increase the information that can be derived per experiment. In practice, the weighting between the goals is continuous and changing depending on where in the hit discovery process that a project currently is located. This is a case of the trade-off between *explore* and *exploit* that is present in decision making.

Still, even producing millions of compounds is but a drip in the ocean, as the total number of synthetically feasible molecules is estimated to be $> 10^{60}$ [8]. Virtual libraries are now a popular alternative as the storage capability of computers has grown to the capacity to store hundreds of billions of virtual compounds for virtual screening, but this is especially sensitive to false positive rates [9], [10]. Recently, several generative models have been developed for library design that are capable of creating *building blocks*, through *de novo* design [11], which might bring library design back into focus when novel compounds

are generated. However, the generative models are capable of producing more suggestions than can feasibly be synthesized by a chemist and such a compound selection method is needed. Moreover, the published models that do provide a ranking of products do not provide their selection from a combinatorial design but rather a cherry-picked selection, which in the worst case can have unique building blocks for all suggestions. Methods for library optimization exist, but assume that all molecules are feasible; the concept that some of the suggested building blocks might be impossible to acquire in practice is not considered, when building block providers with synthesis on demand services have a success rate of 76% [12].

This research focuses on the development of frameworks that can bridge the gap between current generative models and the practical constraints that exist for an actor in drug discovery.

This thesis consist of two papers. **Paper I** is a retrospective study on two combinatorial data sets of high-throughput experimentation for reaction data. Here, we studied the robustness of *Active Learning* [13] for improving the training of predictive machine learning models. We found in the case that active learning could suggest data better than random selection, that the effect was consistent when varying the machine learning model. We also found that even when the model had next to no initial information about the data, active learning still performed at least as good as random selection, and eventually resulted in learning the task with less data required. **Paper II** is the design of a framework for a combinatorial library from *de novo* generation of building blocks, to evaluation of building block availability through retrosynthesis and finally the multi-objective optimization of the library using both quality (exploit) and diversity (explore) metrics. Here we propose the grouping of building blocks by the number of reaction steps needed for acquisition. Further we simulate the case of limited building block availability for a single actor with limited building block stock and show that we can estimate the marginal gain between using the building blocks available to the actor and extending the design to acquisition through synthesis.

Chapter 2

Background

This section covers the chemistry and machine learning background for the papers in this thesis. The chapter starts by covering basic organic chemistry and notation behind the chemical representations that were used in the papers. I then briefly describe the drug discovery process. The following section discusses the areas within early-stage drug discovery which are relevant to this thesis. Finally, I cover the methodology of the computational techniques that were used in the papers.

2.1 Chemistry and chemical representations

In chemistry and cheminformatics, the same molecule can be represented in numerous ways, from the organic chemistry notation to computer based featurization. In this section we summarize the representation methods that are used in the appended papers. This is not an exhaustive list of the available methods.

Molecular Fingerprints are vector representations of a molecule. These can vary from a set of physiochemical descriptors to an encoding of the molecular structure. The most common representations are the Extended-Connectivity Fingerprints (ECFPX) [14], a hashing of the substructures present in the molecule, where X is the the diameter of the structure. As an example, the ECFP4 fingerprint is hashes substructures where for each atom structures up to a "radius" of two atoms away are encoded. The most common version of the ECFP fingerprint is a fixed-length bit vector, which is a folded version of the sparse bit-vector of all possible substructures into a vector constant length. The most common sizes of bit vectors for this fingerprint are 1024 and 2048. As with any folded vector representation, there is a risk of bit-collision and in particular, this risk increases as the size of the molecule increases. In addition, for larger molecules, the ECFP has difficulty capturing some differences between two molecules such scrambled order of amino acids in two peptides of the same length and total composition. Thus, the ECFP fingerprints are suitable mostly for small molecule applications.

SMILES or the Simplified Molecular-Input Line-Entry System [15], is a string representation of the skeletal structure. Starting from an atom in the structure, and traversing along the longest possible path in the molecular graph without revisiting any atom denotes the order of visited atoms. Any branching paths are denoted within brackets and connections such as ring closures are denoted with pairs of digits, where e.g. two atoms that are followed by the digit '1' are connected. An example for the translation of a molecule into a SMILES string is illustrated in Figure 2.1.

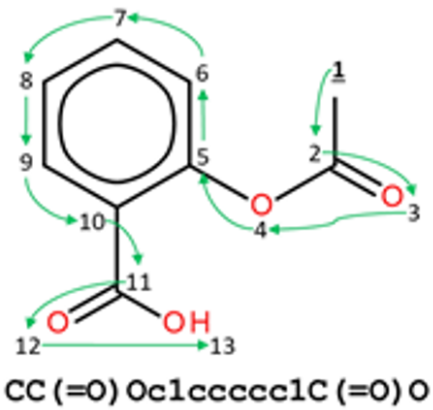

Figure 2.1: Example of the canoncical order to create a SMILES string for an organic molecule. The example shown is for Aspirin. Figure extracted, with permission, from original work by [16].

2.2 The Drug Discovery Process

The drug discovery process is the first step in the process from early stage to released drug. The steps that are included during the discovery stage are as follows:

Target Identification: The process of identifying a molecule or pathway in the body, which might play a crucial role in the disease or condition that the drug is supposed to treat. This target can be e.g. a protein, enzyme or receptor. Classical methods for target identifications are genomics, proteomics, bioinformatics and phenotype-oriented identifications. [17], with recent developments using machine learning [18].

Target Validation: When a target has been identified, a testing process is performed to validate that the target indeed is significant for the disease. Methods that are used for this could be in-silico models, gene knockouts, RNA interference or in-vivo experiments in animal testing [19].

Hit Generation: The process of finding possible molecules that interact with the target to change the activity. Methods for this are various screening methods, library design or *de novo* design [20].

Lead Optimization: The hits that are found during the hit generation might have good interaction with the target, but have undesirable ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties, or perhaps the hit is not selective enough [21]. This is an iterative process where small modifications are made to the structure. These are then examined by experimental ways such as magnetic resonance and mass spectrometry, or by computational methods, pharmacophore studies, molecular docking, molecular dynamics and QSAR [22].

Lead Compound Selection: The most promising hits are selected. Factors that play a role in this selection are based on the pharmacological properties and absence of toxicity, but also chemical properties such as synthesize-ability and possibility of scale-up to production levels [23].

Following this, the candidate drugs leave the discovery stage and enter the drug development stage, which is not covered by this thesis.

2.2.1 Connection to own research

In this section I introduce the drug discovery problems that are related to the papers included in this thesis. These problems are commonly encountered in the Hit Generation and Lead Optimization stages of the drug discovery process.

2.2.1.1 Library design and combinatorial chemistry

Combinatorial library design is a method for producing collections of molecules in an economically and materially efficient manner [20],[2]. Suppose that a number of different reagents, or *building blocks* with the same reactive region are available for reaction synthesis. Combinatorial chemistry is used when multiple building block types are used in the same library design. As an example, if a study has identified that a common central molecule, a *scaffold* [24] has good properties and can be patented, a library could be designed with purpose to explore the attachment of two building blocks, A and B for how good the product binds to a target. The material advantage of combinatorial design becomes clear, since for a 10×10 design, 100 products are produced, whereas the worst case design could use 100 separate building blocks A and B each.

Library design can be either *focused* [3],[4],[5], optimizing around a target area of the chemical space to find variations on a lead compound that optimize some chemical property or *diversity-driven*[6][7], i.e. covering a large part of the chemical space to increase the information that can be derived per

experiment and possibly increase the applicability domain of the molecular property modeling. The multi-objective optimization of focused library goals such as QSAR value and QED score as well the diversity (see 2.3) is a central research question we contribute to in **Paper II**.

2.2.1.2 *de novo* design

The formal definition of *de novo* design is "the design of bioactive compounds by incremental construction of a ligand model within a model of the receptor or enzyme active site, the structure of which is known from X-ray or nuclear magnetic resonance (NMR) data" [11]. This field of research has benefited greatly from development in machine learning [25]. A common problem is that the available training data for a specific target is too small for most most models to learn both the features of the target domain and to generate sensible drug-like molecules at a high rate. Typically this is solved by first training an agent on a larger dataset of valid molecules, such as ZINC [26] or ChEMBL [27], followed by a fine-tuning of the model towards the drug target by transfer learning or reinforcement learning. In **Paper II** we use this technique with the generative model LibINVENT [28] to generate building blocks to attach to a scaffold in order to create the data set we want to optimize around.

2.2.1.3 Synthesis prediction

Synthesis prediction is a field related to the production of a molecule. Synthesis problems are generally of two different categories: forward synthesis prediction and retrosynthesis prediction [29]. Forward synthesis prediction tries to answer questions regarding where a reaction is attempted involving given reactants. Will the reactants interact? What are the resulting product(s)? There are a number of reaction condition variables that can affect the reaction, such as temperature, catalyst, solvent and other additives that could be optimized.

In **Paper I** we examine the performance of models for predicting the reaction yield when reaction conditions are changed for the same reaction. Retrosynthesis instead addresses problems from the product end of the reaction. Given a product, retrosynthesis predictions attempt to compute which products that were used in the formation of the product. Retrosynthesis prediction is used in **Paper II** to evaluate the availability of molecules that are generated through *de novo* design.

2.3 Evaluating molecules for drug discovery

For both *de novo* design and library design a success metric is needed as an optimization goal. These can range from a target lipophilicity, minimization of toxicity or bio-activity. The following are metrics that are used in **Paper II**:

QSAR is a family of modeling approaches which use the assumption that similar molecules should have similar properties [30]. QSAR modeling consist of a regression or classification task with predictor features to describe the

molecule. These features could use physiochemical descriptors or molecular features of the structure, such as molecular fingerprints, a graph of the structure or directly use the SMILES. For **Paper II**, QSAR models are used to model the probability that the molecules can inhibit the dopamine receptor D2 (DRD2).

QED is an estimate of the drug-likeness of a molecule based on the distribution of molecular properties of known drugs [31]. This is not by itself an indicator of the suitability for a molecule as a drug, but shows correlation to features common in small molecule drugs. In general, the QED score favours compounds that are chemically accessible rather than to complex and prefer molecular structures that are not too large. The estimated difference in medians behind molecules that were deemed attractive and those that were considered unattractive was around 0.164.

Chemical Diversity is desired in library design as it represent less redundancy in experiments and higher information gain. There is no convention or formal definition for chemical, or molecular diversity. The main reason is that there are several metrics and properties for which molecules can be compared. These metrics can be based in physical chemistry, bio-activity or molecular structure [32]. Diversity measures most commonly belong to one of the following categories [33]:

- **Distance-based diversity**, based on the pairwise distances between the molecules in a space spanned by the observed metrics or
- **Cell-based diversity**, where the chemical space is divided into distinct regions and diversity is measured as number of occupied cells or,
- **Variance-based diversity**, where diversity is measured as the correlation between the molecules in the chosen metrics.

Paper II in this thesis use a distance-based diversity. It uses Tanimoto similarity [34], also known as the Jaccard index of the ECFP6 fingerprint to define the pairwise similarity between molecules. Most library design methods that optimize diversity based on a similarity metric attempt to find the minimal average of similarities (minAvg), or the minimum of the maximal similarity for each selected molecule [35] [36]. However, due to the number of alternatives in a combinatorial design, brute force solutions are practically infeasible. Thus, the common solution methods are either greedy if the sole objective is diversity, or in the case of multi-objective library designs, based on simulated annealing (SA) [37], or genetic algorithms (GA) [36] [35]. In **Paper II** however, the diversity is instead measured as the logdeterminant of the pairwise similarities. This kernel allows for the use of determinantal point processes (DPPs) to be used as a sampling method (see 2.4.1).

2.4 Computational techniques

In this section, I will describe the determinatal point processes (DPPs). The section then covers the machine learning architectures used in both papers,

and some alternative models that can be used for the same tasks.

2.4.1 Determinantal Point Processes

The Determinantal Point Process is a probabilistic method originally used to measure the repulsion between fermions [38]. The method has gained popularity in the tasks of text summarization and diverse image selection, as it is capable of modeling the trade-off between quality, such as the relevance of the text or image, and the diversity [39]. They can be defined as follows: Let $L \in \mathbb{R}^{n \times n}$ be a positive semi-definite (PSD) kernel. A discrete DPP with kernel L is a probability distribution $\mu : 2^{[n]} \to \mathbb{R}_+$ defined by

$$\mu(S) \propto Det(L_S), \ \forall S \subseteq [n], \tag{2.1}$$

where L_S is a principal submatrix of L indexed by the elements of S. If S is the selection of molecules chosen for the library design, the rows of this matrix are feature vectors that represent each molecule's similarity to the other molecules in the selection. The probability that the DPP would sample a particular selection is proportional to the volume of the hull spanned by the feature vectors. For library design we are interested in selections of a fixed size k, and we condition the DPP such that only selections of size k have a non-zero probability. This version of DPP is called a k-DPP. Gharan and Rezai [40] showed that it was possible to draw samples from a k-DPP by using Gibbs sampling [41] and transitioning from different selection states by exchanging one element of S per step and moving with a probability proportional to $Det(L_S)$. The time complexity per time step is $O(k^3)$

2.4.2 Machine Learning architectures

For the processing of the used chemical representations there are a number of machine learning models that can be used in the studied drug discovery problems. The choice of model is another parameter that can be optimized just as the hyper-parameters internally for the models. For a given representation and fixed available data, the classification accuracy of two different architectures is generally within a couple of percent of each other [42], [43]. Since the research primarily focused on framework development, an exhaustive list of machine learning models was not explored. The following models were used in the framework:

Random Forest This is an ensemble learning method of multiple decision trees that each are created on a random subset of the training data [44]. Each individual decision tree is formed by the nodes of the tree representing an input feature and the leaf nodes representing a classification or regression output. The overall output of the random forest is decided by majority vote between the trees. Random forests are applied in **Paper II** as QSAR models trained on the ECFP6 featurization of the dopamine receptor D2 and also for reaction prediction in **Paper I**. They can be substituted with deep neural networks, but are generally faster to use in both training and prediction and if the test

accuracy is satisfactory preferable to use in time-constrained settings.

Recurrent Neural Networks Recurrent neural networks (RNNs) were designed to work with sequential data and have commonly been used in natural language processing (NLP) tasks. The RNN uses a hidden state h_t for each time step t that the network learns to output which is given as input together with the sequence input, (x_{t+1}, h_t) to the network in the following time step. Thus, the RNN has some information of previous inputs in the sequence which is used in the model predictions. Early RNNs tended to not perform well as the sequences became too large however, and the model was succeeded by the Gated Recurrent Unit [45] (GRU) and Long Short-Term Memory [46] (LSTM) architectures. Several *de novo* design models use GRU and LSTM for the generative task of suggesting new molecules. They perform the task by training on SMILES structures and treating the rules of chemistry the same as a language model treats grammatical rules.

Recently, the Transformer model [47] has been shown to train on the entire sequence instead of processing each item individually, which has made huge improvement for many NLP applications. It has been shown however [16], that the LSTM models that exist already have the capability to generalize well on the small molecule chemical space, as tested on the GDB-13 data set [48].

2.4.3 Active Learning

Active learning is a machine learning approach where the algorithm interacts with a human expert or a labeling oracle to actively select which samples to annotate for training [13]. In traditional supervised learning, a large labeled data set is needed to train a model. However, in active learning, the algorithm starts with a small set of labeled samples and iteratively selects the most informative samples for the expert to label.

The goal of active learning is to maximize the learning efficiency of a model by focusing on the most informative samples, while minimizing the labeling effort and cost. This is particularly useful in situations where labeling data is expensive, time-consuming, or difficult to obtain.

This is analogous to the setting in discovery chemistry, as the set of known molecules (experimentally) in a subset domain of the chemical space is negligibly small compared to the total number of possible molecules that are enumerable.

In **Paper I**, we use active learning with an acquisition strategy for adding new data points known as *Margin* [13]. Margin queries the data point x^* based on the smallest difference in classification probability between the labels

$$x^* = \text{argmin}_x[P_\theta(\hat{y}_1|\ x) - P_\theta(\hat{y}_2|\ x)], \tag{2.2}$$

where $P_\theta(\hat{y}_i|x)$ is the probability that the model assigns to data point x for having label y_i. If the labels are binary and modelled as Bernoulli random variables, then acquisition functions based on maximal variance will select the same points as Margin.

2.5 Research Questions

In this section, we list research questions that were yet to be answered by existing literature that guided the research in this thesis.

Research Question 1: Does the amount of initial data affect the performance gain of active learning for reaction yield prediction models?
In some settings, the amount of initial data that a model has before any points are added through active learning can affect the effectiveness of the active learning model [49], [50]. There has not been a study for active learning in reaction yield prediction that has investigated whether low initial data will cause active learning to provide a different performance gain compared to a case when more data is already available.

Research Question 2: How should *de novo* designed building blocks be treated in library design?
As synthesis on demand of building blocks is limited to a success rate of 76% [12], it would be naive to presume that all suggestions of a *de novo* design generative model can be synthesized in practice. There is a need for a protocol for how the different building blocks should be intepreted.

Research Question 3: How should an actor without access to a large collection of building blocks utilize *de novo* designed building blocks when designing libraries?
In practice an actor, such as a pharmaceutical company, will be operating under a limited budget without access to all synthetically feasible building blocks. Currently, there is little literature on how *de novo* design models can assist in decision making between using available building blocks and purchasing or synthesizing the building blocks suggested by the *de novo* designed libraries.

Chapter 3

Summary of Included Papers

3.1 Paper I: Using active learning to develop machine learning models for reaction yield prediction

In Paper I we conducted a retrospective analysis on two publicly available data sets of chemical reaction experiments to test the robustness of active learning when the initial conditions changed [51], [52]. We also examine the effect of active learning on different machine learning architectures.

Problem

AI-driven synthesis prediction models share the problem of limited access to good reaction data with standardized format. One suggested way to generate new data is to conduct High-Throughput experiments (HTE) to produce thousands of experiments per day [53], [54]. For a constrained combinatorial space in studying reaction conditions, it is possible that redundant experiments are conducted and that the space could have been modeled with a fraction of the experimental data. Active learning is a frequently applied method for sequentially improving a machine learning model by letting the model estimate which data points are expected to yield the most information. Previous studies have shown that active learning is effective for predicting yield [55], but there has been an underlying assumption that active learning performs poorly if the initial amount of known data is low. Furthermore, these studies showed improvements using a single model architecture but demonstrated no indication that the benefit of active learning could be generalized to more architectures.

Contribution

Our study show that

- models trained using active learning to acquire more data improve at least as much as random acquisition even at starting data set configurations of 10 data points in a space of size 4608, and propose that
- the reaction yield prediction task can be simplified to a classification problem for discovery chemistry, and finally that
- the benefit of active learning is consistent across multiple machine learning models for the binary classification task of predicting reaction success.

Methodology

Two neural networks of different levels of complexity were tested against a random forest model and a matrix factorization model [56] to predict the outcome of the chemical reactions on two publicly available combinatorial data sets as a classification task. We chose to view the problem as a classification task where the reaction was labelled as being successful if the reaction yield was greater than 0.20.

We studied the models starting with 10, 100 and 1000 known reactions each with 5 randomly training subsets of the total data. For each model and starting data set, 5 runs of active learning were run using the uncertainty based strategy *margin* as acquisition function. Margin selects data points where the model probability for both labels are as close to 0.5 as possible, indicating that the model is indecisive. Figure 3.1 illustrates how the model during initialization have data points which have probabilities of exactly 0.5, and that the models become more decisive as more data is added.

The active learning strategy was benchmarked against random sampling by examining how many data points where needed to achieve different thresholds of target model performance on a test set of 20% of the total data. The thresholds of AUROC studied were 0.8, 0.85, 0.9, 0.95 and 0.975. We let the acquisition functions add one data point per iteration step and ran all experiments from initial data set until all possible data of the 80% training set had been added.

Contributions

Simon Viet Johansson and Hampus Gummesson Svenson equally performed the main work, and the project was jointly supervised by Esben Bjerrum, Alexander Schliep, Morteza Hagir Chehreghani, Christan Tyrchan and Ola Engkvist.

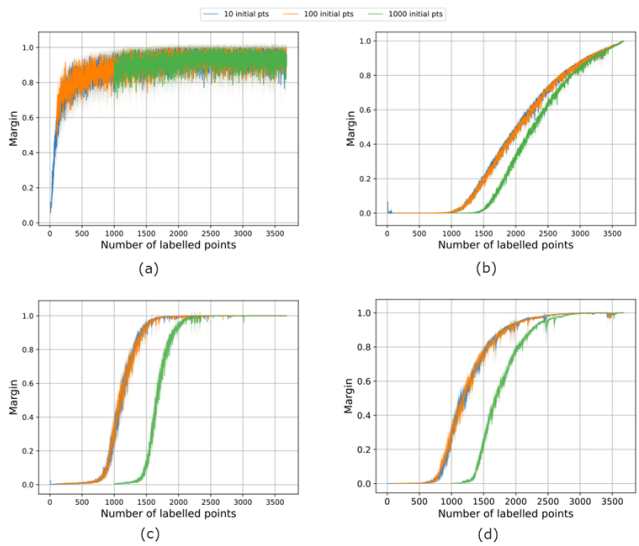

Figure 3.1: The margin between the two labels become increases as points are added. This is shown for the models (a) Matrix factorization, (b) Random forest, (c) Complex NN and (d) Simple NN. Figure extracted, with permission, from original work by [43]

3.2 Paper II: *de novo* generated combinatorial library design

Problem

Combinatorial chemical library design is a method for procuring large amount of chemical data in a materially efficient matter. AI and generative models offer an alternative method to the popular virtual screening for finding hits in drug discovery by generating molecules in a targeted space and can be tuned towards the same goals. However, current generative models only procure *building blocks* rather than a full combinatorial design [28], [57] [58], while methods for combinatorial design offer no verification step that the generated building blocks can actually be synthesized.

Contribution

In this paper we introduce a framework for combinatorial library design, which

- evaluates building blocks generated by de novo design and respect if the building blocks are accessible to the chemist, whereas previous studies used existing virtual libraries or building block catalogues – assuming that the building blocks in the these databases were always available, and which
- can be used by actors with a limited stock of building blocks to estimate the marginal gain in quality of the chemical library they can expect if they were to expand their stock.

For the studied library design, we demonstrated that approximately optimal proposals with the scoring function were possible to make with commercially available building blocks, without further chemical synthesis.

Methodology

Our framework combines the use of the generative model LibINVENT [28] with the retrosynthesis model AiZynthFinder [59] to first generate building blocks and subsequentally evaluate how available they are given the stock of all purchaseable building blocks [60]. LibINVENT was set to train for 1,000 epochs with a batch size of 128 to generate compounds active towards the Dopamine receptor D2 (DRD2) under the constraint that one building block needed to use the Buchwald-Hartwig reaction [61] and the other building block a Amide coupling [62]. After training, 104,991 valid molecules were produced and, after filtering out molecules with an estimated QSAR value of ≤ 0.8 and discarding molecules not following the reaction constraints, 45,928 products remained, from which 32,159 unique carboxylic acids and 2,084 unique aromatic halides were identified.

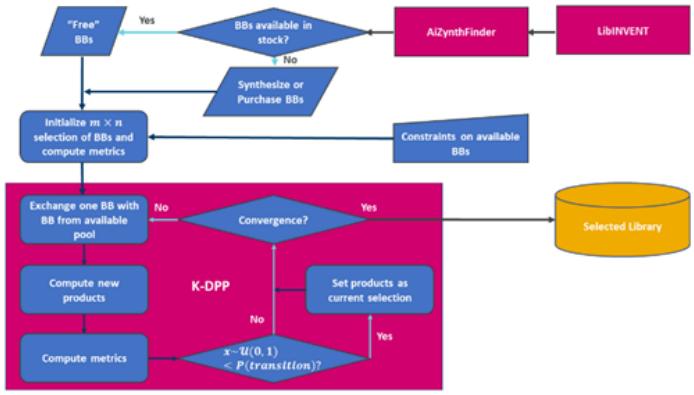

Figure 3.2: Flowchart of the framework used in the paper.

These were then run through AiZynthfinder and the distribution of estimated availability was computed. Building blocks available in 4 reactions or less were kept as candidates for the library, which totalled around 88.7% of the carboxylic acids and 98.3% of the aromatic halides.

The library optimization was performed using k-DPPs, which for this case could be sampled using a Gibbs sampling scheme strictly performing exchange operations. The scoring function used for the decision making process was the average QSAR value from the DRD2 model, the average QED score of the selection and the determinant of the pairwise Tanimoto similarities between the ECFP6 fingerprints. A simulated case for an actor without access to all commercial molecules was conducted by limiting the assumed stock of baseline-available building blocks to a 3% subset of the full dataset, and running AiZynthfinder on this subset to provide a new distribution of availability. A flowchart of the full framework is shown in Figure 3.2.

Contributions

Simon Viet Johansson performed the main work, and the project was jointly supervised by Morteza Hagir Chehreghani, Ola Engkvist and Alexander Schliep.

Chapter 4

Concluding remarks and future direction

In this thesis we have examined two different cases of data acquisition used in drug design.

In **Paper I** we covered the use of active learning to accelerate the improvement of machine learning models, and concluded for the studied combinatorial data sets, that active learning always performed at least as good, if not better, than random selection regardless of the initial amount of known data and the model selection.

In **Paper II**, we studied the design of combinatorial chemical libraries as a process from generative modeling to multi-objective optimization of subset selection. We also simulated the case of an actor with limited access to building blocks and demonstrated the how the framework could be used to estimate the marginal gain in library score from extending the pool of available resources to include building blocks through reaction synthesis.

4.0.1 Future Direction

Since the results of **Paper II** are for a fixed size k, the intuitive extension of the research is to investigate the case of generating multiple libraries and problem of optimizing simultaneously for the 'internal' diversity of the new selection, and the 'global' diversity of all generated libraries. In particular, as library design can be used both for model building and for optimization, it is of relevance to integrate the knowledge learnt from **Paper I** and apply active learning to the library design. Research questions that still need to be answered include the optimal acquisition strategy for building e.g. QSAR models on the library data: given only the data generated from the library design, whether the model performance improve faster from data generated through an 'explore' strategy focused on diversity. Furthermore, if the QSAR model trained is used in the *de novo* generation, the quality of building blocks might improve more if an 'exploit' strategy is used in the library optimization,

but it could also lead to model collapse. It is also possible that QSAR models trained on diverse libraries create a more robust generative model.

Bibliography

[1] A. Thakkar, T. Kogej, J.-L. Reymond, O. Engkvist and E. J. Bjerrum, "Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain," *Chemical Science*, vol. 11, no. 1, pp. 154–168, 2020. DOI: 10.1039/C9SC04944D. [Online]. Available: http://dx.doi.org/10.1039/C9SC04944D (cit. on p. 3).

[2] T. Kodadek, "The rise, fall and reinvention of combinatorial chemistry," *Chemical Communications*, vol. 47, no. 35, pp. 9757–9763, 2011, ISSN: 1359-7345. DOI: 10.1039/C1CC12102B. [Online]. Available: http://dx.doi.org/10.1039/C1CC12102B (cit. on pp. 3, 7).

[3] D. C. Spellmeyer and P. D. J. Grootenhuis, "Chapter 28. recent developments in molecular diversity: Computational approaches to combinatorial chemistry," in *Annual Reports in Medicinal Chemistry*, A. M. Doherty, Ed. Academic Press, 1999, vol. 34, pp. 287–296, ISBN: 0065-7743. DOI: https://doi.org/10.1016/S0065-7743(08)60590-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0065774308605904 (cit. on pp. 3, 7).

[4] F. L. Stahura, L. Xue, J. W. Godden and J. Bajorath, "Molecular scaffold-based design and comparison of combinatorial libraries focused on the atp-binding site of protein kinases11color plate 1, color plate 2 for this article are on pages 51–52," *Journal of Molecular Graphics and Modelling*, vol. 17, no. 1, pp. 1–52, 1999, ISSN: 1093-3263. DOI: https://doi.org/10.1016/S1093-3263(99)00015-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1093326399000157 (cit. on pp. 3, 7).

[5] E. A. Jamois, C. T. Lin and M. Waldman, "Design of focused and restrained subsets from extremely large virtual libraries," *Journal of Molecular Graphics and Modelling*, vol. 22, no. 2, pp. 141–149, 2003, ISSN: 1093-3263. DOI: https://doi.org/10.1016/S1093-3263(03)00154-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1093326303001542 (cit. on pp. 3, 7).

[6] R. Pascual, J. Borrell Ji Fau Teixidó and J. Teixidó, "Analysis of selection methodologies for combinatorial library design," no. 1381-1991 (Print), (cit. on pp. 3, 7).

[7] *Concepts and applications of molecular similarity*. Nashville, TN: John Wiley & Sons, 1990, p. 394, ISBN: 9780471621751 (cit. on pp. 3, 7).

[8] G. Schneider and U. Fechner, "Computer-based de novo design of drug-like molecules," *Nature Reviews Drug Discovery*, vol. 4, no. 8, pp. 649–663, 2005, ISSN: 1474-1784. DOI: 10.1038/nrd1799. [Online]. Available: https://doi.org/10.1038/nrd1799 (cit. on p. 3).

[9] W. P. Walters, "Virtual chemical libraries," *Journal of Medicinal Chemistry*, vol. 62, no. 3, pp. 1116–1124, 2019, ISSN: 0022-2623. DOI: 10.1021/acs.jmedchem.8b01048. [Online]. Available: https://doi.org/10.1021/acs.jmedchem.8b01048 (cit. on p. 3).

[10] N. van Hilten, F. Chevillard and P. Kolb, "Virtual compound libraries in computer-assisted drug discovery," *Journal of Chemical Information and Modeling*, vol. 59, no. 2, pp. 644–651, 2019, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.8b00737. [Online]. Available: https://doi.org/10.1021/acs.jcim.8b00737 (cit. on p. 3).

[11] C. G. Wermuth, C. R. Ganellin, P. Lindberg and L. A. Mitscher, "Glossary of terms used in medicinal chemistry (iupac recommendations 1998)," *Pure and Applied Chemistry*, vol. 70, no. 5, pp. 1129–1143, 1998. DOI: doi:10.1351/pac199870051129. [Online]. Available: https://doi.org/10.1351/pac199870051129 (cit. on pp. 3, 8).

[12] *Made building blocks*, Web Page, 2023. [Online]. Available: https://enamine.net/building-blocks/made-building-blocks (cit. on pp. 4, 12).

[13] L. G. Valiant, "A theory of the learnable," *Commun. ACM*, vol. 27, no. 11, 1134–1142–1134–1142, 1984. DOI: 10.1145/1968.1972. [Online]. Available: https://doi.org/10.1145/1968.1972 (cit. on pp. 4, 11).

[14] M. Hassan, S. Brown Rd Fau Varma-O'brien, D. Varma-O'brien S Fau Rogers and D. Rogers, "Cheminformatics analysis and learning in a data pipelining environment," no. 1381-1991 (Print), (cit. on p. 5).

[15] D. Weininger, "Smiles, a chemical language and information system: 1: Introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, 1988, ISSN: 0095-2338. DOI: 10.1021/ci00057a005 (cit. on p. 6).

[16] J. Arús-Pous, S. V. Johansson, O. Prykhodko *et al.*, "Randomized smiles strings improve the quality of molecular generative models," *Journal of Cheminformatics*, vol. 11, no. 71, 2019. DOI: 10.1186/s13321-019-0393-0. [Online]. Available: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-019-0393-0 (cit. on pp. 6, 11).

[17] M. Sioud, "Main approaches to target discovery and validation," in *Target Discovery and Validation Reviews and Protocols: Volume 1, Emerging Strategies for Targets and Biomarker Discovery*, M. Sioud, Ed. Totowa, NJ: Humana Press, 2007, pp. 1–12, ISBN: 978-1-59745-165-9. DOI: 10.1385/1-59745-165-7:1. [Online]. Available: https://doi.org/10.1385/1-59745-165-7:1 (cit. on p. 6).

[18] Y. Hu, T. Zhao, N. Zhang, Y. Zhang and L. Cheng, "A review of recent advances and research on drug target identification methods," *Current Drug Metabolism*, vol. 20, no. 3, pp. 209–216, 2019, ISSN: 1389-2002/1875-5453. DOI: http://dx.doi.org/10.2174/1389200219666180925091851. [Online]. Available: http://www.eurekaselect.com/article/93197 (cit. on p. 6).

[19] C. Smith, "Drug target validation: Hitting the target," *Nature*, vol. 422, no. 6929, pp. 342–345, 2003, ISSN: 1476-4687. DOI: 10.1038/422341a. [Online]. Available: https://doi.org/10.1038/422341a (cit. on p. 6).

[20] K. H. Bleicher, H.-J. Böhm, K. Müller and A. I. Alanine, "Hit and lead generation: Beyond high-throughput screening," *Nature Reviews Drug Discovery*, vol. 2, no. 5, pp. 369–378, 2003, ISSN: 1474-1784. DOI: 10.1038/nrd1086. [Online]. Available: https://doi.org/10.1038/nrd1086 (cit. on p. 7).

[21] S. K. Ashenden, "Chapter 6 - lead optimization," in *The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry*, S. K. Ashenden, Ed., Academic Press, 2021, pp. 103–117, ISBN: 978-0-12-820045-2. DOI: https://doi.org/10.1016/B978-0-12-820045-2.00007-6. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128200452000076 (cit. on p. 7).

[22] M. P. Barcelos, S. Q. Gomes, L. B. Federico *et al.*, "Lead optimization in drug discovery," in *Research Topics in Bioactivity, Environment and Energy: Experimental and Theoretical Tools*, C. A. Taft and S. R. de Lazaro, Eds. Cham: Springer International Publishing, 2022, pp. 481–500, ISBN: 978-3-031-07622-0. DOI: 10.1007/978-3-031-07622-0_19. [Online]. Available: https://doi.org/10.1007/978-3-031-07622-0_19 (cit. on p. 7).

[23] F. F. Hefti, "Requirements for a lead compound to become a clinical candidate," *B. M. C. Neurosci*, no. 1471-2202 (Electronic), DOI: 10.1186/1471-2202-9-S3-S7 (cit. on p. 7).

[24] J. Arús-Pous, A. Patronov, E. J. Bjerrum *et al.*, "Smiles-based deep generative scaffold decorator for de-novo drug design," *Journal of Cheminformatics*, vol. 12, no. 1, p. 38, 2020, ISSN: 1758-2946. DOI: 10.1186/s13321-020-00441-8. [Online]. Available: https://doi.org/10.1186/s13321-020-00441-8 (cit. on p. 7).

[25] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona and T. Blaschke, "The rise of deep learning in drug discovery," *Drug Discovery Today*, vol. 23, no. 6, pp. 1241–1250, 2018. DOI: 10.1016/j.drudis.2018.01.039. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1359644617303598 (cit. on p. 8).

[26] J. J. Irwin and B. K. Shoichet, "Zinc a free database of commercially available compounds for virtual screening," *Journal of Chemical Information and Modeling*, vol. 45, no. 1, pp. 177–182, 2005. DOI: 10.1021/ci049714+. [Online]. Available: https://doi.org/10.1021/ci049714+ (cit. on p. 8).

[27] A. Gaulton, L. J. Bellis, A. P. Bento *et al.*, "Chembl: A large-scale bioactivity database for drug discovery," *Nucleic Acids Research*, vol. 40, no. D1, pp. D1100–D1107, 2012. DOI: 10.1093/nar/gkr777. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr777 (cit. on p. 8).

[28] V. Fialková, J. Zhao, K. Papadopoulos *et al.*, "Libinvent: Reaction-based generative scaffold decoration for in silico library design," *Journal of Chemical Information and Modeling*, vol. 62, no. 9, pp. 2046–2063, 2022, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.1c00469. [Online]. Available: https://doi.org/10.1021/acs.jcim.1c00469 (cit. on pp. 8, 16).

[29] S. Johansson, A. Thakkar, T. Kogej *et al.*, "Ai-assisted synthesis prediction," *Drug Discovery Today: Technologies*, vol. 32-33, pp. 65–72, 2019. DOI: https://doi.org/10.1016/j.ddtec.2020.06.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1740674920300020 (cit. on p. 8).

[30] C. Hansch, P. P. Maloney, T. Fujita and R. M. Muir, "Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients," *Nature*, vol. 194, no. 4824, pp. 178–180, 1962, ISSN: 1476-4687. DOI: 10.1038/194178b0. [Online]. Available: https://doi.org/10.1038/194178b0 (cit. on p. 8).

[31] G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan and A. L. Hopkins, "Quantifying the chemical beauty of drugs," *Nature Chemistry*, vol. 4, pp. 90–90, 2012. [Online]. Available: https://doi.org/10.1038/nchem.1243http://10.0.4.14/nchem.1243https://www.nature.com/articles/nchem.1243#supplementary-information (cit. on p. 9).

[32] A. Koutsoukas, S. Paricharak, W. R. J. D. Galloway *et al.*, "How diverse are diversity assessment methods? a comparative analysis and benchmarking of molecular descriptor space," *Journal of Chemical Information and Modeling*, vol. 54, no. 1, pp. 230–242, 2014, PMID: 24289493. DOI: 10.1021/ci400469u. eprint: https://doi.org/10.1021/ci400469u. [Online]. Available: https://doi.org/10.1021/ci400469u (cit. on p. 9).

[33] D. K. Agrafiotis, "A constant time algorithm for estimating the diversity of large chemical libraries," *Journal of Chemical Information and Computer Sciences*, vol. 41, no. 1, pp. 159–167, 2001, PMID: 11206368. DOI: 10.1021/ci000091j. eprint: https://doi.org/10.1021/ci000091j. [Online]. Available: https://doi.org/10.1021/ci000091j (cit. on p. 9).

[34] T. T. Tanimoto, *An Elementary Mathematical Theory of Classification and Prediction*. International Business Machines Corporation, 1958. [Online]. Available: https://books.google.se/books?id=yp34HAAACAAJ (cit. on p. 9).

[35] H. Chen, U. Börjesson, O. Engkvist *et al.*, "Prosar: A new methodology for combinatorial library design," *Journal of Chemical Information and Modeling*, vol. 49, no. 3, pp. 603–614, 2009, ISSN: 1549-9596. DOI: 10.1021/ci800231d. [Online]. Available: https://doi.org/10.1021/ci800231d (cit. on p. 9).

[36] V. J. Gillet, W. Khatib, P. Willett, P. J. Fleming and D. V. S. Green, "Combinatorial library design using a multiobjective genetic algorithm," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 2, pp. 375–385, 2002, ISSN: 0095-2338. DOI: 10.1021/ci010375j. [Online]. Available: https://doi.org/10.1021/ci010375j (cit. on p. 9).

[37] D. K. Agrafiotis, "Multiobjective optimization of combinatorial libraries," *Molecular Diversity*, vol. 5, no. 4, pp. 209–230, 2000, ISSN: 1573-501X. DOI: 10.1023/A:1021320124615. [Online]. Available: https://doi.org/10.1023/A:1021320124615 (cit. on p. 9).

[38] O. Macchi, "The coincidence approach to stochastic point processes," *Advances in Applied Probability*, vol. 7, no. 1, pp. 83–122, 1975, ISSN: 0001-8678. DOI: 10.2307/1425855. [Online]. Available: https://www.cambridge.org/core/article/coincidence-approach-to-stochastic-point-processes/1EED58D03316134553E83A9E96501FE1 (cit. on p. 10).

[39] A. Kulesza, B. J. F. Taskar and T. i. M. Learning, "Determinantal point processes for machine learning," vol. 5, no. 2–3, pp. 123–286, 2012, ISSN: 1935-8237 (cit. on p. 10).

[40] S. O. Gharan and A. J. a. p. a. Rezaei, "A polynomial time mcmc method for sampling from continuous dpps," 2018 (cit. on p. 10).

[41] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 6, pp. 721–741, 1984, ISSN: 1939-3539. DOI: 10.1109/TPAMI.1984.4767596 (cit. on p. 10).

[42] K. V. Chuang and M. J. Keiser, "Comment on "predicting reaction performance in c-n cross-coupling using machine learning"," *Science*, vol. 362, no. 6416, eaat8603, 2018. DOI: doi:10.1126/science.aat8603. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aat8603 (cit. on p. 10).

[43] S. Viet Johansson, H. Gummesson Svensson, E. Bjerrum *et al.*, "Using active learning to develop machine learning models for reaction yield prediction," *Molecular Informatics*, vol. 41, no. 12, p. 2 200 043, 2022. DOI: https://doi.org/10.1002/minf.202200043. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/minf.202200043. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.202200043 (cit. on pp. 10, 15).

[44] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324. [Online]. Available: https://doi.org/10.1023/A:1010933404324 (cit. on p. 10).

[45] K. Cho, B. van Merrienboer, C. Gulcehre *et al.*, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, pp. 1724–1734. DOI: 10.3115/v1/D14-1179. [Online]. Available: http://aclweb.org/anthology/D14-1179 (cit. on p. 11).

[46] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735. [Online]. Available: http://www.mitpressjournals.org/doi/10.1162/neco.1997.9.8.1735 (cit. on p. 11).

[47] A. Vaswani, N. Shazeer, N. Parmar *et al.*, *Attention is all you need*, 2017. arXiv: 1706.03762 [cs.CL] (cit. on p. 11).

[48] L. C. Blum and J. L. Reymond, "970 million druglike small molecules for virtual screening in the chemical universe database gdb-13," *Journal of the American Chemical Society*, vol. 131, no. 25, pp. 8732–8733, 2009, ISSN: 0002-7863. DOI: 10.1021/ja902302h (cit. on p. 11).

[49] Y. Zhao and Q. Ji, "An active learning method under very limited initial labeled data," in *2010 IEEE International Conference on Automation and Logistics*, 2010, pp. 524–527. DOI: 10.1109/ICAL.2010.5585339 (cit. on p. 12).

[50] J. D. Bossér, E. Sörstadius and M. H. Chehreghani, "Model-centric and data-centric aspects of active learning for deep neural networks," in *2021 IEEE International Conference on Big Data (Big Data)*, pp. 5053–5062. DOI: 10.1109/BigData52589.2021.9671795 (cit. on p. 12).

[51] D. Perera, J. W. Tucker, S. Brahmbhatt *et al.*, "A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow," *Science*, vol. 359, no. 6374, 429 LP–434, 2018. DOI: 10.1126/science.aap9112. [Online]. Available: http://science.sciencemag.org/content/359/6374/429.abstract (cit. on p. 13).

[52] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, "Predicting reaction performance in c–n cross-coupling using machine learning," *Science*, vol. 360, no. 6385, pp. 186–190, 2018. DOI: 10.1126/science.aar5169. [Online]. Available: https://www.sciencemag.org/lookup/doi/10.1126/science.aar5169 (cit. on p. 13).

[53] E. S. Isbrandt, R. J. Sullivan and S. G. Newman, "High throughput strategies for the discovery and optimization of catalytic reactions," *Angewandte Chemie International Edition*, vol. 58, no. 22, pp. 7180–7191, 2019. DOI: 10.1002/anie.201812534. [Online]. Available: https://doi.org/10.1002/anie.201812534 (cit. on p. 13).

[54] B. Mahjour, Y. Shen and T. Cernak, "Ultrahigh-throughput experimentation for information-rich chemical synthesis," *Accounts of Chemical Research*, 2021. DOI: 10.1021/acs.accounts.1c00119. [Online]. Available: https://doi.org/10.1021/acs.accounts.1c00119 (cit. on p. 13).

[55] N. S. Eyke, W. H. Green and K. F. Jensen, "Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening," *Reaction Chemistry & Engineering*, vol. 5, no. 10, pp. 1963–1972, 2020. DOI: 10.1039/D0RE00232A. [Online]. Available: http://xlink.rsc.org/?DOI=D0RE00232A (cit. on p. 13).

[56] B. R. Beno and J. S. Mason, "The design of combinatorial libraries using properties and 3d pharmacophore fingerprints," *Drug Discovery Today*, vol. 6, no. 5, pp. 251–258, 2001, ISSN: 1359-6446. DOI: https://doi.org/10.1016/S1359-6446(00)01665-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1359644600016652 (cit. on p. 14).

[57] A. Domenico, G. Nicola, T. Daniela, C. Fulvio, A. Nicola and N. Orazio, "De novo drug design of targeted chemical libraries based on artificial intelligence and pair-based multiobjective optimization," *Journal of Chemical Information and Modeling*, vol. 60, no. 10, pp. 4582–4593, 2020, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.0c00517. [Online]. Available: https://doi.org/10.1021/acs.jcim.0c00517 (cit. on p. 16).

[58] V. Bagal, R. Aggarwal, P. K. Vinod and U. D. Priyakumar, "Molgpt: Molecular generation using a transformer-decoder model," *Journal of Chemical Information and Modeling*, vol. 62, no. 9, pp. 2064–2076, 2022, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.1c00600. [Online]. Available: https://doi.org/10.1021/acs.jcim.1c00600 (cit. on p. 16).

[59] S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist and E. Bjerrum, "Aizynthfinder: A fast, robust and flexible open-source software for retrosynthetic planning," *Journal of Cheminformatics*, vol. 12, no. 1, p. 70, 2020, ISSN: 1758-2946. DOI: 10.1186/s13321-020-00472-1. [Online]. Available: https://doi.org/10.1186/s13321-020-00472-1 (cit. on p. 16).

[60] Web Page. [Online]. Available: https://downloads.emolecules.com/free/ (cit. on p. 16).

[61] M. B. Smith and J. March, "March's advanced organic chemistry : Reactions, mechanisms, and structure," in 7th. Somerset: John Wiley & Sons, 2013, pp. 751–755, ISBN: 9781118472217 (cit. on p. 16).

[62] B. Mahjour, Y. Shen, W. Liu and T. Cernak, "A map of the amine–carboxylic acid coupling system," *Nature*, vol. 580, no. 7801, pp. 71–75, 2020, ISSN: 1476-4687. DOI: 10.1038/s41586-020-2142-y. [Online]. Available: https://doi.org/10.1038/s41586-020-2142-y (cit. on p. 16).