HAW HAMBURG

**BACHELORTHESIS**
Deike Maria Flemming

# How can AI support wet lab work?

A literature review of current artificial intelligence techniques for the early stages of the drug discovery pipeline

**FAKULTÄT TECHNIK UND INFORMATIK**
Department Informatik

Faculty of Computer Science and Engineering
Department Computer Science

**HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG**
**Hamburg University of Applied Sciences**

Deike Maria Flemming

# How can AI support wet lab work?

## A literature review of current artificial intelligence techniques for the early stages of the drug discovery pipeline

**Deike Maria Flemming**

## Thema der Arbeit

How can AI support wet lab work? A literature review of current artificial intelligence techniques for the early stages of the drug discovery pipeline

## Stichworte

Artificial Intelligence, Drug Discovery Pipeline

## Kurzzusammenfassung

Künstliche Intelligenz (KI), insbesondere ihre Unterbereiche des maschinellen Lernens und des Deep Learning, ist ein vielversprechendes Werkzeug, das Bioinformatikern und pharmazeutischen Forschern zur Verfügung steht. Sie kann den Entwicklungsprozess von neuen Medikamenten auf vielfältige Weise unterstützen und beschleunigen. Diese Arbeit befasst sich mit Techniken der KI, die während des Forschungsprozesses angewendet werden, um neue Medikamente für bekannte und neu entdeckte Krankheiten zu entdecken.

**Deike Maria Flemming**

## Title of Thesis

How can AI support wet lab work? A literature review of current artificial intelligence techniques for the early stages of the drug discovery pipeline

## Keywords

Artificial Intelligence, Drug Discovery Pipeline

## Abstract

Artificial intelligence, particularly its subsets of machine learning and deep learning, is a promising tool available to bioinformaticians and pharmaceutical researchers. It can support the drug discovery process in numerous ways that help to make the drug discovery process faster and more targeted. This thesis takes a detailed look at the various artificial intelligence techniques that are being proposed and applied during the research process to discover new medications for both well-known and newly discovered diseases and medical conditions.

# Table of Contents

# Declaration of Independent Work

I hereby declare that I wrote the Bachelor's Thesis titled

*"How can AI support wet lab work?*
*A literature review of current artificial intelligence techniques for the early stages of the drug discovery pipeline"*

Independently and only using the declared sources. I marked all content that I used from literature or other sources, such as websites, clearly as quotes and cited the sources.

# Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelor-Thesis mit dem Titel

*"How can AI support wet lab work?*
*A literature review of current artificial intelligence techniques for the early stages of the drug discovery pipeline"*

selbständig und nur mit den angegebenen Hilfsmitteln verfasst habe. Alle Passagen, die ich wörtlich aus der Literatur oder aus anderen Quellen wie z. B. Internetseiten übernommen habe, habe ich deutlich als Zitat mit Angabe der Quelle kenntlich gemacht.

# List of Graphics

# Acknowledgements

I would like to thank Kai von Luck for the support and spirited discussions leading up to the production of this thesis. I would also like to thank Jan Schwarzer and Christian Lins for their advice and support.

# Glossary

This thesis is aimed at readers familiar with the broader terminology of computer science and specifically artificial intelligence as well as its subsets of machine learning and deep learning. These terms are therefore only broadly outlined as they are used in the context of this work. Background on the subject and the more specific terminology in this discipline can be found in literature such as *Machine Learning* by Tom Mitchell, published by MIT Press & The McGraw-Hills Company (1997). This glossary also covers the basic terms required for an appropriate understanding of the biochemical and pharmaceutical aspects of the thesis subject.

## Computer science terminology

### Artificial intelligence (AI)

This thesis will be a review of research papers that look at a number of different artificial intelligence approaches to answering research questions in the realm of biochemistry and pharmacology. George F. Luger puts forth an apt definition of the term in his book *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*: “Artificial intelligence (AI) may be defined as the branch of computer science that is concerned with the automation of intelligent behavior” (p. 1).

However, Luger recognises that this definition raises some almost philosophical questions about the definition and implications of intelligent behaviour. He therefore acknowledges that his “definition of artificial intelligence falls short of unambiguously defining the field” (p.2) and concludes that “[...] this difficulty in arriving at a precise definition of AI is entirely appropriate. Artificial intelligence is still a young discipline, and its structure, concerns, and methods are less clearly defined than those of a more mature science such as physics” (p. 2). Taking into account this ambiguity, he proposes a second definition of the term artificial intelligence: “[T]he collection of problems and methodologies studied by artificial intelligence researchers” (p.2).

### Machine learning (ML)

As a subset of artificial intelligence, machine learning can be defined as “statistical learning and optimization methods that let computers analyze datasets and identify patterns” (M. Tamir, 2020). The important part of this definition that distinguishes

machine learning from other subsets of artificial intelligence is the optimisation aspect that is included in algorithms that are considered to be machine learning algorithms. In his 2020 article on machine learning, Tamir identifies three components that are integral to machine learning:

“

1. ***A decision process****: A recipe of calculations or other steps that takes in the data and returns a “guess” at the kind of pattern in the data your algorithm is looking to find.*
2. ***An error function****: A method of measuring how good the guess was by comparing it to known examples (when they are available). Did the decision process get it right? If not, how do you quantify “how bad” the miss was?*
3. ***An updating or optimization process****: Where the algorithm looks at the miss and then updates how the decision process comes to the final decision so that the next time the miss won’t be as great.*

“

The research papers reviewed for this thesis utilise well-known machine learning algorithms such as:

- Collaborative Metric Learning for Top-K Recommendations
- Support Vector Machine

### Deep learning (DL)

Deep learning is itself a further subset of machine learning. In their 2015 review on deep learning, LeCun et al. distinguish it from the broader field of machine learning by focussing on the data format that can be processed by deep learning algorithms. They state that before deep learning techniques were developed, machine learning algorithms were limited to processing data which had to be extensively preprocessed by humans to extract features that could be analysed by the algorithm to detect patterns in the data (compare p. 436). By contrast, deep learning algorithms are “fed with raw data and [...] automatically discover the representations needed for detection or classification” (p. 436).

In his 2019 book on deep learning Kelleher offers a slightly more narrow definition that equates deep learning techniques with neural network algorithms: “Deep learning is the subfield of artificial intelligence that focus on creating large neural network models that are capable of making accurate data-driven decisions” (p. 1). He concurs that “[d]eep learning is particularly suited to contexts where the data is complex and where there are large datasets available” (p. 1), hinting at the ability of deep learning algorithms to process this data without the need for feature extraction by humans. It is important to mention that this view is slightly more controversial as not all computer scientists agree that a neural network algorithm is automatically a deep learning algorithm. However, for the purposes of this review the terms large neural networks and deep learning can be used interchangeably, as all featured deep learning techniques are indeed neural networks.

Examples for these deep learning techniques that can be found in the reviewed research papers are:

- Convolutional Neural Networks
- Fully-Connected Convolutional Neural Networks
- Graph Neural Networks
- Residual Neural Networks
- Variational Autoencoders
- Graph Convolutional Policy Networks

## Biochemistry terminology

### Molecule

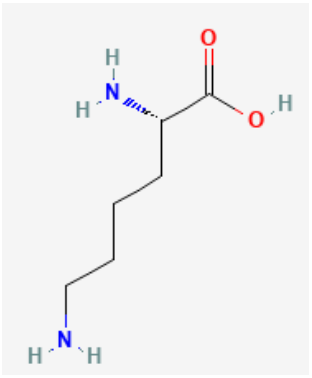
Defined by McNaught and Wilkinson as “[a]n electrically neutral entity consisting of more than one atom”. A molecule can be categorised depending on various factors such as its size, the atoms or atomic groups it is made up of and its chemical reactivity. Examples of such categories that are relevant for this thesis are proteins, peptides and amino acids.

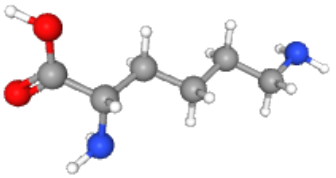
### Representation

Molecules can be represented in numerous formats that intend to describe the positioning of and bonds between the individual atoms that make up the molecule.

Below are three formats of molecular representation commonly used in chemistry and pharmaceutical research.

C(CCN)CC(C(=O)O)N



*Img. 1: Different molecule notation formats for the essential amino acid L-lysine (C6H14N2O2). From left to right: Canonical SMILES, 2D structured graph, 3D structured ball and stick graph. Source for all: PubChem*

### Molecular properties

Molecular properties are physiochemical characteristics that describe how the molecule is structured and how it interacts with its surroundings. Examples for molecular properties can be fairly simple measurements such as “size, shape, lipophilicity, hydrogen bonding capability, and polarity” (Leeson & Young, p. 722) as well as more complex behaviours like binding affinity for certain receptors on the cell surface or the permeability of organic barriers such as cell membranes and the blood-brain-barrier.

### Molecule generation

In the context of this work the term molecular generation refers to an algorithmic approach to generate representations of molecules. It is unrelated to wet lab work or production processes where actual chemical compounds are physically assembled. The process of molecular generation is intended to discover novel molecules that have not previously been examined for their potential pharmaceutical profile.

### Molecule optimisation

The idea of molecule optimisation is closely related to molecule generation. It is also an algorithmic approach and only generates representations of molecules not actual physical compounds. However, instead of creating new molecules from scratch it takes

a known molecule and makes modifications to it. This is usually done in order to add or improve desired molecular properties of a known compound. The resulting optimised molecule will share significant similarities with the original seed compound.

### Amino acid

“Chemically, an amino acid is a molecule that has a carboxylic acid group and an amine group that are each attached to a carbon atom called the α carbon. Each of the 20 amino acids has a specific side chain, known as an R group, that is also attached to the α carbon. The R groups have a variety of shapes, sizes, charges, and reactivities. [...] The sequence and interactions between the side chains of these different amino acids allow each protein to fold into a specific three-dimensional shape and perform biological functions.” (Nature Education, Amino Acid, 2014)

### Peptide

“A peptide is a short chain of amino acids. The amino acids in a peptide are connected to one another in a sequence by bonds called peptide bonds. Typically, peptides are distinguished from proteins by their shorter length, although the cut-off number of amino acids for defining a peptide and protein can be arbitrary. Peptides are generally considered to be short chains of two or more amino acids. Meanwhile, proteins are long molecules made up of multiple peptide subunits, and are also known as polypeptides. Proteins can be digested by enzymes (other proteins) into short peptide fragments.” (Nature Education, Peptide, 2014)

### Protein

McNaught and Wilkinson define proteins as “[n]aturally occurring and synthetic polypeptides having molecular weights greater than about 10000 (the limit is not precise).” A protein is therefore a large molecule made up of multiple amino acids. The functional part of DNA, the genetic code of animals and plants, encodes proteins.

The human genome contains roughly 20,500 protein-encoding genes (Clamp et al., p. 1, 2007). “Proteins serve as structural support inside the cell and they perform many vital chemical reactions. Each protein is a molecule made up of different combinations of 20 types of smaller, simpler amino acids. Protein molecules are long chains of amino acids that are folded into a three-dimensional shape.” (Nature Education, Amino Acid, 2014)

## Ligand

The definition of ligand in biochemistry is broader than the definition applied in the field of inorganic chemistry where it describes “the atoms or groups” that joined only to the one “central atom” (McNaught & Wilkinson, 2019). In biochemistry this central core can be a “polyatomic molecular entity”. However, ligand still describes the atoms or groups that attach to this polyatomic entity.

McNough and Wilkonson add the example that “$H^+$ may be a ligand for proteins and for citrate as well as for $O^{2-}$ . [...] [I]n other circumstances, $AcO^-$ may be the ligand for $H^+$, since the definition makes it clear that the view of which entity is central may change for convenience.” Thus, a ligand is therefore not defined by its atomic makeup but instead by its relative positioning in a molecule. It means that the same atomic group can be a ligand in one molecule and the central polyatomic entity to which ligands attach in another molecule.

# 1. Introduction

## 1.1 Overview

This introductory chapter will discuss the motivation for this work and give some background on the stages of drug development as well as some of the challenges that scientists are faced with during that process. It will also preview some artificial intelligence (AI) techniques that are used to address those challenges as well as the limitations present in those AI based approaches themselves.

The methodology for this literature review is outlined in detail in chapter two. It looks at the goal and constraints for this textual narrative synthesis. Additionally it will describe the selection process for the scientific papers that were included in the analysis.

Chapter three contains the majority of the textual analysis. It looks at the literature included in this review and presents a high level overview of the data extracted from their experiments. This chapter will highlight the results that have been achieved by applying computer based algorithms to different phases of the drug development process. The research papers will be categorised to underline similarities in the approaches and point out where there are important differences. It is also intended to contextualise and connect the research to other works in this realm.

In chapter four there will be a discussion of the limitations that shape this thesis. This will include an overview of the known biases influencing the literature selection and review process. The discussion also offers an outlook on additional research that could be undertaken given more time as well as alternative ways to answer the research question posed in the title.

Finally, the fifth chapter will present the three main conclusions that emerge from the analysis of the literature selection.

## 1.2 Motivation

Artificial intelligence is a very promising tool at the disposal of bioinformaticians and pharmaceutical researchers. It can support the drug discovery process in numerous

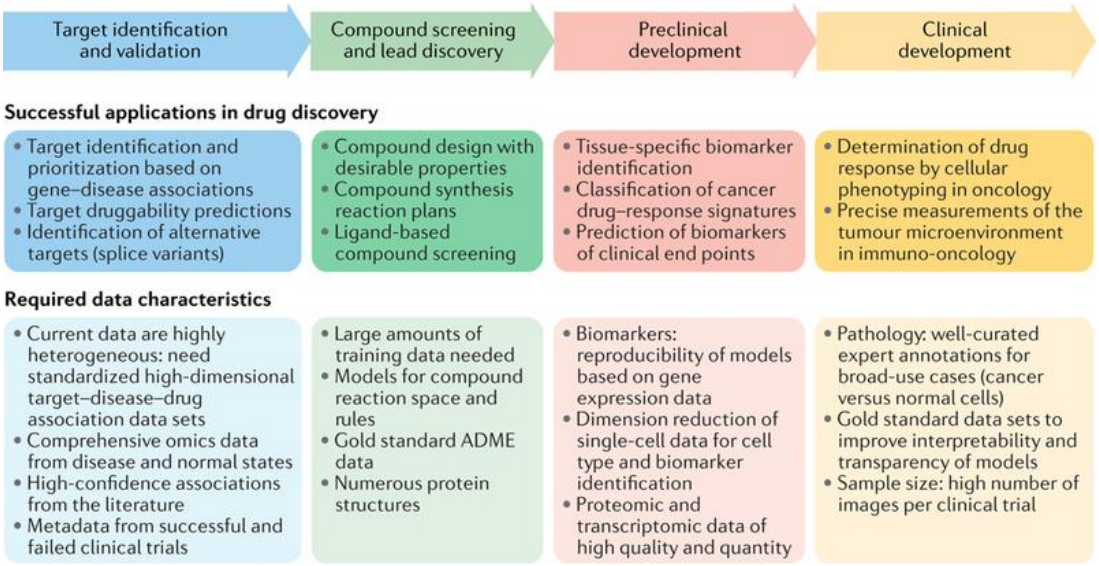
ways that help to make the drug discovery process faster and more targeted. This thesis will take a detailed look at the various artificial intelligence techniques that are being proposed and applied during the research process to discover new medications for both well-known and newly discovered diseases and medical conditions.

Looking at the way tools developed by theoretical computer scientists, mathematicians and statisticians are utilised to answer research questions in another discipline such as pharmacology offers valuable interdisciplinary insight into the strengths of these tools. It demonstrates their transferability and the value they can add to other fields of research when applied in order to solve real life problems experienced in our society.

Today, despite major advances in medicine and technology, there are still large numbers of diseases for which there is no prevention, treatment or cure. For other diseases there are medical interventions available which are unreliable or have undesirable side effects. This means that there is the potential and market for the development of new drugs to target these diseases.

The research process to identify and evaluate these medications is called the drug discovery pipeline. The drug discovery pipeline describes four stages of discovery that lead to the approval of a new medication. As described by Vamathevan et al. these four stages are:

1. Target identification and validation
2. Compound screening and lead discovery
3. Preclinical development
4. Clinical development




*Img. 2: AI applications at all stages of the drug discovery pipeline. Source: Vamathevan et al.*

The four stages described above are usually followed by an approval process which is required for drugs to be marketed legally. This approval process varies from jurisdiction to jurisdiction. A new medication is not guaranteed approval just because it successfully passed all four stages of the drug discovery process. It may also be granted approval in some jurisdictions but not others and different restrictions regarding its application may be applied in different jurisdictions. This approval process is not commonly viewed as part of the drug discovery pipeline.

Although naming the discovery process a pipeline could conjure up the image of a long pipe with a consistent diameter that transports potential new drugs from one end to the other, in reality it functions rather like a funnel or a sieve where a significant number of potential new drugs are eliminated from the process at each stage.

Because the number of potential targets and compounds is so large compared to the number of drugs that actually result in successful clinical trials, there is a significant economic incentive for the pharmaceutical companies to reduce the number of targets and compounds that need to be screened. Aside from the monetary advantage that the pharma industry can gain by streamlining the drug discovery pipeline, there is also the aspect of the timeframe required to complete this process.

AI techniques can help reduce the time it takes to develop a new drug by several years. Some drugs that have already been developed using AI tools and are currently on the market have been developed in as little as a quarter of the time that comparable drug development previously took without the help of AI methods. Any tool that can help reduce the time needed for the development of new drugs can help save lives.

# 2. Methodology

This chapter will lay out the methodology applied to the research process for this literature review. There is a section detailing the criteria used to include papers in the literature selection that will consequently be analysed in more detail in chapter three. It also contains the exact search terms and a description of the process used to prune the search results to the final literature selection.

## 2.1 Inclusion criteria

In order to limit the scope of the research to manageable levels, the search was restricted to recent, completed and peer reviewed works that are available via the ACM Digital Library. Criteria for inclusion in the review process were:

1. Work must be a complete research paper.
2. The research paper must be published in a peer reviewed publication.
3. The paper must have been published between November 2020 and October 2021.
4. The paper must include specific search terms outlined below.

## 2.2 Search terms

The search terms were formulated after a preliminary unstructured review of various papers related to bioinformatics and artificial intelligence. To ensure the paper focuses on both artificial intelligence and drug discovery, the search query consisted of two groups of inputs. The first group aims to include papers that deal with artificial intelligence, whereas the second group of phrases ensures the inclusion of papers that deal with the drug discovery pipeline. Phrases within each group were combined using a boolean OR operator whereas the two groups were combined using a boolean AND operation.

The following groups of terms were used for the search:

### 2.2.1 Group 1

- Machine Learning
- Deep Learning

- Artificial Intelligence

### 2.2.2 Group 2

- Drug Discovery
- Target Identification
- Molecule Generation
- Drug Development
- Target Interaction
- Drug Design
- Drug Target

Furthermore, the second group of terms was required to appear within the papers abstract, whereas the first group was allowed to appear anywhere within the paper's title, abstract or text. This decision was made after an attempt to also restrict the first group to the abstract lead to drastically reduced search results. This is due to the fact that within their abstracts many papers actually specify the exact name of the artificial intelligence technique that the authors applied, as opposed to the more generalised terms from the search group which tend to show up within the introduction, list of key words or conclusion of the paper instead.

## 2.3 Final query

[[All: "machine learning"] OR [All: "deep learning"] OR [All: "artificial intelligence"]] AND [[Abstract: "drug discovery"] OR [Abstract: "target identification"] OR [Abstract: "molecule generation"] OR [Abstract: "drug development"] OR [Abstract: "target interaction"] OR [Abstract: "drug design"] OR [Abstract: "drug target"]] AND [Publication Date: (01/11/2020 TO 31/10/2021)]

## 2.4 Preliminary Search Results

Overall this search yielded 21 research papers. These papers were then scanned for titles that match the subject of this bachelor's thesis. Of the 21 search results, 6 were excluded based on their title. Papers were excluded based on their title when it was obvious that the focus of the paper was not on both bioinformatics and artificial

intelligence. These papers appeared in the search results because they were focussing on specific AI or bioinformatics topics and then listed possible applications of their results matching one or more of the search terms in their abstracts. In a secondary step, the remaining 15 papers' abstracts were reviewed. This confirmed that all 15 papers appeared relevant to the research question of this thesis.

During a thorough analysis of the selected research papers, one last paper was removed from the final literature selection. "An Analytical Review of Computational Drug Repurposing" by Sadeghi and Keyvanpour provides valuable background information for understanding the subject of drug repurposing which represents a special subtype of drug development. It also represents a helpful guideline for a systematic analysis of research on computational models.

However, since the paper was itself a literature review rather than a detailed account of an AI experiment in the realm of drug development, it did not add any additional dimensions to the analysis and comparison of the other papers. It did not neatly fit into any of the categories of characteristics identified within the other literature though this was not due to a novel subject matter representing an additional category but rather because of the format and goal of this particular type of research paper. It was therefore excluded from the in depth review in the following chapter. The final literature selection was therefore 14 research papers strong.

The entire list of search results can be found under *References*. Excluded papers are also listed for the sake of completeness.

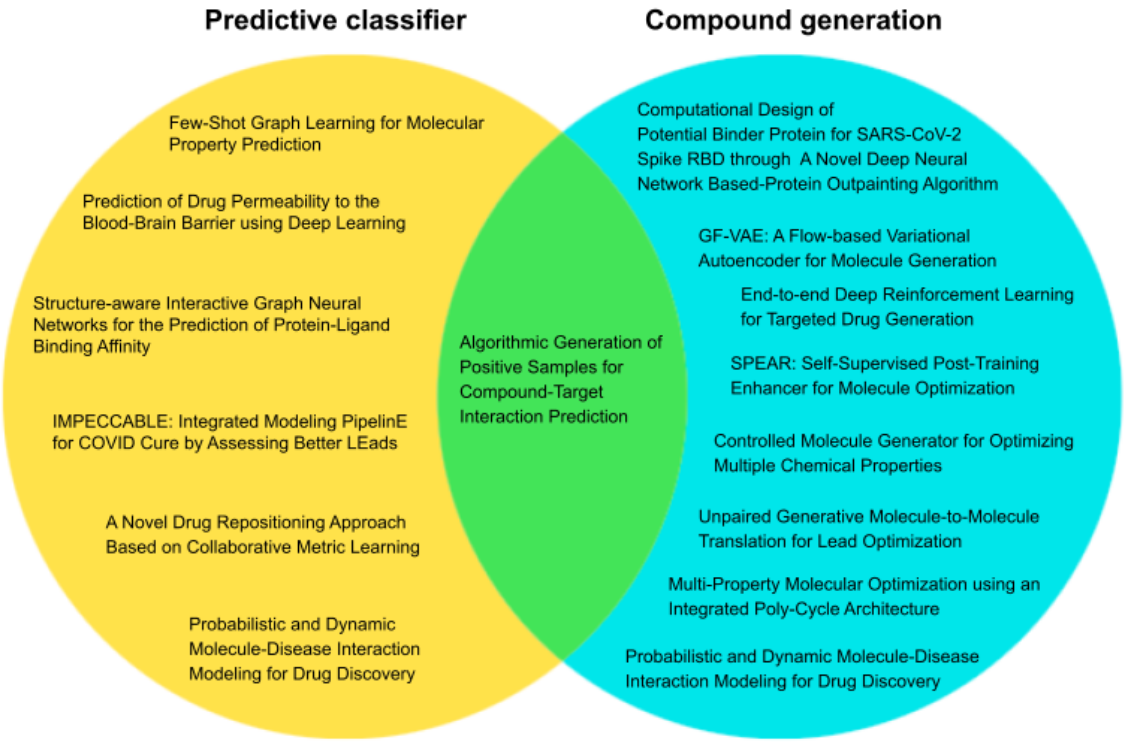
# 3. Literature Review

This chapter will be a review and analysis of the research papers that make up the final literature selection for this thesis. The analysis will examine each paper in order to point out similarities with and important differences to the other works in the selection with regards to two main aspects: the type of result produced by the application of the artificial intelligence technique and the novelty that the described experiments represent in terms of scientific advance within bioinformatics-driven drug development.

The final section of this chapter briefly outlines further investigation that could be completed into the real-life application of the techniques developed by the authors of the analysed works but could not be undertaken within the scope of this thesis.

## 3.1 Result type

When comparing the selected literature on artificial intelligence techniques in drug development, most papers produce one of two major types of results: they either produce a prediction or they generate new data. This subchapter will detail the types predictions that can be made using the AI approaches from the selected papers as well as the types of data that can be generated as both of these major categories can produce a range of types of results. Furthermore, this subchapter will highlight a research paper that does not fit neatly into these categories and outline the reasons for this as well.




*Img. 3: Classification of papers by the result type of their proposed AI technique. Source: D. Flemming*

### 3.1.1 Predictive results

Six of the research papers produce an AI model that can be fed data and make predictions about identically structured but previously unseen data. One of the forms these predictions can take is that of a binary classifier. This type of result is produced by techniques that aim to predict if a protein and a ligand will bind or if a chemical compound can pass the blood-brain barrier.

In “Few-Shot Graph Learning for Molecular Property Prediction” Guo et al. utilise a graph neural network to predict whether a given molecule has a particular molecular property or not. While many of the papers analysed in the subchapter on compound generation results ended up focusing on the same, very narrow set of molecular properties, Guo et al. leverage two different datasets that contain information about very different molecular properties such as “toxicity on 12 biological targets” and side effects on “27 system organ classes” (p. 2563).

Similarly, in their research paper “Prediction of Drug Permeability to the Blood-Brain Barrier using Deep Learning” the authors Atwereboannah et al. describe experiments with two different types of neural networks to achieve a binary classification task. During their research they investigated data from databases for central nervous system drugs in order to make predictions about whether or not a compound can cross the blood-brain barrier.

Another type of prediction that these AI techniques can produce are what can be classified as scores. This is the result that Li et al. produce in their paper on “Structure-aware Interactive Graph Neural Networks for the Prediction of Protein-Ligand Binding Affinity”. Instead of classifying a protein-ligand interaction as binding or not, they aim to predict the strength of that interaction which is called the binding affinity. A higher score therefore indicates a higher binding affinity which is also more desirable for the purposes of drug development.

A research paper that falls into this predictive score category is “IMPECCABLE: Integrated Modeling PipelinE for COVID Cure by Assessing Better LEads” by Al Saadi et al.. This paper is a bit of an outlier in that it does not focus mainly on the artificial intelligence technique that is applied to the data, which in this case centers around the Covid virus, its genetically encoded proteins and potential ligands. Instead, the paper mainly details the architecture used to integrate various stages of the drug development pipeline and the various machine learning techniques applied at each stage. This integration results in a significantly faster workflow that “deliver[s] 100× to 1000× improvement over traditional methods, and [..] speeds up drug discovery by orders of magnitudes” (Al Saadi et al., p. 2).

The actual machine learning and in particular deep learning strategies are only mentioned in passing because the researchers intend to highlight the improvements in efficiency due to their workflow architecture rather than the inner workings of their AI models or their training and validation. Still, the paper outlines the results produced by the machine learning approaches at the different stages of the drug development workflow. Though they apply different constraints and data preprocessing strategies, they all result in predicted scores for things such as binding affinity between a target protein and a ligand. Additionally, these scores are collected as a recommendation list by ranking the predictions.

Another scoring technique is presented in “A Novel Drug Repositioning Approach Based on Collaborative Metric Learning”. Researchers Luo et al. investigated data stemming from observations about drug-disease associations. They look at already observed uses of certain drugs for certain diseases “to infer new uses for existing drugs” (Luo at al., p. 463) which is called drug repurposing. Their technique is able to produce a recommendation list that will take a drug and generate a list of diseases that can potentially be treated with this drug. This recommendation list can be produced for drugs that already have a list of indications as well as for new drugs that have no documented drug-disease associations.

The final paper in this category, “Probabilistic and Dynamic Molecule-Disease Interaction Modeling for Drug Discovery” by Fu et al., actually produces both predictive results and also newly generated molecule compositions. However, unlike the paper described in the subchapter “Hybrid” the two types of results by Fu et al. are not intertwined or in any way dependent on one another. The predictive results they produce are similar to the work by Luo et al. as they also produce a recommendation list of drugs that can be repurposed for a particular disease by ranking the candidates by highest likelihood. The generative results will be detailed in the next subchapter on compound generation results.

The approach proposed by Fu et al. differs from the one outlined by Luo et al. along some technical details regarding the artificial intelligence techniques they use to make their predictions. They also incorporate different data sources and structure their training data differently. While both experiments utilise the chemical compound notation and information from the publicly available database DrugBank, Luo et al. obtain their drug-disease association data from the Comparative Toxicogenomics Database and a 2011 paper by researchers A. Gottlieb, G. Y. Stein and E. Ruppin.

On the other hand, Fu et al. extract data from a publicly available database for clinical trial records. Additionally, they include disease data from “ICD-10 Coding System [...] [which] is a medical classification list by the World Health Organization” (p. 409). Due to the hierarchical nature of the coding system, this allows the researchers to draw out ancestor information about diseases, signs and symptoms (compare p. 409). This allows them to include “time dependency constraints to model evolving drug-disease

relations using a probabilistic deep learning model that can quantify model uncertainty" (Fu et al., p. 404), which the more static approach designed by Luo et al. does not. Fu et al. also use additional chemical compound data from the ZINC Molecule Database which should be mentioned for the sake of completeness but does not represent a significant difference between the two experiment setups as this data is similar to the data from the ZINC database.

During the preliminary research, many of the analysed works focussed on producing binary classifiers for compound target interaction, so answering the question "Will it bind?" with a yes or no. In order to answer that question, there needs to be both a list of target proteins and a list of possible chemical compounds. Binary data produced from observations in wet lab experiments about their interaction forms the basis for training and validation of the AI model that aims to make a binary classification.

So for all of the compounds that have been discovered and are therefore described in these databases but have not yet been tested against all or even any of the relevant target proteins, this type of technique will make a more or less accurate prediction. However, there is also a large number of compound configurations that have not even been described and therefore also have not been tested for their interaction with various target proteins yet. As the research papers selected for this thesis showed, there has been a shift in focus towards approaches that generate these novel and unexplored molecule configurations which the next section will describe in more detail.

### 3.1.2 Compound generation

Within the literature selection for this thesis there are a total of eight research papers that can be grouped together into one category based on the kind of result they produce. These are the works that detail experiments and artificial intelligence techniques that result in the generation of new chemical compounds. Because these are bioinformatics papers, these compounds are not actually physically generated. Instead they are represented in a format such as the SMILES notation or a 2D/3D structured graph.

There is one paper that stands out among the compound generation models as it utilises a unique outpainting approach. In "Computational Design of Potential Binder

Protein for SARS-CoV-2 Spike RBD through A Novel Deep Neural Network Based-Protein Outpainting Algorithm" authors Duan and Sun develop an approach that designs a stable, thermally resistant and easily produces protein as an antiviral blocking agent for the SARS-CoV-2 spike protein RBD. They do not produce a whole protein from scratch. Instead, they rely on previously discovered data about a peptide that has the capability to block the Covid spike protein from binding with human cells. In their approach, this peptide is used as a hotspot motif around which the rest of the protein is designed by the neural network in a method called "outpainting". Outpainting results in the design of a much larger molecule that includes the peptide as the main binding agent for the Covid spike protein but stabilises the peptide for delivery into the organism and also improves the binding capabilities when compared to the stand-alone peptide.

The outpainting method designed by Duan and Sun differs significantly from the techniques detailed in the remaining papers, because it very severely restricts the possible protein molecules that can be generated. This is obviously not an unintended weakness but by design because of their narrowly targeted desire to develop potential drugs for a specific disease. Other researchers approach the question of molecule generation with a broader goal.

For example, in "GF-VAE: A Flow-based Variational Autoencoder for Molecule Generation" Ma and Zhang combine a variational autoencoder with a flow-based model that is used as the decoder in order to generate novel molecular graphs. In an additional experiment, they introduce an optional step for molecular optimisation. During this optimisation step a newly generated molecular structure is modified based on two predefined properties. The makeup of the molecules that can be generated by Ma and Zhang's model is only restricted by the constraints due to the kinds of molecules present in the training data which are molecules consisting of a "maximum [of] 38 atoms" (Ma & Zhang, p. 1185). Within those limitations the model is free to arrange atoms in any form that results in a physically valid structure during the generative first step.

The method described in "End-to-end Deep Reinforcement Learning for Targeted Drug Generation" by Pereira et al. also generates molecular structures, this time in the form of SMILES representations. It is not restricted by the need to include a particular

peptide in the way Duan and Sun's approach is but it is also not as broadly designed as the GF-VAE by Ma and Zhang. The goal for Pereira et al. is to generate molecules that will target a specific κ-opioid receptor. This is achieved by limiting the training data to known ligands for the desired receptor.

Similarly loosely restricted is the approach proposed in "SPEAR: Self-Supervised Post-Training Enhancer for Molecule Optimization" by Fu et al.. However, this research paper focuses solely on the task of molecule optimisation. It is not intended as a stand-alone method but rather designed to be used in concert with a graph-based generative AI model to improve the results of the first model. SPEAR therefore takes a newly generated molecular graph produced by any suitable molecule generation method and further modifies the graph to improve the molecule with regards to the desired chemical property. The only restriction present in the model is the molecule properties for which the generated molecule will be optimised. The paper used three specific properties to train the result enhancer, one of which is the biological reactivity with a particular dopamine receptor. Therefore the generated molecules will only be optimised for these three properties and produce compounds that will interact with the dopamine system. However, given appropriate training data, the enhancer could just as easily be trained for other properties.

In "Controlled Molecule Generator for Optimizing Multiple Chemical Properties" authors Shin et al. design a novel molecule generator that already includes a molecular optimisation step. Their molecule generator takes inputs of known molecules and generates a list of twenty novel molecules that exceed the desired similarity threshold when compared to the input molecule. Much like the SPEAR method, the approach described by Shin et al. defines three chemical properties for which the newly generated molecules are optimised. Since Shin et al. choose the same dopamine receptor as one of their optimisation properties as Fu et al. do, this molecule generator has the same limitations as the SPEAR enhancer.

Researchers Barshatski and Radinsky co-authored two papers in this selection in the same vein as Fu et al. and Shin et al.. In "Unpaired Generative Molecule-to-Molecule Translation for Lead Optimization" they develop an unsupervised approach that generates novel molecules based on a molecule input, also called molecule lead. The output molecules are finally evaluated to see if they represent an improvement over the

molecule lead with regards to the by now familiar molecular properties like the DRD2 biological activity score and drug likeness. This means, like the optimisation methods reviewed for this thesis, Barshatski and Radinsky's molecule generator is also subject to the limitations posed by the properties chosen for this evaluation step and the quality of the data available about the properties.

Based on their research presented in the aforementioned paper, Barshatski and Radinsky team up with their colleague Nordon for "Multi-Property Molecular Optimization using an Integrated Poly-Cycle Architecture". Here they take their generative model one step further and include molecular optimisation for the same DRD2 biological activity and drug likeness properties that are also used in Fu et al.'s and Shin et al.'s approaches. Though Barshatski et al. highlight that the drug likeness score can be viewed as "s a combination of molecular properties such as solubility, ligand efficiency, molecular weight, etc." (p. 3730) and therefore by optimising for this likeness score the model is really optimising for multiple properties at once, in the end this approach carries the exact same limitations as the works by Fu et al. and Shin et al. reviewed for this thesis with regards to the type of result this model can produce.

As already mentioned in the previous section on AI approaches that produce predictive results, the method described in "Probabilistic and Dynamic Molecule-Disease Interaction Modeling for Drug Discovery" also produces new chemical compound structures in addition to the recommendation list for drug repurposing. This is possible because the technique models the interaction between a molecule and a target disease based on data available from published clinical trials. Fu et al. are therefore able to calculate molecule-disease matching score. Utilising this scoring mechanism, the researchers are able to generate a very large list of molecules intended to match a given disease. This large data set is narrowed down by ranking the best 500 molecules based on their drug likeness score. By using test data for which there is a target molecule that has been proven an effective treatment for the disease, Fu et al. can evaluate the 500 highest ranked novel molecules by similarity to the target molecule, the LogP score which evaluates the chemical properties of r
"ring size and synthetic accessibility" (p. 409) and the drug likeness score used by the model itself.

Though not as constrained as the outpainting method by Duan and Sun, this last approach to compound generation by Fu et al. presents different limitations than the other generative approaches. Because it learns from interactions between a wide range of diseases and drugs the model is likely more easily transferable to an unknown disease for which there is not already a known peptide with high levels of interactivity. Unlike the optimisation methods it is also not limited to developing drugs that interact only with a specific opioid or dopamine receptor. However, since the data on which this drug-disease-interaction at the core of this method is modeled comes from clinical trials and even includes weights this data based on the evolution of those results over time it is highly dependent and therefore shaped by the clinical trials which are being conducted. If they concentrate on particular diseases and specific compounds while disregarding others this bias can negatively affect the accuracy of the results for diseases outside of the well-explored selection for which there are clinical trials available.

### 3.1.3 Hybrid

Interestingly, there is one research paper, “Algorithmic Generation of Positive Samples for Compound-Target Interaction Prediction” by Nanor et al., in the analysed selection that represents a hybrid approach. It actually produces both complex data in the form of novel molecule representations and a binary classification. The overall goal of the experiment is a simple binary classification that estimates if a chemical compound will bind to a target protein from the human genome.

In order to achieve this binary classification, an AI model is trained on a dataset that contains positive and negative samples. A positive sample is a combination of a target protein and a chemical compound that will bind whereas a negative sample is a combination that will not bind. Within the publicly available datasets “[t]here exist more negative labeled compound-target pairs than positive labeled ones” (Nanor et al., p. 42). The ratio of positive to negative pairs “is 1:11” (Nanor et al., p. 43). As explained by the authors, this leads to a negative bias in models for compound-target interaction (CTI).

A negative bias means that there will be many false negatives, so compound-target interaction that will be labelled as negative even though in reality the chemical

compound will bind to the protein target. This is problematic for researchers actually working with these types of AI approaches to look for new chemical compounds that do bind as they will be missing potentially medically useful compounds due to these errors.

To improve the F1-score and recall of these predictions, Nanor et al. propose an approach where novel molecules are produced in a preliminary step to address this imbalance in the training datasets. The training data for the AI models is padded with numerous self-generated molecules as a way of balancing the dataset to include the same number of positive and negative samples.

In their experimental setup Nanor et al. use two different datasets balanced in this way to train nine different classifiers and compare the resulting F1-score and recall to the same classifiers trained on the original unbalanced dataset. Maybe surprisingly this approach proved highly effective in improving the F1-score for every single one of the nine AI models in this experiment. The recall score also improved in almost every case. (Compare p. 48, Nanor et al.)

## 3.2 Novelty and Scientific Advance

Evaluating the novelty of these research papers is worthwhile to determine the relevance of artificial intelligence for the purposes of drug development. For a very young discipline that has not been explored thoroughly, this score will be high. Many papers in a representative selection will experiment with artificial intelligence techniques to achieve a goal in ways that have not previously been documented. There will be few closely related works and no replications of the same experiment by other scientists.

A high novelty also indicates that there has not yet been a lot of scientific discourse about the presented research which would help discover problems or errors in the experiment setup and the evaluation of its results. This would make the actual application of these novel approaches to real-life drug development a high risk and

potentially unreliable enterprise for the scientists, research groups and companies invested in the discovery of new drugs.

Once a discipline has had time to mature, there will be lower novelty in the results presented within research papers. Successful approaches have been identified, unsuccessful ones weeded out. New research might focus on validating successful experiments by reproducing them. Papers might investigate new data sources, optimised architecture, a combination of previously explored but not yet combined techniques or new data preparation methods.

As this section of the review demonstrates, the use of AI in drug development is a maturing field. There are many papers that build on and even incorporate extensive previous research. This is also reflected in the literature selection for this thesis. The degree of novelty of the AI techniques that they investigate varies.

### 3.2.1 Probabilistic and Dynamic Molecule-Disease Interaction Modeling for Drug Discovery

This research paper by Fu et al. describes a very novel approach for both the recommendation task for drug repurposing as well as the novel molecule generation for a target disease. Both of these tasks are popular use cases for artificial intelligence applications during the drug discovery process and are therefore explored by other papers in the literature selection for this thesis and also by numerous other researchers in this field. However, no other works neither in the literature selection nor during the preliminary research phase integrated clinical study data in the way that Fu et al. present here. Particularly unique in their approach is how they utilise “time dependency constraints on drug and disease representations” (p. 405). Their approach uses weighting of data so that drug-disease-interaction data which has been superceded by more recent trial results is not given the same gravitas as the newer discoveries during the training phase of the model.

### 3.2.2 Few-Shot Graph Learning for Molecular Property Prediction

The technique for molecular property prediction developed in this paper is rather unique not just among the subset of classification methods but also among the broader

group of approaches focussing on molecular properties. It is unique both because of the datasets the authors chose to train and evaluate their model which no other researchers from the literature selection focussed on and also because of the specific molecular properties that they investigate which are also not used by any other research team but grant this method a much broader potential range of application in pharmaceutical research.

### 3.2.3 Structure-aware Interactive Graph Neural Networks for the Prediction of Protein-Ligand Binding Affinity

Though this work by Li et al. is the only piece of research focussing on the protein-ligand binding affinity in this literature selection, it stands on the shoulders of giants in this field. The authors identify many of these preceding works in their introduction and many more papers like this were identified during the preliminary research phase for this thesis. Even though other works reviewed here also work with graph neural networks as their AI technique of choice, Li et al. point out that they are at the forefront of applying them “from the perspective of polar coordinates for structure-based binding affinity prediction” (p. 976) which none of the other investigated papers contradict.

### 3.2.4 Algorithmic Generation of Positive Samples for Compound-Target Interaction Prediction

This proposed method to increase the accuracy of AI methods that predict compound-target-interaction (CTI) is shaped by the data that is available for training AI models. Although it addresses a common problem shared by all artificial intelligence approaches to this kind of predictive task based on the available data, it takes a very unique approach to solving it. The novelty of this approach is demonstrated by the fact that the authors of the paper could only identify two related works that employ self-generated data to balance training datasets within the whole discipline of bioinformatics, both of which were unrelated to CTI. This approach was also unique within the literature selection for this thesis as well as the preliminary literature review for it. The idea can therefore be categorised as having a high degree of novelty.

### 3.2.5 Prediction of Drug Permeability to the Blood-Brain Barrier using Deep Learning

In the section on related works, Atwereboannah et al. highlight multiple machine learning and deep learning techniques that have previously been applied to produce predictive binary classifiers with a fairly high accuracy. As the authors themselves state, deep learning and in particular neural networks, have previously been used and compared to other artificial intelligence techniques for this particular subject. The references demonstrate that there is a very broad body of research for the particular question of how to utilise machine learning and deep learning for the prediction of a chemical compound's blood-brain barrier permeability. However, owing to the search term design and choice of database none of these related works were included in the literature selection for this thesis nor in the preliminary research, thereby making this paper more of an outlier in the literature selection than it really is in the full body of research on this subject.

### 3.2.6 A Novel Drug Repositioning Approach Based on Collaborative Metric Learning

Compared to other works from the literature selection this paper represents a popular category of AI applications in the narrow field of drug repurposing. While the experiment described in this paper is not an exact reproduction of a previously conducted experiment or study, there are certain conceptual similarities to other works, although the authors are of course justified in calling this particular drug repositioning approach novel since they are not identical methods. Apart from the excluded review paper there is also another research paper in the literature selection that focuses on repurposing existing drugs.

In their introduction Luo et al. also detail a very long list of related works that use very similar approaches to specifically identify new indications for existing drugs. Furthermore, in their evaluation they are able to benchmark their approach against "three state-of-the-art methods: [inductive matrix completion], [neighborhood regularized logistic matrix factorization], and [drug repositioning recommendation system], which have been applied in drug repositioning" (Luo et al., p. 467). This indicates that there is a wide body of research they have built and expanded on, as well as scientific consensus on reliable techniques for achieving this goal.

However, there are multiple novel elements to their approach called CMLDR. As the authors point out, their model is trained on data that describes drug-disease associations. It can also take similarity data into account, but unlike the related works they describe, this approach does not require drug-drug similarity or disease-disease similarity data to make its predictions. Additionally, their experiments have shown that “CMLDR consistently outperforms other competing methods in terms of precision and recall rate at different top ranked predictions, AUPR and AUC values”.

### 3.2.7 IMPECCABLE: Integrated Modeling PipelinE for COVID Cure by Assessing Better LEads

This research paper cannot reasonably be evaluated on the basis of novelty, neither for the pipeline presented by the authors nor for the individual machine learning techniques applied at the various stages of the pipeline. The pipeline itself as well as the improvements in drug development speed cannot be compared with the contents of the other papers in the literature selection, as it is an entirely different subject matter than the data processing and algorithmic analysis described by the other literature.

Additionally, the authors did not include an analysis of related works or any benchmarks produced by comparable frameworks with different artificial intelligence techniques. This means that placing this paper in the proper context of this subject matter, distinguishing the features that represent its uniqueness and discerning how much of a scientific advance was achieved go far beyond the scope of this thesis.

The machine learning and deep learning techniques used within the pipeline could theoretically be assessed and compared to the other works. However, due to the focus of the paper the background and details on the artificial intelligence approaches in use are so sparse that this comparison could only produce unreliable and skewed results. This particular paper will therefore not be judged on its novelty factor.

### 3.2.8 Computational Design of Potential Binder Protein for SARS-CoV-2 Spike RBD through A Novel Deep Neural Network Based-Protein Outpainting Algorithm

The novelty factor of the method described by Duan and Sun is very high. No related works matching the core identifiers of this research paper are identified by the authors. There are also no related works that could be identified, neither within the preliminary research nor within the final literature selection for this thesis.

Like most research, Duan and Sun build on already established techniques. Therefore they describe in detail the previously developed protein folding prediction and protein design methods that they incorporate into their own technique and list works that explain these further. However, what makes their approach stand out very distinctly among the literature selection is the unique way they combine it with the also well-established image inpainting/outpainting technique from the field of computer vision. According to the researchers this technique has not previously been applied to the protein design process used in drug development in a documented experiment, study or research publication making this approach very unique among the evaluated methods.

### 3.2.9 GF-VAE: A Flow-based Variational Autoencoder for Molecule Generation

There is a relatively high novelty factor in Ma and Zhang's work too. Their proposed Graph Flow - Variational AutoEncoder (GF-VAE) is unique in this regard among the analysed literature. While the authors refer to numerous related works that have developed fairly advanced variational autoencoders and flow-based models that will generate molecular structures from representations such as SMILES notation input or molecular graphs, their approach is the only one to combine the two in this way.

The GF-VAE also represents a major advance in the novelty, uniqueness and reconstruction scores that Ma and Zhang use to evaluate their model as well as multiple state-of-the-art variational autoencoders and flow-based models. Additionally, through their experiments they are able to show that the GF-VAE is able to beat the benchmarks set by the other methods while needing significantly less time during the training phase.

### 3.2.10 End-to-end Deep Reinforcement Learning for Targeted Drug Generation

Pereira et al. do not aim for a previously unexplored method but instead focus on improving the process of generating valid SMILES representations of molecules that have a high likelihood of possessing the property of binding to a specific opioid receptor. This is achieved by combining an RNN that produces SMILES strings which are then evaluated for their fitness by an interrelated second neural network which helps improve the first network through reinforcement learning.

It is important to note that while the experiment outlined in the paper is limited to the κ-opioid receptor this approach can be easily modified to work for any desired protein target for which there is a sufficiently big ligand dataset described in the available compound and chemical property libraries. The scientific advance this method enables can therefore be significantly influenced by the chosen target.

### 3.2.11 SPEAR: Self-Supervised Post-Training Enhancer for Molecule Optimization

SPEAR presents a unique method among the analysed literature as it is not meant to be a standalone solution to the problem of molecule generation. Instead it is supposed to work as an add-on to any graph-based molecule generator. This may mean that it will not be outdated as quickly as other methods if new state of the art molecule generation techniques emerge. The molecule generator as designed by Fu et al. will be able to take advantage of these advances as can just be applied in combination with the new technique.

The experiment outlined in the paper also represents significant potential for advance in the field of drug development. Even though the experiments focus on three predefined properties for which the generated molecular graphs are optimised, any property for which there is sufficient training data could be used in this approach.

Overall, the paper investigates a subject that other papers in this literature selection also delve into, albeit with different artificial intelligence methods and model architectures. For example, Fu et al. set themselves the task of leveraging the entire molecule database as opposed to the much smaller subset of labelled molecule pairs which form the basis for other molecule optimisation techniques. They achieve this with

a self-supervised approach whereas Barshatski and Radinsky achieve the same goal with an entirely unsupervised approach in “Unpaired Generative Molecule-to-Molecule Translation for Lead Optimization”. Similarly, in their paper Shin et al. choose the same goal of molecular optimisation, utilising a near identical set of three molecular properties for optimisation.

### 3.2.12 Unpaired Generative Molecule-to-Molecule Translation for Lead Optimization

Much like Shin et al. work with unspecified molecular sequences as the input format, Barshatski and Radinsky work with molecular sequences in the SMILES format as opposed to the Fu et al. for the SPEAR method,, who instead work with graph representations. However, the authors of this work develop a method that is almost identical in its goals, which is to optimise a baseline molecule for selected molecular properties. Indeed, the two selected properties are the same dopamine receptor and drug likeness score that the other groups of researchers exploring molecule optimisation techniques within the literature selection for this thesis also use. It is important to note that only one property can be optimised for during any training and testing cycle.

Because Barshatski and Radinsky, much like Fu et al., manage to train their model on unpaired molecule data they are able to document improvements over previous state of the art results as their technique generates “more successful molecules with [a] higher desired property score” (Barshatski & Radinsky, p. 2562).

### 3.2.13 Controlled Molecule Generator for Optimizing Multiple Chemical Properties

While ultimately this research paper works on the same goal as the SPEAR enhancer and the Unpaired Generative Molecule-to-Molecule Translation for Lead Optimization, there is a major difference in the approach outlined by Shin et al. in this paper. The researchers do not run separate training cycles for each molecular property and then evaluate individual molecular property results as the other two works do. Instead, there is one training cycle where the model takes into account optimisation for all properties

at the same time and the results which are then evaluated against baseline results from other approaches all stem from a single, simultaneous optimisation process.

Nevertheless, for the time being this molecule generator with multiple property optimisation “outperforms all other baseline approaches in both single-objective optimization and multi-objective optimization by a large margin” (Shin et al., p. 152). It has also proved its practicability in an experimental drug case study (compare Shin et al., p. 152), demonstrating its scientific value for real-life drug discovery processes.

### 3.2.14 Multi-Property Molecular Optimization using an Integrated Poly-Cycle Architecture

This work by Barshatski, Nordon and Radinsky represents a continuation of the method described in “Unpaired Generative Molecule-to-Molecule Translation for Lead Optimization”. It focuses on adding a discrete optimisation path to the model architecture for each desired molecular property. While the experiments outlined in this research paper specifically repeat the use of the dopamine receptor and drug likeness score also used by Barshatski and Radinsky in their earlier work as well as by the authors of the other papers in the literature selection that are dealing with molecular optimisation, for this work the focus is optimising both properties simultaneously. It does so with great success when evaluated against two other, graph-based multi-property molecular optimisation methods. Beyond that advancement in the model design the approach described in this research paper shares most of its features with the previous work by Barshatski and Radinsky as well as significant similarities with the other molecule optimisation techniques reviewed for this thesis.

## 3.3 Application Range

It is also worth investigating how broadly the results of the AI techniques studied in the selected research papers can be applied. If a paper presents an approach that allows other scientists to train a comparable model on the same data and then use that model to shorten the time required for the development of a new drug, then the paper presents a scientific advance that will result in patients benefiting from the earlier

availability of drugs to treat their diseases. A similar advantage can be gained if the technique is demonstrated on a specific dataset but can easily be transferred to new datasets without needing to make major adjustments.

On the other hand, if a paper presents an experiment that produces results very narrowly tailored to one particular protein, molecular compound or disease, then extensive further research is required to investigate if the approach is transferable at all. This type of specificity of the application range may not be as valuable to the wider scientific community beyond the original researchers and those already utilising their results.

### 3.3.1 In industry use

When looking at the possibilities for application it is important to highlight the two papers by Al Saadi et al. and Barshatski et al.. Both groups of authors produced an artificial intelligence approach to a problem within the drug development pipeline that was immediately put into use within the pharmaceutical research branches of the institutions supporting their work.

Al Saadi et al. describe in their conclusion how the approach has been “used to screen over 4.2 billion molecules [4] against over a dozen drug targets in SARS-CoV-2, leading to the identification and experimental validation of over 1000 compounds, resulting in over 40 hits that are progressing to advanced testing” and that it “has enabled the [US Department of Energy’s National Virtual Biotechnology Laboratory] to discover a promising anti-viral drug candidate” (p. 11).

Similarly, Barshatski et al. developed an artificial intelligence based method for lead optimisation which at the time of publication was “being deployed for use in the [Isreali Technion - Israel Institute of Technology’s] Targeted Drug Delivery and Personalized Medicine laboratories generating treatments using nanoparticle-based technology” (p. 2554).

None of the other papers mention that their findings are actually applied in a pharmaceutical drug development context. Still, it is possible that other authors’ work may also currently be in industry use and there may be reasons for the authors’

omission of this information. It is conceivable that the authors did not include this information because their approach was adopted after publication either with or without their involvement or even that the authors are bound by the pharmaceutical company's non-disclosure agreement. Without further study and potentially interviews with the respective authors of the papers or industry experts it is not possible to establish if that is the case for any of the artificial intelligence techniques detailed by the literature selection. However, this type of investigation is out of scope for this thesis.

# 4. Discussion

After reviewing the selected 15 research papers with regards to the type of results they produce and their scientific novelty, it is important to highlight the limits and biases for this thesis. This chapter intends to lay out these limits, describe the known biases and outline what further analysis might be undertaken for a more detailed response to the question posed in the title of this thesis.

The most significant factor influencing the content of the literature selection and consequently the quality of this review is the limited domain knowledge of the thesis' author in the broader field of bioinformatics and the more specific pharmaceutical discipline of drug development. As the work on this thesis was designed specifically to increase the familiarity with this subject it is possible that the author's research and subsequent design of the search query was constrained by blindspots not sufficiently illuminated during the preliminary research phase.

Also limiting the analysis in this thesis is the choice of sources. As already described in the chapter on the methodology for this literature review, only one digital library was utilised during the research phase of this work. Although the ACM Digital Library describes itself as "the world's most comprehensive database of full-text articles and bibliographic literature covering computing and information technology" on its website it does not contain every single bioinformatics paper published during the designated time frame for this analysis.

In fact, the same query on the PubMed database yields 15,280 very different results. It is therefore possible that this literature selection has skewed the perception of the author of this thesis regarding the current state of research on the subject of artificial intelligence applications in the early stages of the drug development pipelines. This could affect the accuracy of assessments regarding the level of novelty and the amount of related work for any given paper in the analysed selection. It is also possible that there is research on other artificial intelligence techniques in the realm of drug development potentially with entirely different types of results that were not accessible via the ACM Digital Library but could be found in other databases.

Additionally, the research question could have been interpreted in a much broader way. As is, this thesis focuses on individual AI techniques that are already in use or intended to eventually replace specific resource intensive parts of the scientific discovery process directly related to the chemical compounds that will be used in medications. There is also the possibility that there are AI tools conceivable that improve the drug discovery process in other ways.

Computer scientists could develop tools that can help improve the workload distribution in labs, digitisation of results, the result verification, peer reviewing, publishing and many more. This may help speed up development and regulatory approval times or improve international cooperation and resource distribution potentially far outweighing the advances made by applying artificial intelligence methods to molecular problem solving. Answering the research question in that way would have required a very different approach to the design of the search query and resulted in a very different analysis of the results.

Limiting the search query to terms related to the first two stages of the drug discovery pipeline also constrained the possible analysis and including later stages could have highlighted very different techniques or new applications of the techniques explored in the selected papers.

# 5. Conclusion

After investigating a small part of the published current research on artificial intelligence applications within the early stages of the drug development pipeline, a few conclusions seem to emerge.

1. **Restrictions on the application of the AI techniques are mostly determined by the publicly available data**

   The publicly available datasets in some cases limit the precision of the results produced by the AI technique because there is very little training data, such as the paired molecules used during model training for molecular optimisation tasks. In other cases they limit what kind of goal can be achieved because it is not possible to optimise a molecule for a property for which there is no publicly available data.

   This means that some of the works on the subject of this thesis do not focus directly on developing or improving drugs for particular diseases. Instead, a large part of their efforts are invested in the data preparation that will help balance datasets to improve the accuracy of both predictive and generative AI methods.

2. **Artificial intelligence applications are already in use during drug development.**

   This is confirmed by the research papers on “Multi-Property Molecular Optimization using an Integrated Poly-Cycle Architecture” and “IMPECCABLE: Integrated Modeling PipelinE for COVID Cure by Assessing Better LEads”, detailed in the subchapter in industry use.

   Some of the other works describe what can reasonably be considered a sort of scientific proof-of-concept for AI application in this pharmaceutical field of research. Their approaches are demonstrated on limited data, for example by limiting their method to optimising molecules for a specific and well-known cell receptor or peptide. While these methods demonstrate the potential

improvements with regards to time and resource efficiency that artificial intelligence can lend the drug development process, they must be adapted still if they are to be used for the many receptors or peptides not originally included in the research. However, the two papers mentioned above already tailored their applications to the needs of researchers already working on the development of new drugs and will not just show that improvements due to AI are possible but actually make these advances a reality.

3. **There is a wide range of possible drug development tasks that can be supported or improved by the application of artificial intelligence techniques.**

   While the literature selection appeared to highlight a research focus on molecule optimisation with four papers specifically presenting methods to design molecules with a high binding affinity for the dopamine receptor DRD2, there are actually many more tasks where AI can be applied:

   - *Preliminary ligand screening via binding predictions*
     This can significantly reduce the number of wet-lab experiments that are required to find binding interactions between compounds and a desired target.
   - *Preliminary compound screening for membrane permeability*
     Much like the binding prediction this greatly reduces the number of compounds that need to be tested to find a selection capable of permeating a desired membrane.
   - *Improving molecule design to improve binding capability*
     This work enables easy and fast improvements of known compounds to increase their efficacy and potentially reduce side effects.
   - *Repurposing existing drugs for new diseases*
     Instead of relying on the accidental discovery of a secondary effect that a drug has or randomly trialling large numbers of existing drugs suspected of offering these secondary effects, the AI-supported repurposing both narrows down the number of experiments required to find additional disease applications for a drug and is also capable of recommending drugs for diseases that potentially would have evaded human researchers instincts.

- *Discovering novel compounds for a disease target*
  AI can design a range of novel compounds to target new or previously untreatable diseases.

It seems therefore justified that resources both at public research institutions as well as at commercially oriented pharmaceutical companies are being poured into further research on this subject. As more and more of these results are made public, in the form of research publications for the former and somewhat less detailed press releases for the latter of these institutions, it remains a fascinating subject with plenty more discoveries yet to be made.

# References

## Search results for the literature reviews

### Search query

https://dl.acm.org/action/doSearch?fillQuickSearch=false&target=advanced&ContentItemType=research-article&expand=dl&AfterMonth=11&AfterYear=2020&BeforeMonth=10&BeforeYear=2021&AllField=AllField%3A%28%22machine+learning%22+OR+%22deep+learning%22+OR+%22artificial+intelligence%22%29+AND+Abstract%3A%28%22drug+discovery%22+OR+%22target+identification%22+OR+%22molecule+generation%22+OR+%22drug+development%22+OR+%22target+interaction%22+OR+%22drug+design%22+OR+%22drug+target%22%29&startPage=0&pageSize=50

**Papers printed in bold**: excluded from literature review based on title

### Research Papers


Abena Achiaa Atwereboannah, Wei-Ping Wu, and Ebenezer Nanor. 2021. Prediction of Drug Permeability to the Blood-Brain Barrier using Deep Learning. In *4th International Conference on Biometric Engineering and Applications* (*ICBEA '21*). Association for Computing Machinery, New York, NY, USA, 104–109. DOI:https://doi.org/10.1145/3476779.3476797

**Jianguo Chen, Kenli Li, Zhaolei Zhang, Keqin Li, and Philip S. Yu. 2021. A Survey on Applications of Artificial Intelligence in Fighting Against COVID-19. *ACM Comput. Surv.* 54, 8, Article 158 (November 2022), 32 pages. DOI:https://doi.org/10.1145/3465398**

Ebenezer Nanor, Wei-Ping Wu, Strato Angsoteng Bayitaa, Victor K. Agbesi, and Brighter Agyemang. 2021. Algorithmic Generation of Positive Samples for Compound-Target Interaction Prediction. In *2021 13th International Conference on Machine Learning and Computing* (*ICMLC 2021*). Association for Computing Machinery, New York, NY, USA, 41–49. DOI:https://doi.org/10.1145/3457682.3457689

Tianfan Fu, Cao Xiao, Cheng Qian, Lucas M. Glass, and Jimeng Sun. 2021. Probabilistic and Dynamic Molecule-Disease Interaction Modeling for Drug Discovery. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (*KDD '21*). Association for Computing Machinery, New York, NY, USA, 404–414. DOI:https://doi.org/10.1145/3447548.3467286

Bingya Duan and Yingfei Sun. 2021. Computational Design of Potential Binder Protein for SARS-CoV-2 Spike RBD through A Novel Deep Neural Network Based-Protein Outpainting Algorithm. In *The Fifth International Conference on Biological Information and Biomedical Engineering* (*BIBE2021*). Association for Computing Machinery, New York, NY, USA, Article 7, 1–8. DOI:https://doi.org/10.1145/3469678.3469685

Changsheng Ma and Xiangliang Zhang. 2021. GF-VAE: A Flow-based Variational Autoencoder for Molecule Generation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (*CIKM '21*). Association for Computing Machinery, New York, NY, USA, 1181–1190. DOI:https://doi.org/10.1145/3459637.3482260

Aymen Al Saadi, Dario Alfe, Yadu Babuji, Agastya Bhati, Ben Blaiszik, Alexander Brace, Thomas Brettin, Kyle Chard, Ryan Chard, Austin Clyde, Peter Coveney, Ian Foster, Tom Gibbs, Shantenu Jha, Kristopher Keipert, Dieter Kranzlmüller, Thorsten Kurth, Hyungro Lee, Zhuozhao Li, Heng Ma, Gerald Mathias, Andre Merzky, Alexander Partin, Arvind Ramanathan, Ashka Shah, Abraham Stern, Rick Stevens, Li Tan, Mikhail Titov, Anda Trifan, Aristeidis Tsaris, Matteo Turilli, Huub Van Dam, Shunzhou Wan, David Wifling, and Junqi Yin. 2021. IMPECCABLE: Integrated Modeling PipelinE for COVID Cure by Assessing Better LEads. In *50th International Conference on Parallel Processing* (*ICPP 2021*). Association for Computing Machinery, New York, NY, USA, Article 40, 1–12. DOI:https://doi.org/10.1145/3472456.3473524

**Seyedeh Shaghayegh Sadeghi and Mohammad Reza Keyvanpour. 2021. An Analytical Review of Computational Drug Repurposing. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 18, 2 (March-April 2021), 472–488. DOI:https://doi.org/10.1109/TCBB.2019.2933825**

Tiago Oliveira Pereira, Maryam Abbasi, Bernardete Ribeiro, and Joel P. Arrais. 2021. End-to-end Deep Reinforcement Learning for Targeted Drug Generation. In*2020 4th International Conference on Computational Biology and Bioinformatics* (*ICCBB 2020*). Association for Computing Machinery, New York, NY, USA, 7–13. DOI:https://doi.org/10.1145/3449258.3449260

Tianfan Fu, Cao Xiao, Kexin Huang, Lucas M. Glass, and Jimeng Sun. 2021. SPEAR: self-supervised post-training enhancer for molecule optimization. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* (*BCB '21*). Association for Computing Machinery, New York, NY, USA, Article 27, 1–10. DOI:https://doi.org/10.1145/3459930.3469530

Zhichun Guo, Chuxu Zhang, Wenhao Yu, John Herr, Olaf Wiest, Meng Jiang, and Nitesh V. Chawla. 2021. Few-Shot Graph Learning for Molecular Property Prediction. In *Proceedings of the Web Conference 2021* (*WWW '21*). Association for Computing Machinery, New York, NY, USA, 2559–2567. DOI:https://doi.org/10.1145/3442381.3450112

**Akram Vasighizaker, Li Zhou, and Luis Rueda. 2021. Cell type identification via convolutional neural networks and self-organizing maps on single-cell RNA-seq data. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* (*BCB '21*). Association for Computing Machinery, New York, NY, USA, Article 91, 1–6. DOI:https://doi.org/10.1145/3459930.3471171**

Bonggun Shin, Sungsoo Park, JinYeong Bak, and Joyce C. Ho. 2021. Controlled molecule generator for optimizing multiple chemical properties. In *Proceedings of the Conference on Health, Inference, and Learning* (*CHIL '21*). Association for Computing Machinery, New York, NY, USA, 146–153. DOI:https://doi.org/10.1145/3450439.3451879

**Paidamoyo Chapfuwa, Serge Assaad, Shuxi Zeng, Michael J. Pencina, Lawrence Carin, and Ricardo Henao. 2021. Enabling counterfactual survival analysis with balanced representations. In *Proceedings of the Conference on Health,***

***Inference, and Learning*** **(*CHIL '21*). Association for Computing Machinery, New York, NY, USA, 133–145. DOI:https://doi.org/10.1145/3450439.3451875**

Shuangli Li, Jingbo Zhou, Tong Xu, Liang Huang, Fan Wang, Haoyi Xiong, Weili Huang, Dejing Dou, and Hui Xiong. 2021. Structure-aware Interactive Graph Neural Networks for the Prediction of Protein-Ligand Binding Affinity. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (*KDD '21*). Association for Computing Machinery, New York, NY, USA, 975–985. DOI:https://doi.org/10.1145/3447548.3467311

**Guang-Hui Liu, Bei-Wei Zhang, Gang Qian, Bin Wang, Bo Mao, and Isabelle Bichindaritz. 2020. Bioimage-Based Prediction of Protein Subcellular Location in Human Tissue with Ensemble Features and Deep Networks. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 17, 6 (Nov.-Dec. 2020), 1966–1980. DOI:https://doi.org/10.1109/TCBB.2019.2917429**

**Xiaoyan Zhu, Yingbin Li, Jiayin Wang, Tian Zheng, and Jingwen Fu. 2020. Automatic Recommendation of a Distance Measure for Clustering Algorithms. *ACM Trans. Knowl. Discov. Data* 15, 1, Article 7 (January 2021), 22 pages. DOI:https://doi.org/10.1145/3418228**

Huimin Luo, Jianxin Wang, Cheng Yan, Min Li, Fang-Xiang Wu, and Yi Pan. 2021. A Novel Drug Repositioning Approach Based on Collaborative Metric Learning. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 18, 2 (March-April 2021), 463–471. DOI:https://doi.org/10.1109/TCBB.2019.2926453

**Bram van Dooremaal, Pavlo Burda, Luca Allodi, and Nicola Zannone. 2021. Combining Text and Visual Features to Improve the Identification of Cloned Webpages for Early Phishing Detection. In *The 16th International Conference on Availability, Reliability and Security* (*ARES 2021*). Association for Computing Machinery, New York, NY, USA, Article 60, 1–10. DOI:https://doi.org/10.1145/3465481.3470112**

Guy Barshatski, Galia Nordon, and Kira Radinsky. 2021. Multi-Property Molecular Optimization using an Integrated Poly-Cycle Architecture. In *Proceedings of the 30th*

*ACM International Conference on Information & Knowledge Management* (*CIKM '21*). Association for Computing Machinery, New York, NY, USA, 3727–3736. DOI:https://doi.org/10.1145/3459637.3481938

Guy Barshatski and Kira Radinsky. 2021. Unpaired Generative Molecule-to-Molecule Translation for Lead Optimization. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (*KDD '21*). Association for Computing Machinery, New York, NY, USA, 2554–2564. DOI:https://doi.org/10.1145/3447548.3467120

## Other sources

ACM Digital Library, https://www.acm.org/publications/digital-library. Accessed 28 January, 2022.

George F. Luger (2008) *Artificial Intelligence: Structures and Strategies for Complex Problem Solving (6th. ed.)*. Addison-Wesley Publishing Company, USA.

IUPAC. *Compendium of Chemical Terminology, 2nd ed. (the "Gold Book")*. Compiled by A. D. McNaught and A. Wilkinson. Blackwell Scientific Publications, Oxford (1997). Online version (2019-) created by S. J. Chalk. ISBN 0-9678550-9-8. https://doi.org/10.1351/goldbook.

Kelleher, J. D. (2019). *Deep Learning*. MIT Press, USA.

LeCun, Y., Bengio, Y. & Hinton, G. *Deep learning*. Nature 521, 436–444 (2015). https://doi.org/10.1038/nature14539.

Leeson, P. D., & Young, R. J. (2015). *Molecular Property Design: Does Everyone Get It?*. ACS medicinal chemistry letters, 6(7), 722–725. https://doi.org/10.1021/acsmedchemlett.5b00157.

Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M. F., Kellis, M., Lindblad-Toh, K., Lander, E.S.. *Distinguishing protein-coding and noncoding genes in the human genome*. In Proceedings of the National Academy of Sciences Dec 2007, 104 (49) 19428-19433; DOI: 10.1073/pnas.0709013104 https://www.pnas.org/content/104/49/19428.

National Center for Biotechnology Information. *PubChem Compound Summary for CID 5962, Lysine*. PubChem, https://pubchem.ncbi.nlm.nih.gov/compound/Lysine. Accessed 28 January, 2022.

Nature Education, “Amino Acid”, Scitable by Nature Education, https://www.nature.com/scitable/definition/amino-acid-115/. Accessed 28 January, 2022.

Nature Education, “Peptide”, Scitable by Nature Education, https://www.nature.com/scitable/definition/peptide-317/. Accessed 27 January, 2022.

Tamir, M., (2020). *What is Machine Learning (ML)?* Berkeley School of Information, https://ischoolonline.berkeley.edu/blog/what-is-machine-learning/. Accessed 4 February 2022.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). *Applications of machine learning in drug discovery and development.* Nature reviews. Drug discovery, 18(6), 463–477. https://doi.org/10.1038/s41573-019-0024-5.